Estimating mechanical properties of cloth from videos using dense motion trajectories: Human psychophysics and machine learning

Wenyan Bi
Department of Computer Science, American University, Washington, DC, USA

Peiran Jin
Department of Physics, Georgetown University, Washington, DC, USA

Hendrikje Nienborg
Werner Reichardt Center for Integrative Neuroscience, University of Tübingen, Tübingen, Germany

Bei Xiao
Department of Computer Science, American University, Washington, DC, USA

Humans can visually estimate the mechanical properties of deformable objects (e.g., cloth stiffness). While much of the recent work on material perception has focused on static image cues (e.g., textures and shape), little is known about whether humans can integrate information over time to make a judgment. Here we investigated the effect of spatiotemporal information across multiple frames (multiframe motion) on estimating the bending stiffness of cloth. Using high-fidelity cloth animations, we first examined how the perceived bending stiffness changed as a function of the physical bending stiffness defined in the simulation model. Using maximum-likelihood difference-scaling methods, we found that the perceived stiffness and physical bending stiffness were highly correlated. A second experiment in which we scrambled the frame sequences diminished this correlation. This suggests that multiframe motion plays an important role. To provide further evidence for this finding, we extracted dense motion trajectories from the videos across 15 consecutive frames and used the trajectory descriptors to train a machine-learning model with the measured perceptual scales. The model can predict human perceptual scales in new videos with varied winds, optical properties of cloth, and scene setups. When the correct multiframe was removed (using either scrambled videos or two-frame optical flow to train the model), the predictions significantly worsened. Our findings demonstrate that multiframe motion information is important for both humans and machines to estimate the mechanical properties. In addition, we show that dense motion trajectories are effective features to build a successful automatic cloth-estimation system.

Introduction

In everyday life, we visually estimate material properties of objects when deciding how to interact with them. For example, to grasp a sweater, it is helpful to know its heaviness and stretchability before we come into contact with it. Previous work in material perception has mainly focused on understanding the optical properties of rigid objects, such as surface gloss and translucency (Fleming, Dror, & Adelson, 2003; Fleming & Bülthoff, 2005; Landy, 2007; Motoyoshi, Nishida, Sharan, & Adelson, 2007; Ho, Landy, & Maloney, 2008; Xiao & Brainard, 2008; Kim & Anderson, 2010; Motoyoshi, 2010; Wijntjes & Pont, 2010; Doerschner et al., 2011; Fleming, Jäkel, & Maloney, 2011; Gkioulekas et al., 2013; Xiao et al., 2014). But many materials around us are soft and deformable (e.g., cloth, gels, and liquids). The mechanical properties of these objects determine the way they move and adopt particular shapes in response to external forces. Estimating mechanical properties from visual input is challenging because both external forces and the intrinsic mechanical properties affect appearance. A physics-driven cloth-simulation model involves many parameters and complicated calculations defining...
how a piece of cloth responds to external forces. However, under typical viewing conditions, humans can infer mechanical properties such as stiffness, weight, and elasticity of the object just by looking. It is unlikely that humans can reverse the modeling process; they are more likely to use diagnostic image cues to disambiguate the effects of intrinsic and external factors. Little is known about the image cues that allow for such robust estimation of mechanical properties in complex dynamic scenes.

Some studies have shown that image cues, such as two-frame motion (e.g., optical flow; Kawabe, Maruya, Fleming, & Nishida, 2015; Kawabe & Nishida, 2016), local 3-D structure (Giesel & Zaidi, 2013), and shape deformations (Paulun, Kawabe, Nishida, & Fleming, 2015; Kawabe & Nishida, 2016; Paulun, Schmidt, van Assen, & Fleming, 2017; Schmidt, Paulun, van Assen, & Fleming, 2017), can affect the perception of mechanical properties. However, static cues and two-frame motion cues can be conflicting and accidental, and therefore they are sometimes insufficient to capture the impression of mechanical properties. For example, Figure 1A shows two static images of the same fabric. From the static images, we can already tell a lot about the material properties of the cloth (e.g., transparency and shape deformation). However, we might perceive the cloth in the left image to be stiffer than the one in the right. In contrast, when we view video sequences of how the fabric moves under a wind force (Figure 1B; for a video, see Malcolm, 2017), we might achieve a more consistent impression of its stiffness. This indicates that multiframe motion may help disambiguate the conflicting information and help observers achieve a consistent judgment. In this article, we investigate the role of such long-range motion information, characterized as spatiotemporal coherence over multiple frames, on the perception of mechanical properties of cloth from videos.

Recently, the role of motion in visual perception of material properties has been receiving increasing attention, especially in the studies of liquids and deformable or elastic objects (Davis et al., 2015; Kawabe et al., 2015; Kawabe & Nishida, 2016; Marlow & Anderson, 2016; Paulun et al., 2017; Van Assen, Barla, & Fleming, 2018). Some studies have found that motion can cause changes of other cues (e.g., shape deformation and viewing angle), which are crucial in visual perception of material properties (Warren, Kim, & Husney, 1987; Sakano & Ando, 2010; Kawabe & Nishida, 2016; Paulun et al., 2017; Schmidt et al., 2017). For example, head movements while observing objects leads to a change of the angle of light refraction and reflection, which is critical in the perception of glossiness (Sakano & Ando, 2010). Schmidt et al. (2017) found that shape deformation caused by motion is important in the perception of elasticity of deformable cubes. In a dynamic scene that contains a bouncing ball, the perception of elasticity is mainly based on the relative height information (Warren et al., 1987). Other studies have managed to isolate and quantify motion information (Bouman, Xiao, Battaglia, & Freeman, 2013; Kawabe et al., 2015; Kawabe & Nishida, 2016; Morgenstern & Kersten, 2017). Using optical flow, Kawabe et al. (2015) demonstrated that the visual system utilized image motion speed in the optical-flow field as a cue to estimate liquid viscosity. Additionally, spatial smoothness of motion flow is critical for humans to estimate the liquid flow. With a similar method, Kawabe and Nishida (2016) found that human observers were able to recover the elasticity of computer-rendered jellylike cubes based on the shape-contour deformation alone. This was still true even

![Figure 1. Examples showing the importance of multiframe motion in the perception of the stiffness of cloth. Images come from a YouTube video showing a cloth blowing in the wind (Malcolm, 2017). (A) Two random image frames might provide conflicting information. The cloth in the left image looks stiffer than that in the right, although they are the same fabric. (B) When people see the movements of the cloth across a few frames, they might make a consistent judgment of the stiffness.](http://arvojournals.org/ on 06/15/2018)
when the cube movies were replaced by dynamic random noise patterns, which retained the optical-flow information but not the surface information. The researchers concluded that the elasticity judgment was based on the pattern of image motion arising from the contour and the optical deformations. In these studies, the motion information was typically extracted from two consecutive frames (i.e., two-frame motion). Additionally, there was usually little variation of external force in the scenes, such as pushing a cylinder into an elastic object with a constant pushing force.

