Perceptual transitions between object rigidity and non-rigidity: Competition and cooperation among motion energy, feature tracking, and shape-based priors

Akihito Maruya
Graduate Center for Vision Research,
State University of New York, New York, NY, USA

Qasim Zaidi
Graduate Center for Vision Research,
State University of New York, New York, NY, USA

Why do moving objects appear rigid when projected retinal images are deformed non-rigidly? We used rotating rigid objects that can appear rigid or non-rigid to test whether shape features contribute to rigidity perception. When two circular rings were rigidly linked at an angle and jointly rotated at moderate speeds, observers reported that the rings wobbled and were not linked rigidly, but rigid rotation was reported at slow speeds. When gaps, paint, or vertices were added, the rings appeared rigidly rotating even at moderate speeds. At high speeds, all configurations appeared non-rigid. Salient features thus contribute to rigidity at slow and moderate speeds but not at high speeds. Simulated responses of arrays of motion-energy cells showed that motion flow vectors are predominantly orthogonal to the contours of the rings, not parallel to the rotation direction. A convolutional neural network trained to distinguish flow patterns for wobbling versus rotation gave a high probability of wobbling for the motion-energy flows. However, the convolutional neural network gave high probabilities of rotation for motion flows generated by tracking features with arrays of MT pattern-motion cells and corner detectors. In addition, circular rings can appear to spin and roll despite the absence of any sensory evidence, and this illusion is prevented by vertices, gaps, and painted segments, showing the effects of rotational symmetry and shape. Combining convolutional neural network outputs that give greater weight to motion energy at fast speeds and to feature tracking at slow speeds, with the shape-based priors for wobbling and rolling, explained rigid and non-rigid percepts across shapes and speeds ($R^2 = 0.95$). The results demonstrate how cooperation and competition between different neuronal classes lead to specific states of visual perception and to transitions between the states.

Introduction

Visual neuroscience has been quite successful at identifying specialized neurons as the functional units of vision. Neuronal properties are important building blocks, but there is a big gap between understanding which stimuli drive a neuron and how cooperation and competition among different types of neurons generate visual perception. We try to bridge the gap for the perception of object rigidity and non-rigidity. Both of those states can be stable, so we need to develop ways of understanding how the visual system changes from one state to another. Shifts between different steady states are the general case for biological vision but have barely been investigated.

In the video of Figure 1A, most observers see the top ring as rolling or wobbling over the bottom ring, seemingly defying physical plausibility. In the video of Figure 1B, the two rings seem to be one rigid object rotating in a physically plausible way. Because both videos are of the same physical object rotated at different speeds on a turntable, clearly an explanation is needed.

Humans can sometimes perceive the true shape of a moving object from impoverished information. For example, shadows show the perspective projection of just the silhouette of an object; yet, two-dimensional (2D) shadows can convey the three-dimensional (3D) shape for some simple rotating objects (Wallach & O'Connell, 1953; Albertazzi, 2004). However, for irregular or unfamiliar objects, when the light source is oblique to the surface on which the shadow is cast, shadows often get elongated and distorted in such a way that the casting object is not recognizable. Similarly, from the shadow of an object passing rapidly, it is often difficult to discern whether the shadow

Citation: Maruya, A., & Zaidi, Q. (2024). Perceptual transitions between object rigidity and non-rigidity: Competition and cooperation among motion energy, feature tracking, and shape-based priors. *Journal of Vision*, *24*(2):3, 1–25, https://doi.org/10.1167/jov.24.2.3.

https://doi.org/10.1167/jov.24.2.3 Received August 1, 2023; published February 2, 2024 ISSN 1534-7362 Copyright 2024 The Authors

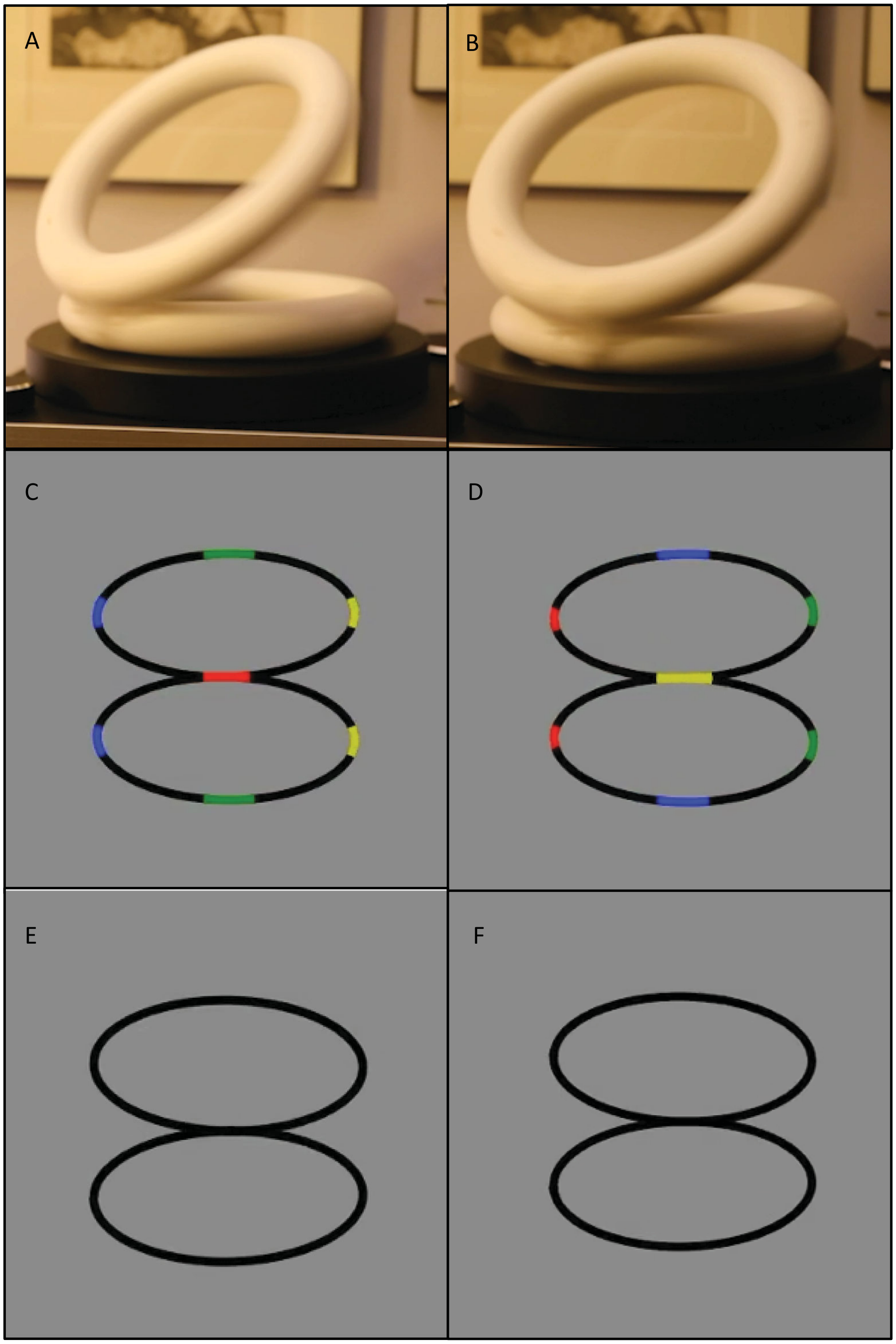

Figure 1. Rotating ring illusion. (**A**) A Styrofoam ring forming an angle with another ring is seen as wobbling or rolling over the bottom ring despite physical implausibility. (**B**) At a slower speed, the rings are seen to rotate together, revealing that they are glued to each other and that the non-rigid rolling was an illusion. Panel (**B**) differs from (**A**) only in turntable rpm. (**C**) Two rings rotate together with a fixed connection at the red segment. (**D**) Two rings wobble against each other, as shown by the connection shifting to other colors. Panels (**E**) and (**F**) are the same as (**C**) and (**D**) except that the colored segments are black like the rest of the rings. (**E**) and (**F**) generate identical sequences of retinal images so that wobbling and rigid rotation are indistinguishable. Movie is available on the journal website.

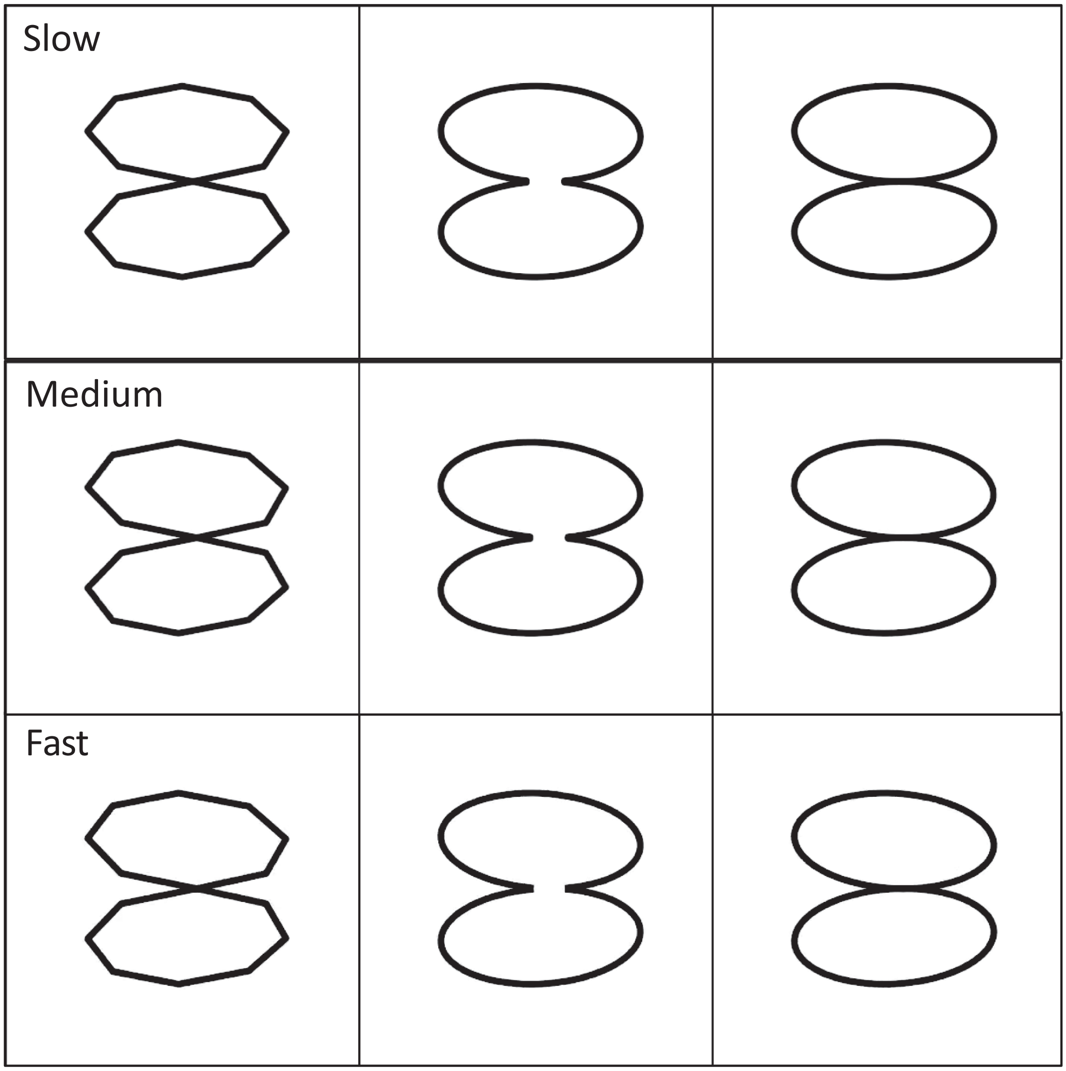

Figure 2. Effect of shape on the rotating-ring illusion. Pairs of rings with different shapes rotate around a vertical axis. When the speed is slow (1°/s), all three shapes are seen as rotating. At the medium speed (10°/s), the circular rings seem to be wobbling against each other, but the other two shapes seem rigid. At the fast speed (30°/s), non-rigid percepts dominate irrespective of shape features. Movie is available on the journal website.

is distorting or the object. Images on the retina are also formed by perspective projection, and they too distort if the observer or the object is in motion, yet observers often correctly see the imaged object as rigid or non-rigid. Examples of rigidity often contain salient features (Shiffrar & Pavel, 1991; Lorenceau & Shiffrar, 1992), whereas rigid shapes without salient features are sometimes seen as non-rigid (Mach, 1886; Weiss & Adelson, 2000; Rokers, Yuille, & Liu, 2006; Vezzani, Kramer, & Bressan, 2014).

We examine how and why salient features help in veridical perception of rigidity when objects are in motion. For that purpose, we use variations of the rigid object in Figures 1A and 1B that can appear rigid or non-rigid depending on the speed and salience of the features. This object is simple by the standards of natural objects but complex compared to stimuli generally used in studies of motion mechanisms. The interaction of motion with shape features has received extensive attention (McDermott, Weiss, & Adelson, 2001; Papathomas, 2002; McDermott & Adelson, 2004; Berzhanskaya, Grossberg, & Mingolla, 2007; Erlikhman, Caplovitz, Gurariy, Medina, & Snow, 2018; Freud, Behrmann, & Snow, 2020), but not the effect of the interaction on object rigidity. By exploring the competition and cooperation among motion-energy mechanisms, feature-tracking mechanisms, and shape-based priors, we present a mechanistic approach to perception of object rigidity.

To introduce controlled variation in the ring pair, we switched from physical objects to computer graphics. Consider the videos of the two ring pairs in Figures 1C and 1D. When attention is directed to the connecting joint, the configuration in Figure 1C is seen as rotating rigidly while the rings in Figure 1D wobble and slide non-rigidly against each other. Note that "rotation" in this paper always refers to "rigid rotation" unless explicitly described as non-rigid. Rigidity and non-rigidity are both easy to see because the painted segments at the junction of the two rings either stay together or slide past each other. The two videos in Figures 1E and 1F are of the same object as the two above, except that the painted segments have been turned to black. Now it is difficult to see any difference in the motion of the two rings because they form identical images, so whether they are seen as rigidly rotating or as non-rigidly wobbling depends on how the images are processed by the visual system. The wobbling illusion of the rigidly connected rings has been used for many purposes. The senior author first saw it over 30 years ago at a tire shop in Texas, where it looked like a stationary horizontal tire was mounted on a high pole and a tire was rolling over it at an acute angle seemingly defying physical laws. The first author pointed out that the wobbling rings illusion was used in a Superman movie to confine criminals during a trial. There are many videos on the internet showing how to make physical versions of the illusion using Styrofoam or wooden rings, but, despite the popularity of the illusion, we could not find any explanation of why people see non-rigidity in this rigid object.

