Title
Experimental and Computational Studies on Human Visual Perception of Structure from Motion and Natural Scenes

Author
Yang, Xiaoyang

Publication Date
2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Experimental and Computational Studies on Human Visual Perception of Structure from Motion and Natural Scenes

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Psychology

by

Xiaoyang Yang

2012

ABSTRACT OF THE DSSERTSATION

Experimental and Computational Studies on Human Visual Perception of Structure from Motion and Natural Scenes

By

Xiaoyang Yang

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2012

Professor Zili Liu, Chair

In this dissertation, we used computational models to answer two questions about human perception. First, what is the underlying computational mechanism of the stereokinetic effect in human structure from motion perception? Second, what is the functional nature of the boundary extension effects in human natural scene perception?

To answer the first question, we extended the motion coherence theory in two-dimensional (2D) space (Yuille and Grzywacz, 1988) and the minimal total motion theory in 3D space (Rokers, Yuille, and Liu, 2006). We framed the underlying computational mechanism as an optimization problem. We proposed that among all the 2D and 3D structure interpretations, the one that gives

rise to the minimal and spatially smoothest motion is preferred by the visual system, and such structure interpretation is perceived by human observers. We also found that it is important to take into account the higher order motion spatial smoothness. The computational model we proposed was able to predict human 2D and 3D structure from motion perception in varies scenarios. We concluded that the perceptual ambiguity of structure and related motion can be resolved using the minimal total motion and spatially smooth motion principle alone, and any additional assumptions are not necessary.

To answer the second question, we designed two visual memory experiments that made use of a modified test procedure that allowed us to explore boundary extension in terms of signal detection theory. We asked questions about the perceived viewing distance change between study and test scenes from two different psychological dimensions: in terms of close or wide or in terms of change or no change. We found that a criterion bias could explain the boundary extension effects when we asked the perceived viewing distance change in terms of close or wide. In contrast, both discrimination sensitivity and bias contributed to the boundary extension effects when we asked the perceived viewing distance change in terms of change or no change. Remarkably, these results could be explained in a straightforward manner by the multisource model (Intraub, 2010; 2012), with a simple assumption that the view-angle of a memorized scene widened.

The dissertation of Xiaoyang Yang is approved.

Dario Ringach

Bosco Tjan

Alan Yuille

Zili Liu, Committee Chair

University of California, Los Angeles

2012

To my parents, my wife Yunzhe Zhao, and my parents-in-law.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I want to express my earnest gratitude to my advisor, Zili Liu. No words can adequately convey my gratefulness to him for his constant guidance, encouragement, patience, and support. I truly appreciate him for giving me this precious opportunity to work with him and learn from him. I learnt so much from his selfless sharing of his knowledge and wisdom.

I owe my heartfelt gratitude to my PhD committee members, Dario Ringach, Bosco Tjan, and Alan Yuille, for their generous support during my graduate study. I want to sincerely thank all of my committee members for their inspiring discussions and valuable suggestions.

Chapter 2 was written with Zili Liu. We would like to thank Alan Yuille for his valuable discussion and suggestions. Chapter 5 was written with Helene Intraub and Zili Liu. We would like to thank Hongjing Lu for her helpful comments.

I am also grateful to the Psychology Department at UCLA and all my good friends and colleagues at UCLA for giving me an incredible graduate school experience.

Finally, I would like to extend my thanks to my parents, my wife, and my parents-in-law. Their unwavering love and faith is what has shaped me to be the person I am today. I would especially like to thank my wife Yunzhe, who makes my life complete in every sense. I feel so blessed to have her by my side.

VITA

2005-06	Research Assistant Hefei National Laboratory for Physical Sciences at the Microscale Hefei, Anhui, China
2006-07	Research Assistant Vision Research Laboratory, University of Science and Technology of China Hefei, Anhui, China
2007	Bachelor of Science in Physics University of Science and Technology of China Hefei, Anhui, China
2008	Master of Arts in Psychology University of California, Los Angeles Los Angeles, California
2008-09	Teaching Assistant Department of Psychology University of California, Los Angeles
2009-10	Graduate Student Researcher Department of Psychology University of California, Los Angeles
2010	Master of Science in Statistics University of California, Los Angeles Los Angeles, California
2010-12	Teaching Associate Department of Psychology University of California, Los Angeles

CHAPTER 1

INTRODUCTION

The visual system combines the prior knowledge with the sensory information to make an inference about the distal stimulus. Computationally, how the prior knowledge is combined with sensory information remains an open question. In this dissertation, I will develop computational models to explain the stereokinetic effect (Musatti, 1924; Wallach and O'Connell, 1953). I will also use the signal detection theory to investigate the boundary extension effects (Intraub and Richardson, 1989). It is important to investigate human perception from a computational perspective because the computational theory is the most fundamental level of any visual processing (Marr, 1982). In this dissertation, I built computational models that can predict phenomena in human visual perception of structure from motion and natural scenes.

In Chapter 1, I explained why a three-dimensional (3D) structure can be perceived when a 2D shape is rotated in the image plane. We tested the hypothesis that a motion interpretation is preferred if it gives rise to a slower and spatially smoother motion field (Yuille and Grzywacz, 1988; Weiss, Simoncelli, and Adelson, 2002). We used a rotating ellipse as an example since it had been studied with in 2D (Weiss, Simoncelli, and Adelson, 2002) and in 3D (Rokers, Yuille, and Liu, 2006), which were never compared, however. We first replicated the model by Weiss, Simoncelli, and Adelson (2002) and confirmed that the motion field from a 2D deforming ellipse interpretation has a smaller total motion than that from a rigidly rotating ellipse interpretation. We then computed the 2D motion field under the interpretation of a 3D wobbling disk, and

found that the total motion was even smaller than that of a 2D deforming ellipse. Finally, we verified Yuille's proof (Rokers, 2006) which showed that the slowest motion field results from 2D, rather than 3D motion, when there is no spatial smoothness constraint. The resultant motion field is not spatially smooth, and is never perceived. Hence, the necessity of smoothness constraint in motion perception is supported. Our results suggested that the perceptual transition from a 2D deforming ellipse to a 3D wobbling disk can be explained by the slow and smooth constraints alone. Neither the rigidity nor the better gestalt of a circle than an ellipse is needed, since these leaves unexplained the specific perceived motion.

In Chapter 3, we further investigated the stereokinetic effect examined in Chapter 2. Human observers, when presented with a rotating ellipse of large aspect ratio, typically briefly perceive a rotating 2D structure at the beginning, followed by the reliable percept of a 3D wobbling disk. When a rotating ellipse with small aspect ratio is presented however, human observers perceive a 2D ellipse instead. Theoretically speaking, a rotating rigid ellipse in a 2D image plane, no matter if it is fat (large aspect ratio) or narrow (small aspect ratio), can be interpreted as a deforming rotating ellipse tilted in 3D (Notice that an ellipse in the image plane is an ellipse with tilted angle 0 and the tilted wobbling 3D disk is also a special case of tilted ellipse) or even all kinds of non-planar structures tilted in 3D. So it is interesting to examine how the visual system solves this ambiguity in structure and related motion. Another interesting phenomenon is that if the 2D ellipse is rocking instead of rotating, it is more difficult for human observers to reliably perceive a wobbling 3D disk. In addition to that, if there is a dot on the 2D ellipse, the position of the dot has a major effect on the perceived structure. More specifically, a wobbling 3D cone is perceived

if the dot is on the ellipse's minor axis whereas a wobbling tilted 3D disk with a dot sliding on the disk is perceived if the dot is on the major axis. In this chapter, we answered three questions: 1) why a tilted wobbling 3D disk is preferred among all alternative planar structure interpretations; 2) why it is more difficult to perceive a 3D wobbling disk if the ellipse is rocking or if it is narrow; and 3) why the dots at different positions on the 2D ellipse lead to different structure percepts. We developed a computational model that combined the slow and smooth priors in 3D so that the structure ambiguity in 3D can be solved. The computational model we proposed provided answers to all the three questions we asked above using one unique principle, that is, the 3D structure interpretation that gives rise to the slowest and spatially smoothest motion is preferred by the visual system. We also demonstrated that it is necessary to take into account the higher order motion spatial smoothness.

In Chapter 4, we investigated another type of stereokinetic stimulus consisted of two objects. Specifically, we measured the perceived depth of a tilted cylinder percept generated from two rotating circles. We designed two perceptual tasks and one visuomotor task to quantify this stereokinetic effect, and the measurements from all three tasks had qualitatively identical characteristics. In addition to quantifying this stereokinetic effect, we also investigated the computational mechanism underlying this phenomenon. We asked the question: Why human observers do not perceive a cylinder of either zero depth, or on the other extreme, of infinite depth, but instead a cylinder of a finite depth? Previous theories mostly focused on using the rigidity prior to explain the stereokinetic effect (Wallach and O'Connell, 1953). However, the rigidity prior cannot explain the phenomenon in this study. To explain the stereokinetic effect

investigated in the current study, we developed a computational model from Yuille and Grzywacz's (1988) motion coherence theory in 2D and Rokers, Yuille, and Liu's (2006) minimal motion principle in 3D. We framed the computational question as an optimization problem, and we hypothesized that the 3D structure interpretation that has the minimal and spatially smoothest motion in 3D is preferred by the visual system. And we demonstrated that such preferred structure interpretation will give rise to a finite depth of the perceived cylinder. The computational model's predictions are consistent with the empirical results, indicating that the visual system is taking into account the slowness and the spatially smoothness of the motion field to achieve a unique optimal structure interpretation.

In Chapter 5, we investigated the boundary extension in the signal detection theory framework. After viewing a natural scene, people often remember having seen more of the world than was originally visible, an error referred to as boundary extension. Despite the large number of studies on this phenomenon, performance has never been considered in terms of signal detection theory. We reported two visual memory experiments that made use of a modified test procedure that allowed us to explore boundary extension in terms of signal detection theory. In Exp.1, participants first studied pictures presented at close or wide view-angles. At test the same view was never shown, instead the closer or wider counterparts of the studied pictures were presented and participants rated each on a six-point scale to indicate how much closer or wider the view appeared to be. In Exp.2, at test, either the alternate view (as in Exp. 1) or the identical view was presented. Participants rated whether a test image was exactly the same as or different from the studied in view-angle. We found that a criterion bias could explain the boundary extension

effects in Exp.1, whereas in Exp.2 both discrimination sensitivity and bias contributed to the boundary extension effects. Remarkably, these results could be explained in a straightforward manner by the multisource model proposed by Intraub (2010, 2012), with a simple assumption that the view-angle of a memorized scene widened.

To summarize, the work in this dissertation shed light on the computational theory for human visual perception of structure from motion and natural scenes. We showed that computationally the visual processing can be framed as an optimization problem. The work in this dissertation supported the idea that the visual percept is an optimal solution achieved by the visual system's combining of the prior knowledge with the sensory input information.

References

Intraub, H., & Richardson, M. (1989). Wide-angle memories of close-up scenes. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 15, 179-187.

Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information*. Freeman, New York.

Musatti, C. (1924). Sui fenomeni stereocinetici. *Archivio Italiano di Psicologia*. 3, 105-120.

Rokers, B., Yuille, A. & Liu, Z. (2006). The perceived motion of a stereokinetic stimulus. *Vision Research*. 46, 2375-2387.

Rokers, B. (2006). The role of prior knowledge in perception, learning, and recognition. *PhD Thesis, University of California Los Angeles*.

Weiss, Y., Simoncelli, E. & Adelson, E. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*. 5(6), 598-604.

Yuille, A. & Grzywacz, N. (1988). A computational theory for the perception of coherent visual motion. *Nature*. 333, 71 – 74.

Wallach, H. & O'Connell, D. (1953). The kinetic depth effect. *Journal of Experimental Psychology.* 45, 205-217.

Intraub, H. (2010). Rethinking Scene Perception: A Multisource Model. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 52, pp. 231-264). Burlington: Academic Press.

Intraub, H. (2012). Rethinking visual scene perception. *Wiley Interdisciplinary Reviews: Cognitive Science. 3*(1), 117-127.

CHAPTER 2

A COMPUTATIONAL EXPLANATIN OF THE STEREOKINETIC EFFECT

2.1. Introduction

The visual system has the capability of perceiving a three-dimensional (3D) structure from a 2D stimulus. One example is the Musatti effect, or the stereokinetic effect (Musatti, 1924, 1975; Wallach and O'Connell, 1953), that is, a rotating ellipse is perceived as a "true circle gyrating in three-dimensional space" (Duncan, 1975). Duncan (1975) further suggested that in order that a disk in 3D is perceived, the ellipse has to be fat (with aspect ratio larger than .49). To describe the phenomenon in more details, when presented with a rotating fat ellipse, human observer first perceives a 2D deforming ellipse, then reliably perceive a 3D wobbling circular disk. Once the percept of a 3D disk is achieved, it becomes nearly impossible for human observers to perceive the original 2D deforming ellipse. On the other hand, if the rotating ellipse is narrow (with small aspect ratio), the stimulus remains to be perceived as a 2D rigid ellipse, rather than a 2D deforming ellipse or a 3D wobbling disk. In this chapter, we aimed to explain these phenomena with a single principle, and we proposed a computational theory that can predict the empirical observations.

Inferring an object's structure and motion from a 2D image sequence is a difficult problem. First of all, there are infinite 3D structures and related motion that can give rise to the identical 2D

projection. The possible 3D structures include an infinite number of ellipses, tilted in 3D space and rotating around the line of sight, generating the same movie. This ambiguity is stemmed from the information loss due to the image projection from 3D space to 2D plane. When object motion is involved, different structure interpretations will give rise to different motion fields, and consequently the property of the motion field provided the visual system with additional cues about the object structure. In this chapter we investigated if the motion field property alone is sufficient to provide a unique solution to the object structure inference problem.

As a matter of fact, making inference about object structure from motion is still a difficult problem. Even if the structure and motion are restricted to a 2D plane, there are still multiple possible interpretations. As discussed above, a rotating 2D ellipse can be perceived as either a rigid ellipse or a deforming one. This ambiguity comes from the intertwined nature of motion and structure perception. More specifically, the inferred structure determines the motion correspondence, and on the other hand the inferred motion determines the inferred object structure. Weiss, Simoncelli, and Adelson (Weiss, 1998; Weiss, Simoncelli, and Adelson, 2002) approached this puzzle by proposing a "slow and smooth" principle developed from the motion coherence theory (Yuille and Grzywacz, 1988). That is, an object structure interpretation that gives rise to a slowest and spatially smoothest motion flow is favored by the visual system. Note that the classic aperture problem illusion can be reinterpreted as the visual system favors the slowest possible motion. Intuitively, in the ellipse example we discussed at the beginning of the chapter, since local curvature of a fat ellipse is similar from one location to it neighbors, there is much ambiguity to establish motion correspondence (Ullman, 1979) compared to the scenario

when a narrow ellipse is involved. Hence, according to "slow and smooth", the motion correspondence is established by taking into account the motion slowness motion and spatially smoothness. Weiss and his colleagues rephrased the motion coherence theory under the framework of probability distribution and Bayesian theory. They assumed that local motion measurements have independent Gaussian noise with standard deviation σ_l and the slow prior is $N(0,\sigma_p{}^2)$. The maximum a posteriori (MAP) estimation of the 2D motion is determined by the spatial-temporal derivatives of the image sequence as well as σ_l / σ_p:

$$\begin{pmatrix} v_x \\ v_y \end{pmatrix} = -\begin{pmatrix} \sum I_x{}^2 + \sigma_l{}^2/\sigma_p{}^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y{}^2 + \sigma_l{}^2/\sigma_p{}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum I_x I_t \\ \sum I_y I_t \end{pmatrix} \qquad (2.1.)$$

where I_x, I_y and I_t are the spatial and temporal derivatives of the image sequence and the summations are taken over all locations translate together. In the mathematical sense, the slow and smooth prior is independent of the rigidity assumption. Weiss and colleagues were able to predict a wide range of motion perception related phenomena in 2D using the "slow and smooth" principle.

Rokers, Yuille, and Liu (2006) were perhaps the first to generalize this "slow and smooth" principle from 2D to 3D. They studied a similar rotating ellipse stimulus consisted of dashed lines, and the stimulus was perceived to be a 3D wobbling disk. Theoretically, there are two motion components on the wobbling disk: the rotation perpendicular to the image plane and the

spinning on the disk plane (Figure 2.2.). Rokers *et al*. studied the perceived spinning component experimentally and computationally. Specifically, they asked the participants to adjust the spinning component on the disk so that no spinning was visible. In the meanwhile, they also theoretically derived the optimal spinning so that a minimum motion is achieved. Remarkably, the theoretical prediction of the spinning that gave rise to slowest motion quantitatively matched the empirical data obtained in Rokers *et al*. (2006). This was evidence for the first time showing that humans used the "slow and smooth" principle in 3D as well.

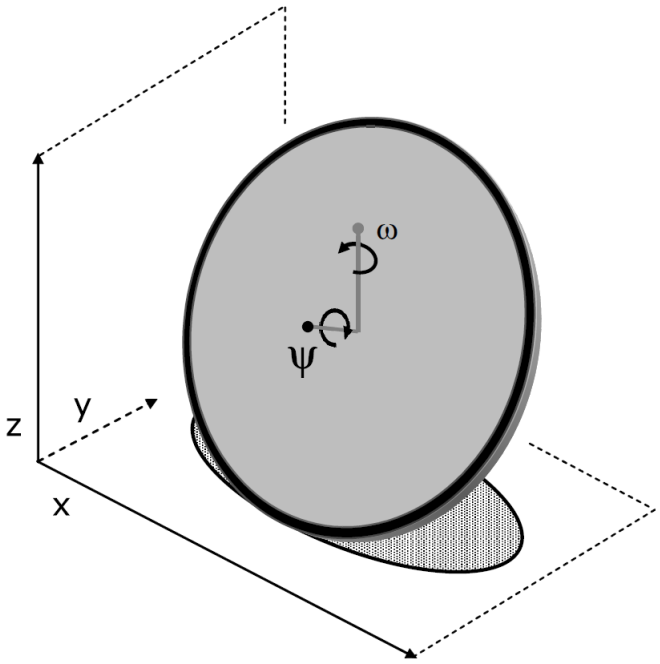

Figure 2.1. Schematic illustration of the percept generated by a rotating ellipse. The stimulus is an ellipse on the x-y plane, and the ellipse it is rotating around its center at angular speed ω. The observer is looking down at the stimulus from above (the observer's line of sight is parallel to the

z axis). The generated percept is a tilted 3D disk, which is not only rotating along z axis, but also spinning around its surface at angular speed ψ. Figure adapted from Rokers, *et al*. (2006).

Rokers *et al.* (2006) investigated the minimal total motion principle in 3D on a stereokinetic stimulus, and demonstrated that among all the 3D disk interpretations, the one with minimum total motion is preferred by the visual system. However, it remains an open question why human observers perceived a 3D wobbling disk rather than a 2D ellipse when a fat ellipse is presented, and why a 2D ellipse is perceived when a narrow ellipse is presented. In order to answer this question from the perspective of minimal total motion principle, we need to know which structure interpretation gives rise to a minimal total motion: the 3D wobbling circular disk, the 2D rigid ellipse, or the 2D deforming ellipse. We hypothesized that the visual system preferred the structure interpretation with minimal total motion, and consequently such structure is perceived by human observers. Specifically, the 3D disk interpretation has the minimal total motion when the ellipse is fat, so that a tilted 3D wobbling circular disk is perceived upon a fat ellipse stimulus. In contrast, the 2D ellipse interpretation has the minimum total motion when the ellipse is narrow, and as a result a 2D ellipse is perceived upon a narrow ellipse stimulus. Besides, the aforementioned total motion calculated by Rokers *et al.* (2006) assumed rigid motion, which is ensured by the circular disk interpretation. However, when 2D interpretations are also taken into account, the rigidity is no longer guaranteed, and as a result the resultant motion field is not necessarily spatially smooth. In this chapter, in addition to the discussion about minimal motion principle in 3D, we will also investigate the importance of spatially smooth motion field.

2.2. Methods

We used as an example an ellipse with an angular speed *1 deg/s* and with standard notations. The total motion is defined as:

$$F = \int (v_x{}^2 + v_y{}^2)\rho(s)ds \qquad (2.2.)$$

Here $s = s(x, y)$ is the ellipse parametric equation in this example. The identical rotation of the 2D ellipse can be interpreted as: (1) a 2D non-rigid rotating ellipse; (2) a 2D rigid rotating ellipse; and (3) the 2D projection of a 3D rigid rotating circular disk. We want to compute the total motions under different interpretations in a fair way so that we can make a comparison.

2.2.1. Compute 2D non-rigid ellipse slow and smooth optical flow

We put a 64×64 equally spaced grid on each image frame, and these grid points are the centers of 64×64 windows covering the entire image frame without overlapping. We assume that within each window motion is translational. So Equation (2.1.) can be used to compute the local MAP motion under slow prior.

In order to implement the smooth constraint, we assume that the interaction (smoothness) is Gaussian with standard deviation σ_i. Yuille and Grzywacy (1989) proposed that the in order that

the motions computed locally are representative of the overall probability distribution of the motion flow, the density of the local motion measurements need to be large enough relative to the spatial scale of the interaction σ_i. More specifically, the local motion are representative if $\rho\pi\sigma_i^2 >> 1$. In our case, we selected the standard deviation of the interaction to be three times of the window size, so that $\rho\pi\sigma_i^2 = 28.3$.

The only free parameter left for the MAP solution is σ_l / σ_p, which is the ratio between the spatial scale of the local measurement noise and the slow prior. To have a better sense of what that free parameter means, assume that the distribution of the prior is fixed, the more accurate the local measurements (i.e., the smaller σ_l) the closer the MAP motion to the center of the local measurement in the velocity space, and an example is the contrast-induced biases in speed perception (Figure 2.2.(a) and 2.2.(b)). On the other hand, assume that the spatial scale of the local measurement noise is fixed, the stronger the slower prior (i.e., the smaller σ_p) the closer the MAP motion to the center of slow prior, that is, slower motion estimation (Figure 2.2.(b) and 2.2.(c)). When both the local measurement noise and slower prior are taken into account, it is their ratio that affects the MAP estimate of the motion.

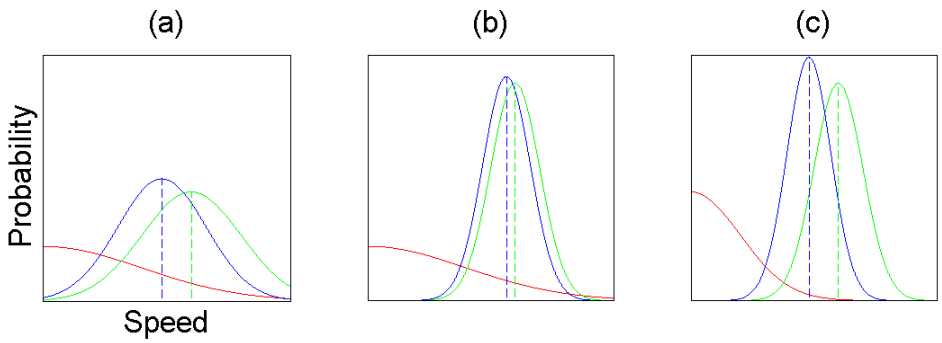

Figure 2.2. The MAP estimate of motion depends on the ratio between the spatial scale of the local measurement noise and the slow prior (σ_l / σ_p). In (a), (b), and (c), the Gaussian distribution in red depicts the slow prior (leftmost) and the Gaussian distribution in green is the local measurement likelihood (rightmost). The curve in blue is the posterior distribution (middle). The prior distribution centers at zero. The dashed lines indicate the centers of the other two distributions. Specifically, the center of the likelihood distribution is the physical speed of the stimulus, whereas the center of the posterior distribution is the MAP estimate of perceived speed, which is slower than the physical one. (a) A weak prior (large σ_p) and an inaccurate local measurement (large σ_l); (b) A weak prior (large σ_p) and an accurate local measurement (small σ_l); (c) A strong prior (small σ_p) and an accurate local measurement (small σ_l).

A number of empirical studies have been done to determine the proper range of this free parameter. Hurlimann *et al.* (2002) found that empirically $N(0,\sigma_p{}^2)$ is flat and σ_l / σ_p is at the level of 10^{-3}, though individual difference exists. Stoker and Simoncelli (2006) used a two alternative force choice (2AFC) protocol to test one dimensional human motion perception and

used the empirical data to fit the shape of the prior distribution as well as the local measurement likelihood function. They found that for high contrast stimulus moving at the speed level of 1 deg s^{-1} the standard deviation of the local measurement noise (σ_l) is around 0.2. On the other hand, the prior probability density at the speed level of 1 deg s^{-1} for such stimulus is at level of 10^{-2}, which means if the prior distribution is assumed to follow $N(0,\sigma_p{}^2)$ then σ_p is at the level of 40, and as a result σ_l / σ_p is at the level of 10^{-3}. Everything else being the same, the motion flow increases as σ_l / σ_p decreases, and we selected σ_l / σ_p = 0.01 to be conservative, i.e., if a smaller σ_l / σ_p is chosen, the total motion for the slow and smooth estimate will be larger than that obtained in our simulation.

In our example, we used image sequence of 128×128 pixels $\times 5$ frames in which an ellipse with major semi-axis 43 pixels and aspect ratio of either 0.8 or 0.2 is rotating counter-clockwise at angular speed of 1 deg s^{-1}. The width of the ellipse is 2 pixels on average and the contrast is 100% (the back ground with lowest pixel value and the ellipse with highest).

2.2.2. Compute 2D rigid ellipse optical flow

Here we briefly derive the motion flow under the interpretation of a rigid rotating 2D ellipse. The ellipse in this example is:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a\cos\theta \\ b\sin\theta \end{pmatrix} \tag{2.3.}$$

For the rigid rotation at angular speed ω, the only motion component on the elliptical contour is the rotation, which is:

$$\begin{pmatrix} v_x \\ v_y \end{pmatrix} = \begin{pmatrix} -a\omega\cos\theta\sin\omega t - b\omega\sin\theta\cos\omega t \\ a\omega\cos\theta\cos\omega t - b\omega\sin\theta\sin\omega t \end{pmatrix} \tag{2.4.}$$

Following the definition in Equation (2.4.), the motion flow under the interpretation of a rigid rotating ellipse is:

$$F_{Rigid\_Ellipse} = \int_0^{2\pi} \omega^2 (a^2\cos^2\theta + b^2\sin^2\theta)\rho(\theta)d\theta \tag{2.5.}$$

where $\rho^2(\theta) = \left(\frac{ds}{d\theta}\right)^2 = a^2\sin^2\theta + b^2\cos^2\theta$.

In all the numerical integrations here and after, we consider in a 2D Euclidean space with the origin at the center of the image. We put exactly the same 64 × 64 equally spaced grid as we used in the slow and smooth simulation on the center image frame, and these grid points are the centers of 64 × 64 windows covering the entire image frame without overlapping. If a window covers part of the ellipse, then the average horizontal coordinate and the average vertical

coordinate of all the pixels belonging to the ellipse and covered by this window is computed. The total motion integration element is computed at the obtained average coordinates using the total motion solution we got as the local movement measurement. The smoothing (interaction) is defined exactly the same as that was used in the slow and smooth simulation, namely, Gaussian distributions with the standard deviation (σ_i) to be three times of the window size. In a word, using slow and smooth motion flow simulation method as a benchmark, all other motion flows were calculated accordingly and identically, except that the appropriate local motion vectors under different interpretations were used instead.

It is noteworthy that when no rigidity constraint is applied, a motion flow field with even smaller total motion (Eq. 2.6.) can be obtained (derived by Yuille. Rokers, 2006):

$$F_{\min} = \int_0^{2\pi} \left(\left(-\omega b \sin\theta + \frac{a^2 b \omega \sin\theta}{\rho^2(\theta)} \right)^2 + \left(\omega a \cos\theta - \frac{a b^2 \omega \cos\theta}{\rho^2(\theta)} \right)^2 \right) \rho(\theta) d\theta \qquad (2.6.)$$

This minimal motion is identical to the motion flow acquired by greedily searching for the nearest neighbor in the second frame for every point on the contour in the first frame. However, this motion field that gives rise to the minimal total motion is not spatially smooth (Figure 2.4(b)), and is never perceived by human subjects. As a result it is sensible to conclude that the human visual system not only take into account the minimal motion prior, but also prefers the spatially smooth motion field.

2.2.3. Compute 3D rigid circular disk optical flow

In the 3D scenario, the circular disk is rotating around the image plane normal as well as spinning around its surface, so the total motion computation should take into account both the rotation and spinning. Assume that the circular disc is spinning at angular speed ψ, and we want to find the minimal total motion with respect to ψ. We denote the spinning at which the total motion reaches its minimum using ψ_m. The 3D motion flow projected onto the 2D image plane, and integrated along the 3D contour. It is the conventional practice that integration along the interpreted (or 3D) contour is used when considering an extremum principle (in this case minimal total motion) for shape from contour (Brady and Yuille, 1984). Under the interpretation of a wobbling 3D disk is (Rokers, Yuille, and Liu, 2006):

$$F_{Rigid\_Disk\_2D\_Projecion} = \int_0^{2\pi} \left({\psi_m}^2 \rho^2(\theta) + 2ab\omega\psi_m + \omega^2 \left(a^2 \cos^2\theta + b^2 \sin^2\theta \right) \right) a d\theta \quad (2.7.)$$

Solving the integral we get:

$$F_{Rigid\_Disk\_2D\_Projecion} = a\pi\left(\psi^2 + \omega^2\right)\left(a^2 + b^2\right) + 4\pi a^2 b\omega\psi \quad (2.8.)$$

We solve for ψ_m by setting $\frac{dF}{d\psi_m} = 0$, and found:

$$\psi_m = -\frac{2ab\omega}{a^2 + b^2} \qquad (2.9.)$$

Notice that for the slow and smooth optic flow computation in section 2.2.1, any local motion vector is computed following only the slow principle, and after that the local motions were smoothed using Gaussian bases. So in order that the optic flow comparison between section 2.2.1 and here is fair, we only consider the slow principle when computing the local motion vectors and the vectors were then smoothed using exactly the same method as we used in previous sections. We used exactly the same method as was used for $F_{Rigid\_Ellipse}$ to compute the total motion in the 3D scenario, except that the appropriate local motion vectors under different interpretations were used instead.

2.2.4. Minimal motion principle versus dimensionality preference

In the previous sections we showed that the minimal total motion principle alone can explain the perceptual preference of 3D disk interpretation over 2D ones when a fat ellipse is shown, and 2D rigid ellipse over its counterparts when a narrow ellipse is shown. However, it might be argued that a preference of dimensionality can also give explanation to this example. In this section, we further elucidate that it is the minimal motion rather than the preference of a percept in certain dimensionality that plays significant role. We provide evidence in two steps. First, we show another example in which a 2D percept with minimal motion is preferred by the visual system

over its corresponding 3D counterpart, indicating that it is not the preference of 3D percept that leads to the interpretation of an ambiguous visual stimulus. Second, we further discuss why the visual system does not prefer to go to an even higher dimensionality (4D) in order to find a better interpretation of the visual stimulus discussed in the previous sections.

As we discussed above, a rotating ellipse on the image plane can be interpreted as either a 2D rotating ellipse or a 3D tilted disk. Analogously, a rotating 2D circle can be interpreted as either a 2D rotating disk or a 3D rotating tilted ellipse. Assume that a tilted ellipse standard following standard notations is rotating at speed ω around an axis perpendicular to the image plane and crossing a point O1 on the image plane. At the same time, the ellipse is spinning at speed ψ around an axis orthogonal to the image plane and crossing the center of the ellipse. The distance between two axes is *d*, so the time dependence of the projected circle in the image plane can be parameterized as:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\omega t & -\sin\omega t \\ \sin\omega t & \cos\omega t \end{pmatrix} \begin{pmatrix} b\cos\psi t + d \\ b\sin\psi t \end{pmatrix} \tag{2.10.}$$

So the velocity of the points on the contour can be obtained by differentiation with respect to time *t*:

$$\begin{pmatrix} v_x \\ v_y \end{pmatrix} = \begin{pmatrix} -b(\psi+\omega)\sin(\psi t+\omega t) - d\omega\sin\omega t \\ b(\psi+\omega)\cos(\psi t+\omega t) + d\omega\cos\omega t \end{pmatrix} \tag{2.11.}$$

The total motion is the integral of velocity along the contour. If the integration is along the 2D projected contour in the image plane, any tilted ellipse will have the same integral, and consequently leaves the visual system ambiguity. However, if the integration is along the 3D contour, the integral depends on the shape of the interpretation, and a unique solution favoring minimal total motion is possible. As a result, we consider the integral along the 3D contour. Substitute θ for ψt so the total motion is:

$$F_{Tilted\_Ellipse} = \int_0^{2\pi} \left(b^2 (\psi + \omega)^2 + d^2 \omega^2 + 2bd\omega(\psi + \omega)\cos\theta \right) \rho(\theta) d\theta \qquad (2.12.)$$

When the interpretation is a 2D circle, $\rho(\theta)d\theta = bd\theta$, so:

$$F_{Tilted\_Ellipse} = 2\pi b \left(b^2 (\psi + \omega)^2 + d^2 \omega^2 \right) \qquad (2.13.)$$

We solve for solve for ψ_m by setting $\frac{dF}{d\psi_m} = 0$, and found:

$$\psi_m = -\omega \qquad (2.14.)$$

And the minimal motion is just the rotational component of the projection:

$$F_{Tilted\_Ellipse} = 2\pi b d^2 \omega^2 \qquad (2.15.)$$

On the other hand, if the interpretation is a 3D tilted ellipse, the contour will be longer than that in the 2D circle case and $\rho(\theta)d\theta$ will be larger than $bd\theta$. As a result the total motion is larger than what we have obtained under the 2D circle interpretation. Since the 2D circle interpretation has the minimal total motion among all the possible 2D and 3D interpretations, it is favored by the visual system. Here we see another example in which a 2D interpretation with minimal motion is favored over any 3D counterparts. To summarize, an interpretation with the minimal total motion is favored by the visual system, no matter it is a 2D or 3D.

