STRUCTURE FROM MOTION WITHOUT PROJECTIVE CONSISTENCY

By

XIAOLI HE

A thesis submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Psychology

Written under the direction of

Dr. Manish Singh

And approved by

_____

_____

_____

New Brunswick, New Jersey

May 2016

ABSTRACT OF THE THESIS

Structure-From-Motion without Projective Consistency

By XIAOLI HE

Thesis Director:

Dr. Manish Singh

Structure-from-motion (SFM) studies have shown that people are good at perceiving 3D structure in dynamic dot displays consistent with rigid object rotation. However, observers can perceive volumetric structure even when image motion is inconsistent with rigid rotation. As an extreme case, in dynamic figure-ground displays containing textural motion, observers perceive one set of regions as rotating in 3D, despite constant dot speed everywhere (projectively inconsistent with 3D rotation; Froyen et al. 2013, JOV; Tanrikulu et al., 2016, JOV). It is unclear, however, to what extent this extreme "tolerance" is due to the figure-ground competition, which induces the assigned figural region as rotating in those displays. Here we used standard SFM displays, depicting a single object in isolation, and manipulated the discrepancy of image motion from 3D rigid rotation. We started not with 3D objects, but with a 2D velocity field within a vertically oriented ellipse. For an ellipsoid rotating about its principal axis, its orthographic-projected speed profile is a cosine function along each orthogonal "rib". We manipulated the proportion α of cosine speed versus constant speed (range 0-1), and the motion direction θ relative to the orthogonal ribs (range 0-60°). In experiment 1,

ii

observers used a 7-point scale to rate the degree to which the display depicted a volumetric object. In experiment 2, observers adjusted the depth to match the SFM displays. Both experiments shows that the volumetric percepts increased significantly with α, and were surprisingly tolerant to deviations from the projectively correct $\alpha=1$, $\theta=0$. For α, volumetric ratings increased between 0-0.6 but plateaued beyond 0.6. The effect of θ was surprisingly small, with even $\theta=45°$ yielding high volumetric ratings. Then we applied a rigidity-based computer vision model to our displays, and compared the model prediction with observers' data. The comparison shows that the rigidity-based model does not predict human's volumetric percepts correctly. In addition, the predicted motion from the model has large deviations from the display itself in terms of rotation axis. Thus even in standard SFM displays, the 3D percept was surprisingly tolerant to discrepancies from projectively correct rigid object motion. These results argue for a more nuanced view of 3D interpretation in which strict projective consistency plays a less prominent role than in conventional SFM accounts.

Acknowledgement

I would like to express my very great appreciation to my advisor Manish Singh for his guidance throughout the experiment and completion of this thesis. I would like to offer my special thanks to Jacob Feldman, for his many helpful comments and discussions, and to Vicky Froyen for the help with the model. I would thank my committee member Dr. Thomas Papathomas for his helpful comments and suggestions. I would also thank for my lab mates in Perceptual Science Lab and Visual Cognition Lab for their helpful discussions. In the end, I would thank my dear parents, my brother and my friends for being so understanding and supportive.

Table of Contents

List of illustrations

1. INTRODUCTION

Structure from motion (SFM) is a phenomenon that our visual system can perceive the three-dimensional structure from two-dimensional images sequences that contain only 2D motion information. For decades, researchers standardly used 2D projections of simulated rotating 3D objects to study the phenomenon. For example, one of the first SFM studies used 2 frames of projection from a wireframe object that consisted of connected line segments (Wallach, 1953). Later on, researchers used disconnected elements such as isolated points, terminations of line segments, or texture elements, and added more frames and elements to the display. There are two characteristics in common in the above SFM displays: first, no 3D motion information was conveyed by each individual frame; and second, such elements eliminated other cues like occluding contour and shading. Therefore, there are many possible 3D structures and 3D motions that are consistent with the frame sequences—though some interpretations are simpler, and some can be very complicated.

This leads to the interesting and important question: what constraint does the visual system use in SFM? Different constraints have been proposed to interpret human SFM. Most of them focus on explaining how people can perceive 3D structure from image sequences that are projected from rigid object rotation. One of the most well known constraints is the rigidity assumption (Ullman, 1979).

Based on empirical research, Ullman (1979) proposed the rigidity assumption: "*Any set of elements undergoing a 2-D transformation which has a unique interpretation as a rigid body moving in space, should be interpreted as such a body in motion.*" The

rigidity assumption assumes that a rigid interpretation is uniquely determined (up to a depth reversal) by applying the rigidity constraint. He also showed that the rigidity assumption was computable from his structure-from-motion theorem. The theorem states that if rigidity is satisfied in three distinct orthographic views of four non-planar points, the 3D structure of the four points can be uniquely determined (including the reflection about the image plane). Based on the above assumptions, Ullman developed the original rigidity-based algorithm: in order to estimate the structure of the whole object, we can divide the object into many small groups of four elements, and test rigidity within each group, and then combine the results into a global object. In summary, this model assumes that as long as the 2D display is projectively consistent with a rigid object, the 3D structure and motion can be uniquely determined. Since its publication, the rigidity assumption has had great impact on other location-based models in human SFM research (Clocksin, 1980; Husain, Treue, Anderseon, 1989). The location-based models code the frames in terms of the (x,y) locations of all the dots, and assume that the visual system recovers the 3D structure by analyzing the changes in the dots' locations. The rigidity model is also widely used in computer vision (Tomasi & Kanade, 1992; Morita & Kanade, 1997; Anandan & Irani, 2002).

The above location-based models can't deal with correspondence problem — when the dots have limited lifetime. Specifically, with limited lifetime, it's difficult to determine which dots correspond to which in the previous frame. In another word, the appearance and disappearance of dots makes it impossible to track individual dots. Another type of models based on velocity field can avoid the correspondence problem. Such models assume that the visual system uses motion information such as image

velocity field to compute the 3D structure. For example, Braunstein and Andersen (1984) ran a series of experiments and proposed the constant angular velocity heuristic. Instead of testing dots' consistency directly with rigidity interpretation, they assumed that the visual system analyzes the projected image motion profile (i.e. 2D speed profile) of the dots to test for consistency with 3D rotation. They found that if the dots' 2D speed profile was a cosine function along each horizontal cross-section (assuming that the object is centered at the origin), people could perceive the rotation of sphere with high accuracy (Figure 1). Therefore, they proposed that the visual system prefers the interpretation of a 3D shape rotating with constant angular velocity.
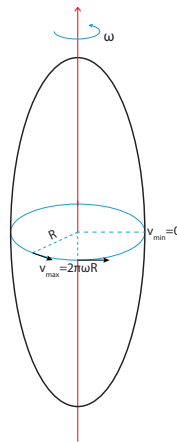
Figure 1: The orthographic image motion profile (i.e. 2D speed profile) of a rotating ellipsoid should be a cosine function along each horizontal cross-section (assuming that the ellipse is centered at the origin).

The location-based models and the velocity-based models take different inputs and use different algorithms, but both types of models assume projective consistency with rigidly moving objects. In particular, these models perform quite well with SFM displays that are projected from rigid rotating objects. However, the generalizability of

the algorithms based on the rigidity constraint is unclear given that people can also handle displays that have certain deviations from rigidity motion, such as bending, stretching, twisting and flowing (Todd, 1982). Therefore, researchers attempted to relax this assumption and proposed some new algorithms to handle certain types of non-rigid motions.