By varying the speed of rotation, we discovered that at slow speeds both the painted and non-painted rings appear rigidly connected, and at high speeds both appear non-rigid. This is reminiscent of the differences in velocity requirements of motion-energy and feature-tracking mechanisms (Lu & Sperling, 1995; Zaidi & DeBonet, 2000), where motion-energy mechanisms function above a threshold velocity and feature tracking only functions at slow speeds. Consequently, we processed the stimuli through arrays of motion-energy and feature-tracking units and then trained a convolutional neural network (CNN) to classify their outputs, demonstrating that the velocity-based relative importance of motion-energy versus feature tracking could explain the change in percepts with speed. To critically test this hypothesis, we manipulated the shape of the rings to create salient features that could be tracked more easily, and we used physical measures such as rotational symmetry to estimate prior expectations for wobbling and rolling of different shapes. These manipulations are demonstrated in Figure 2, where at medium speeds the rings with vertices and gaps appear rigidly rotating but the circular rings appear to be wobbling, whereas all three appear rigid at slow speeds and non-rigidly connected at fast speeds. (We established that the multiple images seen at fast speeds were generated by the visual system and were not in the display, as photographs taken with a Canon T7 with a 1/6400-second shutter speed showed only one pair of rings.) Taken together, our investigations revealed previously unrecognized roles for feature tracking and priors in maintaining veridical percepts of object rigidity.

Perceived non-rigidity of a rigidly connected object

Shapes from motion models generally invoke rigidity priors (Ullman 1979; Andersen & Bradley, 1998) with a few exceptions (Bregler, Hertzmann, & Biermann, 2000; Fernández, Watson, & Qian, 2002; Akhter, Sheikh, Khan, & Kanade, 2010). Besides the large class of rigid objects, there is also a large class of articulated objects in the world, including most animals whose limbs and trunks change shape to perform actions. If priors reflect statistics of the real world, it is quite likely that there is also a prior for objects consisting of connected parts to appear non-rigidly connected while parts appear rigid or at most elastic (Jain & Zaidi, 2011). This prior could support percepts of non-rigidity even when connected objects move rigidly—for example, the rigid rotation of the ring pair. To quantify perception of the non-rigid illusion, we measured the proportion of times different shapes of ring pairs look rigidly or non-rigidly connected at different rotation speeds.

Methods

Using Python, we created two circular rings with a rigid connection at an angle. The rigid object rotates around a vertical axis oblique to both rings. The videos in Figure 3 show the stimuli that were used in this study to test the role of features in object rigidity. There were nine different shapes. The original circular ring pair is "Circ ring." A gap in the junction is "Circ w gap." The junction painted red is "Circ w paint." Two octagons were rigidly attached together at an edge ("Oct on edge") or at a vertex ("Oct on vertex"). Two squares were attached in the same manner ("Sqr on vertex") and ("Sqr on edge"). The junction between two ellipses (ratio of the longest to shortest axis was 4:3) was parallel to either the long or the short axis leading to "Wide ellipse" and "Long ellipse," respectively. A tenth configuration where the circular rings physically wobbled ("Circ wobble") was the only non-rigid configuration.

The rotating ring pairs were rendered as if captured by a camera at a distance of 1.0 m and at either the same height as the junction (0° elevation) or at 15° elevation (see equations in the Stimulus generation and projection section of the Appendix). We varied the rotation speed, but linear speed is more relevant for motion-selective cells, so the speed of the joint when

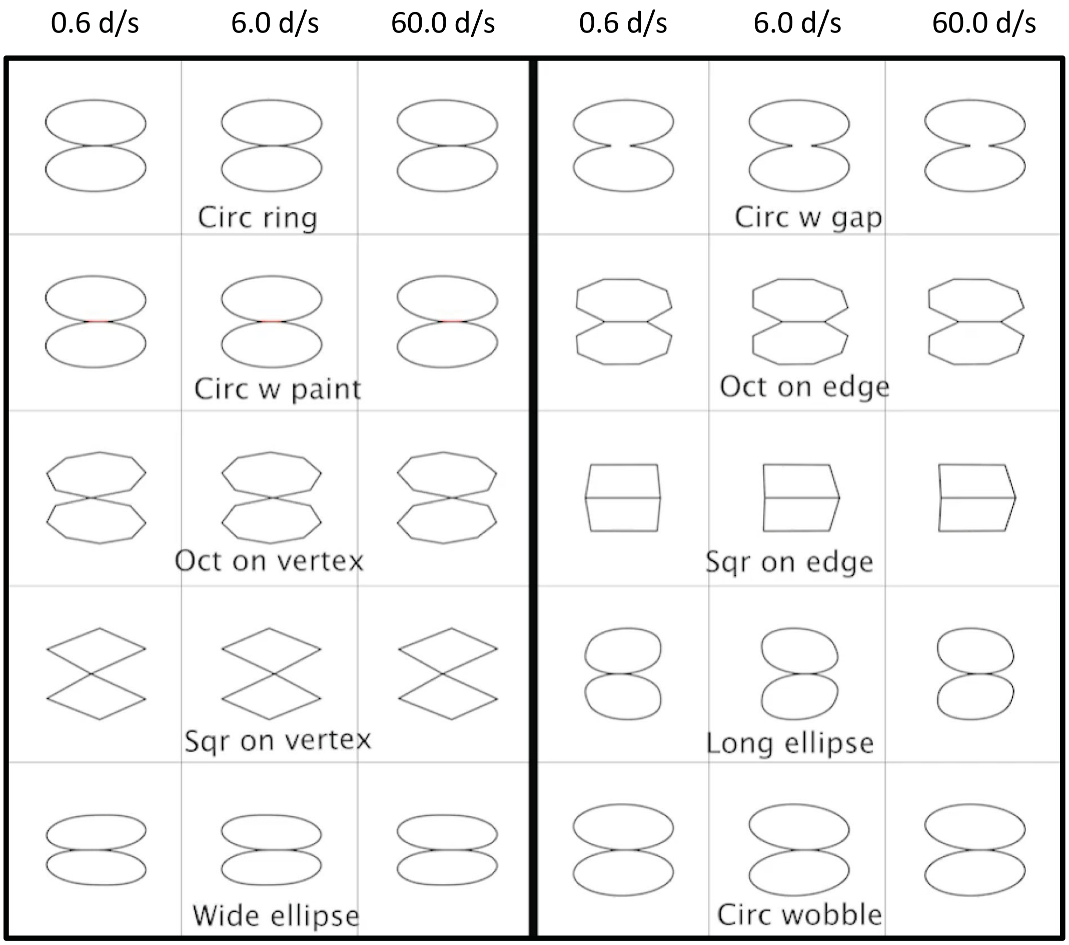

Figure 3. Shapes showing the effects of features on object rigidity. Rows provide the names of the shapes of the stimuli. Columns show the three speeds of rotation. Movie is available on the journal website.

passing fronto-parallel was 0.6°/s, 6°/s, or 60°/s. The diameters of the rings were 3 degrees of visual angle (dva) or 6 dva, so the corresponding angular speeds of rotation were 0.03, 0.3, and 3 cycles per second (cps) for the 3-dva rings and 0.015, 0.15, and 1.5 cps for the 6-dva rings. 3D cues other than motion and contour were eliminated by making the width of the line uniform and removing shading cues. In the videos of the stimuli (Figure 3), each row indicates a different shape of the stimulus, and the column represents the speed of the stimulus from 0.6°/s (left) to 60°/s (right).

The videos were displayed on a VIEWPixx/3D (VPixx Technologies, Saint-Bruno-de-Montarville, QC, Canada) at 120 Hz. MATLAB (MathWorks, Natick, MA) and Psychtoolbox were used to display the stimulus and run the experiment. The data were analyzed using Python and MATLAB. The initial rotational phase defined by the junction location was randomized for each trial, as was the rotational direction (clockwise or counterclockwise looking down at the rings). An observer's viewing position was fixed by using a chin rest so that the video was viewed at the same elevation as the camera position. The observer was asked to look at the junction between the rings and to report by pressing buttons whether or not the rings were rigidly connected. The set of 120 conditions (10 shapes × 2 sizes × 3 speeds × 2 rotation directions) was repeated 20 times (10 times at each viewing elevation). Measurements were made by 10 observers with normal or corrected-to-normal vision. Observers gave written informed consent. All experiments were conducted in compliance with a protocol approved by the institutional review board at SUNY College of Optometry, in compliance with the tenets of the Declaration of Helsinki.

Results

Figures 4A and 4B show the average proportion of non-rigid percepts for each observer for each shape at the three speeds (0.6°/s, 6.0°/s, and 60.0°/s) for the 3-dva and 6-dva diameter sizes. Different colors indicate different observers, and the symbols are displaced

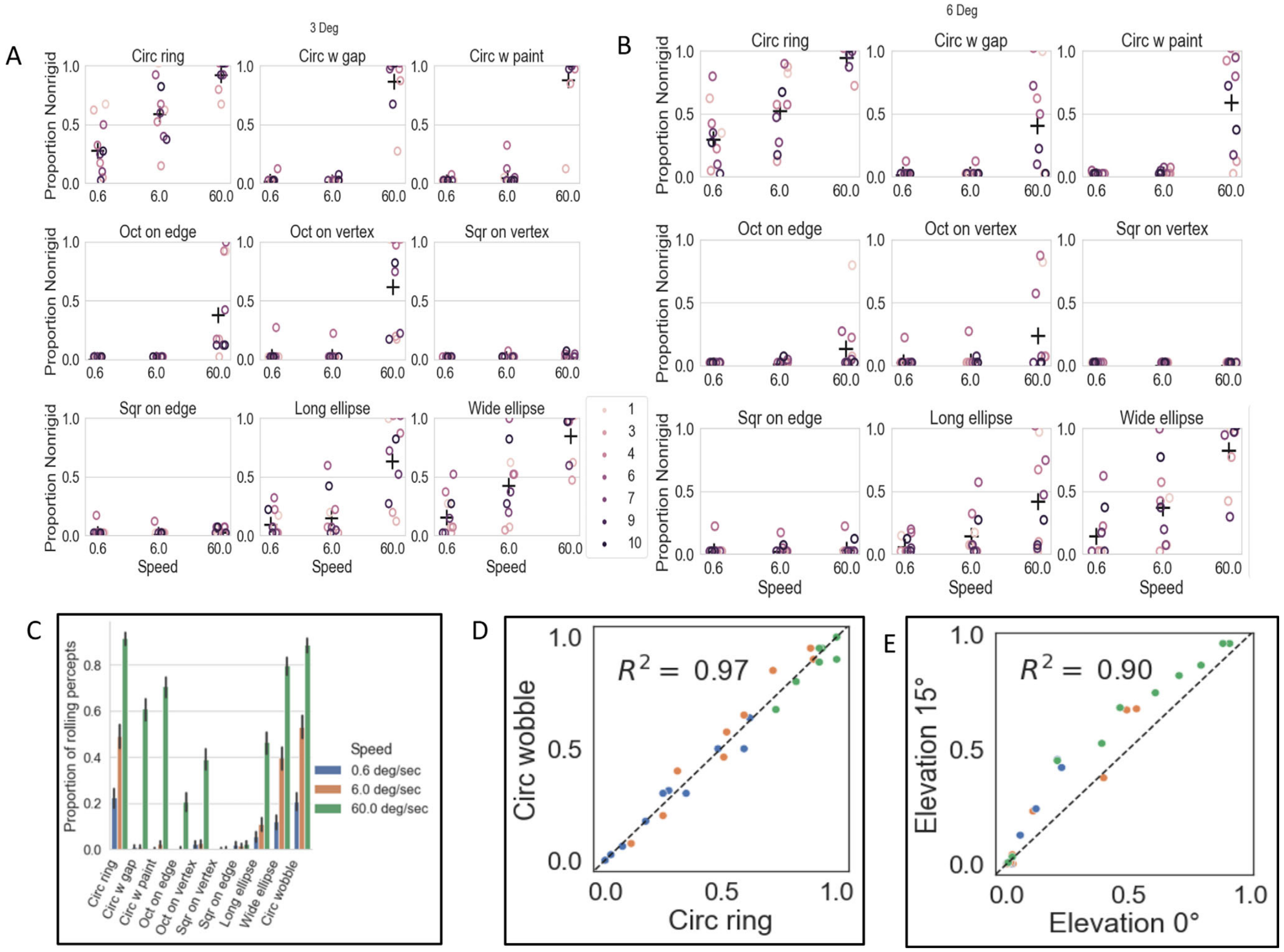

Figure 4. Non-rigid percepts. Average proportions of reports of non-rigidity for each of 10 observers for each shape at three speeds (0.6°/s, 6.0°/s, and 60.0°/s) for diameters of 3 dva (**A**) and 6 dva (**B**). Different colored circles indicate different observers, and the average of all of the observers is shown by the black cross. (**C**) Histograms of non-rigid percepts averaged over the 10 observers. The error bar indicates 95% confidence intervals after 1000 bootstrap resamples. (**D**) Average proportion of non-rigid percepts for the rotating and wobbling circular rings for 10 observers and three speeds. Similarity is shown by closeness to the unit diagonal and $R^2 = 0.97$. (**E**) Average proportion of non-rigid percepts for 0° elevation versus 15° elevation. Proportions are similar ($R^2 = 0.90$) but slightly higher for the 15° elevation.

horizontally to avoid some of them becoming hidden; the dark cross represents the mean. The results for the two sizes are similar, except that there is a slightly greater tendency to see rigidity for the larger size. The combined results for the two sizes, averaged over the 10 observers, are shown as histograms in Figure 4C. For the circular rings, there is a clear progression toward non-rigid percepts as the speed increases: At 0.6°/s, a rigid rotation is perceived on average around 25% of the time. As the speed of rotation is increased, the average proportions of non-rigid percepts increase to around 60% at 6.0°/s and around 90% at 60.0°/s. These results provide empirical corroboration for the illusory non-rigidity of the rigidly rotating rings. Introducing a gap or painted segment in the circular rings increases the percept of rigidity, especially at the medium speed. Turning the circular shapes into octagons with vertices further increases rigidity percepts, and making the rings squares almost completely abolishes non-rigid percepts. If the circular rings are stretched into long ellipses that too reduces non-rigid percepts, but, if they are stretched into wide ellipses, it has little effect, possibly because perspective shortening makes the projections of the ellipses close to circular. The results are averaged for the 10 observers in the histograms in Figure 4C, which will be used as a comparison with the model simulations. For all configurations other than the squares, non-rigid percepts increase as a function of increasing speed. The effect of salient features is thus greater at the slower speeds. The effect of speed provides clues for modeling the illusion based on established mechanisms for motion energy and feature tracking. Figure 4D shows the similarity between the results for the rotating and wobbling circular rings as the dots are close to the unit diagonal and $R^2 = 0.97$, which is not surprising given that their images are identical. Figure 4E shows that there is a slight tendency to see more non-rigidity at the 15° viewing elevation than the 0° elevation, but the $R^2 = 0.90$ meant that we could combine the data from the two viewpoints in the figures for fitting with a model.

