The perceptual dimensions of natural dynamic flow

Yaniv Morgenstern

Department of Psychology, University of Minnesota, Minneapolis, MN, USA
Department of Psychology, Justus-Liebig-Universität Giessen, Giessen, Germany

Daniel J. Kersten

Department of Psychology, University of Minnesota, Minneapolis, MN, USA

We measured perceptual judgments of category, material attributes, affordances, and similarity to investigate the perceptual dimensions underlying the visual representation of a broad class of natural dynamic flows (sea waves, smoke, and windblown foliage). The dynamic flows were looped 3-s movies windowed with circular apertures of two sizes to manipulate the level of spatial context. In low levels of spatial context (smaller apertures), human observers' judgments of material attributes and affordances were inaccurate, with estimates biased toward assumptions that the flows resulted from objects that were rigid, "pick-up-able," and not penetrable. The similarity arrangements showed dynamic flow clusters based partly on material, but dominated by color appearance. In high levels of spatial context (large apertures), observers reliably estimated material categories and their attributes. The similarity arrangements were based primarily on categories related to external, physical causes. Representational similarity analysis suggests that while shallow dimensions like color sometimes account for inferences of physical causes in the low-context condition, shallow dimensions cannot fully account for these inferences in the high-context condition. For the current broad data set of dynamic flows, the perceptual dimensions that best account for the similarity arrangements in the high-context condition are related to the intermolecular bond strength of a material's underlying physical structure. These arrangements are also best related to affordances that underlie common motor activities. Thus, the visual system appears to use an efficient strategy to resolve flow ambiguity; vision will sometimes rely on local, image-based, statistical properties that can support reliable inference of external physical causes, and other times it uses deeper causal knowledge to interpret and use flow information to the extent that it is useful for everyday action decisions.

Introduction

In daily life, our visual system receives a constant flow of image patterns; an important subset of these arise from the dynamics of material substances, such as sea waves, smoke, foliage, and cloth. Such image patterns are complex functions of intrinsic physical properties (e.g. stiffness), photometric properties that affect how light is scattered and reflected, and the viewing conditions of illumination and relative motion between the viewer and substance. It is remarkable that given the complexity, high rate, and dimensionality of the retinal input (~10 Mbps; Koch et al., 2006), the visual system manages to almost instantaneously transform dynamic flow information into useful decisions and actions, such as deciding whether and how to interact with the underlying substance. These behaviors depend on perceptual inferences that span a range of abstraction in the causes of flows, involving processes that boil down a high-dimensional input to a relatively small set of perceptual dimensions that support useful tasks. The purpose of our study is to identify key perceptual dimensions that are used in comparing natural flows within and across categories, with a view to understanding how these dimensions may support a broader range of tasks.

We can get initial insight into the perceptual dimensions of flow by considering how natural flows are generated, in particular by using information from physics-based computer-graphics research that synthesizes dynamic flows. The generation of realistic flows begins with models of structure that can be described as going from "deep" to "shallow" causes (Figure 1), roughly corresponding to the standard distinction between distal and proximal stimuli. The deepest level models the interaction of external forces, such as gravity and wind, and internal forces, such as a material's surface tension and resistance to stretching,

Citation: Morgenstern, Y., & Kersten, D. J. (2017). The perceptual dimensions of natural dynamic flow. *Journal of Vision*, *17*(12):7, 1–25, doi:10.1167/17.12.7.

doi: 10.1167/17.12.7 Received February 25, 2017; published October 12, 2017 ISSN 1534-7362

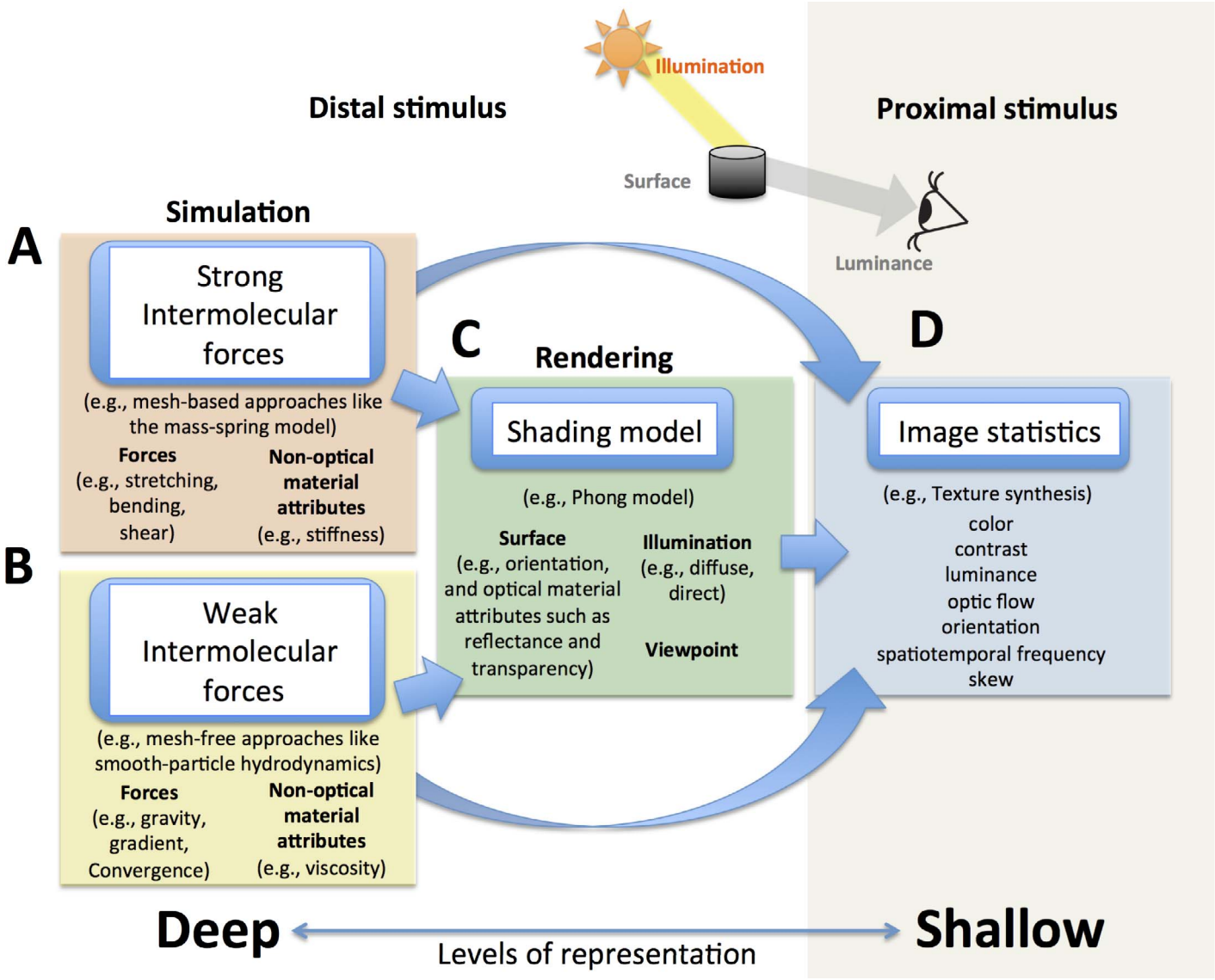

Figure 1. The image generation and perceptual representation of dynamic flows falls along a shallow-to-deep continuum. Dynamic flows can be synthesized using deep or shallow approaches. On the deep end, the movement of dynamic flows that have strong intermolecular bonds—such as cloths, hairs, and other deformable solids—can be simulated with (A) mesh-based approaches (e.g., mass-spring models), while those with weaker intermolecular bonds, such as smoke or fast-moving shallow water, can be simulated with (B) mesh-free approaches (e.g., smoothed-particle hydrodynamics). (C) In the next step of this deep image-synthesis technique, the lighting, viewpoint, shape, and optical properties are rendered through a shading model (e.g., Phong, 1975) that produces an image frame of the flow. (D) On the shallow end of the continuum, the flow can be synthesized based on image features such as color, contrast, spatiotemporal frequency, orientation, and optic flow (e.g., Portilla & Simoncelli, 2000). A perceptual systems representation of dynamic flows can also fall at any point along this continuum. For example, an observer might infer nonoptical properties from (A) or (B), like stiffness and viscosity, with a deep representation—say from the characteristic way the particles move from one frame to the next (A, B)—or a shallower representation—say from the generated shape at the rendering step (short arrows going from (A, B) to (C)); or from local, image-based, statistical properties (long arrows going from (A, B) to (D)). Shallow representations provide an account consistent with the response properties of early-level visual neurons that behave by responding to simple images features. Deep representations require processes that compute material attributes, including nonoptical dimensions. The visual system can use its deep representations to generate actions or affordances, like how to pick up an object, in the world. Computing shallow representations can be quick and efficient; computing deep representations can in theory be complex and demanding (e.g., inferring the position and movement of every one of 150,000 or so strands of hair), but human vision likely relies on computational heuristics.

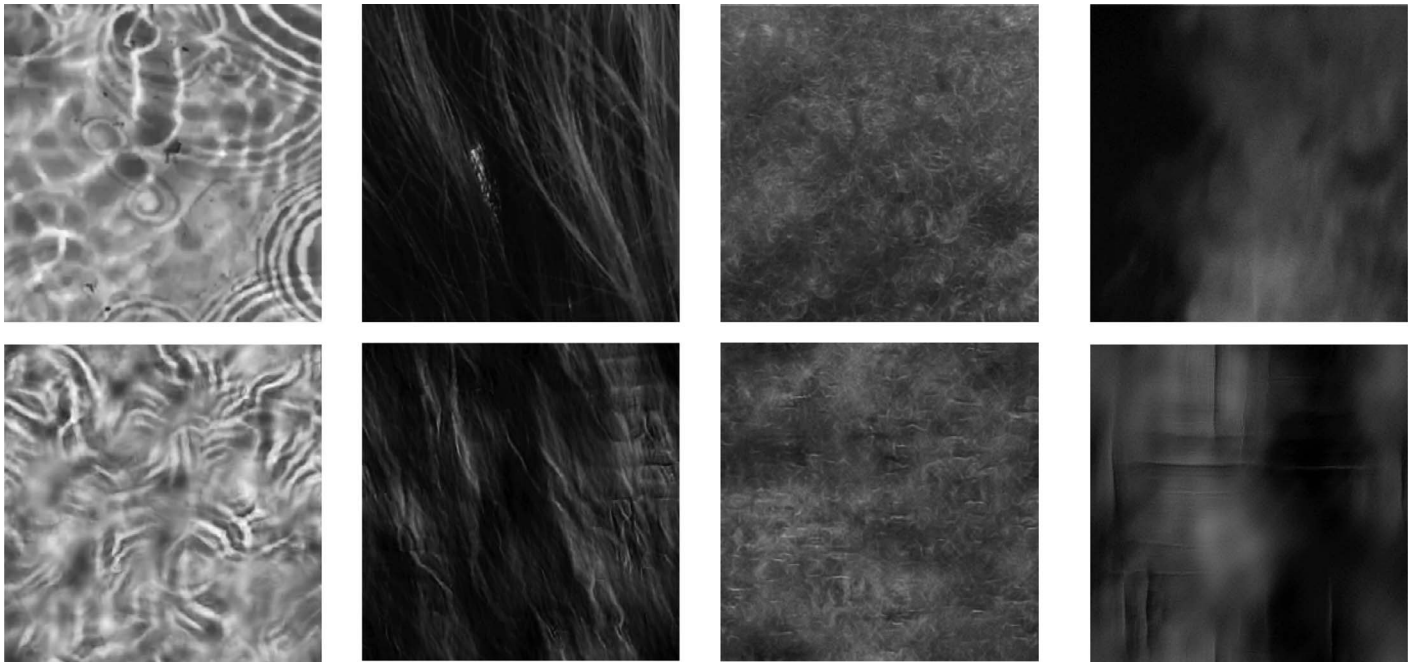

Figure 2. Examples of image synthesis with a shallow approach. Photographs of textures (top row) and their synthesized counterparts (bottom row). The synthesized images were originally white-noise patterns that were transformed with the Portilla–Simoncelli (2000) algorithm to match the spatial statistics of the original photograph. While these textures to a large extent resemble the original textures, they miss the details captured by physics-based approaches, resulting in images that do not always appear how one would expect a photograph of the material to look. Photographs from the last two columns are single frames courtesy of the DynTex movie database (Péteri et al., 2010).

bending, and twisting. These internal forces determine the intermolecular bond strength of the material's underlying physical structure, with solids tending to have stronger intermolecular bonds than liquids and gases. Less deep are midlevel generative photometric processes that depend on configuration, viewpoint, shape, and material properties (e.g., reflectance, specularity, and transparency), followed by even shallower modes of processing that capture the effects of deep and midlevel processes on the resulting image patterns and statistics.