Todd's (1982) trajectory-based algorithm attempts to discriminate between rigid and non-rigid objects. The algorithm is based on the image trajectories of each element's projected motion in space and time. It assumed that different types of motion such as rigid motion and non-rigid motion have their specific geometric relations among their trajectories. Normally, the trajectories from rigid motion have certain geometric patterns. For example, for a vertical orientated ellipsoid rotating about its primary axis, the trajectories of dots lying on the surface will be a series of horizontal circles with different radii. He claimed that people can discriminate rigid from non-rigid motion based on trajectory cue. He applied the algorithm to SFM displays that were projected from rigid and non-rigid motion, and compared the algorithm predictions with human discriminations. The algorithm succeeded in distinguishing rigid from non-rigid motion as human did. However, it didn't attempt to compute the non-rigid 3D shapes people actually perceive.

Later on, Ullman (1983) also developed an incremental rigidity scheme to deal with displays containing motion that includes some deviations from perfect rigidity, such as an object undergoes certain degrees of distortion while rotating. The schema tried to maintain an internal rigid model by allowing minimum non-rigid deformations over successive frames. The scheme can tolerate a certain degree of deviation from rigidity,

but it is time consuming. Moreover, although it attempts to relax the rigidity assumption, it's so sensitive to noise that it doesn't always have a stable interpretation especially when the initial rigid interpretation is incorrect (Braunstein, 1994).

Instead of using location-based constraints such as incremental rigidity or trajectory patterns, Jain & Zaidi (2011) showed that relative velocity plays a dominant role in SFM. They used random-dot displays containing projections of narrow cylinders that rotated simultaneously about the vertical and depth axes. In addition, the cylinders were either rigid or flexing, so the deformation in the displays resulted from two types of 3D motion: either rigid or non-rigid motion. Their experiments showed that human observers had equal sensitivity in discriminating the shape of cross-sections from rigid and non-rigid displays. The model they proposed was based on motion perspective and three differential invariants (*curl*, *div* and *def*) decomposed from the first order velocity field (Koenderink & Van Doorn, 1986). Their computational model performed similarly as human observers. In addition, they showed that the rigidity-based models such as Todd's (1982) trajectory algorithm couldn't predict human percepts well.

In summary, all these models assume that the human perception is projectively consistent with image data. At first, researchers tried to explain people's ability in perceiving 3D structure from displays that were projected from rigid object rotation. Various models were proposed based on the rigidity constraints. Later on, it was found that people can also perceive the 3D structure from certain types of non-rigid object rotation, because such non-rigid motions as bending a tube and biological motion are quite common in daily life. These phenomena questioned the previous models, especially the location-based models. The rigidity-based models tried to relax the original rigidity

assumption, but they still didn't perform well with non-rigid 3D motion. This suggests that the visual system does not work in the same way as the rigidity-based models propose. Therefore, people began to use velocity field as model input, so as to further explore humans' ability to estimate different non-rigid 3D motions. These SFM models underwent huge changes over the decades, but all of them have one universal assumption: in order to perceive 3D structure, the 2D motion sequences should be projectively consistent with 3D motion, whether rigid or non-rigid. But what happens when the universal assumption is violated? There is some evidence showing that people are good at perceiving volumetric structure when projective consistency is violated.

Surprising evidence comes from the figure/ground study of Froyen et al. in 2013. As Figure 2 shows, these displays contained alternating light and dark regions with different shapes of contours in terms of their convexity and symmetry. Inside each region, there were random-dot textures moving horizontally at constant speed, but the motion direction is opposite in alternating regions. The pattern of texture motion was not projectively consistent with a 3D rotating object. However, people still had strong percept of rotating columns in one set of regions (light or dark). The study showed that even if the image motion is projectively inconsistent with 3D rotation, people can still have strong 3D percepts.
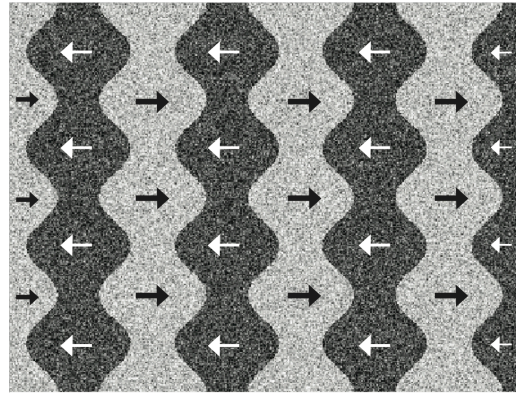
Figure 2: An example of display used in the figure/ground study of Froyen et al. in 2013. The arrows show the motion direction of the textures.

In traditional SFM studies, the displays were generated from 3D objects—real or simulated, with rigid or non-rigid motion. So projective consistency is a necessary constraint for current SFM models. However, the above figure/ground studies show that when the 2D image motion has large deviations from projective consistency, people can still have strong 3D percepts. But a natural question is whether this occurred only due to figure-ground competition across each contour. In the rotating-column displays, the cue about figure-ground competition is accretion/deletion along each contour, and there are two possible interpretations. The traditional interpretation is that accretion/deletion is a strong cue to background (Kaplan, 1969; Thompson, Mutch, & Berzins, 1985): when a translating texture accretes/deletes from a boundary, it is perceived as appearing/disappearing behind the surface defined by the boundary. Accretion/deletion can also be interpreted as a cue to 3D objects. It is compatible with self-occlusion by a rotating 3D object—deletion can be a result of texture moving to the backside of an object, while accretion can be a result of texture moving from the backside of the object to the front side. The second interpretation leads to the rotating percept. In the rotating-

columns displays, accretion/deletion was present on both sides of each contour, so each side can be perceived as either figure or background. However, both sides cannot be in the back at the same time. Since there are two possible interpretations for each side, one side can be interpreted as a rotating object in the front, and another side as background. Therefore, it's unclear whether the 3D percept is due to figure/ground competition or due to the effect of SFM.

Ramachandran, Cobb and Rogers-Ramachandran (1988) reported a visual demonstration, which suggests that projective consistency can be violated even in the absence of figure-ground competition. They used random-dot display that was projected from a rotating cylinder, so the display was projectively consistent with a rotating cylinder, and subjects perceived it as a rotating cylinder. However, when the display was seen through a triangular aperture, subjects reported perceiving a rotating cone. In this case, the velocity field was inconsistent with a rotating cone, but the shape of the aperture influenced the interpretation of the display. However, Ramachandran et al. didn't systematically manipulate the degree of discrepancy from projective consistency, or investigate how much discrepancy the visual system can tolerate.

2. EXPERIMENTS

As noted above, different models embody different visual constraints to estimate 3D structure from 2D image motion. One assumption that these models share is projective consistency: the inferred 3D structure must be projectively consistent with the image motion. However, in the context of figure/ground study mentioned above (where both sides of each border contained accretion/deletion), observers get a strong percept of 3D rotation even though the 2D speed profile is projectively inconsistent with 3D rotation. Can this happen with standard SFM displays where the region is always perceived as figurual and the figure/ground cue will not help with the 3D percept?

In order to understand the internal constraints that give rise to a 3D percept, we used dynamic random dots displays that were similar to traditional SFM displays: there were white dots moving inside an elliptical area (see Figure 3). However, in order to quantitatively manipulate the degree of discrepancy of the image motion from physical 3D rotation, we manipulated the speed profile of the dots' motion in two respects: first, the speed profile is defined as a linear combination of cosine speed and constant speed, and we manipulated α— the proportion of cosine speed that is projected from 3D rotation (higher speed in the center and going down to zero at the edges, see the section *Manipulation of speed* for details). If $\alpha=0$, the dot speed is constant (linear motion); if $\alpha=1$, the speed profile is cosine, consistent with 3D rotation (See Figure 5). Therefore, as α increases, the projective discrepancy with 3D rotation decreases. Second, we also manipulated the direction θ of the dot motion. If $\theta=0$, the direction is horizontal, and the motion is consistent with a rotating ellipsoid. As θ increases, the projective discrepancy

with 3D rotation increases. With 6 levels of α and 5 levels of θ, we have 30 different conditions overall (not including the counterbalancing of motion direction). In the following two experiments, we used the above stimuli in two different tasks—a rating task and a method of adjustment—to explore the influence of projective consistency on subjects' 3D percepts.