Motion-energy computations

The first question that arises is why observers see non-rigid motion when the physical motion is rigid. To gain some insight into the answer, we simulated responses of direction-selective motion cells in primary visual cortex (Hubel, 1959; Hubel & Wiesel, 1959; Hubel & Wiesel, 1962; Movshon, Thompson, & Tolhurst,

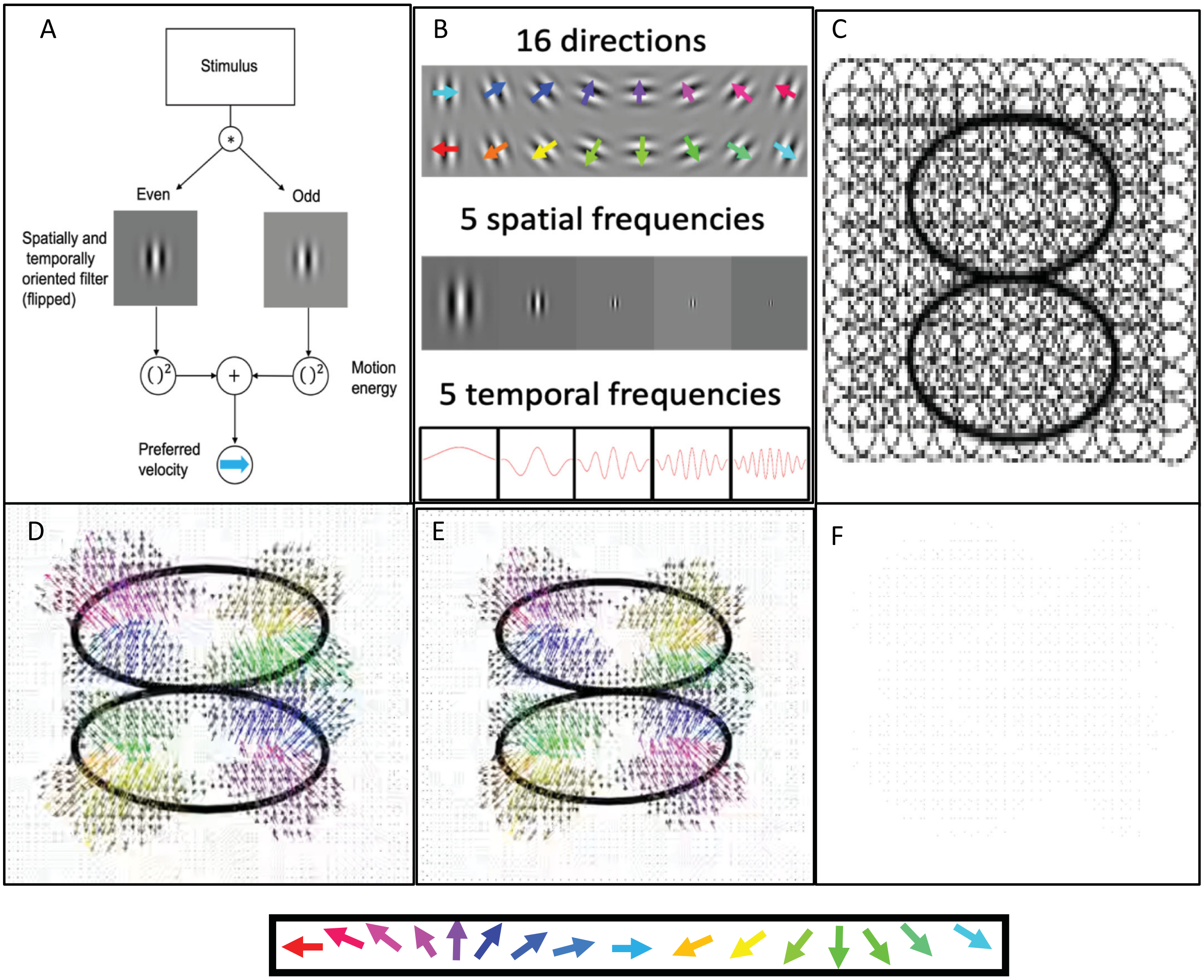

Figure 5. Motion-energy mechanism. (**A**) Schematic diagram of a motion-energy unit: The moving stimulus is convolved with two filters that are odd and even symmetric spatially and temporally oriented filters; the outputs are squared and summed to create a phase-independent motion-energy response. (**B**) Motion-energy units used in the model. At each spatial location there were 16 preferred directions, five spatial frequencies, and five temporal frequencies. (**C**) An array of 51,984,000 motion-energy units uniformly covering the whole stimulus were applied to the change at each pair of 199 frames. At each location, the preferred velocity of the highest responding motion-energy unit from the array was selected as the population response. Motion vectors from physically rotating (**D**) and wobbling (**E**) ring pairs are predominantly orthogonal to contours instead of in the rotation direction. (**F**) The difference between the two vector fields is negligible. Because the flows for physically rotating and wobbling circular rings are almost identical, other factors must govern the perceptual shift from wobbling to rotation at slower speeds. Movie is available on the journal website.

1978a; Movshon, Thompson, & Tolhurst, 1978b) with a spatiotemporal energy model (Watson & Ahumada,1983; Adelson & Bergen, 1985; Watson & Ahumada, 1985; Van Santen & Sperling,1985; Rust, Mante, Simoncelli, & Movshon, 2006; Bradley & Goyal, 2008; Nishimoto & Gallant, 2011; Nishimoto et al., 2011). A schematic diagram of a motion-energy filter is shown in Figure 5A, and the equations are presented in the Appendix (Motion energy section). At the linear filtering stage, two spatially and temporally oriented filters in quadrature phase were convolved with a sequence of images. Pairs of quadrature filters were squared and added to give phase-invariant motion energy. Responses of V1 direction-selective cells are transmitted to extrastriate area MT, where cells include component and pattern cells (Movshon, Adelson, Gizzi, & Newsome, 1985; Movshon & Newsome, 1996; Rust et al., 2006). Component cells have larger receptive fields than V1 direction-selective cells, but their motion responses are similar.

Each motion-energy unit was composed of direction-selective cells with five temporal frequencies (0, 2, 4, 6, and 8 cycles/video) × five spatial frequencies (4, 8, 16, 20, and 30 cycles/image) × 8 orientations (from 0° to 180° every 22.5°) × 2 directions except for the static 0-Hz filters leading to 360 cells at each location (Figure 5B). Because the stimulus that we used in the analysis was 3.0 dva × 3.0 dva × 3.0 s, the spatial frequencies equate to 1.3, 2.7, 5.3, 6.7, and 10.0 cycles/deg, and the temporal frequencies to 0, 0.67, 1.3, 2.0, and 2.7 cps. An array of 360 cells × (380 × 380) pixels = 51,984,000 motion-energy units uniformly covering the whole stimulus were applied to the change at each pair of 199 frames (Figure 5C). Because at every location many direction-selective cells respond at every instant, the responses have to be collapsed into a representation to visualize the velocity response dynamically. We use a color-coded vector whose direction and length at a location and instant respectively depict the preferred velocity of the unit that has the maximum response, akin to a winner-take-all rule (note that the length of the vector is not the magnitude of the response but the preferred velocity of the most responsive unit).

We begin by analyzing the circular ring pair, because that shows the most change from rigid to non-rigid. Figure 5D indicates the motion-energy field when the circular ring pair is rigidly rotating, and Figure 5E shows when the two rings are physically wobbling. There are 200 × 200 × 199 (height × width × time frame) = 7,960,000 2D vectors in each video. In both cases, for most locations and times, the preferred velocities are perpendicular to the contours of the rings. The response velocities in the rotating and wobbling rings look identical. To confirm this, Figure 5F subtracts Figure 5E from Figure 5D and shows that the difference is negligible. This vector field could thus contribute to the perception of wobbling or rotating or even be bistable. Because the vector directions are mostly not in the rotation direction, it would seem to support a percept of wobbling, but, to provide a more objective answer, we trained a CNN to discern between rotation and wobbling and fed it the motion-energy vector field to perform a classification.

Motion-pattern recognition CNN

For training the CNN, we generated random moving dot stimuli from a 3D space, and these dots either rotated around a vertical axis or wobbled at a random speed (0.1°/s–9°/s). The magnitude of wobbling was selected from −50° to 50°, and the top and the bottom parts wobbled against each other with a similar magnitude as the physically wobbling ring stimulus that was presented to the observers. These 3D motions were projected to the 2D screen with camera elevations ranging from −45° to 45°. At each successive frame, the optimal velocity at each point was computed and the direction perturbed by Gaussian noise with sigma = 1° to simulate noisy sensory evidence. Two examples of the 9000 vector fields (4500 rotational and 4500 wobbling) are shown in Figure 6A. As the examples show, it is easy for humans to discern the type of motion. The 9000 vector fields were randomly divided into 6300 training fields and 2700 validation fields. Random shapes were used to generate the training and validation random-dot motion fields, with the wobbling motion generated by dividing the shape into two parallel to the *XZ* plane at the midpoint of the *Y*-axis.

The neural network was created and trained with TensorFlow (Abadi et al., 2016). For each pair of frames, the motion field is fed to the CNN as vectors (Figure 6B). The first layer of the CNN contains 32 filters, each of which contains two channels for the horizontal and the vertical components of vectors. These filters are cross-correlated with the horizontal and vertical components of the random-dot flow fields. Each filter output is half-wave rectified and then max pooled. The second layer contains 10 filters. Their output (half-wave rectified) is flattened into arrays and followed by two fully connected layers that act like Perceptrons (Gallant, 1990). For each pair of frames, the last layer of the CNN provides a relative confidence level for wobbling versus rotation calculated by the softmax activation function (Sharma, Sharma, & Athaiya, 2017). Based on the higher confidence level for the set of frames in a trial, the network classifies the motion as rotation or wobbling (see Convolutional neural network section in the Appendix). After training for 5 epochs, the CNN reached 99.78% accuracy for the training dataset and 99.85% for the validation test

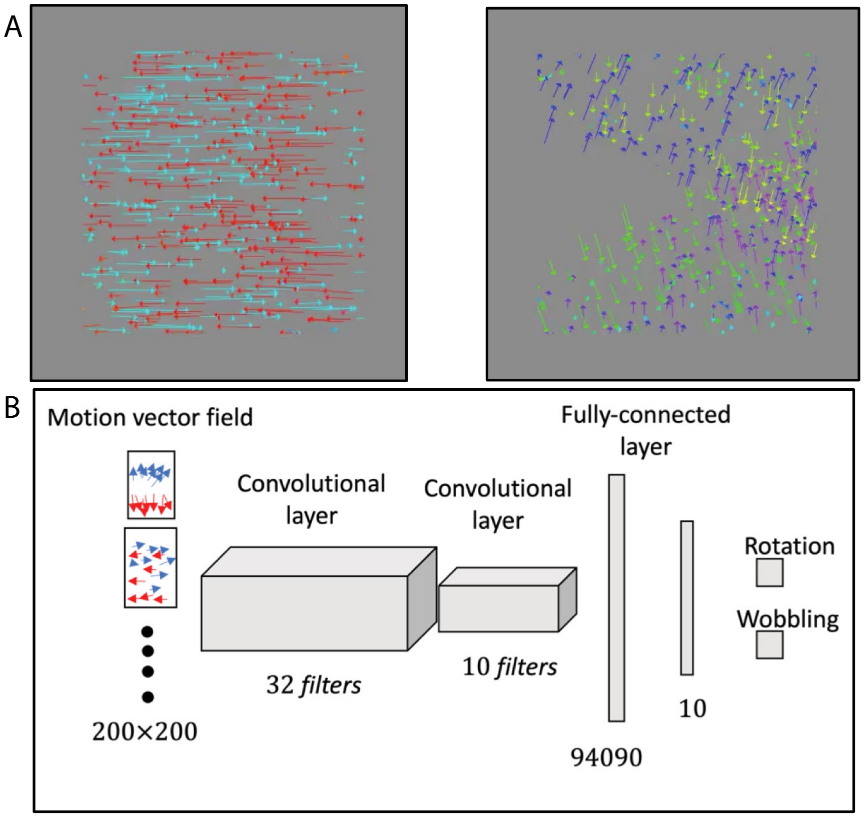

Figure 6. CNN for classifying patterns of motion vectors as rotating or wobbling. (**A**) Two examples of the 9000 vector fields from random dot moving stimuli that were used to train and validate the CNN. (Left) A rotating vector field. (Right) A wobbling vector field. The 9000 vector fields were randomly divided into 6300 training and 2700 validation fields. (**B**) The network consists of two convolutional layers followed by two fully connected layers. The output layer gives a confidence level between rotation and wobbling on a scale of 0.0 to 1.0. Movie is available on the journal website.

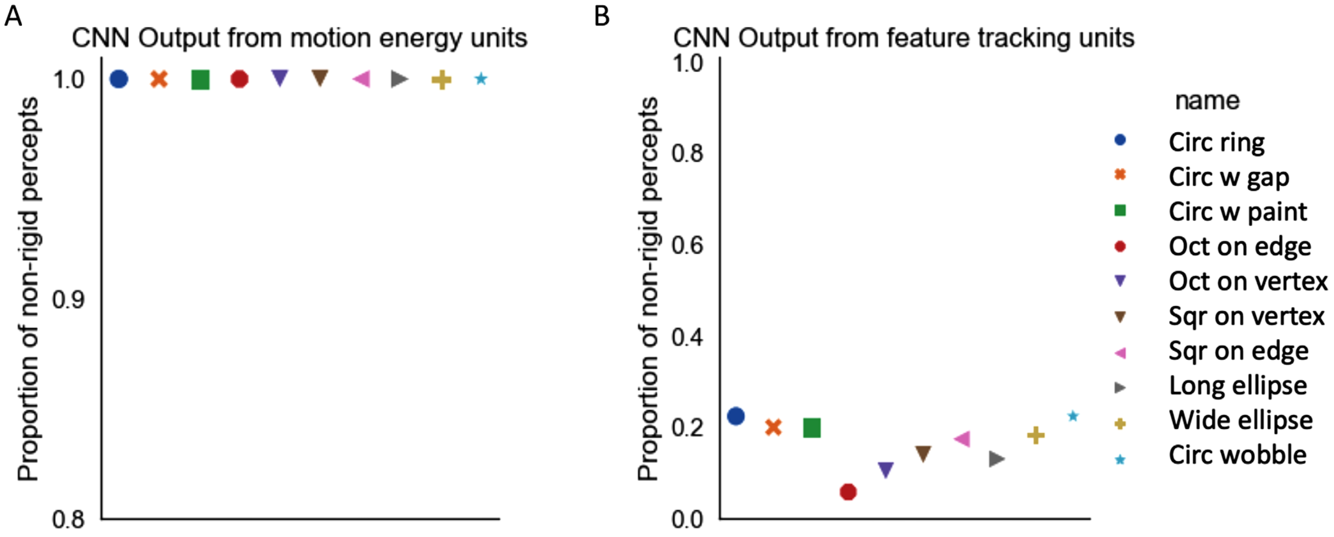

Figure 7. CNN output. (**A**) Proportion of non-rigid percepts for CNN output from motion-energy units for each shape. Symbol shape and color indicate ring-pair shape. For all ring shapes, the proportion of non-rigid classifications was 0.996. (**B**) The average CNN output based on the feature-tracking vector fields being the inputs for different stimulus shapes shows the higher probabilities of rigid percepts.

dataset. We use the CNN purely as a pattern recognizer without any claims to biological validity, but its output does resemble position-independent neural responses to the pattern of the velocity field in the medial superior temporal (MST) or superior temporal sulcus (STS) cortical areas (Sakata, Shibutani, Ito, & Tsurugai, 1986; Tanaka et al., 1986; Duffy & Wurtz, 1991; Zhang, Sereno, & Sereno, 1993; Pitzalis et al., 2010).