Computationally intensive physics-based approaches are usually necessary to provide convincing realism, suggesting that the visual system is well tuned to the characteristic dimensions, arising from both shallow and deep causes, of many natural flows. However, the human visual system has neither the need nor the computational resources to infer the parameters of a full generative model. For example, it is unnecessarily deep to require vision to keep track of every one of the 150,000 or so strands of hair on a head and how they interact with other strands, gravity, friction, elasticity, and so on (for a review on hair simulation, see Ward et al., 2007). In fact, while some "deep" properties are easily perceived (e.g., the elasticity of silly putty or the relative viscosity of water versus honey), other causal factors are not at all obvious to human perception (e.g., how the folding, coiling, and meandering of a thin thread of honey depend on contact and speed; Bergou, Audoly, Vouga, Wardetzky, & Grinspun, 2010).

One solution to the computational problem of deep causal modeling has been to characterize flows at shallower, descriptive levels of representation, recognizing the fact that human vision can often deal with complex patterns using lower dimensional statistical summaries, as has been exploited in static textures with stationary statistics (e.g., Portilla & Simoncelli, 2000, figure 2). Dynamic flows that exhibit stationarity properties in space and time (e.g., sea waves, smoke, foliage), have been generated by a number of different image-based approaches (e.g., Doretto, Chiuso, Wu, & Soatto, 2003; Kwatra, Schödl, Essa, Turk, & Bobick, 2003; Lizarraga-Morales, Guo, Zhao, Pietikäinen, & Sanchez-Yanez, 2014). However, as with the modeling of static textures, these synthesized flows can miss important perceptual details that are captured by physics-based models, which suggests that human visual inferences of causal parameters span a range of abstraction from shallow to deep depending on the task.

As an example, human vision can predict the path of a liquid, suggesting the use of deep knowledge that approximately simulates physical properties (e.g., Bates, Yildirim, Tenenbaum, & Battaglia, 2015; Kubricht et al., 2016; see Figure 1A and 1B). For other tasks, vision might estimate the deeper nonoptical

Figure 3. The role of spatial context in material perception (see Movie 1). An image of Movie 1 at a small aperature size. At the smallest aperture size, the perceptual dimensions available are shallow. Observers tend to disagree on the flow's underlying material. As the aperture size grows, a sense of fluidity emerges, suggesting that inferring deeper perceptual dimensions requires greater context. At the largest aperture size, observers can use both shallow and deep perceptual dimensions and tend to be in greater agreement as to the flow's underlying material. (Original movie clip courtesy of Thomas Porett).

properties using shallow, but sufficiently diagnostic spatiotemporal image statistics (e.g., Bouman, Xiao, Battaglia, & Freeman, 2013; see Figure 1D). An example is vision's use of shallow knowledge, such as image-based dynamic deformations, to infer optical properties such as transparency (Kawabe, Maruya, & Nishida, 2015).

In summary, vision's representation of images can be posited to fall along a continuum from "shallow" to "deep" (Figure 1). At one end of the continuum is the distal stimulus; at the other end is the proximal stimulus. If vision represents an image by the proximal stimulus, then the same object or material in the world would typically have different representations (due to changes in lighting, viewpoint, etc.). A completely shallow representation may be sufficiently invariant in some cases (e.g., predicting lightness based on contrast in some situations) but would not work in other instances (e.g., recognizing objects or material attributes such as viscosity across changes in lighting, viewpoint, material properties, etc.). If vision represents an image at the distal end of the continuum, then it knows every physical aspect of the object, a completely deep representation. However, it is computationally too expensive for vision to represent every aspect of an object (e.g., inferring the position and movement of every one of 150,000 or so strands of hair). Thus, in general, human vision represents image information along the shallow–deep continuum, at a location determined by the trade-offs between computation and the requirements of the behavioral task.

In the following experiment, we have sought to characterize key dimensions that underlie the perception of dynamic flows. Given a potentially large range of useful dimensions from shallow to deep, we began by measuring human similarity arrangements of dynamic flows over a database of 86 natural flows to determine dominant perceptual dimensions. In order to gain understanding into how comparisons depend on shallower versus deeper causes, flows were viewed through either small or large apertures. The rationale is that interpretability increases with the larger spatial context provided by an increased number of pixels, allowing us to tap perceptual dimensions from shallow to deep causes (Figure 3, Movie 1). For example, we might expect perceptual similarity extracted from small apertures to be based on shallow dimensions, such as color or spatiotemporal statistics, whereas larger apertures provide more information to infer deeper causes of flow patterns, such as nonoptical material attributes (e.g., viscosity, elasticity, rigidity), optical attributes (e.g., transparency), and category labels (e.g., smoke). In anticipation of the importance of color, we also included a larger aperture condition in which color information was removed.

In order to evaluate how well candidate dimensions could explain flow similarities, we posited a set of dimensions from shallow to deep. Shallow dimensions were features based on easily computable image measures, such as color, that are known to be important for image synthesis and perception (Table 1). To aid in identifying higher level or deeper dimensions that are not easily computable from the images, we also asked observers to assign nonoptical material attributes to the flows that are known to be important in computer graphics (Tables 1 and 2) and human vision (e.g., Paulun, Schmidt, van Assen, & Fleming, 2017; van Assen & Fleming, 2016), and to categorize the flows by name. In order to evaluate how well action requirements could explain flow similarities, we also asked observers to rate the flows on affordance properties (e.g., penetrability; Table 2).

Methods

Participants

Observers ($N = 50$) with normal or corrected-to-normal visual acuity were enrolled in the University of Minnesota Research Experience Program to receive extra credit in their undergraduate psychology classes. The observers provided informed written consent under an experimental protocol that was approved by the institutional review board at the University of Minnesota.

Stimuli

Dynamic flows were looped 3-s movie clips. The movie clips were cropped versions of original videos that were captured with an iPhone or adapted from the DynTex database (Péteri, Fazekas, & Huiskes, 2010),

Dimension	Features	Rationale
Shallow		
Color (Color)	Marginal statistics for LAB color space	Color has an important role in low-level vision, but also in higher level vision (e.g., Hansen, Olkkonen, Walter, & Gegenfurtner, 2006). LAB color space is a color-opponent space that approximates human vision. Color is important for synthesizing images that can be recognized (e.g,, fire and smoke cannot be differentiated in grayscale).
Spatiotemporal (XYT)	Marginal statistics, local autocorrelation, cross correlation, and phase correlation	Spatiotemporal image statistics can explain different aspects of material perception (e.g., Bouman, Xiao, Battaglia, & Freeman, 2013; Kawabe, Maruya, & Nishida, 2015; Motoyoshi et al., 2007) and can be used to synthesize textures that appear like their original texture (e.g., Portilla & Simoncelli, 2000).
Optic flow (Flow)	Marginal statistics for speed, overall motion direction, absolute curl, absolute divergence, gradient, and Laplacian	In physics-based computer graphics, the Navier–Stokes equations use the gradient, divergence, and Laplacian of particle systems in 3-D to simulate the flow of liquids and gases. Patterns of optic flow have been used to account for material perception from dynamic flow (e.g., Doerschner et al., 2011; Kawabe, Maruya, Fleming, & Nishida, 2015).
Deep		
Material attributes (A&A)	Compressibility, elasticity, rigidity, and viscosity (based on human attribute estimates)	These attributes are motivated from models in physics-based computer graphics known for producing realistic simulations for deformable solids, liquids, and gases. Some of these attributes are also known to be important in human vision (e.g., Paulun, Schmidt, van Assen, & Fleming, 2017; van Assen & Fleming, 2016).
Category (Category)	Features related to material category: water, nonwater liquids, cloth, solids, plants, other fluids Features related to simulation model: strong, intermediate, and weak intermolecular-bond strength, wind Features related to conceptual theme: human-made, natural, food, water objects (based on human category judgments)	Higher level features could be important for similarity arrangements.

Table 1. The shallow and deep dimensions extracted from the dynamic flows and their relative importance in material perception or image synthesis. We define a perceptual dimension as a feature or a group of features arising from a common quality or theme that represents aspects of the image (shallow dimensions) or physical scene that generated the image (deep dimensions). The shallow dimensions were computed from readily available image-processing tools. The deep dimensions were based on human observers' estimates on attributes and material identity (see the Appendix for further details).

Attribute	Description
Nonoptical material attribute	
Elasticity	Elasticity is the ability of an object or material to return to its normal shape after being stretched or compressed. Materials high on elasticity strain when stretched and quickly return to their original state once the stress is removed (e.g., latex). Materials low on elasticity do not quickly return to their original state (e.g., gum). Please assign a rating from 0 to 100 on the material's elasticity, where 0 is the lowest possible elasticity and 100 is the highest possible elasticity.
Compressibility	Compressibility describes the material's ability to be forced into less space. Consider filling a cylinder with this material. Imagine closing the cylinder with a piston that can move downward in the cylinder. The more compressible the material, the further the piston can move downwards. For example, air is more compressible than oil, so the piston in an oil cylinder will not move much, while the piston in the air cylinder will move downward some distance. Please assign a rating from 0 to 100 on the material's compressibility, where 0 is the lowest possible compressibility and 100 is the highest possible compressibility.
Rigidity	Rigidity is an object's ability to resist being altered by force. Consider dropping an object of that material. A rigid object would show very little squash or deformation when it hits the surface (e.g., a wooden or metal spoon). A flexible object would stretch as it is thrown and squash when it hits the surface (e.g., gum or cloth). Please assign a rating from 0 to 100 on the object's rigidity, where 0 is the lowest possible rigidity and 100 is the highest possible rigidity.
Viscosity	Viscosity is the property of a fluid that resists the force tending to cause the fluid to flow. Consider pouring a fluid made of this material down a tilted surface such as a slide. A fluid high on viscosity will tend to resist flowing down the slide (e.g., honey, motor oil). A fluid low on viscosity will tend to not resist the force of flow down the slide (e.g., water). Please assign a rating from 0 to 100 on the fluid's viscosity, where 0 is the lowest possible viscosity and 100 is the highest possible viscosity.
Affordance	
Penetrability	Penetrability describes the ease of passing through a material. Consider thrusting your fist onto a surface made from this material. The material is highly penetrable if your fist easily pierces the surface (e.g., air) and less penetrable if it harder to pierce the surface (e.g., a brick wall). Please assign a rating from 0 to 100 on the material's penetrability, where 0 is the lowest possible penetrability and 100 is the highest possible penetrability.
Pick-up-ability	Pick-up-ability is the ability to apprehend and also to move an object or material with our hands through grasping or cupping. An object or material high on pick-up-ability would be easy to apprehend and move (e.g., a spoon). An object or material low on pick-up-ability would not be easy to apprehend or move (e.g., air). Please assign a rating from 0 to 100 on the object/material's pick-up-ability, where 0 is the lowest possible pick-up-ability and 100 is the highest possible pick-up-ability.

Table 2. The attributes and descriptions observers used to guide their ratings.

the National Park Service, Shutterstock, YouTube, or Vimeo. The cropped videos captured a small portion of a scene whose foreground showed movement dominated by the flow of a material substance (e.g., hair, foliage, snow, water, wood). The 86 dynamic flows used in these experiments included fluids (such as water, paint, lava, caramel, fire, and smoke), hair, plants, cloths, and other solids. The data set (not including the Shutterstock videos, due to their copyright) can be downloaded at https://sites.google.com/site/yanivmorgenstern/stimuli.

We divided the 86 movies into two data sets as follows. Similar kinds of dynamic flows with only two occurrences within the 86 flows (e.g., fire, lava, and milk bubbles) were divided between the two sets. The remaining flows were assigned randomly, half to Data set 1 and the other half to Data set 2. The type of flows ranged from materials whose particles are held together

by strong intermolecular forces (e.g., textiles, plants) to substances held together with weaker intermolecular forces (e.g., steam, snow, sand; see Figure 1). These flows also fit the two broad categories of simulation techniques, from mesh-based to mesh-free approaches, which are used to create different types of flows in computer graphics. A total of 32 flows in these videos were dynamic textures that were judged by the experimenter (YM) to have local statistics that were stationary in space and time (e.g., water movies from Clips 7, 8, and 16 in Data set 1). The remaining flows were judged to be nonstationary in both space and time (e.g., highly viscous flows, like the honey in Clip 28).

Aperture manipulation

Each observer participated in one of three context manipulations: the small color condition ($n = 16$), the large color condition ($n = 16$), and the large gray condition ($n = 18$). By windowing the movies with a small (diameter = 1.83°, 78 pixels) or a large (diameter = 10.76°, 460 pixels) circular aperture, we varied the level of spatial context for the stimuli in the small and large conditions. By converting the large-context movies into grayscale, we varied the color context for the stimuli in the large color versus the large gray condition.

Procedure

The experimental procedures were run in MATLAB 2013a using the multiarrangement code provided by Kriegeskorte and Mur (2012) and adapted for the Psychophysics Toolbox (Brainard, 1997; Kleiner et al., 2007). The experiments were run on a 27-in. iMac (3.4 GHz Intel Core i7) with a resolution of 2,560 × 1,440 pixels. Observers were seated approximately 57 cm from the screen, at which distance a single pixel subtended 0.0235°. The experiment was completed in three sessions lasting approximately 1 hr each. In the first part of the experiment, participants performed similarity judgments using a multiarrangement method. In the second part, they rated each flow along several perceptual dimensions. In the final part, they were asked to label the flows.