2.1 Experiment 1: Rating Task

In this rating experiment, we examined the subjects' 3D volumetric percepts using a continuous 7-point scale. We manipulated the discrepancy from projective consistency in the display and wanted to see how it influenced their volumetric percepts.

2.1.1   Method

Subjects

Eight Rutgers University graduate students with normal vision participated in the study. All were naive to the purpose of the experiment, and were paid for their participation.

Stimuli and Design

The stimuli for this experiment were 2D dynamic random-dot displays in a vertically oriented elliptical area (See Figure 3; the boundary of the ellipse was not part of the stimulus). The ellipse was 2.13° wide and 5.58° height. There are 165 white dots inside the black area, and the dot radius is 0.03°. The initial positions of the dots were randomly picked inside the ellipse in the first frame. The motion was generated by changing the dots' location across frames. From frame to frame, the dots moved in the same fixed direction (determined by θ), and each dot's speed was defined by a function

S(r), r represents the relative location in the 2D plane; we call this 2D speed profile. (See the following paragraphs for details.) Each display consisted of 90 frames and each frame lasted 1/30s, adding up to 3s.

We also did some control to clear up other possible confounding factors.

1) In most conditions, the speed of individual dots varied over frames, which led to dot density changes across the ellipse. To eliminate possible effects of texture density change, we kept the density uniform from frame to frame using a method proposed by Sperling et al. (1989): we divided the ellipse area into 10×10 grid of sub-squares, and kept the dot density in each sub-square constant and uniform in each frame by adding new dots and deleting old ones.

2) In some conditions, the dots at the boundary had essentially zero speed (see manipulation below), which resulted in an obvious illusory boundary. To minimize the influence of the illusory boundary, we introduced a 'threshold' to the dots' speed so that when speed is too close to 0, the dot would be forced to move with a very small speed. In addition, the background was filled with darker dots with the same density.

Figure 3: The 2D dynamic display used in the first experiment

Manipulation of speed

Suppose there is a vertically oriented ellipsoid rotating about its primary axis with certain angular speed ω, and we randomly sprinkle dots on its surface, so that the dots rotate together with the ellipsoid (Figure 4, left). If we make an orthographic projection of the dots (Figure 4, right), there are two main features about their image motion: 1) the motion direction of the dots is horizontal, i.e. it is orthogonal to the object's rotation axis, 2) Along the motion direction, its speed profile is a cosine function. Let's assume that there is a dot D at location (x,y) in the projection plane (see Figure 4, right). Let r be the

distance to the ellipse's primary axis, and R be the radius of an arbitrary cross-section. Then its 2D speed is a function of r:

$$C(r) = 2\pi\omega R \times cos(0.5\pi\frac{r}{R})$$

(1)

Here, the 2D speed C(r) is projectively consistent with a rotating ellipsoid. For example, when the dot reaches the primary axis of the ellipse, its speed will reach the maximum $2\pi\omega R$. When the dot reaches the two boundaries, its speed will reduce to 0. Let's call it the cosine speed profile.
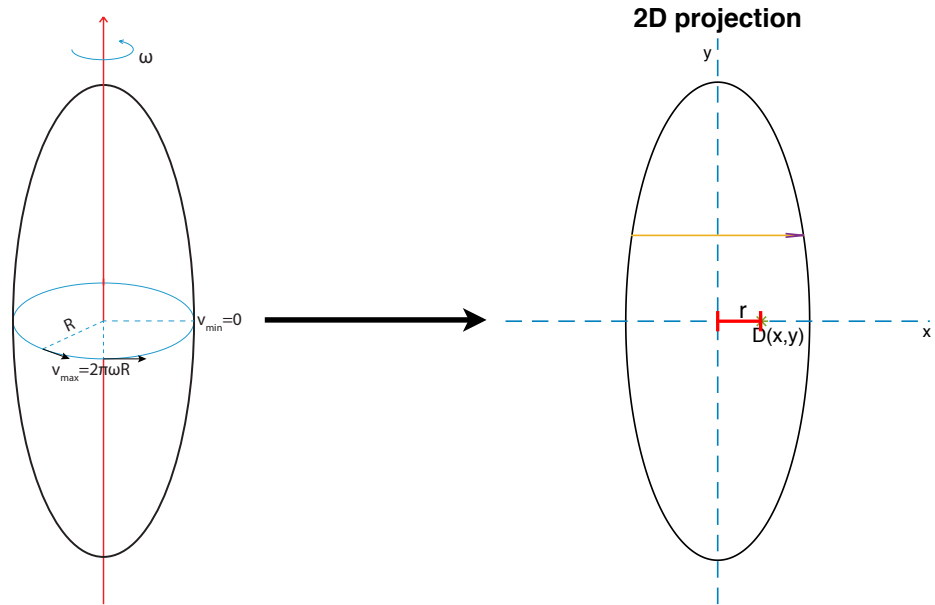
Figure 4: An ellipsoid rotating about its primary axis with angular speed ω, and its orthographic projection. R: the radius of an arbitrary cross-section. r: distant to the primary axis. So the dot's speed is determined by its location in the projection plane (R and r).

In order to explore the importance of projective consistency, we manipulated the proportion of the cosine speed profile and the motion direction quantitatively. Corresponding to the above two features, we manipulate:

1) Cosine weight α: the actual speed profile of dots S(r), is a linear mixture of the cosine speed profile C(r), and constant speed component L. α is the proportion of C(r). As a result, the actual speed S of the dot at location (x,y) is:

$$S(r) = \alpha C(r) + (1 - \alpha)L \tag{2}$$

r is the distance from the dot to the ellipse's primary axis. α can vary from 0 to 1. When α is 1, the speed is consistent with the rotating ellipsoid; when α is 0, the speed is constant (See Figure 5).

2) Motion direction θ: We also manipulated the motion direction relative to horizontal. θ can vary from 0 to 60°. When θ is 0°, it's consistent with the rotating ellipsoid.

Therefore, when α=1 and θ =0°, the speed profile is consistent with a rotating ellipsoid; when α=0, the speed profile is consistent with 2D translation. If α and θ is between the two extremes, there is no simple physical rotation that is consistent with the speed profile. Those are the conditions where we want to see how people's 3D percepts change.

In the two experiments, there were 6 levels of α equally spaced from 0 to 1, and there were 5 levels of θ equally spaced from 0 ° to 60 °.

In the first experiment, we also had a third independent variable determining constant speed L in equation (2), which we called 'constant speed component'. We used two values: in one condition, L was set to the peak speed of the cosine speed profile (in the condition of α =1 and θ=0°, and the peak speed was 0.84°/s); in the other condition, L was set to the average speed of the cosine speed profile (0.56°/s). Figure 5 shows the two conditions. When we combined two speeds L and C together, the actual speed S was faster in the first condition.