If we consider an even higher dimensionality (4D), there are several reasons why the visual system does not seek for a 4D interpretation over the lower dimensional ones. First, analogous to the discussion of a 2D disk interpretation versus a 3D tilted ellipse example, having the same projected motion on the 2D image plane, total motion integrated along the 4D contour which is longer will have at least the same magnitude as that obtained along the 3D contour, and consequently the 4D (and even higher dimensional) interpretation is not favored based on minimal total motion principle. Second, it is well accepted that the visual system is not likely to adopt an interpretation assuming an accidental view point of a visual stimulus and statistically it is almost zero probability that even the simplest symmetric 4D object, namely, a 4D sphere (3-sphere) will give rise to a projection of a disk from a specific view point (for the projection and visualization of a 3-sphere, refer to Peter, 1979). And in general a 4D object needs an accidental and rarely happened view point in order that a disk projection is obtained, and thus a 4D interpretation is not preferred by the visual system based on the non-accidental view principle.

To summarize, both principles leads to the conclusion that going to higher dimensionality beyond 3D will not provide a better interpretation, and it is consistent with empirical evidence.

2.3. Results

As discussed in the previous section, we used as an example an ellipse rotating in the image plane around its center with an angular speed 1 deg s^{-1}. The total motion is defined as the sum of each velocity components (x and y) squared at all positions in the image where motion measurement is available. We discuss two types of ellipses, specifically, a fat one with aspect ratio 0.8 and a narrow one with aspect ratio 0.2.

The identical rotation of the 2D ellipse can be interpreted as: (1) a 2D non-rigid rotating ellipse; (2) a 2D rigid rotating ellipse; and (3) the 2D projection of a 3D rigid rotating circular disk. Empirical study showed that when a fat ellipse is shown, human observers perceive shortly a deforming 2D non-rigid ellipse, followed by the stable and similar percept of a 3D rotating rigid circular disk tilted to the image plane (Rokers, Yuille, and Liu 2006). However, when a narrow ellipse is shown, human observers tend to perceive rigid rotating 2D ellipse (Weiss, 1998). We hypothesized that given this ambiguous 2D stimulus the visual system's preference of the stable percepts (rigid 3D disk in fat ellipse case and rigid 2D ellipse in the narrow case) can be explained by the minimal total motion principle alone. In order to test our hypothesis, we want to

compute the total motions under different interpretations in a fair way so that we can make a comparison.

Mathematically, neither the 2D nor 3D motion can be deterministically inferred based on the 2D stimuli, so additional constraints have to be applied by the visual system to reach a unique motion and structure percept. In the 2D scenario, either the rigidity or the slow and smooth constrain can be applied with sensory input to achieve the Bayesian motion estimate. The rigidity assumption predicts a rotating rigid ellipse percept, which is not perceived by human observer. And the slow and smooth constraint leads to a deforming non-rigid ellipse interpretation which is perceived by human observer for a short period of time. In the 3D scenario, the circular disk is rotating around the image plane normal as well as spinning around its surface, so the total motion is the sum of the rotation and spinning. Although the total motion in 3D disk interpretation can be analytically solved, neither the 2D rigid motion nor 2D slow and smooth motion of an ellipse has analytical solution. We designed a way to compute the total motions under these three interpretations in a fair way so that comparisons among them are possible.

In our example, we used image sequence of 128 × 128 pixels × 5 frames. The fat ellipse has a major semi-axis of 43 pixels and aspect ratio of 0.8 is rotating counter-clockwise at angular speed of 1 deg s^{-1}. The narrow ellipse has the same major semi-axis and rotation, whereas the aspect ratio is 0.2. The width of the ellipse is 2 pixels on average and the contrast is 100% (the back ground with lowest pixel value and the ellipse with highest). We put a 64 × 64 equally spaced grid on each image frame, and these grid points are the centers of 64 × 64 windows. In

the interpretation of a slowly and smoothly rotating 2D ellipse, it is worth noting that the total motion depends on a free parameter σ_l / σ_p which describes the relative emphasis on the slow prior (Figure 2.3.). We used a conservative selection of the free parameter to control the slow regulation (details in Methods).

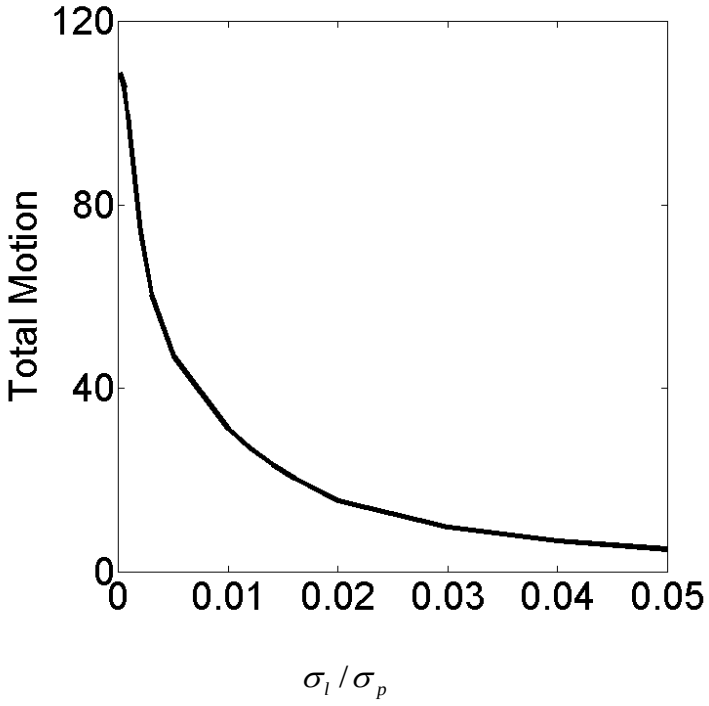

Figure 2.3. The slow and smooth total motion for a rotating ellipse (aspect ratio = 0.8) as a function of the free parameter σ_l / σ_p, where σ_l and σ_p are standard deviations of the slow and smooth prior Gaussian probability distributions, respectively. This calculation was replicating Weiss (1998), where a value in the order of 10^{-3} was typically used. In the current paper, the value was conservatively selected as 0.01.

2.3.1. Fat ellipse with aspect ratio 0.8

The three motion interpretations for a rotating ellipse movie will be: 1) a rotating rigid ellipse, 2) a deforming rotating ellipse (Weiss et al., 2002), and 3) a wobbling 3D circular disk (Rokers et al., 2006). The total motion in 1) and 3) could be analytically derived and their numerical calculations can be arbitrarily precise. However, the total motion in 2) can only be simulated with a free parameter, which is the variance ratio of slow and smooth prior probability distributions that are modeled as Gaussians. We replicated the simulations of Weiss (1998), chose the parameters conservatively so that the total motion flow was even slower than in Weiss (1998) (Fig. 2.3.), and replaced the motion correspondence in order to calculate the motion flows in 1) and 3). We now show the calculation results for the ellipse with an aspect ratio of 0.8 first, and then show results from all aspect ratios.

For the 2D non-rigid slow and smooth scenario, the total motion is 31.2 (unit is the squared velocity, as defined in Equation (2.2.)). And such structure interpretation has a spatially smooth motion field (Figure 2.4.) from the conservative free parameter selection.

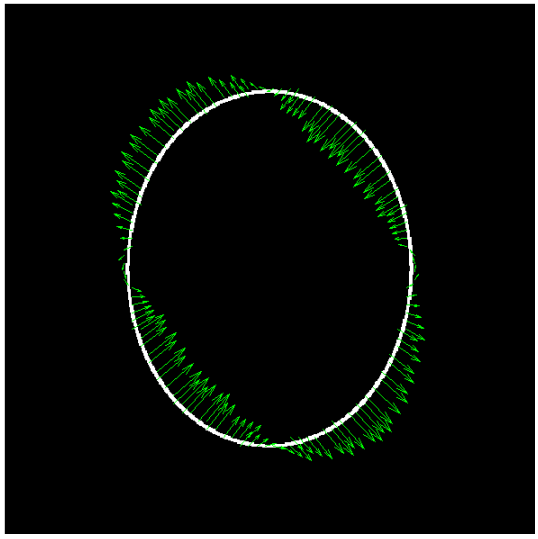

Figure 2.4. The motion field of a rotating ellipse under slow and smooth prior following the interpretation of a 2D deforming ellipse. The computation replicated Weiss (1998) and Weiss *et al*. (2002). The recovered motion field is spatially smooth.

For the 2D rigid ellipse motion scenario the total motion is 108.8 with a spatially smooth motion field (Figure 2.5(a)). Yuille (2006) proved that a minimal 2D motion can be obtained without the rigidity assumption, and the field is essentially obtained by computing the minimal motion locally at every point where motion information is available. This motion is computationally identical to that from greedily search for the spatially closest correspondence. The total motion from this greedy search is the minimal motion in the 2D scenario (it is even smaller than the motion from the 3D disk interpretation), which is 2.7. However, the motion in this case has a spatially unsmooth motion field (Figure 2.5(b)), and such motion is never perceived by the human observer.

(a) (b)

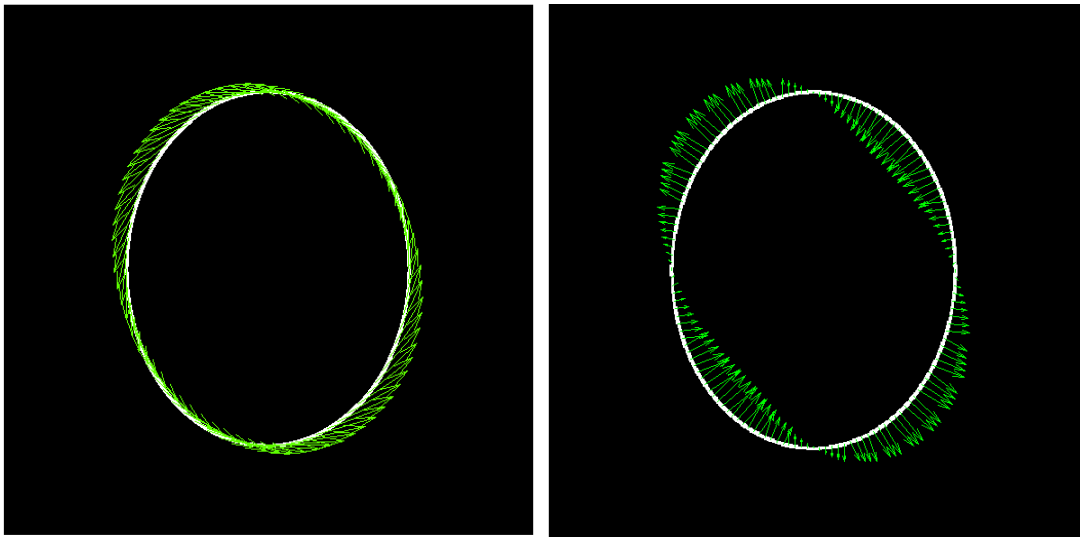

Figure 2.5.The motion field of a rotating ellipse under: (a) rigidity prior and (b) non-rigid greedy nearest neighbor search, following the interpretation of 2D structure. The motion from greedy nearest neighbor search has a spatially non-smooth motion field, and is never perceived by human observer.

In the 3D scenario, the 2D projection of the 3D total motion for a rigid circular disk is smaller for any integration ways. From the comparison of the total motion from different interpretations it is clear that the rigid 3D disk percept give rise to the minimal total motion with smooth motion field (Figure 2.5.). Notice that here we used conservative selection of the free parameter σ_l / σ_p to compute the slow and smooth 2D motion, and the total motion in this case will be even larger if the value of the parameter is set to be close to the previous empirical findings.

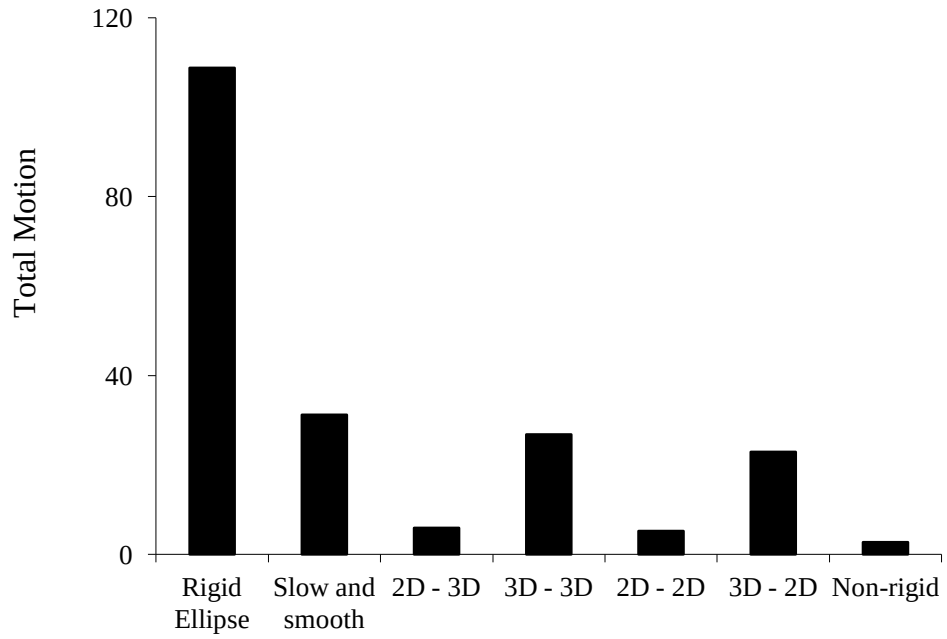

Figure 2.6. The total motion of a rotating ellipse stimulus (aspect ratio = 0.8) calculated numerically and identically in three structure interpretations. The only difference was that the motion vector field was different between the conditions. Human observers reliably perceive a tilted 3D wobbling disk, which is consistent with the model prediction that a 3D disk interpretation gives rise to the minimum total motion among all smooth motion fields.

2.3.2. Narrow ellipse with aspect ratio 0.2

We used exactly the same conservative free parameter selection as in the previous section. For the 2D non-rigid slow and smooth scenario, the total motion is 94.3 and for the 2D rigid ellipse

motion scenario the total motion is 38.3, both with a spatially smooth and qualitatively similar motion field (Figure 2.7.). In the 3D scenario, the 2D projection of the 3D total motion for a rigid circular disk is 66.0, 3D-3D motion was 71.5, 2D-2D motion was 32.2 and 2D-3D motion was 34.3.

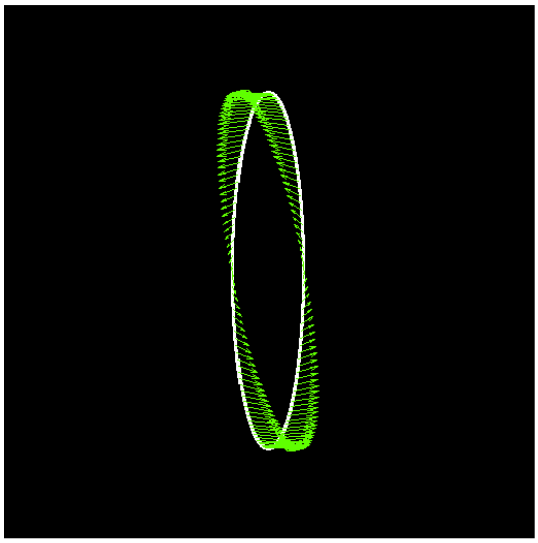

Figure 2.7.The motion field of a rotating narrow ellipse under the rigid 2D interpretation. The motion field is spatially smooth.

From the comparison of the total motion from different interpretations it is clear that the rigid 2D ellipse percept give rise to the minimal total motion with smooth motion field (Figure 2.8.) if the motion of the disk is integrated along the 3D contour.

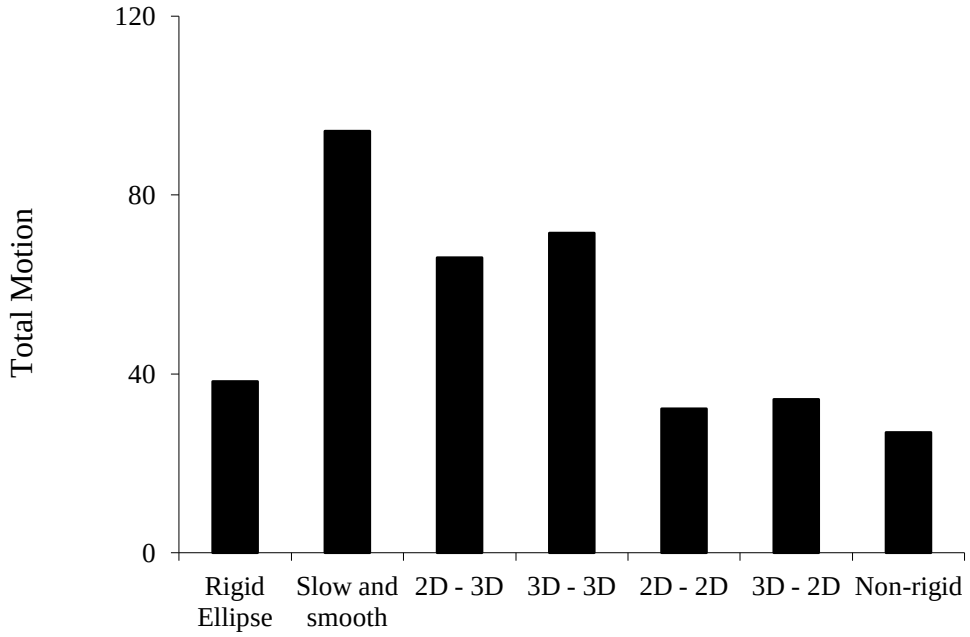

Figure 2.8. The total motion of a rotating ellipse (aspect ratio = 0.2) calculated numerically and identically in three structure interpretations. The only difference was that the motion vector field was different between the conditions. Human observers reliably perceive rigid rotating 2D ellipse, which is consistent with the model prediction that a 2D rigid ellipse interpretation gives rise to the minimum total motion among all smooth motion fields.

2.3.3. Total motion for ellipses with aspect ratio of 0.1 to 0.9

We used exactly the same conservative free parameter selection as in the previous sections. We found that (Figure 2.9.) the total motion under 2D rigid ellipse interpretation monotonically

increases with the increasing aspect ratio. The total motion in 2D non-rigid slow and smooth scenario fluctuates at small aspect ratio, and then decreases as the ellipse gets larger. The total motion in the 3D disk scenario first increases then decreases with the raise of the ellipse aspect ratio. In general, for narrow ellipse with small aspect ratio, the rigid 2D ellipse interpretation leads to the minimal total motion, indicating that a rigid rotating 2D ellipse percept is preferred, which is consistent with human percept. On the other hand, for fat ellipse with large aspect ratio, the rigid 3D disk interpretation has the minimal total motion, so that a tilted rigid rotating disk percept is favored, which is again consistent with human percept. For the ellipses with aspect ratios between the two extremes, the non-rigid slow and smooth interpretation results in the minimal total motion, indicating that it is possible to perceive such percept at certain aspect ratio. Please note that the slow and smooth total motion is achieved with a conservative selection of the free parameter. As a result, the slow and smooth total motion computed with a free parameter suggested by empirical results can be larger than the data we got in our simulation (As an example, refer to Figure 2.3.).

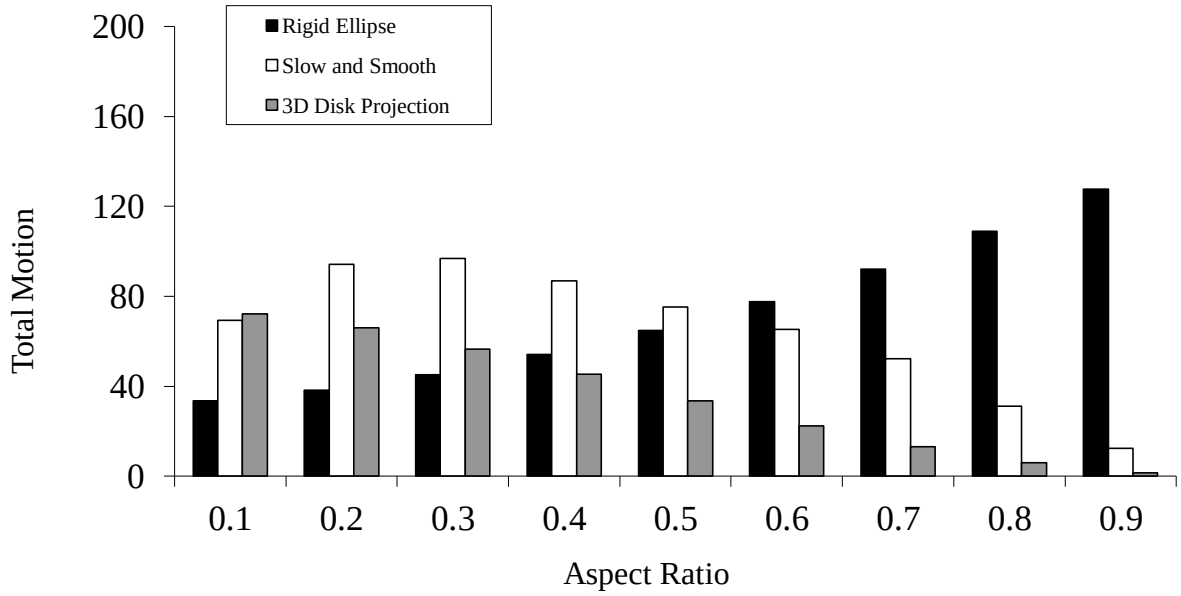

Figure 2.9. The total motion under three different interpretations (rigid 2D, non-rigid 2D, and rigid 3D) for ellipses with aspect ratio from 0.1 (narrow) to 0.9 (fat). At the right end of the graph, the tilted 3D wobbling disk is predicted by the model, which is consistent with the human observer's report that a disk is reliably perceived for fat ellipse. In the left end of the graph, the rigid 2D rotating ellipse is predicted by the model, which is consistent with human percept of a rigid narrow ellipse. Between the two ends, a deforming ellipse is predicted by the model, which is also consistent with human observations.

2.4. Discussion

In this chapter, we investigated the visual system's capability of perceiving a 3D structure from a 2D stimulus. When presented with a rotating fat ellipse, human observers reliably perceive a 3D

wobbling circular disk. However, if the rotating ellipse is narrow, a 2D rigid ellipse, rather than a 2D deforming ellipse is perceived. In this chapter, we aimed to explain these phenomena with a single principle, and we proposed a computational theory that can predict the empirical observations.

Inferring an object's structure and motion from a 2D image sequence is a difficult problem, and this difficulty is stemmed from the information loss due to the image projection from 3D space to 2D plane. Motion field provided the visual system with additional cues about the object structure. But it remains an open question if the motion field property alone is sufficient to provide a unique solution to the object structure inference problem. We found that motion alone can solve the ambiguity of the structure interpretation in 3D.

Yuille and Grzywacz (1988) first proposed the motion coherence theory to solve the structure ambiguity from motion in 2D plane. Weiss, Simoncelli, and Adelson (Weiss, 1998; Weiss, Simoncelli, and Adelson, 2002) rephrased the motion coherence theory into the Bayesian framework, and proposed a "slow and smooth" principle. However, all these work only deal with structure from motion in 2D. Rokers, Yuille, and Liu (2006) generalized this "slow and smooth" principle from 2D to 3D. However, no comparison on motion field across different structure interpretations was performed by Rokers *et al*., so it remains an open question why human observers perceived a 3D wobbling disk rather than a 2D ellipse when a fat ellipse is presented, and why a 2D ellipse is perceived when a narrow ellipse is presented.

We hypothesized that the visual system preferred the structure interpretation with minimal total motion, and consequently such structure is perceived by human observers. In order to test this hypothesis we computed the total motion of the following three structure interpretations in a fair manner: the 3D wobbling circular disk, the 2D rigid ellipse, or the 2D deforming ellipse. We found that the 3D disk interpretation has the minimal total motion when the ellipse is fat, and the 2D rigid ellipse interpretation has the minimum total motion when the ellipse is narrow. It is also noteworthy that if local minimal motion is allowed without taking into account the spatial smoothness of the motion field, an even smaller total motion can be achieved. However, the motion field that gives rise to this smaller total motion is not spatially smooth, and is never perceived by human observers. This indicated that human visual system not only take into account the slowness of the total motion, but also the spatially smoothness.

In summary, we hypothesized that given an ambiguous 2D stimulus, the visual system favors a percept with the minimal total motion and a smooth motion field. Any additional assumptions such as shape compactness, rigidity, or preference for 3D are not necessary in order that a unique percept is reached. We demonstrated this principle using an example of a rotating ellipse. When a rotating 2D fat ellipse is presented, human observers stable percept is a rigid 3D wobbling circular disk tilted to the image plane. Whereas when a rotating 2D narrow ellipse is shown, human observers perceive rigid 2D ellipse, showing that human observers do not simply prefer more symmetric (or compact) shape over less symmetric one. Additionally, when a rotating 2D disk is sown, human observers perceive a rigid 2D disk instead of a tilted 3D ellipse or even higher dimensional contours, revealing that the preference for certain dimensional percept cannot

explain this behavior phenomenon. We demonstrated in this chapter that all these human percepts can be explained by the minimal total motion and smooth motion field principle. We made a fair comparison among the 2D non-rigid, 2D rigid, and 3D rigid motions by developing a way to numerically solve the 2D rigid as well as the 3D rigid motion integrations so that the approach is mathematically comparable to the image processing method used for 2D non-rigid solution. We found that the 3D rigid percept gives rise to minimal smooth total motion in the fat ellipse case, whereas the 2D rigid elliptical rotation generates the minimal smooth total motion in the narrow ellipse case. In the meanwhile, the interpretation that gives rise to a spatially smooth motion field is preferred. These results indicated that the perceptual ambiguity of motion and related structure can be resolved using the minimal total motion and smooth motion field principle alone, and any additional assumption is not necessary.

References

Brady, M. & Yuille, A. (1984). An extremum principle for shape from contour. *IEEE Trans Pattern Anal,* 6(3): 288 – 301.

Duncan, F. (1975). Kinetic art: On my psychokinematic objects. *Leonardo, 8,* 97-101.

Hildreth, E. & Ullman, S (1982). The measurement of visual motion. *MIT A.I.Memo* 699.

Hildreth, E. (1983). The computation of velocity field. *MIT A.I.Memo* 734.

Hildreth, E. (1984). *The measurement of visual motion*. Cambridge, MA: MIT Press.

Horn, B. & Schunck, B. (1981). Determining optical flow. *Artificial Intelligence*. 17, 185 – 203.

Hurlimann, F., Kiper, D. & Carandini, M. (2002). Testing the Bayesian model of perceived speed. *Vision Research*, 42, 2252 – 2257.

Lucas B. & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *Proceedings of Imaging Understanding Workshop*. 121 – 130.

Peter, M. (1979). Dante and the 3-sphere. *American Journal of Physics*. 47(12), 1031 – 1035.

Rokers, B., Yuille, A. & Liu, Z. (2006). The perceived motion of a stereokinetic stimulus. *Vision Research*. 46, 2375-2387.

Rokers, B (2006). The role of prior knowledge in perception, learning, and recognition. *PhD Thesis, University of California Los Angeles*.

Simoncelli, E., Adelson, E. and Heeger, D. (1991). Probability distributions of optical flow. *Proc Conf on Computer Vision and Pattern Recognition*. 310 – 315.

Stocker, A. & Simoncelli, E. (2005). Constraining a Bayesian model of human visual speed perception. *In Advances in Neural Information Processing Systems (NIPS)*, Vol. 17.

Stocker, A. & Simoncelli, E. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4), 578 – 585.

Ullman, S. (1979). The interpretation of structure from motion. *Proceedings of the Royal society of London. Series B, Biological Sciences*, 203(1153), 405 – 426.

Ullman, S. (1984). Maximizing rigidity: The incremental recovery of 3D structure from rigid and non-rigid motion. *Perception*, 13(3), 255 – 274.

Ullman, S. and Yuille, A. (1987). Rigidity and smoothness of motion. AIM-989.

Weiss, Y. (1998). Thesis (Ph.D.)--Massachusetts Institute of Technology, Dept. of Brain and Cognitive Sciences.

Weiss, Y., Simoncelli, E. & Adelson, E. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*. 5(6), 598-604.

Yuille, A. & Grzywacz, N. (1988). A computational theory for the perception of coherent visual motion. *Nature*. 333, 71 – 74.

Yuille, A. & Grzywacz, N. (1989). A mathematical analysis of the motion coherence theory. *International Journal of Computer Vision*, 3, 155 – 75.

Yuille, A. & Bülthoff, H. (1996). Bayesian decision theory and psychophysics. *In Perception as Bayesian Inference*. Knill, D.C. and Richards, W. (Eds.). Cambridge University Press: New York, 123 – 161.

CHAPTER 3

SLOW AND HIGH ORDER MOTION SMOOTHNESS IN 3D

3.1. Introduction

Human visual system has the capability of recovering three-dimensional (3D) structure from 2D sensory input. Because of such capability, human can behave properly in a 3D environment. Accommodation, convergence, binocular disparity, linear perspective projection, familiar size, texture and occlusion are important depth cues for the visual system to make inference about the 3D structure. It is surprising that human visual system is even capable of recovering 3D structure when all of the cues above are not applicable. As a matter of fact, human observers are able to recover 3D structure from impoverished monocularly presented stimulus. For example, human observers perceive a tilted wobbling 3D disk when they are monocularly exposed to a rotating 2D ellipse stimulus. In this example, the only information available for the visual system to make inference about the 3D structure is the motion and shape of the 2D contour. The human visual system recovers the 3D structure so quickly and accurately, so it appears that the problem of making 3D structure inference from 2D sensory input seems to be trivial. Actually, such problem is surprisingly a very complicated puzzle because in theory, there are infinite possible interpretations about the object's 3D structure, each giving rise to a particular motion interpretation.

The inference of the structure and the motion of the stereokinetic object is an ill-posed inverse problem, so additional priors have to be imposed in order that a definitely answer can be found. Since the measurement of object motion is directly relevant to the interpretation of object structure, previous research has been focused on both priors regarding the object shape and the object motion. The interesting question is what assumptions the human visual system deploys that turns a problem with many alternative solutions into one with a unique answer. The rigidity assumption is one candidate that directly provides constraint to the object shape (Ullman, 1979, 1983), namely, an interpretation with maximal rigidity is favored by the visual system. And since the object structure is inherently connected with the perceived object motion, the slow and smooth assumption (Horn and Schunck, 1981; Hildreth and Ullman, 1982; Yuille and Grzywacz, 1988; Weiss, Simoncelli, and Adelson, 2002) is another example of the priors, that is, the structure interpretation in which a slow and smooth motion field can be obtained is favored by the visual system. Specifically, the slow and smoothness prior comes from the insight that discontinuous motion is rare in the world if no object collusion happens. In the mathematical sense, the smoothness constrain penalizes the magnitude of the optical flow gradient as well as the sum of squared of the Laplacians of the optical flow. More generally, smoothness constrain penalizes all orders of derivatives on the optical flow, known as the motion coherence theory (Yuille and Grzywacz, 1988), so the regularization term is:

$$\lambda \int \sum_{m=0}^{\infty} c_m \left| \nabla^m \vec{v} \right|^2 \qquad (3.1.)$$

Here $\vec{v}$ is the estimated optical flow at a certain location, $\lambda \geq 0$ and $c_m \geq 0$ are constants, and ∇^m essentially is the m[th] order gradient along all orthogonal directions (starting at the 0 order). And the integration is over the entire field where local motion information is available. The constant can be set as $c_m = \sigma^{2m}/(m!2^m)$ to obtain a Gaussian interaction. So the only free parameter remains is λ which mediates the relative impact strength on the global motion estimate from the local measurements and the smoothness regularization. For the zero order derivatives, the penalty term is restricting the magnitude of the optical flow so that it was explicitly defined as the slow constrain (Weiss, 1998; Weiss, Simoncelli and Adelson, 2002). And the first order term restricts the gradient of the optical flow and as a result referred to as the smoothing term. So the motion measurement can be formally described as minimizing the following energy function:

$$E(\vec{v}) = \sum_i (\vec{v}(\vec{r}_i) - \vec{v}_l(\vec{r}_i))^2 + \lambda \int \sum_{m=0}^{\infty} c_m \left| \nabla^m \vec{v} \right|^2 \qquad (3.2.)$$

Here $\vec{v}(\vec{r}_i)$ is estimated velocity field and $\vec{v}_l(\vec{r}_i)$ is the local velocity measurement or data. And $c_m = \sigma_l^{2m}/(m!2^m)$, where σ_l is the standard deviation of the local measurement noise assuming Gaussian distribution. Specifically, larger σ_l indicates more noise in the local velocity measurement whereas smaller σ_l means less noise. Notice that the second order regularization vanishes in the following situations: the pure translational motion parallel to surface, the pure translational motion along surface normal and the pure rotation about surface normal. Free parameter λ quantifies the emphasis on slow and smooth regularization relative to local measurement, and σ depicts the weight between slow and smoothness constrains. And in general,

the first term of the equation's right-hand side penalizes incompatibility with the data and the second term emphasized slow and smooth prior. Weiss and Adelson (1998) rephrased the motion coherence theory under the framework of probability distribution and Bayesian theory. They assumed that local motion measurements have independent Gaussian noise with standard deviation σ_l and the slow prior is $N(0,\sigma_p{}^2)$. The MAP of the 2D motion is determined by the spatial-temporal derivatives of the image sequence as well as σ_l / σ_p (Weiss, 1998):

$$\begin{pmatrix} v_x \\ v_y \end{pmatrix} = -\begin{pmatrix} \sum I_x{}^2 + \sigma_l{}^2/\sigma_p{}^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y{}^2 + \sigma_l{}^2/\sigma_p{}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum I_x I_t \\ \sum I_y I_t \end{pmatrix} \qquad (3.3.)$$

where I_x, I_y and I_t are the spatial and temporal derivatives of the image sequence and the summations are taken over all locations translate together. In mathematical sense, the slow and smooth prior is independent of the rigidity assumption, so we want to investigate if the slow and smooth motion prior alone suffice or the rigidity prior is necessary in order that a definitely answer consistent with human percept can be obtained for the motion and structure of a stereokinetic object.