Similarity arrangements

There are a large range of potentially useful perceptual dimensions that observers can use to make material judgments and comparisons (Figure 1). We explored which dimensions are useful by using a multiarrangement method that, within a testing trial, allows observers to arrange the 2-D distances between a subset of dynamic flow stimuli based on perceived dissimilarity. Compared to other approaches, the multiarrangement method quickly acquires judgments reflecting higher dimensional dissimilarity structures by allowing the placement of one item to reveal multiple similarity judgments with other items. In the first trial, the multiarrangement method presents all stimuli (from Data set 1 or 2) as animated icons (scaled versions of the flows) in a circular arrangement around an arena (Supplementary Figure S1 and Movie S1; diameter = 22.38°, 966 pixels). The icons were placed at regular angular intervals in random order. Observers used the drag and drop operations of the computer's mouse to arrange these icons on the computer screen according to their similarity. Specifically, observers were told that the distance between two objects represents their similarity, where similar objects are put close together and dissimilar objects are put further apart. Observers were not explicitly instructed by which similarity criteria to arrange the icons. The subsequent trials showed a subset of the stimuli from Trial 1 based on an algorithm that selects stimuli with a lower dissimilarity signal-to-noise ratio (i.e., stimuli that tend to be placed nearby one another) and also takes into account the trial cost (i.e., the time taken to arrange the subset; Kriegeskorte & Mur, 2012).

On the right of the arena, the dynamic flows for the current and last icon selection were presented at their actual sizes. The observers were instructed to judge the similarity of these proper-size movies since they were shown at the correct resolution. A checkerboard frame highlighted the selected icon and the proper-size dynamic flow (on the right).

Once the arrangements were complete, the observers pressed the Return key to go the next trial. The subsequent trials presented a subset of the dynamic flows from the first trial based on the lift-the-weakest algorithm described by Kriegeskorte and Mur (2012). The arrangements ended after 25 min. had passed. On average, observers completed 12.8 trials, with the final result being pairwise dissimilarities (in terms of distances) for the set of dynamic flows. These dissimilarities were assembled as a representational dissimilarity matrix (RDM), which had height and width corresponding to the number of dynamic flows presented on Trial 1 and was symmetric along the diagonal.

The multiarrangement method was used to acquire observers' arrangements for Data sets 1 and 2 in separate sessions (selected in a random order). After the arrangements, observers were asked to report the strategy they used to arrange objects according to their similarity.

Ratings and identification

We related the similarity arrangements to image features, nonoptical material attributes, affordances, and categorical grouping of the flows into classes having shared characteristics (see Estimating affordances and shallow and deep perceptual dimensions in the Appendix). These nonoptical material attributes and affordances were estimated from human-observer ratings of the dynamic flows. The categorical groupings were based on human-observer dynamic flow identification.

Rating dynamic flows on affordances and material attributes: Observers rated the flows on four material attributes that are related to parameters used for simulating material flows in computer graphics (nonoptical material attributes) but could also be important for guiding our actions or affordances with these flows (Table 2). The other two attributes were motor activities related to a large range of action decisions (affordances). The attributes were rated on a continuous scale from 0 to 100. Observers were given the option of responding "not applicable" if they thought a particular flow could not be rated on that attribute.

The ratings for one subject in the large color experiment and one in the small color experiment were removed because they showed ratings negatively correlated to the remaining subjects, suggesting that they had inverted (or misunderstood) the meaning of the scales.

Dynamic flow identification: Observers were asked to identify flows from a series of options arranged in terms of a hierarchal tree structure (Supplementary Figure S2). Observers were presented with a dynamic flow (on the right of the screen) and asked to click on the label (on the left) that best characterized the texture. At first the labels showed the top layer of the hierarchy, which were general (e.g., *animate* or *inanimate*). The next screen again displayed the flow with a series of options, but instead of labels indicating the top layer of the hierarchy, observers were shown the next more specific layer; for example, if *inanimate* had been clicked on in the first screen, a following screen would show the labels *fiber*, *light emitting*, *liquid*, *particles*, *solid*, *vapor*, *other*, and *go back*, for which the observer would again be prompted to indicate the best suited label. This would continue until the entire tree branch was traversed or the observer clicked on *other*. If observers clicked on *other* the subsequent screen would allow them to type in the dynamic flow category label. Observers were instructed to go as deep into the tree structure as possible before clicking on the *other* option. The *go back* option would present them with the previous screen (i.e., the preceding level on the hierarchy).

The ratings and identification responses were converted into model similarity matrices. The dimensions were evaluated by comparing each model similarity matrix to the human similarity judgments (see Appendix for details).

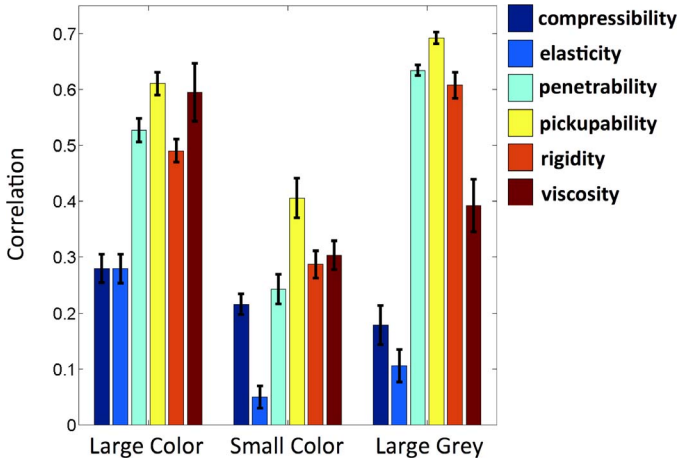

Figure 4. Mean correlations between observers for each rating scale. Error bars show standard errors of the *r* scores. Especially at large apertures, observer ratings of penetrability, pick-upability, rigidity, and viscosity tend to be highly correlated, suggesting that the meanings of these attributes and affordances were interpreted in the same way. Thus, these attributes could be a meaningful way to interpret material identity.

Results

Part I describes the results of the judgments of material attributes, affordances, and flow identification. We show that ratings of attributes tend to become more consistent and distinct across observers under higher levels of context. We then analyze the relationships between these judgments and show that under the highest degrees of flow-identification uncertainty, observers have strong biases towards rigidity. Part II describes the results of the similarity judgments. The results are analyzed in terms of shallow and deep dimensions that contribute to the patterns of similarity.

Part I: Judgments of material attributes and affordances

Are the attributes meaningful?

We begin by analyzing observers' estimates of material attributes and affordances that underlie many important action decisions. For simplicity, we will refer to both material attributes and affordances as *attributes*. One possibility is that observers interpreted the meaning of the attributes differently, resulting in a dynamic flow that has a high value for one subject and a low value for another. On the other hand, if observers tended to agree

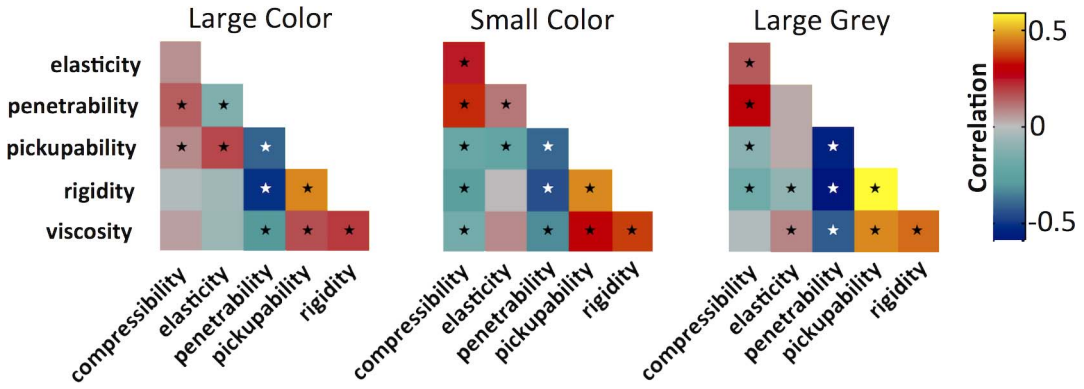

Figure 5. Correlation matrix relating estimates of material attributes and affordances to each other. Colors indicate the correlation coefficient, as specified by the color bar. The correlations between most attributes are highest in the small color and large gray conditions. The attributes appear to decorrelate somewhat at higher contexts, suggesting that they become more distinct. Stars indicate that the correlation in the cell is significantly different from zero ($p < 0.05$).

on the ratings for a given flow, the attributes could be a meaningful way to estimate a material's identity. Figure 4 shows the mean correlation between observers for each rating scale. Consistency across observers tended to be highest for penetrability, pick-up-ability, rigidity, and viscosity. The higher correlations for these scales suggest they may be more meaningful attributes to evaluate material identity for this data set than compressibility and elasticity. Another finding is that the correlation for penetrability, pick-up-ability, and rigidity is higher in the large gray condition than in the large color condition, while the correlation for viscosity decreases. This suggests that under grayscale, observers tend to see flows as arising from stiffer objects. Figure 4 also shows that the correlation for these scales tended to increase when viewing the flows under a larger aperture, suggesting that reliably estimating these attributes requires greater context.

Are the attributes distinct?

To what extent do these attributes reveal different aspects of the material? To some degree, some attributes will correlate with one another. For example, the highest positive correlation is between pick-up-ability and rigidity (Figure 5), and intuitively one would expect these to be highly correlated, because rigid objects are easier to grip and pick up than nonrigid objects. The lowest correlations were for penetrability with pick-up-ability and rigidity (Figure 5). This again makes sense, because things that are penetrable, like water, are harder to pick up and tend to be less rigid. The correlation trends between the attributes are highly correlated across all aperture conditions; the correlations between the attributes across the large color and small color conditions in Figure 5 is 0.76 (Pearson; $p < 0.01$); across large color and large gray, 0.91 (Pearson; $p < 0.01$); and across small color and large gray, 0.94 (Pearson; $p < 0.01$). However, the attributes tend to be least correlated to one another in the large color condition. This indicates that observers can treat the attributes as more distinct in larger contexts.

Attribute estimates are biased for highly ambiguous dynamic flows

In order to recognize materials, the visual system reduces its high-dimensional input into a smaller dimensional summary. Sometimes vision will only need to rely on local, image-based properties; for example, fire has a distinctive pattern of colors that flows upward. Other times vision will need deeper knowledge, for example, to decide whether textiles with similar optical properties are made from silk or hemp. Local image properties are available at small apertures, while nonoptical properties, such as the attributes in Table 2, are better evaluated at larger contexts (Figures 4 and 5).

In this section, we examine the relationship between attribute estimation and material identification as a function of context. We begin by evaluating percent correct identification of the flows by setting the mode of the label responses in the large color condition as the ground-truth labels for the dynamic flows (Figure 6A). (Note that in this evaluation of percent correct we do not take into account semantic similarities within the labels—that is, responses such as *lake* and *river* were treated as different labels in scoring percent correct.) Thus, percent correct reflects consistency across observers for the ground-truth label. Consistency was highest for the large-aperture conditions.

Does the poorer label-identification performance in the small color condition correspond with an inability to reliably estimate material attributes? To get a handle on this question, we separately analyzed the attribute

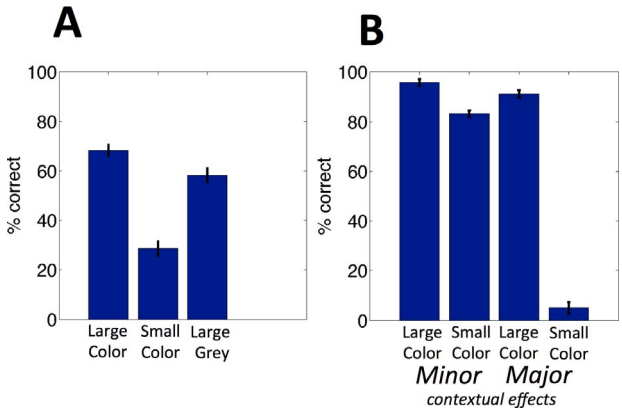

Figure 6. Flow identification. Percent correct dynamic flow identification across observers for (A) the 86 stimuli in the small color, large color, and large gray conditions, and (B) two subsets of the stimuli in (A) that are grouped depending on whether the contextual effects are large (major contextual effects) or small (minor contextual effects). The mode of observer responses for the large color condition were taken as ground truth. The accuracy does not take into account semantic similarities within the labels (i.e., responses such as *lake* and *river* were treated as different labels when scoring percent correct). The standard errors were based on the standard deviation of the samples within each group. Observers are generally better at identifying flows in the larger aperture conditions. However, for a subset of the experimental stimuli, observers did not require more context to identify some flows (e.g., minor-contextual-effects group).