Therefore, based on α, θ and L, we had 60 (6×5×2) different conditions. For each condition, we had 4 repetitions including 2(up/down) × 2(right/left) motion directions for counterbalancing. There were 240 trials in total.

$$S(r)= \alpha \times C(r) + (1- \alpha) \times L$$
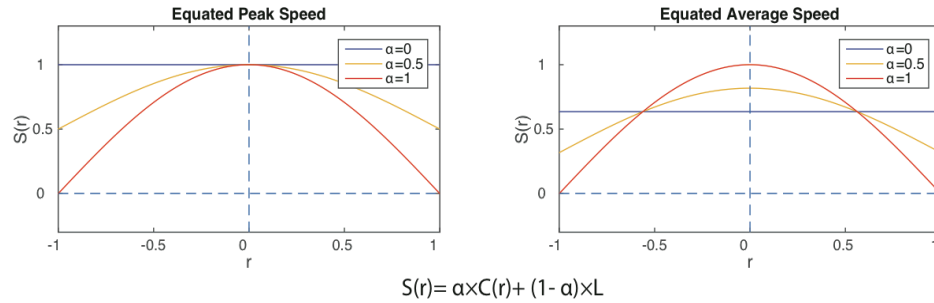
Figure 5: Actual speed of dots as a function of r with different α values and constant speed component (L). The left panel shows when L was set to the peak speed of the cosine speed profile; the right panel shows when L was set to the average speed of the cosine speed profile.

Procedure

The experiment was run in a quiet dark room. Subjects sat at distant of 105 cm from a HP monitor (refresh rate: 100Hz, with resolution 1280 × 1024 pixels) connected to a power Mac. The experiment was presented using Psychtoolbox on a mac version MATLAB (R2007b) (Brainard, 1997; Kleiner et al., 2007).

Before the experimental trial, instructions were shown on the screen. Then to get familiar with the rating scale, subjects were shown two examples that might be rated as 1 and 7. For '7', we used a classical structure-from-motion dynamic dots display in a cylinder, and for '1', we used the same display but the dots' speed inside the cylinder was constant. After watching the two examples, subjects pressed any key to enter the practice trials to get acquaint with the procedure and the scale. There were 18 practice trials with different α and θ values. After the practice trials, subjects were shown the two examples again, and then pressed any key to enter the actual experiment.

At the beginning of each trial, there was a fixation at the center of the screen for 500ms. Then the dynamic display described above was shown for 3s, followed by a flickering mask for 750ms (250ms for fixation and 500ms for the backward mask). A 7-point scale was shown after the mask, with the question '*To what extent do you perceive the shape as a rotating volumetric object*' on the top, and a slider at the bottom representing the scale. The slider is a bar with a continuous change of luminance. Dark color means low ratings, and bright color means high ratings. Numbers from 1 to 7 were also shown under the slider. 1 means 'definitely flat', 7 means 'definitely a rotating volumetric object'. Selecting a response somewhere between these two extremes would

mean that you get some sense of a volumetric object in the display, but it is somewhat weak. Subjects used the mouse to move the curser along the slider, and could choose any continuous number between 1-7 (e.g. 2.75, 3.4) by mouse click. There was no time limit. It went into the next trial automatically after a click was made. The 240 trials were divided into three 10-minute sessions, with 80 trials in each. Subjects could take a short break between sessions. The order of 240 trials was randomized for each subject. It took approximately 40 minutes for subjects to complete the whole experiment.

2.1.2 Results and Discussion

The graph in Figure 6 left plots the volumetric ratings as a function of α and θ. The x-axis represented different α values. Different lines represented different θ values.

A 3-way repeated measurement ANOVA was applied to the ratings, with independent variables α, θ and L. The main effects of the three independent variables were significant, with $F_{\alpha}(5,35)=66.464$, $p<.0005$, $F_{\theta}(4,28)=54.698$, $p<.0005$, $F_L(1,7)=7.411$, $p=0.029<.05$. The volumetric ratings increased significantly with α, and decreased significantly as θ became larger. The volumetric ratings were significantly higher when L was set to the average speed. However there's no interaction between the independent variables. Furthermore, we ran pairwise comparisons with Bonferroni correction for different levels of α and θ, respectively. For α, the ratings increased significantly when α increased from 0 to 0.6, but the effect was not significant beyond 0.6. The effect of θ was relatively small. The ratings decreased significantly when θ changed from 0° to 45°, and there was no significant difference between 45° and 60°.

The analysis showed that as α increased and θ decreased, volumetric ratings increased. However, by the time that α was 0.6, the ratings had already been asymptotic. Note that α=0.6 corresponds to a large deviation from projective consistency. And even when θ was as large as 45°, which was highly inconsistent with 3D rotation, the volumetric ratings were still high. The results suggest that projective consistency is not as important as previously thought for the visual system.

In addition, we suspected that different observers might use the same rating scale differently. Therefore we normalized each subject's data with respect to their own average ratings across all the conditions. The normalized volumetric ratings are shown in Figure 6 right panel. Then we also applied the same tests to the normalized rating. The results were the same as for the raw data.
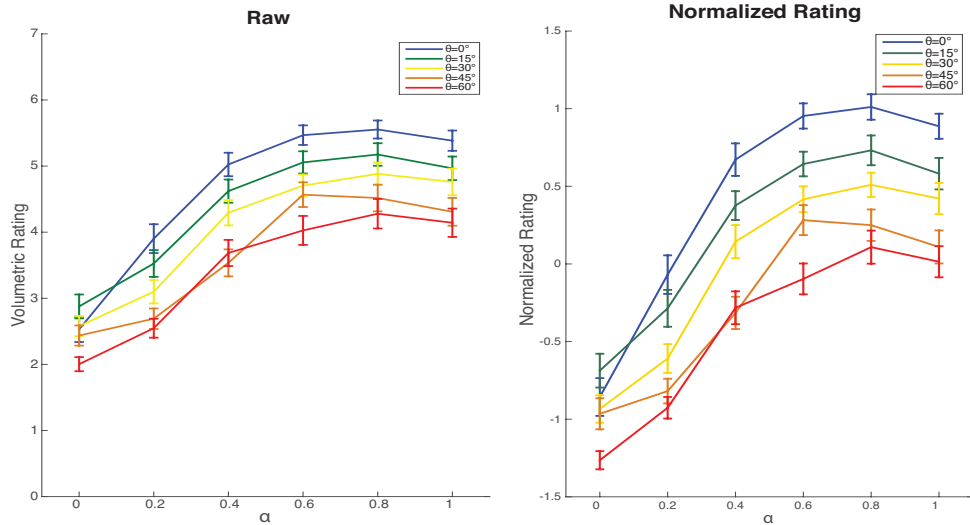
Figure 6: Results for the rating task as a function of α and θ. Left panel: raw ratings; Right panel: normalized ratings (Normalized by each subject's average rating across all the conditions.)

2.2 Experiment 2: Adjustment Task

In the above rating experiment, we examined the subjects' 3D volumetric percepts by using a 7-point scale. The simple task gives us an idea about how people's 3D percepts change as a function of α and θ. However, there are two possible problems with the rating task. First, subjects might have different understandings about 'volumetric percepts'. Second, they might use the same rating scale differently. So the same number on the scale might mean different 3D percepts for different subjects. This would introduce inter-subject differences that are not due to perceptual differences per se. To address the above concerns, we used another psychophysical task–method of adjustment. We asked the subjects to adjust the perceived depth of the 3D objects to match the given SFM displays. This makes the task more perceptually.

2.2.1 Method

<u>**Subjects**</u>

Eight Rutgers University graduate students with normal vision participated in the study. They were naive to the purpose of the experiment. All of them were paid for their participation.