We calculated motion-energy vector fields for rotations of all nine shapes and classified them with the CNN. At each time frame, the network reports a confidence level from 0 to 1 between wobbling and rotation. The proportion of non-rigid percepts of the CNN is derived by the average of the confidence level for wobbling across time frames. Classification of the motion-energy vectors led to 99.6% percepts of wobbling for all the shapes at all the speeds (Figure 7A). This raises the question as to what makes observers perceive rigid rotation for slow speeds and different proportions of wobble for different shapes at medium speeds. To understand these departures from the motion-energy–based predictions, we examined feature-tracking mechanisms, motion illusions that are unsupported by sensory signals, and prior assumptions based on shape and geometry.

Feature-tracking computations

Motion-energy signals support wobbling for a rigidly rotating ring pair because the preferred direction of local motion detectors with limited receptive field sizes is normal to the contour instead of along the direction of object motion, known as the aperture problem (Stumpf, 1911; Wallach, 1935; Todorović, 1996; Wuerger, Shapley, & Rubin, 1996; Bradley & Goyal, 2008). In many cases, the visual system resolves this ambiguity by integrating local velocities (Adelson & Movshon, 1982; Heeger, 1987; Recanzone, Wurtz, & Schwarz, 1997; Simoncelli & Heeger, 1998; Weiss, Simoncelli, & Adelson, 2002; Rust et al., 2006) or tracking specific features (Shiffrar & Pavel, 1991; Lorenceau & Shiffrar, 1992; Stoner & Albright, 1992; Wilson, Ferrera, & Yo, 1992; Ben-Av & Shiffrar, 1995; Lorenceau & Shiffrar, 1999; Pack, Livingstone, Duffy, & Born, 2003). In fact, humans can sometimes see unambiguous motion of shapes without a consistent motion-energy direction by tracking salient features (Cavanagh & Anstis, 1991; Lu & Sperling, 2001)—for example, where the features that segment a square from the background such as gratings with different orientations and contrasts change over time while the square moves laterally. There is no motion-energy information that can support the movement of the square; yet, we can reliably judge the direction of the movement. Consequently, to understand the contribution of salient features to percepts of rigidity, we built a feature-tracking network.

Figure 8A shows a schematic diagram of the network. We used two ways to extract the direction of pattern motion: tracking of extracted features (Lu & Sperling, 2001; Sun, Chubb, & Sperling, 2015) and motion energy combined into units resembling MT pattern-direction selective cells (Adelson & Movshon, 1982; Simoncelli, Adelson, & Heeger, 1991; Weiss et al., 2002; Rust et al., 2006). In the first module of the model (Figure 8A, bottom), features such as corners and sharp curvatures are extracted by the Harris corner detector (Harris & Stephens, 1988) but could also be extracted by end-stopped cells (Hubel & Wiesel, 1965; Dobbins, Zucker, & Cynader, 1987; Pasupathy & Connor, 1999; Pack et al., 2003; Rodríguez-Sánchez & Tsotsos, 2012). The extracted features are tracked by taking the correlation between successive images and using the highest across-image correlated location to estimate the direction and speed of the motion of each feature. The process is similar to that used to model the effects of shape in deciphering the rotation direction of non-rigid 2D shapes (Cohen, Jain, & Zaidi, 2010), and to measure the efficiency of stereo-driven rotational motion (Jain & Zaidi, 2013). Feature-tracking models used to dominate the computer vision literature, and, after locating features with nonlinear operations on the outputs of spatial filters at different scales, the correspondence problem was generally solved by correlation (Del Viva & Morrone, 1998). The equations for feature-tracking detection of motion as used in this study are presented in the Appendix (Feature tracking section).

The second module in feature tracking combines the outputs from motion-energy units into pattern direction–selective (PDS) mechanisms (Figure 8A, top). The red circle indicates the receptive field of one pattern-selective unit subtending 0.9°. The vectors within the receptive field are the motion-energy outputs that are inputs of the unit. The vectors emerging from the upper ring point down and to the right (indicated in green) and illustrate the aperture problem caused by a small receptive field size. The ambiguity created by the aperture problem is represented by the upper-panel likelihood function, with the probability distribution over the velocity components in horizontal and vertical directions (V_x and V_y) represented by a heatmap with yellow indicating high-probability and blue low-probability velocities. The spread in the high-likelihood yellow region is similar to Wallach's illustration using two infinitely long lines (Wallach, 1935). The lower-panel likelihood represents the aperture problem from motion-energy vectors from the bottom ring. To model an observer who estimates pattern velocity with much less uncertainty, the local uncertain measurements are multiplied together by a narrow prior for the slowest motion to obtain a

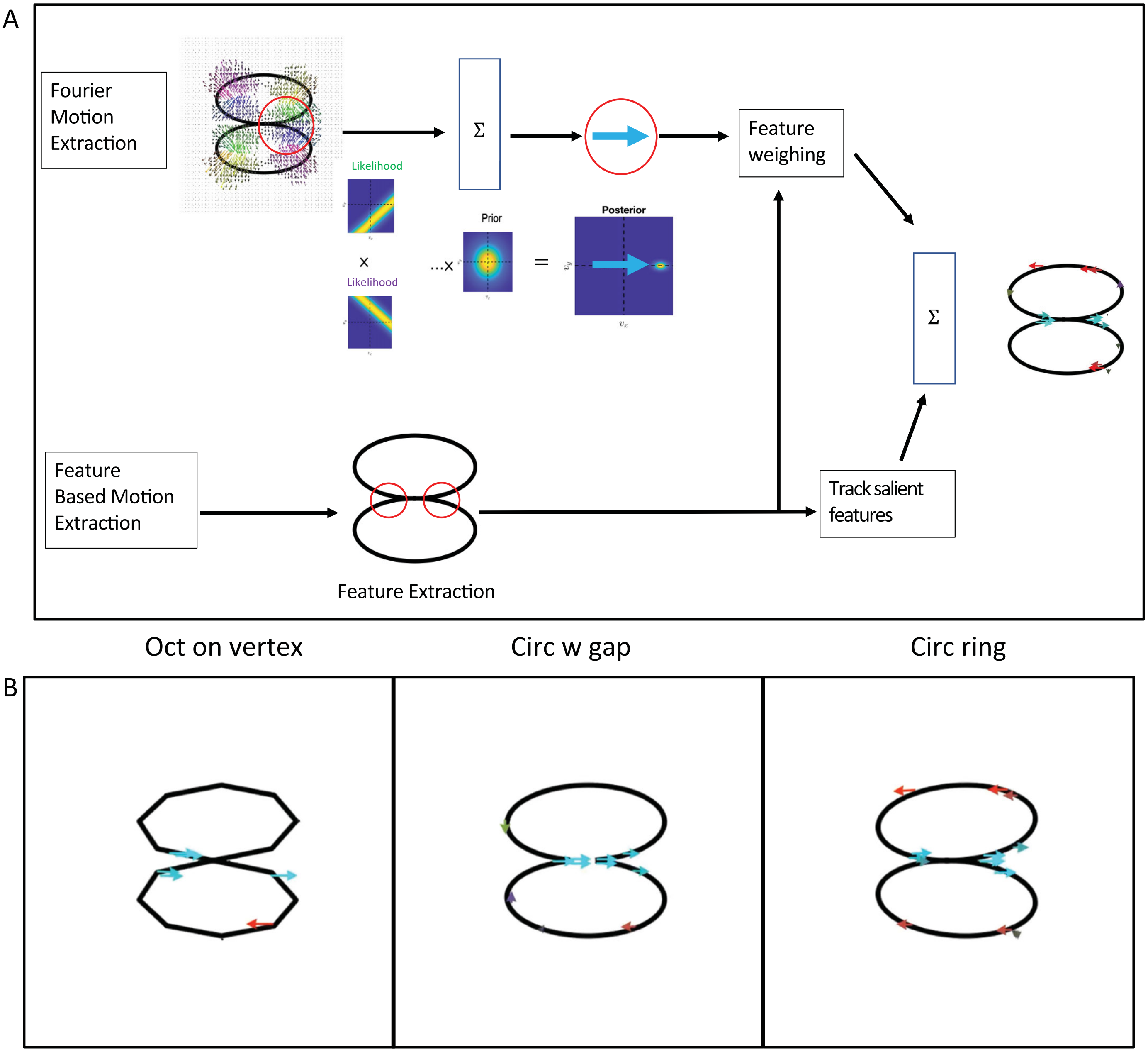

Figure 8. Feature-tracking mechanism. (**A**) Two feature-tracking streams simulating MT PDS units (top) and feature extraction–based motion (bottom). (**A**, top) The inputs are the vectors obtained from the motion-energy units, and each motion-energy vector creates a likelihood function that is perpendicular to the vector. Likelihood functions are combined with the slowest motion prior by Bayes' rule. The output vector at each location is selected by the MAP. (A, bottom) Salient features (corners) are extracted, and a velocity of the feature is computed by picking up the highest correlated location in the succeeding image. The outputs from two streams are combined. (**B**) The preferred velocity of the most responsive unit at each location is shown by a colored vector using the same key as Figure 6. Stimulus name is indicated on the bottom of each section. Most of the motion vectors point right or left, corresponding to the direction of rotation. Movie is available on the journal website.

velocity that maximizes the posterior distribution (maximum a posteriori probability [MAP]) (Simoncelli et al., 1991; Weiss et al., 2002). When the ring stimuli are run through arrays of pattern-direction selective cells, there is a response in the rotation direction at the joint and at some corners, but most of the responses are orthogonal to the contours as would be expected where locally there is a single contour (Zaharia, Goris, Movshon, & Simoncelli, 2019) and would support wobbling classifications from the CNN. Feature tracking either requires attention or is enhanced by it (Cavanagh, 1992; Lu & Sperling, 1995a;

Treue & Maunsell, 1999; Treue & Trujillo, 1999; Thompson & Parasuraman, 2012), so we used the corner detector to identify regions with salient features that could be tracked and simulated the effect of attention by attenuating the gain of PDS units falling outside windows of feature-based attention. The attention-based attenuation suppresses pattern-selective responses to long contours and preferentially accentuates the motion of features such as corners or dots (Noest & Van Den Berg, 1993; Pack, Gartland, & Born, 2004; Bradley & Goyal, 2008; Tsui, Hunter, Born, & Pack, 2010).

At the last stage of Figure 8A, the vector fields from the pattern-selective units and feature-tracking units are summed together. Figure 8B shows videos of samples of the outputs of the feature-tracking module for three examples of the ring stimuli, and the shape of the stimulus is shown at the bottom of each video. At the connection of the two rings, the preferred direction is mainly right or left, corresponding to the rotational direction. Depending on the phase of the rotation, the lateral velocities are also observed at sharp curvatures as the projected 2D contours deform significantly depending on the object pose angle (Koch, Baig, & Zaidi, 2018; Maruya & Zaidi, 2020a; Maruya & Zaidi, 2020b).

The combined vector fields were used as the inputs to the previously trained CNN for classification as rotating or wobbling. The results are shown in Figure 7B as probabilities of non-rigid classifications. The feature-tracking vector fields generate classification proportions from 0.1 to 0.2 indicating rigidity, suggesting that feature tracking could contribute to percepts of rigidity in the rotating rings.

Combining outputs of motion mechanisms for CNN classification

Psychophysical experiments that require detecting the motion direction of low-contrast gratings superimposed on stationary grating pedestals have shown that feature tracking happens only at slow speeds and that motion energy requires a minimum speed (Lu & Sperling, 1995b; Zaidi & DeBonet, 2000). We linearly combined the two vector fields attained from motion-energy units and feature-tracking units with a free weight parameter that was a function of speed, and the combined vector fields were fed to the trained CNN to simulate the observer's proportion of non-rigidity as a function of different shapes and different speeds (see Combining motion mechanisms section in the Appendix). We tried weights between 0.0 and 1.0 every 0.01 increment at each speed to minimize the mean square error (*MSE*) from the observers' average. The optimal weights are shown in Figure 9A. The weight for the feature tracking decreases and the weight for the motion energy increases as a function of rotation speed, consistent with published direct psychophysical results on the two motion mechanisms. In Figure 9B, the green bars show that the average proportion of non-rigid classifications generated by the CNN output across speeds is very similar to the average percepts (blue bars). However, the average across all shapes hides the fact that the proportion of non-rigid classifications from the CNN explains only a moderate amount of variance in the proportions of non-rigid percepts if examined for the complete set of shapes in Figures 9C and 9D ($R^2 = 0.64$). The scattergram shows that the CNN classification is systematically different from the perceptual results at the fastest speed where the prediction is flat across different shapes while the observers' responses vary with shape. Next, we examined possible factors that could modify percepts as a function of shape.

Priors and illusions

The video in Figure 1A shows that wobbling is not the only non-rigid percept when the circular rings are rotated, as the top ring also seems to be rolling around its center. Unlike the motion-energy support for wobbling and the feature-tracking support for rotation, there are no motion-energy or feature-tracking signals that would support the rolling percept (see Figures 5D and 5E and Figure 8B), which would require local motion vectors tangential to the contours of the rings. To illuminate the factors that could evoke or stop the rolling illusion, we show a simpler 2D rolling illusion. In the video in Figure 10A, a circular untextured 2D ring translated horizontally on top of a straight line is perceived predominantly as rolling like a tire on a road. The perception of rolling would be supported by motion signals tangential to the contour (Figure 10B), but the local velocities extracted from the stimulus by an array of motion-energy units are predominantly orthogonal to the contour as expected from the aperture effect (Figure 10C), whereas feature tracking extracts motion predominantly in the direction of the translation (Figure 10D), both of which should counter the illusion of clockwise rolling. Hence, the rolling illusion goes against the sensory information in the video. Note that, if an observer sees the ring as translating as well as spinning, then the translation vector will be added to the tangential vectors and Johansson's vector analysis into common and relative motions will be needed (Johansson,1994; Gershman, Tenenbaum, & Jäkel, 2016). Figure 10D just illustrates the tangential

