Discerning Non-rigid 3-D Shapes from Motion Cues

Anshul Jain and Qasim Zaidi

Graduate Center for Vision Research, SUNY College of Optometry, New York, NY 10036

Corresponding authors:
Qasim Zaidi and Anshul Jain
Graduate Center for Vision Research
33W 42nd Street, New York City
NY 10036, USA

E-mail: qz@sunyopt.edu
       anshuljain@gmail.com

Phone: 212-938-5542
Fax: 212-938-5537
Abstract: Many organisms and objects deform non-rigidly when moving, requiring perceivers to separate shape changes from object motions. Surprisingly, the abilities of observers to correctly infer non-rigid volumetric shapes from motion cues have not been measured, and structure-from-motion models predominantly use variants of rigidity assumptions. We show that observers are equally sensitive to discriminating cross-sections of flexing and rigid cylinders from motion cues, when the cylinders are rotated simultaneously around the vertical and depth axes. A computational model based on motion perspective (i.e. assuming perceived depth inversely proportional to local velocity) predicted the psychometric curves better than shape-from-motion factorization models using shape or trajectory basis functions. Asymmetric percepts of symmetric cylinders, arising due to asymmetric velocity profiles, provided additional evidence for the dominant role of relative velocity in shape perception. Finally, we show that inexperienced observers are generally incapable of using motion cues to detect inflation/deflation of rigid and flexing cylinders, but this handicap can be overcome with practice for both non-rigid and rigid shapes. The empirical and computational results of this study argue against the use of rigidity assumptions in extracting 3-D shape from motion, and for the primacy of motion deformations computed from motion shears.

3D shape from motion | non-rigid 3D shapes | optic flow

Introduction

All animals, fish, insects, birds, and reptiles, and many mechanical objects such as cranes, earthmovers, harvesters, and bicycles change shape as they move. Some shape changes, e.g. bending and straightening of legs, create forward propulsion, others such as torso flexes and head bobs serve to maintain balance, while others like head turns may be unrelated to object motion. Humans seem to be quite good at disentangling shape changes from object motion, and in characterizing motion as tumbling, rolling, swaying, stretching, leaping, spinning, flapping, dancing, kicking, bucking, jerking, sliding, gliding, tripping, or shaking. A large number of studies have examined human perception of rigid 3-D shapes from motion cues (e.g. 1, 2-5), but very few have examined non-rigid shape perception (6-8), and these have not dealt with what shapes are perceived.

Structure-from-motion models fall into three broad theoretical classes. In the incremental rigidity scheme (9, 10), the visual system assumes that the object is rigid or approximately rigid in order to extract the 3-D shape. The perceptual system creates a 3-D model of the object that is continuously updated to minimize non-rigid deformations across frames. This inverse optics approach seeks to compute the exact Euclidian (up to a scale and rotation) structure that would project to the 2D images. However, it is very sensitive to measurement noise and thus requires accurate frame-by-frame feature tracking. In addition, often, the human perceptual system does not extract a rigid Euclidian or Affine structure that corresponds to the moving object (11, 12).

The most common retinal motions arise from the movements of observers, and in this case, local retinal velocities are inversely proportional to distances in a given scene (3, 13, 14). Similarly retinal velocities arising from an object translating in front of a stationary observer are inversely proportional to distances of different parts of the object. Models that assume that this principle is used by the visual system to extract depth from relative velocities will be called motion-perspective models. Psychophysical (15-17) and physiological (18-21) evidence for relative velocity detectors suggests that they could play an intermediate role in computing 3D shape, and electrophysiological studies have implicated cortical area MT, which contains such neurons, as having a significant role in computing 3D structure from motion (22-26). This approach has been shown to be in general agreement with human perception of rigid objects (3, 27, 28), but has not been tested on non-rigid motion.