In addition to liquids and deformable or elastic cubes, cloth is another unique yet extremely common type of deformable material. Previous findings on the perception of liquid might not be directly related to cloth. Shape and optical properties might be the dominant cues for the perception of viscosity of liquids (Paulun et al., 2015; Van Assen & Fleming, 2016). Unlike liquids and elastic or deformable cubes, cloth often forms wrinkles and folds under applied forces, and the wrinkles and folds can appear differently across the time course of the forces (as illustrated in Figure 2A). Thus, static information such as 2-D shape outline might not reliably reveal the mechanical properties of cloth.

A few studies have focused on cloth recognition and inference from dynamic scenes (Bouman et al., 2013; Aliaga, O’Sullivan, Gutierrez, & Tamstorf, 2015; Davis et al., 2015). For example, in the study by Aliaga et al. (2015), observers were asked to categorize the hybrid cloth video, which looks like one category of cloth (e.g., cotton) but moves like another (e.g., silk). The researchers found that the appearance, rather than the motion, dominated the categorical judgment, except for fabrics with extremely characteristic motion dynamics (i.e., silk). More recently, Yang, Liang, and Lin (2017) combined appearance and motion information to classify cloth. Specifically, they combined the image-signal feature extraction method (i.e., convolutional neural network) with the temporal-sequence learning method (i.e., long short term memory) to learn the mapping from visual input to material categorization. However, they did not explicitly test whether the model could be related to human perception. Moreover, they did not examine whether motion cues were important for humans to categorize cloth. In addition to categorization, humans often need to estimate the value of a particular property (e.g., how heavy is the cloth?) or compare a property of two objects (e.g., which one is heavier?) during daily activities such as online shopping. In a study by Bouman et al. (2013), which focused on machine estimation of cloth properties, observers from the Amazon Mechanical Turk were asked to estimate the stiffness and mass of cloth examples in real scenes. Results showed that the observers’ responses were well correlated with the log-adjusted physical-parameter values when the video stimuli were presented. This correlational relationship was less obvious when observers were asked to make judgments from still images. This finding supports the importance of motion in visual estimation of material properties. However, a single still image inherently contains much less information than a 10-s video clip. It is therefore possible that the observed better performance in the video condition than in the image condition was simply due to the fact that a single still image does not contain sufficient information for the purposes of estimating material properties. In the same article, the authors additionally trained a machine-learning model to predict the physical properties of the fabrics. Nevertheless, it is unknown whether their algorithm could be generalized to new dynamic scenes and whether multiframe motion information was included.

In this paper, we used computer-graphic generated high-fidelity cloth animations as stimuli to evaluate the effects of multiframe motion information on estimating the bending stiffness of the cloth. We particularly focused our attention on the stiffness because it is one prominent mechanical property of cloth as well as a common property across a variety of deformable objects. In the recent computer vision literature, densely sampled motion trajectories have been found to be a very effective cue for successful action recognition (Wang, Kläser, Schmid, & Liu, 2011; Wang, Kläser, Schmid, & Liu, 2013; Rubinstein, Liu, & Freeman, 2012). Inspired by these algorithms, we extracted multiframe correspondence of feature points across multiple frames by computing the dense trajectory descriptors from the cloth animation videos. We then trained a machine learning model with only these descriptors to predict the human perceptual scales of the bending stiffness. We found that the human observers can recover the differences in the bending stiffness of cloth samples from two different dynamic scenes, and that tracking feature points consecutively over multiple frames was important for both human and machine to infer the bending stiffness. In addition, we provided a dataset consisting of high-fidelity cloth videos with systematically varied mechanical properties and different textures that are accessible for other studies. The codes and demo videos of this paper are available online (https://github.com/BumbleBee0819/Estimating_mechanical_properties_of_cloth).

**Experiment 1a: Perceptual scale of bending stiffness**

In Experiment 1a, we measured how the perceived stiffness of a hanging fabric moving under an oscillating wind changed as a function of its physical
We used the maximum-likelihood difference-scaling (MLDS) method to estimate the function relating a physical parameter (i.e., the bending stiffness) to its corresponding perceptual score (Maloney & Yang, 2003; Knoblauch, Maloney, 2008; Wiebel, Aguilar, & Maertens, 2017). We varied two different scene parameters in separate blocks: the optical properties of the cloth (matte or glossy) and the scene setups (free corner or pinned corner; see Figure 2B). Across stimuli, we randomly sampled wind direction and strength from a fixed range so that each stimulus had slightly different movements.

**Materials and methods**

**Observers**

Five observers (four women, one man; mean age = 27.6 years, SD = 4.5 years) participated in the free-corner scene setup (Scene 1). Another seven observers participating in the pinned corner scene setup (Scene 2).
(five women, two men; mean age = 24.7 years, SD = 2.1 years) participated in the pinned-corner scene setup (Scene 2). All observers reported typical visual acuity and color vision. There were no significant differences between the two groups with respect to demographic data (p > 0.1). All observers took part in the experiment on a voluntary basis and were not paid for their participation.

**Stimuli**

*Scenes*: Figure 2 shows examples of frames from the video stimuli. Stimuli consisted of computer-rendered animations of cloth under oscillating wind force. We used two different scenes in the experiment (Figure 2B). Scene 1 consisted of a piece of cloth hanging on a rod and being blown by unknown oscillating winds. The top of the cloth was pinned onto the rod. Scene 2 was similar to Scene 1 except the lower two corners were initially pinned to another rod and the left corner was released after 80 frames. This simple change could cause the wind to interact with the cloth in different ways, and hence increase variability in the cloth’s movement.

*Wind forces*: The wind forces were different in the two scenes. Figure 2C (left panels) shows how the strength of the oscillating wind varies as a function of time in both scenes.

In Scene 1, a wind source was randomly placed on either the left front or the right front of the fabric for each video clip (see Figure 2C, upper right panel). The wind was oscillating in the horizontal plane perpendicular to the fabric. The initial wind direction was randomly chosen for each video to be between 30° and 90°. To better display how the cloth responded to external forces, the wind was on and off throughout the animation sequences (Figure 2C, upper left panel). At the beginning of the animation, the wind was on with a strength of 300 and an added noise level of 5, then it was turned off after 30 frames, and then it was restarted at the 120th frame. This was repeated for three cycles. In addition to the oscillating wind, we included a turbulence force field that could create more ripples on the cloth. The strength of the turbulence wind was 10 (1/30 of the main wind), with a noise level of 3.