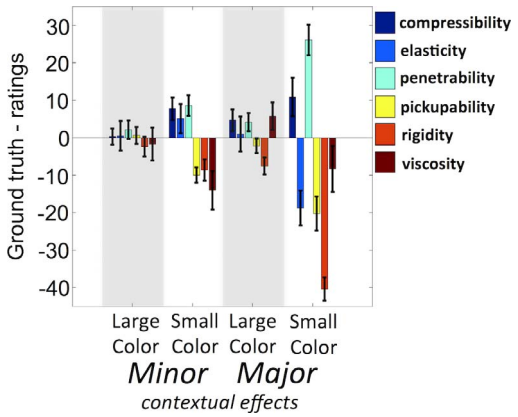

Figure 7. Estimating material attributes and affordances. The mean ratings relative to ground truth for stimuli from the minor- and major-contextual-effects subsets. The ground-truth response for each stimulus was taken to be the median attribute response across observers in the large color condition. The standard errors were based on 100 bootstrapped samples. Under the highest degrees of ambiguity (stimuli from the major-contextual-effects subset of the small color condition), observers tend to have strong biases towards more rigid and pick-up-able, and less penetrable, objects.

estimates and labeled responses from two subsets of the 86 stimuli. The first subset consists of six dynamic flows that were identified with high consistency with respect to ground truth across observers for the large color condition (>80% accuracy) and high consistency for the small color condition (>85% accuracy). We call this subset of stimuli the minor-contextual-effects group. For these flows, the additional spatial context provided by the larger aperture did not help observers much in assigning labels. The second subset consists of five flows with high consistency relative to ground truth for the large color condition (>80% accuracy) and low consistency for the small color condition (<15% accuracy). We call this subset of stimuli the major-contextual-effects group. For these flows, the additional context substantially improved flow identification (Figure 6B). Supplementary Figures S3 and S4 show the label responses for the major- and minor-contextual-effects groups, respectively.

The mean attribute responses for these two subgroups relative to ground-truth responses (the median attribute response across observers in the large color condition for a particular flow) are shown in Figure 7. In the minor-contextual-effects group, observer responses hover around ground truth for most attributes across all conditions. In the major-contextual-effects group, observer responses tend to hover around ground truth for the large color condition but are biased in the small color condition. The biases along the most meaningful attributes (those that are rated most consistently across observers; Figure 4) show higher rigidity and pick-up-ability estimates than ground truth and lower penetrability estimates. This pattern is also evident in the correlation trends between these three attributes: positive correlation between rigidity and pick-up-ability and negative correlation between penetrability and both rigidity and pick-up-ability (Figure 5). These biases are consistent with a prior on rigid objects (Grzywacz & Hildreth, 1987; Ullman, 1979). A closer look at the attribute scores in the minor-contextual-effects subgroup shows that the small color estimates, which are near ground truth, are also biased toward the direction of this prior. (Note that elasticity shows a bias in the opposite direction, but elasticity ratings are also less reliable across observers; see Figure 4.) In other words, the small color estimates in the minor-contextual-effects group are slightly more rigid and pick-up-able and less penetrable than ground truth. That these biases appear in response to the most ambiguous stimuli (i.e., the major-contextual-effects subgroup) and sway estimates in less ambiguous scenes show that they are important attributes and assumptions in human vision.

One important way that the visual system may use these assumptions is to integrate them with other perceptual dimensions (e.g., color, orientation) to identify material labels. Qualitatively, some of the

errors in labeling could be due to biases that substances are more rigid and less penetrable than their ground-truth labels. In the small color condition for minor-contextual-effects stimuli, for example, the mistakes in identifying *fire* are *lava*, *sun*, and *torch*, which may be interpreted as substances that are stiffer than fire (e.g., if one interprets the sun as a celestial body and a torch as being composed of wood; see Supplementary Figure S3). As another example, the cilia of aquatic plants are confused for more rigid body parts. Similarly, for the major-contextual-effects subgroup, in the small color condition *caramel* (last row in Supplementary Figure S4) is identified as many other substances with a similar tan color, such as *honey*, *sand*, and *dust*, suggesting that shallow dimensions such as color can sometimes dominate flow recognition. Some of these labels are also more rigid than *caramel* (e.g., *copper*, *wood*, *torch*, and *leather*), suggesting that the rigidity prior may play a role in these decisions.

Part II: The role of affordances and shallow and deep perceptual dimensions in dynamic flow similarity perception

In the following, we explore the role of affordances and shallow and deep perceptual dimensions on human similarity judgments of dynamic flow.

Similarity arrangements

Figure 8 shows the dissimilarities (in terms of distances) assembled as an RDM (on the right), with height and width corresponding to the number of stimuli in the data set and symmetric along the diagonal. The positions of the stimuli along the RDM were arranged by the experimenter to approximately reflect the intermolecular bond strength of the dynamic flows' underlying materials, with weak intermolecular forces (i.e., particle-based flows like steam, snow, and sand) on one side of the continuum and strong intermolecular forces (e.g., cloths and solids) on the other side. This ordering to some degree also approximately arranges the stimuli such that similar materials are positioned close to each other. In the large color condition, the RDM pooled across observers ($n = 16$) shows that flows more similar to each other in terms of their material category and the strength of their intermolecular forces have smaller dissimilarities (Figure 8A; blue regions). On the other hand, the most dissimilar flows (in yellow) tended to come from stimulus pairs whose difference in intermolecular force strength was large. This result can be visualized by arranging the stimuli in two dimensions using MDS (multidimensional scaling) such that the pairwise distances approximately reflect the distances in the RDM (left image of Figure 8A). These results were consistent with observers' subjective reports: Most observers tended to rely primarily on higher level categories (e.g., object, material, man-made, plants, foods) and secondarily on appearance, grouping flows based on material attributes (e.g., viscosity, fluffiness) and shallow features (e.g., color, motion, spatial frequency). Similar results were found with the 43 dynamic flows from Data set 2 (Supplementary Figure S5A) and in grayscale (Supplementary Figure S6).

In the small color condition, the RDM pooled across observers ($n = 16$) still showed some perseveration of grouping based on material category (e.g., blue regions in Figure 8B's RDM for water and other liquids), but this was much less prevalent than in the large-aperture conditions. The small color RDM also did not tend to have its largest dissimilarities depend on the disparity between intermolecular bond strength. The similarity arrangements visualized by using MDS (Figure 8B, left) show that groupings were based primarily on color. In their subjective reports, most observers stated using primarily color and motion, and secondarily category. Similar results were found under small apertures with the 43 dynamic flows from Data set 2 (Supplementary Figure S5A).

Evaluating shallow and deep perceptual dimensions

We compared these arrangements to dimensions along the shallow-to-deep continuum. The dimensions toward the shallow end were related to the response properties of neurons in the early visual system, which included statistical summaries from color (from LAB color space), multiple spatial and temporal scales (based on a multiscale pyramid decomposition), and optic flow (e.g., magnitude, curl, divergence, gradient; Table 1). These shallow dimensions were extracted from the dynamic flows with freely available image-processing tools (Portilla & Simoncelli, 2000; Sun, Roth, & Black, 2010). The deeper dimensions included nonoptical material properties, affordances, and categories (Tables 1 and 2). We used human-observer responses (see Methods) to estimate these dimensions, since reliable machine-vision methods do not exist or were not readily available. This led to five types of perceptual dimensions: color (Color), spatiotemporal (XYT), optic flow (Flow), the affordances and non-optical material attributes (A&A) listed in Table 2, and categories (Category), which are categorical groupings of the flows into classes having shared characteristics and based on observers' flow-identification responses (see Estimating perceptual dimensions and affordances in the Appendix for further details).

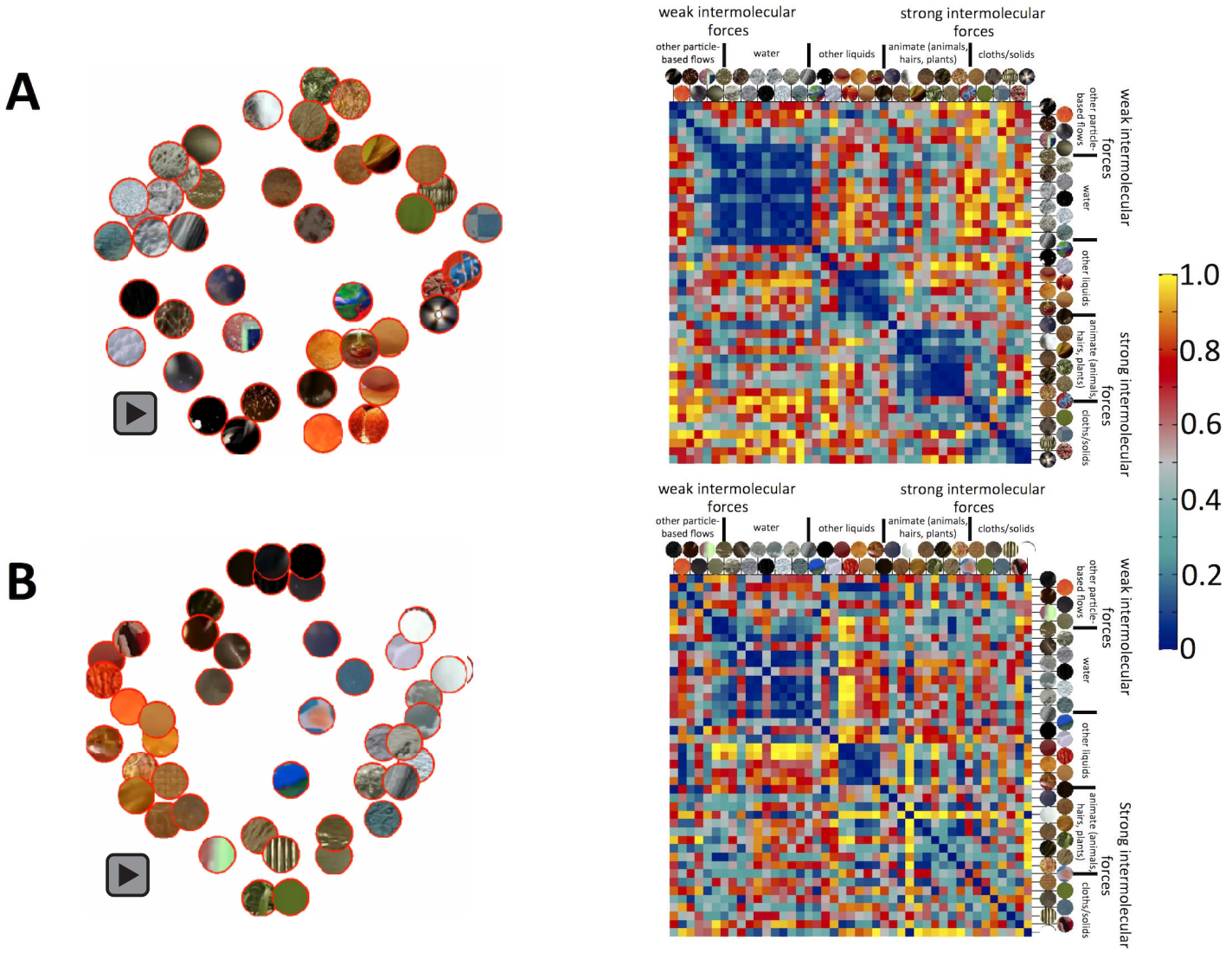

Figure 8. Stimulus arrangements and representational dissimilarity matrices (RDMs) for the pooled data in the (A) large color and (B) small color conditions for Data set 1. For each pair of stimuli, each RDM (right) color-codes the dissimilarity. The experimental stimuli have been arranged (on the left) such that their pairwise distances approximately reflect the distances in the RDM (multidimensional scaling; dissimilarity: distances, criterion: metric stress). (A) is associated with Movie 2A, and (B) with Movie 2B. In each arrangement, dynamic flows placed close together were also arranged this way in the experiment. The correlations between the high-dimensional RDMs and the two-dimensional Euclidean distances in the figure are 0.83 (Pearson) and 0.84 (Spearman) for the small-aperture condition and 0.78 (Pearson and Spearman) for the large-aperture condition, suggesting that the 2-D visualization (on the left) captures much of the variance. The RDMs are separately rank-transformed and scaled into [0, 1]. (See Supplementary Figure S5 for Data set 2 and Supplementary Figure S6 for the large gray condition.) In (A) the large-aperture conditions, similar stimuli (depicted in blue on RDM) tended to come from the same material category and to have similar strength in their intermolecular forces. Dissimilar stimuli in (A)—depicted in yellow on the RDM—tended to come from stimuli with contrasting strength in their intermolecular forces. In (B) the small color condition, the RDMs show that stimuli within the same category are sometimes similar (in blue). However, large disparities between the strength of the intermolecular forces do not tend to lead to the strongest dissimilates. The MDS visualization on the left shows that color dominates the small color arrangement.

Each type of perceptual dimension could consist of multiple features (e.g., the Category dimension type consisted of 14 features; see Table 1), so we combined the features belonging to a perceptual dimension by regressing them to the pooled human similarity arrangement. We also used regression to combine the shallow perceptual dimensions (i.e., the features in Color, XYT, and Flow were combined and called Shallow), deeper perceptual dimensions (i.e., the features in A&A and Category were combined and called Deep), and all features (i.e., the features from every perceptual dimension were combined and called

All). We compared the fitted models to the human similarity arrangements using representational similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008; for further details, see Representational similarity analysis in the Appendix). This evaluation was done separately for the large color (Figure 9A and Supplementary Figure S7A), small color (Figure 9B and Supplementary Figure S7B), and large gray (Supplementary Figure S8) conditions.