The human visual system possesses the ability of inferring a unique and stable 3D structure of an object from its 2D motion, and the phenomenon that a 3D percept is perceived from a moving 2D stimulus is called the stereokinetic effect (Duchamp, 1920s, more details see Shearer and Gould, 1999; Musatti, 1924; Wallach, 1953). The percept of 3D disk from a rotating ellipse we

described above was referred to as the Musatti effect, that is, a rotating ellipse is perceived as a "true circle gyrating in three-dimensional space" (Duncan, 1975). Duncan (1975) further claimed that in order that a "true circle" in 3D is perceived, the ellipse has to be fat (more specifically, ellipse with eccentricity less than .87, which is equivalent to with aspect ratio larger than .49). Human observers, when presented with a fat rotating ellipse, typically briefly perceive a moving 2D pattern at the beginning, followed by stable percept of a 3D wobbling disk. When a narrow ellipse is presented, human observers perceive a 2D ellipse instead. Theoretically speaking, a rotating rigid ellipse on a 2D image plane, no matter if it is fat or narrow, can be interpreted as a deforming rotating ellipse tilted in 3D (Notice that an ellipse on the image plane is an ellipse with tilted angle 0 and the tilted wobbling disk is also a special case of a ellipse tilted in 3D) or even all kinds of non-planar shapes tilted in 3D. If the visual system only relies on the sensory input from the 2D stimulus, there is no way to achieve one unique solution among all the alternatives. Empirically, on the other hand, human observers appear not to suffer from this ambiguity, and always reach the stable percept of a wobbling disk tilted in 3D when a fat ellipse is present and the stable percept of a 2D ellipse when a narrow ellipse is presented. Another interesting observation is that if the 2D ellipse is rocking instead of rotating, it is more difficult for human observers to reliably perceive a wobbling 3D disk. Besides, when there is a dot on the 2D ellipse, the position of the dot will have a major effect on the perceived structure, more specifically, a wobbling 3D cone is perceived if the dot is on the minor axis whereas a wobbling tilted 3D disk with a dot sliding on it is perceived if the dot is on the major axis. In this chapter, we aimed to answer three questions: 1) why a tilted wobbling 3D disk is preferred among all alternative planar interpretations; 2) why it is more difficult to perceive the 3D disk if the ellipse

is rocking or if the ellipse is narrow; And 3) why the dots at different positions on the 2D ellipse will lead to totally different percepts.

Yuille and Grzywacz's (1988) motion coherence theory and Weiss, Simoncelli, and Adelson's (2002) slow and smooth Bayesian framework tried frame the 2D motion and structure from motion question into an optimization problem, and proposed that among all possible 2D motion interpretations, the one that gives rise to slowest and spatially smoothest motion is preferred by the visual system. Yuille and Grzywacz (1989) further suggested that to deal with motion that not only composed of pure translation and rotation, higher order regularization terms are necessary. However, previous work focused on 2D motion and structure and as a result could not resolve the 3D motion and structure problem. In this chapter, we developed a computational model that that combines the slow and smooth priors in 3D so that the structure from motion problem in 3D can be solved. The computational model we proposed provided answers to all three questions we described above using one unique principle, that is, the 3D interpretation that gives rise to the slowest and spatially smoothest motion is preferred by the visual system. When motion smoothness is concerned, previous usually took into account the zero order (slowness) and the first order (1^{st} order spatial smoothness) motion smoothness regulation term for computational simplicity, although Yuille and Grzywacz (1989) theoretically elaborated the significance of higher order regulation terms. In this chapter, we also demonstrated that it is necessary to take into account the higher order motion spatial smoothness so that the visual system can determine the optimal structure interpretation.

3.2. Question 1: Why a tilted wobbling 3D disk is perceived?

3.2.1. Methods

An ellipse rotating on the image plane can be interpreted as a deforming rotating ellipse tilted in 3D space. Assume that the stimulus is a 2D ellipse on the x-y plane with the semi-major axis *a* on x axis and semi-minor axis *b* on y axis, and the ellipse can be expressed as:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a\cos\alpha \\ b\sin\alpha \\ 0 \end{pmatrix} \tag{3.4.}$$

And the ellipse on the x-y plane can be interpreted as the projection of a tilted 3D ellipse with tilted angle of θ (the angle between the semi-minor axes of the x-y plane ellipse and the tilted ellipse), so that the tilted ellipse can be expressed as:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a\cos\alpha \\ b\sin\alpha \\ b\sin\alpha\tan\theta \end{pmatrix} \tag{3.5.}$$

There is a rotation ω around the z axis, so considering the rotation:

$$\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} \cos\omega t & -\sin\omega t & 0 \\ \sin\omega t & \cos\omega t & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a\cos\alpha \\ b\sin\alpha \\ b\sin\alpha\tan\theta \end{pmatrix} \tag{3.6.}$$

From Equation (3.6.), when only the rigid rotation is considered, the tilted ellipse at time t can be expressed as:

$$\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} a\cos\omega t\cos\alpha - b\sin\omega t\sin\alpha \\ a\sin\omega t\cos\alpha + b\cos\omega t\sin\alpha \\ b\sin\alpha\tan\theta \end{pmatrix} \tag{3.7.}$$

So the rigid motion is simply the temporal derivative of Equation (3.7.), which is:

$$\begin{pmatrix} v_x(t) \\ v_y(t) \\ v_z(t) \end{pmatrix} = \begin{pmatrix} -a\omega\sin\omega t\cos\alpha - b\omega\cos\omega t\sin\alpha \\ a\omega\cos\omega t\cos\alpha - b\omega\sin\omega t\sin\alpha \\ 0 \end{pmatrix} \tag{3.8.}$$

In addition to the rigid rotation, we assume that there is a spin along the contour, and the spin is always in the direction of the contour tangent that cancels part of the rigid rotation:

$$\vec{\psi} = \frac{\phi}{\sqrt{\frac{b^2\sin^2\alpha}{\cos^2\theta} + a^2\cos^2\alpha}} \begin{pmatrix} -\frac{b\sin\alpha}{\cos\theta} \\ a\cos\theta\cos\alpha \\ a\sin\theta\cos\alpha \end{pmatrix} \tag{3.9.}$$

So the motion is the rigid rotation combined with the spin:

$$\begin{pmatrix} v_x(t) \\ v_y(t) \\ v_z(t) \end{pmatrix} = \begin{pmatrix} -a\omega \sin \omega t \cos \alpha - b\omega \cos \omega t \sin \alpha \\ a\omega \cos \omega t \cos \alpha - b\omega \sin \omega t \sin \alpha \\ 0 \end{pmatrix} + \frac{\phi}{\sqrt{\frac{b^2 \sin^2 \alpha}{\cos^2 \theta} + a^2 \cos^2 \alpha}} \begin{pmatrix} -\frac{b \sin \alpha}{\cos \theta} \\ a \cos \theta \cos \alpha \\ a \sin \theta \cos \alpha \end{pmatrix} \quad (3.10.)$$

Recall that the regularization theory proposed that the motion measurement is the motion field that gives rise to a minimum value of the loss function:

$$E(\vec{v}) = \sum_i (\vec{v}(\vec{r}_i) - \vec{v}_l(\vec{r}_i))^2 + \lambda \int \sum_{m=0}^{\infty} c_m \left| \nabla^m \vec{v} \right|^2 \quad (3.11.)$$

Here in our example given the ellipse on the x-y plane, $E(\vec{v})$ is a function of tilted angle θ and spin ϕ. We first calculated the values for each of the regularization terms up to the 3rd order defined in Equation (3.1.) as a function of the spin ϕ for different tilted angles. This calculation will give us a better understanding about the significance of motion smoothness regulation terms at different orders. Then we combined the penalty resulted from each order regularization term, and calculated the lost function. We calculated the loss function value at different tilted angle θ, given the ellipse on the x-y plane, and checked which tilted angle gives rise to the global minimum of the loss function value.

3.2.2. Results

For a given x-y ellipse, we first calculated the values for each of the regularization terms up to the 3rd order defined in Equation (3.1.) as a function of the spin ϕ under different interpretations of tilted angles. We did this so that we can have a good understanding about the significance of the regularization terms on the overall lost function. More specifically, if the regularization term preferred a rigid structure interpretation, then a spin close to zero will give rise to a minimum in the term. In contrast, if a regularization terms focuses more on the slowness of the total motion, then a spin that can maximally cancel the rotational component gives rise to the minimum in this term. We followed the definition in Equation (3.1.), and used in all the following computations σ = 2, which is about the mid-point of the suggested range of the parameter selection (Yuille and Grzywacz 1988, 1989). Figure (3.1.), (3.2.), (3.3.), and (3.4.) showed the lost function value for different order regulation terms under structure interpretations of tilted ellipses at different angles, from smaller tilted angle to larger ones, given an x-y elliptical stimulus with aspect ratio 0.8.

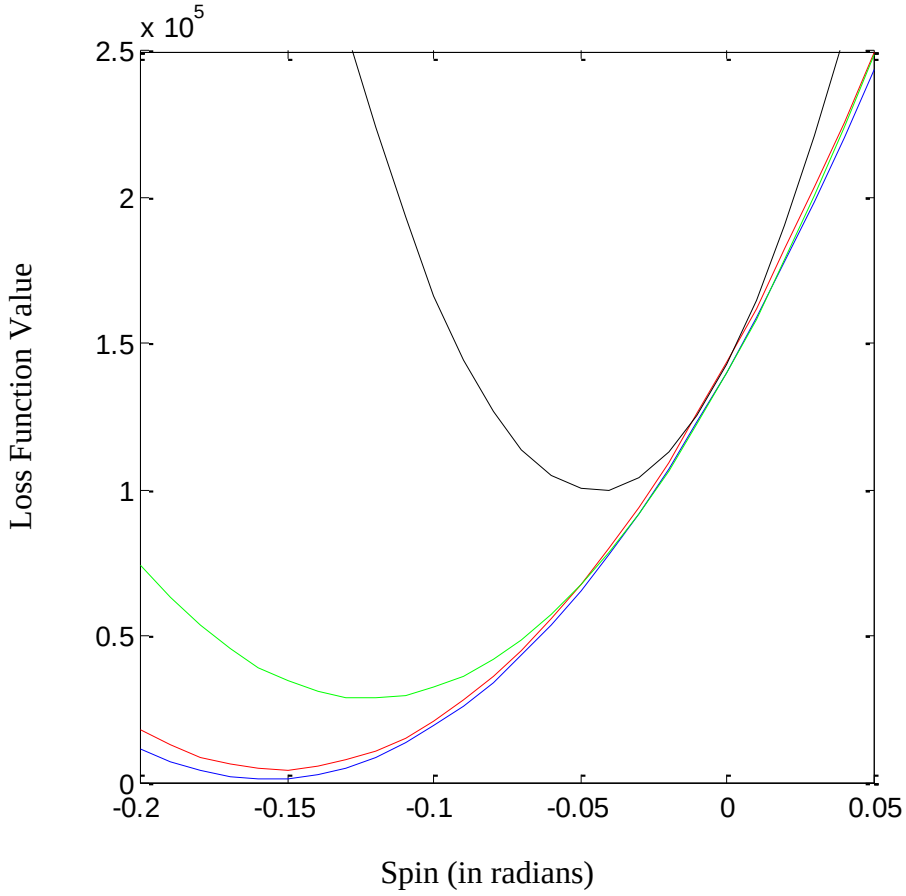

Figure 3.1. The values of 0 to 3 orders of regularization terms assuming that the x-y plane ellipse is projected from an ellipse tilted at angle 0 degree (2D ellipse on the image plane). Blue: 0 order derivation (total motion); Red: 1st order derivative of motion; Green: 2nd order derivative of motion; Black: 3rd order derivative of motion. Lower order regularization terms tend to prefer the interpretation that gives rise to a slower total motion (so that high spinning to cancel the rotation is preferred) whereas higher order regularization terms tend to prefer a rigid structure (so no spinning is preferred).

We can clearly see from Figure (3.1.) that with a certain range of spin θ the 0 order regularization term (blue) and 1st order regularization term (red) can be very small. However, at this point, the motion is highly non-rigid which leads to high penalty from the 2nd and 3rd order regularization terms (especially the 3rd order term). This graph demonstrated the significance of the higher order regularization terms, especially when the motion is not simply consist of translation and rigid rotation. Many previous works emphasized only the first order regularization (for example, Horn and Schunck 1981) or only zero and first order regularizations (for example Weiss *et al.*, 2002). The current study showed that the 3rd order regularization term can play a significant role in the motion estimation, especially when the object motion has non-rigid components. Figure (3.2.) below showed similar pattern when the interpretation is a tilted ellipse with a small tilted angle (θ = 22.9^{o}).

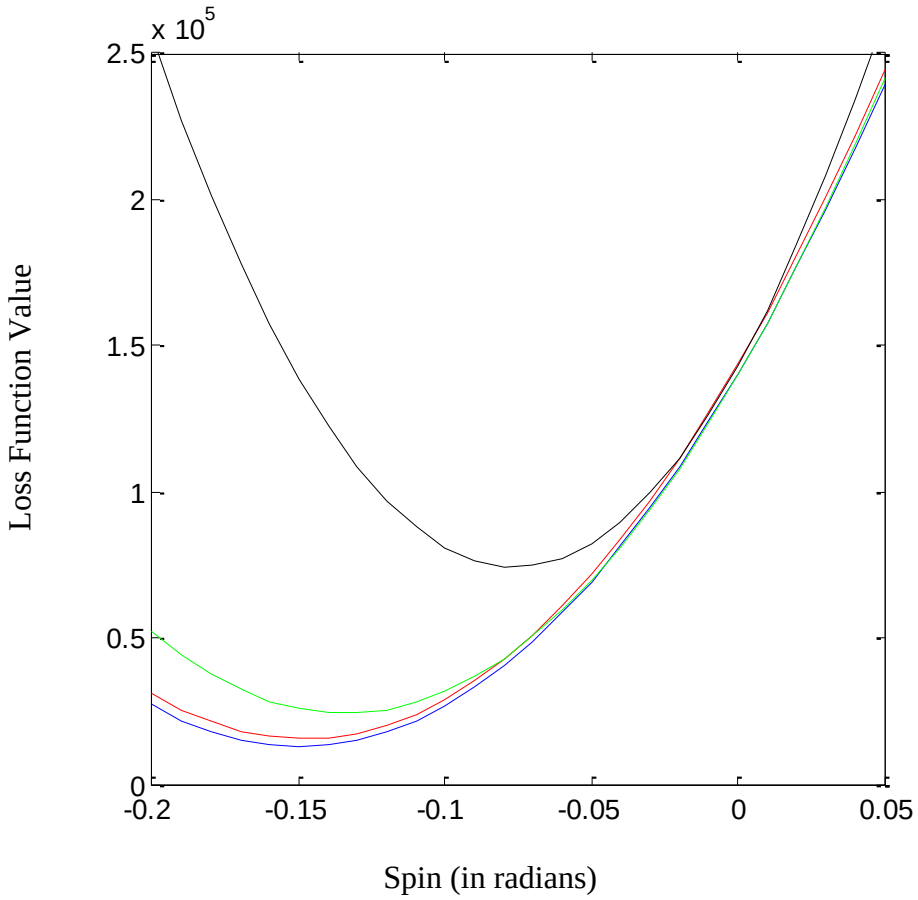

Figure 3.2. The values of 0 to 3 orders of regularization terms assuming that the x-y plane ellipse is projected from an ellipse tilted at angle 22.9 degree. Legends follow Figure (3.1.). Similarly, lower order regularization terms tend to prefer the interpretation that gives rise to a slower total motion (so that high spinning to cancel the rotation is preferred) whereas higher order regularization terms tend to prefer a rigid structure (so no spinning is preferred).

We can see from Figure (3.3.) that when the interpretation is a tilted disk (tilted angle $\theta = 36.9^{o}$), the higher order regularization term no longer put penalty to the spin, more specifically, since under this interpretation, the spin will not introduce non-rigidity, so that the higher order

regularizations will tolerate the spin. More importantly, different orders of regularization terms reach global minimum with the same spin magnitude.

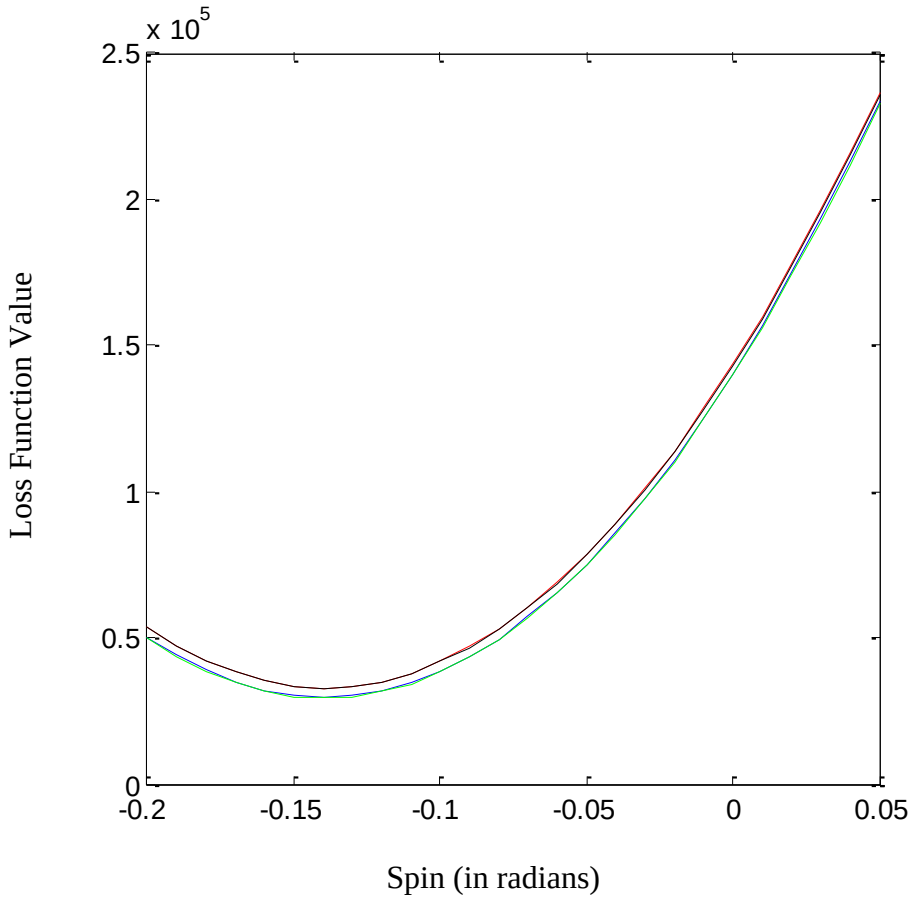

Figure 3.3. The values of 0 to 3 orders of regularization terms assuming that the x-y plane ellipse is projected from an ellipse tilted at angle 36.9 degree (tilted disk). Legends follow Figure (3.1.). This interpretation (a 3D tilted wobbling disk) is special because both the lower order and the higher order regularization terms reach the global minimum at the same optimal spinning.

We can also see that when the interpretation is an ellipse with a large angle (tilted angle θ = 55.4^{o}) as shown in Figure (3.4.), the higher order regularization term again put high penalty to the non-rigidity introduced by the spinning, and favors no spin. In general, we can see that the interpretation of a tilted disk is a very special interpretation where the minimum of "slow" regularization meets the minimum of different orders of "smooth" regularizations.

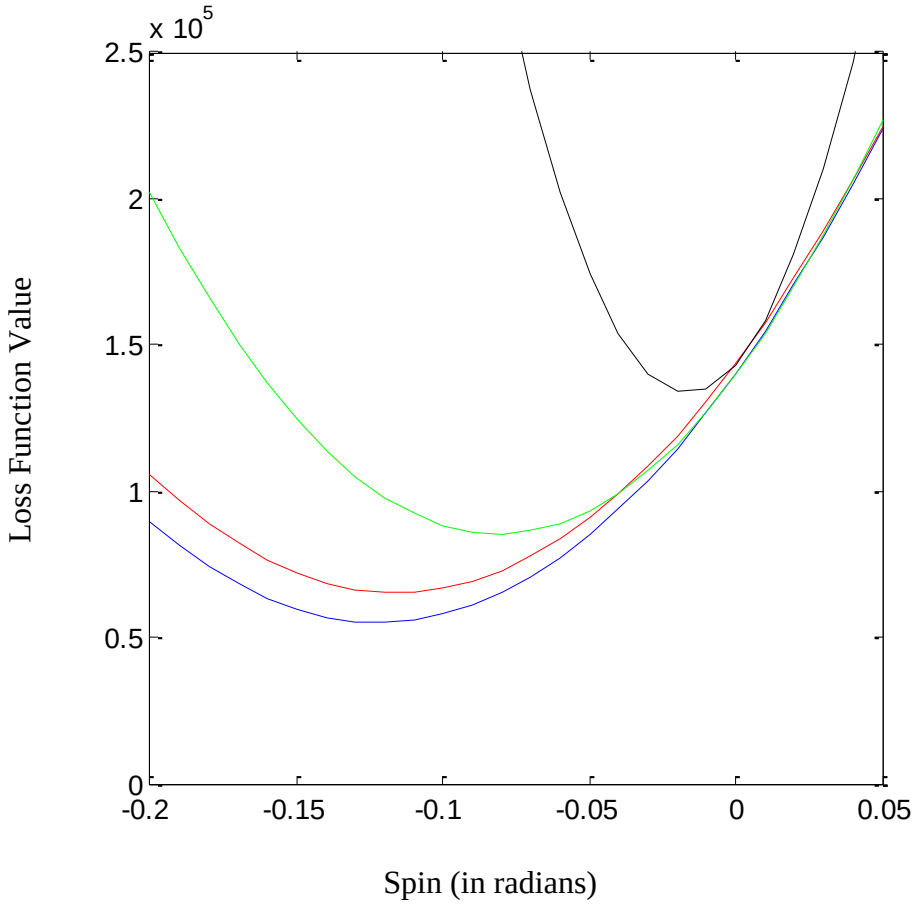

Figure 3.4. The values of 0 to 3 orders of regularization terms assuming that the x-y plane ellipse is projected from an ellipse tilted at angle 54.4 degree. Legends follow Figure (3.1.) Similarly,

lower order regularization terms tend to prefer the interpretation that gives rise to a slower total motion (so that high spinning to cancel the rotation is preferred) whereas higher order regularization terms tend to prefer a rigid structure (so no spinning is preferred).

To summarize, for all elliptical structure interpretations, the lower order regularization terms tend to prefer the interpretation that gives rise to a slower total motion (so that high spinning to cancel the rotation is preferred) whereas higher order regularization terms tend to prefer a rigid structure (so that no spinning is preferred). This might provide an answer to why if only slow and 1^{st} order smooth is taken into account, then a 2D deforming ellipse is preferred for fat ellipse as suggested by Weiss (1998). In general, the higher order motion smoothness regularization term put high penalty on non-rigid structure interpretation. After understanding what regularization term plays a significant role in the motion estimate, we then calculated the combined regularization cost functions value defined by Equation (3.1.). Figure (3.5.) shows the loss function value when the x-y plane image is a circle, the minimum point is at tilted angle $\theta = 0^{o}$, namely the static circle on the x-y plane is the optimal solution from the motion coherence theory in 3D, consistent with human observer's experience.

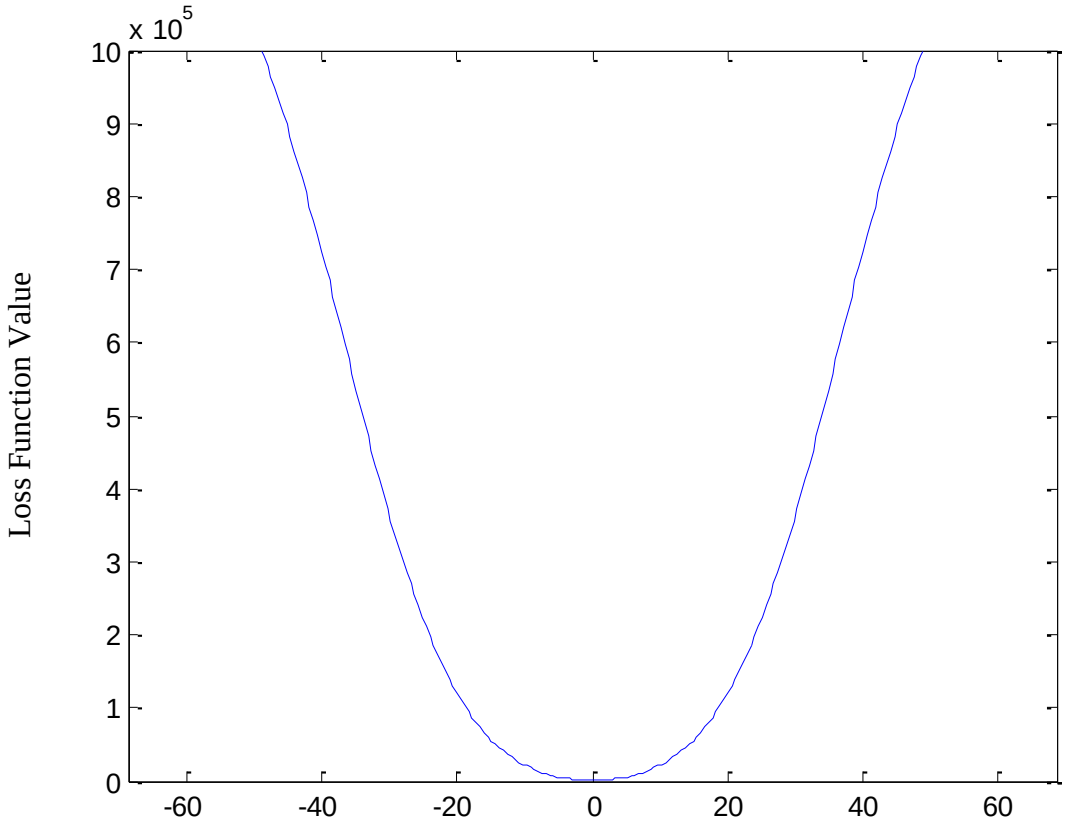

Figure 3.5. The loss function value assuming that the x-y plane circle is projected from an ellipse tilted at different angles. The loss function is a weighted combination of 0 to 4 order regularization terms, and the weights were defined following Grzywacz and Yuille, 1988. The global minimum is located right at tilted angle $\theta = 0^o$, and the loss value increases monotonically as the tilted angel deviates from zero. The model predicts that a static 2D circle is perceived, which is consistent with the human percept.

When the x-y plan image is not a circle, but an ellipse, human observers reliably report perceiving a 3D tilted wobbling disk, after a brief percept of 2D ellipse. Figure 3.6. shows the loss function value when the x-y plane image is an ellipse with aspect ratio 0.8. It is clear that the minimum point is no longer at tilted angle $\theta = 0^o$, instead the minimum points are at $\theta = \pm 33.2^o$. And when the interpretation is a tilted disk, the tilted angle is $\theta = \pm 36.9^o$. It is noteworthy that the theoretical predicted θ giving rise to the minimal loss function value depends on the selection of free parameter σ (in our case we selected $\sigma = 2$), more specifically, smaller σ will put more weights on lower order regularization terms and thus favors the interpretation on the x-y plane, whereas larger σ will but more weights on higher order regularization terms and thus favors the interpretation of a tilted disk. In our case, if a larger σ is selected (value up to 4 is reasonable, more details in Yuille and Grzywacz, 1989) our theoretical predicted interpretation will be even more close to the prediction of a tilted disk. Our theoretical prediction is quantitatively very close to the interpretation of a tilted disk, and it predicts the empirical observation that when a rotating ellipse is presented human observers reliably perceive a tilted wobbling disk.

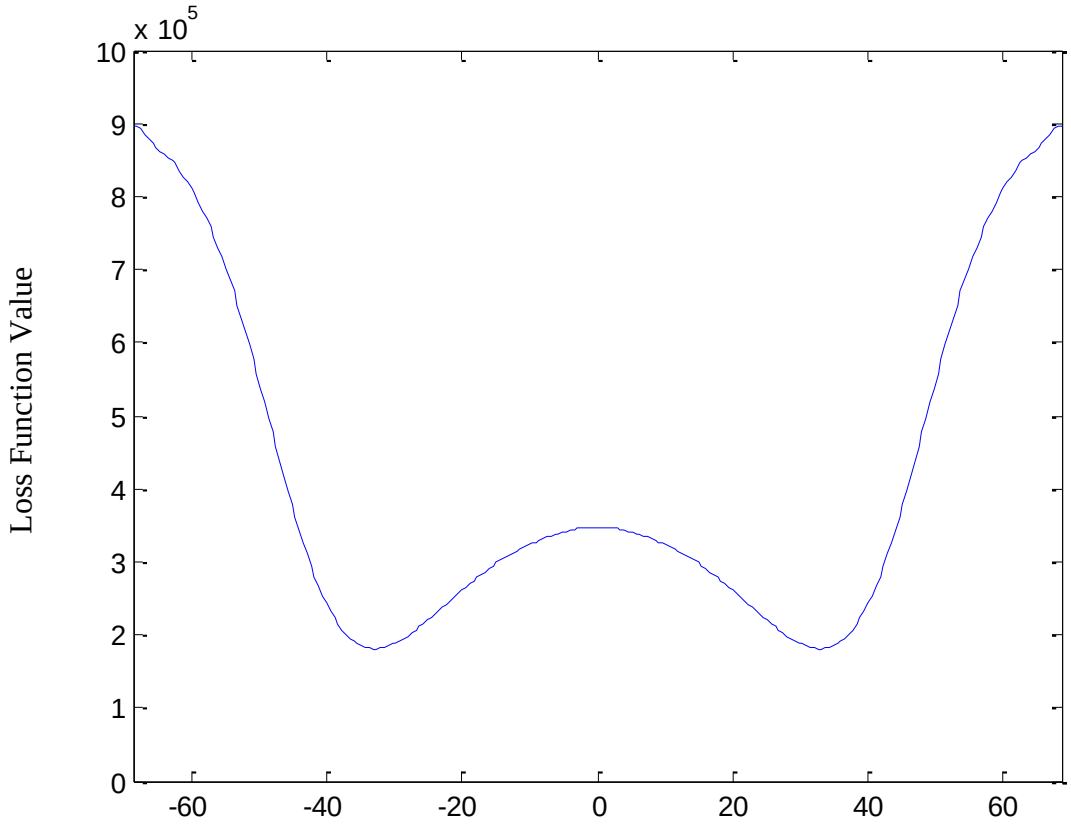

Figure 3.6. The loss function value assuming that the x-y plane ellipse (with aspect ratio 0.8) is projected from an ellipse tilted in space at different angles. The loss function is defined in a same way as described before. The global minimum are located at $\theta = \pm 33.2^{o}$, and the loss value first decreases monotonically as the tilted angel deviates from zero until the global minimum points are reached, and then increase monotonically as the tilted angle is further increased. The model predicts that a static 3D wobbling disk is perceived, which is consistent with the human percept.

3.3. Question 2: Why a 3D disk is difficult to see if the ellipse is narrow or rocking?

3.3.1. Methods

We first calculated the loss function for narrower ellipse to make a comparison to the loss function obtained from fat ellipse. The definition of the loss function was identical to that was described above. We then hypothesized that the when the 2D ellipse is presented, the visual system starts with an interpretation that the structure is an ellipse on the 2D image plane. Then the visual system keeps updating the interpretation so that the loss function value will keep decreasing until a global minimum is reached, and the stable precept is the one correspond to the interpretation with globally minimum loss.

To simulate the inference process, we adopted the gradient descent algorithm (Morse and Feshbach, 1953) with fixed step size. Assuming that the interpretation starts at angle $\theta = 0^o$, which is a 2D ellipse on the image plane, then θ is updated in the direction that decreases the loss function value defined in Equation (3.1.) following the update rule:

$$\theta_{n+1} = \theta_n - \Delta\theta \frac{\partial E}{\partial \theta} \qquad (3.12.)$$

where $\Delta\theta$ is the step size and we used $\Delta\theta = .1^{o}$ in this study. It is sensible to assume that the more step it takes for the gradient descent algorithm to converge, the more difficult it is for the visual system to achieve a stable percept, so the number of steps can be a measurement of how difficult it is to form a stable percept.

3.3.2. Results

When the x-y plan image is a narrower ellipse, though it is still likely that human observers can perceive a tilted wobbling disk after longer observation, it is more difficult for them to achieve such a 3D percept, instead, a 2D rotating ellipse is easier to perceive. Figure 3.7. shows the loss function value when the x-y plane image is a narrower ellipse with aspect ratio 0.5. It is clear that the minimum point is still not at tilted angle $\theta = 0^{o}$, instead the minimum points are at $\theta = \pm 58.4^{o}$. And when the interpretation is a tilted disk, the tilted angle is $\theta = \pm 60^{o}$.