Under general conditions, i.e. motions of observers and objects that include rotations, and shape changes of objects, motion leads to deformations of the pattern of retinal-image velocities (optic flow). Once translation parallel to the image plane has been factored out, a particularly useful parsing of the optic flow is in terms of the kinematic differential invariants curl, div and def (29). Of these only the def component carries information about the structure of the...
scene, and has formally been linked to local slants and tilts (29-32). Building on this formal work, experimental studies have shown that that perceived slant is proportional to the def component of the optic flow (33), a def based heuristic model has been suggested for computing local slants and tilts of planar stimuli (4, 34 for review), and def has been used in a template model for volumetric rigid stimuli rotating along arbitrary axes (35). Physiological studies on nonhuman primates and imaging studies on humans have found neurons in area MST (36-38), STPa (39, 40), and STS (41), that preferentially respond to the curl and div components of optic flow but not to the def component. Lagae et al. (42) reported a few cells that preferentially respond to the def component in area MSTd. In addition, def can be composed from a pair of motion shears or compressions, and a number of studies have found cells in area MT that can signal motion shear (18-21).

Non-rigid structures have been studied extensively in one domain, i.e. point-light depictions of biological motion (43), where humans can identify gender (44), emotion (45) and other attributes from moving lights placed at the joints of articulation. How this is done by the brain is an open question, but some neural models postulate the use of “snapshot” neurons that are tuned to different poses of a deforming humanoid shape (46). However, it is unlikely that snapshot neurons exist for very many objects in the large class of deforming shapes.

In machine vision, moving rigid shapes are most often extracted using variants of Tomasi and Kanade’s factorization of the image-stream matrix into a 3-D shape matrix and an orthonormal rotation matrix (47). For non-rigid shapes, the shape matrix changes on every frame, but since the shapes generally change continuously with smooth movements, the factorization can be extended by either constraining form as a linear combination of a small set of basis-shapes derived from the images, or by constraining motion as a linear combination of a small set of basis trajectories, e.g. oscillations (48-51).

The current study measured observers’ ability to discern non-rigid shapes from motion cues, and examined different models for explaining these capabilities. We began with asking observers to discriminate deviations from perfect circularity for rotating rigid and flexing cylinders. We found that observers were as sensitive to deviations from circularity for non-rigid shapes as for rigid shapes. A motion perspective model was better than a trajectory basis model in predicting average performance. In addition, we found that when perspective projection of symmetric stimuli created asymmetric retinal velocity fields, observers reported asymmetric percepts. Finally we examined how performance degrades in the presence of multiple non-rigid deformations, i.e. for cylinders that inflated and deflated while flexing in the image plane or depth.

Figure 1: Sample frames for simultaneous rotation about the vertical and depth axes for (A) a rigid cylinder, (B) a cylinder flexing in the image plane, and (C) a cylinder flexing in depth. The regular grids of dots in the figure are only for illustration purposes, in the experiments, dots were randomly placed after surface generation to remove density and texture cues to shape, and cylinders were presented behind a frame so that the curved edges were not visible to the observer.

Results

Experiment 1 – Discriminating Shapes using Motion Cues: Observers were instructed to report whether the cross-sections of rigid and flexing cylinders (Figure 1) were shallower or deeper than perfect circles. The cylinders were rotated simultaneously about the vertical axis in the image plane and the depth axis orthogonal to the image plane. The diameter of the cylinder in
the image plane was kept constant, but the cylinder’s depth was varied to set the aspect ratio at one of [0.7071, 0.8409, 1, 1.1892, 1.4142] selected randomly on each trial. Observers ran 20 blocks of 45 trials each spread over two sessions. Each session lasted about 45 minutes.