Shallow dimensions best account for the small-aperture condition

The role of shallow and deep dimensions in the small color condition is unclear from the attribute estimates and label responses. In the small color condition, observers' attribute estimates are less reliable (Figure 4) and biased (Figure 7), and their ability to recognize materials greatly diminishes (Figure 6A). However, the small color similarity arrangements show some material grouping (e.g., water in Figure 8B), and sometimes identification is accurate (minor-contextual-effects group, Figure 6B). How much do observers rely on shallow and deep dimensions in their arrangements? To explore what dimensions observers rely on under these conditions, we compared several model RDMs (Kriegeskorte et al., 2008) to the human similarity judgments (Figure 9B). The performance of the fitted Color model (Color), Shallow model (Shallow), and All model (All) approaches the noise ceiling, suggesting that these models almost fully explain the similarity judgments. The pairwise model comparisons show that these three models outperform models representing deep dimensions. This finding suggests that shallow models, in particular the Color dimension, can explain variance in the similarity judgments that deeper models cannot explain. This is consistent with observers' subjective reports and the similarity arrangements visualized with MDS (Figure 8B). The fact that the performance of the Color model approaches the noise ceiling indicates that there is not much room for model improvement. This means that the single-subject similarity judgments do not seem more similar to each other than to the color model. The Category model, on the other hand, was far from the noise ceiling and not significantly correlated to the human arrangements. Thus, one possible explanation for material categorization and accurate identification of some flows at smaller apertures is observers' reliance on shallow dimensions for inferences as to material identity.

Deeper dimensions best account for large-aperture conditions

In the large color condition, observers' attribute and affordance estimates (Figure 4) become more consistent, as do their inferences on material category (Figure 6). How important are these attributes and category labels in determining flow similarity? In the large color condition, the model RDMs consisting of features arising from deeper dimensions (Category, Deep, and All) approached the noise ceiling (Figure 9A); the Deep model is not much different from the All model, while the Shallow model is significantly different. However, most model RDMs consisting of deeper features do not significantly outperform the models containing the shallow features (Shallow, Color, XYT, and Flow). This suggests that, while features along the deeper end of the continuum better explain the similarity judgments in the larger aperture condition, the shallow features also play a role.

What are the important deep perceptual dimensions?

The shallow dimensions evaluated here are sufficient for textures synthesis (Figure 2; Portilla & Simoncelli, 2000) and have been used to account for a range of perceptual phenomena (e.g., Doerschner et al., 2011; Kawabe, Maruya, & Nishida, 2015; Motoyoshi, Nishida, Sharan, & Adelson, 2007). The shallow dimensions have a significant role in the large-aperture arrangements, for both data sets in the large color and large gray conditions, but they do not account for as much variance in human similarity arrangements as deep dimensions. The Shallow models are sometimes significantly less predictive of the human similarity data than the Deep models (Data set 1 in the large gray condition; Supplementary Figure S8A) and always significantly less predictive than the All models, while the Deep models are never significantly different from the All models. These findings suggest that the perception of dynamic flow similarity is deeper than can be accounted for by shallow explanations that are evaluated here.

Intuitively, in the large-aperture condition it makes sense that deep dimensions provide better accounts of the similarity arrangements. The human similarity arrangements show higher level groupings, such as ones based on food products that do not have many shallow dimensions in common. The large-aperture similarity arrangements visualized by MDS show the flows arranged into several of these groups or clusters. Careful inspection of nearby flows shows their tendency to have similarities within shallow dimensions (e.g., color, motion), which shows that shallow features do play a role in the large-aperture similarity arrangements. Those flows that are farther apart have very little in common in terms of shallow dimensions and reveal an overarching principle for the global arrangements.

To get an idea of the principles that underlie the global arrangements, we explored how well each feature in the Category dimension accounts for the

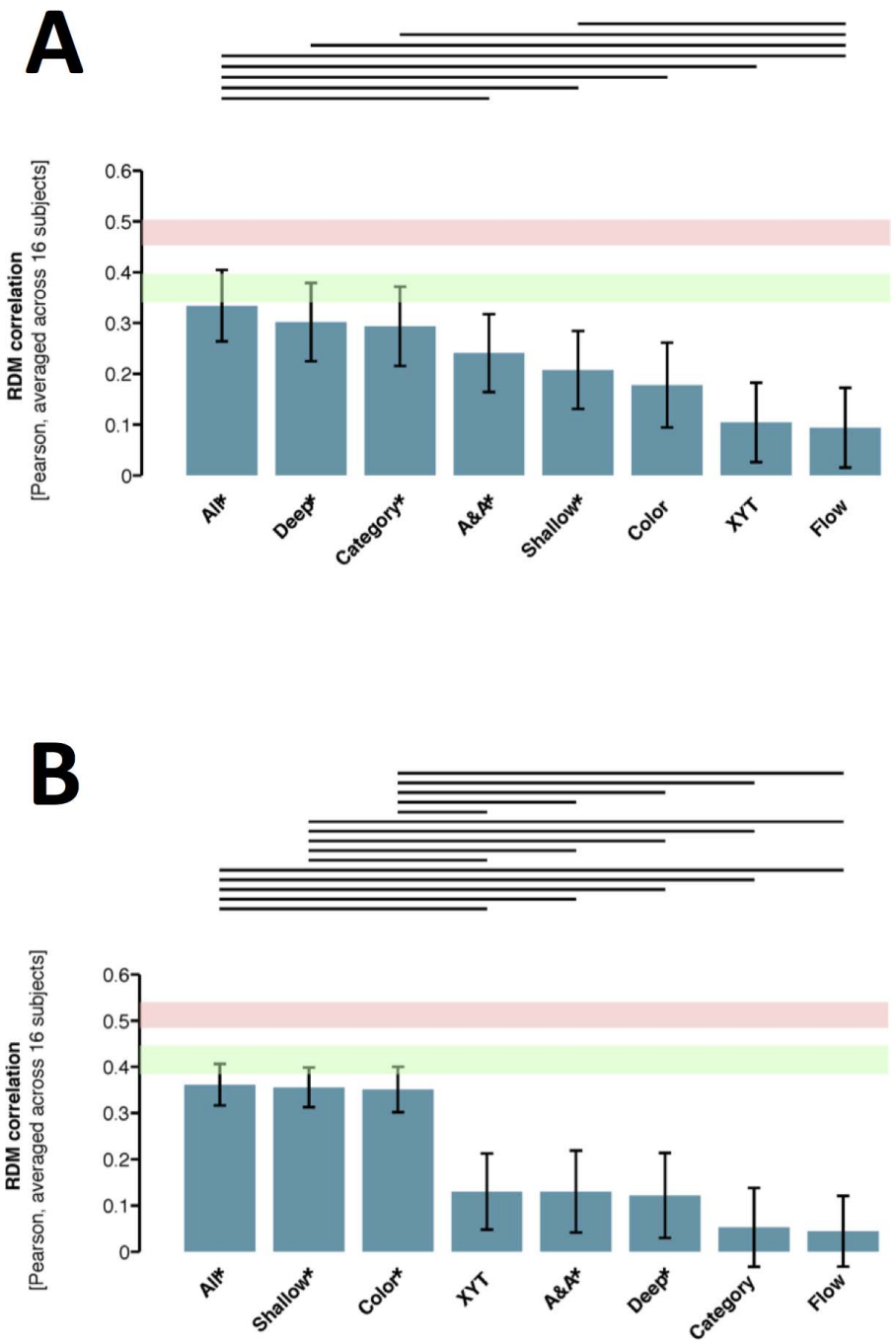

Figure 9. Model performance for similarity judgments in the (A) large and (B) small color conditions for Data set 1. The deep dimensions tend to explain the similarity arrangements better in the large-aperture condition, while the shallow dimensions, in particular color, tend to better account for the small-aperture similarity arrangements. The bar graphs show the correlations between the similarity-judgment RDM and each of the feature (or model-prediction) RDMs. Significant correlations between a feature RDM and the similarity-judgment RDM are indicated by an asterisk (stimulus-label randomization test, $p < 0.05$ corrected for family-wise error). Significant differences between models in how well they can account for the similarity judgments are indicated by the black horizontal lines plotted above the bars (stimulus-bootstrap test, $p < 0.05$ corrected for family-wise error). Error bars show the standard error of the mean based on bootstrap resampling of the stimulus set. The noise ceiling, indicated by the red and green horizontal bars, is the expected RDM correlation achieved by the (unknown) true model, given the noise in the data. The red bar represents the high noise ceiling, calculated by taking the correlation between each subject's RDM and the average of all subject RDMs. The green bar represent the low noise ceiling, calculated by taking the correlation between each subject's RDM and the average of the RDMs belonging to the remaining subjects. The noise-ceiling bars are centered on their mean (computed across subjects) with a width that corresponds to their standard error. All models are based on a weighted combination of features. Similar results were found for Data set 2 (see Supplementary Figure S7) and the large gray condition (Supplementary Figure S8) .

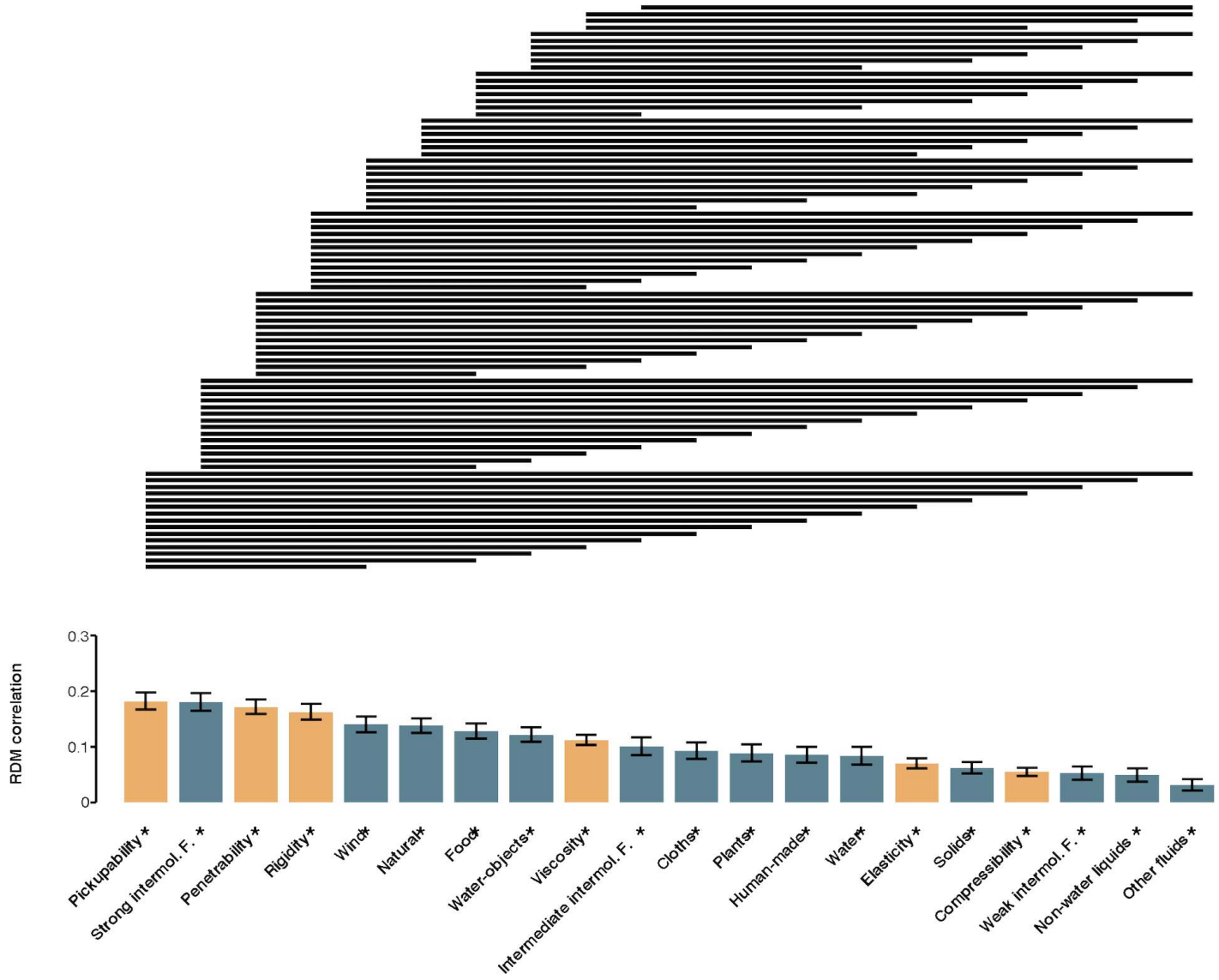

Figure 10. Relationship of category features to human similarity arrangements. The bar graphs show the correlations between the similarity-judgment RDM and each of the category-dimension RDMs (in blue) and action and attribute RDMs (in orange) for data pooled across the large-aperture conditions (both Data sets 1 and 2 from the large color and large gray conditions). The categorical features (in blue) best related to human similarity are those that indicate material properties relevant to simulation, such as the strength of the material's intermolecular bonds or wind. The affordances (in orange) are best related to human similarity, suggesting that our everyday action decisions underlie the degree of the visual system's representation of dynamic flow. Significant correlations between a feature RDM and the similarity-judgment RDM are indicated by an asterisk (signed-rank test, subject as random effect, $p < 0.05$). Significant differences between models in how well they can account for the similarity judgments are indicated by the black horizontal lines plotted above the bars (subject bootstrap test, $p < 0.05$ corrected for family-wise error). Error bars show the standard error of the mean based on human-model correlations across subjects. Supplementary Figures S9 and S10 show similar correlation trends across data sets and conditions for the category features and attributes and affordances, respectively.

similarity arrangements. Each feature divides the flows that belong to a class that have shared characteristics from those that do not. Thus, features from the Category dimension with higher correlations with human similarity provide better explanations of the arrangements as two clusters of flows than features with lower similarity. Thus, these features hint at a more general or global arrangement strategy.