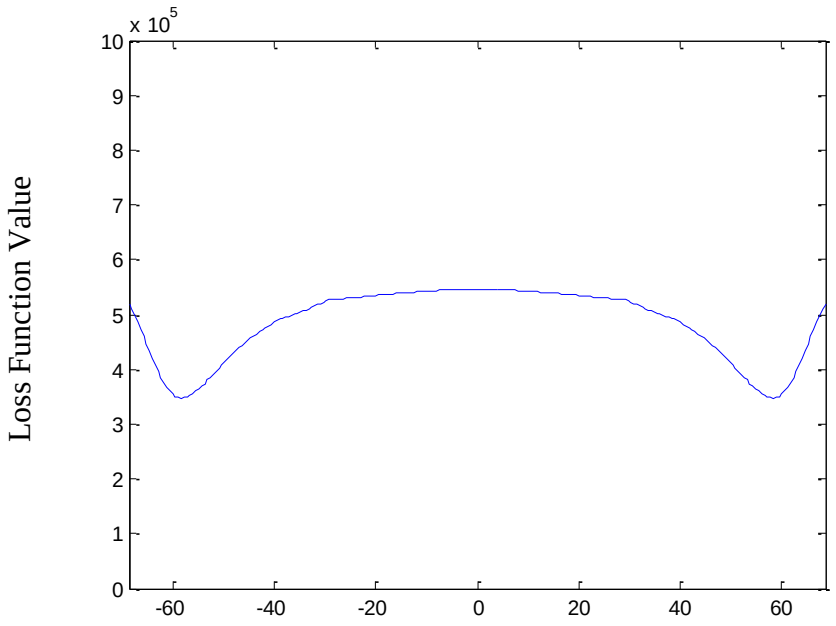

Figure 3.7. The loss function value assuming that the x-y plane ellipse (with aspect ratio 0.5) is projected from an ellipse tilted in space at different angles. The loss function is defined in a same way as described before. The global minimum are located at $\theta = \pm 58.4^{o}$, and the loss value first decreases monotonically as the tilted angel deviates from zero until the global minimum points are reached, and then increase monotonically as the tilted angle is further increased. The model predicts that a static 3D wobbling disk is perceived given that the visual system is provided sufficient time to reach the global minimum. Human observers need longer observation time and feel more difficult to perceive a 3D wobbling disk, instead a 2D rotating ellipse is easier to perceive. This discrepancy is explained in details below.

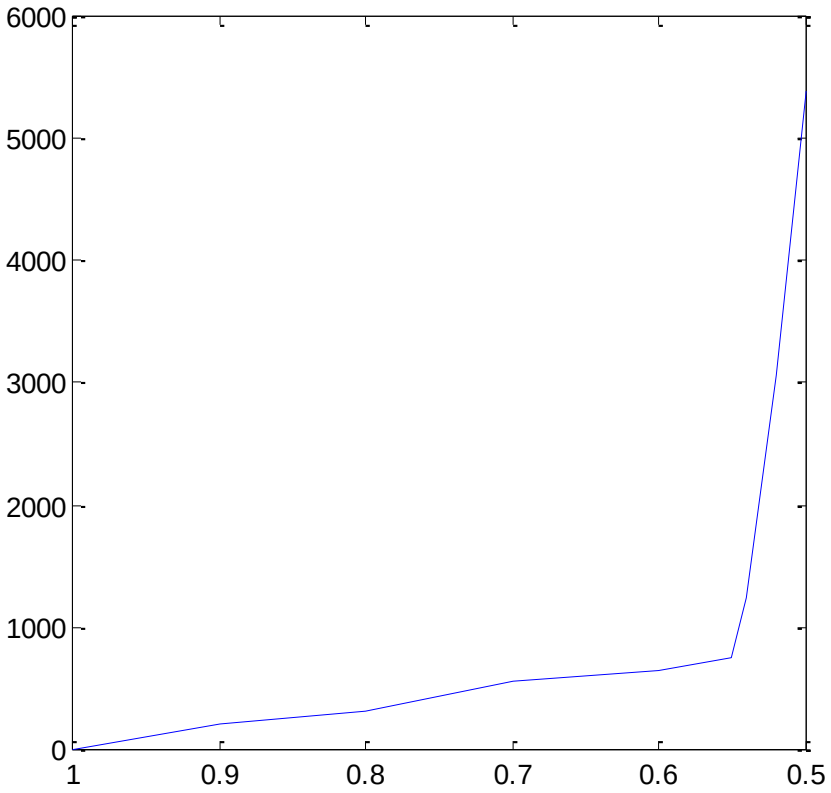

Figure 3.8. The number of iterations for a gradient descent algorithm to reach global minimum when the x-y plane image is ellipse with different aspect ratios. Number of iterations increases as the aspect ratio gets smaller, more specifically, when the aspect ratio is smaller than 0.6, the number of interactions increases dramatically as aspect ratio decreases. Number of iterations in the simulation can indicate the time it takes for a biological visual system to achieve an optimal visual interpretation of structure and motion. The simulation is consistent with the empirical evidence that it takes longer (and more difficult) for human observers to perceive a tilted 3D wobbling disk if the x-y plane ellipse is narrow.

Why it takes longer and more difficult for human observers to perceive a 3D tilted wobbling disk in a narrow ellipse case? If we hypothesize that the visual system is trying to minimize the loss function value starting from a interpretation of ellipse on the x-y image plane (tilted angle $\theta = 0^o$) following a method similar to the gradient descent (Morse and Feshbach, 1953), since compared to the loss function shown in Figure 3.6., the loss function in Figure 3.7. decreases slower from the point of tilted angle $\theta = 0^o$, more specifically, starting from the point of $\theta = 0^o$ the gradient of the loss function for a narrow ellipse is smaller than that for a fat ellipse. Using the gradient descent algorithm described in the previous section, we found that it takes the algorithm more steps to converge to the global minimum for narrower ellipse than that for fat ellipse. Figure (3.8.) showed the number of iterations that is needed when a global minimum of the loss function is reached, if starting from the interpretation that the tilted angle $\theta = 0^o$. It is clear that when the ellipse is narrower, it takes longer to reach the optimal solution. Actually when the aspect ratio is small enough (for example, smaller than 0.6), the number of iterations needed to converge to global minimum increases dramatically as the x-y plane ellipse aspect ratio gets smaller. This fact indicates an optimal solution (tilted 3D disk) for very narrow ellipse is practically impossible to achieve if starting from the interpretation of an x-y plane ellipse. Empirically, Duncan (1975) found that in order that a rotating ellipse can be perceived as a wobbling 3D disk, the eccentricity of the ellipse should be 0.87, which is equivalent to requiring that the aspect ratio of the ellipse being larger than 0.49. It is clear in Figure (3.8.) that the number of steps needed for a gradient descent algorithm to converge to a global minimum slowly increases as the aspect ratio decreases from 1.0, until around 0.55. Then around the point of

aspect ratio 0.50, the steps needed starts increasing dramatically as the aspect ratio decreases, which means practically it is impossible for the visual system to perceive a 3D wobbling disk beyond aspect ratio of 0.5. Our computational model quantitatively matches the empirical results by Duncan (1975).

The same logic can explain the interesting empirical fact that even for ellipse with aspect ratio that human observers can reliably perceive a tilted 3D wobbling disk when the ellipse is continuously rotating, it is more difficult for the human observers to perceive the disk if the ellipse is rocking. Our computational model and gradient descent simulation shed lights on this phenomenon as well. More specifically, if we assume that human observers have an initial interpretation that the shape is 2D (tilted angle $\theta = 0^{o}$), and then evaluates the loss function value and aim to decrease the loss by adopting new interpretations (different tilted angle). As we have shown above, the global minimum of the loss function lies on the tilted angle that gives rise to a tilted 3D wobbling disk (Figure 3.6. and 3.7.), and at the same time it takes time to reach the global minimum (as number of steps shown in Figure 3.8). When the ellipse is continuously rotating, it is possible for the visual system to actively update the interpretation to decrease the loss and ultimately reach the global minimum. However, when the ellipse is rocking, the updated interpretation is disturbed when the ellipse stopped and reversed rotation abruptly so that it is hard for the visual system to update the interpretation until a global minimum is achieved. In the meanwhile, it is still not clear what the interpretation will be right after the point when motion abruptly changes direction.

3.4. Question 3: Why the dot position on the ellipse affects structure perception?

3.4.1. Methods

In this section, we aimed to model the percept when there is a dot on either the major or minor axis of the ellipse. Empirically, when the dot is on the minor axis, human observers perceive a tilted rigid 3D cone whereas when the dot is on the major axis, human observers perceive a tilted rotating 3D disk with a dot sliding on the disk.

Following the definitions from Equation (3.4.) to (3.10.), a tilted disk can be expressed as:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a\cos\alpha \\ b\sin\alpha \\ c\sin\alpha \end{pmatrix} \tag{3.13.}$$

where $c = \sqrt{a^2 - b^2}$. There is a rotation ω around the z axis, so considering the rotation:

$$\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} \cos\omega t & -\sin\omega t & 0 \\ \sin\omega t & \cos\omega t & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a\cos\alpha \\ b\sin\alpha \\ c\sin\alpha \end{pmatrix} \tag{3.14.}$$

From Equation (3.14.), when only the rigid rotation is considered, the tilted ellipse at time t can be expressed as:

$$\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} a\cos\omega t\cos\alpha - b\sin\omega t\sin\alpha \\ a\sin\omega t\cos\alpha + b\cos\omega t\sin\alpha \\ c\sin\alpha \end{pmatrix} \tag{3.15.}$$

So the rigid motion is simply the temporal derivative of Equation (3.15.), which is:

$$\begin{pmatrix} v_x(t) \\ v_y(t) \\ v_z(t) \end{pmatrix} = \begin{pmatrix} -a\omega\sin\omega t\cos\alpha - b\omega\cos\omega t\sin\alpha \\ a\omega\cos\omega t\cos\alpha - b\omega\sin\omega t\sin\alpha \\ 0 \end{pmatrix} \tag{3.16.}$$

In addition to the rigid rotation, we assume that there is a spin along the contour, and the spin is always in the direction of the contour tangent that cancels part of the rigid rotation:

$$\begin{pmatrix} \psi_x \\ \psi_y \\ \psi_z \end{pmatrix} = \frac{\phi}{\sqrt{\frac{b^2\sin^2\alpha}{\cos^2\theta} + a^2\cos^2\alpha}} \begin{pmatrix} -\frac{b\sin\alpha}{\cos\theta} \\ a\cos\theta\cos\alpha \\ a\sin\theta\cos\alpha \end{pmatrix} \tag{3.17.}$$

When the dot is on the minor axis, the position taking into account rotation is:

$$\begin{pmatrix} x_0(t) \\ y_0(t) \\ z_0(t) \end{pmatrix} = \begin{pmatrix} -d\sin\omega t \\ d\cos\omega t \\ h \end{pmatrix} \qquad (3.18.)$$

where d is a constant describing the distance from the dot to the ellipse center, and h is a variable depicting the perceived distance from the dot to the x-y image plane.

Given that human observers perceive a rigid 3D cone when the dot is on the minor axis, it is sensible to assume that the visual system 'fills in' implied surface in the space between the tip and the base of the perceived cone, so that the inferred structure is no longer an isolated tilted disk and a dot. As a result, we assumed that while making inference about the 3D structure of the stimulus, the visual system also take into account motion on the implied contours in the space between the dot and tilted disk. Figure (3.9.) demonstrated the perceived tilted cone. The black contour and the black dot are perceived base and tip of the cone from the 2D ellipse with a dot on it, and the three green contours are implied contours on the surface of the cone. Note that there are two motion components for the 3D cone, one is a rotation ω along the z-axis, and the other is the spinning ψ along an axis going through the center of the tilted disk and normal to the tilted disk plane. In Figure (3.9.) the 3 contours in planes parallel to the tilted disk plane, and are of a quarter, a half and three quarters of the cone's height away to the disk plane.

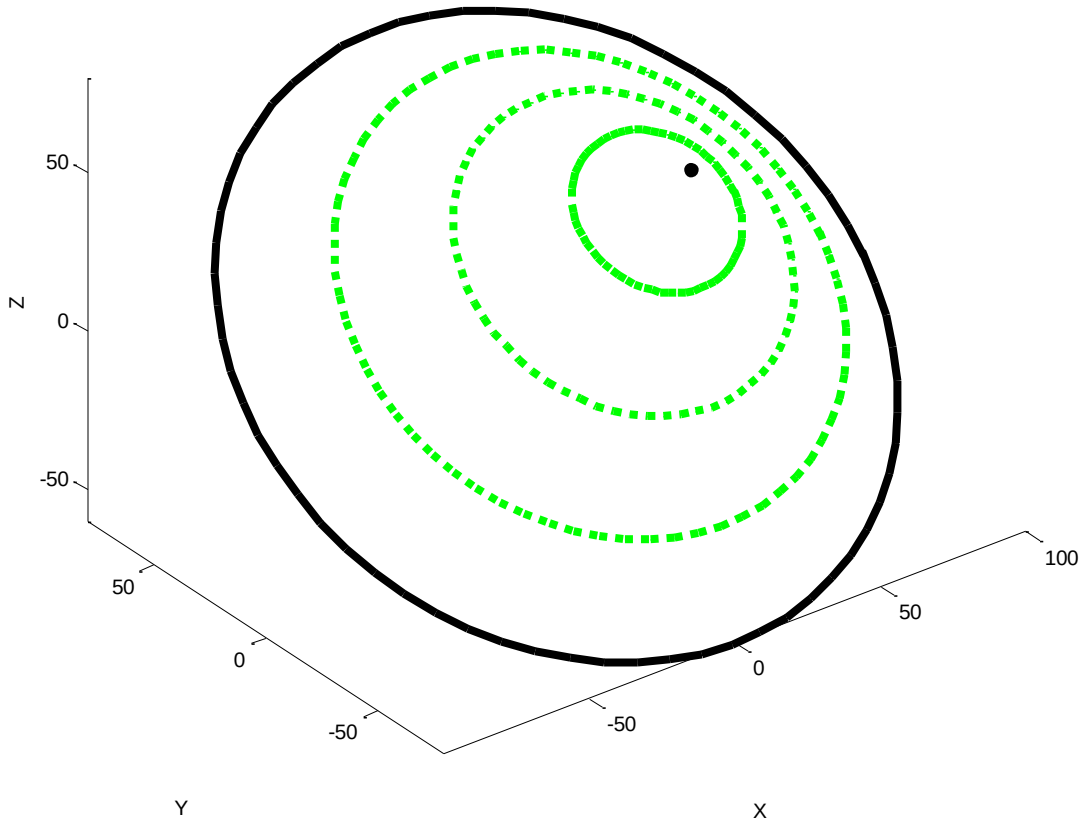

Figure 3.9. The perceived tilted cone from a rotating ellipse with a dot on the minor axis. The stimulus is an ellipse with aspect ratio *.8* on the x-y plane. A dot is located on the minor axis with distance *.6a* from the ellipse center, where *a* is the ellipse's major half axis. The black contour and the black dot are the perceived base and tip of the cone from the 2D ellipse with a dot on it, and the three green contours are implied contours on the surface of the cone. Note that there are two motion components for the 3D cone, one is a rotation ω along the z-axis, and the other is the spinning ψ along an axis going through the center of the tilted disk and normal to the tilted disk plane. The 3 contours in planes parallel to the tilted disk plane, and are of a quarter, a half and three quarters of the cone's height away to the disk plane.

So the position of the green contour in the midway is:

$$\begin{pmatrix} x_2(t) \\ y_2(t) \\ z_2(t) \end{pmatrix} = \frac{\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} + \begin{pmatrix} x_0(t) \\ y_0(t) \\ z_0(t) \end{pmatrix}}{2} \tag{3.19.}$$

And the rotational components of the same contour in the mid-way are:

$$\begin{pmatrix} v_{2x}(t) \\ v_{2y}(t) \\ v_{2z}(t) \end{pmatrix} = \begin{pmatrix} v_x(t) \\ v_y(t) \\ v_z(t) \end{pmatrix} \frac{\sqrt{x_2(t)^2 + y_2(t)^2}}{\sqrt{x(t)^2 + y(t)^2}} \tag{3.20.}$$

In Equation (3.18.) and (3.19.) $x(t)$, $y(t)$, and $z(t)$ represent position of the tiled disk, which are defined in Equation (3.15.). $v_x(t)$, $v_y(t)$ and $v_z(t)$ are the rotational components for the tilted disk, which are defined in Equation (3.16.). $x_0(t)$, $y_0(t)$ and $z_0(t)$ represent the position of the dot, which are defined in Equation (3.18.).

To calculate the spinning component of this green contour, we first calculate the spinning center on the same green contour plane:

$$\begin{pmatrix} c_{2x}(t) \\ c_{2y}(t) \\ c_{2z}(t) \end{pmatrix} = \begin{pmatrix} \dfrac{bc\left((2b-d)\dfrac{c}{b}+h-c\right)\sin\omega t}{2a^2} \\ -\dfrac{bc\left((2b-d)\dfrac{c}{b}+h-c\right)\cos\omega t}{2a^2} \\ -\dfrac{b^2\left((2b-d)\dfrac{c}{b}+h-c\right)\cos\omega t}{2a^2} \end{pmatrix} \tag{3.21.}$$

So the spinning component of this contour is:

$$\begin{pmatrix} \psi_{x2} \\ \psi_{y2} \\ \psi_{z2} \end{pmatrix} = \begin{pmatrix} \psi_x \\ \psi_y \\ \psi_z \end{pmatrix} \frac{\sqrt{\left(x_2(t)-c_{2x}(t)\right)^2+\left(y_2(t)-c_{2y}(t)\right)^2+\left(z_2(t)-c_{2z}(t)\right)^2}}{\sqrt{x(t)^2+y(t)^2+z(t)^2}} \tag{3.22.}$$

And the motion of this contour is a combination of rotational and spinning components. Similarly we can get the motion of the other two green contours.

If the dot is on the major axis of the ellipse, then the position of the dot is:

$$\begin{pmatrix} x_0(t) \\ y_0(t) \\ z_0(t) \end{pmatrix} = \begin{pmatrix} d\cos\omega t \\ d\sin\omega t \\ h \end{pmatrix} \tag{3.23.}$$

And we can compute the motion for three implied contours following similar logic above.

Recall that the regularization theory proposed that the motion measurement is the motion field that gives rise to a minimum value of the loss function:

$$E(\vec{v}) = \int \sum_{m=0}^{\infty} c_m \left| \nabla^m \vec{v} \right|^2 \qquad (3.24.)$$

Here in our example given the ellipse on the x-y plane, $E(\vec{v})$ is a function of cone height *h* and spin ϕ in Equation (3.17). We first calculated the values for each of the regularization terms up to the 3rd order defined in Equation (3.24.) as a function of the spin ϕ for different cone height *h* and for each *h* we select the spin that gives rise to a global minimum of the loss function. Then we compare the loss function value with optimal spin across different cone height *h* to determine the optimal cone height that gives rise to the minimum loss function value. We hypothesized that the optimal cone height *h* in our computational model is preferred by the visual system.

3.4.2. Results

Figure (3.10) showed the loss function when the dot is on the minor axis, as a function of the cone height *h*. The ellipse's major half axis was 100 (arbitrary unit), minor half axis was 80, and distance between the dot and the ellipse center was 60. The model predicted that the cone height that gives rise to a global minimum of the loss function is $h_{model} = 75$. Form straightforward

geometrical derivation, we know that the symmetric cone that gives rise to the 2D projection on the x-y plane has a height $h_{geo} = 80$. Prediction from our computational model is quantitative close to the geometrical solution.

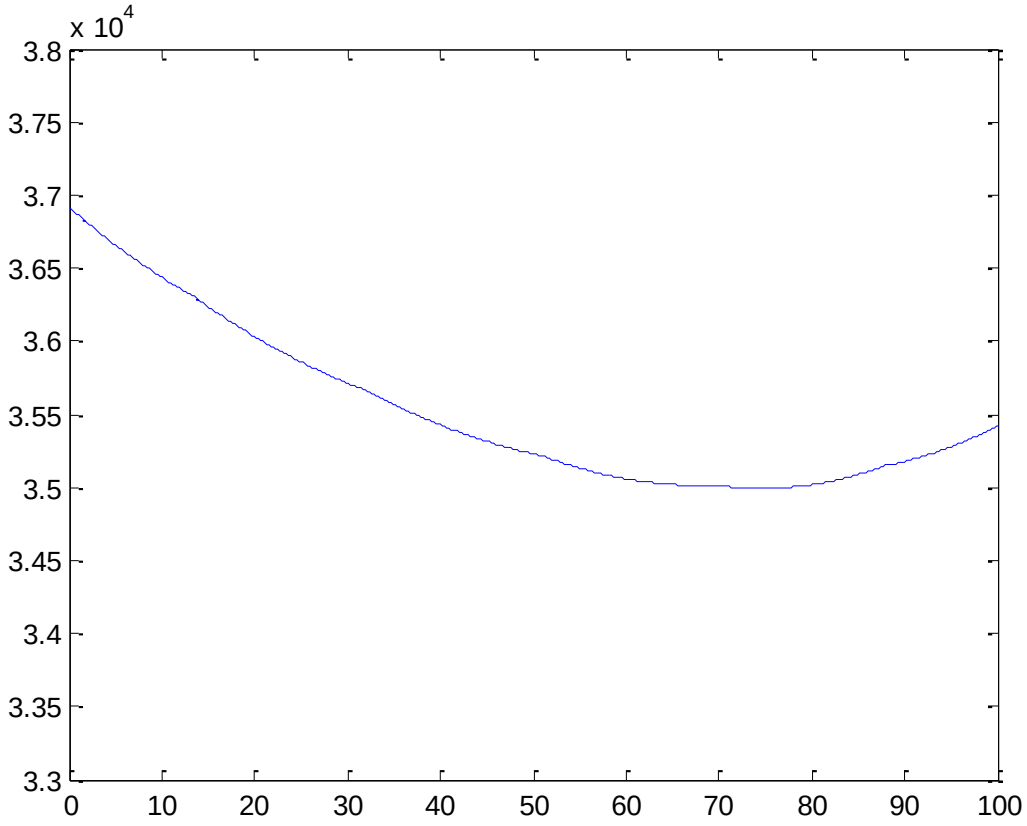

Figure 3.10. Loss function as a function of cone height (dot on minor axis). The stimulus is an ellipse on the x-y plane with a dot on the minor axis, and $\sigma = 2$ was used in the simulation. The ellipse has major half axis a = 100, aspect ratio = .8, the distance from ellipse center to the dot d = 60. The computational model predicted $h_{model} = 75$, and geometrical derivation results in $h_{geo} = 80$. Prediction from our computational model is quantitative close to the geometrical solution.

When manipulating the distance between the dot and the ellipse center, we found that when the distance was 40 $h_{model} = 33$, and geometrical derivation results in $h_{geo} = 53$, and when the distance was 20, $h_{model} = 14$, and geometrical derivation results in $h_{geo} = 27$. The predictions from our model reflected the qualitative trend of results from geometrical derivation. When manipulating the aspect ratio of the ellipse while fixing the distance between the dot and the ellipse center to be 60, we found that when aspect ratio was .7 $h_{model} = 12$, and geometrical derivation results in $h_{geo} = 19$, and when the aspect ratio was 20, $h_{model} = 9$, and geometrical derivation results in $h_{geo} = 15$. Again the predictions from our model reflected the qualitative trend of results from geometrical derivation. To summarize, when the dot is on the minor axis of the ellipse, our computational model predicted that a tilted 3D cone is the optimal structure interpretation, which is consistent with human observer's percept.

When the dot is on the ellipse's major axis, the computational model predicted $h_{model} = 0$ for any d. The loss function as a function of cone height is shown in Figure (3.11.). This means that the optimal solution is when the dot is on the tilted disk. And given that the dot has smaller spinning component relative to the rotational component compared to that on the contour, and that the motion is a combination of the spin and rotation (note that the spinning and rotational components are in the opposite directions), so the computation model predicted that the optimal structure interpretation is a dot slipping on a tilted wobbling disk. The model prediction is again consistent with human observer's percept.

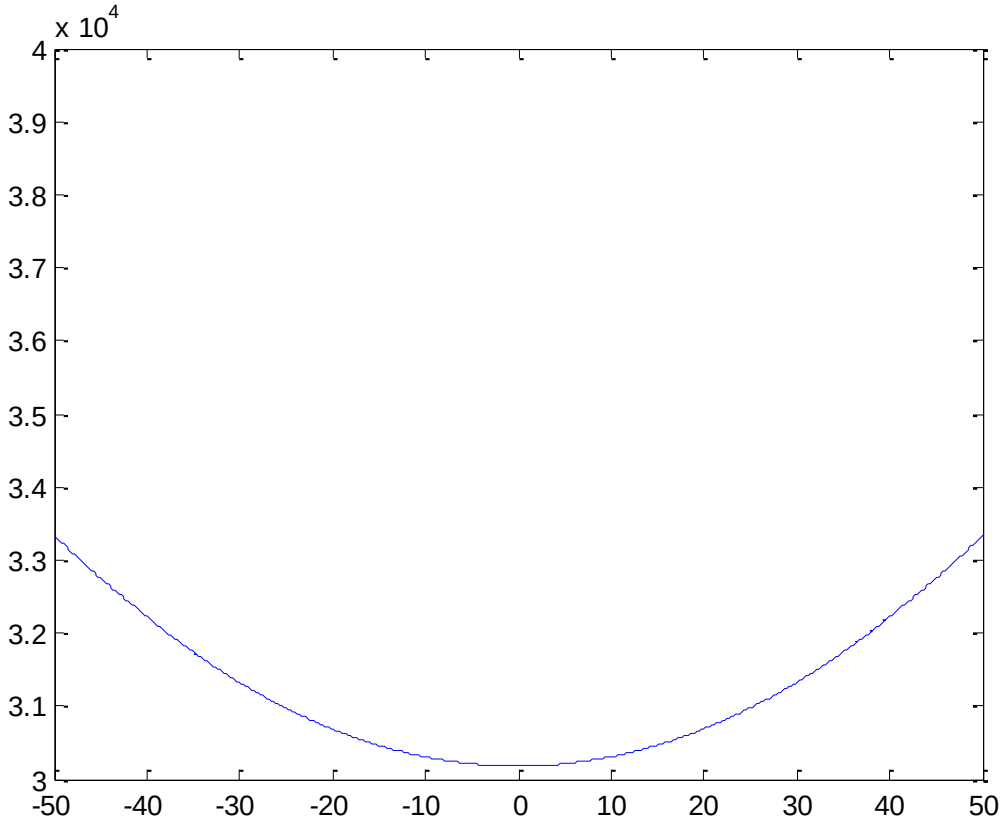

Figure 3.11. Loss function as a function of cone height (dot on major axis). The stimulus is an ellipse on the x-y plane with a dot on the major axis, and $\sigma = 2$ was used in the simulation. The ellipse has major half axis a = 100, aspect ratio = .8, the distance from ellipse center to the dot d = 60. The computational model predicted $h_{model} = 0$, meaning the dot is on the tilted disk. And given that the dot has a smaller spin component than on the contour, so the dot is perceived to be slipping on the tilted disk at a faster speed than the disk. The model prediction is consistent with human observer's percept.

3.5. Discussion

To answer the question why a tilted 3D wobbling disk is perceived when a rotating ellipse is presented, we invested the motion properties of the 3D percepts generated by a 2D stereokinetic stimulus. We hypothesized that the perceived 3D shape is an optimal structure interpretation with certain motion properties. More specifically, we hypothesized that among all the possible 3D interpretations, the one with slowest and spatially smoothest motion field is proffered. We framed the 3D structure from motion question into an optimization problem. To solve the 3D structure and motion problem from the stereokinetic stimuli, additional priors have to be imposed in order that a unique answer can be found. The rigidity assumption on 2D is widely used that directly provides constraint to the object shape (Ullman, 1979, 1983). And 2D regularizations on motion are also popular (Yuille and Grzywacz, 1988; Weiss, Simoncelli, and Adelson, 2002). We showed that 3D motion regularization provide sufficient constraints for solving the motion and structure of 3D object. We also showed that when higher order smoothness of motion is adopted, there is no need to further impose the rigidity assumption to solve the structure from motion.

Human observers, when presented with the stereokinetic stimulus, typically perceive a rigid moving 2D pattern at the beginning and in the end a stable percept of a 3D moving object is achieved. In our study, we investigated a rotating rigid ellipse on a 2D image plane, which leads to the percept of a rigid ellipse rotating on the image plane when the aspect ratio is small and a tilted wobbling disk in 3D when the aspect ratio is large. We showed that the interpretation of a

wobbling 3D disk is the global optimal solution when the ellipse is fat, which is consistent with human observer's percept. On the other hand, when the ellipse is narrow, although the global optimal is still a tilted 3D disk, our computational model suggests that it takes longer for the visual system to reach the global optimal. Siegel and Andersen (1988) found that the temporal integration is important in the structure from motion perception for both human and monkey. So longer processing time required to converge for narrow ellipse suggested that it is more difficult for human to perceive a 3D disk in this scenario, which is also consistent with human observer's experience. Following the same logic, our model also predicted that given a rocking ellipse stimulus, the human visual system may not have sufficient time to achieve an optimal solution that gives rise to a 3D wobbling disk interpretation.

Yuille and Grzywacz (1988) proposed the regularization theory to explain human motion perception in 2D. They hypothesized that, among all possible motion interpretations, the one that gives rise to slowest and spatially smoothest motion best explains human percept. Weiss, Simoncelli, and Adelson (2002) rephrased the regularization theory using a Bayesian probabilistic framework, and explicitly introduced the idea of "slow and smooth" prior, and essentially their framework is special case of the regularization theory taking into account the zero (slow) and first order (smooth) of the regularization terms. Yuille and Grzywacz (1989) suggested that to deal with motion that not merely composed of translation and rotation, higher order regularization terms are necessary. In our study, we found that lower order regularization terms focused more on the slowness of total motion, whereas the higher order regularization terms penalize heavily the spatially non-smoothness of the motion field, and consequently the

non-rigid structure interpretation. To achieve a motion and structure interpretation that is moving slowly and smoothly, higher order regularization terms needed to be imposed. In addition to that, it is important to understand that the optimal solution is not achieved immediately upon the visual stimuli. Instead, it takes time for the visual system to update the structure and related motion interpretation until an optimal solution is found. Based on the empirical evidence and the computational modeling, we believe that the slow and higher order spatially smooth prior account for both the 2D and 3D percept generated by this stereokinetic stimulus, and that the perceived shape and motion by human observer is an optimal solution from our computational model.

References

Duncan, F. (1975). Kinetic art: On my psychokinematic objects. *Leonardo, 8,* 97-101.

Horn, B. & Schunck, B. (1981). Determining optical Flow. *Artificial Intelligence*. 17, 185-204.

Musatti, C. (1924). Sui fenomeni stereocinetici. *Archivio Italiano di Psicologia* 3 105-120.

Morse, P. & Feshbach, H. *(*1953*). Methods of Theoretical Physics, Volume 1.* McGraw-Hill.

Movshon, J. *et al.* (1985). The analysis of moving visual patterns. *Pattern Recognition Mechanisms.* Vol. 54. 117-151.

Rokers B., Yuille, A., & Liu, Z. (2006). The perceived motion of a stereokinetic stimulus. *Vision Research.* 46, 2375-2387.

Siegel, R. & Andersen, R. (1988). Perception of three-dimensional structure from motion in monkey and man. *Nature*. Vol. 331, No. 6153. 259-261.

Shearer, R. &. Gould, S. (1999). Of Two Minds and One Nature. *Science.* Vol. 286 no. 5442: 1093-1094.

Ullman, S. (1979). *The interpretation of Visual Motion*, the MIT Press.

Ullman, S. (1983). Maximizing rigidity: The incremental recovery of 3-D structure from rigid and rubbery motion. *Massachusetts Institute of Technology, Artificial Intelligence Laboratory Memo* 721.

Wallach, H. & O'Connell, D. (1953). The kinetic depth effect. *Journal of Experimental Psychology.* 45, 205-217.

Weiss, Y. (1998). Ph.D. Thesis -- Massachusetts Institute of Technology, Dept. of Brain and Cognitive Sciences.

Weiss, Y. *et al*. (2002). Motion illusions as optimal percepts. Nature Neuroscience. 5(6), 598-604

Yuillle, A. & Grzywacz, N. (1988). A computational theory for the perception of coherent. *Nature*. 333, 71 – 74.

Yuille, A. & Grzywacz, N. (1989). A Mathematical Analysis of the Motion Coherence Theory. *International Journal of Computer Visoin*. 3, 155-175.

CHAPTER 4

QUANTIFYING AND MODELING A STEREOKINETIC PERCEPT

4.1. Introduction

The goal of visual perception is to make accurate and precise inference about the environment so that an organism can behave properly and survive. Visual perception is an inverse problem, namely, the information from the external world is insufficiency and as a result cannot lead to a unique inference that can direct the proper behavior. Because of the insufficient sensory input information, ambiguity inevitably exists in visual perception. The perception of depth is an example for such ambiguity in visual perception. A two-dimensional (2D) image on the retina conveys indirect depth information about the 3D environment, but any direct measurement of the depth in space is inevitably lost in the process of the projecting information from the external environment to the retina. Although the projection from outside world to retina follows straightforward optics principle, the reversal puzzle, namely recovering the 3D depth from a 2D retina image is much more complex. The complexity of such puzzle finds its roots in the fact that there are infinite 3D structures that can give rise to exactly the same 2D projection on the retina. The human visual system, on the other hand, does not suffer from the ambiguity from the projection and information loss. In most cases, the human visual system recovers the 3D depth of the external environmental accurately.

In the meanwhile, although the human visual system can always achieve a unique inference about the external world, such inference is not guaranteed to be consistent with the physical environment, and visual illusion happens in these scenarios. A good example is the Benussi effect (Benussi, 1916), specifically, when a circle is rotating on an image plane and the center of the circle is displaced from the rotating axis, the circle seems to be at different depth as the image plane (either floating above or recede below). Musatti (1924) independently discovered this illusion around the same time. In addition to the science society, some artists independently discovered similar visual illusion and used in their arts creations (Duchamp, 1920s; Duncan, 1970s; For review, Shearer and Gould, 1999). If there are multiple eccentric circles rotating on the plane, each of the circles will be perceived at a certain depth. As a matter of fact, it is the same inference process that enables the human visual system to solve the 3D depth ambiguity as the one that makes the visual system subject to the visual illusions described above. Specifically, this inference process consists of combining certain assumptions with the 2D sensory information on retina. The assumptions applied by the visual system are critical in overcoming the inherently ambiguous problem of depth. Depth perception is thus a heuristic process in which inference is not only based on well-defined input sensory information, but prior assumptions as well. In most cases, the prior assumption leads to veridical visual inference that makes us behave properly in the environment and in other cases it leads us to visual illusions.

