Published in final edited form as:
Int J Comput Vis. 2013 July 1; 103(3): 348–371. doi:10.1007/s11263-013-0609-0.

Recognizing Materials using Perceptually Inspired Features

Lavanya Sharan,
Disney Research, Pittsburgh, 4720 Forbes Avenue, Lower Level, Suite 110, Pittsburgh, PA 15213, lavanya@disneyresearch.com

Ce Liu,
Microsoft Research New England, One Memorial Drive, Cambridge, MA 02142, celiu@microsoft.com

Ruth Rosenholtz, and
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, rruth@mit.edu

Edward H. Adelson
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, adelson@csail.mit.edu

Abstract

Our world consists not only of objects and scenes but also of materials of various kinds. Being able to recognize the materials that surround us (e.g., plastic, glass, concrete) is important for humans as well as for computer vision systems. Unfortunately, materials have received little attention in the visual recognition literature, and very few computer vision systems have been designed specifically to recognize materials. In this paper, we present a system for recognizing material categories from single images. We propose a set of low and mid-level image features that are based on studies of human material recognition, and we combine these features using an SVM classifier. Our system outperforms a state-of-the-art system [Varma and Zisserman, 2009] on a challenging database of real-world material categories [Sharan et al., 2009]. When the performance of our system is compared directly to that of human observers, humans outperform our system quite easily. However, when we account for the local nature of our image features and the surface properties they measure (e.g., color, texture, local shape), our system rivals human performance. We suggest that future progress in material recognition will come from: (1) a deeper understanding of the role of non-local surface properties (e.g., extended highlights, object identity); and (2) efforts to model such non-local surface properties in images.

Keywords

material recognition; material classification; texture classification; Mechanical Turk; perception

1 Introduction

It is easy for us to distinguish lustrous metal from shiny plastic, crumpled paper from knotted wood, and translucent glass from wax. This ability to visually discriminate and identify materials is known as material recognition, and it is important for interacting with surfaces in the real world. For example, we can tell if a banana is ripe, if the handles of a paper bag are strong, and if a patch of sidewalk is icy merely by looking. It is valuable to build computer vision systems that can make similar inferences about our world [Adelson, 2001]. Systems that can visually identify what a surface is made of can be useful in a number of scenarios: robotic manipulation, robotic navigation, assisted driving, visual quality assessments in manufacturing, etc. In this work, we have taken the first step towards building such systems by considering the problem of recognizing high-level material categories (e.g., paper, plastic) from single, uncalibrated images.

Previous work on material appearance has focused on two main themes: the modeling of surface reflectance properties [Sato et al., 1997, Marschner et al., 1999, Yu et al., 1999, Debevec et al., 2000, Matusik et al., 2000, Tominaga and Tanaka, 2000, Boivin and Gagalowicz, 2001, Nishino et al., 2001, Ramamoorthi and Hanrahan, 2001, Cula et al., 2004, Debevec et al., 2004, Marschner et al., 2005, Pont and Koenderink, 2005, Romeiro et al., 2008, Romeiro and Zickler, 2010] and the recognition of 3-D textures [Dana et al., 1999, Dana and Nayar, 1998, Cula and Dana, 2004a, Nillius and Eklundh, 2004, Caputo et al., 2005, Varma and Zisserman, 2005, Caputo et al., 2007, Varma and Zisserman, 2009]. Models of surface reflectance such as the bidirectional reflectance distribution function (BRDF) [Nicodemus, 1965] and its variants [Dana et al., 1999, Jensen et al., 2001] are popular in computer graphics because they allow for convincing simulations of real-world materials (e.g., wood [Marschner et al., 2005] and human skin [Marschner et al., 1999, Debevec et al., 2000]). These models describe how light interacts with a surface as a function of the illumination and viewing angles, and they can be estimated from images of a surface although not from a single, uncalibrated image [Ramamoorthi and Hanrahan, 2001, Marschner et al., 1999]. Surface reflectance properties are often correlated with material categories. For example, wooden surfaces tend to be brown, and metallic surfaces tend to be shiny. However, surfaces that belong to different material categories can exhibit similar reflectance properties (e.g., plastic, glass, and wax can all be translucent; see Figure 1a), and therefore, information about surface reflectance does not necessarily determine the material category.