Figure 10 shows the relationship between each feature from the Deep dimension and the human similarity data. The features from the Category dimension (in blue) that are most related to the large-aperture arrangements (pooled over data set and color conditions) tend to come from physical properties related to simulation (e.g., whether flows tend to have strong intermolecular bonds or not) or other higher

level conceptual categories (e.g., whether the flow is natural or not). Strong intermolecular bonds best account for the human arrangements in the large-aperture condition (leftmost feature in Figure 10 in blue), and this is true across data sets and conditions (Supplementary Figure S9). These results suggest that flows with similar physical consistency (i.e., whether their intermolecular bond strength is strong or weak), whether or not the material category is the same, tend to be grouped closer together than flows with dissimilar consistency. Thus, flow similarity is guided by perceptual information that is physically deeper than optical material properties (Figure 1).

Affordances can guide depth of the perceptual representation

The depth of the visual system's representation of dynamic flow should be sufficient to carry out its basic functions (i.e., to gain information about the physical world that is useful for navigating, recognizing objects, and planning future actions). The deep dimensions that best account for the similarity arrangements of a broad set of natural dynamic flows were related to the strength of a flow's underlying intermolecular forces. If this depth is necessary for useful action decisions, then there should be affordances that are about equally related to the human similarity arrangements. Figure 10 shows that affordances (penetrability, and pick-up-ability in orange) are about equally correlated with the human similarity arrangements as categorical groupings based on whether the flow's intermolecular bond strength was strong or not. Thus, material inferences based on deep features can reveal information relevant to common motor decisions. On the other hand, inferences based on shallow features, like color, are poorly related to motor decisions (Figure 9B; low correlation between A&A dimension and human similarity in the small color condition). Moreover, in the small color condition, observers' affordances tend to be biased relative to ground truth (Figure 7), further suggesting that shallow features do not provide the necessary visual information relevant for many action decisions. These results are consistent with the idea that the depth of the visual system's inferences on the physical causes of images is guided by important action decisions.

Discussion

Previous studies on the visual perception of optical material properties have accounted for visual phenomena with representations that spanned from shallow (e.g., gloss perception: Marlow, Kim, & Anderson, 2012; Motoyoshi et al. 2007; distinguishing shiny versus matte: Doerschner et al., 2011) to deep (e.g., lightness perception: Brainard & Maloney, 2011; Knill & Kersten, 1991). Previous works on the perception of nonoptical material properties has relied more heavily on shallow representations (e.g., viscosity: Kawabe, Maruya, Fleming, & Nishida, 2015; elasticity: Kawabe & Nishida, 2016; stiffness: Bi & Xiao, 2016), unless an observer is made to simulate a future outcome, such as predicting the path of liquid flow (Bates et al., 2015; Kubricht et al., 2016), rather than estimate the material property itself.

The contribution of the ratings and similarity analysis as a function of small and large apertures was to understand the relative roles of shallow versus deep representations in perceiving and comparing flows. We show that shallow dimensions dominate similarity arrangements in the small-aperture condition, when flows are highly ambiguous as to the external physical causes. In these conditions, our experiments suggest that the visual system cannot reliably estimate the strength of the underlying flow's intermolecular forces; instead, observers' responses are consistent with a built-in, prior, assumption that the flow's underlying physical structure is strong and rigid, and use color to aid identification. Shallow dimensions also play a role in the large-aperture arrangements, but not as much as deeper dimensions that partially reveal the strength of flow's intermolecular forces. These deeper dimensions are about as highly correlated with the human similarity arrangements as some affordances, such as pick-up-ability and penetrability, suggesting that inferences on deep dimensions are necessary to guide our everyday action decisions. That these affordances cannot reliably be inferred from the flow in the small-aperture conditions (observers instead use built-in prior assumptions; see Figure 7) suggests that, overall, shallow dimensions are used when they provide sufficient information to complete the task at hand. When shallow representations are insufficient for a task or action, the visual system relies on representations of the deeper underlying physical causes of the images.

The computation of deeper representations may require spatial and temporal inhomogeneity

Dynamic spatial textures, such as sea waves, have stationary statistics—that is, they can be characterized by local statistics that change little over space and time. By taking advantage of these stationarity properties, like various texture-synthesis algorithms (e.g., Portilla & Simoncelli, 2000), the visual system can represent a texture in terms of local statistics (Figure 2). In the small-aperture condition, consistent with using statio-

narity properties for visual representation, shallow perceptual dimensions better describe perceptual judgments of similarity. These shallow dimensions are sometimes sufficient to group flows that are based on similar categories, such as whether they are flowing water or windblown foliage (Figure 8 and Supplementary Figure S5).

In the large-aperture condition, many dynamic flows in the data set become inhomogeneous in their pixel statistics across space and time. These nonstationary image sequences (e.g., viscous fluids, hair or a flag blown by wind) have important departures from homogeneity, and our analysis suggests that such departures provide contextual cues that vision can use to infer aspects of the scene, such as the strength of a material's underlying intermolecular forces. This also suggests that information represented in shallow dimensions is inadequate for inferring many aspects of the world. Consistent with this idea, Figure 2 shows that synthesized textures qualitatively appear more rigid than their original texture, suggesting that deeper dimensions are needed to adequately model intermolecular surface structure.

The role of shallow perceptual dimensions in the large-aperture condition

Our stimulus set contained flows that spanned the entire spectrum of physical causes at the deepest level, from flows that consist of weak to strong intermolecular forces (Figure 1), and the RDM analysis revealed shallow and deep dimensions consistent with observers' comparisons across Data sets 1 and 2 and in color and grayscale. With the present data set, however, the shallow dimensions could not fully account for human similarity arrangements in the large-aperture conditions, possibly because dynamic flows varied in many dimensions (intermolecular forces, other nonoptical and optical material properties, viewpoint, illumination, orientation, etc.), causing vastly different image sequences. Thus, the shallow features between different flows varied greatly, causing observers to sometimes rely on one shallow feature, such as color, to group Stimulus A next to Stimulus B but then use another shallow feature from Stimulus A, such as motion direction, to group it next to Stimulus C. This led to an arrangement in the large-aperture condition that was globally dominated by the deepest causal features (Figure 1; intermolecular force) rather than by different kinds of shallow features that were used primarily for local arrangements.

The shallow dimensions evaluated in the present study were hand-selected features that are commonly used in explaining visual phenomena (see Table 1), produce synthetic images that appear natural (e.g., see Figure 2), and are related to the response properties of neurons in the early visual cortex. One could imagine there exist more complex image features that better account for the human similarity measurements or other aspects of the distal stimulus. Rather than hand selecting these features, an alternative method of feature selection is machine learning. For example, convolutional neural networks could be used to learn features for a task, such as flow recognition, producing a larger set of features, some of which may be more reflective of the human similarity arrangements or other aspects of the distal stimulus (e.g., Bell, Upchurch, Snavely, & Bala, 2015).

The role of deep perceptual dimensions in the large-aperture condition

There are at least several reasons why intuitively important deep features (e.g., elasticity and compressibility) were not rated consistently across observers (Figures 4 and 10). One possibility is that the duration needed to estimate these features was not adequate. For example, Kawabe And Nishida (2016) have shown that increasing the simulated movie frame duration (the time of a single movie frame) of a falling cube from 33 to 266 ms also increases the impression of its elasticity. Another possibility is that looping the video clip acted like temporal noise that introduced oscillatory behaviors inconsistent with the normal elastic motion of the objects, leading to noisy and inconsistent ratings on some of the attributes. Another possibility is that the chosen stimuli did not exhaust the perceptual range over those dimensions. To further explore the role of shallow dimensions in inferences of physical causes and also determine the importance of other relevant deep features (like elasticity, surface reflectance, and illumination), future work can use a larger stimulus set across a smaller subset of dynamic flows (e.g., flowing water or windblown textiles; Bi & Xiao, 2016). Focusing on more specific stimulus sets based on intermediate or weak intermolecular bonds may reveal the importance of other attributes, such as elasticity, and the actions decisions that guide inferences on these attributes.

The present analysis is missing several deep features known to be important in human perception, including optical material properties (e.g., reflectance: Brainard & Maloney, 2011; Knill & Kersten, 1991). However, some of the model RDMs are close enough to the noise ceiling (Figure 9) to suggest that adding additional deep features, such as those based on optical material properties, will have a minor overall effect. Furthermore, given that flows in the present data set consist of materials with a broad set of reflectance values (e.g., textiles), global similarity groupings based on optical material properties is expected to be minimal.

The depth of visual representation depends on task

In reality, vision relies on both shallow and deep representations of images, and this will depend on task. Consider, for example, the image of a highly localized edge. Shallow representations will be sufficient if local information unambiguously signals what is needed to successfully complete the task (e.g., edge orientation or color). They will not, however, be sufficient for some tasks. For example, imagining the physical cause of a highly localized edge quickly reveals that there is a high degree of uncertainty as to whether that edge is due to a discontinuity in depth, a change in surface pigment, a shadow, a texture, or a reflection from a shiny surface. The visual system resolves this ambiguity by relying on prior assumptions about external physical states or by combining the localized image with other contextual information (e.g., other sparsely sampled image patches) to arrive at a probable interpretation of the scene. Similarly, in the present set of experiments, we find that with identification, observers judged highly ambiguous flows in the small color condition by incorporating their built-in prior assumptions that materials were rigid with the color information from the image (Supplementary Figure S4). On the other hand, in the small color similarity arrangements, the prior on rigidity—a deep feature—was not as highly correlated with the arrange-ments (Figure 9), suggesting that observers relied more heavily on image features in this task.

Conclusion

Shallow perceptual dimensions like color can sometimes account for perceptual similarity and material inference of natural dynamic flow (e.g., fire can be identified by its distinctive pattern of colors that flows upward). Other times, when shallow perceptual dimensions are ambiguous about the underlying causes, vision relies on deeper perceptual dimensions that reveal the generative physical causes of images (e.g., a solid can be differentiated from a liquid or gas by the strength of its intermolecular forces). We find that perceived dynamic flow similarity and inferences based on deeper dimensions require greater visual spatial context and enable the estimation of important action decisions, such as pick-up-ability and penetrability, while inferences based on shallow dimensions, like color, do not. Thus, visual inference of material from dynamic flow appears to fall along a shallow-to-deep continuum, with the depth of the representation guided by behaviorally important action decisions.

Keywords: flow perception, material perception, dynamic textures, natural image statistics, visual inference

Acknowledgments

We thank Kendrick Kay for helpful discussions on data analysis, and Shinho Cho, Erik Wingerson, Hanlin Zhu, and the reviewers for helpful comments. We would also like to thank Renaud Péteri, Sándor Fazekas, Mark J. Huiskes, the National Park Service, Jakob Op den Brouw, Randy Perry, Philip Moore, Nicole Alfonzo, Thomas Porett, Idan Radai, Dan Meyer, Simon Bolz, Age of Rockets Production and Design, R&A Collaborations (Richard Foot and Arron Fowler), Jean Slosberg, theFilmArtist, Dmitrii Lezine, Boris Godfroid, Justin Lewis, and Sebastian Sadowski for allowing us to use portions of their films as stimuli. The work was funded by Office of Naval Research Grant N000141210883 to DJK.

Commercial relationships: none.
Corresponding author: Yaniv Morgenstern.
Email: yaniv.morgenstern@psychol.uni-giessen.de.
Address: Department of Psychology, Justus-Liebig-Universität Giessen, Giessen, Germany.

References

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. W. (2015). Humans predict liquid dynamics using probabilistic simulation. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 172–177). Austin, TX: Cognitive Science Society.

Bell, S., Upchurch, P., Snavely, N., & Bala, K. (2015). Material recognition in the wild with the materials in context database. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3479–3487). Washington, DC: IEEE.

Bergou, M., Audoly, B., Vouga, E., Wardetzky, M., & Grinspun, E. (2010). Discrete viscous threads. *ACM Transactions on Graphics*, *29*(4), 1–10.

Bi, W. Y., and Xiao, B. (2016). Perceptual constancy of mechanical properties of cloth under variation of external force. In *SAP '16: Proceedings of the ACM symposium on applied perception* (pp. 19–23). New York: ACM.