Figure 2A shows the average percent of trials perceived as deeper than a perfect circular cross-section for the four types of cylinders (two rigid and two flexing). Points of subjective circularity (PSC) estimated from the smooth fits are shown in Figure 2B. Observers’ judgments of aspect-ratio were generally slightly shallower than veridical, i.e. PSCs were greater than 1.0 (F(3,28) = 5.6947, p<0.01), except for the significantly smaller PSC of the cylinder flexing in the image plane (p<0.05). More importantly, the four psychometric functions had similar slopes (F(3,28) = 0.9435, p = 0.43). This result demonstrates that non-rigidity does not affect the sensitivity of human observers in discriminating shapes extracted from motion cues. Since rigidity assumptions are not useful in discerning the shapes of flexing cylinders, and the similarity between the slopes of the psychometric curves make it unlikely that entirely separate neural processes are used for flexing than for rigid cylinders, these results suggest that percepts of rigidity may arise as a result of a more general prior assumption. Even though observers were able to perform this task in a consistent manner (% perceived deeper varied monotonically with the aspect ratio, and mean PSCs were within 20% of veridical), the sensitivity observed here is substantially lower than that for discriminating elliptical departures from circular shapes in 2-D (52).

**Computational Analyses:** In order to understand how observers were accomplishing the experimental tasks, we analyzed our stimuli using the global motion-perspective approach (3, 13, 14) and, the trajectory-basis approach (49).

**Motion Perspective Model:** Using the heuristic that the image velocity of every point is inversely related to the distance of the point from the optical center of the observer (27), we computed a *velocity contrast metric* (VCM) for each narrow cross-sectional slice as, $VCM = (V_{max} - V_{min}) / (V_{max} + V_{min})$, i.e., the normalized difference between the maximum and minimum velocity along the cross-sectional slice (Figure 3A). To compute the VCM for the entire trial, the VCM was averaged across each cross-sectional slice on a single frame and then across all the frames. *Trajectory-basis Model:* We applied the algorithm of Akhter et al. (49) to compute the structure of the deforming cylinder. Taking advantage of the fact that the cylinders were deforming smoothly, the 3-D path taken by each dot on the cylinder was modeled as a linear combination of simple trajectories formed by Discrete Cosine Transforms. The number of harmonics used as basis trajectories was empirically determined to be three for the two flexing cylinders. Once the structure was determined (up to an arbitrary scale and rotation) we aligned the cylinder along its length and computed the aspect ratio on each frame. The aspect ratio for
the trial was computed by averaging the aspect ratio across all frames. The schematic in Figure 3B shows how a complex smooth trajectory can be modeled as a linear combination of cosine oscillations. It should be noted that the shape-basis approach (50) recovers similar non-rigid structures from our stimuli. We chose the trajectory basis functions due to the greater potential for generalization: the same trajectory basis functions can be used for a wide range of 3D shapes, but a novel shape basis needs to be computed each time for a different shape.

\[ \text{VCM} = \frac{V_{\text{max}} - V_{\text{min}}}{V_{\text{max}} + V_{\text{min}}} \]

**Figure 3:** The two modeling approaches: (A) Computation of the *velocity contrast metric* (VCM) for a cross-sectional slice on the cylinder. (B) A complex but smooth feature trajectory modeled as a linear combination of sinusoidal trajectories.

The simplistic motion perspective model explains the observers’ bias, namely, the tendency to perceive non-rigid cylinder flexing in the image plane deeper than rigid cylinders, quite well. The order of the curves is very similar (albeit the biases are slightly exaggerated) to the order in the psychophysical data (Figure 2C). The VCM is highest for the cylinder flexing in the image plane and smallest for the faster-rotating rigid cylinder. Interestingly, the faster rotating rigid cylinder was perceived as shallower than the slower rotating rigid cylinder, indicating that perceived depth magnitude depends more on velocity contrast than on absolute speed. These simulations suggest that observers rely on relative velocity information to infer 3D shapes, at least for these qualitative judgments. On the other hand, the aspect ratio computed by the trajectory-basis model is very similar for all the cylinders, and slightly higher than the veridical aspect ratio (Figures 2D). Thus this model does not explain the perceptual biases observed in the psychophysical data, particularly the bias to perceive cylinders flexing in the image plane as deepest.