Previous studies suggested that there are multiple cues and assumptions about the 3D depth that can be used by the visual system. One of the important principles used for monocular depth perception is the linear perspective principle. Based on linear perspective, the projected image

size of an object is inversely proportional to the distance from the observer to object. It is noteworthy that the application of linear perspective principle is directly related to the perceived viewing distance (and the perceived size of the distant object) which is not necessarily the same as the physical measurements. The second important assumption that can assist the visual system to recover the depth is the rigidity prior (Ullman, 1983): all else being equal, if there is an interpretation that is consistent with the motion of a rigid objects, such interpretation will be most likely accepted by the visual system. The rigidity prior is consistent with the statistical distribution of the object shape in the real world. Since a majority of the real world objects preserve rigidity, it is sensible for the visual system to assume that an object of interest is also a rigid object. For a strictly rigid object, the relative distance between different parts of the object remain the same as the object moves. In perception studies, the rigidity is commonly defined as a continuous measurement rather than a yes-or-no decision. In general, a shape that is more rigid is preferred by the rigidity prior.

Figure 4.1. A linear perspective projection example. The projected image size of an object is inversely proportional to the distance from the object to the observer. In this example, the

horizontal distance between the left and right handrails are identical no matter at a point closer to the observer or at a point further away. However, the distance looks longer at closer point whereas it looks shorter at further away points, because of the linear perspective projection.

In addition to linear perspective and rigidity, Wallach and O'Connell (1953) described a depth precept derived from motion cues, which they called the kinetic depth effect (KDE). In KDE, the depth is recovered from the rigidity heuristic and motion cues, and motion is estimated given such rigid shape interpretation. Gibson (1966) also emphasized that that depth information can also arise from motion, more specifically, form points moving at different retinal velocities. And the pattern of the apparent retinal velocity is called the optic flow. In general, this is called depth from motion. For example, if an object is moving, object parts will move in different manners because the parts are at different depth from the observer, and the difference in motion for different object parts is called the motion parallax. The depth from motion problem seems straightforward if the motion measurement is well defined, but the measurement of motion turns out to be a difficult question in the first place. The biggest challenge is the correspondence problem, namely, the problem of deciding in the retina image which part at time point t_1 correspondent to which part at time t_2. When the retina image has salient features, it is less difficult to solve the correspondence problem. However, if no salient feature is available, correspondence problem can bring ambiguity to motion perception. One example is the aperture problem. When a human observer is looking at the stimuli through an aperture the motion is always perceived to be perpendicular to the stripes no matter what the “real” motion is (Figure 4.2.). In general, the measurement of depth and the measurement of motion are tightly

intertwined questions, because the depth determines the correspondence for motion perception and in the meanwhile motion decides the depth interpretation.

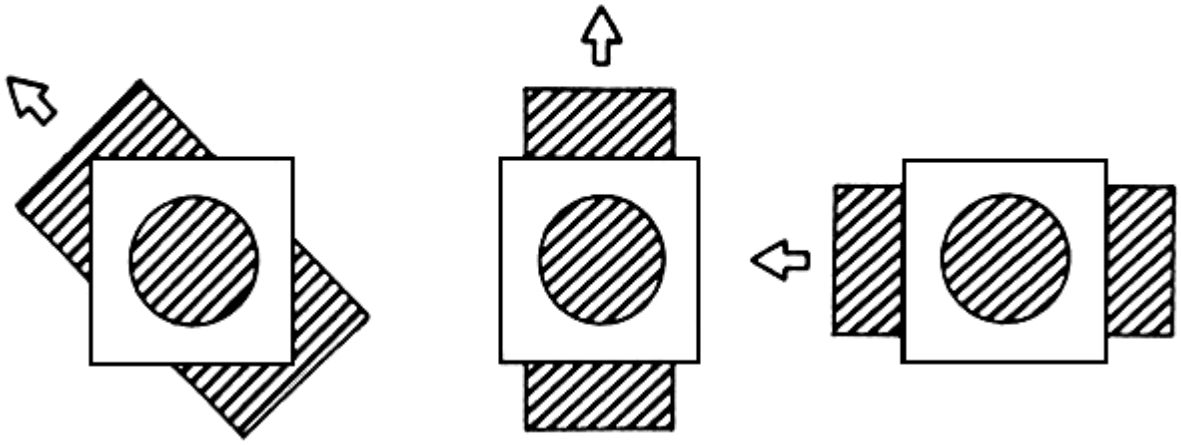

Figure 4.2. Schematic illustration of the aperture problem. For any of the three "real" motion of the board, if a human observer looks through the circular aperture, the observer will report that a motion perpendicular to the stripes is perceived. Adapted from Movshon et al. (1985).

In the current study, we planned to quantify a stereokinetic effect similar to the one Benussi and Musatti discovered. More specifically, when two circles are rotating on an image plane and both circular centers are displaced from the rotation axis, the two circles seem to be each at a different depth. The two circles form a tilted cylinder with a specific depth, and we plan to quantify the depth of the perceived cylinder. Quantitative empirical studies on the stereokinetic effect relied on the participant's reported estimate of the depth, for example, using ruler (Zanforlin, 1988). Such measurements usually involve the observer's motor system which may complicate the

interpretation of the measurement (Goodale and Milner, 1992; Króliczak et al., 2006; Knill, 2005; Hartung et al., 2005).

Rokers, Yuille, and Liu (2006) used a new measurement to quantify a stereokinetic stimulus without asking the subject heavily deploy the motor system. Instead, the task was pressing buttons to adjust a visual stimulus. Their quantitative study showed reliable results with small between and within subject variance. In the current study, we developed a new method to measure the perceived depth of the rotating circles that has heavy reliance on the motor system. More specifically we designed a stimulus that composed of two eccentric rotating circles, one of which has its radius adjustable. The observers adjust the radius of one circle until they perceive a uniformly rotating cylinder, namely, the two circles are perceived to be same size in 3D. Based on linear perspective, the circle that perceived to be further away will have a smaller retina image, and the relative depth can be calculated from the ratio between the two circular radii on the retina image assuming that the absolutely distance from observer to one circle in known. Since the ratio between the two radiuses is identical to the ratio of the two radiuses on the image plane, the relative depth can be calculated from the physical radius ratio.

The interesting fact is that human observers reliably report that they perceive a cylinder of a definite depth. Why they do not perceive a cylinder of zero depth? And why on the other extreme, they do not perceive a cylinder of infinite length? These questions touched the underlying essence of the stereokinetic effect's mechanisms. In addition to the empirical studies aiming to quantify the stereokinetic effect, researchers also have been trying to explore the underlying

mechanism of the phenomenon in the hope that a computational theory will explain the stereokinetic effect. Previous theories tried to explain this effect using the rigidity prior (Wallach and O'Connell, 1953). However, the rigidity prior cannot provide the visual system with a unique solution in the stereokinetic stimulus investigated in the current study. The rigidity prior only gives rise to a family of rigid truncated cone-shape interpretations, but there are still infinite possibilities on the relative depth between circles. On the other hand, human observers not only perceive a unique truncated cone (cylinder is a special truncated cone), but also perceive a cone with a definite depth. Why such depth is perceived remains an open question.

Yuille and Grzywacz (1988) proposed the regularization theory claiming that a motion interpretation with the slowest and spatially smoothest 2D motion is preferred by the visual system. Rokers, Yuille, and Liu (2006) further demonstrated that perceptual ambiguity in a stereokinetic stimuli composed by a single ellipse can be resolved using slow motion constraints in 3D. In the current study, we hypothesized that the visual system groups the two circles to form a cylinder percept, and the cylinder structure interpretation that gives rise to the slowest and spatially smoothest motion in 3D is preferred by the visual system, and such preferred interpretation will give rise to a definite depth of the perceived cylinder. From the hypothesis, we developed a computational model that explains the mechanism of the stereokinetic effect. We tested the computational model with the empirical finding from the current study and found that the model qualitatively predicted the empirical results.

4.2. Experiments

4.2.1. Experiment 1: Radius ratio adjustment and distance pointing

4.2.1.1. Participants

19 students and faculty (9 female, 10 male) from UCLA participated in the experiment.

4.2.1.2. Apparatus and stimuli

The computer graphical stimuli are rendered using OpenGL and PsychToolBox 3 in MATLAB environment. Stimuli are displayed on a 20 inch CRT monitor with resolution of 1280 × 960 pixels and refresh rate of 75 Hz. A vision cube is used to mask of the edges of the screen, and a chin rest is also used to fix the viewing distance at 40 cm.

A light-emitted diode (LED) sensor was attached to the tip of the participant's index finger, and its position was accurately tracked by a Precise Position Tracking (PPT) recording system manufactured by WorldWizard (accuracy = ±.5 cm, precision = ±.1 cm). This system can track the position of the LED at a very high frequency (latency < 20 milliseconds), and the participants pressed a button to control the recording system.

Two white circles are displayed on a black background, rotating around a common axis. One of the two circles is always of a fixed size, and the size is randomly chosen from five possible radii. The radius of the other circle is rendered by adding to the fixed radius a random real number (may be positive or negative). The participants observe the stimuli monocularly using their dominant eyes.

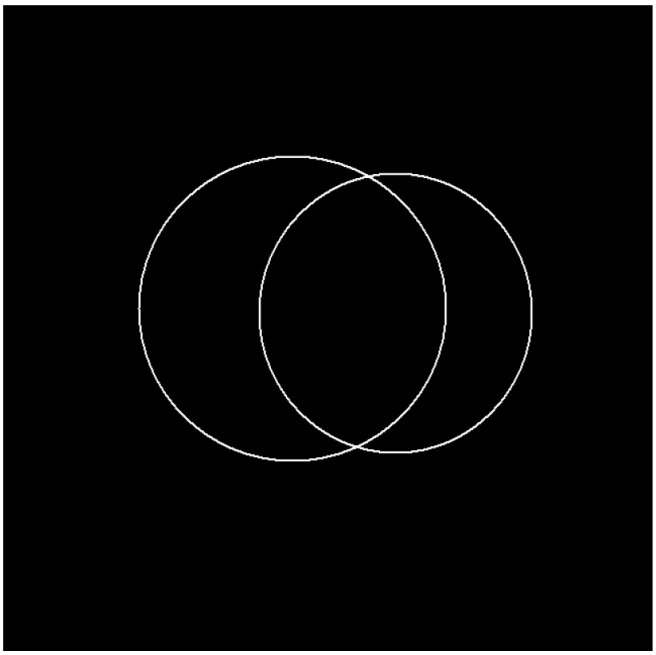

Figure 4.3. Experiment stimuli consist of two rotating white circles in a dark background. The two circles are rotating along a common axis perpendicular to the image plane going through the mid-way of the two circular centers. The participant observes the stimulus monocularly using the dominant eye through a viewing tube in a dark testing room.

4.2.1.3. Procedure

In each experimental trial, participants first adjusted the size of the variable circle so that a uniform cylinder can be perceived. It was emphasized to the participants that, by uniform cylinder, we meant that the two circles were perceived to be of identical size in 3D (they were perceived to be at different distance from the participants). More specifically, because of the 3D percept from stereokinetic effect, two circles of identical size but at different distance from observer will be have different sizes on the retina because of linear perspective projection (more generally, on any projection plane perpendicular to the participant's sightline). Besides, all participants were asked to indicate if the variable circle is perceived to be in front (close to the observer) or at back. Since all the participants reliably indicated that the larger circle is always in front, we define the larger circle to be the one in front and the smaller one at back in our analysis.

After the participants had finished adjusting the radius of the variable circle so that a uniform cylinder was perceived, they were asked to use their index fingers to point to the front end of the perceived cylinder, and they were told make sure that their index finger tips were at same distance as the cylinder end to themselves. A blue LED was attached to the finger tip of the participant, and the participants controlled the PPT system described in the previous session by clicking a button to record the LED's position. The LED's position was recorded by taking 5 recordings within one second, and the median of five groups of tracking coordinates will be used as a measurement. And then they were asked to point to the back end of the cylinder and the

position of the LED on the finger tip was recorded as well. There are 60 trials in the experiment, and it takes the participant about 45 minutes to finish the test.

4.2.1.4. Results

We first analyzed the ratio between the smaller circle radius and the larger circle radius. Suppose that the front end of the perceived uniform cylinder is at distance *D* from the observer, and the two circles are of radius r_1 and r_2 respectively (Figure 4.4.), then the height of the perceived cylinder (*h*) can be obtained follow the linear perspective projection principle:

$$h = \left(\frac{r_2}{r_1} - 1 \right) D \qquad (4.1.)$$

Define $r = \frac{r_1}{r_2}$ (so that *r* is always smaller than 1), we get:

$$h = \left(\frac{1}{r} - 1 \right) D \qquad (4.2.)$$

From (4.2.) we know that with a constant *D*, the depth of the perceived cylinder can be obtained from participants' radius adjustment ratio (*r*), more specifically, the larger the *r*, the shorter the perceived cylinder.

The reported ratio of the two radius was *.904 ± .003* (standard error), which is significantly different from 1 (*t(18) = 35.56, p < .001*). That means participants reliably perceive a 3D uniform cylinder in the experiment, because a ratio of 1 should be expected if percept was 2D. The ratio for trials with inter-circular distance (ICD) 2.24 cm was *.920 ± .003*, and the ratio for trials with ICD = 4.48 cm was *.888 ± .004*. The effect of the ICD was highly significant: *F(1, 18) = 57.921, p < 0.001* (see Figure 4.5.). The fact that condition of ICD = 4.48 cm has smaller ratio means that the perceived cylinder is significantly longer than that in the ICD = 2.24 cm condition, assuming that the viewing distance are identical in both conditions.

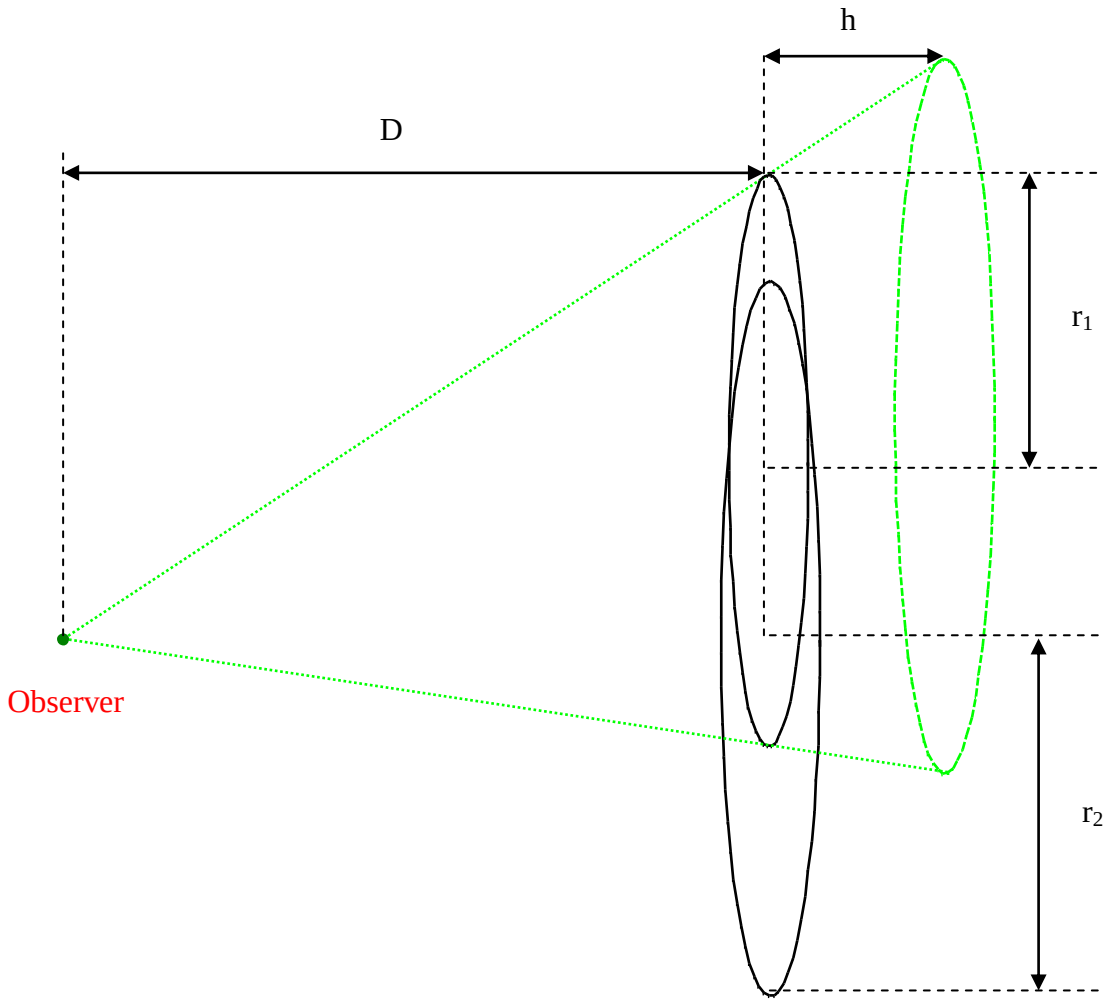

Figure 4.4. Schematic illustration of the percept from the stimuli of two rotating circles. The two circles are at the same distance from the observer. They are rotating around an axis perpendicular to the screen plane at rotational speed ω. Without loss of generality, the larger circle (r_1) was perceived to be at distance D from the observer whereas the smaller circles (r_2) was perceived to be a projection of the hypothetical circle at distance $D + h$ from the image plane (green dashed circle). The dashed green line starting from the observer depicts the linear perspective projection.

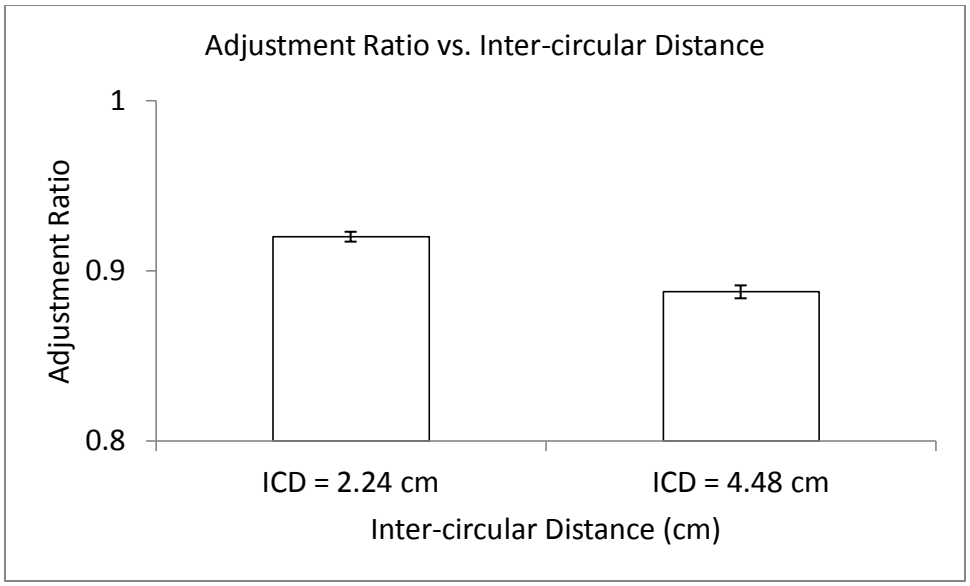

Figure 4.5. Adjustment radius ratio for the inter-circular distance factor (N = 19). Error bar represents standard error. Note that the vertical axis starts from .8 and ends at 1.0. Adjustment ratio under condition ICD = 4.48 cm is significantly smaller than that under condition ICD = 2.24 cm, indicating that the perceived cylinder is longer in ICD = 4.48 condition, assuming that the viewing distance are identical in both conditions.

The adjustment ratio for trials with fixed circle radius 2.85 cm was *.902 ± .005*, for trials with radius 3.70 cm was *.897 ± .006*, for trials with radius 4.56 cm was *.904 ± .006*, for trials with radius 5.41 cm was *.909 ± .006*, and for trials with radius 6.27 cm was *.908 ± .007*. The effect of the radius was significant: $F(4, 72) = 3.035$, $p = 0.023$ (see Figure 4.6.). The pattern of the ratio depending on circle radius is not clear, more specifically, all radii except 3.70 cm has statistically same adjustment ratio. Compared to the effect of ICD on the adjustment ratio, the effect of circle

radius was minor. In addition to the main effects of the ICD and circle radii factors, the interaction was significant: $F(4, 72) = 14.029, p < 0.001$ (see Figure 4.7.).

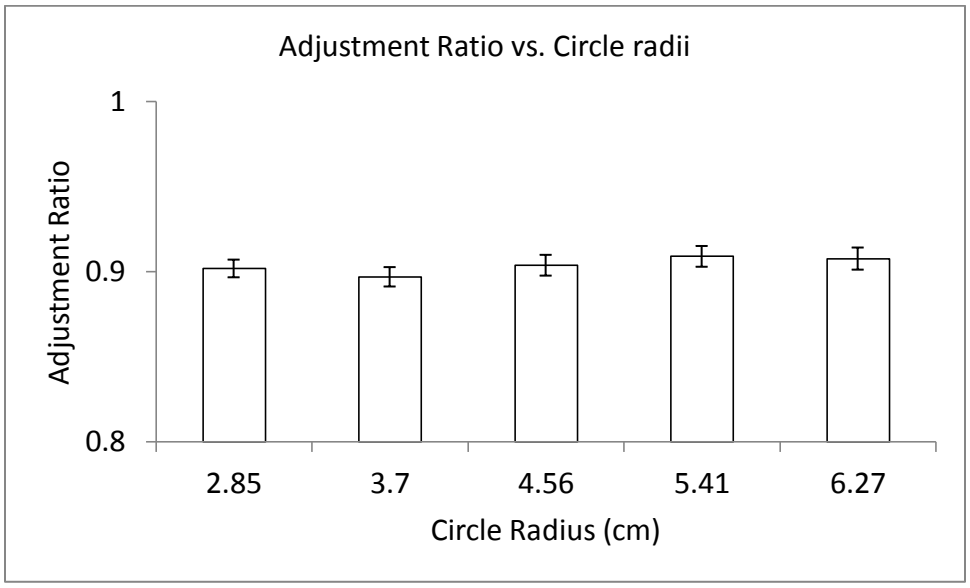

Figure 4.6. Adjustment radius ratio for the circle radii factor (N = 19). Error bar represents standard error. Note that the vertical axis starts from .8 and ends at 1.0. There is no clear trend to demonstrate if the adjustment ratio is changing monotonically with the circle radii factor.

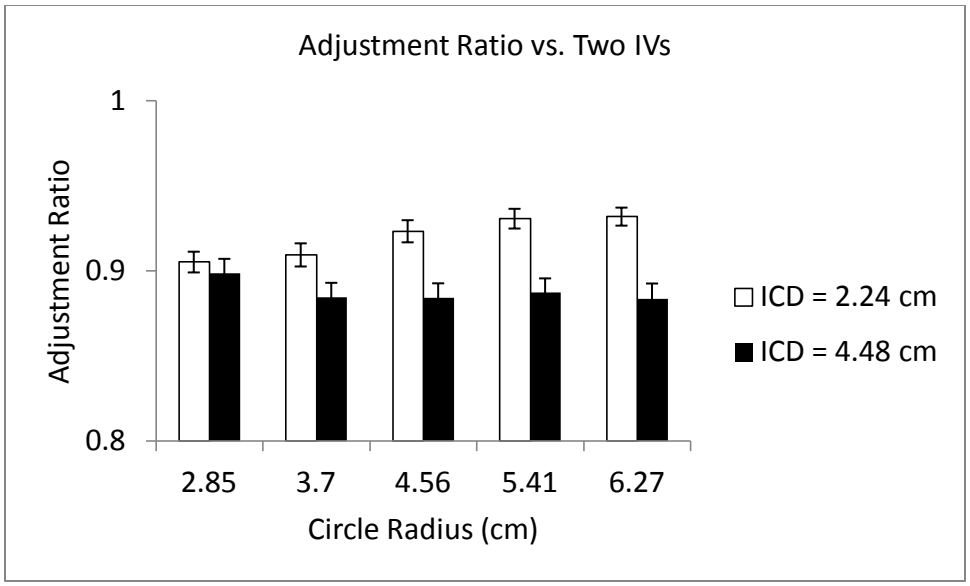

Figure 4.7. Adjustment radius ratio vs. two independent variable (IVs): ICD and circle radii (N = 19). Error bar represents standard error. Note that the vertical axis starts from .8 and ends at 1.0. It is clear that the adjustment ratio decreases as ICD gets larger, whereas there is no clear trend to demonstrate if the adjustment ratio is changing monotonically with the circle radii factor.

As we have discussed, the adjustment ratio can be used to compute the perceived cylinder depth by assuming that the center of the perceived cylinder is at a known viewing distance. From equation (4.2.) it is clear that the viewing distance serves as a constant, and we assumed that the viewing distance is identical to the physical viewing distance from the observer to the monitor screen, which is 40 cm. The converted cylinder depth for the 17 participants was $4.094 \pm .118$ (cm). The converted cylinder depth for trials with ICD = 2.24 cm was $3.566 \pm .126$ (cm), and the ratio for trials with ICD = 4.48 cm was $4.822 \pm .171$ (cm). The effect of the ICD was highly significant: $F(1, 18) = 55.551$, $p < 0.001$. The converted cylinder depth for trials with fixed

circle radius 2.85 cm was 4.*182* ± .*232* (cm), for trials with radius 3.70 cm was 4.*397* ± .*256* (cm), for trials with radius 4.56 cm was 4.*096* ± .*273* (cm), for trials with radius 5.41 cm was 3.*865* ± .*273* (cm), and for trials with radius 6.27 cm was 3.*930* ± .*292* (cm). The effect of the radius was significant: $F(4, 72) = 2.882$, $p = 0.028$ (see Figure 4.8.). Compared to the effect of ICD on the converted cylinder depth, the effect of circle radius was minor. The converted cylinder depth increases as ICD gets larger, whereas there is no clear trend to demonstrate if the converted cylinder depth is changing monotonically with the circle radii factor. In addition to the main effects of the ICD and circle radii factors, the interaction was significant: $F(4, 72) = 13.581$, $p < 0.001$.

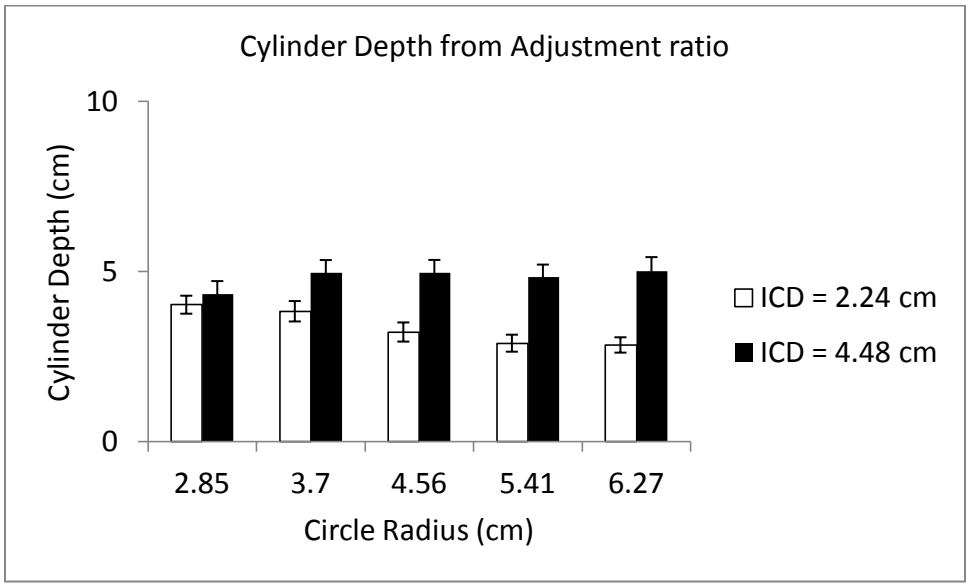

Figure 4.8. Converted cylinder depth (in centimeters) vs. two independent variable (IVs): ICD and circle radii (N = 19). The center of the perceived cylinder was assumed to be at the physical

viewing distance of 40 cm. Error bar represents standard error. Note that the vertical axis starts from 0 and ends at 10 cm. It is clear that the converted cylinder depth increases as ICD gets larger, whereas there is no clear trend to demonstrate if the converted cylinder depth is changing monotonically with the circle radii factor.

In addition to the ratio adjustment measurement, 17 of the 19 participants (7 females, 10 males) did the task of pointing to both ends of the perceived cylinder after ratio adjustment in each trial. The observers' finger tip positions (more specifically, the LED attached to their index finger tip) were recorded as a measurement of the perceived distance from themselves to the stimuli. The pointed front circle distance for the 17 participants was 31.*669* ± .*652* (cm). The pointed front circle distance for trials with ICD = 2.24 cm was 32.*267* ± .*938* (cm), and the pointed distance for trials with ICD = 4.48 cm was 31.*072* ± .*906* (cm). The effect of the ICD was significant: *F(1, 16) = 8.187, p = 0.011*. The pointed front circle distance for trials with fixed circle radius 2.85 cm was 32.*415* ± *1.480* (cm), for trials with radius 3.70 cm was 31.*750* ± *1.410* (cm), for trials with radius 4.56 cm was 31.*775* ± *1.545* (cm), for trials with radius 5.41 cm was 31.*221* ± *1.437* (cm), and for trials with radius 6.27 cm was 31.*185* ± *1.489* (cm). The effect of the radius was significant: *F(4, 64) = 3.910, p = 0.007*. And there was no interaction. The larger the ICD the closer the pointed front circle distance and the larger the circle radius the smaller the distance, but all the difference are small, that is, the deviation of distance in different conditions are around 1 cm. Given the accuracy of the PPT recording system (±5 cm,), we can see that there was no very big variance in the pointed distance for the front circle.

The pointed back circle distance for the 17 participants was 41.*326* ± .*729* (cm). The pointed front circle distance for trials with ICD = 2.24 cm was 40.*951* ± *1.045* (cm), and the pointed distance for trials with ICD = 4.48 cm was 41.*700* ± *1.020* (cm). The effect of the ICD was not significant, though the trend was the distance was larger when the ICD was larger. The pointed back circle distance for trials with fixed circle radius 2.85 cm was 41.*913* ± *1.672* (cm), for trials with radius 3.70 cm was 41.*054* ± *1.671* (cm), for trials with radius 4.56 cm was 41.*798* ± *1.691* (cm), for trials with radius 5.41 cm was 41.*227* ± *1.536* (cm), and for trials with radius 6.27 cm was 40.*636* ± *1.656* (cm). The effect of the radius was not significant and there was no interaction.

The average of the pointed distance for front circle and that for the back circle can be used as a measurement of the viewing distance from the observer to the center of the perceived cylinder. The viewing distance for the 17 participants was 36.*474* ± .*669* (cm), indicating that the center of the perceived was actually closer to the participant compared to the physical viewing distance of 40 cm (*t(16) = 5.274, p < .001*). The viewing distance with ICD = 2.24 cm was 36.*609* ± .*969* (cm), and the pointed viewing distance with ICD = 4.48 cm was 36.*338* ± .*927* (cm). The effect of the ICD was not significant. The viewing distance for trials with fixed circle radius 2.85 cm was 37.*164* ± *1.529* (cm), for trials with radius 3.70 cm was 36.*402* ± *1.489* (cm), for trials with radius 4.56 cm was 36.*787* ± *1.573* (cm), for trials with radius 5.41 cm was 36.*105* ± *1.424* (cm), and for trials with radius 6.27 cm was 35.*911* ± *1.536* (cm). The effect of the radius was significant: *F(4, 64) = 3.282, p = 0.016*. The fact that the larger the circle radius the smaller the distance indicated that participants tend to perceive larger circles to be closer to themes. But still

all the difference was small relative to the recording system recording accuracy of ±.5 cm. And there was no interaction.

The difference of the pointed distance for back circle and that for the front circle can be used as a measurement of the perceived cylinder depth. The perceived cylinder depth for the 17 participants was 9.*752* ±.*341* (cm), which is longer than the converted depth from the adjustment ration of 4.*094* ±.*118* (cm). It is noteworthy that the converted depth was calculated using the physical viewing distance of 40 cm, and if we use the viewing distance from the average pointing distance of the front and back circle in the experiment which was 36.*474* ±.*669* (cm), the converted cylinder depth will be even shorter than 4.*094* ±.*118* (cm). The perceived cylinder depth with ICD = 2.24 cm was 8.*684* ±.*430* (cm), and the perceived cylinder depth with ICD = 4.48 cm was 10.*820* ±.*505* (cm). The effect of the ICD was significant: $F(1, 16) = 27.570$, $p < 0.001$. The perceived cylinder depth for trials with fixed circle radius 2.85 cm was 9.*498* ±*788* (cm), for trials with radius 3.70 cm was 9.*305* ±.*830* (cm), for trials with radius 4.56 cm was 10.*023* ±.*775* (cm), for trials with radius 5.41 cm was 10.*484* ±.*744* (cm), and for trials with radius 6.27 cm was 9.*451* ±.*691* (cm). The effect of the radius was significant: $F(4, 64) = 2.650$, $p = 0.041$ (Figure 4.9.). There was no clear pattern showing how the perceived cylinder depth changed monotonically with the circle radii. And there was no interaction.