Brainard, D. H., & Maloney, L. T. (2011). Surface color perception and equivalent illumination mod-

els. *Journal of Vision*, *11*(5):1, 1–18, doi:10.1167/11.5.1. [PubMed] [Article]

Bouman, K. L., Xiao, B., Battaglia, P., & Freeman, W. T. (2013). Estimating the material properties of fabric from video. In P. Kellenberger (Ed.), *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1984–1991). Washington, DC: IEEE Computer Society.

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

Diedrichsen, J., Ridgway, G. R., Friston, K. J., & Wiestler, T. (2011). Comparing the similarity and spatial structure of neural representations: A pattern-component model. *NeuroImage*, *55*(4), 1665–1678.

Doerschner, K., Fleming, R. W., Yilmaz, O., Schrater, P. R., Hartung, B., & Kersten, D. (2011). Visual motion and the perception of surface material. *Current Biology*, *21*(23), 2010–2016, doi:10.1016/j.cub.2011.10.036.

Doretto, G., Chiuso, A., Wu, Y. N., & Soatto, S. (2003). Dynamic textures. *International Journal of Computer Vision*, *51*(2), 91–109.

Grzywacz, N. M., & Hildreth, E. C. (1987). Incremental rigidity scheme for recovering structure from motion: Position-based versus velocity-based formulations. *Journal of the Optical Society of America A*, *4*(3), 503–518.

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, *9*(11), 1367–1368.

Jozwik, K. M., Kriegeskorte, N., & Mur, M. (2016). Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia*, *83*, 201–226.

Kawabe, T., Maruya, K., Fleming, R.W., Nishida, S. (2015). Seeing liquids from visual motion. *Vision Research*, *109*, 125–138.

Kawabe, T., Maruya, K., & Nishida, S. (2015). Perceptual transparency from image deformation. *Proceedings of the National Academy of Sciences, USA*, *112*(33), E4620–E4627.

Kawabe, T., & Nishida, S. Y. (2016). Seeing jelly: Judging elasticity of a transparent object. J. In Editor (Ed.), *Proceedings of the ACM Symposium on Applied Perception* (pp. 121–128). New York: ACM.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*(14), 1.

Knill, D. C., & Kersten, D. (1991) Apparent surface curvature affects lightness perception. *Nature*, *351*, 228–229.

Koch, K., McLean, J., Segev, R., Freed, M. A., Berry, M. J., Balasubramanian, V., & Sterling, P. (2006). How much the eye tells the brain. *Current Biology*, *16*(14), 1428–1434.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, *3*, 245.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Kubricht, J., Jiang C., Zhu, Y., Zhu, S.-C., Terzopoulos D., & Lu, H. (2016). Probabilistic simulation predicts human performance on viscous water-pouring problem. In A. Papafragou et al. (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 1805–1810). Philadelphia, PA: Cognitive Science Society.

Kwatra, V., Schödl, A., Essa, I., Turk, G., & Bobick, A. (2003). Graphcut textures: Image and video synthesis using graph cuts. *ACM Transactions on Graphics*, *22*(3), 277–286.

Lawson, C. L., & Hanson, R. J. (1995). *Solving least squares problems* (Vol. 15). Philadelphia, PA: Society for Industrial and Applied Mathematics.

Lizarraga-Morales, R. A., Guo, Y., Zhao, G., Pietikäinen, M., & Sanchez-Yanez, R. E. (2014). Local spatiotemporal features for dynamic texture synthesis. *EURASIP Journal on Image and Video Processing*, *2014*(1), 1–15.

Marlow, P., Kim, J., Anderson, B. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, *22*(20), R865–R866.

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E.H. (2007). Image statistics and the perception of surface qualities. *Nature*, *447*(7141), 2006–2009.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, *10*(4), e1003553.

Paulun, V. C., Schmidt, F., van Assen, J. J. R., & Fleming, R. W. (2017). Shape, motion, and optical cues to stiffness of elastic objects. *Journal of Vision*, *17*(1):20, 1–22, doi:10.1167/17.1.20. [PubMed] [Article]

Péteri, R., Fazekas, S., & Huiskes, M. J. (2010).

DynTex: A comprehensive database of dynamic textures. *Pattern Recognition Letters*, *31*(12), 1627–1632.

Phong, B. T. (1975). Illumination for computer generated pictures. *Communications of the ACM*, *18*(6), 311–317.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–70.

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *International conference on image processing* (Vol. 3, pp. 444–447). Washington, DC: IEEE Computer Society.

Sun, D., Roth, S., & Black, M. J. (2010). Secrets of optical flow estimation and their principles. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2432–2439). Washington, DC: IEEE.

Ullman, S. (1979). *The interpretation of visual motion*. Cambridge, MA: MIT Press.

van Assen, J. J. R., & Fleming, R. W. (2016). Influence of optical material properties on the perception of liquids. *Journal of Vision*, *16*(15):12, 1–20, doi:10.1167/16.15.12. [PubMed] [Article]

Ward, K., Bertails, F., Kim, T. Y., Marschner, S. R., Cani, M. P., & Lin, M. C. (2007). A survey on hair modeling: Styling, simulation, and rendering *IEEE Transactions on Visualization and Computer Graphics*, *13*(2), 213–234.

Appendix

Estimating perceptual dimensions and affordances

We use the phrase *perceptual dimension* to refer to a feature or a group of features arising from a common quality or theme that represents aspects of the image or physical scene that generated the image. Color, for example, is a perceptual dimension, and two of its features can be the mean and standard deviation of a single pixel from the R, G, or B color layer. In contrast, an affordance is a *possible action* with the image or scene.

In order to evaluate how well candidate dimensions could explain flow similarities, we posited a set of dimensions from shallow to deep. Shallow dimensions were features based on easily computable image measures, such as color, that are known to be important for image synthesis and perception (Table 1). To aid in identifying higher level or deeper dimensions that are not easily computable from the images, we also asked observers to assign nonoptical material attributes to the flows that are known to be important in physics-based computer graphics, and to categorize the flows by name. In order to evaluate how well our action decisions could explain flow similarities, we also asked observers to rate the flows on affordance properties (e.g., penetrability).

We analyzed the dynamic flows, extracting perceptual dimensions along the shallow-to-deep continuum. The dimensions near the shallow end were related to the response properties of neurons in the early visual systems (e.g., color, information summarized from a multiscale spatiotemporal pyramid decomposition, and optic flow) and computed with common image-processing tools. Perceptual dimensions near the deeper end of the continuum included intermediate-level material properties (e.g., viscosity), and categories. We used human-observer responses to estimate these dimensions, since reliable machine-vision methods do not exist or were not readily available.

In the experiments, the dynamic flows were viewed within a circular aperture. The shallow dimensions of these flows were extracted from a square region within the circular aperture. For the small-aperture condition (aperture diameter = 1.83°, 78 pixels), one side of this square region amounted to 1.13° (48 pixels). For the large-aperture condition (aperture diameter = 10.76°, 460 pixels), one side of the square region amounted to 7.50° (320 pixels).

Following are descriptions of how we extracted these perceptual dimensions.

Color

The RGB color images for each movie frame were converted to LAB color space. We computed the mean, variance, skew, and kurtosis by marginalizing across space (i.e., for each movie frame). Then these marginalized spatial statistics were marginalized across time (i.e., across the movie frames), but separately for each color layer, producing the mean, variance, skew, and kurtosis for each of the four marginalized spatial statistics. This led to 48 features (4 statistical summaries across frames × 4 statistics × 3 color layers) for the color dimension.

Multiscale spatiotemporal statistics

The spatiotemporal statistical features were based on a pyramid decomposition (Simoncelli & Freeman, 1995) that breaks up an image into a high-pass component, a number (N_f) of frequency sub-bands (which are further separated into N_o orientation

bands), and a low-pass component. We took this decomposition and summarized the dynamic flows in terms of the core features that the Portilla–Simoncelli model (Portilla & Simoncelli, 2000) uses to synthesize novel textured images of the original image. These core sets of features are based on the original image and its pyramid decomposition's marginal statistics, local autocorrelation, cross correlations with other sub-bands and orientations, and cross-scale phase statistics.

Marginal statistics: The marginal statistics characterize the pixel-intensity distribution of the original image and the pyramid decomposition at different spatiotemporal scales. The statistics include mean, variance, skew and kurtosis of the original image and of the magnitude (absolute value) at each sub-band and orientation, including the low- and high-pass components. Thus, the total number of marginal statistical features was 4 (mean, variance, skew, and kurtosis for the original image) $+ 4 \times (N_f \times N_o + 2$ (high- and low-pass components)).

Local autocorrelation: The local autocorrelation characterizes the salient spatial frequencies and the regularity of the images, as represented by periodic or globally oriented structures (Portilla & Simoncelli, 2000). Here the local autocorrelation is taken to be a region of 9 pixels for each of the N_f frequency sub-bands. We include the real autocorrelation for each frequency sub-band (with all orientations included) and the autocorrelation of the magnitude of each sub-band and orientation band. By taking each 9-pixel autocorrelation to be a single feature (rather than each of the 9 pixels being a unique feature), we have as the total number of local autocorrelation features N_f (real autocorrelation at each band across orientations) and $N_f \times N_o$ (autocorrelation of magnitude at each sub-band and orientation band).

Cross correlation: The correlation of sub-band magnitudes of an image's pyramid decomposition has been previously used to represent structures such as edges, bars, and corners in image textures. Here we take these products in three ways:

a. We cross-correlate each sub-band magnitude with its cousins—that is, those of the other orientations at the same scale. The number of pairwise cousin products for each scale is $N_o(N_o - 1)/2$. From taking the $N_o(N_o - 1)/2$ cousin products along a single scale to be a single feature, there are N_f features.
b. We also cross-correlate the sub-band magnitudes with their parents—that is, all orientations at the next (coarser) scale (until there are no more parents). The number of child–parent products is the product of the number of orientations at each scale ($N_o \times N_o$). From taking the $N_o \times N_o$ child–parent products across one scale as a single feature, there are $N_f - 1$ features.
c. Finally, we also cross correlate the real sub-band image with its cousins—that is, other orientations at the same scale. The number of pairwise cousin products for each scale is $N_o(N_o - 1)/2$. We treat all the pairwise cousin products for a given scale as a single feature; thus, there are N_f features.

In total, there are $3 \times N_f - 1$ features.

Phase correlation: The phase correlation distinguishes edges from lines, and helps in representing gradients due to shading and lighting effects (Portilla & Simoncelli, 2000). The phase correlation is based on the child–parent cross correlation of the real part of the child with both the real and imaginary parts at all orientations at the next coarser level. The number of cross products is the N_o (number of orientations at the finer scale) $\times 2$ (1 for real part and 1 for the imaginary part) $\times N_o$ (the number of orientations at the coarser scale). We treat all child–parent products across one scale as a single feature; thus, there are $N_f - 1$ features.

Spatiotemporal statistics for dynamic flows

A movie clip of a dynamic flow can be visualized as a three-dimensional volume, which has two spatial dimensions (x and y) and a temporal dimension (t). We used the pyramid decomposition introduced by Portilla and Simoncelli (2000) to separately summarize statistics across the spatial dimensions and the two space-time dimensions (x and t and also y and t) for grayscale versions of the dynamic flows. Along the x-y dimensions, we decomposed images into N_f spatial scales with N_o orientations. We took multiple decomposed samples across the opposite dimension (i.e., time or frames). For the large-aperture condition, $N_f = 4$ and $N_o = 4$. For the small-aperture condition, $N_f = 2$ and $N_o = 4$. In total, we took 64 samples. Along the x-t and y-t dimensions, we decomposed images along the opposite spatial dimension, into N_f spatiotemporal sub-bands with N_o orientations. For the large-aperture condition, $N_f = 4$ and $N_o = 4$. The number of samples—pixels along the opposite spatial dimension—was 320. For the small-aperture condition, $N_f = 3$ and $N_o = 4$. The number of samples was 48. We then summarized the three types of pyramid decomposition (x-y, y-t, and x-t) in terms of the multiscale spatiotemporal features. Since there were many samples for each dimension (e.g., 64 samples along x-y), the final feature value for a dynamic flow was taken to be the average of that feature across the sampled dimension.