Textures, both 2-D and 3-D [Pont and Koenderink, 2005], are an important component of material appearance. Wooden surfaces tend to have textures that are quite distinct from those of polished stones or printed fabrics. However, as illustrated in Figure 1b, surfaces made of different materials can exhibit similar textures, and therefore, systems designed for texture recognition [Leung and Malik, 2001, Varma and Zisserman, 2009] may not be adequate for material category recognition.

It is tempting to reason that the problem of recognizing material categories can be reduced to the well-studied problem of recognizing object categories. Coffee mugs tend to be made of ceramic, cars of metal and glass, and chairs of wood. If object recognition methods can be used to detect object categories (e.g., coffee mugs) in images, then inferring the material category (e.g., ceramic) should be straight-forward. Despite the promising correlations between object categories and material categories, it is important to note that their relationship is not one-to-one, especially for man-made objects. A given category of man-made objects can be made of different materials (Figure 2a), and different categories of man-made objects can be made of the same material (Figure 2b). In fact, most object recognition systems [Belongie et al., 2002, Dalal and Triggs, 2005, Liu et al., 2009] that consider man-made objects tend to either rely on material-invariant features such as shape or ignore material information altogether.

As existing techniques of reflectance modeling, texture recognition, and object recognition cannot be directly applied to our problem of material category recognition, we start by gathering the ingredients of a familiar recipe in visual recognition: (i) an annotated image database; (ii) diagnostic image features; and (iii) a classifier. In this work, we introduce the Flickr Materials Database (FMD) [Sharan et al., 2009] to the computer vision community, study human material recognition on Amazon Mechanical Turk, design and employ features based on studies of human perception, combine features in both a generative model and a discriminative model to categorize images, and systematically compare the performance of our recognition system to human performance.

We chose the Flickr Materials Database [Sharan et al., 2009] because most databases popular in the computer vision community fail to capture the diversity of real-world material appearances. These databases are either instance databases (e.g., CURET [Dana et al., 1999]) or texture category databases with very few samples per category (e.g., KTH-TIPS2 [Caputo et al., 2005]). The high recognition rates achieved for these databases (e.g., > 95% texture classification accuracy for CURET [Varma and Zisserman, 2009]) highlight the need for more challenging databases. FMD was developed with the specific goal of capturing the appearance variations of real-world materials, and by including a diverse selection of samples in each category, FMD avoids the poor intra-class variation found in earlier databases. As shown in Figure 2b, FMD images contain surfaces that belong to one of ten common material categories: Fabric, Foliage, Glass, Leather, Metal, Paper, Plastic, Stone, Water, and Wood. Each category includes one hundred color photographs (50 close-up views and 50 object-level views) of 512×384 pixel resolution. FMD was originally developed to study the human perception of materials, and as we will see later in the paper, human performance on FMD serves as a challenging benchmark for computer vision algorithms.

Unlike the case of objects or scenes, it is difficult to find image features that can reliably distinguish material categories from one another. Consider Figure 2b; surfaces vary in their size, 3-D shape, color, reflectance, texture, and object information both within and across material categories. Given these variations in appearance, it is not obvious which features are diagnostic of the material category. Our strategy has been to: (i) conduct perceptual studies to understand the types of low and mid-level image information that can be used to characterize materials [Sharan et al., 2009]; and (ii) use the results of our studies to propose a set of image features. In addition to well-known features such as color, jet, and SIFT [Koenderink and van Doorn, 1987, Lowe, 2004], we find that certain new features (e.g., histogram of oriented gradient features measured along and perpendicular to strong edges in an image) are useful for material recognition.