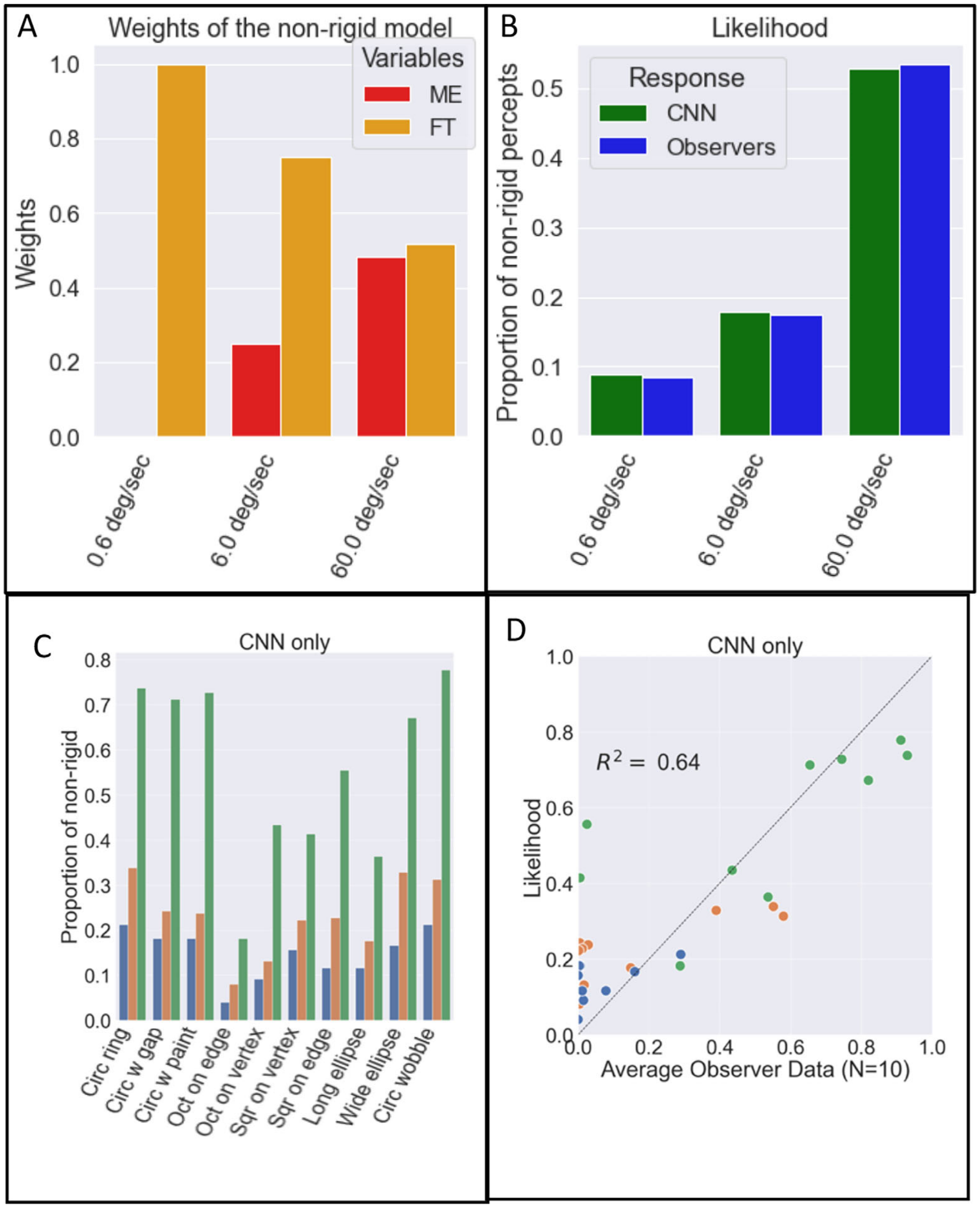

Figure 9. Combining motion-energy and feature-tracking outputs. (**A**) Estimated optimal weights of inputs from the motion-energy mechanism (red) and the feature-tracking mechanism (yellow) as a function of rotation speed over all shapes. (**B**) CNN rigidity versus non-rigidity classifications as a function of rotation speed. The trained CNN output from the linear combination of two vector fields is the likelihood, which is denoted by the green bar, whereas the blue bar indicates the average of the 10 observers' responses. (**C**) Proportion of non-rigid percepts from the likelihood function of the CNN as a function of the speed of the stimulus for different shapes. Different colors show different speeds of stimulus (blue, 0.6°/s; orange, 6.0°/s; green, 60.0°/s). (**D**) Likelihood of non-rigidity output plotted against the average of 10 observers' reports. Although $R^2 = 0.65$, at the fast speed, the model predicts similar probability of non-rigidity for shapes where the observers' percepts vary. Thus, the model does not capture some important properties of observers' percepts as a function of the shape of the object.

motion signals that will have to be added to the translation vector. This illusion could demonstrate the power of prior probabilities of motion types. To identify factors that enhance or attenuate the illusion, we performed experiments on the two-ring configurations.

Quantifying the rolling illusion

We quantified the perception of rolling in the original ring illusion by using the same ring pairs as in Figure 3 (0.6°/s, 6.0°/s, 60.0°/s; 3°), but now on each

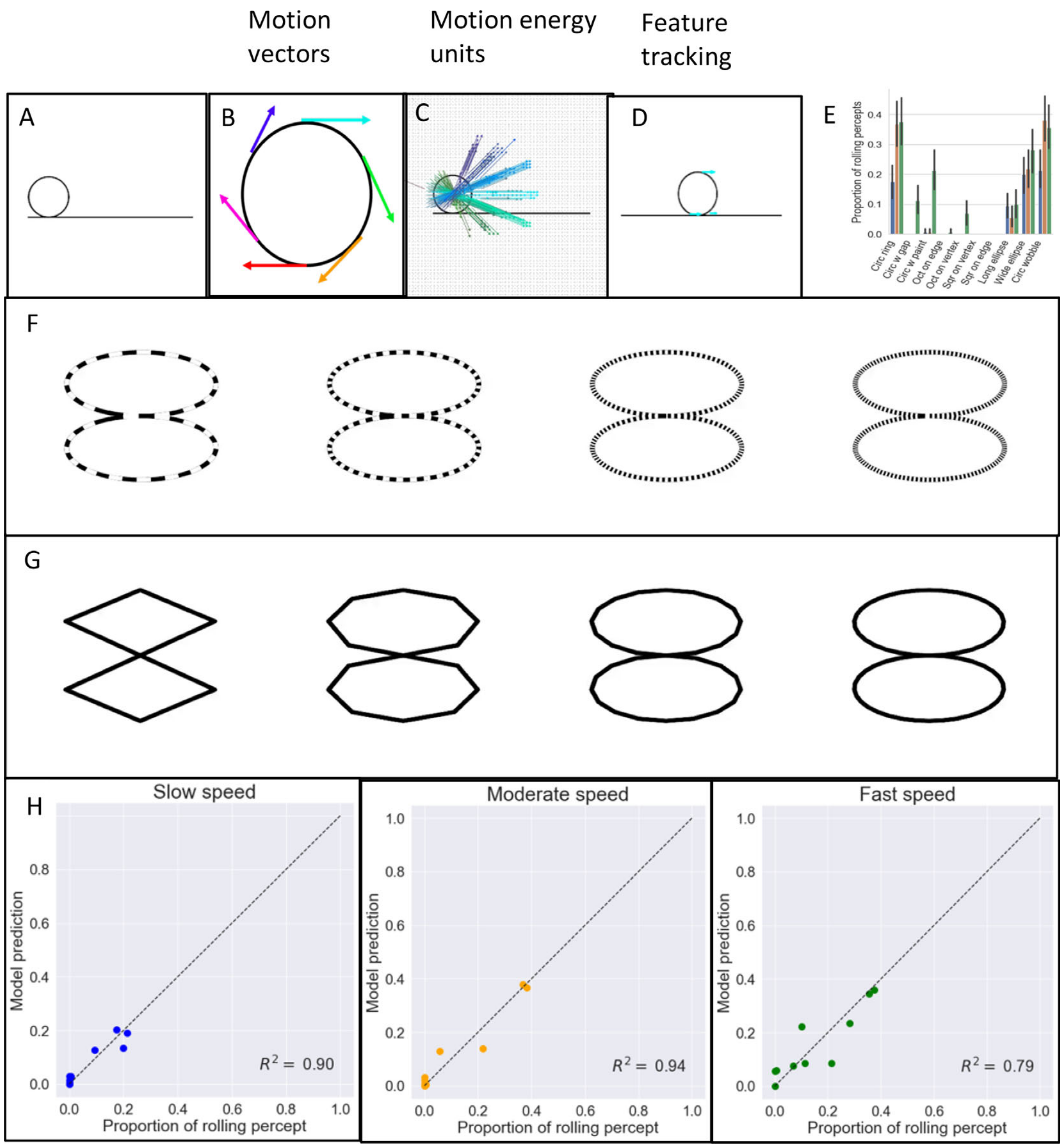

Figure 10. Rolling illusion. (**A**) A 2D circle on a line, translating from left to right but perceived as rolling clockwise. (**B**) To perceive rolling, local units that detect motion tangential to the contour are required. (**C**, **D**) Local motion signals from motion energy and feature tracking, respectively. In both cases, the vectors are inconsistent with the required tangential vectors. (**E**) Average proportion of rolling percepts (eight observers). The color of the bar indicates the speed of the stimulus (blue, 0.6°/s; orange, 6.0°/s; green, 60.0°/s). The shape of the stimulus is indicated on the *X*-axis. The proportion of rolling percepts increased with speed and decreased when salient features were added to the rings. (**F**) Rolling illusion and rotational symmetry. The non-rigidity (rolling) percepts increase with the order of rotational symmetry from left to right. (**G**) The relationship between rolling illusion and the strength of feature. As the number of corners increase from left to right, they become less salient as the increasingly obtuse angles become more difficult to extract, and accordingly the percept of rolling increases. (**H**) Model prediction with rotational symmetry and average strength of features versus average proportion of rolling percepts for slow (left), moderate (middle), and fast (right) speeds ($R^2 = 0.90$, 0.94, and 0.79, respectively). Movie is available on the journal website.

trial the observers were asked to respond “yes” or “no” to the question of whether the rings were rolling individually around their own centers. The results are plotted in Figure 10E (for 20 repetitions/condition and eight observers). The frequency of trials seen as rolling increased with speed and decreased when gaps, paint, or corners were added to the rings, with corners leading to the greatest decrease. We think this illusion demonstrates the power of prior expectations for rolling for different shapes. The prior probability for rolling could reflect the rotational symmetry of the shape, as circular rings with higher order rotational symmetries are more likely to be seen as rolling and wobbling in Figure 10F. In addition, the more features, such as corners, that are seen as not rolling, the more they may attenuate the illusion as shown in Figure 10G. It is possible that aliasing increases with the degree of symmetry at high speeds and features may be less effective at high speeds if they get blurred in the visual system, explaining greater rolling and wobbling at higher speeds. The degree of rotational symmetry of a circle is infinite, so we reduced it to the number of discrete pixels in the circumference and regressed the proportion of rolling percepts against the log of the order of rotational symmetry of each shape and the mean strength of features, $\bar{h}$ (the average value of h defined in Equation A18). The two factors together predicted rolling frequency with $R^2 = 0.90$, 0.94, and 0.79 for slow, medium, and fast speeds, respectively (Figure 10H).

Adding prior assumptions to motion mechanism–based CNN classification for rigid and non-rigid perception of rotating ring pairs

The first model showed that to completely explain the variation of the illusion where rigid rings are perceived as non-rigidly connected as a function of speed and shape, other factors have to be considered besides the outputs of bottom-up motion mechanisms. The degree of rotational symmetry may supply not only a prior for rolling but also for wobbling as a priori a circular ring is more likely to wobble than a square ring. We thus set up a prior dependent on the number of rotational symmetries and the average strength of the detected features. The posterior probability of a motion class is thus conditional on the motion-vector fields, the speed that determines the relative weights of the motion-energy and feature-tracking motion fields, and the shape-based priors. These factors are combined with weights and the posterior is computed by using Bayes' rule (see Final model section in the Appendix). When the weights were estimated by using gradient descent to minimize the MSE, the three factors together predicted proportions of non-rigid percepts with $R^2 = 0.95$ for all three speeds combined (Figure 11).

Discussion

We began this article by stating that our aim was to bridge the gap between understanding neuronal stimulus preferences and understanding how cooperation and competition among different classes of neuronal responses generate visual perception, as well as to identify the factors that govern transitions between stable states of perception. We now discuss how we have attempted this and how far we have succeeded.

We used a variant of a physical illusion that has been widely used but never explained. A pair of circular rings

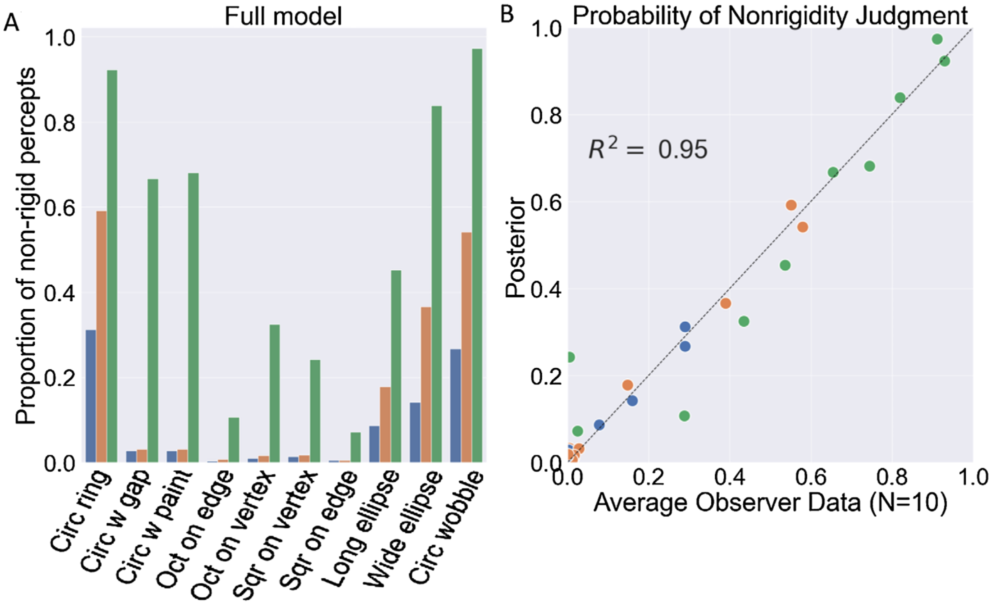

Figure 11. Final model. (**A**) The proportion of non-rigid percept classifications as a function of the speed and shape of the stimulus from the final model (combining shape-based rolling and wobbling priors with the CNN likelihoods). Different colors indicate different speeds of the stimulus (blue, 0.6°/s; orange, 6.0°/s; green, 60.0°/s). (**B**) The predictions of the final model are plotted against the average observer percepts. The model explains the observer's percepts ($R^2 = 0.95$).

rigidly connected appear non-rigidly connected when rotated. By varying rotation speed, we discovered that, even for physical objects, percepts of rigidity dominate at slow speeds and percepts of non-rigidity dominate at high speeds. We then varied shapes using computer graphics and discovered that the presence of vertices or gaps or painted segments promoted percepts of rigid rotation at moderate speeds where the circular rings looked non-rigid, but at high speeds all shapes appeared non-rigid.

By analyzing the stream of images for motion signals, we found that the non-rigid wobble can be explained by the velocity field evoked by the rigidly rotating object in an array of motion-energy units because of the limited aperture of the units. This explanation, of course, depends on the decoding of the motion field. In the absence of knowledge of biological decoders for object motion, it is possible to infer the degree of rigidity from the velocity field itself. Analytic tests of the geometrical rigidity of velocity fields can be based on the decomposition of the velocity gradient matrix or on an analysis of the temporal derivative of the curvature of moving plane curves, but limitations of both approaches have been noted (Todorović, 1993). We did not attempt to recreate the complete percept or to model elastic versus articulated motion (Jasinschi & Yuille, 1989; Aggarwal, Cai, Liao, & Sabata, 1998, Jain & Zaidi, 2011) but restricted the analysis to distinguish wobbling from rigid rotation. For this purpose, we trained a CNN that could make this distinction for many velocity fields. The CNN indicated an almost 100% confidence in wobble from the velocity fields of all shapes and at all speeds, pointing to the need for other mechanisms that take speed and shape into account to explain the results.