<u>**Stimuli and Design**</u>

Figure 7 shows the snapshot of the adjustment task. We used the same SFM displays with the same α and θ values as Experiment 1. The parameters (such as the dimension and luminance) of the displays were also the same. There were only two differences: 1) experiment 1's results showed that constant speed component L did not interact with α and θ, therefore we only used one level of the L which corresponds to the

average speed of the cosine speed profile (0.56°/s) 2) there were no darker dots outside of the ellipse area and the background was black. As shown in Figure 7, the SFM display was shown on the left of the screen. At the same time, on the right of the screen, there were three different views of a same adjustable half ellipsoid: top-down view, side view, and mesh views. The height of the top-down view and the width of the side view corresponded to the depth of the mesh surface. The three views changed synchronically with the subjects' depth adjustment. The dimensions of the three views were scaled by a factor of 2/3 of the SFM displays. The depth of the mesh surface was adjustable by the subjects. Blue lines indicated the 0 longitude and 0 latitude. The view angle of the mesh surface is [1,4,0.5]. The initial depth of the mesh view was randomly picked. It could be adjusted from 0 to as much as 3 times of the mesh surface's width.
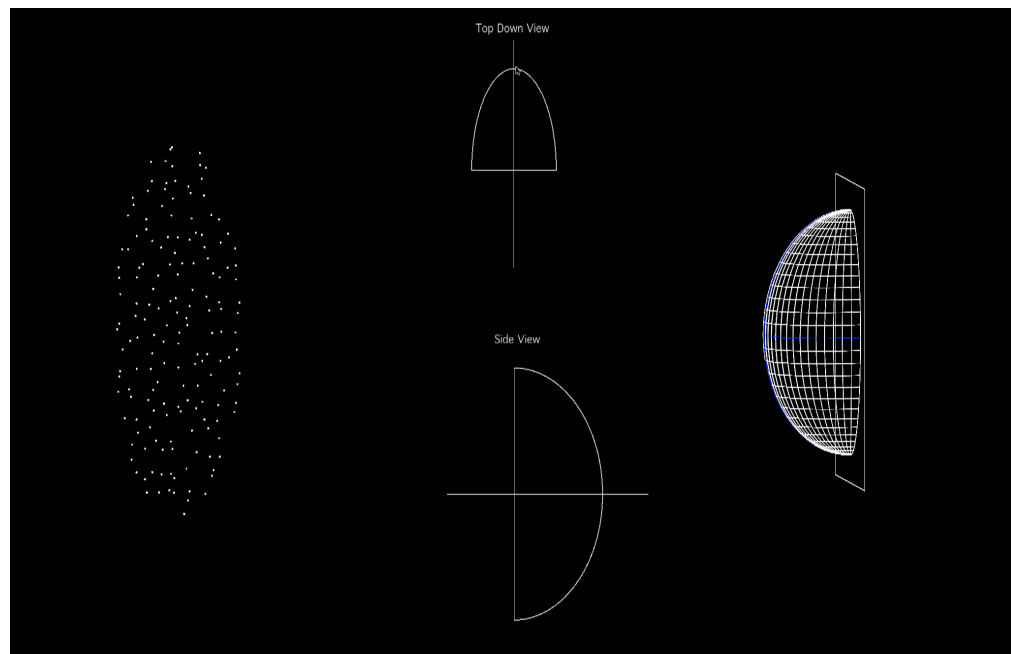
Figure 7: A snapshot from the adjustment task

Before the actual experiment, subjects had 9 practice trials to become familiar with the adjustment procedure. The stimuli for practice trials were similar to the actual experiment. The only difference was that the practice displays used a circular region rather than an ellipse (Thus when $\alpha =1$, $\theta =0$, the display would be consistent with a rotating sphere; See Figure 8). The radius of the circle in the image was the same as the width of the ellipse, which was 2.13°. The sizes of three views on the right were still 2/3 of the display on the left. There were 9 practice trials with 3 α values (0, 0.6 and 1) × 3 θ values (0°, 30° and 60°).
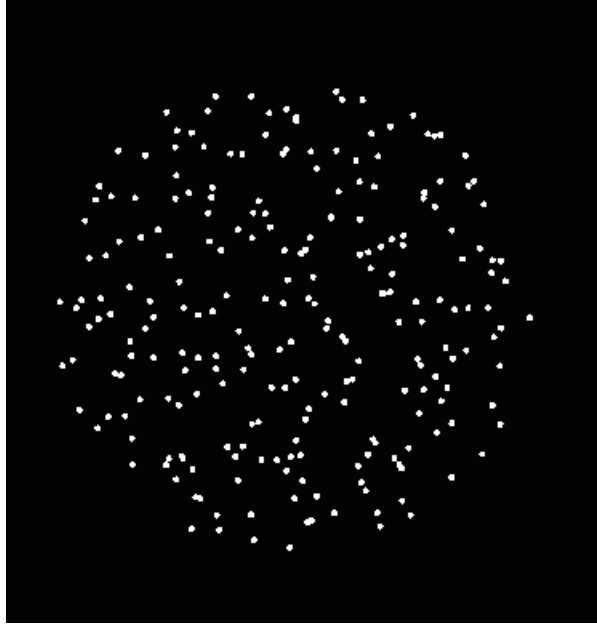
Figure 8: Stimuli for the practice trials

In summary, in the second experiment, there were 6 levels of α: 0, 0.2, 0.4, 0.6, 0.8, 1, and 5 levels of θ: 0°, 15°, 30°, 45° and 60°. The average speed was 0.56°/s. In total, we had 30 (6×5) different conditions. And for each condition, we had 8 repetitions including 2(up/down) × 2(right/left) motion directions for counterbalancing. There were

2 exact repetitions for each motion direction. Thus there were 240 trials for one participant.

Procedure

The experiment was run in a dark room. Subjects sat at 105 cm from an iMac monitor (85 Hz, 1280pxl × 1024 pxl) connected to an iMac. The experiment was presented using Psychtoolbox on mac version of MATLAB_R2014b (Brainard, 1997; Kleiner et al., 2007).

Before the experimental trial, instructions were shown on the screen. The task of the subjects was to adjust the depth of the ellipsoid (how deep & narrow, or how flat, it is) in order to match as closely as possible the 3D shape they perceived when looking at the SFM display. Subjects ran the practice trials first, and then pressed any key to enter the actual experiment.

At the beginning of each trial, there was a fixation at the center of the screen for 500ms. Then the dynamic display and the three views described above were shown, subjects adjusted the depth of the three views simultaneously by mouse, and press 'space' key when they finish adjusting. The ratio of the adjusted depth over the horizontal width of the ellipse was recorded. The display lasted until subjects finished adjusting and pressed the 'space'. There was no time limit for each trial. It went into next trial automatically after subject pressed the 'space' key. The order of 240 trials was randomized for each subject, and they were divided into three 10-minute sessions, with 80 trials in each. Subjects took a short break between different sessions. It took approximately 45 minutes to finish the whole experiment.

2.2.2   Results and Discussion

Figure 9 plots the volumetric ratings as a function of α and θ. The x-axis represents different α values. Different lines represent different θ values. The normalized depth is the ratio of the adjusted depth over the width of the ellipse. So when normalized depth is 1, the mesh ellipsoid has a circular cross-section. When normalized depth is smaller than 1, the ellipsoid will be flatter, and when it is larger than 1, the ellipsoid will be deeper.

A 2-way repeated measurement ANOVA was applied to the normalized depth settings, with independent variables α and θ. The main effects of α was significant, with $F_\alpha(5,35)=9.99$, p<.0001. In general, the volumetric percept increased significantly as α got larger. The trend of θ is similar to experiment 1 but the effect is not significant, with $F_\theta(4,28)=1.65$, p=.195. There was no interaction with α and θ, with $F_{(\alpha, \theta)}(20,140)=0.7317$, p=.787. Furthermore, we ran pairwise comparisons with Bonferroni correction for different levels of α. The adjusted depth increased significantly when α changed from 0 to 0.6, but plateaued beyond 0.6.