We evaluated two different classifiers in this paper: a generative latent Dirichlet allocation (LDA) model [Blei et al., 2003] and a discriminative SVM classifier [Burges, 1998]. For both types of classifiers, we quantized our features into dictionaries, concatenated dictionaries for different features, and converted images into “bags of words”. In the generative case, the LDA model learned the clusters of visual words that characterize different material categories. In the discriminative case, the SVM classifier learned the hyperplanes that separate different material categories in the space defined by the shared dictionary of visual words. Both classifiers performed reasonably well on the challenging Flickr Materials Database, and they outperformed a state-of-the-art material recognition system [Varma and Zisserman, 2009]. The SVM classifier performed better than the LDA model, and therefore, we selected the SVM classifier for our recognition system. To avoid confusion, we will use “our system” or “our SVM-based system” to denote the SVM classifier trained with our features. When we discuss the LDA model, we will use “our aLDA model” or “our LDA-based system” to denote an augmented LDA model trained on our features. The augmentation of the standard LDA model [Blei et al., 2003] takes two forms: (i) we concatenate dictionaries for different features; and (ii) we learn the optimal combination of features by maximizing the recognition rate.

As a final step, we compared the performance of our SVM-based system to the that of human observers in various ways. First, we compared the categorization accuracy on the original images, and we show that human performance on the original FMD images serves as a challenging benchmark for computer vision systems. Next, we compared the categorization accuracy on images that were modified to emphasize either “non-local features” such as outline shape and object identity (which are not explicitly modeled in this paper) or “local features” such as color, texture, and local shape (which are explicitly modeled in this paper).[1] When non-local features are emphasized, at the expense of local features, the performance of our system lags human performance by a large margin. When local features are emphasized, at the expense of non-local features, the performance of our system is comparable to human performance. Based on these results, we can conclude that humans are better at utilizing non-local surface information for material recognition than our system. On the other hand, our computer vision system is nearly as effective as humans at utilizing local surface information for material recognition.

2 Related work

We now review prior work from the fields of computer graphics, computer vision, and human vision. This includes attempts to: (i) characterize real-world materials in computer graphics; (ii) recognize textures, materials, and surface reflectance properties in computer vision; and (iii) understand the perception of materials in human vision.

2.1 BRDF estimation

In the field of computer graphics, the desire to create convincing simulations of materials such as skin, hair, and fabric has led to several formalizations of the reflectance properties of materials. These formalizations are of great importance because they allow for realistic depictions of materials in synthetic scenes. A popular formalization, the bidirectional reflectance distribution function (BRDF) [Nicodemus, 1965], specifies the amount of light reflected at a given point of a surface for any combination of incidence and reflection angles. As BRDFs are functions of four or more variables, BRDF specifications can turn into large and unwieldy lookup tables. For this reason, BRDFs are often approximated by parametric models to enable efficient rendering [He et al., 1991, Phong, 1975, Ward, 1992]. Parametric BRDF models can represent the reflectance properties of several common materials effectively (e.g., plaster and concrete [Koenderink et al., 1999, Oren and Nayar, 1995]), but they cannot capture the full range of real-world reflectance phenomena.

As an alternative to parametric BRDF models, empirically measured BRDFs are used when renderings of complex, real-world materials are desired. The BRDF of a surface of interest (e.g., a copper sphere [Matusik et al., 2000]) is measured in the laboratory using a specialized setup that consists of light sources, light-measuring devices like cameras, and mechanical components that allow BRDF measurements for a range of angles. A number of techniques have been developed to recover the BRDF of a surface from photographs that are acquired in such setups [Boivin and Gagalowicz, 2001, Debevec et al., 2000, 2004, Marschner et al., 1999, Matusik et al., 2000, Nishino et al., 2001, Ramamoorthi and Hanrahan, 2001, Sato et al., 1997, Tominaga and Tanaka, 2000, Yu et al., 1999]. These techniques typically assume some prior knowledge of the illumination conditions, 3-D shape, and material properties of the surfaces being imaged, and on the basis of such prior knowledge, they are able to estimate BRDF values from the image data.