In Scene 2, two oscillating wind sources were created: one on the left rear side and the other on the right front side (see Figure 2C, lower right panel). The lower left panel of Figure 2C shows the time course of the two winds. The initial strength of the left wind force was set to 200, with a noise level of 5, and that of the right was set to 50 with a noise level of 5. The initial angles of these two winds were random, and the oscillating angle ranged from −75° to 75° on the x- and y-planes and −45° to 45° on the z-plane. Similar to Scene 1, a turbulence force field with strength 10 and noise level 3 was added to the center of the cloth.

**Rendering**: All the cloth animations were rendered using the Blender Cycles Render Engine (Blender v. 2.76). We created two appearances for the cloth: cotton and silk. They differed in texture, thickness, surface reflectance, and roughness (examples are shown in Figure 2B). The scene was lit by four objects, which emitted light from the top, front, left side, and right side of the cloth. We modeled the cloth as a triangular mesh and used a mass spring model (Provat, 1995) to define the cloth’s interactions with external forces. The cloth–object collision was determined using the algorithm proposed by Mezger, Kimmerle, and Etzmuß (2002). We used three parameters to describe the intrinsic mechanical properties of cloth in Blender: bending stiffness, mass, and structural stiffness. These three parameters controlled the stiffness, heaviness, and elasticity of the cloth, respectively. We varied only the bending-stiffness values and kept the mass and structural stiffness fixed across all conditions. Thirteen different bending-stiffness values {0.005, 0.01, 0.1, 1, 5, 10, 25, 40, 80, 110, 180, 300, 450} were sampled. Based on the values of the preset materials in Blender and our observations, we determined that this range of bending stiffness would cover all relevant cloth categories. We kept the mass value constant at 0.7 and structural stiffness constant at 10 across all videos. Spring damping was set to 0, and both air damping and velocity damping were set to 1. Each animation lasted 12.5 s at 24 frames/s and all were saved as 1,280 × 720 pixel .mov files. See Supplementary Movies S3–S6 for illustration.

**Procedure**

Stimuli were presented on an LED display (27-in. iMac). Observers were seated about 70 cm away from the screen in a dark experimental chamber.

We used MLDS with the method of triads (Maloney & Yang, 2003; Knoblauch & Maloney, 2008) to measure the psychometric function relating changes in physical bending-stiffness values to changes in perceived stiffness by humans. On each trial, observers were presented with video triads (see Figure 3) and asked to judge whether the difference between the center video and the left video, in terms of the stiffness of the cloth, was greater or less than the difference between the center video and the right video. They indicated their choice by pressing the P or the Q key. Observers were explicitly told to ignore the differences in wind and focus only on the material properties of the cloth. On any given trial, the three videos in the triads always had different bending-stiffness values, and the stiffness of the center videos was always in between the left and right ones. Therefore, the stiffness of the three
videos was either in ascending (left < center < right) or descending (left > center > right) order.

We used 13 bending-stiffness values to construct the triads. The total number of unique triads was \( \binom{13}{3} \) for each texture and scene condition. Trials were randomized for each observer, scene, and texture. Observers had unlimited time to perform the task. The whole experiment took around 2 hr.

**Results**

Perceptual scales were computed for each condition and each observer separately using the MLDS package for R from Knoblauch and Maloney (2008). Figure 4 shows the estimated perceptual scale for each observer as a function of physical bending stiffness, along with the mean across all observers, which were estimated by MLDS using the GLM (generalized linear model) implementation (McCullagh, 1984). The upper panels show the perceptual scale for Scene 1, and the lower panels for Scene 2. For the majority of the parameter range, the perceptual scale increases as the bending stiffness increases in a log-linear fashion, indicating that observers are able to distinguish different bending-stiffness values. Observers performed equally well in both scenes.

Second, we looked closer at how the perceptual scale was related to the physical bending stiffness. We fitted the perceptual-scale parameters for each experimental condition (scene and texture) with a log function \( \Psi(x) = \)
\[ \ln(x) + b, \] where \( x \) represents the physical bending-stiffness value and \( \Psi(x) \) represents the perceived stiffness by human observers. We concatenated all observers’ data together for this analysis. We found a significant linear relationship between the log-adjusted physical bending stiffness and the perceptual scores in all conditions, \( R^2 > 0.82, p < 0.001 \).

Finally, we assessed the agreement between perceptual scales across the two texture conditions (Figure 4C and 4F, cotton versus silk). We built a global fit model with physical parameters as the independent variable and textures as the indicator variable, and then performed the analysis of covariance, which is a general linear model blending analysis of variance and regression (Howell, 2012). The results showed no interaction between the physical bending stiffness and the textures for either Scene 1, \( F(1, 126) = 0.002, p > 0.1 \), or Scene 2, \( F(1, 126) = 0.337, p > 0.1 \), indicating that the texture did not affect the shape of the perceptual scales. However, the textures affected the intercept of the perceptual scales, where the perceived stiffness of the cotton was higher than that of the silk in Scene 1, \( F(1, 127) = 6.09, p < 0.05 \) (see Figure 4C) but not in Scene 2, \( F(1, 127) = 0.002, p > 0.1 \) (see Figure 4F).

Discussion

Experiment 1a reveals that the perceived stiffness and log-adjusted physical bending stiffness are linearly correlated, and that textures do not influence observers’ sensitivity to the differences in stiffness. In Scene 1, we found that the optical appearance had a significant effect on the overall ratings of stiffness, such that the cotton cloth appeared to be stiffer than the silk. No such effect was found in Scene 2. This could be due to the fact that there were more movements of the cloth in Scene 1, so that the specular reflections on the silk had a bigger influence on the perceived bending stiffness. Except for this small difference, the perceptual scales are strikingly similar across textures and scenes. This indicates that humans can invariantly infer the differences between bending-stiffness levels despite differences in scenes and textures.

Experiment 1b: Perceptual scale with scrambled videos

Experiment 1a showed that observers are able to discriminate bending stiffness of moving fabrics across different textures and scenes. To estimate the bending stiffness, observers could use image cues, such as shape silhouettes, reflections, textures, and motion. In Experiment 1b, we were interested in investigating whether the correct temporal coherence between frames is necessary for such estimation. To do so, we created videos that contain randomly scrambled frame sequences from the original videos (i.e., scrambled video) and performed the same MLDS experiment as in Experiment 1a. The scrambled video does not contain the correct ground-truth motion sequences, but the contents of individual frames are meaningful. If the observers cannot distinguish bending stiffness equally well from the scrambled videos as from original videos, we would know that the longish correlation in multiframe motion information is important in estimating mechanical properties.