**Experiment 2 – Asymmetric Percepts from Symmetric Cylinders:** The cylinders presented in Experiment 1 were perfectly symmetric around the central transversal plane orthogonal to the image plane, however, while making observations in the experiment, the authors noted that some cylinders appeared asymmetric. Experiment 2 was designed to measure this systematically. The stimuli were exactly the same as in Experiment 1, except that the aspect ratio was fixed at one of \([0.7071, 1, 1.4142]\) selected pseudo-randomly on each trial. On half the trials, selected pseudo-randomly, the cylinder rotated about both the depth and vertical axes (Compound Motion condition as in Experiment 1), and on the other half, the cylinder rotated only about the vertical axis (Simple Motion condition). The observers’ task was to report whether the cylinder had a symmetric or asymmetric profile by pressing one of two keys. Observers ran 12 blocks of 54 trials each. The entire session lasted about an hour.

**Figure 4A** shows the percentage of trials perceived as asymmetric as a function of aspect ratio for the rigid and flexing cylinders for the two types of motion. Cylinders were perceived as asymmetric more often when undergoing complex motion than when undergoing simple motion (\(F(1,5) = 22.574, p<0.01\)) and shallow cylinders were perceived as asymmetric more often than deep cylinders (\(F(2,10) = 4.915, p<0.05\)). Further, there was also a main effect of non-rigidity (\(F(2,10) = 7.610, p<0.01\)), particularly cylinders flexing in the image plane were perceived as asymmetric more often than the other two cylinders.
To understand the perceived asymmetry, we looked at the velocity profiles generated by the stimuli. Figure 5A and 5B show the velocity profiles for rigid cylinders under simple and complex motion, respectively. As the figure shows, the velocity profiles are symmetric about the horizontal midline for simple motion, but distinctly asymmetric for complex motion. The peak velocity is shifted above the midline. This shift in peak is a direct consequence of the compound rotation present during the trial. When we reversed the phase of the two simultaneous rotations the peak velocity and consequently the perceived asymmetry were shifted below the horizontal midline.

![Velocity profiles for plane-flex cylinder with an aspect ratio of 0.71 under simple motion (A) and complex motion (B).](image)

To judge whether the two models could simulate the perceived asymmetry, we calculated an asymmetry metric (AM) as the mean-squared error between the veridical symmetric shape and the shapes computed by the trajectory basis and motion perspective models. The shapes were normalized to have the same range and magnitude before computing the error. Figure 4B and 4C show the computed AM for the two models. Neither of these models explains the psychophysical data perfectly. The motion perspective model shows that asymmetry decreases with increasing aspect ratio as observed in the data and also shows higher asymmetry for complex motion than for simple motion as found in the experimental data, however the asymmetry for cylinders flexing in depth was highest unlike the experimental data. The trajectory basis model does predict that cylinders flexing in the image plane appear most asymmetric, but the bias is grossly exaggerated because the cylinders flexing in depth and the rigid cylinders have very little asymmetry for high aspect ratios, even for complex motion. The results of this experiment provide further evidence that observers rely on relative velocity measurements to compute 3D shape, both to estimate qualitative properties as in Experiment 1, and to estimate...
global shape properties like symmetry. This provides an interesting counterpart to results showing that global properties like symmetry can help to resolve ambiguities in non-rigid motion (53).

**Experiment 3 – Detection of multiple non-rigidities:** In the natural world, objects often undergo multiple deformations simultaneously, e.g. a stalking animal inhales and exhales. To see if observers can detect multiple simultaneous non-rigidities, we examined whether observers could detect inflation and deflation of rigid and flexing cylinder by smoothly varying the aspect ratio of the cylinder during the trial (the cross-section in the image plane was kept constant). Any given trial began with a cylinder with aspect ratio of 1.0 (perfectly circular cross-section) and then either inflated by a fixed amount in the first half of the trial followed by a deflation by the same amount to a perfectly circular cylinder or deflated by a fixed amount followed by an inflation by the same amount. The order of inflation and deflation was randomized across trials. The observers’ task was to report whether the trial consisted of an inflation followed by a deflation or vice versa. This experiment was conducted in two parts; in the first part we used a constant stimulus paradigm using relatively high signal values to assess the performance of observers on the task; only those observers who could perform the task in a consistent manner participated in the second part. In the second part, we used a QUEST procedure (54) to determine 82% correct thresholds for the rigid and flexing cylinders. The degree of inflation/deflation was controlled as the percentage change from unit aspect ratio.