The obvious second mechanism to explore was feature tracking, whose output can depend on the salience of features. For this purpose, we used the Harris corner detector that is widely used in computer vision, but also pattern motion selective units. The velocity field from this mechanism was judged by the trained CNN to be compatible with rigid rotation, with little variation based on shape. The output of the CNN can be considered as classifying the information present in a velocity field without committing to a particular decoding process. By making the empirically supported assumption that motion energy requires a minimum speed and that feature tracking functions only at slow speeds, the CNN output from the two vector fields could be combined to explain the empirical results with an $R^2 = 0.64$, mainly by accounting for the speed effects. An inspection of empirical versus predicted results showed the need for mechanisms or factors that were more dependent on shape.

In both the physical and graphical stimuli, observers see an illusion of the rings rolling or spinning around their own center. This illusion is remarkable because there are no sensory signals that support it, shown starkly by translating a circular ring along a line. The illusion is however suppressed by vertices, gaps and painted segments, suggesting that a powerful prior for rolling may depend on rotational symmetry or jaggedness of the shape. The addition of this shape-based prior to the model, leads to an $R^2 = 0.95$, which suggests that we have almost completely accounted for the most important factors.

All of the shape changes in our stimuli reduce the non-rigidity illusion. A shape change that separates the disks in height so that they do not physically touch seems to slightly increase the percept of non-rigidity, as would be expected by the loss of some feature-tracking information at the junction. The closest illusion to ours was devised by Wallach (1976): When a 2D figure made of two overlapping circular disks is rotated rigidly about its center, the two disks are perceived to slide over each other in non-rigid motion. If the circular disks are changed to polygons, the disks can be perceived to be yoked together or move independently (Hupé & Rubin, 2000). It is obvious that the 2D illusory percepts of independently moving rings can be explained by our model's motion-energy component and the veridical yoked percept by feature tracking. Rotating 2D ellipses also generate the remarkable Benussi–Musatti stereokinetic illusions (Musatti, 1924). Stereokinetic illusions are generated by unchanging 2D ellipses, whereas in our stimuli the 2D projections of the 3D rings change with rotation angle. The resulting illusions are qualitatively different, as narrow 2D ellipses are seen as more rigid than broad ellipses in the stereokinetic displays (Weiss & Adelson, 2000), whereas 3D rings that are wide in the horizontal (rotation) direction, and thus project to narrower ellipses, are more likely to be seen as non-rigid than the 3D rings that are long in the vertical direction and thus project to broad ellipses (Figures 3 and 4).

In the absence of direct physiological evidence, we have thought about possible neural substrates for components of our model. The neural substrate for motion-energy mechanisms is well established as direction-selective cells in primary visual cortex (V1) that project to multiple areas containing motion-sensitive cells, such as MT, MST, and V3, that have larger receptive fields. The neural substrate for feature tracking is much less certain. We have found no references to electrophysiological or imaging studies of cortex for tracking of visual features. There are a handful of cortical studies of second-order motion that may be relevant, because feature tracking has been proposed as a mechanism for detecting the direction of second-order motion (Seiffert & Cavanagh, 1998; Derrington & Ukkonen, 1999). Motion of contrast-modulated noise gives rise to blood oxygen level–dependent (BOLD) signals in areas V3, VP, and MT (Smith, Greenlee, Singh, Kraemer, & Hennig, 1998), but the stimuli used do not require the tracking of extracted features, as motion direction is predicted well by a filter–rectify–filter model (Chubb & Sperling, 1988). So, the best we could do is suggest that, if sensitivity to feature tracking is measured in neurons using variations in shapes and speeds (as we have done), it may be better to start in the medial superior temporal area MST and regions in the posterior parietal cortex (Erlikhman et al., 2018; Freud et al., 2020).

The neural substrate for priors and where sensory information combines with stored knowledge is even more elusive. A number of functional magnetic resonance imaging (fMRI) studies have shown the effects of expectations that are generated by varying frequencies of presentation during the experiment (Esterman & Yantis, 2010; Kok, Jehee, & De Lange, 2012; Kumar, Kaposvari, & Vogels, 2017), and fMRI studies have used degraded stimuli to separate image-specific information in specific brain regions from the traditional effects of priming (Gorlin et al., 2012; González-García & He, 2021), but we have found only one neural study on the influence of long-term priors on visual perception (Hardstone et al., 2021). This study used electrophysiological monitoring with electrodes placed directly on the exposed surface of the cerebral cortex of neurological patients to measure gross electrical activity while viewing a Necker cube as percepts switched from a configuration compatible with viewed from the top to a configuration compatible with viewed from the bottom. Granger causality was used to infer greater feedback from temporal to occipital cortex during the viewed from the top phase, supposedly corresponding to a long-term prior (Mamassian & Landy, 1998; Troje & McAdam, 2010), and greater feedforward drive was inferred during the viewed from the bottom phase. Percepts of the Necker cube are more variable than this dichotomy (Mason, Kaszor, & Bourassa, 1973) and can be varied voluntarily by choosing the intersection to attend (Peterson & Hochberg, 1983; Hochberg & Peterson, 1987), revealing that the whole configuration is not the effective stimulus for perception. This, coupled with the lack of characterization of the putative feedback signal, suggests that the neural locus of the effect of long-term priors still remains to be identified. The use of shape-based priors in our experiments with their graded effects could provide more effective stimuli for this purpose.

Our study could provide a new impetus to neural studies of perceptual phase transitions. There is a clear shift of phase from rigidity at slow speeds to non-rigidity at fast speeds that could be attributed to the activity of different sets of neurons, motion-energy units functioning above a threshold velocity versus feature-tracking units that do not function at high velocities. However, at medium velocities, not only does rigidity versus non-rigidity depend on shape but the percept is also bistable and depends on the attended portion of the stimulus, just like the Necker cube. Rigidity is more likely to be seen when attention is directed to the junction, as observers were instructed to do in the experiment, but the probability of non-rigidity increases as attention is directed to the contour at the top or bottom of the display. Phase transitions have been extensively modeled in physics, building on the Kadanoff–Wilson work on scaling and renormalization (Kadanoff, 1966; Wilson, 1971), but these models assume all units to be identical, only local influences, conservation, and symmetry (Goldenfeld & Kadanoff, 1999), which are unrepresentative of neural circuits. Brain areas in the frontoparietal cortex that are active during perceptual phase transitions have been identified by several fMRI studies using binocular rivalry and a variety of bistable stationary and moving stimuli (Brascamp, Sterzer, Blake, & Knapen, 2018), but almost nothing has been identified at the level of single neurons or specified populations of neurons. In the absence of direct measurements on single neurons, models have used abstract concepts of adaptation linked to mutual inhibition and probabilistic sampling linked to predictive coding (Gershman, Vul, & Tenenbaum, 2012; Block, 2018). The stimuli used in our study could provide more direct tests of mechanistic models of competition and cooperation between groups of neurons, because we have identified the properties of neuronal populations that are dominant in each phase.

To summarize, we have shown how visual percepts of rigidity or non-rigidity can be based on the information provided by different classes of neuronal mechanisms, combined with shape-based priors. We further showed that the transition from perception of rigidity to non-rigidity depends on the speed requirements of different neuronal mechanisms.

Keywords: object rigidity, non-rigidity, motion illusion, feature tracking, shape priors

Acknowledgments

Supported by a grant from the National Eye Institute, National Institutes of Health (EY035085).

Commercial relationships: none.

Corresponding author: Qasim Zaidi.
Email: qz@sunyopt.edu.
Address: Graduate Center for Vision Research, State University of New York, New York, NY, USA.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv,* https://doi.org/10.48550/arXiv.1603.04467.

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A, 2*(2), 284–299.

Adelson, E. H., & Movshon, J. A. (1982). Phenomenal coherence of moving visual patterns. *Nature, 300*(5892), 523–525.

Aggarwal, J. K., Cai, Q., Liao, W., & Sabata, B. (1998). Nonrigid motion analysis: Articulated and elastic motion. *Computer Vision and Image Understanding, 70*(2), 142–156.

Akhter, I., Sheikh, Y., Khan, S., & Kanade, T. (2010). Trajectory space: A dual representation for nonrigid structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(7), 1442–1456.

Albertazzi, L. (2004). Stereokinetic shapes and their shadows. *Perception, 33*(12), 1437–1452.

Andersen, R. A., & Bradley, D. C. (1998). Perception of three-dimensional structure from motion. *Trends in Cognitive Sciences, 2*(6), 222–228.

Ben-Av, M. B., & Shiffrar, M. (1995). Disambiguating velocity estimates across image space. *Vision Research, 35*(20), 2889–2895.

Berzhanskaya, J., Grossberg, S., & Mingolla, E. (2007). Laminar cortical dynamics of visual form and motion interactions during coherent object motion perception. *Spatial Vision, 20*(4), 337–395.

Block, N. (2018). If perception is probabilistic, why does it not seem probabilistic? *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1755), 20170341.

Bradley, D. C., & Goyal, M. S. (2008). Velocity computation in the primate visual system. *Nature Reviews Neuroscience, 9*(9), 686–695.

Brascamp, J., Sterzer, P., Blake, R., & Knapen, T. (2018). Multistable perception and the role of the frontoparietal cortex in perceptual inference. *Annual Review of Psychology, 69*, 77–103.

Bregler, C., Hertzmann, A., & Biermann, H. (2000). Recovering non-rigid 3D shape from image streams. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000* (Vol. *2*, pp. 690–696). Los Alamitos, CA: IEEE Computer Society.

Cavanagh, P. (1992). Attention-based motion perception. *Science, 257*(5076), 1563–1565.

Cavanagh, P., & Anstis, S. (1991). The contribution of color to motion in normal and color-deficient observers. *Vision Research 31*, 2109–2148.

Chubb, C., & Sperling, G. (1988). Driftbalanced random stimuli: A general basis for studying non-fourier motion perception. *Journal of the Optical Society of America A, 5*(11), 1986–2007.

Cohen, E. H., Jain, A., & Zaidi, Q. (2010). The utility of shape attributes in deciphering movements of non-rigid objects. *Journal of Vision, 10*(11):29, 1–15, http://www.journalofvision.org/content/10/11/29.

Del Viva, M. M., & Morrone, M. C. (1998). Motion analysis by feature tracking. *Vision Research, 38*(22), 3633–3653.

Derrington, A. M., & Ukkonen, O. I. (1999). Second-order motion discrimination by feature-tracking. *Vision Research, 39*(8), 1465–1475.

Dobbins, A., Zucker, S. W., & Cynader, M. S. (1987). Endstopped neurons in the visual cortex as a substrate for calculating curvature. *Nature, 329*(6138), 438–441.

Duffy, C. J., & Wurtz, R. H. (1991). Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large field stimuli. *Journal of Neurophysiology, 65*(6), 1329–1345.

Erlikhman, G., Caplovitz, G. P., Gurariy, G., Medina, J., & Snow, J. C. (2018). Towards a unified perspective of object shape and motion processing in human dorsal cortex. *Consciousness and Cognition, 64*, 106–120.

Esterman, M., & Yantis, S. (2010). Perceptual expectation evokes category-selective cortical activity. *Cerebral Cortex, 20*(5), 1245–1253.

Fernández, J. M., Watson, B., & Qian, N. (2002). Computing relief structure from motion with a distributed velocity and disparity representation. *Vision Research, 42*(7), 883–898.

Freud, E., Behrmann, M., & Snow, J. C. (2020). What does dorsal cortex contribute to perception?. *Open Mind, 4*, 40–56.

Gallant, S. I. (1990). Perceptron-based learning algorithms. *IEEE Transactions on Neural Networks, 1*(2), 179–191.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation, 24*(1), 1–24.

Gershman, S. J., Tenenbaum, J. B., & Jäkel, F. (2016). Discovering hierarchical motion structure. *Vision Research, 126*, 232–241.

Goldenfeld, N., & Kadanoff, L. P. (1999). Simple lessons from complexity. *Science, 284*(5411), 87–89.

González-García, C., & He, B. J. (2021). A gradient of sharpening effects by perceptual prior across the human cortical hierarchy. *Journal of Neuroscience, 41*(1), 167–178.

Gorlin, S., Meng, M., Sharma, J., Sugihara, H., Sur, M., & Sinha, P. (2012). Imaging prior information in the brain. *Proceedings of the National Academy of Sciences, USA, 109*(20), 7935–7940.

Hardstone, R., Zhu, M., Flinker, A., Melloni, L., Devore, S., Friedman, D., . . . He, B. J. (2021). Long-term priors influence visual perception through recruitment of long-range feedback. *Nature Communications, 12*(1), 6288.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In C. J. Taylor (Ed.), *Proceedings of the Alvey Vision Conference* (pp. 23.1–23.6). Manchester, UK: Alvey Vision Club.

Heeger, D. J. (1987). Model for the extraction of image flow. *Journal of the Optical Society of America A, 4*(8), 1455–1471.

Hochberg, J., & Peterson, M. A. (1987). Piecemeal organization and cognitive components in object perception: Perceptually coupled responses to moving objects. *Journal of Experimental Psychology: General, 116*(4), 370.

Hubel, D. H. (1959). Single unit activity in striate cortex of unrestrained cats. *The Journal of Physiology, 147*(2), 226.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology, 148*(3), 574–591.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*(1), 106–154.

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology, 28*(2), 229–289.

Hupé, J. M., & Rubin, N. (2000). Perceived motion transparency can override luminance/color cues which are inconsistent with transparency. *Investigative Ophthalmology & Visual Science*, *41*(4), S721–S721.

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

Jain, A., & Zaidi, Q. (2013). Efficiency of extracting stereo-driven object motions. *Journal of Vision, 13*(1):18, 1–14, https://doi.org/10.1167/13.1.18.

Jasinschi, R., & Yuille, A. (1989). Nonrigid motion and Regge calculus. *Journal of the Optical Society of America A, 6*(7), 1088–1095.

Johansson, G. (1994). Configurations in event perception: An experimental study. In G. Jansson, S. S. Bergström, & W. Epstein (Eds.), *Perceiving events and objects* (pp. 29–122). Mahwah, NJ: Lawrence Erlbaum Associates. (Reprinted in modified form from G. Johansson's 1950 doctoral dissertation.)

Kadanoff, L. P. (1966). Scaling laws for Ising models near T_c. *Physics Physique Fizika, 2*(6), 263.

Koch, E., Baig, F., & Zaidi, Q. (2018). Picture perception reveals mental geometry of 3D scene inferences. *Proceedings of the National Academy of Sciences, USA, 115*(30), 7807–7812.

Kok, P., Jehee, J. F., & De Lange, F. P. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron, 75*(2), 265–270.

Kumar, S., Kaposvari, P., & Vogels, R. (2017). Encoding of predictable and unpredictable stimuli by inferior temporal cortical neurons. *Journal of Cognitive Neuroscience, 29*(8), 1445–1454.

Lorenceau, J., & Shiffrar, M. (1992). The influence of terminators on motion integration across space. *Vision Research, 32*(2), 263–273.

Lorenceau, J., & Shiffrar, M. (1999). The linkage of visual motion signals. *Visual Cognition, 6*(3-4), 431–460.

Lu, Z. L., & Sperling, G. (1995a). Attention-generated apparent motion. *Nature, 377*(6546), 237–239.

Lu, Z. L., & Sperling, G. (1995b). The functional architecture of human visual motion perception. *Vision Research, 35*(19), 2697–2722.

Lu, Z. L., & Sperling, G. (2001). Three-systems theory of human visual motion perception: review and update. *Journal of the Optical Society of America A, 18*(9), 2331–2370.