[1] In this paper, we use the terms “local features” and “non-local features” relative to the size of the surface of interest and *not* the size of the image. The images we will consider in this paper correspond to the spatial scale depicted in Figure 3b. For this scale, features such as color, texture, and local shape are considered local features, whereas features such as outline shape and object identity are considered non-local features.

The work on image-based BRDF estimation is relevant, although not directly applicable, to our problem. BRDF estimation techniques try to isolate reflectance-relevant, and therefore, material-relevant information in images. However, these techniques ignore texture and geometric shape properties that can also contribute to material appearance. Moreover, BRDF estimation as a means to recover the material category is not a feasible strategy. For single images acquired in unknown conditions like the ones in Figure 1b, estimating the BRDF of a surface is nearly impossible [Ramamoorthi and Hanrahan, 2001, Marschner et al., 1999] without simplifying assumptions about the 3-D shape or the material properties of the surface [Tominaga and Tanaka, 2000, Boivin and Gagalowicz, 2001, Romeiro et al., 2008, Romeiro and Zickler, 2010]. In this work, we want to avoid *estimating* a large number of physical parameters as an intermediate step to material category recognition. Instead, we want to *measure* image features (a large number of them, if necessary) that can capture information relevant to the high-level material category.

2.2 3-D texture recognition

Textures result either from variations in surface reflectance (i.e., wallpaper textures) or from variations in fine-scale surface geometry (i.e., 3-D textures) [Dana and Nayar, 1998, Koenderink et al., 1999]. A surface is considered a 3-D texture when its surface roughness can be resolved by the human eye or by a camera. Consider Figure 2b; the two surfaces in the Stone category and the first surface in the Wood category are instances of 3-D textures. Dana *et al.* were the first to systematically study 3-D textures. They created CURET [Dana et al., 1999], a widely used image database of 3-D textures. CURET consists of photographs of 61 real-world surfaces (e.g., crumpled aluminum foil, a lettuce leaf) acquired under a variety of illumination and viewing conditions. Dana *et al.* modeled the appearance of these surfaces using the bidirectional texture function (BTF) [Dana et al., 1999]. Like the BRDF, the BTF is a high-dimensional representation of surface appearance, and it is mainly used for rendering purposes in computer graphics.

In addition to modeling 3-D textures [Dana and Nayar, 1998, Dana et al., 1999, Pont and Koenderink, 2005], there has been interest in recognizing 3-D textures from images [Cula and Dana, 2004b, Nillius and Eklundh, 2004, Caputo et al., 2005, Varma and Zisserman, 2005, 2009]. Filter-based and patch-based image features have been employed to recognize instances of 3-D textures with great success. For example, Cula *et al.*'s system, which uses multi-scale Gaussian and Difference of Gaussians (DoG) filtering [Cula and Dana, 2004b], and Varma and Zisserman's system, which uses image patches as small as 5×5 pixels [Varma and Zisserman, 2009], both achieve accuracies greater than 95% at distinguishing CURET surfaces from each other. Caputo *et al.* have argued that the choice of classifier and the choice of image database influence recognition performance much more than the choice of image features [Cula and Dana, 2004b]. Their SVM-based recognition system is equally successful for a variety of image features (~90% accuracy), and their KTH-TIPS2 database, unlike CURET, contains multiple examples in each 3-D texture category, which makes it a somewhat more challenging benchmark for 3-D texture recognition.

It is important to understand the work on 3-D texture recognition and its connection to the problem of material category recognition. Most 3-D texture recognition techniques were developed for the CURET database, and as a consequence, they have focused on recognizing *instances* rather than classes of 3-D textures. The CURET database contains 61 real-world surfaces, and the 200+ photographs of each surface constitute a unique 3-D texture category. For example, there is one sponge in the CURET database, and while that specific sponge has been imaged extensively, sponge-like surfaces as a 3-D texture *class* are poorly represented in CURET. This aspect of the CURET database is not surprising as it was developed for rendering purposes rather than for testing texture recognition algorithms. Other 3-D texture databases such as the Microsoft Textile Database [Savarese and Criminisi, 2004] (one surface per 3-D texture category) and KTH-TIPS2 [Caputo et al., 2005] (four surfaces per 3-D texture category) are similar in structure to CURET and, therefore, lack intra-class variations in 3-D texture appearance.