![Figure 6: Results from Experiment 3. (A) Percentage correct responses for distinguishing different amounts of inflation and deflation (different colors for nine observers). Observers who did not reach threshold for 90% deformation, were not tested on less extreme deformations. Rigid, depth-flex and plane-flex cylinders are represented using squares, circles and diamonds respectively. (B) Detection thresholds measured for inflation and deflation on the three types of cylinder averaged across four observers capable of doing the task.](image)

Figure 6A shows the percentage accuracy averaged across inflation and deflation for 9 naïve observers as a function of signal strength. Only two out of 9 observers could perform the task consistently implying that the task is inherently difficult and may require extensive training. The difficulty may be related to problems in detecting instantaneous stretching of objects along the line of sight from motion cues alone (55). We selected these two observers, plus author AJ and another observer who was extensively trained on the task but was uninformed about the purposes of the experiment, and measured accuracy thresholds for inflation and deflation independently. Figure 6B shows the 82% accuracy thresholds averaged across the 4 observers. Overall, thresholds for deflation were higher than thresholds for inflation (F(1,18) = 5.33, p<0.05). More importantly, there was no systematic difference in thresholds between rigid and flexing cylinders (F(2,18) = 0, p=0.99), nor a significant interaction (F(2,18) = 0.18, p=0.83).

Figure 7 shows the performance of the two computational models for the rigid cylinder undergoing inflation and deflation. The motion perspective model follows the variation in aspect ratio quite well (A and B) but the trajectory basis model does not (C and D). We observed no improvement in the performance of the trajectory basis model by using a larger number of basis trajectories. The results show that detecting changes in relative velocities over time would be a successful strategy for this task. Performance in the task may be limited by the inability to detect...
accelerations (56, 57). However, after training, four observers could perform the task reliably, which suggests that observers can learn to associate changes in relative velocities with changes in shape.

![Figure 7](image)

**Figure 7**: Simulated performance of the two computational models while following inflation/deflation of the cylinders. The left column (A and C) shows model predictions for a rigid cylinder undergoing 90% inflation while the right column (B and D) shows model predictions for a rigid cylinder undergoing 90% deflation. The top row shows predictions based on the motion perspective model while the bottom row shows the predictions based on the trajectory basis model. Models made similar predictions for flexing cylinders.

**Discussion**

Motions in the retinal image are predominantly due to movements of the observer’s body and head, and in this case, most objects have to be inferred to be rigid and stationary despite distortions in the retinal images. There is some evidence for a rigidity prior that biases scene percepts during self-motion (58), and it is possible that this prior is generalized to object generated retinal motions. The most significant findings of the experiments in this study, are that humans are equally sensitive at discriminating non-rigid and rigid shapes from motion cues, show similar biases in perceiving departures from non-rigid and rigid symmetries, and are equally sensitive to expansion and contraction of non-rigid and rigid shapes. These results suggest that the human perceptual system uses similar mechanisms for extracting rigid and non-rigid shapes from motion cues, and it is unlikely that a rigidity assumption per se is paramount in this process. As general alternatives to the rigidity assumption, a number of regularization principles have been proposed, such as minimal mapping (9, 59), smoothest motion (59), or motion coherence (60), but a model that generates general non-rigid percepts on the basis of minimization principles remains to be constructed (61).