Mach, E. (1886). Beitra¨ge zur Analyse der Empfindungen. Jena: Gustav Fischer. In: C. M. Williams (Trans.), *English translation: Contributions to the analysis of the sensations* (p. 1897). Chicago: The Open Court.

Mamassian, P., & Landy, M. S. (1998). Observer biases in the 3D interpretation of line drawings. *Vision Research, 38*(18), 2817–2832.

Maruya, A., & Zaidi, Q. (2020a). Mental geometry of three-dimensional size perception. *Journal of Vision, 20*(8):14, 1–16, https://doi.org/10.1167/jov.20.8.14.

Maruya, A., & Zaidi, Q. (2020b). Mental geometry of perceiving 3D size in pictures. *Journal of Vision, 20*(10):4, 1–16, https://doi.org/10.1167/jov.20.10.4.

Mason, J., Kaszor, P., & Bourassa, C. M. (1973). Perceptual structure of the Necker cube. *Nature, 244*(5410), 54–56.

McDermott, J., & Adelson, E. H. (2004). The geometry of the occluding contour and its effect on motion interpretation. *Journal of Vision, 4*(10):9, 944–954, https://doi.org/10.1167/4.10.9.

McDermott, J., Weiss, Y., & Adelson, E. H. (2001). Beyond junctions: Nonlocal form constraints on motion interpretation. *Perception, 30*(8), 905–923.

Movshon, J.A., Adelson, E.H., Gizzi, M.S., & Newsome, W.T. (1985). The analysis of moving visual patterns. In C. Chagas, R. Gattass, & C. Gross (Eds.), *Pattern recognition mechanisms (Pontificiae Academiae Scientiarum Scripta Varia)* (Vol. *54*, pp. 117–151). Rome: Vatican Press.

Movshon, J. A., & Newsome, W. T. (1996). Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. *Journal of Neuroscience, 16*(23), 7733–7741.

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978a). Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *The Journal of Physiology, 283*(1), 53–77.

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978b). Receptive field organization of complex cells in the cat's striate cortex. *The Journal of Physiology, 283*(1), 79–99.

Musatti, C. L. (1924). Sui fenomeni stereocinetici. *Archivio italiano di psicologia, 3*, 105–120.

Nishimoto, S., & Gallant, J. L. (2011). A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies. *Journal of Neuroscience, 31*(41), 14551–14564.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology, 21*(19), 1641–1646.

Noest, A. J., & Van Den Berg, A. V. (1993). The role of early mechanisms in motion transparency and coherence. *Spatial Vision, 7*(2), 125–147.

Pack, C. C., Gartland, A. J., & Born, R. T. (2004). Integration of contour and terminator signals in visual area MT of alert macaque. *Journal of Neuroscience, 24*(13), 3268–3280.

Pack, C. C., Livingstone, M. S., Duffy, K. R., & Born, R. T. (2003). End-stopping and the aperture problem: two-dimensional motion signals in macaque V1. *Neuron, 39*(4), 671–680.

Papathomas, T. V. (2002). Experiments on the role of painted cues in Hughes's reverspectives. *Perception, 31*(5), 521–530.

Pasupathy, A., & Connor, C. E. (1999). Responses to contour features in macaque area V4. *Journal of Neurophysiology, 82*(5), 2490–2502.

Peterson, M. A., & Hochberg, J. (1983). Opposed-set measurement procedure: A quantitative analysis of the role of local cues and intention in form perception. *Journal of Experimental Psychology: Human Perception and Performance, 9*(2), 183.

Pitzalis, S., Sereno, M. I., Committeri, G., Fattori, P., Galati, G., Patria, F., . . . Galletti, C. (2010). Human V6: The medial motion area. *Cerebral Cortex, 20*(2), 411–424.

Recanzone, G. H., Wurtz, R. H., & Schwarz, U. (1997). Responses of MT and MST neurons to one and two moving objects in the receptive field. *Journal of Neurophysiology, 78*(6), 2904–2915.

Rodríguez-Sánchez, A. J., & Tsotsos, J. K. (2012). The roles of endstopped and curvature tuned computations in a hierarchical representation of 2D shape. *PLoS One, 7*(8), e42058.

Rokers, B., Yuille, A., & Liu, Z. (2006). The perceived motion of a stereokinetic stimulus. *Vision Research, 46*(15), 2375–2387.

Rust, N. C., Mante, V., Simoncelli, E. P., & Movshon, J. A. (2006). How MT cells analyze the motion of visual patterns. *Nature Neuroscience, 9*(11), 1421–1431.

Sakata, H., Shibutani, H., Ito, Y., & Tsurugai, K. (1986). Parietal cortical neurons responding to rotary movement of visual stimulus in space. *Experimental Brain Research, 61*(3), 658–663.

Seiffert, A. E., & Cavanagh, P. (1998). Position displacement, not velocity, is the cue to motion detection of second-order stimuli. *Vision Research, 38*(22), 3569–3582.

Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. *Towards Data Science, 6*(12), 310–316

Shiffrar, M., & Pavel, M. (1991). Percepts of rigid motion within and across apertures. *Journal of Experimental Psychology: Human Perception and Performance, 17*(3), 749.

Simoncelli, E. P., Adelson, E. H., & Heeger, D. J. (1991). Probability distributions of optical flow. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 310–315). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research, 38*(5), 743–761.

Smith, A. T., Greenlee, M. W., Singh, K. D., Kraemer, F. M., & Hennig, J. (1998). The processing of first-and second-order motion in human visual cortex assessed by functional magnetic resonance imaging (fMRI). *Journal of Neuroscience, 18*(10), 3816–3830.

Stoner, G. R., & Albright, T. D. (1992). Neural correlates of perceptual motion coherence. *Nature, 358*(6385), 412–414.

Stumpf, P. (1911). Über die Abhängigkeit der visuellen Bewegungsrichtung und negativen Nachbildes von den Reizvorgangen auf der Netzhaut. *Zeitschrift fur Psychologie, 59*, 321–330.

Sun, P., Chubb, C., & Sperling, G. (2015). Two mechanisms that determine the Barber-Pole Illusion. *Vision Research, 111*, 43–54.

Tanaka, K., Hikosaka, K., Saito, H. A., Yukie, M., Fukada, Y., & Iwai, E. (1986). Analysis of local and wide-field movements in the superior temporal visual areas of the macaque monkey. *Journal of Neuroscience, 6*(1), 134–144.

Thompson, J., & Parasuraman, R. (2012). Attention, biological motion, and action recognition. *NeuroImage, 59*(1), 4–13.

Todorović, D. (1993). Analysis of two-and three-dimensional rigid and nonrigid motions in the stereokinetic effect. *Journal of the Optical Society of America A, 10*(5), 804–826.

Todorović, D. (1996). A gem from the past: Pleikart Stumpf's (1911) anticipation of the aperture problem, Reichardt detectors, and perceived motion loss at equiluminance. *Perception, 25*(10), 1235–1242.

Treue, S., & Maunsell, J. H. (1999). Effects of attention on the processing of motion in macaque middle temporal and medial superior temporal visual cortical areas. *Journal of Neuroscience, 19*(17), 7591–7602.

Treue, S., & Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature, 399*(6736), 575–579.

Troje, N. F., & McAdam, M. (2010). The viewing-from-above bias and the silhouette illusion. *i-Perception, 1*(3), 143–148.

Tsui, J. M., Hunter, J. N., Born, R. T., & Pack, C. C. (2010). The role of V1 surround suppression in MT motion integration. *Journal of Neurophysiology, 103*(6), 3123–3138.

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

Van Santen, J. P., & Sperling, G. (1985). Elaborated Reichardt detectors. *Journal of the Optical Society of America A, 2*(2), 300–321.

Vezzani, S., Kramer, P., & Bressan, P. (2014). Stereokinetic effect, kinetic depth effect, and structure from motion. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 521–540). Oxford, UK: Oxford Library of Psychology.

Wallach, H. (1935). Über visuell wahrgenommene Bewegungsrichtung. *Psychologische Forschung, 20*(1), 325–380.

Wallach, H. (1976). *On perception*. New York: Times Book Company.

Wallach, H., & O'Connell, D. N. (1953). The kinetic depth effect. *Journal of Experimental Psychology, 45*(4), 205.

Watson, A. B., & Ahumada, A. (1983). *A look at motion in the frequency domain*. Silicon Valley, CA: Ames Research Center, National Aeronautics and Space Administration.

Watson, A. B., & Ahumada, A. J. (1985). Model of human visual-motion sensing. *Journal of the Optical Society of America A, 2*(2), 322–342.

Weiss, Y., & Adelson, E. H. (2000). Adventures with gelatinous ellipses—constraints on models of human motion analysis. *Perception, 29*(5), 543–566.

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

Wilson, K. G. (1971). Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture. *Physical Review B, 4*(9), 3174.

Wilson, H. R., Ferrera, V. P., & Yo, C. (1992). A psychophysically motivated model for two-dimensional motion perception. *Visual Neuroscience, 9*(1), 79–97.

Wuerger, S., Shapley, R., & Rubin, N. (1996). "On the Visually Perceived Direction of Motion" by Hans Wallach: 60 years later. *Perception, 25*(11), 1317–1367.

Zaharia, A. D., Goris, R. L., Movshon, J. A., & Simoncelli, E. P. (2019). Compound stimuli reveal the structure of visual motion selectivity in macaque MT neurons. *eNeuro, 6*(6):ENEURO.0258–19.2019.

Zaidi, Q., & DeBonet, S. J. (2000). Motion energy versus position tracking: Spatial, temporal and chromatic parameters. *Vision Research, 40*(26), 3613–3635.

Zhang, K., Sereno, M. I., & Sereno, M. E. (1993). Emergence of position-independent detectors of sense of rotation and dilation with Hebbian learning: An analysis. *Neural Computation, 5*(4), 597–612.

Appendix

Stimulus generation and projection

We generated 3D rotating, wobbling, and rolling stimuli by applying the equations for rotation along each X-axis, Y-axis, and Z-axis in a 3D space to all points on the rendered objects. The rotational matrix around each axis, $\boldsymbol{R}_{axis}(\theta)$ is:

$$\boldsymbol{R}_X(\theta) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix} \quad \text{(A1)}$$

$$\boldsymbol{R}_Y(\theta) = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \quad \text{(A2)}$$

$$\boldsymbol{R}_Z(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{(A3)}$$

If $\vec{\boldsymbol{o}}$ is the initial position of a 3D point on the object lying on the XY plane, the location of the point ($\overrightarrow{\boldsymbol{o_{rot}}}$) on a rotating object inclined at an angle of ϕ from the ground plane and angular velocity of ω is expressed by Equation A4:

$$\overrightarrow{\boldsymbol{o_{rot}}} = \boldsymbol{R}_Z(\omega t)\,\boldsymbol{R}_Y(\phi)\,\vec{\boldsymbol{o}} \quad \text{(A4)}$$

The wobbling object is described by Equation A5:

$$\overrightarrow{\boldsymbol{o_{wbl}}} = \boldsymbol{R}_Z(\omega t)\,\boldsymbol{R}_Y(\phi)\,\boldsymbol{R}_Z(-\omega t)\,\vec{\boldsymbol{o}} \quad \text{(A5)}$$

The rolling object is described by Equation A6:

$$\overrightarrow{\boldsymbol{o_{rll}}} = \boldsymbol{R}_Z(\omega t)\,\boldsymbol{R}_Y(\phi)\,\boldsymbol{R}_Z(\omega t)\,\vec{\boldsymbol{o}} \quad \text{(A6)}$$

The projected image $\vec{\boldsymbol{i}}$ of the X, Y, and Z components of $\overrightarrow{\boldsymbol{o_{mov}}}$ to the screen at time t is calculated by Equation A7:

$$\vec{\boldsymbol{i}} = \begin{pmatrix} \frac{xf_c}{d_c - y} \\ \frac{zf_c}{d_c - y} \\ t \end{pmatrix} \quad \text{(A7)}$$

where f_c is the focal length of the camera, and d_c is the distance from the camera to the object.

The difference in the equations between rotation and wobbling or rolling are the rotations around the center of the object: $\boldsymbol{R}_Z(-\omega t)\vec{O}$ or $\boldsymbol{R}_Z(\omega t)\vec{O}$. These rotations are not discernible for circular rings, so rotating, wobbling, and rolling circular rings generate the same images on the screen.

Motion energy

To understand the response of the motion-energy mechanism to the rotating rings, we generated arrays of motion-energy filters that were convolved with the video stimulus. Each filter was based on a pair of odd and even symmetric Gabors in quadrature pairs. We first computed the ith pair of Gabor filters at each position and time:

$$\begin{aligned} G_{i,\ odd}(x, y, t) &= \exp\left(-\frac{(x - c_{x,i})^2 + (y - c_{y,i})^2}{2\sigma_{s,i}{}^2} - \frac{(t - c_{t,i})^2}{2\sigma_{t,i}{}^2}\right) \\ &\quad \times \sin\left((x - c_{x,i})\,f_{x,i} + (y - c_{y,i})\,f_{y,i} \right. \\ &\quad \left. + \ (t - c_{t,i})\,f_{t,i}\right) \end{aligned} \quad \text{(A8)}$$

$$\begin{aligned} G_{i,\ even}(x, y, t) &= \exp\left(-\frac{(x - c_{x,i})^2 + (y - c_{y,i})^2}{2w_{s,i}{}^2} - \frac{(t - c_{t,i})^2}{2w_{t,i}{}^2}\right) \\ &\quad \times \cos\left((x - c_{x,i})\,f_{x,i} + (y - c_{y,i})\,f_{y,i} \right. \\ &\quad \left. + \ (t - c_{t,i})\,f_{t,i}\right) \end{aligned} \quad \text{(A9)}$$

where $c_{x,i}$ and $c_{y,i}$ are the center of the filter in space, and $c_{t,i}$ is the center in time; $\sigma_{s,i}$ and $\sigma_{t,i}$ are the spatial and temporal standard deviations of the Gaussian envelopes; and $f_{x,i}$, $f_{y,i}$, and $f_{t,i}$ are the spatial and temporal frequency of the sine component of the Gabor, referred to as the preferred spatial and temporal frequencies. Each filter, G_i, was convolved with the video $I(x, y, t)$, then the responses of the quadrature pair were squared and added to give a phase-independent response, $ME_i(x, y, t)$. At each location (x, y, t), the preferred spatial and temporal frequencies, $\widehat{f_x},\ \widehat{f_y},\ and\ \widehat{f_t}$ of the filter that gave the maximum response $ME_i(x, y, t)$ was picked:

$$\widehat{f_x}, \widehat{f_y}, \widehat{f_t} = \underset{f_{x,i}, f_{y,i}, f_{t,i}}{\operatorname{argmax}}\ (ME_i(x, y, t)) \quad \text{(A10)}$$

Then, the speed, $S(x, y, t)$, and the direction of the velocity, $\theta(x, y, t)$, were calculated as

$$S(x, y, t) = \frac{\widehat{f_t}}{\sqrt{\widehat{f_x}^2 + \widehat{f_y}^2}} \quad \text{(A11)}$$

$$\theta(x, y, t) = \arctan\left(\frac{\widehat{f_y}}{\widehat{f_x}}\right) \quad \text{(A12)}$$