Materials and methods

Stimuli, design, and procedure

Stimuli were the same 13 silk videos from Scene 1 in Experiment 1 (upper right panel of Figure 2B), but with the frame sequences of each video randomly permuted. The experimental design and procedure were the same as in Experiment 1a. The observers first finished the MLDS experiment with the scrambled videos for the silk textures in Scene 1, and then were asked to finish the same sets of conditions for the unscrambled versions as well.

Observers

Four new observers (three women, one man; mean age = 24 years, \( SD = 1.83 \) years) participated in the scrambled- and original-video experiments. All observers reported typical visual acuity and color vision.

Results and discussion

We did the same MLDS data analysis as in Experiment 1a for this experiment. The perceptual scale was less correlated with the physical bending stiffness in the scrambled-video condition (Figure 5A; \( R^2 = 0.50 \)), than in the original-video condition (Figure 5B; \( R^2 = 0.80 \)). A paired \( t \) test on the residuals of the regression line fitted to the perceptual scale against the physical bending stiffness revealed that the residuals were much higher in the scrambled-video condition (\( M = 0.62, SD = 0.25 \)) than in the original-video condition (\( M = 0.56, SD = 0.32 \), \( t(51) = 2.82, p < 0.01 \)).

The results show that the perceived stiffness correlated worse with the physical bending stiffness in observation of scrambled videos than original videos. The observers could still distinguish the stiffest fabric from the most flexible one from the scrambled videos, but they had a hard time when the differences in stiffness between the two fabrics being compared were small. Even though, on average, observers can distin-
guish larger stiffness difference in scrambled videos almost as well as they can in unscrambled videos, there is much more individual variation. The observers seemed less confident of their decisions. This suggests that longish correlation in multiframe motion is critical in recovering the bending stiffness of cloth from dynamic scenes, even though static cues such as shape deformation could provide some information.

Computational modeling

In Experiments 1a and 1b, we found that the observers could distinguish the bending stiffness of cloth with varying physical stiffness values from dynamic scenes, and when the ground-truth motion sequences were scrambled, performance became worse. This suggests that the longish correlation in multiframe motion plays a role in the visual discrimination of mechanical properties. In this section, we provide computational evidence for this observation. To do so, we trained a machine-learning algorithm on the perceptual scales obtained from one texture condition and used it to predict the perceptual scale of the other texture condition. It should be noted that the wind forces in the two texture conditions were different, creating different movement patterns of the cloth samples. If multiframe motion information was discriminative in estimating the stiffness of cloth under dynamic scenes, then the model trained with only multiframe motion information should be able to recover people’s sensitivity to differences in the stiffness of cloth.

To extract the multiframe motion information, we used the method of extracting dense motion trajectories from studies of automatic action recognition (Wang et al., 2011; Wang et al., 2013) and trained a regression model with dense motion features alone to estimate the human perceptual scales. We first extracted the dense-trajectory features to represent the multiframe motion information, and then we encoded these features using the Fisher vector (FV). The input of the regression model was the concatenation of the FVs of each feature, and the attached label was the mean perceptual scale that was obtained from Experiment 1.

Method

Dense-trajectory motion features

In existing work on motion and material perception, two-frame optical flow is often computed (e.g., Doerschner et al., 2011; Kawabe et al., 2015) and the statistics extracted from the flow fields used to describe motion information. However, the two-frame optical-flow algorithm cannot accurately capture motion information across multiple frames. Here, inspired by recent advances in the field of computer vision in action recognition, we used dense-trajectory features to capture multiframe spatiotemporal information from our videos (Wang et al., 2011; Wang et al., 2013). A previous article (Rubinstein, Liu, & Freeman, 2012) illustrates the importance of tracking dense interest points over multiple frames. As figure 1 in that article shows, if only two frames are tracked using the optical-flow algorithm, the trajectories are broken and the trajectories for the same pixels can be hard to follow over.
time. This might cause information loss. By contrast, the trajectories of the pixels obtained by tracking multiple frames using the dense-trajectory algorithm are much smoother. This might explain the importance of long-range correspondences over multiple frames for characterizing the cloth motion we observed in Figure 1.

The first step in computing dense trajectories is the dense sampling of interest points, which we carried out at a fixed number of frames and on multiple spatial scales (see Figure 6A1 and 6B1). Each interest point $P_t = (x_t, y_t)$ at frame $t$ was tracked to the next frame $t + 1$ by median filtering in a dense optical-flow field. Points of subsequent frames were concatenated to form a trajectory: $(P_t, P_{t+1}, \ldots)$. In a smooth region without any structure, it is impossible to track any point. The algorithm will remove points in these areas. Therefore, probably similar to human vision, this algorithm does not rely on motion information in smooth regions. The details can be seen in figure 2 of Wang et al. (2011).

Given a trajectory of length $L$, its shape can be described by a sequence $S = (\Delta P_t, \ldots, \Delta P_{t+L-1})$ of displacement vectors. In addition to the Trajectory Shape descriptor, a histogram of optical flow (HOF: Figure 6A2 and 6B2), a histogram of gradient (HOG: Figure 6A3 and 6B3), a horizontal-motion boundary histogram (MBHx: Figure 6A4 and 6B4), and a vertical-motion boundary histogram (MBHy: Figure 6A5 and 6B5) were constructed over the spatiotemporal volume aligned with the trajectories to present more appearance and motion information. Specifically, HOG focused on static appearance, whereas HOF and MBH captured the local motion information. These five descriptors were finally combined to serve as the motion descriptor in this experiment.

**Data sets and models**

We fitted our data using a support vector-machine regression model that was optimized with dual stochastic gradient descent. With this same method, we built three models in this experiment. Table 1 summarizes the training and testing data sets for these three models.

Eleven of 13 videos of Scene 1 with cotton textures from Experiment 1a were used as training data for the regression model. The physical bending stiffnesses of these 11 videos were $\{0.005, 0.01, 1, 5, 10, 25, 40, 80, 180, 300, 450\}$. For each video, we chose six clips with random durations ranging from 1.25 to 2.69 s. Thus, our training data set contained 66 cotton video clips of different durations. Because the wind forces were varied throughout the original video, each video clip included in the training data set also contained unique...
wind forces. We used the mean perceptual scale across observers as the ground truth for the training data (Figure 4C, blue line).