The emphasis on recognizing instances rather than classes of 3-D textures has meant that most 3-D texture recognition techniques cannot handle the diversity of real-world material appearances, of the sort shown in Figure 2. Materials such as fabric, stone, and wood can exhibit large variations in their 3-D texture appearance, and for our purposes, we require a system that can ignore such texture variations within material categories (e.g., wool vs. silk). As far as we are aware, 3-D texture recognition techniques that are sensitive to texture variations across material categories (e.g., fabric vs. stone) are not specific enough to ignore texture variations within material categories. For this reason, the work on 3-D texture recognition cannot be directly applied to our problem of material category recognition. Moreover, many of the images in the Flickr Materials Database were acquired at a scale where the 3-D texture cannot be resolved, and 3-D texture recognition is of limited value for these images. Finally, it is important to note that the Flickr Materials Database (FMD) is a far more challenging database for visual recognition algorithms than CURET. Consider the plots in Figure 4. The 61 CURET categories are more separable than the 10 FMD categories, even in a two-dimensional projection of Varma and Zisserman's patch-based feature space [Varma and Zisserman, 2009].

2.3 Recognizing specific reflectance properties

In addition to the work on image-based BRDF estimation that was described in Section 2.1, there have been attempts to recognize specific aspects of the BRDF such as albedo and surface gloss. Dror *et al.* used histogram statistics of raw pixels to classify photographs of spheres as white, grey, shiny, matte, and so on [Dror et al., 2001]. Sharan *et al.* employed histogram statistics of raw pixels and filter outputs to estimate the albedo of real-world surfaces such as stucco [Motoyoshi et al., 2007, Sharan et al., 2008]. Materials such as skin and glass have been identified in images [Forsyth and Fleck, 1999, McHenry and Ponce, 2005, McHenry et al., 2005, Fritz et al., 2009] by recognizing certain reflectance properties associated with those materials (e.g., flesh-color or transparency). Khan *et al.* developed an image editing method to alter the reflectance properties of objects in precise ways [Khan et al., 2006]. Many of these techniques for measuring specific aspects of reflectance rely on restrictive assumptions (e.g., surface shape [Dror et al., 2001] or imaging conditions [Motoyoshi et al., 2007, Sharan et al., 2008]), and as such, they cannot be applied easily to the images in the Flickr Materials Database. In addition, as demonstrated in Figure 1a, material category recognition is not simply reflectance recognition; knowing the albedo or gloss properties of a surface may not constrain the material category of the surface.

2.4 Human material recognition

Studies of human material recognition have focused on the perception of specific aspects of surface reflectance such as color and albedo [Bloj et al., 1999, Boyaci et al., 2003, Brainard et al., 2003, Gilchrist et al., 1999, Maloney and Yang, 2003]. While most of this work has considered simple stimuli (e.g., gray matte squares), recent work has made use of stimuli that are more representative of the real world. Photographs of real-world surfaces [Robilotto and Zaidi, 2004, Motoyoshi et al., 2007, Sharan et al., 2008] as well as synthetic images created by sophisticated graphics software [Pellacini et al., 2000, Fleming et al., 2003, Nishida and Shinya, 1998, Todd et al., 2004, Ho et al., 2008, Xiao and Brainard, 2008] have been employed to identify the cues underlying real-world material recognition. Nishida and Shinya [Nishida and Shinya, 1998] and others [Motoyoshi et al., 2007, Sharan et al., 2008] have shown that image-based information like the shape of the luminance histogram is correlated with judgments of diffuse and specular reflectance. Fleming *et al.* have argued that the nature of the illumination affects the ability to estimate surface gloss [Fleming et al., 2003]. Berzhanskaya *et al.* have shown the perception of surface gloss is not spatially uniform, and that it is influenced by the proximity to specular highlights [Berzhanskaya et al., 2005]. For translucent materials like jade and porcelain, cues such as the presence of specular highlights, coloring, and contrast relationships are believed to be important [Fleming and Bülthoff, 2005].