Perceived Depth(Normalized by Ellipse Width)
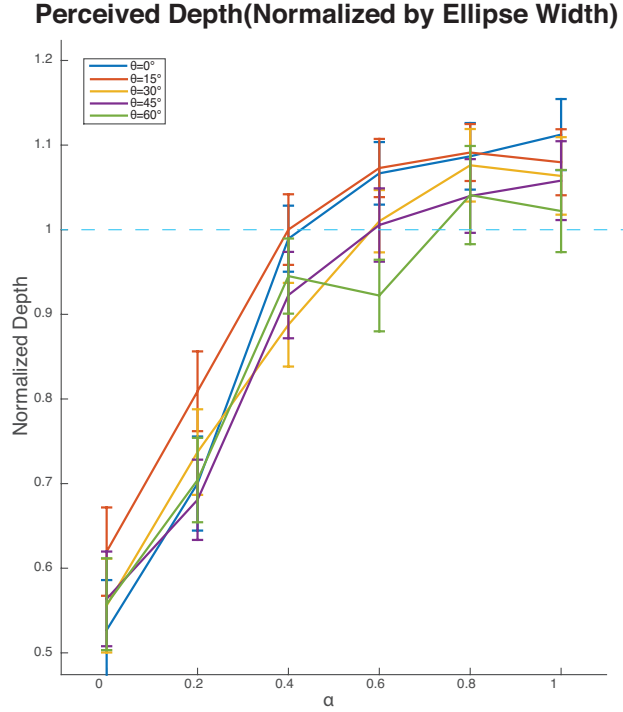
Figure 9: The mean (standard error) of the normalized depth with different α and θ

The results show that the effect of α is consistent with the first experiment. Subjects' volumetric percept increased as α increased from 0 to 0.6, but plateaued beyond 0.6, which suggests that people has high tolerance with projective consistency in terms of α. Moreover, when we used a more perceptual task in experiment 2, there was no effect of motion direction θ. Note that when θ is high, the image motion is highly inconsistent with 3D rotation. It suggests that the 3D percept relies more on α. The angle between the motion direction and the axis seems less important.

Combining the results from the two experiments, we can conclude that the visual system has high tolerance for deviations from projective consistency, and that projective consistency plays a less prominent role in SFM than previously thought

3. COMPARISON AGAINST THE RIGIDITY-BASED MODEL

In the above two experiments, we showed that projective consistency was not a strict constraint in SFM. One possible argument is that when α is not 1, the visual system will just treat the constant speed component (the deviation from projective consistency) as noise. It's hard to simply attribute the high tolerance to α and θ to noise, since when α was far away from 1 and θ is relatively large, the 3D percept was still strong. But to test for this possibility, we applied a classic rigidity based SFM model from computer vision that used factorization techniques to our displays (Tomasi & Kanade, 1992). The model is based on SVD factorization, and therefore it can deal with certain degrees of noise by only considering the three greatest singular values from SVD factorization.

The model takes the sequence of frames with 2d locations of dots as input. Based on SVD factorization, the model predicts the underlying 3D shape matrix and also the rotation matrix from frame to frame. Because the model assumes rigidity, the predicted 3D shape remains the same from frame to frame, and the 3D rotation matrix can vary.

Here, we compared the model predictions about the SFM displays used in the experiment with the human data in the following two ways:

3.1 Discrepancy between the model and the rating data

The model outputs a 3D shape matrix, and we calculated the orthogonal projection matrix from the matrix. Then we compared the discrepancy of the 2D matrix with the original 2D display (input) by calculating the Error Sum of Squares (SSE). If the model can predict the display well, the SSE should be very low. Therefore, higher SSE means larger discrepancy, and larger discrepancy means worse prediction. Given the 2D

display, if we have an observer that follows the rigidity assumption just like the model does, they should have the best volumetric rating for α=1 and θ=0°. And if α gets smaller and θ becomes larger, the discrepancy from rigidity of the display becomes larger, and the observer should have worse (or lower) volumetric ratings. In other words, the higher discrepancy (SSE) will imply lower ratings for the rigidity-based observer. However, Figure 10 shows that there is no systematic pattern for α and θ. This is inconsistent with human data from experiment 1, where the ratings showed a systematic dependence on α and θ (Recall Figure 4). The comparison suggests that rigidity assumption may not be the constraint or the only constraint that human observer used when rating our displays.
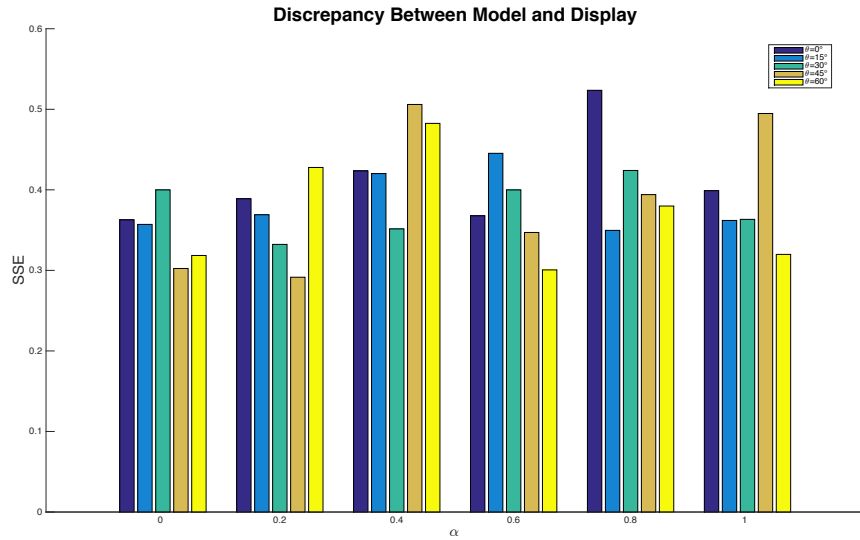
Figure 10: Discrepancy (SSE) between the model and the display. x-axis indicates different α values, and different color bars indicate different θ values.

3.2 Depth comparison between the model and adjustment data

From the predicted 3D shape matrix, we also calculated the depth of the half ellipsoid–let's call it 'predicted depth'. Then we compared the predicted depth with the

human perceived depth from the adjustment task. Figure 11 shows the predicted depth and the perceived depth from the adjustment task over different α and θ values.

Comparing the two panels in Figure 11, There are mainly three differences between model prediction and human data: 1) subjects overestimated the depth than the model when α is 0. In this condition, the speed of dots is constant, which is consistent with 2D translation, and the model predicts zero depth. However, subjects still perceived some depth for constant speed. Moreover, the perceived depth was almost half the depth of the ellipsoid with circular cross-section. 2) When θ increases from 0° to large degrees (such as 45° and 60°), the predicted depth increases dramatically, and when α is above 0.4, θ is 60°, the predicted depth exceeds the circular cross-section. This means that when the motion direction deviates highly from horizontal, the model overestimates the depth compared to human observers. However, results from experiment 2 showed that θ had no effect on the perceived depth, which means that human observers didn't care about the motion direction when performing the task. 3) The maximum perceived depth in the human data is around the circular cross-section, which suggests that human observers didn't overestimate the depth compared to the model.