We included two testing data sets for the regression model. The first testing data set (testing data1) contained the 11 silk videos in Scene 1 from Experiment 1a. The second testing data set (testing data2) contained silk and cotton videos with the other two bending-stiffness levels that the model had not seen. We applied the same clipping procedure as we did for the training data to create the testing video clips, resulting in 66 video clips for testing data1 and 12 for testing data2. The details of our experiments are discussed later. In sum, the main differences between the training and testing data are the length of the videos, the wind forces, and the optical appearance of the cloth. Due to this difference, the testing videos (silk, Figure 6A1) had very different trajectories from the training ones (cotton, Figure 6B1). In this article, we chose cotton as training data and silk as testing because they are very different in their trajectories when moving. However, we think our modeling method can still work if it is trained on other fabrics (see Results later).

The scrambled model was built using the same method except that the training and testing data were from the scrambled videos. This model provided a baseline measurement when the longish temporal correlation was removed. Lastly, we built a random model with randomly generated numbers (from 0 to 1) to serve as the chance-level baseline.

Model implementation

Figure 7 shows the pipeline of our framework of estimating perceptual scales of cloth from videos. First, we extracted the dense trajectories from both the training and testing data sets, with the parameters set to default according to the source code of Wang et al. (2011). The following element is five descriptors concatenated one by one:

- Trajectory: $2 \times \left[ \text{trajectory length} \right] \text{(default 30 dimension)}$

  - HOG: $8 \times \left[ \text{spatial cells} \right] \times \left[ \text{temporal cells} \right] \text{(default 96 dimension)}$

  - HOF: $9 \times \left[ \text{spatial cells} \right] \times \left[ \text{temporal cells} \right] \text{(default 108 dimension)}$

  - MBHx: $8 \times \left[ \text{spatial cells} \right] \times \left[ \text{temporal cells} \right] \text{(default 96 dimension)}$

  - MBHy: $8 \times \left[ \text{spatial cells} \right] \times \left[ \text{temporal cells} \right] \text{(default 96 dimension)}$

Next, we randomly subsampled 5,000 points from each motion descriptor for each movie clip. We then used principal-components analysis to reduce the dimensions of the motion descriptors by half: $5,000 \times \left( 15 + 48 + 54 + 48 + 48 \right)$. This number of the dimensionality reduction was decided empirically in the original algorithm.

We used a generative Gaussian mixture model (GMM) to fit the distribution of features $x \in \mathbb{R}^D$ extracted from the video, and determined the GMM parameters like mixture weight $\omega_k$, mean vector $\mu_k$, and standard-deviation vector $\sigma_k$ to best fit the features. Then the FVs were calculated from the fitted GMM models (Perronnin, Sánchez, & Mensink, 2010; Sánchez, Perronnin, Mensink, & Verbeek, 2013). In this experiment, we used $K = 256$ Gaussians to represent the trajectory features. The final dimensionality of the FV is $2 \times D \times K$, where $D$ is the dimensionality of the descriptor (i.e., $15 + 48 + 54 + 48 + 48$) and $K$ is the number of GMM components (i.e., 256). Finally, we applied power and $L_2$ normalization to the FVs, as in Wang et al. (2013). To combine different types of descriptors, we concatenated the normalized FVs into a single long vector with the

<table>
<thead>
<tr>
<th>Model name</th>
<th>Stiffness levels</th>
<th>Data type</th>
<th>Size</th>
<th>Testing data 1</th>
<th>Stiffness levels</th>
<th>Data type</th>
<th>Size</th>
<th>Testing data 2</th>
<th>Stiffness levels</th>
<th>Data type</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>0.005, 0.01, 1, 5, 10, 25, 40, 80, 180, 300, 450</td>
<td>Original, cotton</td>
<td>66</td>
<td>Same as training</td>
<td>Original, silk</td>
<td>66</td>
<td>0.1, 110</td>
<td>Original, silk + cotton</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scrambled</td>
<td>0.005, 0.01, 1, 5, 10, 25, 40, 80, 180, 300, 450</td>
<td>Scrambled, cotton</td>
<td>66</td>
<td>Same as training</td>
<td>Scrambled, silk</td>
<td>66</td>
<td>0.1, 110</td>
<td>Scrambled, silk + cotton</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>—</td>
<td>Random points</td>
<td>66</td>
<td>—</td>
<td>Random points</td>
<td>66</td>
<td>—</td>
<td>Random points</td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Data sets for different models in the main testing. Notes: All training data come from Scene 1. All testing data come from Scene 2.
dimension of 54,528. This concatenated FV would be used as the input to our models. In the testing stage, we used the same coefficients for performing principal-components analysis and training the GMM as those in the training process. The computation of GMM and FV was done using the VLFeat package in MATLAB (Vedaldi & Fulkerson, 2010).

Results

In this section, we demonstrate the effectiveness of dense-trajectory features in predicting the perceptual scale of bending stiffness of cloth in videos. The results are summarized in Table 2 (Main test). In Figure 8A, we have plotted the predicted scale of stiffness from the regression model versus the ground-truth physical parameters and compared it with the perceptual scale obtained in Experiment 1a (silk texture in Scene 1) by human observers. This figure shows that the predicted scale and the perceptual scale are highly correlated ($R^2 = 0.81$). Thus, the regression model is able to differentiate cloths with different bending stiffnesses in the videos as well as humans can. To test whether the modeling method can be generalized to training data with other fabrics, we swapped the training and testing data, and trained the model with perceptual scales obtained with the silk fabric and tested on the data from the cotton fabric. The predicted scale and the perceptual scale are also highly correlated, indicating that our finding is likely to be generalized to other fabrics too.

To provide a baseline for evaluating the predictive performance of the regression model, we trained a