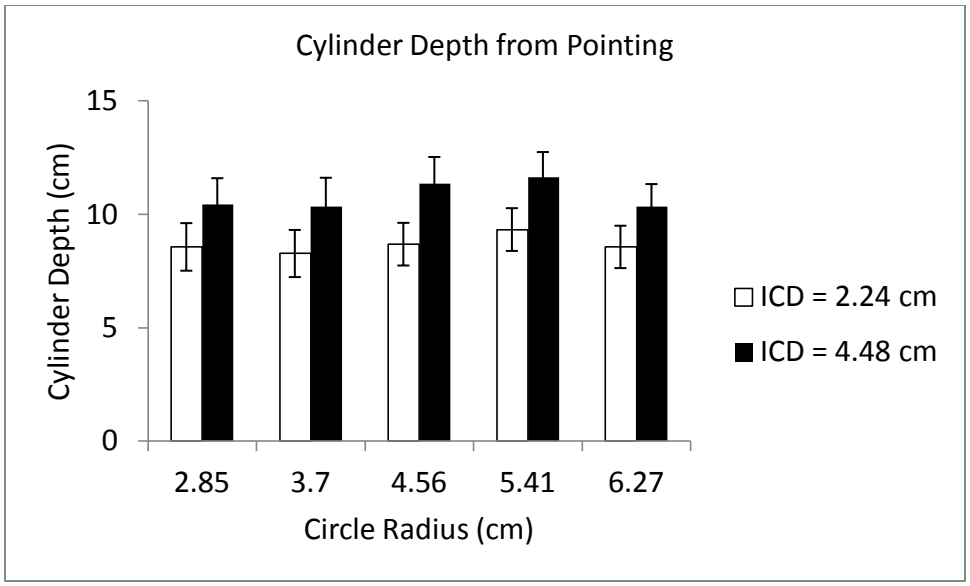

Figure 4.9. Perceived cylinder depth (in centimeters) from pointing vs. two independent variable (IVs): ICD and circle radii (N = 17). Error bar represents standard error. Note that the vertical axis starts from 0 and ends at 15 cm. It is clear that the perceived cylinder depth increases as ICD gets larger, whereas there is no clear trend to demonstrate if the converted cylinder depth is changing monotonically with the circle radii factor.

4.2.2. Experiment 2: Radius ratio adjustment and stereo matching

4.2.2.1. Participants

27 students from UCLA participated in the experiment.

4.2.2.2. Apparatus and stimuli

The computer graphical stimuli are rendered using OpenGL and PsychToolBox 3 in MATLAB environment. Stimuli are displayed on a 20 inch CRT monitor with resolution of 1280 × 960 pixels and refresh rate of 75 Hz. A vision cube is used to mask of the edges of the screen, and a chin rest is also used to fix the viewing distance at 40 cm.

The participant observed the stimuli through a viewing tube and the participant's head was restricted by a chinrest. The participant wore a pair of red/green stereoscopic goggle, more specifically, in front of the participant's right-eye there was a green light pass filter and in front of the participant's left eye there was a red light pass filter. On the monitor screen 40 cm from the observer, two green circles were displayed on a black background, rotating around a common axis. One of the two circles was always of a fixed size, and the size was randomly chosen from two possible radii. A red circle was on the right of the green fixed size circle, with the same radius as its green counterpart, so that the circle pair formed a stereographical circle. The center for the red circle is vertically at the same level as the fixed size green circle whereas horizontally .337 cm to the right. So the perceived viewing distance of the fixed size stereo circle can be obtained from:

$$h = \left(\frac{d}{d_i - d} \right) D \qquad (4.3.)$$

In equation (4.3.), d_i is the interpupillary distance (IPD), and we used the average IPD for the 27 participants which equals to 6.04 cm in our analysis. d is the binocular stereo disparity of the green and red circle, which was -.037 cm (without loss of generality and for the convenience in following derivations, we defined the red on the right to be negative). D is the viewing distance from the observer to the monitor screen, and it was 40 cm in this experiment. So the fixed size cylinder is perceived to be 2.11 cm in front of the monitor screen, which is 37.89 cm from the observer. The radius of the other circle is rendered by adding to the fixed radius a random real number (may be positive or negative). This circle can only be seen monocularly by the observer's right eye because of the red pass filter on the left eye.

The green circular pair stimulus when measured through the red pass filter had the CIE coordinate of (0.05, .51, .32). When measured through the green pass filter the luminance was 0.19 cd/m^2 with color coordinate (.19, .58). The green diamonds stimuli when measured through the red pass filter had CIE coordinate of (0.05, .46, .32). When measured through the green pass filter the luminance was 0.19 cd/m^2 with color coordinate (.20, .64). The red diamonds stimuli when measured through the red pass filter, the CIE coordinate was (0.19, .66, .30). When measured through the green pass filter the luminance was 0.06 cd/m^2 with color coordinate (.23, .42). The photometer reading of the background was 0.03 cd/m^2 with color (.50, .27) through red pass filter and 0.05 cd/m^2 with color (.05, .68) through green pass filter. The measurements were made using the Minolta CS-100 photometer.

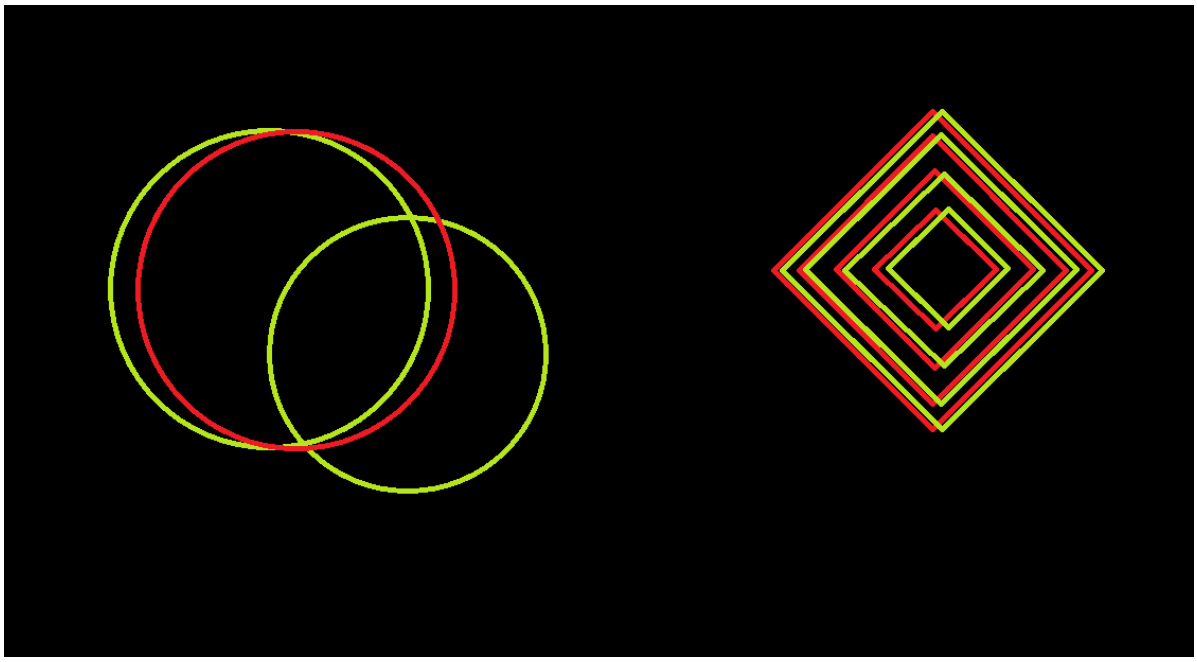

Figure 4.10. Experiment stimuli used in Experiment 2. The two circles were rotating along a common axis perpendicular to the image plane going through the mid-way of the two circular centers. The circle with fixed radius had a red counterpart on its right with fixed stereo disparity of .337 cm. With the viewing distance from the observer to the screen to be 40 cm, the fixed size stereographic circle was perceived to be 37.89 cm away from the observer. It is noteworthy that during the ratio adjustment task, only the circles were present. After the participant had finished the ratio adjustment so that a uniform cylinder was perceived, they pressed a button so that the stereographical diamond set on the other side of the screen was displayed. Participants then adjusted the binocular stereo disparity between the green and red counterparts in the diamond set to first match the distance of the cylinder's front end then that of the cylinder's back end. The

binocular stereo disparity of the diamond set was recorded as a measurement of the perceived viewing distance of the cylinder's two ends.

4.2.2.3. Procedure

In each experimental trial, participants first adjusted the size of the variable circle so that a uniform cylinder was perceived. It was emphasized to the participants by instructions that, by a uniform cylinder, we meant that the two circles were perceived to be of identical size in 3D (they were perceived to be at different distance from the participants). More specifically, because of the 3D percept from stereokinetic effect, two circles of identical size but at different distance from observer will be have different sizes on the retina because of linear perspective projection (more generally, on any projection plane perpendicular to the participant's sightline). Besides, all participants were asked to indicate if the variable circle is perceived to be in front (close to the observer) or at back. All the participants reliably indicated that the larger circle was always in front, and as a result we define the larger circle to be the one in front and the smaller one at back in our analysis.

After the participant had finished the ratio adjustment task so that a uniform cylinder was perceived, they were asked to press a button to bring out a binocular stereographical diamond set on the other side of the screen (Figure 4.10.). The participants were able to adjust the binocular disparity between the green and red counterparts in the stereograph to make the diamond to be

perceived as closer to the participant or further away. The participants first adjusted the diamond set so that it matches the distance of perceived cylinder's front end, then they repeated the same task but to match the distance of the perceived cylinder's back end. The binocular stereo disparity was recorded as a measurement of the perceived distance of the cylinder's two ends. There are 60 trials in the experiment, and on average it took the participant about 45 minutes to finish the test.

4.2.2.4. Results

The adjustment ratio for the 27 participants was *.83 ± .01*. When the fixed circle radius was 2.85 cm, the adjustment ratio for ICD = 1.12 cm was *.85 ± .01*, and for ICD = 3.36 cm the adjustment ratio was *.82 ± .01*. And the difference was significant: $t(26) = 2.70$, $p = 0.01$. When the fixed circle radius was 6.27 cm, the adjustment ratio for ICD = 2.24 cm was *.83 ± .01*, and that for ICD = 4.48 cm was *.81 ± .01*. And the difference was also significant: $t(26) = 2.11$, $p = 0.04$. The results showed that with the same circle radius, the larger the ICD, the smaller the adjustment ratio which is consistent with the results obtained in Experiment 1.

Given that the stereographical circle was 37.89 cm away from the observer, the converted cylinder depth can be calculated using equation (4.2.). For the 27 participants the converted cylinder depth was *8.33 ± .35* (cm). When the fixed circle radius was 2.85 cm, the converted cylinder depth for ICD = 1.12 cm was 6.95 *± .55* (cm), and that for ICD = 3.36 cm was *8.62* ±

.88 (cm). And the difference was significant: *t(26) = 2.35, p = 0.03*. When the fixed circle radius was 6.27 cm, the converted cylinder depth for ICD = 2.24 cm was *8.46 ± .66* (cm), and that for ICD = 4.48 cm was *9.28 ± .65* (cm). The difference in the circle radius = 6.27 cm condition was not significant.

The distance from the observer to the front end of the cylinder for the 27 participants was *37.62 ± .29* (cm), which is statistically the same as the theoretically anticipated viewing distance of 37.89 cm. This result indicated that all the participants were capable of matching the distance of the stereographical circle with the distance of the stereographical diamond set. When the fixed circle radius was 2.85 cm, the converted cylinder depth for ICD = 1.12 cm was *37.63 ± .22* (cm), and for ICD = 3.36 cm was 3*7.25 ± .34* (cm). The difference between viewing distances in two ICDs is not significant. When the fixed circle radius was 6.27 cm, the converted cylinder depth for ICD = 2.24 cm was 3*7.98 ± .49* (cm), and for ICD = 4.48 cm was 37.*60 ± .38* (cm). The difference between viewing distances in two ICDs is not significant either. These results demonstrated that the participants' stereo vision is reasonably good and they can match the distance of cylinder's front end with the diamond set very well.

When analyzing the viewing distance of the back end of the cylinder, we found that 4 of the 27 participants had trouble matching the viewing distance of the diamond set and the back circle, more specifically, they adjusted the viewing distance of the diamond to be more than 100 cm away from the observer. The data from these 4 participants were excluded because they were not able to match the viewing distance of the cylinder's back end. The distance from the observer to

the back end of the cylinder for the remaining 23 participants was *54.56 ± 2.02* (cm). When the fixed circle radius was 2.85 cm, the converted cylinder depth for ICD = 1.12 cm was *50.16 ± 1.35* (cm), and for ICD = 3.36 cm was 53*.41 ± 2.56* (cm). The difference between viewing distances in two ICDs is not significant. When the fixed circle radius was 6.27 cm, the converted cylinder depth for ICD = 2.24 cm was 57*.06 ± 2.57* (cm), and for ICD = 4.48 cm was 57*.59 ±2 .88* (cm). The difference between viewing distances of the cylinder's back end in two ICDs is not significant. The trend was that if the circle radius was the same, the perceived viewing distance of the cylinder's back end was further away with larger ICD.

The difference between the viewing distance of the back end and that of the front end of the cylinder is the measurement of the depth of the perceived cylinder. The data from those 4 participants excluded in the previous analysis was also excluded in the perceived depth analysis. The perceived cylinder depth for the remaining 23 participants was *17.13 ± 2.09* (cm). When the fixed circle radius was 2.85 cm, the converted cylinder depth for ICD = 1.12 cm was *12.58 ± 1.41* (cm), and for ICD = 3.36 cm was 16*.15 ± 2.72* (cm). The difference between viewing distances in two ICDs is not significant. When the fixed circle radius was 6.27 cm, the converted cylinder depth for ICD = 2.24 cm was 19*.48 ± 2.61* (cm), and for ICD = 4.48 cm was 20*.31 ±2 .93* (cm). The difference between viewing distances of the cylinder's back end in two ICDs was not significant. The trend was that if the circle radius was the same, the perceived cylinder depth was longer with larger ICD.

4.3. Computational model

4.3.1. The model

Without loss of generality, we put the two circles on the x-y plane and assume that one of the two circles is centered at origin (*0, 0, 0*) with radius r_1 whereas the other circle is centered at (*0, d, 0*) with radius r_2. The two circles are rotating around the z axis at rotational speed ω and the observer is looking from position (*0, d, - D*). The circle centered at (*0, d, 0*) was perceived to be at distance h from the image plane (see Figure 4.4. for more details). Because of linear perspective projection, the perceived circle is centered at (*0* , $\left(\frac{h}{D}+1\right) d$, *h*), and with radius:

$$r_p = \left(\frac{h}{D}+1\right) r_2 \qquad (4.4.)$$

The circle on the x-y plane can be described as:

$$\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} r_1 \cos\alpha \\ r_1 \sin\alpha \\ 0 \end{pmatrix} \qquad (4.5.)$$

There is a rotation ω around the z axis, so considering the rotation:

$$\begin{pmatrix} x_1(t) \\ y_1(t) \\ z_1(t) \end{pmatrix} = \begin{pmatrix} \cos\omega t & -\sin\omega t & 0 \\ \sin\omega t & \cos\omega t & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} r_1 \cos\alpha \\ r_1 \sin\alpha \\ 0 \end{pmatrix} \tag{4.6.}$$

By taking the temporal derivative of Equation 4.3., we get the motion of the first circle:

$$\begin{pmatrix} v_{1x}(t) \\ v_{1y}(t) \\ v_{1z}(t) \end{pmatrix} = \begin{pmatrix} -r_1\omega\sin\omega t\cos\alpha - r_1\omega\cos\omega t\sin\alpha \\ r_1\omega\cos\omega t\cos\alpha - r_1\omega\sin\omega t\sin\alpha \\ 0 \end{pmatrix} \tag{4.7.}$$

The other circle perceived away from the x-y plane can be described as:

$$\begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} = \begin{pmatrix} r_p \cos\alpha \\ r_p \sin\alpha + \left(\frac{h}{D}+1\right)d \\ h \end{pmatrix} \tag{4.8.}$$

Similarly, there is a rotation ω around the z axis, so considering the rotation:

$$\begin{pmatrix} x_2(t) \\ y_2(t) \\ z_2(t) \end{pmatrix} = \begin{pmatrix} \cos\omega t & -\sin\omega t & 0 \\ \sin\omega t & \cos\omega t & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} r_p \cos\alpha \\ r_p \sin\alpha + \left(\frac{h}{D}+1\right)d \\ h \end{pmatrix} \tag{4.9.}$$

By taking the temporal derivative of Equation 4.6., we get the motion of the perceived circle:

$$\begin{pmatrix} v_{2x}(t) \\ v_{2y}(t) \\ v_{2z}(t) \end{pmatrix} = \begin{pmatrix} -r_p\omega\sin\omega t\cos\alpha - r_p\omega\cos\omega t\sin\alpha - \left(\frac{h}{D}+1\right)d\omega\cos\omega t \\ r_p\omega\cos\omega t\cos\alpha - r_p\omega\sin\omega t\sin\alpha - \left(\frac{h}{D}+1\right)d\omega\sin\omega t \\ 0 \end{pmatrix} \quad (4.10.)$$

We proposed that the motion and structure interpretation in 3D that gives rise to the slowest and spatially smoothest motion is perceived by the visual system. Mathematically, we define the loss function of a certain motion interpretation as:

$$E(\vec{v}) = \int \sum_{m=0}^{\infty} c_m \left|\nabla^m \vec{v}\right|^2 \quad (4.11.)$$

We calculated the loss function including the 0 order cost and the 1st order cost. The 0 order cost is the magnitude of motion in 3D, or a measurement of the slowness of the motion. The 1st order cost is the first order gradient of the motion in x, y, and z direction, which is a measurement of the spatial smoothness of the motion. The constant c_m controls the relative strength of the slowness term and the spatially smoothness term's contribution to the loss function. We followed Yuille and Grzywacz's (1988) definition on the constant, which was widely used in the computational work on motion perception:

$$c_m = \frac{\lambda^{2m}}{m!2^m} \quad (4.12.)$$

We used $\lambda = 2$, which in the range of the appropriate parameter setting (Yuille and Grzywacz, 1989).

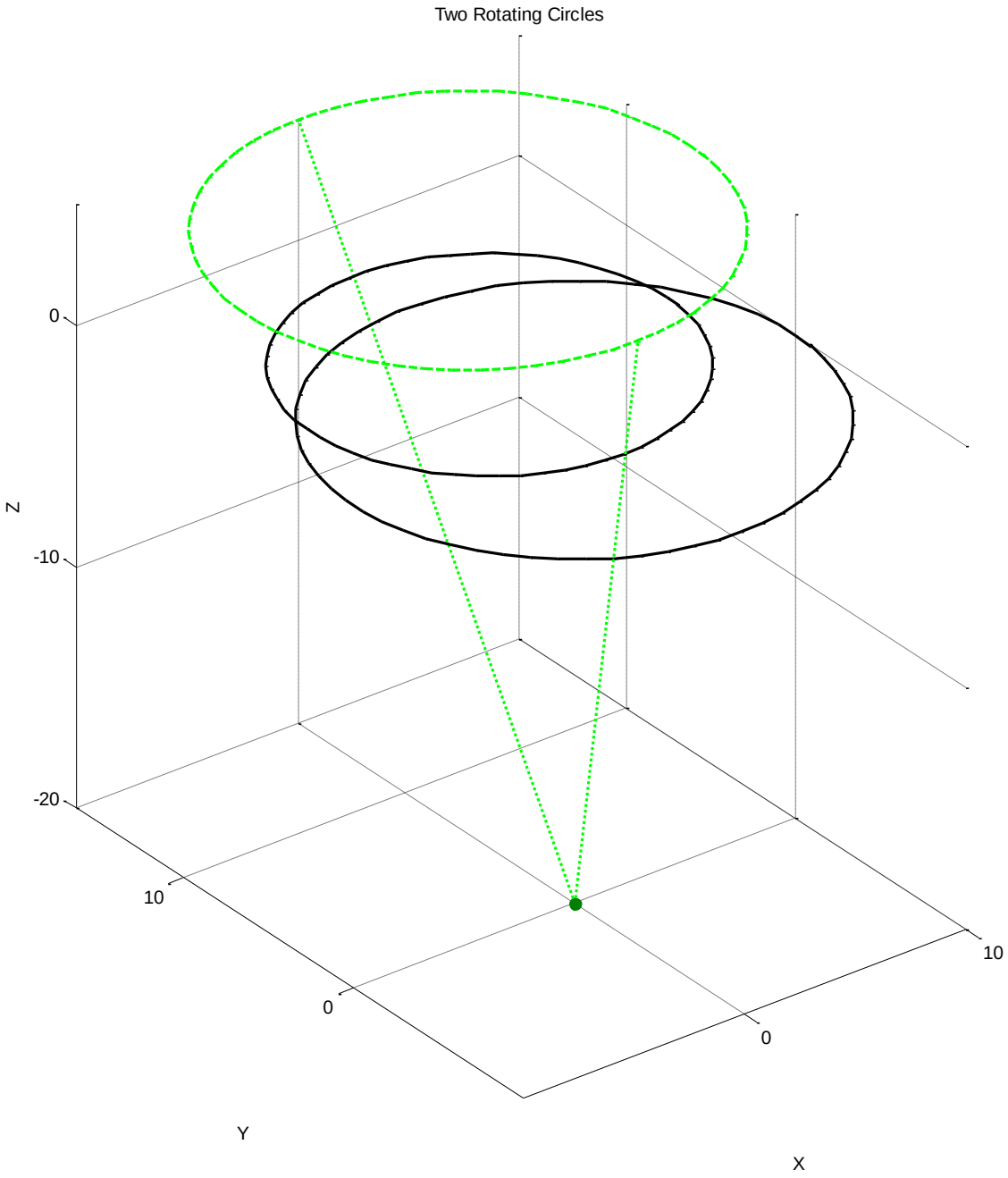

Figure 4.11. Schematic illustration of the perception from two rotating circles. The two circles are on the x-y plane: the circle with stereo disparity information (larger black solid circle) is centered at origin (*0, 0, 0*) whereas the other circle (smaller black solid circle) is centered at (*0, d, 0*). The two circles are rotating around the z axis at rotational speed ω and the observer (the

black dot at the bottom) is looking from position (*0, 0, - D*). The smaller back circles was perceived to be a projection of the hypothetical circle at distance h from the image plane (green dashed circle). The dashed green line starting from the observer depicts the linear perspective projection. The center of the hypothetical circle is at (0 , $\left(\frac{h}{D} + 1\right) d$, h).

Before looking at the simulation results, here we explained the intuition of the computational model. The optimal motion and structure interpretation is the one that gives rise to the global minimum of the loss function. First for simplicity, let us assume that the loss function only takes into account the motion slowness term, and that means only the magnitude of the motion is considered in the computational model. From equation (4.7.) and equation (4.10.), we can see that the magnitude of motion increase as the perceived cylinder depth (h) increases. As a result, the globally optimal solution will be a cylinder that has zero depth. On the other hand, if we assume that the loss function only takes in to account the motion smoothness term, and that means only the first order gradient was considered by the computational model. By definition it is clear that when the motion field is the same, the longer the cylinder depth (h), the smaller the first order motion gradient in the z direction, and thus the smaller then first order cost term. As a result, the globally optimal solution will be a cylinder that has infinite depth. To summarize, motion slowness constraint term will prefer a cylinder that is as short as possible, but such short cylinder will have spatially non-smooth motion field (in other words, the corresponding structure form such motion interpretation will be non-rigid) and as a result such interpretation is not preferred by the motion smoothness constraint term. Since the motion slowness and spatially smoothness terms are both taken into account in the computational model, a balance between the

two constraints will lead to an interpretation that is neither a cylinder of 0 depth nor one with infinite depth, but one with a definite depth.

4.3.2. Model predictions

We investigated the model predicted cylinder depth as a function of circle radius and ICD. Figure (4.12.) showed the loss function value (vertical axis) as a function of the interpreted cylinder depth (horizontal axis). As we have discussed, the optimal cylinder depth that gives rise to the global minimum is neither zero depth nor infinite depth, but a definitely depth of 3.0 cm in this case (circle radius = 2.85 cm, ICD = 1.12 cm, viewing distance = 40.00 cm). It is noteworthy that the loss function decreases quickly as the cylinder depth (h) increases from 0, a result mainly because of the quick improving of spatial smoothness of the interpreted motion field. As the interpreted cylinder depth gets longer, the spatial smoothness change less dramatically compared to increasing from $h = 0$. Figure (4.13.) showed the loss function value as a function of h when the circle radius = 2.85 cm, ICD = 1.12 cm and viewing distance = 40.00 cm, and the computational model predicts the cylinder depth of 8.20 cm.

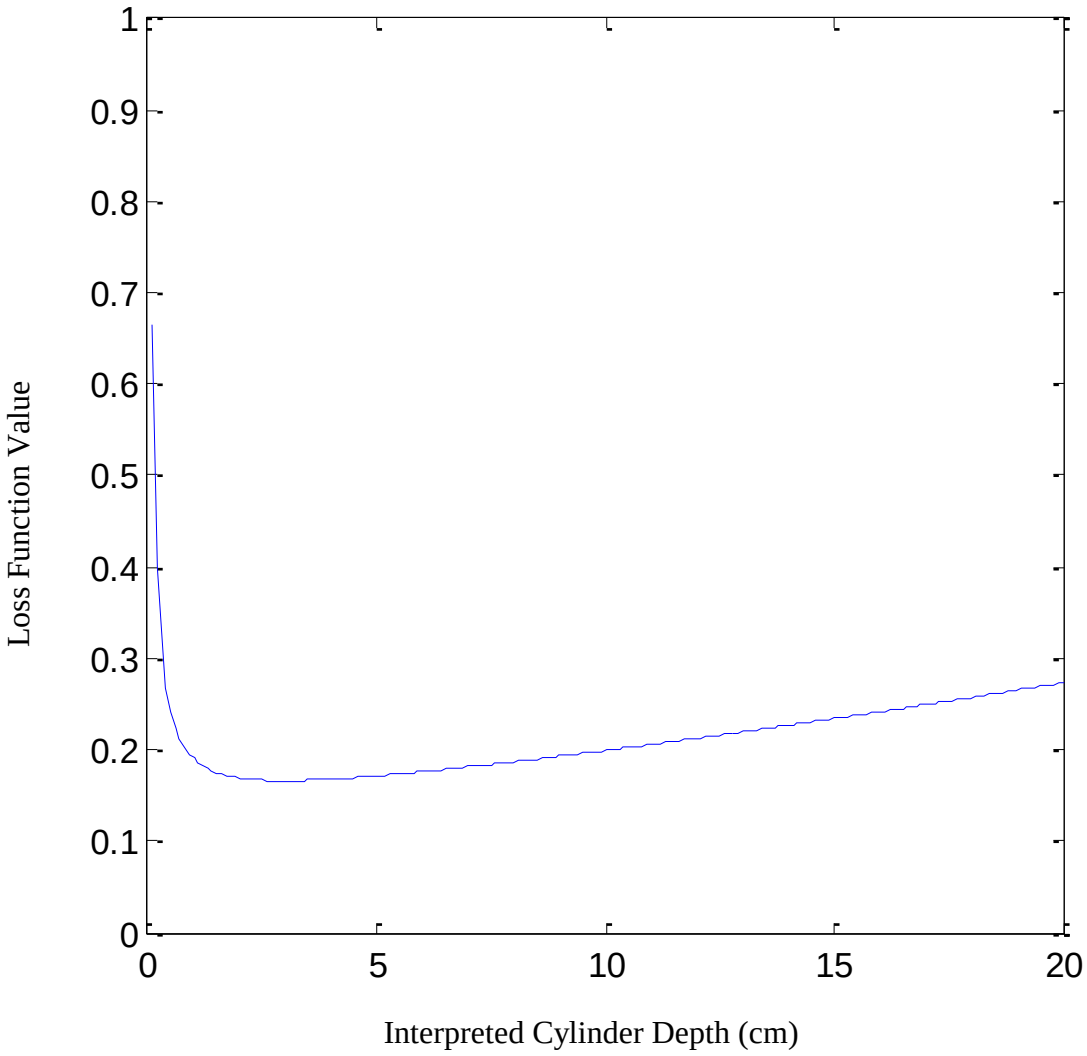

Figure 4.12. Loss function value as a function of the interpreted cylinder depth. The vertical axis represented the loss function value and the horizontal axis is the interpreted cylinder depth. Circle radius = 2.85 cm, ICD = 1.12 cm, viewing distance = 40.00 cm, and $\lambda = 2$. The model predicts that the optimal cylinder depth is $h_{model} = 3.00$.

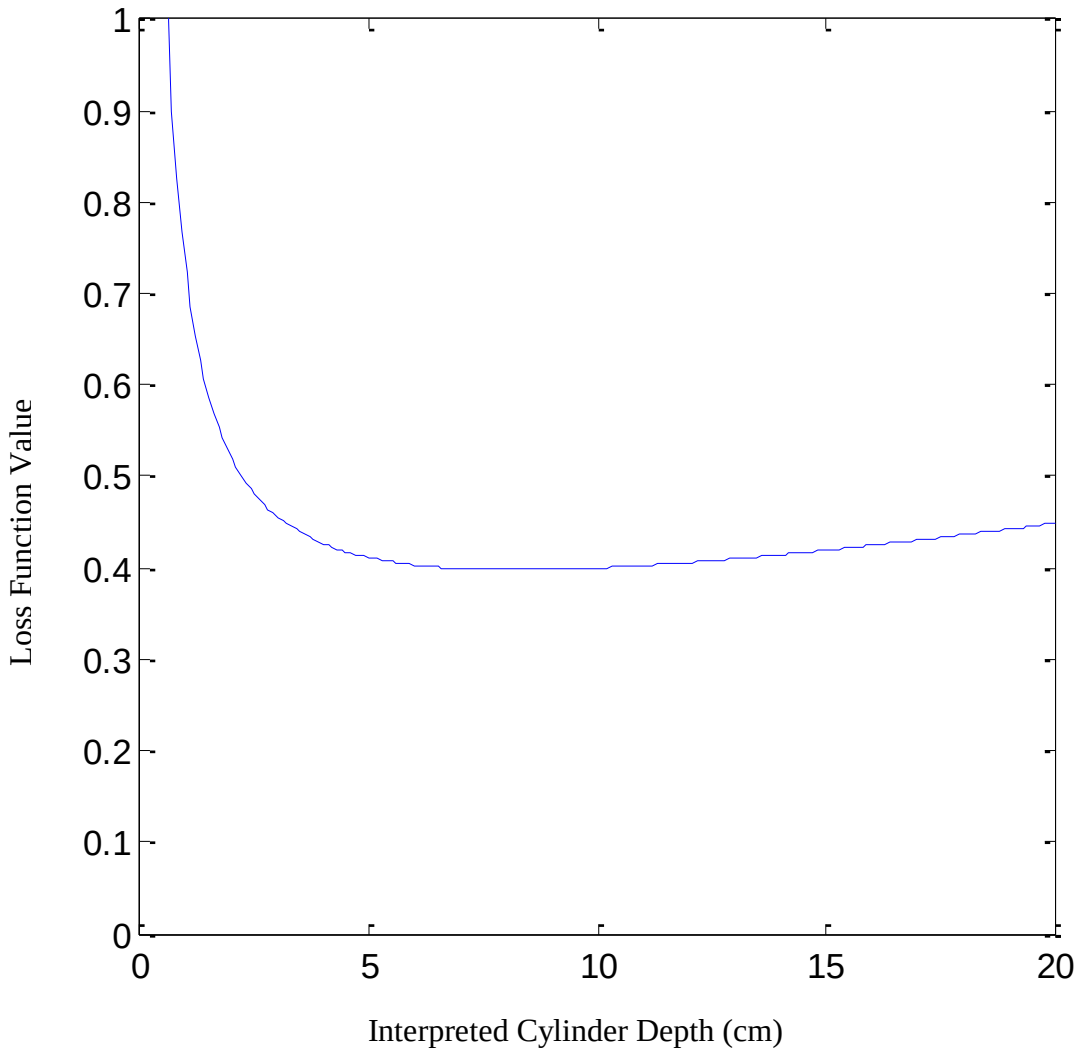

Figure 4.13. Loss function value as a function of the interpreted cylinder depth. The vertical axis represented the loss function value and the horizontal axis is the interpreted cylinder depth. Circle radius = 2.85 cm, ICD = 3.36 cm, viewing distance = 40.00 cm, and $\lambda = 2$. The model predicts that the optimal cylinder depth is $h_{model} = 8.20$.

Figure (4.14.) showed the loss function value as a function of the interpreted *h*. Similarly, the optimal cylinder depth that gives rise to the global minimum is neither zero depth nor infinite depth, but a definitely depth of 2.90 cm in this case (circle radius = 6.27 cm, ICD = 2.24 cm, viewing distance = 40.00 cm). Figure (4.15.) showed the loss function value as a function of *h* when the circle radius = 6.27 cm, ICD = 4.48 cm and viewing distance = 40.00 cm, and the computational model predicted a cylinder depth of 5.40 cm. Again, when the circle radius are the same (both are 6.27 cm), the computational model predicts longer cylinder with larger ICD.

In Experiment 1, empirical data showed that ICD had a major impact on the converted cylinder depth ($F(1, 18) = 55.551, p < 0.001$), and larger ICD leads to longer converted cylinder. In Experiment 2, when the fixed circle radius was 2.85 cm, the converted cylinder depth for ICD = 3.36 cm (*8.62* ± *.88* cm) was significantly longer ($t(26) = 2.35, p = 0.03$) than the depth for ICD = 1.12 cm (6.95 ± *.55* cm). When the fixed circle radius was 6.27 cm, the converted cylinder depth for ICD = 4.48 cm (*9.28* ± *.65* cm) was also longer than that for ICD = 2.24 cm (*8.46* ± *.66* cm), though the difference was not statistically significant. Note that when the circle radius is the same, the computational model predicts longer cylinder with larger ICD, which is qualitatively consistent with the empirical findings obtained in Experiment 1 and 2.