In summary, the comparison of depth shows that 1) the rigid SFM model can't explain why human observers had some volumetric percepts when the speed was constant; 2) The predicted depth from the model increases as θ increases, while the perceived depth doesn't rely on θ; 3) The model overestimates the depth with large θ, which is inconsistent with the perceived depth from human data.
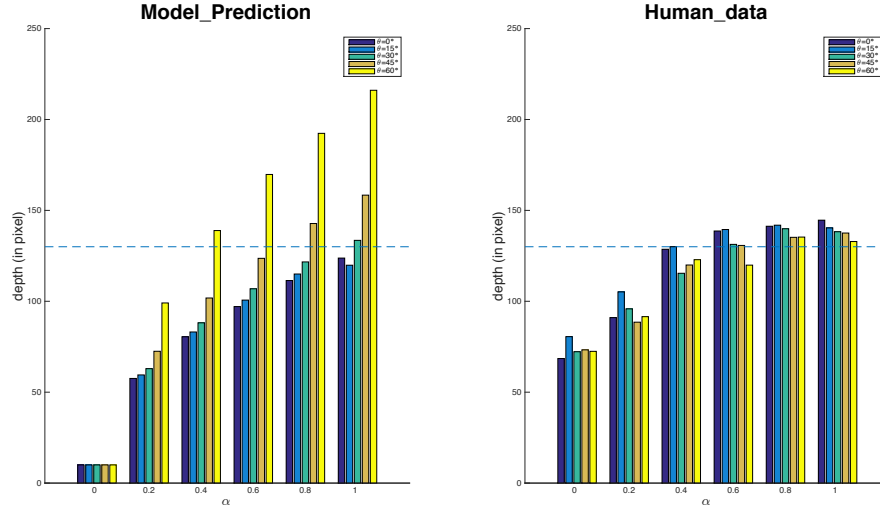
Figure 11: The predicted depth from the model (left) versus the perceived depth from the adjustment task (right). X-axis indicates different α values, and different color bars indicate different θ values. The dash line showed the circular cross-section.

3.3 Predicted direction of rotation axis from the model

Why does the model predict differently from human perception? One possible explanation is that when θ is large or α is small, i.e., the display's discrepancy from a physical possible rigid rotating object gets larger, the model needs to do something to maintain the rigidity assumption. We suspected that the model achieved this by continually changing the predicted axis of rotation from frame to frame—which would be very different from a human observer's percepts. To test for this, we did a more detailed analysis from the recovered 3D shape: we recovered the rotation axis from each frame. The results were shown in Figure 12. The left panel shows the orientation vectors of the rotation axis over each frame. Each intersection point on the mesh grid represents one combination of α and θ. The x-axis indicates different θ values; y-axis indicates different

α values. At the same time, the mesh grid is parallel to the 2D projection plane, so the z-axis corresponds to the line of sight that is orthogonal to the projection plane. For example, if we have a rotating ellipsoid as Figure 4 shows, the direction of rotation axis should be parallel to y-axis.
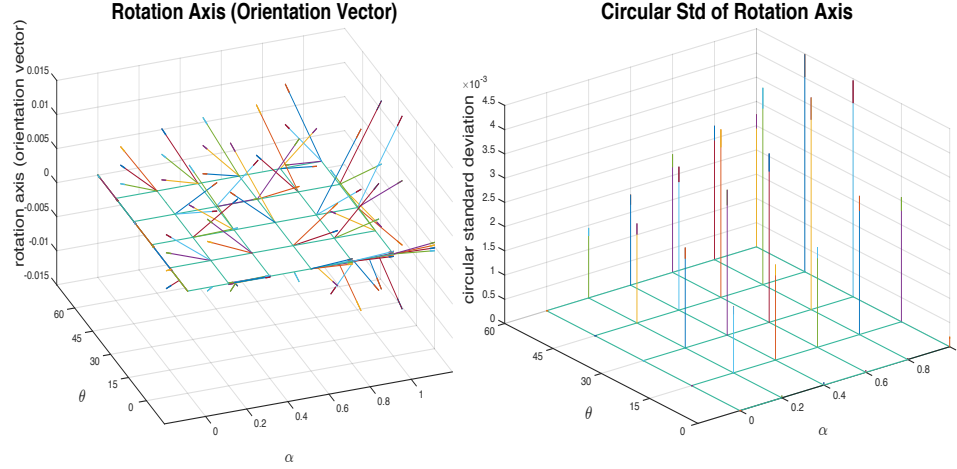
Figure 12: The predicted orientation of rotating axis (left) and change of the axis (right) over different conditions.

The figure shows that when θ=0°, the directions of rotation axis are generally consistent with a rotating ellipsoid. However, as θ gets larger, the rotation axis becomes less stable over different frames. There is no obvious pattern for α.

To see the patterns of change more clearly, we calculated the circular standard deviation of rotation axis over different frames, which were shown in the right panel of Figure 12. Now the trend is clearer: for a fixed α, as θ increases, the variance of the rotation axis becomes larger; for a fixed θ, the variance of the rotation axis also increases as α increases.

The change of rotation axis corroborates the above guess. Of the 30 displays with different α and θ combinations, only six have physical possible motion: when α=1, θ=0°, the moving pattern of the display is consistent with a rigid rotating ellipsoid; when α=0, regardless of θ, the display is consistent with 2D translation. The other 24 conditions have different degrees of discrepancy from rigidity– as θ and α increases, the discrepancy from rigidity also increases. For the rigid SFM model, to keep the shape constant over frames, it has to counteract the effect of the discrepancy by means of a rotation axis change over frames. Therefore, as α and θ increases together, the discrepancy increases, and the rotating axis change also increases. The result was corroborated by the circular STD change in the right panel of Figure 12. Consider again in Figure 11, as θ increases, the predicted depth also increases. From the rotation axis analysis here, we can see that even the predicted depth is not meaningful given that the predicted rotating axis keeps changing.

4. GENERAL DISCUSSION

In human structure-from-motion literature, a vital and interesting question is what assumption the visual system uses to recover the underlying 3D structure. There are many models such as Ullman (1979)'s rigidity assumption and the incremental rigidity assumption (Ullman, 1985), Braunstein's (1984) constant velocity assumption and others. One assumption in common underlying these models is that they regard projective consistency as the necessary and sufficient assumption for human to recover the underlying 3D structure. In another word, to perceive the correct 3D structure, the 2D image motion should be consistent with 3D rotation. However, our results show that it plays a less dominate role than expected in SFM.

With our two manipulations, α and θ, we quantitatively manipulated the degree of discrepancy between image motion and projection that are consistent with physical rotation: as α gets away from 1 and as θ gets larger, the discrepancy increases. If projective consistency is the only and vital assumption visual system uses, as the discrepancy increases, their performance should get worse. However, our results from the two tasks show that this is not simply the case. In one aspect, when the motion direction is correct ($\theta = 0°$), as α increases from 0 to 1, the proportion of cosine speed increases, which means that the image motion becomes more consistent with the physical rotation. And as α reaches 1, it should be completely consistent with physical rotation. However, the volumetric percepts and the depth adjustment plateaued beyond α was 0.6, which suggests that for human observers, the volumetric percepts when $\alpha = 0.6$ was as strong as $\alpha = 1$. Therefore, it suggests that in our case, the volumetric perception doesn't require strict cosine motion profile. There might be some other constraints for the visual system.

In another aspect, the two experiments showed small effect of the motion direction θ. The 3D percepts were strong in spite of the fact that as θ got larger, the image motion becomes less physically possible. This suggests that the 3D perception is not influenced by motion direction as much as α. From the above two aspects, we can conclude that subjects' 3D percepts are more sensitive to cosine speed profile and less sensitive to the direction of motion; and overall, α and θ shows that projective consistency plays a less prominent role in SFM than previously thought.

Even when the deviations of α and θ are so huge, the 3D percepts are still strong compared to the percepts of projectively consistent display, so it's hard to attribute the effects to random noise. We further tested this assumption by the rigidity-based SFM model from computer vision. The model is based on rigidity assumption and it is able to deal with certain degrees of noise from strict rigidity. If the effects of α and θ are due to noise, the rigidity-based SFM model should be able to simulate human observers percepts well. However, our comparisons showed that the model behaved differently from human observers. This suggests that projective consistency alone is not a necessary and sufficient assumption for human SFM.