<table>
<thead>
<tr>
<th>Section name</th>
<th>Model name</th>
<th>$R^2$</th>
<th>Predicting error (SD)</th>
<th>Significance</th>
<th>$R^2$</th>
<th>Predicting error (SD)</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main test</td>
<td>Regression</td>
<td>0.81</td>
<td>0.11 (0.09)</td>
<td>$&lt; 0.001$</td>
<td></td>
<td>0.14 (0.13)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Scrambled</td>
<td>0.15</td>
<td>0.29 (0.14)</td>
<td>$&lt; 0.001$</td>
<td></td>
<td>0.43 (0.28)</td>
<td>$&lt; 0.005$</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>0.12</td>
<td>0.39 (0.25)</td>
<td></td>
<td></td>
<td>0.38 (0.23)</td>
<td></td>
</tr>
<tr>
<td>Validation 1</td>
<td>Combined regression</td>
<td>0.77</td>
<td>0.12 (0.10)</td>
<td>$&lt; 0.001$</td>
<td>0.84</td>
<td>0.11 (0.09)</td>
<td>$&lt; 0.005$</td>
</tr>
<tr>
<td></td>
<td>Combined scrambled</td>
<td>0.13</td>
<td>0.29 (0.20)</td>
<td></td>
<td>0.51</td>
<td>0.23 (0.11)</td>
<td></td>
</tr>
<tr>
<td>Validation 2</td>
<td>Regression</td>
<td>0.81</td>
<td>0.11 (0.09)</td>
<td>$&lt; 0.001$</td>
<td></td>
<td>—</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Two-frame regression</td>
<td>0.12</td>
<td>0.25 (0.24)</td>
<td></td>
<td></td>
<td>—</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Results summary of all tests. Notes: $R^2$ is calculated from the model prediction and the ground truth.
random model which utilized features that were randomly valued between 0 and 1. In contrast to the regression model, the prediction of the random model was poorly correlated with the perceptual scale ($R^2 = 0.12$; Figure 8B). We then did a paired $t$ test on the absolute predicting error ($|\tilde{y} - y|$) of the two models. Results revealed that the mean predicting error of the regression model ($M = 0.11, SD = 0.09$) was significantly lower than that of the random model ($M = 0.39, SD = 0.25$), $t(65) = 7.93, p < 0.0001$, demonstrating that the regression model trained with multiframe spatiotemporal information was able to predict perceptual scales of bending stiffness.

To further test the hypothesis that correct multi-frame motion is necessary in estimating stiffness from videos, we evaluated the performance of the scrambled model. Figure 8C shows that the prediction from the scrambled model is poorly correlated with the ground-truth perceptual scale ($R^2 = 0.15$).

To test whether the regression model is significantly better than the other two models, we did a one-way analysis of variance on the absolute predicting error ($|\tilde{y} - y|$). Results (Figure 8D) revealed significant differences among the regression model ($M = 0.11, SD = 0.09$), the scrambled model ($M = 0.29, SD = 0.14$), and the random model ($M = 0.39, SD = 0.25$), $F(2, 195) = 41.58, p < 0.0001$. A Bonferroni post hoc test revealed that the predicting error of the regression model was significantly smaller than that of the scrambled model, which was also smaller than that of the random model ($ps < 0.01$). Thus, the scrambled model still performs better than chance level. One possibility is that cues such as appearance and shape were preserved in the scrambled videos that could be indicative of the bending stiffness.

To test whether our model can be generalized to predict new stiffness values, we assessed the model predictions on two more stiffness levels that had never been seen during the training process: a soft cloth (bending stiffness = 0.1) and a stiff cloth (bending stiffness = 110; i.e., testing data 2). As shown in Table 3, the ground-truth perceptual scales for these two stiffness levels are 0.03 and 0.83, respectively. A Mann–
investigated how perceptual impressions are linked to the properties of cloth in more than one scene. Here, we trained another regression model (a combined regression model) with training data containing 110 cotton video clips from both Scene 1 (66 video clips) and Scene 2 (44 video clips). The training labels were the corresponding average perceptual scale for each scene (i.e., Figure 4C and Figure 4F, blue line). Because Scene 1 and Scene 2 differ in scene setups and wind forces, incorporating videos from both scenes will lead to different model input space and hence yield different models. The testing data contained 66 silk videos from Scene 1 and 22 silk videos from Scene 2. Table 4 summarizes the training and testing data sets for this validation. To evaluate the contribution of multiframe motion information, we trained another scrambled model (a combined scrambled model) using the same approach except that the training and testing data came from the scrambled videos instead of the original ones.

Results of this validation test are summarized in Table 2 (Validation 1). Figure 9 plots the comparison of predictions from the combined regression model and the combined scrambled model. The predictions from the combined regression model for the two scenes are plotted in Figure 9A and 9B together with the corresponding ground-truth perceptual scale (black line in Figure 9). Model predictions corresponded well with the perceptual scales in both Scene 1 ($R^2 = 0.77$) and Scene 2 ($R^2 = 0.84$). Similarly, predictions by the combined scrambled model are shown in Figure 9C and 9D. In this case, model predictions did not correlate well with the perceptual scales in either scene (Scene 1: $R^2 = 0.13$; Scene 2: $R^2 = 0.51$). Moreover, in both scenes the predicting error ($[\bar{y} - y]$) of the combined regression model (Scene 1: $M = 0.12$, $SD = 0.10$; Scene 2: $M = 0.11$, $SD = 0.09$) was smaller than that of the combined scrambled model (Scene 1: $M = 0.29$, $SD = 0.20$; Scene 2: $M = 0.23$, $SD = 0.11$). A paired $t$ test shows that the observed difference is significant—Scene 1: $t(65) = 6.25$, $p < 0.0001$; Scene 2: $t(21) = 3.28$, $p < 0.0005$.

These results were consistent with those of the computational modeling section, verifying that multiframe motion is important in estimating material properties of cloth in more than one scene.

**Further validation**

In this section, we aim to verify the findings from the computational modeling section that multiframe motion is necessary in predicting perceived stiffness in more than one dynamic scene. Here, we trained another regression model (a combined regression model) with training data containing 110 cotton video clips from both Scene 1 (66 video clips) and Scene 2 (44 video clips). The training labels were the corresponding average perceptual scale for each scene (i.e., Figure 4C and Figure 4F, blue line). Because Scene 1 and Scene 2 differ in scene setups and wind forces, incorporating videos from both scenes will lead to different model input space and hence yield different models. The testing data contained 66 silk videos from Scene 1 and 22 silk videos from Scene 2. Table 4 summarizes the training and testing data sets for this validation. To
physical variables. Using MLDS, we derived perceptual scales of cloth stiffness in two different dynamic wind scenes (free-corner and pinned-corner) and with two different textures (silk and cotton; Experiment 1a). We found that in both of the two dynamic scenes, the perceived bending stiffness of cloth was linearly correlated with log-adjusted physical bending stiffness, and optical properties (e.g., textures and thickness) did not influence humans’ perceptual scale of stiffness. This indicates that observers have a robust estimation of the bending stiffness of cloth under variation of external forces and optical appearance. In Experiment 1b, we investigated the effect of correct motion sequences on the perceptual scale of bending stiffness. Using the same scene, we randomly scrambled the frame sequences of the videos and discovered that the observers’ perceptual scales were much less correlated with the physical values. Together, the results of Experiments 1a and 1b illustrate that multiframe motion information is important for viewers to assess cloth stiffness in dynamic scenes.