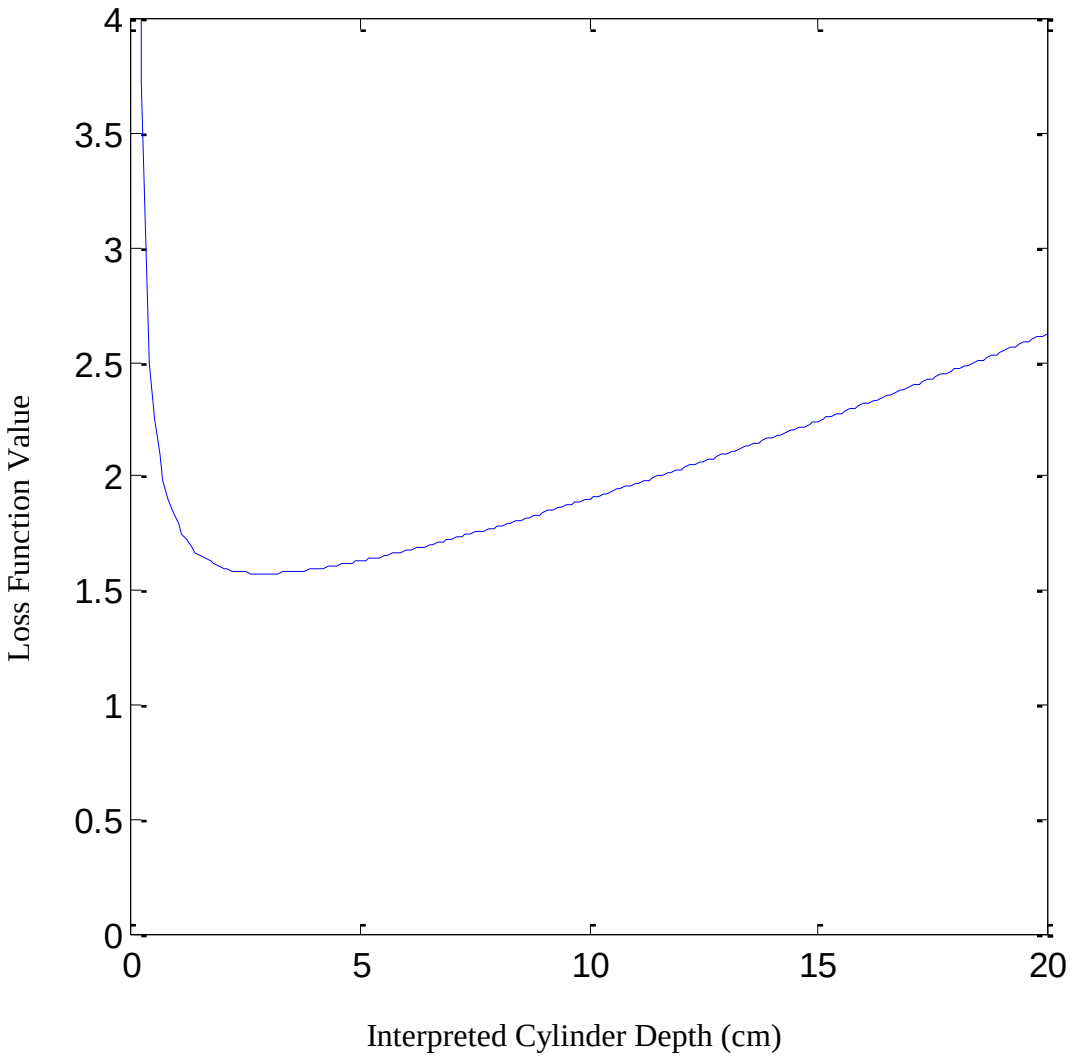

Figure 4.14. Loss function value as a function of the interpreted cylinder depth. The vertical axis represented the loss function value and the horizontal axis is the interpreted cylinder depth. Circle radius = 6.27 cm, ICD = 2.24 cm, viewing distance = 40.00 cm, and $\lambda = 2$. The model predicts that the optimal cylinder depth is $h_{model} = 2.90$.

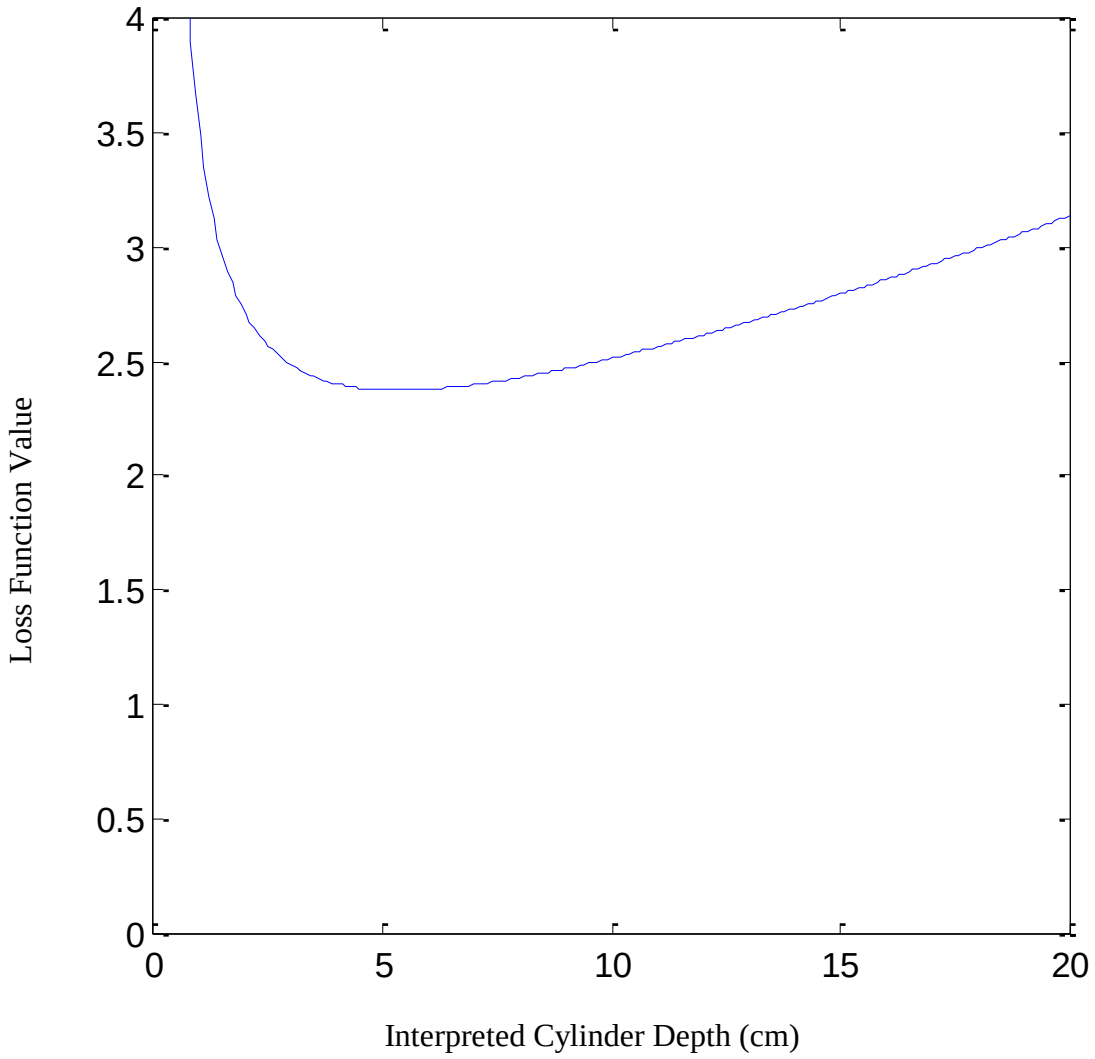

Figure 4.15. Loss function value as a function of the interpreted cylinder depth. The vertical axis represented the loss function value and the horizontal axis is the interpreted cylinder depth. Circle radius = 6.27 cm, ICD = 4.48 cm, viewing distance = 40.00 cm, and $\lambda = 2$. The model predicts that the optimal cylinder depth is $h_{model} = 5.40$ cm.

Another question was how the perceived cylinder depth varies with the circle radius change, or how the visual angle of the stimuli affects the perceived cylinder when the viewing distance is fixed. In Experiment 1, we have found that even though the effect of the circle radius on the converted cylinder depth was significant ($F(4, 72) = 2.882$, $p = 0.028$), but compared to the effect of ICD on the converted cylinder depth ($F(1, 18) = 55.551$, $p < 0.001$), the effect of circle radius was minor. A closer look at the conditions of circle radius = 2.85 cm ($h = 4.182 \pm .232$ cm) and the condition of circle radius = 6.27 cm ($h = 3.930 \pm .292$ cm) showed that the cylinder depth difference under these two conditions was not statistically significant. In Experiment 2, we found that if the circle radius was smaller (2.85 cm), the converted cylinder depth increases more (from $6.95 \pm .55$ cm to $8.62 \pm .88$ cm, or 1.67 cm increase) than if the circle radius was larger (from $8.46 \pm .66$ cm to $9.28 \pm .65$ cm, or .82 cm increase if circle radius = 6.27 cm) when the ICD increases identical amount (increase 2.24 cm). Further analysis showed that if we use the circle radius as a normalization factor, then ICD increase of 2.24 cm was .36 increase of the radius = 6.27 cm and .79 increase of the radius = 2.85 cm. When the ICD increase was .36 of the circle radius the cylinder depth increase was .82 cm and when the ICD increase was .79 radius the depth increase was 1.67 cm. Combining the findings in Experiment 1 and Experiment 2, we hypothesized that the increase of ICD/radius ratio will lead to an increase of the cylinder depth, whereas the circle radius alone would not affect the cylinder depth.

Figure (4.16.) showed the model predicted cylinder depth as a function of ICD/radius ratio under two different circle radius conditions. The horizontal axis is the ICD/radius ratio and the vertical axis is the model predicted cylinder depth. The upper and lower curves are quantitatively close in

shape, indicating that the circle radius alone does not affect the predicted cylinder depth, which is consistent with the empirical findings from Experiment 1 and 2. The predicted cylinder depth increases monotonically as the ICD/radius ratio increases, and the increase is not linear. This is also consistent with the findings on the converted cylinder depth from the two empirical experiments described in this chapter. To summarize, the computational model predicted the characteristics of the human percept we observed in the empirical studies.

It is also noteworthy that we used $\lambda = 2$ in all the simulations in this chapter. A larger λ value will make the loss function receive more penalty from a spatially non-smooth motion field whereas a smaller λ value will make the loss function receive more penalty from the fast motion. Consequently, the computational model with a larger λ value will predict a longer cylinder depth whereas a smaller λ value will predict a shorter cylinder. Nevertheless, different λ values will not change the qualitative predictive power of the computational model on the characteristics of the human percept on this stereokinetic stimulus, although quantitatively the parameter value will affect the simulation results.

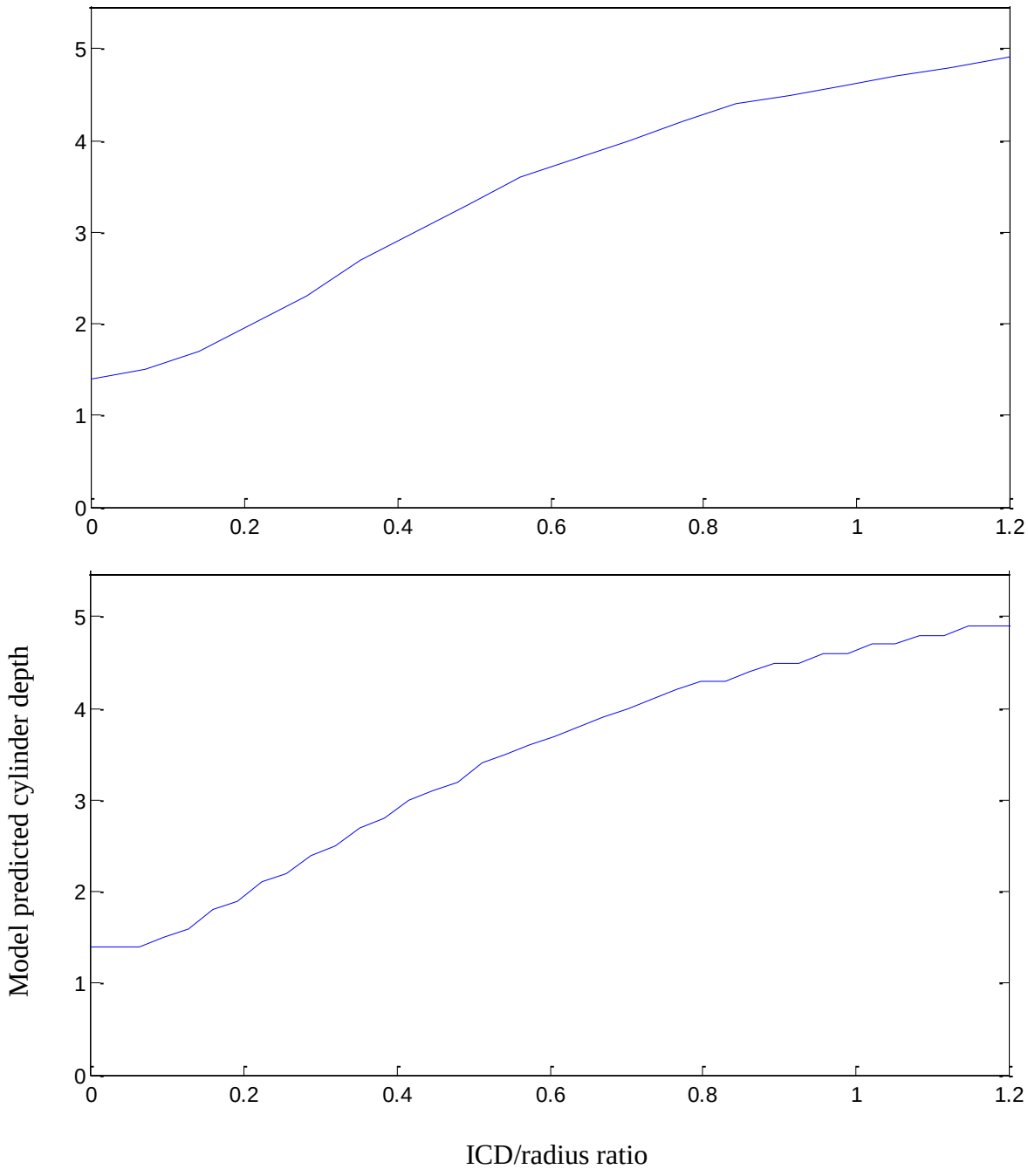

Figure 4.16. Model predicted cylinder depth as a function of the ICD/radius ratio. Viewing distance = 40.00 cm, and λ = 2. Upper: circle radius = 2.85 cm; lower: circle radius = 6.27 cm. The horizontal axis is the ICD/radius ratio and the vertical axis is the model predicted cylinder

depth. The circle radius does not affect the predicted cylinder depth. The predicted cylinder depth increases monotonically as the ICD/radius ratio increases.

4.4. Discussion

In the current study, we quantified a stereokinetic effect, more specifically, we measured the perceived depth of a tilted cylinder consisted of two rotating circles. In Experiment 1, we manipulated the ICD and circle radius as independent variables and asked the participants to adjust the size of one circle so that the two circles forms a uniform cylinder. In addition to using the circle radii ratio to calculate the perceived cylinder depth based on linear perspective projection, we also investigated the participant's capability of pointing to the perceived distance of both ends of the cylinder. Both the circle radii ratio and the finger printing measurements indicated that human observers reliably perceived a tilted 3D cylinder from the stereokinetic stimuli, and the perceived cylinder depth was affected by both the ICD and circle radius, while the effect of circle radius was minor. We also found that the perceived cylinder depth measured by circle radii ratio and linear perspective projection was not quantitatively the same as the perceived cylinder depth measured by the finger printing, and such inconsistency cannot be contributed to the difference between the physical viewing distance and the perceived viewing distance. Previous studies in perception and motor control suggested that the measurement from a perceptual task (circle radius adjustment) is arguably not necessarily identical to the measurement from a visuomotor task (finger printing) on the same stimuli, due to the nature of

the separate visual pathways (Goodale and Milner, 1992; Króliczak et al., 2006; Knill, 2005; Hartung et al., 2005). In Experiment 2, we modified the stimuli used in Experiment 1 so that the perceived viewing distance is well defined and the complication from motor control system was ruled out by introducing another perceptual task to measure the perceived cylinder depth. Again we found that circle radius ratio data showed that longer ICD leads to longer perceived cylinder. The binocular stereo disparity measurement also suggested the same trend, though the result was not significant. To summarize, in Experiment 1 and 2, we designed two perceptual tasks and one visuomotor task to quantify a stereokinetic effect, the measurements from all three tasks had qualitatively identical characteristics.

In addition to quantifying this stereokinetic effect, we also investigated the computational mechanism underlying the phenomenon. We asked the question: Why human observers do not perceive a cylinder of either zero depth, or on the other extreme, of infinite length, but instead a cylinder of a definite depth? These questions touched the underlying essence of the stereokinetic effect's mechanisms. Previous theories mostly focused on using the rigidity prior to explain the stereokinetic effect (Wallach and O'Connell, 1953). However, the rigidity prior cannot explain the phenomenon in this study. More specifically, the rigidity prior can only give rise to a set of rigid truncated cone-shape interpretations, but possible depth range from zero to infinity. To explain the stereokinetic effect investigated in the current study, we developed a computational model from Yuille and Grzywacz's (1988) regularization theory in 2D and Rokers, Yuille, and Liu's (2006) minimum motion principle in 3D. We framed the computational question as a optimization problem, and hypothesized that the 3D structure interpretation that has the minimal

loss function value is the one preferred by the visual system. We formulated the loss function so that it takes into account both the minimum motion and the spatially smooth motion in 3D. So the structure interpretation that gives rise to the slowest and spatially smoothest motion in 3D is preferred by the visual system, and we demonstrated that such preferred structure interpretation will give rise to a definite depth of the perceived cylinder. Our computational model predicted that a definite cylinder depth is the optimal solution, and longer ICD will lead to longer perceived cylinder depth. The ratio between circle radius and ICD on the other hand, does not affect the perceived cylinder depth. These simulation results from our computational model are consistent with the empirical study results obtained in this study, indicating that the visual system is taking into account both the slowness and the spatially smoothness of the motion in 3D to achieve a unique structure interpretation for this stereokinetic stimulus.

In conclusion, we designed three tasks to quantify the human perception on a stereokinetic effect. The empirical results from all three measurements qualitatively converge. In addition to the empirical studies, we also investigated the underlying computational mechanism of stereokinetic effect. We hypothesized that the visual system takes into account the motion slowness and the spatially smoothness to interpret the stereokinetic stimulus. We developed a computational model that can predict all the empirical findings we got in the current study. Our work suggested that the essential mechanism of this stereokinetic effect is that the human visual system prefers a structure interpretation that has a slowest and spatially smoothest motion in 3D.

References

Benussi, V. (1916). Versuche zur Analyse taktil erweckter Scheinbewegungen. *Archly fiir die gesamte Psychologie*. 36: 59-135.

Hartung, B. *et al.* (2005) Is prior knowledge of object geometry used in visually guided reaching? *Journal of Vision.* 5, 504-514.

Gibson, J. (1966). *The Senses Considered as Perceptual Systems,* Houghton-Mifflin, Boston.

Goodale M. & Milner A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences.* 15, 20-25.

Króliczaka, G. *et al.* (2006). Dissociation of perception and action unmasked by the hollow-face illusion. *Brain Research.* 1080, 9-16.

Knill, D.C. (2005). Reaching for visual cues to depth: The brain combines depth cues differently for motor control and perception. *Journal of Vision.* 5, 103-115.

Musatti, C. (1924). Sui fenomeni stereocinetici. *Archivio Italiano di Psicologia*. 3 105-120.

Movshon, J. *et al.* (1985). The analysis of moving visual patterns. *In Pattern Recognition Mechanisms.* Vol. 54. 117-151.

Rokers B., Yuille, A., & Liu, Z. (2006). The perceived motion of a stereokinetic stimulus. *Vision Research.* 46, 2375-2387.

Shearer, R. &. (1999). Of Two Minds and One Nature. *Science.* Vol. 286 no. 5442: 1093-1094.

Ullman, S. (1979). *The interpretation of Visual Motion*, the MIT Press.

Ullman, S. (1983). Maximizing rigidity: The incremental recovery of 3-D structure from rigid and rubbery motion. *Massachusetts Institute of Technology, Artificial Intelligence Laboratory Memo* 721.

Wallach, H. & O'Connell, D. (1953). The kinetic depth effect. *Journal of Experimental Psychology*. 45, 205-217.

Weiss, Y. (1998). Ph.D. Thesis -- Massachusetts Institute of Technology, Dept. of Brain and Cognitive Sciences.

Weiss, Y. *et al*. (2002). Motion illusions as optimal percepts. Nature Neuroscience. 5(6), 598-604

Yuillle, A. & Grzywacz, N. (1988). A computational theory for the perception of coherent. *Nature*. 333, 71 – 74.

Yuille, A. & Grzywacz, N. (1989). *A Mathematical Analysis of the Motion Coherence Theory*. International Journal of Computer Visoin. 3, 155-175.

Zanforlin M. (1988). *Stereokinetic phenomena as good gestalts*. Gestalt Theory. 10, 187–214.

CHAPTER 5

BOUNDARY EXTENSION: INSIGHTS FROM THE SIGNAL DETECTION THEORY

5.1. Introduction

After viewing a photograph of a natural scene, human participants tend to remember having seen more of the world than was shown, as if the boundaries of the view had extended outward in memory. This error is called boundary extension (Intraub and Richardson, 1989). A direct behavioral consequence of this effect is that when a scene is first shown from a close-angle and then a wider-angle view, the perceived change of the viewed scene is less than when the sequence order is reversed. This asymmetry between the perceived wide-close and close-wide changes has been one of the signature effects in boundary extension studies.

Despite the robustness of this phenomenon across numerous durations and types of memory tests (see Hubbard, Hutchison, and Courtney (2010) and Intraub (2010) for reviews), whether boundary extension is due to, in signal detection terms, criterion bias, or discrimination sensitivity, or both, remains uncertain.

To illustrate this uncertainty, consider the following experiment by Park, Intraub, Yi, Widders, and Chun (2007). The stimulus images were paired. Each pair showed a same natural scene, one with closer-angle view and the other with wider-angle view. In the study phase of the

experiment, participants were shown one image from each pair such that half of these images were close up views and half wider angle views. In the subsequent test phase, half of the studied scenes were shown exactly as before, and half were shown in the opposite version (e.g., a close-up stimulus and wider-angle test picture). Participants rated whether the test scene was too close or too wide with respect to the studied version of that scene. The rating results replicated the diagnostic asymmetry typical in boundary extension studies. When close-wide and wide-close changes between study and test were presented, results indicated that the change from close to wide was rated as a smaller change than the change from wide to close. Although the same pair of pictures was presented, the difference between them was rated differently depending upon the order of presentation. The typical interpretation of this pattern of results is that boundary extension in memory for the close-up stimuli causes it to be more similar to the wide-angle test picture than vise versa. However, these results cannot distinguish whether the asymmetrical response was due to a criterion bias favoring a wider view of the initial scene, or to discrimination sensitivity in visual memory, or both. The current study was designed to answer this question.

5.2. Experimental design and theoretical assumptions

The key design of the new experiment was to create two probability distributions that correspond to the standard "noise" and "signal" distributions in signal detection theory, as follows. The close-wide paired images in Park et al. (2007) were randomly divided into two halves, such that

no pair was in the same half. One half was designated as study images, and the other half as test images. This guaranteed that a studied image was never shown in the test, so that the viewing distance change from study to test was either wider or closer. Consequently, there were two distributions regarding viewpoint change from study to test: close-wide and wide-close. The division between the study and test images was also randomized across participants, such that the two distributions on average will be symmetric with one another, even if each distribution was asymmetric in itself. This property of mirror-symmetry is therefore a property of the physical stimuli, and has nothing to do with visual memory (Figure 5.1.).

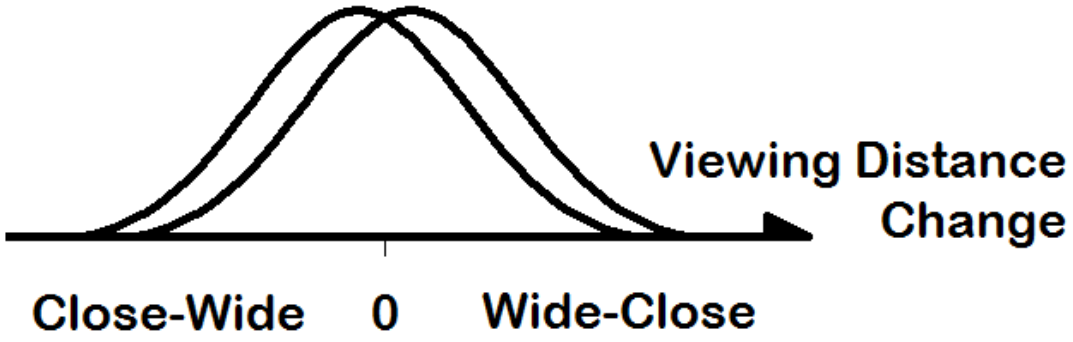

Figure 5.1. A schematic illustration of the objective probability distributions of viewing distance change: close-wide and wide-close, before any brain processing. The mirror symmetry between the two distributions can be obtained by randomly assigning study and test images across participants, given sufficiently large number of participants. We assumed that, after brain processing, each of the two distributions was a Gaussian, an assumption that could be

empirically verified. Our research question was: does the brain processing break the mirror symmetry such that the two Gaussians have different variances?

When visual memory is involved, however, there are two possible consequences. The first is that the mirror-symmetry between the two distributions is broken. This means that the visual memory has effectively changed the shape of one distribution relative to the other, resulting in a shape change of the Receiver Operating Characteristic (ROC). Consequently, the area under the ROC will also change, in general. Because the area under the ROC is one definition of discrimination sensitivity, this may mean that discrimination sensitivity has changed (note that another definition of discrimination sensitivity, *d'*, is undefined because of the asymmetry between the two distributions).

The second possibility is that the two distributions remain mirror symmetric with each other. This leads to the following two alternatives. 1) The distance between the two distributions is changed, as a result of the memory process. For example, the close-wide viewing distance change could be registered in memory as less than the opposite change of the same magnitude, giving rise to a sensitivity change that is specific to the asymmetry between wide-close and close-wide memorization. 2) There is no distance change between the two distributions, and that criterion bias is responsible for the boundary extension effect. Naturally, the two possibilities above are not mutually exclusive, so it is possible that both sensitivity change and bias are responsible for boundary extension.

To further explain the possible distance change between the two distributions, we illustrate this with two Gaussians distributions of equal variance. This distance change is simply discrimination sensitivity *d'* change, which can result from either the distance change between the Gaussian centers or from the variance change even when the distance between the two centers is unchanged. The latter case can be illustrated by the following simple example. Let us assume that the internal noise in the memory system is additive Gaussian with zero-mean. Then the resultant distributions are two wider Gaussians, whereas the distance between the two centers remains unchanged. As a result, *d'* is reduced.

The above example is simple in the sense that it is expected that internal noise in the memory system will decrease sensitivity as compared to an ideal observer without internal noise (Geisler, 1989; Knill and Kersten, 1991). What is difficult to know is whether there is any asymmetric change when the two distributions are mapped from the physical to the psychological dimension of the change between the study and test view. For example, if as a result of the memory process one distribution is shifted more with respect to the origin along the axis than the other distribution, the resultant sensitivity change will be due to the memory process that codes wide-close and close-wide viewpoint changes asymmetrically. However, because the mapping of the origin from the physical to the psychological dimension is unknown, we cannot determine whether the mirror symmetry between the two distributions is broken. That is to say, without the broken symmetry, we cannot be sure whether discrimination sensitivity is responsible for the boundary extension effects.

Therefore, if the mirror symmetry between the two distributions remains after the mapping, any discrimination sensitivity change may or may not have anything to do with the boundary extension effect. In other words, if the two distributions remain symmetric with each other, we will not have enough evidence to determine whether the boundary extension effect is due to sensitivity change or due to bias.

We now specify a single assumption underlying the upcoming test of the hypothesis. We assumed that the wide-close and close-wide distributions obtained *after* the visual memory test were Gaussians with possibly unequal variance. This assumption could be partially experimentally verified, because two Gaussian distributions always give rise to a linear ROC in the z-space. This verification is partial because two Gaussian distributions are a sufficient but not a necessary condition.

The research question then became whether the ROC could be better explained by two Gaussians of equal or unequal variance, because these two possibilities amounted to whether the slope of the ROC was unitary or not. The data were collected in the following rating experiment with six scales to obtain the ROC: whether the scene in test was wider or closer with respect to the studied scene. From the data, we could also quantify the subject's bias.

To anticipate, in Exp.1, we verified that the ROC was well approximated by two Gaussian distributions, but could not reject the hypothesis that the two Gaussians shared the same variance. We therefore ruled out the possibility that boundary extension is due solely to a

sensitivity change, because the response was strongly biased. However, we could not rule out the possibility that, accompanying the bias, there was also sensitivity change, because any sensitivity change that retained symmetry is undetectable by our method. In Exp.2, we kept the original Park et al. (2007) experiment nearly unchanged. The only thing different was that we changed the nature of the decision, instead of asking for a rating of how much closer or wider the test view looked, we asked for an old/new rating. This change made it possible for the data to be analyzed in signal detection terms, whereas the data in Park et al. (2007) could not be analyzed in signal detection terms. There, we found indeed that the boundary extension effects were due to both discrimination sensitivity change and criterion bias. In the discussion, we consider what these results imply about scene memory and in what way they are consistent with a recent model that characterizes scene representation in terms of multiple sources of information (Intraub, 2010, 2012).

5.3. Experiment 1: close-wide rating experiment

5.3.1. Stimuli

The stimuli were 121 pairs of color photographs similar to that in Park et al. (2007). It included many of the same single-object scenes and others of the same kind. Each pair included both a closer and a wider angle view of the same single object on a natural background. The resolution was 640 × 480 pixels. For any given participant, 96 pairs from the 121 available were randomly

selected as the experimental stimuli. These 96 pairs were split up into two groups, one of which contained 48 wide and 48 close views. The second group contained the remaining 96 partner images. One group was randomly assigned to each participant as the study photos (stimuli), and the second group as the test photos. For the next participant, the selection and assignment were again randomized.

5.3.2. Procedure

The experiment consisted of three blocks. Each block had a study and a test phase. In a study phase, 18 wide and 18 close images were shown. The presentation began with a 1000 ms green fixation dot at the center of the screen, and then a photo was presented for 500 ms. Each photo was followed by a 500 ms image mask, then a white fixation dot for 4000 ms, and the cycle continued until all 36 stimuli were shown. The participants were instructed to spread their attention across each image and remember it in as much detail as possible, including the objects, their layout in the scene, and the background. The participants were informed that the background was as important to remember as the foreground object and they were instructed to try to remember the image photographically.

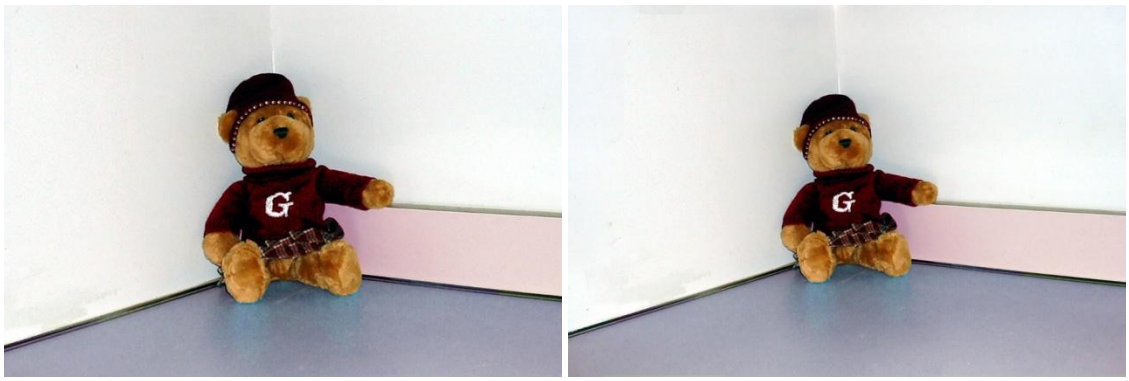

Figure 5.2. Examples used in experimental instructions illustrating a close up (left) and a wide angle scene (right) between the first study and test phases.

After the first study phase and prior to the test phase, participants were shown an example of a closer and a wider image of the same scene to illustrate the meaning of "viewing distance" in the instruction for the memory task (Figure 5.2.). During the test phase, the 36 photos that were partners of the 36 studied images were shown. Each test image was shown with unlimited time, and with a six-point rating scale underneath. Below the scales were the following terms: `Much Closer', `Closer', `A Little Closer', `A Little Further', `Further', and `Much Further'. Thus a test picture always differed from its studied counterpart. It was either too wide or too close. It should be noted that not until this illustration (Figure 5.2.) in the first block did participants know the nature of the memory test (as is common in boundary extension tasks). However in the following blocks, they knew exactly what would be tested. It took about 45 minutes for the participants to complete all three blocks.

5.3.3. Participants

Seventy-one psychology undergraduate students from University of California Los Angeles (UCLA) participated for course credits.

5.3.4. Apparatus

The display was a 17 inch Dell E773c CRT monitor, with a resolution of 1024 × 768 pixels, 32 bit color, and 85 Hz refresh rate. The images were rendered using MatLab (Math Works, Inc.) and Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) . The experiment was conducted in a dim room lit from indirect natural light. The viewing distance was 57 cm.