In summary, the two experiments show that subjects' volumetric percepts plateau beyond α of 0.6, and are still strong when θ is as large as 45°. This suggests that human 3D percepts are surprisingly resilient to deviations from projective consistency, with speed profiles that are substantially constant-velocity (up to about 40%) still perceived as 3D. The model comparisons also convinced that the effects of α and θ are not simply due to noise, and that projective consistency is not necessary for human SFM.

There have been lots of studies that tried to explore what constraints human use in SFM, and researchers assume that projective consistency is a necessary constraint for the visual system. However, our results showed that people had vivid 3D percepts when projective consistency was violated (most of our displays were not consistent with 3D rotating objects). And their judgments of the volumetric shapes had systematic patterns with the degree of deviation from projective consistency. These results suggested that the visual system has large tolerance on the discrepancy of projective consistency, and therefore projective consistency may not be as vital as previously thought in SFM.

Moreover, our studies have a few advantages over previous studies. Firstly, compared with figure/ground studies mentioned in the introduction (Froyen et al., 2013), our studies used standard SFM displays where the ellipse region is always perceived as figure, so that it eliminated the possibility that the 3D perception was due to the figure/ground cues. Secondly, compared to Ramachandran et al. (1988)'s rotating cone study, our study manipulated the projective consistency in a more systematic way.

In addition, our studies also proposed a new way to generating SFM displays and explore the constraints used in SFM. The traditional method usually generates the displays from the projections of simulated or real 3D objects, while our method generated the SFM displays directly from 2D velocity field that defined the image motion of the dots. Although there are studies that also use velocity information in their computational models (Koenderink & Van Doorn, 1986; Jain & Zaidi, 2011, et al.), they normally follow the traditional way of stimuli generation and therefore presume projective consistency. But here, we used 2D velocity field directly, so it doesn't require an underlying 3D object. The advantage of our method is that it has more freedom to

systematically manipulate projective consistency. One may argue that the traditional way can also manipulate projective consistency by adding noise after generating the projections. However, our results suggest that the effects can't be simply explained by random noise, because in the display, the velocity field had large and systematic deviation from projective consistency, but human observers had the plateau effect when the proportion of cosine speed is beyond 0.6. If it's due to random noise, the 3D percepts should decrease gradually as the noise increases. Moreover, the prediction from the rigidity-based SFM model (which can deal with certain levels of noise) is different from human data.

The role of projective consistency is challenged by our results, as well as some previous studies in terms of motion and shape (Ramachandran, Cobb & Rogers, 1988). The current studies manipulated the degree of projective consistency by manipulating the velocity field. In particular, we used the same ellipse-like contour, and manipulated the 2D speed profile by the proportion of cosine speed profile (α) and motion direction (θ). But with the speed profile method, we can also manipulate the degree of projective consistency in a different way—by manipulating the shape of contour. The manipulation of shape might be a better and stronger way to explore the role of projective consistency. Here are three main reasons: firstly, the shape manipulation will break projective consistency more dramatically. When a rotationally symmetric object is rotating about its symmetric axis, the occluding contour remains constant; but if it is not rotationally symmetric, the contour will vary in time as it rotates. If we keep the shape of contour fixed over rotation, some shapes will have higher degree of projective inconsistency than others. For example, asymmetric shapes will have more inconsistency than symmetric

shapes. Secondly, our current studies, as well as previous related studies, suggest that there might be some interactions between the contour shape and velocity information. For example, in Ramachandran (1988)'s displays, subjects reported perceiving a rotating cone when a triangular aperture was added, despite the constant speed of the texture. In the figure/ground study (Froyen et al. in 2013), compared with asymmetric/concave areas, the symmetric/convex areas are more likely to be perceived as rotating columns despite the constant texture speed. Thirdly, traditional SFM computational models often ignore the shape information, which leaves a rich area for us to explore. Therefore it would be interesting to explore how the change of contour shape influences the 3D percepts.

Convexity and symmetry are two important properties of shape. The figure/ground study showed that convex and symmetric areas were more likely to be perceived as rotating objects (Froyen et al., 2013). This raised the question of whether it would be true of SFM displays where the moving region is always interpreted as figure, and the figure/ground cue cannot help to promote the 3D percept. Therefore, it's worthwhile to explore the influence of convexity and symmetry on SFM. In particular, whether the change of contours' degree of symmetry and convexity will change the strength of SFM percepts, and if so, will it change in a systematic way and why? We can apply the speed profile method to contours with different degrees of symmetry or convexity, and generate a series of SFM displays with different values of proportion of cosine motion—α (α can be from 0 to 1). Then, we can explore the minimum values of α required for people to perceive the displays as rotating 3D objects. The minimum values of α is as an index of deviation of projective consistency. If symmetric/convex shapes have lower α than asymmetric/concave shapes, it will demonstrate that the two properties

of shape indeed influence the interpretation of motion information, and it again implies that projective consistency is not necessary for SFM.

BIBLIOGRAPHY

Anandan, P., & Irani, M. (2002). Factorization with uncertainty. *International Journal of Computer Vision*, *49*(2-3), 101-116.

Brainard, D. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4), 433–436.

Braunstein, M. L., & Anderson, G. J. (1984). Shape and depth perception from parallel projections of three-dimensional motion. *Journal of Experimental Psychology: Human Perception and Performance*, 10(6), 749.

Braunstein, M. L. (1994). Structure from motion. *Visual detection of motion*, 367-393.

Clocksin, W. F. (1980). Perception of surface slant and edge labels from optical flow: a computational approach. *Perception*, 9(3), 253-269.

Froyen, V., Feldman, J., & Singh, M. (2013). Rotating columns: Relating structure-from-motion, accretion/deletion, and figure/ground. *Journal of vision*, 13(10), 6.

Husain, M., Treue, S., & Andersen, R. A. (1989). Surface interpolation in three-dimensional structure-from-motion perception. *Neural Computation*, 1(3), 324-333.

Kaplan, G. A. (1969). Kinetic disruption of optical texture: The perception of depth at an edge. *Perception & Psychophysics*, 6(4), 193-198.

Koenderink, J. J., & Van Doorn, A. J. (1986). Depth and shape from differential perspective in the presence of bending deformations. *JOSA A,* 3(2), 242-249.

Jain, A., & Zaidi, Q. (2011). Discerning nonrigid 3D shapes from motion cues. *Proceedings of the National Academy of Sciences*, 108(4), 1663-1668.

Kleiner, M., Brainard, D. H., & Pelli, D. G. (2007). What's new in Psychtoolbox-3? *Perception*, 36, ECVP Supplement 14.

Morita, T., & Kanade, T. (1997). A sequential factorization method for recovering shape and motion from image streams. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *19*(8), 858-867.

Ramachandran, V. S., Cobb, S., & Rogers-Ramachandran, D. (1988). Perception of 3-D structure from motion: The role of velocity gradients and segmentation boundaries. *Attention, Perception, & Psychophysics*, 44(4), 390-393.

Tanrıkulu, Ö. D., Froyen, V., Feldman, J., & Singh, M. (2016). Geometric figure–ground cues override standard depth from accretion-deletion. *Journal of Vision*, *16*(5), 15-15.

Thompson, W. B., Mutch, K. M., & Berzins, V. (1985). Dynamic occlusion analysis in optical flow fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (4), 374-383.

Todd, J. T. (1982). Visual information about rigid and nonrigid motion: A geometric analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 8(2), 238.

Tomasi, C., & Kanade, T. (1992). Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2), 137-154.

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

Ullman, S. (1983). Ullman, S. (1984). Maximizing rigidity: The incremental recovery of 3-D structure from rigid and nonrigid motion. *Perception*, *13*(3), 255-274.

Wallach, H., & O'connell, D. N. (1953). The kinetic depth effect. *Journal of experimental psychology*, 45(4), 205.