Optic flow

We extracted optic-flow fields F (F_x for horizontal vector and F_y for vertical vector) for successive frames

(using MATLAB code provided by Sun et al., 2010) from grayscale versions of the dynamic flows. We summarized these optic flow fields as follows:
Speed: We summed the scalar of the motion vector extracted from optic-flow fields marginalized across space. We calculated the mean, variance, skew, and kurtosis for these values marginalized over time.
Direction: We summed the scalar of the motion vector extracted from optic-flow fields marginalized across space and time. We computed the angle relative to a reference vector pointing toward (0,1).
Absolute curl: The vector field's curl represents its magnitude of rotation. The curl is computed by subtracting the differences in the values of the vector field along the axis orthogonal to the vector components:

$$F_{curl} = \nabla \times F = \frac{\partial F_x}{\partial y} - \frac{\partial F_y}{\partial x}.$$

Its discrete form is

$$F_{curl}(i,j,t) = \frac{F_x(i,j+1,t) - F_x(i,j-1,t)}{2} - \left(\frac{F_y(i+1,j,t) - F_y(i-1,j,t)}{2}\right).$$

We computed the mean, variance, skew, and kurtosis for the absolute curl of the vector field marginalized across space, then calculated the mean, variance, skew, and kurtosis of these values marginalized over time.
Absolute divergence: The vector field's divergence represents the extent to which there is more flow exiting a region of space than entering it. In fluid simulation, converging flows—such as a narrowing river—act like funnels that cause the overall flow velocity to increase. Diverging flows, on the other hand, spread the particles out, causing the flow speed to decrease. The divergence is calculated by summing the differences in the values of the vector field along the axis parallel to the vector components:

$$F_{div} = \nabla \cdot F = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y}.$$

The discrete form is

$$F_{div}(i,j,t) = \frac{F_x(i+1,j,t) - F_x(i-1,j,t)}{2} + \frac{F_y(i,j+1,t) - F_y(i,j-1,t)}{2}.$$

We computed the mean, variance, skew, and kurtosis for the absolute value of the divergence of the vector field across space, then calculated the mean, variance, skew, and kurtosis of these values over time.
Gradient: The vector field's gradient represents the direction of the greatest rate of increase. In fluid simulation, the gradient describes the movement of particles with pressure change; high-pressure regions push low-pressure regions, just like concentrated pressure on dough will force the dough to spread out to lower pressure regions. The gradient is computed as the slope of the variation in scalar of motion vector components along the axis parallel or orthogonal to the vector:

$$F_{grad} = \vec{\nabla} F = \frac{\partial F}{\partial x} I + \frac{\partial F}{\partial y} J,$$

where I and J are standard unit vectors. The discrete form is

$$F_{grad}(i,j,t) = \left[\left(\frac{F(i+1,j,t) - F(i-1,j,t)}{2}\right)^2 + \left(\frac{F(i,j+1,t) - F(i,j-1,t)}{2}\right)^2\right]^{\frac{1}{2}}$$

We computed the gradient along the horizontal F_{grad_x} and vertical F_{grad_y} vector fields across space. We also combined them as a root sum of squares:

$$F_{grad\_xy} = \sqrt{F_{grad\_x}^2 + F_{grad\_y}^2}.$$

We computed the mean, variance, skew, and kurtosis for F_{grad_x}, F_{grad_y}, and F_{grad_xy} of the vector fields across space, then calculated the mean, variance, skew, and kurtosis of these values over time.
Laplacian: The Laplacian operator represents the divergence of the gradient of a vector field:

$$F_{lap} = \nabla \cdot \vec{\nabla} F = \frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2}$$

The discrete form is

$$F_{lap}(i,j,t) = -4F(i,j,t) + F(i+1,j,t) + F(i-1,j,t) + F(i,j+1,t) + F(i,j-1,t).$$

In fluid simulation, the Laplacian operator describes how a particle moves relative to its neighbors. We computed the Laplacian along the horizontal F_{lap_x} and vertical F_{lap_y} vector fields across space. We also combined them as a root sum of squares:

$$F_{lap\_xy} = \sqrt{F_{lap\_x}^2 + F_{lap\_y}^2}.$$

We computed the mean, variance, skew, and kurtosis for F_{lap_x}, F_{lap_y}, and F_{lap_xy} of the vector fields across space, then calculated the mean, variance, skew, and kurtosis of these values over time.

Optic-flow statistics for dynamic flows

We assumed that observers attend to larger changes in motion magnitude for similarity judgments. Thus,

we analyzed only flow vectors (computed over a two-frame distance) that had a magnitude > norminv(0.9) × SD, where SD is the standard deviation of the magnitudes of all flow vectors in a given frame. This also serves to denoise the flow data. Moreover, given the multiscale nature of neurons in the early visual cortex and the tendency for improvements in optic-flow outputs with blur, we extracted optic-flow fields for the dynamic flow movie clips at several coarser spatial scales. For the large-aperture conditions, the optic flow was computed for movies whose length was resized from 7.5° to 3.75°, 1.88°, 0.94°, and 0.47°. For the small-aperture conditions, it was computed for movies whose length was resized from 1.13° to 0.56°, and 0.28°.

Material and affordance attributes

Observers rated the dynamic flows along six attributes (Table 2). The attributes closely related to the material properties are viscosity, elasticity, rigidity, and compressibility, while those related more to important action decisions are pick-up-ability and penetrability. To compute a given flow's feature value, we first scaled each observer's estimated attributes from 0 to 1. The flow's feature value was then the average of the scaled observers' ratings for that flow (missing values or "not applicable" rating scores were ignored). This provided an overall pooled feature value for each flow and each attribute.

Categories

We created categorical features based on observers' dynamic flow identification (Supplementary Figure S2). Within each feature, a flow was assigned a value of 1 if the feature was present and 0 if it was absent. There were a total of 14 categorical features that fell into three broad classes based on whether the stimulus belonged to some material category, had some aspect of the simulation model (Figure 1A and 1B), or was based on some conceptual theme.

Features related to the material category: The following categorical groups are related to a flow's identified material category. Since materials from the same class have similar properties, objects within a category will have similar simulation, rendering, and image-statistic properties (Figure 1).

a. Water. Observers categorized the flow as some form of water. See Supplementary Figure S2 for what is considered a *water* classification.
b. Nonwater liquids. Observers categorized the flow as some form of liquid other than water (e.g., milk).
c. Cloth. Observers categorized the flow as a cloth or flag.
d. Solids. Observers categorized the flow as a solid. See Supplementary Figure S2 for what is considered a *solid* classification.
e. Plants. Observers categorized the flow as a plant.
f. Other fluids. Observers categorized the flow as a fluid other than a liquid, such as vapor, fire, or smoke.

Features related to the simulation model: The following categorical groups are related to properties from the simulation stage. This category will to some extent also reflect the rendering and image-statistics stage, as materials that fall into the same simulation category will also have similar properties (Figure 1). However, as opposed to the material-category variables just discussed, the simulation-model variables will group stimuli across material categories, so there will be greater variation in rendering properties and image statistics.

a. Strong intermolecular bonds. Observers categorized the flow as a solid, cloth, plant, web, hair, or Jell-O.
b. Intermediate intermolecular bonds. Observers categorized the flow as water, nonwater liquid, or lava.
c. Weak intermolecular bonds. Observers categorized the flow as vapor, fire, smoke, steam, or an object consisting of many tiny particles (e.g., snow, dust, powder, sand).
d. Wind. The previous three categories are related to material properties. Wind and gravity, on the other hand, contribute to the forces that guide the object's behavior during simulation (Figure 1). The wind category represents these external forces. Here observers categorized the flow as something that is blown by the wind. This includes land plants, cloths, hair, and webs.

Features related to the conceptual theme: The following categorical groups are related to higher level scene analysis.

a. Human-made. Observers categorized the flow as something that is human-made. This includes human-made liquids (coffee, Coca-Cola, etc.), cloths, and solids. Humans are partially responsible for making the objects underlying these flows.
b. Natural. Observers categorized the dynamic flow as something that is natural (i.e., not artificial, or objects such as cloths that are human-made). This includes liquids (such as water or milk), animate objects, webs, fire, and smoke.
c. Food. Observers categorized the flow as something that is edible and nonliving (e.g., honey, Nutella, water, snow, oil).

d. Water objects. Observers categorized the dynamic flow as something that is made of water or found in water. This includes water, aquatic plants, and snow.

Adding category features for Color, XYT, Flow, and A&A dimensions

In addition to real-valued features for the groups already discussed, we added categorical features for Color, XYT, Flow, and A&A. Each feature was converted to three additional categorical features that categorically grouped the data into whether they had low, medium, or high values. For example, in the low-value category feature, the real values that had a low value were set to 1 and the remaining values were set to 0. The thresholds for what was considered low, medium, or high were determined by separating the range of real-valued outputs across stimuli for the chosen feature into three roughly equal-sized groups. We found that these categorical features used as additional regressors improved model correlations with human similarity arrangements.

Grouping perceptual dimensions

We created three additional perceptual dimensions by grouping the perceptual dimensions listed above as follows:

a. Shallow. This perceptual dimension was created by merging the low-level features (color, multi-scale spatiotemporal statistics, and optic flow).
b. Deep. This perceptual dimension was created by combining the higher level features (affordances and material attributes and categories).
c. All. This perceptual dimension included all features.

Representational similarity analysis

Creating model RDMs

In order to compare the models to the similarity judgments, the model responses were transformed into the same space as the human responses. For each model feature, we computed for each pair of dynamic flows the square root of the squared difference between their values on that feature (i.e., the Euclidean distance). Once the computation for all pairwise comparisons was complete, the dissimilarities were assembled into the upper triangular portion of their RDM and normalized to have unit sums of squares.

Reducing features before regression

Averaging highly correlated features: Following in the footsteps of Jozwik, Kriegeskorte, and Mur (2016), to increase the stability of the weights estimated during regression, we iteratively combined high-correlated ($r > 0.9$) vectors, alternately computing pairwise correlations between the vectors and averaging highly correlated vector pairs, until all pairwise correlations were below threshold.

Removing features with outliers: We removed features with outliers, since they tended to be poorly correlated with the human arrangements, but also because of their potential influence on model selection with the leave-one-out cross-validation procedure discussed later. For each feature, we computed the average distance of a dynamic flow with all other flows. This led to a 43-element vector per dataset, $\vec{d^i}$. The feature was removed if its deviation $\left(\vec{d^i} - \text{median}\left(\vec{d^i}\right)\right)/\left(\text{MAD}\left(\vec{d^i}\right)\right)$ from the median (normalized by its median absolute deviation) was greater than a threshold. The threshold was determined as follows. For each feature, we computed the maximum deviation from the median (normalized by its median absolute deviation). The threshold value was the 95th percentile of these maximum values across all features.

Non-negative least-squares fitting of the representational models

We used linear methods to find the optimal weighted sum of model-RDM features that predict the measured similarity representations (Diedrichsen, Ridgway, Friston, & Wiestler, 2011; Jozwik et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014). To use these methods, first we had to transform the human and model RDMs from Euclidean distances to squared differences, since squared differences can be added across dimensions. In our case, both the human and feature RDMs are in terms of Euclidean distances:

$$d^{ij} = \sqrt{\left(f^i_k - f^j_k\right)^2} = \sqrt{\left(\Delta f^{ij}_k\right)^2},$$

where d^{ij} is the distance between stimuli i and j on feature k, f^i_k is the value on feature k for stimulus i, and f^j_k is the value on feature k for stimulus j.

By squaring the Euclidean distances to get the squared differences between stimuli on a given feature, we can apply traditional linear feature-combination methods, since now our features, in terms of squared differences, can be added:

$$\left(d^{ij}\right)^2 = w^2_1\Delta f^2_1 + w^2_2\Delta f^2_2 + \ldots + w^2_k\Delta f^2_k,$$

where w_k is the weight given to feature k and Δf^2_k is the squared difference between stimuli i and j on feature k.

We converted the human and feature RDMs from Euclidean distances to squared differences and then estimated the RDM weights with a non-negative least-

squares fitting algorithm (Jozwik et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Lawson & Hanson, 1995) in MATLAB (function lsqnonneg). In order to prevent positive bias of the model performance estimates due to overfitting to a particular set of stimuli, model-prediction accuracy was estimated by cross validation with a subset of the dynamic flows held out on each fold. For each cross-validation fold, we selected 43 of the 44 dynamic flows as the training set and used the corresponding pairwise dissimilarities for estimating the model weights. The model weights were then used to predict the pairwise dissimilarities for the left-out dynamic flow. This procedure was repeated until every flow was left out and predictions were obtained for all pairwise dissimilarities. Finally we converted these pairwise dissimilarities from squared differences to Euclidean distances and correlated them with the human similarity arrangements.

Inferential analysis on model performance

We used the representational-similarity-analysis toolbox for inferential analyses (Nili et al., 2014). We quantified model performance by measuring the Pearson correlation between the human dissimilarities and the dissimilarities predicted by the models. For each model, we computed the correlation coefficient between each subject's data RDM and the RDM predicted by the model. Figure 9 and Supplementary Figures S7 and S8 show the subject-average correlation coefficients for the fitted models.

We first determined whether each of the model-prediction RDMs is significantly related to each subject-average data RDM using a stimulus-label randomization test (10,000 randomizations per test). The test simulates the null hypothesis that the RDMs are unrelated (i.e., zero correlation). We conclude that the model-prediction and data RDMs are significantly related if the actual correlation falls within the top tail of the simulated null distribution. We used Bonferroni correction to adjust the alpha value for multiple comparisons. Next we tested for differences in model performance. We performed pairwise model comparisons using bootstrap resampling of the stimulus set (1,000 bootstrap resamplings per test). This simulates the variability of model performance across random samples of stimuli. If the simulated distribution of model-performance differences is significantly greater than zero, we conclude that the actual model performances significantly differ from each other. We corrected for multiple comparisons by adjusting alpha with Bonferroni correction.