5.3.5. Results

5.3.5.1. Testing behavioral asymmetry

We first computed response accuracy by converting the six-scale rating data into the binary close and wide responses (three scales each). The accuracy of the wide-close condition was 0.85, and that of the close-wide condition was 0.68. A t-test (all t-tests in this paper were two-tailed) confirmed that this difference was highly significant ($t(70) = 7.62, p < 0.001$). This mirrors the asymmetry observed in earlier studies.

Recall that in the experimental instructions, boundary expansion or contraction were never explicitly mentioned during this phase of the experiment, participants were not informed that boundaries were manipulated from study to test in Block 1. In Blocks 2 and 3, however, the participants' uncertainty about the change from study to test was likely reduced. Consequently, we asked whether there was any behavioral change as a result of this reduced uncertainty, and measured the asymmetry between close-wide and wide-close changes. In Block 1, the difference between the wide-close and close-wide response accuracies was 0.28 ± 0.03 (standard error). The t-test confirmed that this difference was statistically significant ($t(70) = 9.01$, $p < 0.001$). In Blocks 2, this difference became 0.13 ± 0.03, and was also statistically significant ($t(70) = 4.49$, $p < 0.001$). In Block 3, the corresponding numbers were 0.10 ± 0.03, $p = 0.001$. The reduced asymmetry from Block 1 to 2 was statistically significant: $t(70) = 3.90$, $p < 0.001$. However, the reduced asymmetry from Block 2 to 3 was no longer statistically significant ($t(70) < 1$). This implies that participants may have attended more selectively to the photo boundaries in Blocks 2 and 3, which reduced the size of the effect.

This reduced uncertainty also co-varied with sequential order, so we could not rule out the role of instruction-independent learning from data in this experiment. Nevertheless, Intraub and Bodamer (1993) used a between-subjects design to compare participants who were explicitly told about boundary extension and challenged to prevent it and participants without explicit instructions. They found that advance knowledge indeed attenuated, but did not eliminate

boundary extension. Therefore, sequential order in this experiment is unlikely to be the only explanation.

5.3.5.2. Testing distribution asymmetry

We now address the question whether this behavioral asymmetry was due to mirror-asymmetry between the two distributions. In other words, we asked whether the two distributions, wide-close and close-wide, shared the same variance when both were assumed to be Gaussian. Without loss of generality, we assumed that the wide-close distribution was $N(0, 1)$, and the close-wide distribution was $N(\mu, \sigma)$. The question was therefore whether σ was unitary.

There are two methods one can use to answer this question. The first is to plot the ROC in the z-space, which should be linear if the two distributions are Gaussian. Then the question becomes whether the slope of this ROC line is unitary. Testing of a unitary slope, however, turned out to be technically difficult. This is because a substantial number of hit and false-alarm rates across participants were 1's and 0's, causing the corresponding z-scores to be infinity. The standard correction is to subtract or add a small number (e.g., 1/(2n), where n is the number of signal or noise trials) to avoid the infinity. However, this correction is arbitrary and hence problematic in hypothesis testing. We nevertheless carried out this analysis using the 1/(2n) correction in a hope to obtain converging results with the second method. We first checked the goodness of linear fit from each participant's data. The mean $R^2 = 0.94 \pm 0.005$ (range: 0.82 to 1.00),

indicating reasonably good fit for all participants. The average linear slope ($1/\sigma$) thus calculated was 0.98 ± 0.016. The *t* test could not reject the null hypothesis that $\sigma = 1$ ($t(70) = 1.46$, $p = 0.15$) (Figure 5.3.).

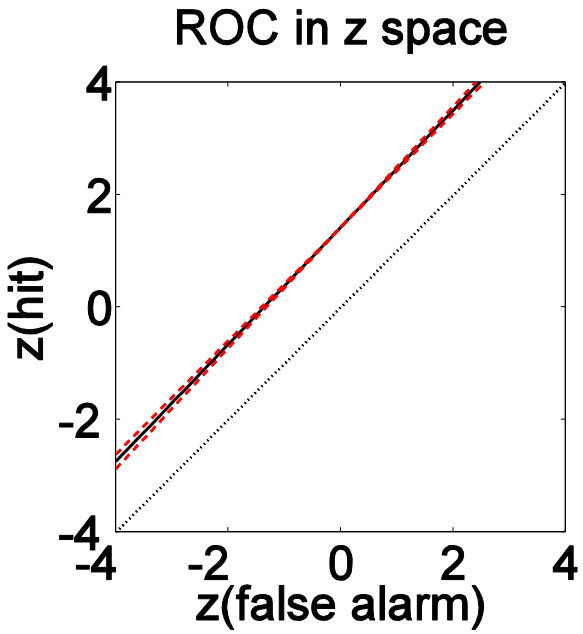

Figure 5.3. The solid black line is ROC in z-space, the dashed red lines are 95% confidence intervals. In order to avoid the problem of an infinity z value, when a rate number was 1 or 0, the rate was corrected by subtracting or adding 1/(2n), respectively, where n was the number of "signal" or "noise" trials.

The second method is to fit the ROC in the hit and false-alarm rate space. The fitting is nonlinear, but does not suffer from the infinity problem above and hence needs no arbitrary corrections. We were aware of the well-known "error-in-variables" problem (Griliches and Ringstad, 1970). That is, both the hit and false-alarm rates had error bars. Accordingly, we

minimized the sum of the squared shortest distance from each datum point to the parameterized ROC model curve. In the unitary square of hit and false-alarm rate space, the fitting residuals had a mean of 0.0009, with a standard error of 0.00014. The average σ thus obtained was 0.95, with a standard error 0.04. A *t* test could not reject the null hypothesis that σ was unitary (*t(70) = 1.29, p = 0.20*) (Figure 5.4.).

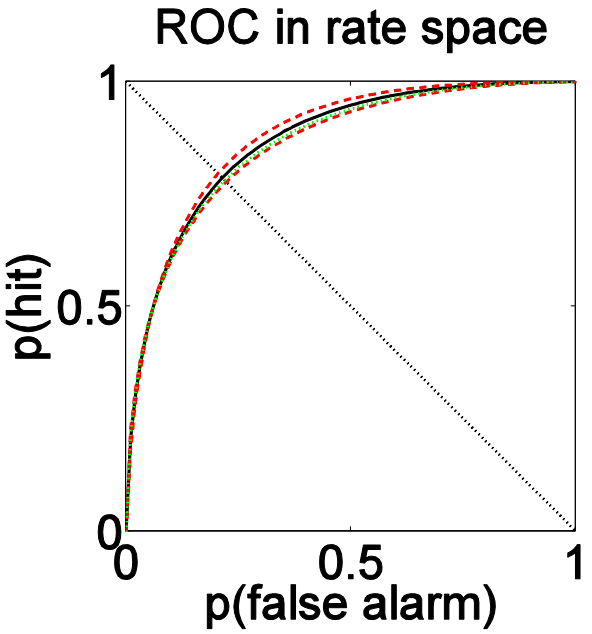

Figure 5.4. The solid black line is ROC in hit and false-alarm rate space. The dashed red lines indicate the 95% confidence interval. The dashed green line is obtained when the "signal" distribution is assumed to have a unitary variance, as compared to the black line where the variance of the "signal" distribution was calculated from the fitting data.

Given that the two Gaussian distributions shared comparable variance, the bias free criterion was where the two distributions intercepted. Any deviation from it would be the bias. The mean distance from the origin to the intersection was 0.80 ± 0.05. The participants' decision criteria were recovered from their false alarm rates. The mean distance from the origin of $z = 0$ to the criterion was 1.13 ± 0.05. This difference from bias-free location was statistically significant ($t(70) = 8.04, p < 0.001$) (Figure 5.5.).

To conclude Exp.1, we found no evidence that the mirror-symmetry between the two distributions, "noise" and "signal," was broken. This means that any possible sensitivity change that retained the mirror symmetry could not be found by our method. Examples of such change from the physical to psychological dimension mapping include relative distance change between the two distributions, and equal increase or decrease of the variance of both distributions. On the other hand, the participants' bias could fully explain the results of boundary extension if no discrimination sensitivity was changed. It is worthwhile to emphasize that this bias in signal detection terms is not necessarily the same as a preference to respond to any stimulus image as being too close. We will elaborate this in the discussion.

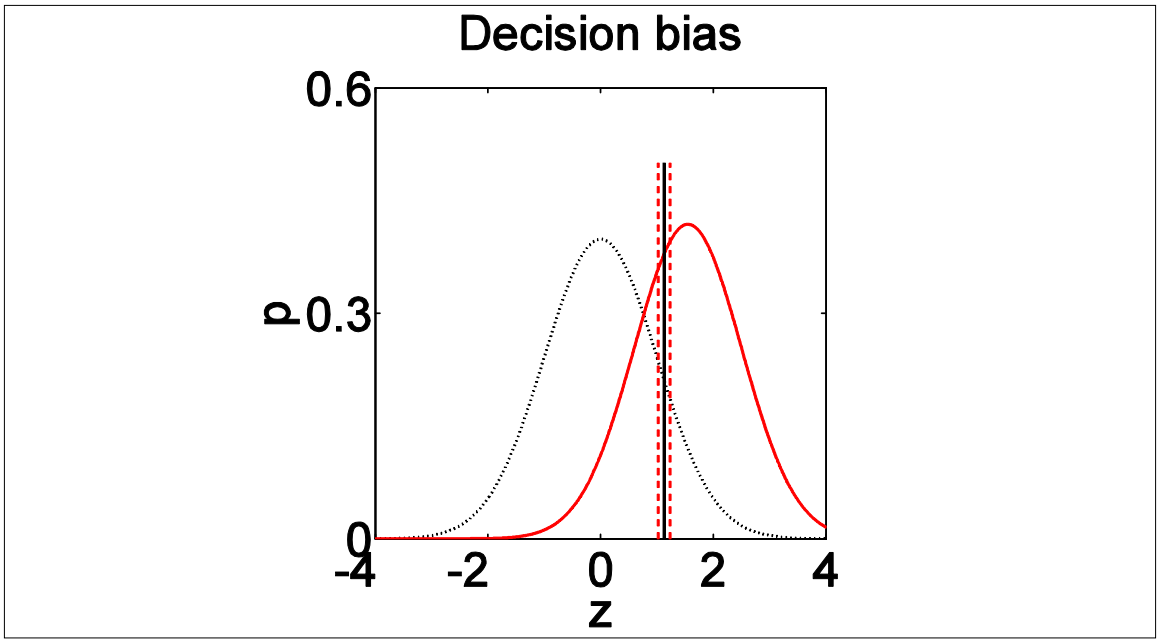

Figure 5.5. The average decision criterion location and its 95% confidence interval.

5.4. Experiment 2: old-new experiment

In Exp.1, we studied boundary extension effects along the dimension of view-angle change, and found no evidence that discrimination sensitivity was responsible for the effects. However, that dimension (from too close to too wide) is but one of many possible dimensions along which one can study visual memory of natural scenes. Another possible dimension, commonly studied in memory research in general and in boundary extension in particular, is whether a test scene is the same as or different from the study scene. Exp.2 studied boundary extension along this dimension. We first provide an overview of experimental design before describing the details. During the study, half of the images were close and half were wide. During test, half of the

images were identical to the studied view and half were changed to the alternate view so that if the stimulus was close the test item was wide, and if the stimulus was wide the test item was close. To be concrete, we first consider the close studied images. During test, either the same images were shown, or their wider counterparts were shown. Participants responded either "old" (image unchanged from study) or "new" (image changed from study). Therefore, the two objective stimulus categories and two subjective responses made up the standard 2 × 2 contingency table in signal detection theory (Green and Swets, 1974). So discrimination sensitivity *d'* can be calculated. Likewise, *d'* could also be calculated for wide studied images. With these two sensitivities, one can ask whether there is any difference between them.

Technically, though, *d'* may not be definable, because the "signal" and "noise" distributions may not be Gaussian or may not be of equal variance. Accordingly, a rating experiment, rather than a binary old-new experiment, was used to obtain an ROC function. In this way, the Gaussian and equal variance assumptions could be verified, as in Experiment 1.

5.4.1. Experimental design

The stimulus images, apparatus, design, procedure, experimental instructions, and experimental duration were nearly identical to Experiment 1. Here we specify only the differences.

1) After the study phase and prior to the test phase in Block 1, participants were shown the same example in Figure (5.2.) of a closer and a wider image. The only difference was that participants were instructed to respond "old" (at one of three scale points) if a test

image was exactly the same as in the study, or "new" if the test was wider or closer than the studied.

2) Unlike Experiment 1, in which the test views always differed from the study view, in this experiment half the time they were the same, and half the time they differed as describe earlier. The memory task was to rate whether the test image was exactly the same as the studied image. A six point rating scale was provided from 'Sure Old' (-3), 'Guess Old' (-1), 'Guess New' (+1), to 'Sure New' (+3).

5.4.2. Participants

Twenty-four undergraduate students from UCLA participated for course credits.

5.4.3. Results

We first looked at the overall accuracy by collapsing the three "old" and three "new" responses into binary responses. The overall accuracy of the task was 0.61, with standard error ± 0.02 (the reported errors below will also be standard errors).

We now look at the data in more detail. When the studied images were wide views, the hit rate in test was 0.73 (± 0.02), and the false alarm rate was 0.45 (± 0.03). In comparison, when the studied images were close views, the corresponding rates were 0.72 (± 0.03) and 0.57 (± 0.02). Apparently, the signature asymmetry between close and wide study images was mainly reflected

in the higher false alarm rate for the close studied images. This means that wider test images were more often mistakenly identified as exactly the same as the studied, closer images, which is the boundary extension effect.

We now compute discrimination sensitivities. Without loss of generality, we assumed that the wide-wide distribution was *N(0, 1)*, and the wide-close distribution was *N(μ, σ)*. We similarly assumed two distributions for the close-close and close-wide. The research question remains whether the two discrimination sensitivities thus separately obtained, measured in either *d'* or area under the ROC curve, were the same. It should be noted that whether the wide-wide and close-close distributions are identical in shape or not is unknown, but is irrelevant to the research question.

We used the first method to obtain each participant's ROC in the z-space. The mean R^2 for linear fitting was 0.90 ± 0.02, indicating reasonably good fit for all participants. The average σ calculated from linear slope for wide-close was 1.18 ± 0.10, which was marginally significantly different from unitary (*t(23) = 1.83, p = 0.08*). The average σ for close-wide was 1.17 ± 0.09, which was also marginally significantly different from unitary (*t(23) = 1.88, p = 0.07*). Because of the marginal significance, we decided to use the area under the ROC to calculate discrimination sensitivities. The areas were 0.683 ± 0.019 and 0.603 ± 0.022, and the difference was statistically significant, *t(23) = 2.52, p < 0.02*. (The *d'* values would have been 0.77 ± 0.09 and 0.43 ± 0.09, giving rise to a significant difference between them, *t(23) = 2.73, p = 0.009.)*

Using the second method of ROC curve fitting in the hit and false-alarm rate space, the mean residual error was 0.006 ± 0.002, indicating reasonably good fit for all participants. The average σ for the wide-close trials was *1.20 ± 0.14*, the *t* test could not reject the null hypothesis that $\sigma = 1$, $t(23) = 1.41$, $p = 0.17$. The average σ for the close-wide trials was *1.23 ± 0.13*, the *t* test could not reject the null hypothesis that $\sigma = 1$ either, $t(23) = 1.73$, $p = 0.10$. The areas under the ROC became *0.684 ± 0.019* and *0.610* ± 0.026, and the difference was statistically significant, $t(23) = 2.13$, $p < .05$. (If the equal-variance model were applied, the resultant *d'* values would have be 0.79 ± 0.10 and 0.39 ± 0.12, giving rise to a statistically significant difference, $t(23) = 2.78$, $p = 0.01$.) An important point here was that discrimination sensitivities were statistically significantly different, irrespective of the measures used.

We looked next at the decision criteria for close and wide studied images, respectively. We again defined a bias-free criterion as the intersection between the "signal" and "noise" distributions. In the case of close studied images, these two distributions correspond to close-close and close-wide distributions, respectively. The bias-free criterion obtained from the rate-space fitting was 0.21 ± 0.14, and was 0.28 ± 0.15 from the z-space linear fitting. The actually criterion calculated from the participants' false-alarm rate was 0.63 ± 0.08. There was therefore indeed bias ($t(23) = 3.45$, $p = 0.002$; and $t(23) = 2.91$, $p = 0.008$), in that a wider test image was more likely to be considered as the same as the closer studied image, in agreement with the boundary extension effect.

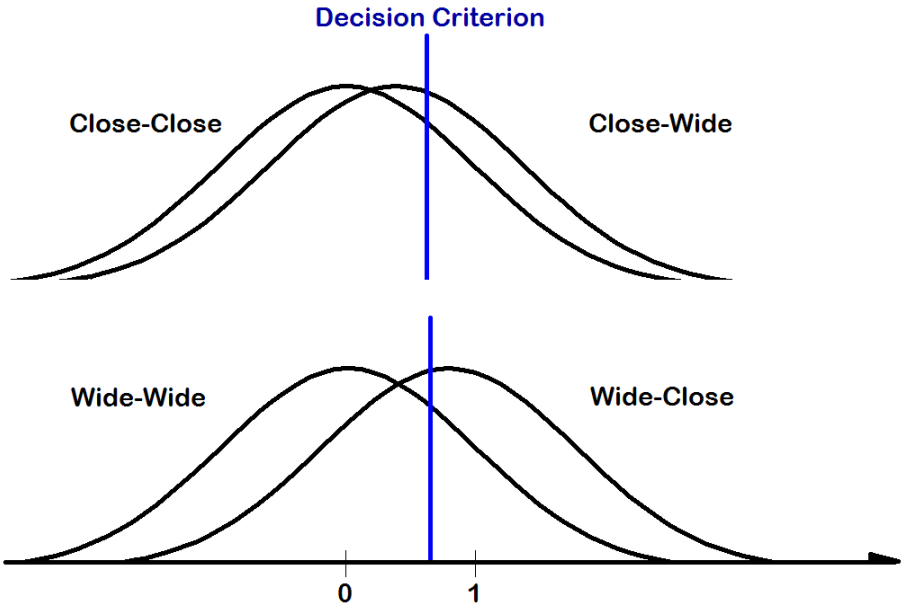

Figure 5.6. Top: The recovered "noise" (close-close) and "signal" (close-wide) distributions for close studied images, and the participant' decision criterion. Bottom: The corresponding distributions and criterion in the case of wide studied images. Along the horizontal z-axis, the two "noise" distributions are centered at the origin per convention. The two decision criteria were located at approximately the same location. Nevertheless, the "signal" distribution in the case of close studied images was closer to the "noise" distribution, resulting both in a smaller d' and a statistically significant bias as compared to the case of wide studied images, where the bias was not significant.

In the case of wide studied images, the bias-free criterion obtained from the rate-space fitting was 0.52 ± 0.12, and was 0.55 ± 0.10 from the z-space linear fitting. The actually criterion calculated from the participants' false-alarm rate was 0.65 ± 0.08. This bias was statistically not significant ($t(23) = 1.48$, $p = 0.15$; and $t(23) = 1.12$, $p = 0.27$). It is interesting to note that the

criterion locations in the two cases were very similar to each other (0.63 and 0.65, respectively). We will discuss the possible implications in the next section.

5.5. Discussion

We modified the typical boundary rating procedure used to test boundary extension in order to assess the functional nature of the boundary error as suggested by signal detection theory. In both experiments, we were able to replicate boundary extension in memory for photographs of scenes. However, in Exp.1 when participants had to determine whether the change in the expanse of the view from study to test was close-up to wider-angle or wider-angle to close-up, we obtained no evidence that boundary extension was due to a change in discrimination sensitivity. Rather, decision criterion bias (in signal detection terms) could well explain the behavioral results. In comparison, in Exp.2 participants determined whether a test image was exactly the same as studied or different in viewing angle (either wider or closer). Here, as compared to the wider studied images, the closer ones gave rise to a sensitivity reduction and a bias shift, both of which promoting boundary extension effects.

Although these results are straightforward in signal detection terms, their implications in terms of inferring the nature of scene representations in memory remain an open question. This is because signal detection theory (SDT) is based on functional characterizations of uncertainties among various categories that need to be separated, irrespective of the mechanistic processes

involved. SDT is therefore agnostic about the origin of the sensitivity or criterion. Nevertheless, a widespread misconception exists that equates the bias to "decision" bias or "response" bias, suggesting that the bias is necessarily high level in nature. However, as Georgeson (2012) pointed out, bias could be also perceptual, or lower level. Georgeson used the motion aftereffect as an example, whose perceptual nature is questioned by few. Motion aftereffect is nevertheless characterized as a shift of the psychometrical function, but not a slope change. This shift is indistinguishable from a response bias. This example hence illustrates that bias in SDT is not necessarily equivalent to high level decision or response bias. It remains an open question nevertheless what exactly bias and sensitivity mean in psychological terms. In what follows, we speculate what our experimental results imply in boundary extension effects.

One theoretical explanation of boundary extension, is provided by the multisource model of scene representation (Intraub 2010; 2012).. The model assumes that the representation of a visual scene is formed by visual stimulus information, amodal completion, as well as expectations and constraints from contextual scene classifications. The definition of amodal completion is that visual fragments in a scene perceptually complete behind occlusion to connect into a single object (Michotte, 1954). In the context of amodal completion, Lu and Liu (2008, 2009) used an experimental technique similar to the current study to investigate memory representations of objects and scenes, and their results were consistent with the multisource model. In the context of boundary extension, this aspect of amodal perception is better characterized as amodal continuation. Here is an example to illustrate. When a photo of a natural scene is viewed, the photo necessarily has a boundary, making the scene limited in spatial

expanse. The multisource model assumes that when a scene with limited spatial expanse is viewed, it is analogous to viewing the scene through a window with the surrounding scene being occluded. The memory system automatically fills in, to a limited spatial extent, the missing boundary scene using the available sensory data and generic knowledge of natural scenes, for example, grassland should continue with similar texture statistics, and a partially visible object at the boundary should be a complete object. Therefore, if the traditional amodal completion can be considered as spatial interpolation, boundary extension may be analogously considered as spatial extrapolation.

In this way, according to the multisource model, the memorized scene is not a photographic replica, but expanded beyond the initially visible (yet artificial) boundary. We now start from this hypothesis to interpret results in Exp.1 and 2. From the outset, another assumption is apparently needed. We assume that the representation of a studied scene has an extended boundary, but that the representation of a test scene has no boundary extension. This assumption is reasonable because, in both experiments, a test image disappeared only after a participant had selected a rating response. Therefore, no expansion of the test image representation is expected because the constantly available and unambiguous stimulus information should overwhelm all other sources. In other words, if study and test images shared similar boundary extension, there should be no difference in behavioral results between wide-close and close-wise study-test sequences.

In Exp.1, we assume that the representation of a studied image was extended in such a way that the effective viewing distance was lengthened. For simplicity, we start by considering this lengthening as a constant. Now, let us consider the horizontal axis in Figure (5.7.) as the physical viewing distance change from study to test. The corresponding subjectively perceived horizontal axis is simply shifted to the right by a constant. This means that the mirror symmetry was not broken between the two distributions, close-wide and wide-close. As a result of this constant shift, the decision criterion was effectively closer to the distribution on the right (wide-close), giving rise to the boundary extension effects.

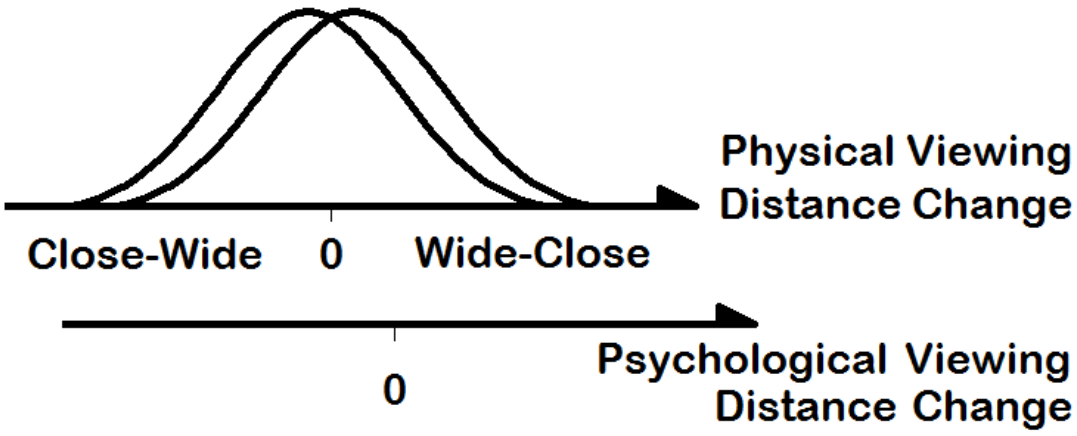

Figure 5.7. Schematic illustration of what might have happened in Exp.1. During to boundary expansion of the representation of a studied image, the studied image's viewing distance was effectively lengthened. For illustration and for simplicity, we assume that this lengthening is constant (see text when this lengthening is not considered as a constant). Therefore, the ***perceived*** viewing distance change, shown as the horizontal axis at the bottom, shifted to the right with respect to the physical viewing distance change from study to test. This means that the

symmetry between the close-wide and wide-close distributions was not broken. In effect, there was a decision criterion shift along the axis. This simple mechanism could explain the results in Exp.1.

More generally, the extension of the representation of a studied image may not be a constant, but depends on the view-angle. This means that the effective lengthening of the viewing distance varies as a function of the view-angle. The lengthening of all the close study images follows some distribution. So the psychological distribution of close-wide viewing distance change is the convolution of this distribution and the physical close-wide distribution. We can similarly obtain the psychological wide-close distribution. In theory, these two psychological distributions may or may not share the same variance, since the lengthening of the psychological close and wide distributions may not necessarily share the same variance. In practice, however, our data indicate that the two variances are nearly identical. The distance between the psychological close-wide and wide-close distributions may also differ from the corresponding distance in the physical dimension. So the combined consequence of greater variance and a possible change in relative distance may change the discrimination sensitivity index *d'*. What is certain is that these two psychological distributions will move to the right with respect to their physical counterparts, effectively creating a bias (when everything else is held unchanged) promoting boundary extension.

In Exp.2, the relevant axis was changed to a psychological dimension of same or different. Because of the constant shift, the hit rates for the close-close and wide-wide conditions should be

comparable, which is confirmed by the data (.72 and .73, respectively). However, in the wide-close condition, the perceived difference became larger because the wide studied images became even wider. In comparison, in the close-wide condition, the perceived difference became smaller because the close studied images became wider. In effect, and relatively speaking, this means that the wide-close distribution moved away from the wide-wide distribution, whereas the close-wide distribution moved toward the close-close distribution. Hence, the discrimination sensitivity was higher for the wide than for the close studied images. Figure (5.8.) illustrates this explanation.

Given that the close and wide study images were randomly interleaved, that all test conditions were also randomly interleaved, and that wider and closer viewing angles were relative terms, it is sensible for the participants to hold a single decision criterion location. This location was also consistent with the behavioral boundary extension effects. Nevertheless, we do not have a theoretical explanation why this criterion was necessarily located there, as opposed to be, for example, midway between the two original distributions before boundary extension happened. In other words, boundary extension could still occur by the sensitivity difference alone, without resorting to bias.

In conclusion, we have designed two new experiments to study boundary extension using signal detection theory. Although these new designs made only small changes to rating task often used to test boundary extension in the literature, it should be noted that these changes made it possible, for the first time, for boundary extension to be examined in terms of sensitivity and

bias. Our two new experiments were also very similar to each other. Yet, with a minimal change, one can ask whether the study-test view-angle change was too close or too wide, or whether there was any change at all. With this small change, one effectively asked a question from a different psychological dimension: the perceived viewing distance change either in terms of close or wide or in terms of change or no change. Interestingly, simply by changing the question being asked, one could find that discrimination sensitivity was unneeded or playing a major role in the behavioral boundary extension effects. It is remarkable, we believe, that a single, simple assumption could explain most of the diverse results in a straightforward fashion. This assumption, that the memory representation of a natural scene extends its boundary by a constant extent, is consistent with the multisource model of visual scene memory by Intraub (2010, 2012).

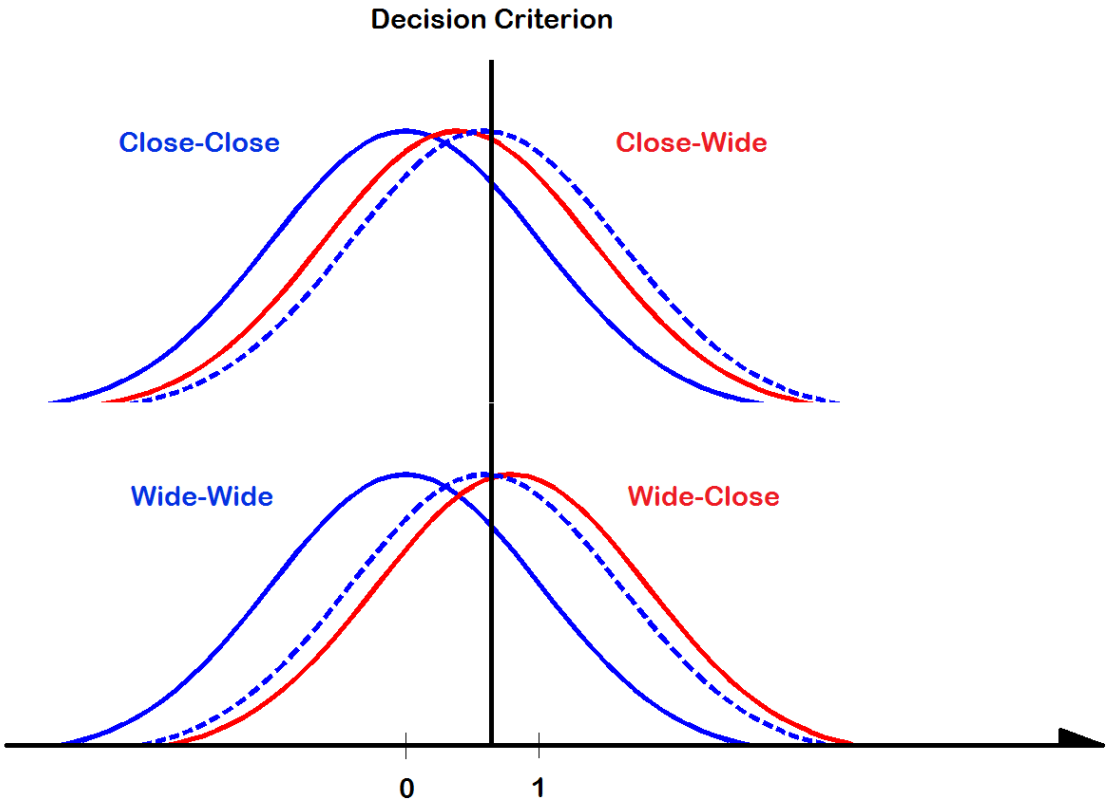

Figure 5.8. Schematic illustration of what might have happened in Exp.2, where the horizontal axis indicates difference between study-test image matching. Due to boundary expansion of the representation of a studied image, the study-test matching between close-close and wide-wide conditions would change. But the changes were comparable, giving rise to comparable hit rates in these two conditions (shown as the two blue distributions on the far left, in top and bottom panels). Interestingly, the close-wide matching would reduce the matching difference, making the psychological close-wide distribution shifted to the left (top panel, dashed blue to red). In contrast, the wide-close matching would enhance the matching difference, making the psychological wide-close distribution shifted to right (bottom panel, dashed blue to red). These

two shifts could explain the discrimination sensitivity difference between the close and wide study image conditions.

References

Brainard, D. J. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433-436.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychol Rev, 96*(2), 267-314.

Georgeson, M. (2012). *Sensory, perceptual and response biases: the criterion concept in perception*. Paper presented at the Vision Sciences Symposia. from http://www.visionsciences.org/symposia2012_1.html

Green, D. M., & Swets, J. A. (1974). *Signal Detection Theory and Psychophysics*. Huntington, New York: Robert E. Krieger Publishing Company.

Griliches, Z., & Ringstad, V. (1970). Error-in-the-Variables Bias in Nonlinear Contexts. *Econometrica, 38*(2), 368-370.

Hubbard, T. L., Hutchison, J. L., & Courtney, J. R. (2010). Boundary extension: findings and theories. *Q J Exp Psychol (Hove), 63*(8), 1467-1494.

Intraub, H. (2010). Rethinking Scene Perception: A Multisource Model. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 52, pp. 231-264). Burlington: Academic Press.

Intraub, H. (2012). Rethinking visual scene perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*(1), 117-127.

Intraub, H., & Bodamer, J. L. (1993). Boundary extension: fundamental aspect of pictorial representation or encoding artifact? *J Exp Psychol Learn Mem Cogn, 19*(6), 1387-1397.

Knill, D. C., & Kersten, D. (1991). Ideal Perceptual Observers for Computation, Psychophysics and Neural Networks. In R. Watt (Ed.), *Pattern Recognition by Man and Machine*: MacMillan Press, London.

Lu, H., & Liu, Z. (2008). When a never-seen but less-occluded image is better recognized: Evidence old-new memory experiments. *Journal of Vision, 8*(7(31)), 1-9.

Lu, H., & Liu, Z. (2009). When a never-seen but less-occluded image is better recognized: Evidence from same-different matching experiments and a model. *Journal of Vision, 9*(4), 1-12.

Michotte, A. (1954). *La perception de la causalite*. Louvain: Publications Universitaires de Louvain.

Park, S., Intraub, H., Yi, D. J., Widders, D., & Chun, M. M. (2007). Beyond the edges of a view: boundary extension in human scene-selective visual cortex. *Neuron, 54*(2), 335-342.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10*, 437-442.