Using computational modeling, we provided further evidence for the effect of multiframe motion information on estimating bending stiffness. Specifically, we trained a machine-learning model with only features extracted from the multiframe motion fields of the videos. The model predictions were highly correlated with the human perceptual scales, and the results could be generalized to new bending-stiffness values and new dynamic scenes. When multiframe motion information was removed, such that the model was trained and tested with scrambled videos, the model’s performance dropped dramatically. These findings were consistent with Experiment 1b, suggesting that multiframe motion information is robust in estimating stiffness of cloth in dynamic scenes.

### Multiframe motion information is robust in recovering mechanical properties of deformable objects

Motion information influences material perception in different ways. For example, specular motion facilitates 3-D shape estimation (Dövencioglu, Ben-Shahar, Barla, & Doerschner, 2017), frame-by-frame
optical flow is indicative of viscosity of liquids (Kawabe et al., 2015), motion pattern arising from contour and optical deformation is important for judging elasticity of jellylike objects (Kawabe & Nishida, 2016), head motion affects perception of glossiness (Sakano & Ando, 2010), and so on. In this article, we reveal the effect of multiframe motion information on estimating cloth stiffness. Our study is the first to explicitly test the hypothesis that multiframe temporal correlation is important in perception of mechanical properties. We believe that it is an important extension of previous findings as well as a new framework to test how motion information takes effect.

Although more experimental evidence is needed, motion appears to affect the estimation of both mechanical and optical properties. Specifically, relative motion between observers and objects seems to be critical in judging optical properties, such as glossiness (Sakano & Ando, 2010; Doerschner et al., 2011; Tani et al., 2013). The movements of the objects, in the form of either frame-by-frame motion or multiframe motion trajectory—or how the shape outline changes over time—are important in judging mechanical properties (Kawabe et al., 2015; Kawabe & Nishida, 2016; Schmidt et al., 2017). As to the estimation of cloth properties, optical properties seem to dominate categorical judgments (Aliaga et al., 2015), whereas motion information might be important in estimating mechanical properties. Recently, in the field of computer vision, increasing attention has been paid to recognizing cloth properties from videos. Bouman et al. (2013) developed an algorithm for predicting mechanical properties of a cloth from videos. They excluded the surface information, such as textures and colors, from the input for the model training. However, it is unknown whether their algorithm can be generalized to new dynamic scenes and whether multiframe motion information is included. Most recently, Yang et al. (2017) utilized the appearance changes of the moving cloth to categorize the fabrics. They combined the image-signal feature-extraction method (i.e., the convolutional neural network) with the temporal-sequence learning method (i.e., the long short term memory) to learn the mapping from visual input to material categorization. However, they did not explicitly test whether the model could be related to human perception. Though these studies provide additional evidence for the importance of dynamic information in understanding material properties, they did not specifically test the role of multiframe motion information in predicting human perception of mechanical properties.

In this article, we have provided direct perceptual and computational evidence for the important role of multiframe motion in estimating mechanical properties by comparing performances in original and scrambled videos. It might be argued that not only multiframe but also two-frame motion information is removed in scrambled videos. To make our findings more convincing, we used the same training method as in the computational section to train another regression model on motion descriptors extracted from two-frame dense motion trajectories (i.e., a two-frame regression model). The number of parameters in the two-frame and 15-frame models is the same. This is due to the fact that the number of Gaussian distributions in the GMM determines the dimension of the FVs, and we assigned the same number of Gaussian distributions (256) to capture the data in both the two-frame and 15-frame models. Table 4 (validation 2) summarizes the data sets for this validation test, and the results are shown in Table 2 (Validation 2). Figure 10 shows that when compared to the model trained with motion information extracted from 15 consecutive frames (Figure 10A), the model that was trained with two-frame...
motion information (Figure 10B) not only did much worse in predicting human perceptual scale ($R^2 = 0.81$; two-frame: $R^2 = 0.12$) but also yielded significantly larger predicting errors (15-frame: $M = 0.11$, $SD = 0.09$; two-frame: $M = 0.25$, $SD = 0.24$), $t(65) = 4.52$, $p < 0.0001$. This result shows that two-frame motion information is not sufficient for predicting stiffness of cloth, thereby demonstrating the important role of long-range motion information. However, trajectories longer than 15 frames are not guaranteed to improve the performance. Specifically, we find that the performance of the machine-learning model increases as a function of the number of sampled frames up to $L = 16$ frames (i.e., 2, 4, 8, 16). But trajectories longer than 15 frames (e.g., 32) decrease the performance. This is consistent with the findings in the original dense-trajectory article (Wang et al., 2011). As discussed by those authors, this might be because longer frames will lead to a higher chance to drift from the initial position during the tracking process or to cross shot boundaries.

We have shown the limitation of motion features computed from two consecutive frames. To illustrate the importance of tracking multiple frames from videos in the dense trajectory, we did another test where we simply increased the interval of sampling frames. First, we sampled every three frames (0, 3, 6, 9, etc.) for the two-frame dense-trajectory features and used these to train the model (long-interval two-frame dense-trajectory model). Figure 11A compares the results of this model with those of the consecutive two-frame dense-trajectory model shown in Figure 10B. We found that the predictions from sampling every three frames for the two-frame model became worse ($R^2 = 0.05$ vs. 0.12).

Second, to address the question whether dense sampling over an extended number of frames was informative, we sampled over a 15-frame period but only every three frames (0, 3, 6, 9, etc.), effectively sampling over a 43-frame period but less densely. We then compared the predictions of this model to those of the original 15-frame dense-trajectory model shown in Figure 10A. Figure 11B shows that the predictions from sampling every three frames for the 15-frame model also became worse than the 15-frame dense sampling ($R^2 = 0.45$ vs. 0.81). This finding indicates that merely providing information on longer spatiotemporal scales is insufficient, but that dense sampling of frames over longer periods improves performance.

The importance of individual motion features

To understand the importance of each of the dense-trajectory descriptors (i.e., trajectory shape, HOF, HOG, MBH), we trained additional models with an individual descriptor and calculated the $R^2$ of the predicted scale by each model with the human perceptual scale. We find that trajectory shape ($R^2 = 0.50$) and HOF ($R^2 = 0.57$) are of equal importance, each accounting for more than 50% of the variance in the human perceptual scale. By contrast, both HOG ($R^2 = 0.35$) and MBH ($R^2 = 0.35$) are poor predictors of human perceptual scales. This is also true when the model predicts new stiffness values that it did not see during the training. HOF yielded the lowest predicting error ($M = 0.22$, $SD = 0.16$), which was slightly better than trajectory shape ($M = 0.25$, $SD = 0.05$). In contrast, the predicting errors of both HOG ($M = 0.31$,
$SD = 0.05$) and MBH ($M = 0.32$, $SD = 0.24$) were much higher. Together, these analyses reveal that trajectory shape and HOF are considerably more important than HOG and MBH in estimating cloth stiffness.