Thus, the vector field, $\boldsymbol{q_{ME}}$, attained from motion-energy units in the horizontal and vertical components of the velocity, $\boldsymbol{q_{ME}}$, will be

$$\boldsymbol{q_{ME}}(x, y, t) = \begin{pmatrix} S(x, y, t) * \cos\theta(x, y, t) \\ S(x, y, t) * \sin\theta(x, y, t) \end{pmatrix} \quad \text{(A13)}$$

Convolutional neural network

The CNN has two convolutional layers and one fully connected layer with the softmax activation function. For the first convolutional layer, each vector field at time t is cross-correlated with 32 filters with the size of 3×3, and a nonlinear rectification is applied:

$$a_i^1 = Max_i\left(ReLU\left(\sum_{c,m,n} w_i^{1,c}[x+m, y+n] \times Q_S[x, y, t|V_S] + b_i^{1,c}\right)\right) \quad \text{(A14)}$$

where c is the horizontal and vertical components of the velocity, w_i^c is the ith weight of 32 filters, b_i^c is the ith bias, $ReLU(x) = \max(x, 0)$ is the rectified linear activation function, and Max_i is the max pooling with the size of 2×2. Then, in the second layer, each output of a_i^1 is cross-correlated with 10 3×3 filters followed by the $ReLU$ function:

$$a_{i,j}^2 = ReLU\left(\sum_{m,n} w_j^2\left[x+m, y+n\right] a_i^1\left[x, y\right] + b_j^2\right) \quad \text{(A15)}$$

Finally, $a_{i,j}^2$ is flattened to be a vector, $\overrightarrow{a^2}$, and the output of the CNN is computed by the fully connected layer with the softmax activation function:

$$f_{CNN}^t = softmax\left(\boldsymbol{W}^3\overrightarrow{\boldsymbol{a}^2} + \boldsymbol{b}^3\right) \quad \text{(A16)}$$

where the *softmax* activation function is calculated by

$$softmax(Z)_i = \frac{e^{z_i}}{\sum e^{z_j}} \quad \text{(A17)}$$

Feature tracking

Our feature-tracking mechanism simulation has two modules. In the first module, at corners and sharp curvatures, the image intensity changes along different directions, and the Harris corner detector exploits this property to extract salient features. First, we computed the change in the intensity value of a part of image by sifting a small image patch in all directions and taking a difference between the patch and the shifted one. Suppose that $I(x, y)$ is the image intensity at the (x, y) position and consider a small image patch (receptive field) with a Gaussian window, $(x, y) \in W$ (where W is a Gaussian window of size 5×5 pixels). If the window is shifted by $(\Delta x, \Delta y)$, the sum of the squared difference (SSD) between two image patches will be

$$\begin{aligned} &E\left(\Delta x, \Delta y\right) \\ &\quad = \sum_{x,y} W\left(x, y\right)\left(I\left(x, y\right) - I\left(x + \Delta x, y + \Delta y\right)\right)^2 \end{aligned}$$

From the first-order Taylor expansion,

$$\begin{aligned} &I\left(x + \Delta x, y + \Delta y\right) \\ &\quad \approx I\left(x, y\right) + I_x\left(\mathrm{x, y}\right)\Delta x + I_y\left(\mathrm{x, y}\right)\Delta y \quad \text{(A18)} \end{aligned}$$

The SSD will be approximated by

$$\begin{aligned} E\left(\Delta x, \Delta y\right) &\approx \sum_{x,y} W\left(x, y\right)\left(I\left(x, y\right) - \left(I\left(x, y\right) \right.\right. \\ &\qquad \left.\left. + I_x\left(\mathrm{x, y}\right)\Delta x + I_y\left(\mathrm{x, y}\right)\Delta y\right)\right)^2 \\ &= \sum_{x,y} W\left(x, y\right)\left(I_x\left(\mathrm{x, y}\right)\Delta x \right. \\ &\qquad \left. + I_y\left(\mathrm{x, y}\right)\Delta y\right)^2 \qquad \text{(A19)} \end{aligned}$$

which can be written in matrix form:

$$\begin{aligned} &= \left(\Delta x \ \Delta y\right)\sum_{x,y} W\left(x, y\right)\begin{pmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \\ &= \left(\Delta x \ \Delta y\right)\boldsymbol{M}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \qquad \text{(A20)} \\ &= \left(\Delta x \ \Delta y\right)\boldsymbol{U\Lambda U^T}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \end{aligned}$$

where $\mathbf{U}$ is an orthonormal matrix containing the two eigenvectors of $\mathbf{M}$, and $\mathbf{\Lambda}$ is a diagonal matrix, where $\mathbf{\Lambda} = \left(\begin{smallmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{smallmatrix}\right)$, with the two eigenvalues. Eigenvalues quantify change in the image intensity values along the eigenvectors, and they differ based on image properties. Figure A1 shows those properties. At a uniform region on the left, there is no change in the image intensity at the receptive field in accordance with the displacement of the receptive field (λ_1 and λ_2 are close to zero). When the receptive field is close to the edge, the image intensity changes in a direction perpendicular to the contour but not along the contour (λ_1 has a large value, but λ_2 is small). However, at the corner, the image intensity differs in all directions (both λ_1 and λ_2 have a large value). We quantified features (corners) by the following equation:

$$\begin{aligned} h &= \lambda_1\lambda_2 - K(\lambda_1 + \lambda_2)^2 \\ &= \det\left(\mathbf{\Lambda}\right) - K\, tr(\Lambda)^2 \quad \text{(A21)} \end{aligned}$$

where tr is the trace of the matrix, and the threshold K is set at 0.05 as a default by MATLAB. We extracted local features if $h > 0.05$. Suppose that the extracted local feature that satisfies $h > 0.05$ is $L_F(cx_t, cy_t)$ where cx_t and cy_t are the center of a 5×5 pixel extracted feature at time t, then this local feature is cross-correlated with the succeeding frame to extract the next location of the

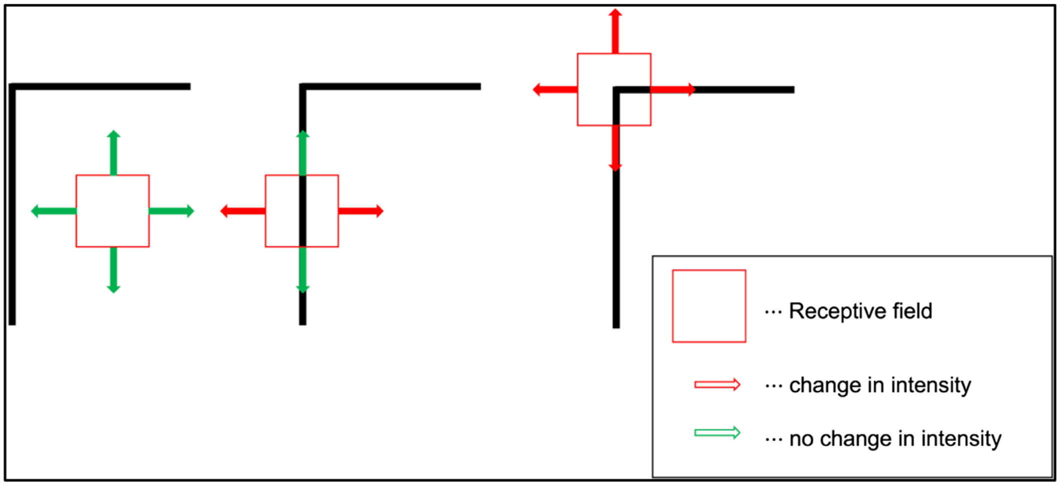

Figure A1. Feature selection and change in the image intensity across different regions. The red square shows a receptive field at a flat region (left), at an edge (middle), and at a corner (right). For the flat region, a small change in the receptive field location does not change the image intensity (shown in green arrows). At the edge, moving along the edge direction does not change the overall image intensity, but, except for that direction, the image intensity shifts, especially along the direction perpendicular to the edge. At the corner, the overall image intensity changes in every direction.

feature:

$$cx_{(t+1)}, cy_{(t+1)} = \underset{x,y}{\operatorname{argmax}} \left(\sum_u \sum_v I\left(u, v, t+1\right) L_F(x+u, y+v) \right) \quad \text{(A22)}$$

Then, the vector fields $\boldsymbol{Q_H}$ are extracted by

$$\boldsymbol{Q_H}\left(cx_t, cy_t, t\right) = \begin{pmatrix} cx_{(t+1)} - cx_t \\ cy_{(t+1)} - cy_t \end{pmatrix} \quad \text{(A23)}$$

The second module in feature tracking combines the outputs from motion-energy units into PDS mechanisms (Figure 7A, top). The input of this unit is the motion-energy vector flow, given by Equations A11 and A12. Considering only the locations (cx_t, cy_t), the center of locations extracted from the Harris corner detector, for each vector, the ambiguity resides along a perpendicular line from the vector, and we assumed the observer's measurement is ambiguous. Thus, the ith measurement distribution given the local velocity, $\vec{\boldsymbol{v}}$, is represented by a Gaussian distribution:

$$p\left(S_i,\ \Theta_i | \vec{\boldsymbol{v}}\left(cx_t, cy_t, t\right)\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left(\sin\Theta_i v_x + \cos\Theta_i v_y - S_i\right)^2\right) \quad \text{(A24)}$$

Then, we combine the observer's measurement with the slowest motion prior, represented as follows:

$$p\left(\vec{\boldsymbol{v}}\right) \propto \exp\left(-\frac{v_x^2 + v_y^2}{2{\sigma_p}^2}\right) \quad \text{(A25)}$$

where $\sigma_p = 2$ (arbitrarily chosen). The posterior probability over a velocity would be computed by using Bayes' rule:

$$p\left(\vec{\boldsymbol{v}} | S_1,\ \Theta_1, \ldots, S_n,\ \Theta_n\right) \propto p\left(\vec{\boldsymbol{v}}\right) p(S_1,\ \Theta_1, \ldots, S_n,\ \Theta_n | \vec{\boldsymbol{v}})$$

By assuming conditional independence, $S_1,\ \Theta_1 \perp S_2,\ \Theta_2\ \ldots \perp S_n,\ \Theta_n | \vec{v}$, the equation becomes:

$$p\left(\vec{\boldsymbol{v}} | S_1,\ \Theta_1, \ldots, S_n,\ \Theta_n\right) \propto p\left(\vec{\boldsymbol{v}}\right) \prod_i^n p(S_i,\ \Theta_i | \vec{\boldsymbol{v}}) \quad \text{(A26)}$$

Then, because the local velocity was estimated by taking the MAP, the velocity field from PDS becomes

$$Q_{PDS}\left(cx_t, cy_t, t\right) = \underset{\vec{v}}{\operatorname{argmax}} \left(p\left(\vec{\boldsymbol{v}}\right) \prod_i^n p(S_i,\ \Theta_i | \vec{\boldsymbol{v}}) \right) \quad \text{(A27)}$$

The vector fields $\boldsymbol{Q_{FT}}$ from these two feature-tracking units are combined by a vector sum:

$$Q_{FT} = Q_{ME} + Q_{PDS} \quad \text{(A28)}$$

Combining motion mechanisms

We linearly combined the motion-energy and feature-tracking vector fields with different combinations of weights that summed to 1.0, and the combined vector fields were fed to the trained CNN to classify the observer's proportion of non-rigidity as a function of different speeds. Suppose Q_S is the combined field and V_S is the speed of the stimulus. Q_S is computed by the weighted sum of two velocity field such that

$$Q_S(x, y, t|V_S) = w(V_S)\,Q_{ME} + (1 - w(V_S))\,Q_{FT} \quad \text{(A29)}$$

where $w(V_S)$ is a weight function that depends on the speed of stimulus. The higher $w(V_S)$ is, the more the vector field gets closer to $\boldsymbol{Q_{ME}}$. Conversely, if $w(V_S)$ is lower, the vector field more resembles the rigid rotation. The likelihood of C_M (classification of motion as wobbling or rotation) is estimated by the trained CNN as a function of Q_S and V_S:

$$L(C_M) = p(Q_S|C_M, V_S, S_h) = \bar{f}_{CNN}(Q_S) \quad \text{(A30)}$$

where $\bar{f}_{CNN}$ is computed by the average of f^t_{CNN} (Equation A16) across all time.

Final model

The first model showed that, to completely explain the variation of the illusion where rigid rings are perceived as non-rigidly connected as a function of speed and shape, other factors have to be considered besides the outputs of the two bottom-up motion mechanisms. In this section, we add prior assumptions to motion mechanism–based CNN classifications for rigid and non-rigid perceptions of the rotating ring pairs. The degree of rotational symmetry may supply not only a prior for rolling but also for wobbling, as a priori a circular ring is more likely to wobble than a square ring. Suppose that $S_h = (n_s/\bar{h})$, where the number of rotational symmetries is n_s and the average strength of the detected corner is $\bar{h}$. The posterior probability of a motion class, C_M, given the vector fields, rotation speed, and object shape is computed by using Bayes' rule:

$$p(C_M|Q_S, V_S, S_h) = \frac{p(C_M)\,p(Q_S, V_S, S_h|C_M)}{p(Q_S, V_S, S_h)}$$

By factorizing the conditional probability:

$$p(C_M|F_S, V_S, S_h) = \frac{p(C_M)\,p(V_S|C_M)p(S_h|C_M, V_S)p(Q_S|C_M, V_S, S_h)}{p(Q_S, V_S, S_h)}$$

$$\begin{aligned} p(C_M|F_S, V_S, S_h) &= \frac{p(C_M)\,p(V_S|C_M)p(S_h|C_M, V_S)\bar{f}_{CNN}(Q_S)}{p(Q_S, V_S, S_h)} \\ &= \frac{p(C_M)\,p(V_S, C_M)\,p(S_h, C_M, V_S)\,\bar{f}_{CNN}(Q_S)}{p(C_M)\,p(C_M, V_S)\,p(Q_S, V_S, S_h)} \\ &= \frac{p(C_M|S_h, V_S)p(S_h, V_S)\,\bar{f}_{CNN}(Q_S)}{p(Q_S|S_h, V_S)p(S_h, V_S)} \\ &= \frac{p(C_M|S_h, V_S)\bar{f}_{CNN}(Q_S)}{p(Q_S|S_h, V_S)} \end{aligned} \quad \text{(A31)}$$

Because the strength of features and the rotational symmetry seemed to be related to the percept of non-rigidity/rolling as shown in the rolling illusion experiment despite that there is no vector field that supports it, the conditional prior $p(C_M|S_h, V_S)$ is estimated by the following equation:

$$p(C_M|S_h, V_S) = \varsigma\left(\vec{W}(V_S)^T S_h + b(V_S)\right) \quad \text{(A32)}$$

where $b(V_S)$ and $\vec{W}(V_S)$ are a 2 × 1 weight vector and bias, both of which are dependent on the speed of the stimulus; ς is a sigmoid function:

$$\varsigma(x) = \frac{1}{1 + e^{-x}} \quad \text{(A33)}$$

Thus, the posterior becomes

$$p(C_M|Q_S, V_S, S_h) = \alpha\varsigma\left(\vec{W}(V_S)^T S_h + b(V_S)\right) \times \bar{f}_{CNN}(Q_S) \quad \text{(A34)}$$

where α is proportional to $p(Q_S|S_h, V_S)$ thus depending on the speed of the rotation; $\vec{W}$, b, and α are estimated by using gradient descent to minimize the MSE.