The computational simulations in this study found that the motion perspective model predicted human performance better than the trajectory space model, suggesting that the human perceptual system uses relative motions to compute 3D structure. In addition, simulated symmetric cylinders were perceived as asymmetric whenever the velocity profile was asymmetric, providing further evidence that the relative velocity profile determined the perceived 3D shape. The motion-perspective model is essentially based on the computation of motion shears at different spatial scales. We have presented it as a simple explanation of our empirical results, while recognizing that it will need to be elaborated for general situations. In a simple case like our rotating rigid cylinders, local motions are predominantly parallel, as are the shears. In general, variations in speed as a function of depth and distance from the rotation axis, lead to 2-D deformations in the perspective image (31), that can be captured by 2-D combinations of local
and then rotated back to the initial position. The stimuli lasted for 1
in that order. The cylinder rotated until it reached an orientation with
image
was rotated
perspective
removing texture and density cues to 3D shape
horizontal
Stimuli and Task
Stimuli were generated using the CRS Toolbox for MATLAB (The Mathworks™, USA) and were displayed on a Sony CRT monitor using Cambridge Research Systems’s ViSaGe system (Cambridge Research Systems, UK) controlled by a Dell GX620. The monitor’s resolution was set to 1024x768 pixels and the refresh rate was set to 120 Hz. The experiments were conducted in compliance with the standards set by the IRB at SUNY College of Optometry. Subjects gave their informed consent prior to their inclusion in the study and were paid for their participation.

Methods
Observers: Eight observers participated in Experiment 1, six in Experiment 2, and eleven observers participated in Experiment 3. Both authors participated in Experiments 1 and 2; author AJ also participated in Experiment 3. Observers other than the authors were informed about the purpose of the experiment only after it was completed. All subjects had normal or corrected-to-normal vision. The experiments were conducted in compliance with the standards set by the IRB at SUNY College of Optometry. Subjects gave their informed consent prior to their inclusion in the study and were paid for their participation.

Apparatus: Stimuli were generated using the CRS Toolbox for MATLAB (The Mathworks™, USA) and were displayed on a Sony CRT monitor using Cambridge Research Systems’s ViSaGe system (Cambridge Research Systems, UK) controlled by a Dell GX620. The monitor’s resolution was set to 1024x768 pixels and the refresh rate was set to 120 Hz. The experiments were conducted in a dark room. The observers viewed the stimuli monocularly from a distance of 100 cm, using a chin-rest to stabilize head position.

Stimuli and Tasks: Stimuli consisted of white dots randomly placed on an opaque black horizontal cylinder on a black background after rendering the 3-D surface (Figure 1), thus removing texture and density cues to 3D shape. The cylinders were presented in proper perspective for the position of the observer’s eyes. The 7.36 deg long and 2.12 deg wide cylinder was rotated simultaneously at $\pi$ rad/s about the vertical and at $2\pi / 3$ rad/s about the depth axis (the cylinder did not spin on its axis and 3D shape was not perceived with rotation only in the image plane). The starting orientation was set at $-\pi / 6$ roll and $-\pi / 4$ yaw for all trials, applied in that order. The cylinder rotated until it reached an orientation with $+\pi / 6$ roll and $+\pi / 4$ yaw, and then rotated back to the initial position. The stimuli lasted for 1.0 s. The cylinder was either
rigid or flexed smoothly in the depth or the image plane, using Equations 1 and 2 respectively (Videos S1 – S3):

\[ z(k) = z_0 - (x_0/20)^2 * \sin(4\pi k/N) \]  
\[ y(k) = y_0 - (x_0/30)^2 * \sin(4\pi k/N) \]

where,
\[ k = \text{current frame} \]
\[ N = \text{total frames} \]
\[ x_0, y_0, z_0 = \text{initial position} \]

Rotation was applied after computing the flexing. We also included a condition where the rigid cylinder rotated at a higher speed (1.5 times) so that the mean speed of the dots approximately matched the mean speed for the non-rigid cylinders. The rotating cylinder was only visible through a window 6.36 deg wide so that observers could not see the ends of the cylinder. The central 0.8 deg section of the cylinder was occluded to prevent observers from using just the approximately rigid central section.

Acknowledgements
This work was supported by NEI grants EY07556 & EY13312 to QZ.

References
13. Helmholtz Hv (1910/1925) Treatise on Physiological Optics (Vol. III) (Optical Society of America, New York); trans 1910 JPCSoG.