These results are in line with our main findings because among the four dense-trajectory descriptors, trajectory shape and HOF are the main descriptors of the local motion information, while HOG mostly captures the appearance, even though HOF is also affected by spatial information since it is restricted by how the interest points are sampled. The contribution of MBH might be underestimated in the current study, due to the fact that it mainly encodes the effect of camera motion, and the camera position is fixed in our videos. We believe camera motion would be inevitable when estimating cloth properties from real videos (e.g., video of a fashion show); thus, future studies that include camera motions might find the MBH feature to be more relevant.

Influence of optical properties on perception of mechanical properties

There have been a few recent works addressing the influence of optical properties on material perception, such as viscosity of liquids and classification of cloth (Aliaga et al., 2015; Paulun et al., 2015; Van Assen & Fleming, 2016; Xiao, Bi, Jia, Wei, & Adelson, 2016). Our goal was to evaluate the role of multiframe motion in the perception of mechanical properties, but this does not exclude the role of appearance. In fact, dense trajectories encode both motion and appearance information (e.g., the HOG feature). In addition, our rendered cloth samples only contain two types of appearance: silk (shiny, smooth, and thin) and cotton (matte, rough, and thick). Our data showed a small but significant effect of appearance on the average values of the perceptual scale, indicating that on average, observers perceive the silk cloth to be more flexible than the cotton (Figure 4). We believe the optical properties still dominate cloth categorization, but motion plays an equal, if not more important, role in estimating mechanical properties. Hence, our results are largely consistent with previous work on the influence of optical properties on material perception.

Humans can use different cues under different contexts

Humans are able to estimate mechanical properties of objects under variation of shape, size, optical appearance, and external forces (Bi & Xiao, 2016; Schmidt et al., 2017). Here we provide additional evidence for this suggestion. In Experiments 1a and 1b, we found that all observers performed very well ($R^2 > 0.8$) and there was little individual difference (Figures 4A, 4B, 4D, 4E, and 5B), indicating that observers could successfully use the motion information for estimation. In Experiment 1b, when observers made judgments from scrambled videos, there were some individual differences in their performance (Figure 5A). This could be due to the fact that different observers used different cues when the motion information was absent.

Comparing the model predictions with human perceptual scales under the scrambled video condition supported these observations. When multiframe motion was removed, the regression model performed no better than chance level (Figure 8C vs. 8B). By contrast, although the observers’ performance dropped dramatically, they could still distinguish the stiffest fabric from the most flexible one (Figure 5A), suggesting that observers potentially use other cues such as shape outline or appearance for judgment.

In addition to the availability of the image cues, the choice of tasks might affect the cues that humans use for material perception. For example, previous work indicates that observers predominately use optical cues (textures, glossiness, colorfulness, etc.) for material-categorization tasks (Fleming, Wiebel, & Gegenfurtner, 2013; Aliaga et al., 2015). In addition to the current study, others have also found that motion information is more important for estimation of mechanical properties (Bi & Xiao, 2016; Kawabe & Nishida, 2016; Yang et al., 2017). Future studies should evaluate the interactions between task and image cues for material perception.

Intuitive physics and multiframe motion information

Recent research has proposed that people reason about complex environments using approximate and probabilistic mental simulations of physical dynamics (Hamrick, Battaglia, & Tenenbaum, 2011; Battaglia, Hamrick, & Tenenbaum, 2013; Hamrick et al., 2016; Kubricht et al., 2016; Kubricht, Holyoak, & Lu, 2017). Even though we did not explicitly test the model of intuitive physics, we found evidence that viewers could use multiframe motion to infer mechanical properties of cloth. In addition, we find that such inference is robust across different scene setups and different wind forces. It is possible that multiframe motion cues are diagnostic of the causal relation between the object’s shape deformation and the applied force. During our experiments, observers could combine the low-level image statistics through learning and exposure (learning based) and their prior belief of the noisy generative physics (knowledge based) to make an inference of cloth mechanical properties. Future experiments and models are needed to test whether it is possible for humans to reason about the outcome of deformable...
objects in dynamic scenes they have not seen before using a probabilistic simulation model, and whether such a model is affected by different temporal parameters (e.g., the length of the video).

Using computer-vision methods to understand human perception

It is extremely difficult to create psychophysical stimuli that isolate motion information for cloth perception. Kawabe et al. (2015) used the noise videos simulated by optical-flow fields to isolate the motion information for liquids. This method might not be suitable for creating cloth stimuli, however, because the fine folds and creases would not be revealed in such noise simulations, and thus the stimuli will appear unnatural to observers. In the current study, we used machine learning as an alternative approach to examine the influence of multiframe motion information. The regression model was trained with only the dense-trajectory descriptors, which can capture the motion information in the videos efficiently and outperform the state-of-the-art approaches, at least in action recognition (for the evaluations, see Wang et al., 2011, section 5.2). Results showed that our model does well at predicting the human perceptual scales, and moreover, it can predict the situations when humans failed, such as when the video frames were scrambled.

Our results demonstrate that combining machine learning and human perception is a promising method to understand which image features humans utilize in estimating material properties across variations of scene setups. The recent advances in using deep neural networks trained to learn physical properties from videos suggest that it is possible to visualize features in different layers of the neural networks for a variety of tasks (Zeiler & Fergus, 2014). Recent work has addressed the robustness of the two-stream ConvNet for action recognition (Simonyan & Zisserman, 2014; Feichtenhofer, Pinz, & Zisserman, 2016), where one stream processes spatial information while the other deals with motion information. These studies show that motion provides critical additional information for visual recognition. Future studies can benefit from these methods to understand the contribution of optical properties and dynamics in material perception.

Conclusion

This article reveals that humans can recover the scale of bending stiffness of cloth from dynamic videos. We also find that optical appearance (e.g., textures, thickness, and roughness) and types of external forces do not influence observers’ sensitivity to the differences in stiffness. Most importantly, this article is the first to directly demonstrate that multiframe motion information is important for both humans and machines to estimate cloth stiffness. The methods of combining human perceptual studies and machine learning used here provide a successful paradigm of evaluating image cues on material perception.

Keywords: multiframe motion, material properties, machine learning, rendering, psychophysics, cloth perception

Acknowledgments

Commercial relationships: none.
Corresponding author: Wenyan Bi.
Email: wb1918a@student.american.edu.
Address: Department of Computer Science, American University, Washington DC, USA.

References


motion and the perception of surface material. *Current Biology, 21*(23), 2010–2016.