How can we relate these perceptual findings to computer vision, specifically to material category recognition? One hypothesis of human material recognition, known as 'inverse optics', suggests that the human visual system estimates the parameters of an internal model of the 3-D layout and illumination of a scene so as to be consistent with the 2-D images received at the eyes. Most image-based BRDF estimation techniques in computer vision and computer graphics can be viewed as examples of 'inverse optics'-like processing. A competing hypothesis of human material recognition argues that in real-world scenes, surface geometry, illumination distributions, and material properties are too complex and too uncertain for 'inverse optics' computations to be feasible. Instead, the visual system might use simple rules like those suggested by Gilchrist *et al.* for albedo computations [Gilchrist et al., 1999] or simple image-based information like orientation flows or statistics of luminance [Fleming and Bülthoff, 2005, Fleming et al., 2004, Motoyoshi et al., 2007, Nishida and Shinya, 1998, Sharan et al., 2008]. Techniques that have been developed for recognizing 3-D textures and surface reflectance properties [Dror et al., 2001, Varma and Zisserman, 2005, 2009] can be viewed as examples of the simpler processing advocated by this second school of thought.

The work in human vision that is most relevant to this paper is our work on human material categorization [Sharan et al., 2009]. We conducted perceptual studies using the images in the Flickr Materials Database (FMD) and established that human observers can recognize high-level material categories accurately and quickly. By presenting the original FMD images and their modified versions (e.g., images with only color or shape information) to observers, we showed that recognition performance cannot be explained by a single cue such as surface color, global surface shape, or surface texture. Details of these studies can be found elsewhere [Sharan et al., 2009]. Our perceptual findings argue for a computational strategy that utilizes multiple cues (e.g, color, shape, and texture) for material category recognition. In order to implement such a strategy in a computer vision system, one has to know which image features are important and how to combine them. Our work on human material categorization examined the contribution of visual cues like color but not the utility of specific image features that are commonly used by computer vision algorithms. To test the utility of standard image feature types, we conducted a new set of perceptual studies that are described in Section 3.

2.5 Material category recognition systems

In an early version of this work [Liu et al., 2010], we had used only the LDA-based classifier, and we had reported an accuracy of 44.6% at categorizing FMD images. In this paper, we report categorization accuracies for both the LDA-based classifier and the SVM classifier. In addition, we present the results of training our classifiers on original FMD images and testing them on distorted FMD images, in an effort to understand the utility of various features and to relate the performance of our system to human perception.

There has been some followup work since [Liu et al., 2010]. In particular, Hu *et al.* presented a system for material category recognition [Hu et al., 2011] in which kernel descriptors that measure color, shape, gradients, variance of gradient orientation, and variance of gradient magnitude were used as features. Large-margin distance learning was used to reduce the dimensionality of the descriptors, and efficient match kernels were used to compute image-level features for SVM classification. Hu *et al.*'s system achieves 54% accuracy on FMD images, averaged across 5 splits of FMD images into training and test sets. We will show later in Section 6 that our standard SVM-based system achieves higher accuracies (55.6% for unmasked images, 57.1% for masked images, averaged across 14 random splits) than Hu *et al.*'s system.

3 Studying human material perception using Mechanical Turk

In order to understand which image features are useful for recognizing material categories, we turned to human perception for answers. If a particular image feature is useful for distinguishing, say, fabric from paper, then it is *likely* that humans will make use of the information that is contained in that image feature. By presenting a variety of image features (in visual form) to human observers and measuring their accuracy at material category recognition, we can identify the image feature types that are correlated with human responses, and therefore, the image feature types that are *likely* to be useful in a computer vision system for recognizing material categories.

There are two types of image features that are popular in visual recognition systems – features that focus on object properties and features that measure local image properties.[2] Similar to Sharan *et al.* [Sharan et al., 2009], we used the original FMD images and distorted them in ways that emphasized these two types of image features. We then asked users on Amazon's Mechanical Turk website to categorize these distorted images into the ten FMD categories. The conditions used in our Mechanical Turk studies differ from those of Sharan *et al.* as all of our distorted images were created by automatic methods (e.g., by performing bilateral filtering) rather than by hand, and the perceptual data was obtained from a large number of Mechanical Turk participants (Turkers) instead of a small set of laboratory participants.

To assess object-based image features, we created images that emphasize global object shape and object identity information and minimize color and reflectance information. We used two types of processing to obtain such images. First, we performed bilateral filtering (surface blur filter in Adobe Photoshop, radius = 5 pixels, threshold = 50) to remove surface information while preserving strong edges in images. Next, we subtracted the bilateral filtered results from grayscale versions of the original images to emphasize details of surface structure, an operation similar to high-pass filtering. Finally, pixels outside the region of interest were suppressed using binary masks (that are available with FMD) for both the bilateral filtered and "high-pass filtered" results. Examples of these two types of processing are shown in Figures 5b & c. One can see that few color and reflectance details remain after the processing steps. The highlight on the plastic toy is nearly gone (Figures 5b & c), and the surface structure of the wine glasses is made more visible (Figure 5c.)

To assess local image features, we created images that emphasize local surface information and minimize global surface information. Like Sharan *et al.* [Sharan et al., 2009], we used a nonparametric texture synthesis algorithm [Efros and Freeman, 2001] to generate locally preserved but globally scrambled images from the material-relevant regions in the FMD images. We used two different window sizes, 15×15 and 30×30 pixels, as shown in Figures 5d & e. It is hard to identify any objects or large surfaces in these texture synthesized images even though, at a local scale, these images are nearly identical to the original images.

[2] For the spatial scales depicted in FMD images, object properties such as outline shape are "non-local" in nature. Meanwhile, local image properties such as color or texture can vary across the surface of interest, and hence, they are "local" in nature.

Images in all five experimental conditions, shown in Figure 5, were presented to the users of Amazon's Mechanical Turk website (Turkers). For each condition, the 1000 images in FMD were divided into 50 non-overlapping sets of 20 images each. Each Turker completed one set of images for one experimental condition. A total of 2,500 Turkers participated in our experiment, 500 per experimental condition and 10 per set. Instructions and sample images preceded each set. For example, Turkers were given the following instructions along with four pairs of sample images in the texture synthesized conditions (Figures 5d & e):

> The images presented to you were transformed from their original versions by scrambling the regions containing the object(s). For example, here are four original images and their transformed versions. Your task is to label the material category of the original image based on the transformed image given to you. Select your response from the choices provided to you. Remember, if you want to change your response for an image, you can always come to back to it using the navigation buttons.

Turkers were allowed as much time as needed to complete the task. We paid the Turkers $0.15 per set, which translated to an average hourly wage of $12 per hour. This hourly wage is comparable to that of laboratory studies (~ $10 per hour).

Results for all five conditions are shown in Figure 6. Chance performance corresponds to 10% accuracy in each plot. The best performance, not surprisingly, is with the original FMD images (84.9%). In the bilateral filtering condition, where only the strong edges in images are preserved, the performance averaged across categories drops to 65.3%. This performance implies a strong correlation between object identities and material categories; participants can guess the underlying material category even when they are given mostly object-relevant information. Performance is similar in the residual of bilateral filtering or the "high-pass" condition (64.8%) indicating that sometimes high-pass image information is sufficient for material category recognition. When object information is suppressed and only local image information is presented, there is a large decrease in categorization performance (38.7% for the 15×15 patches, 46.9% for 30×30; Figures 6d and e). This decrease in performance from the 84.9% recognition rate obtained on the original FMD images and the trend of improved performance on larger patch sizes indicates that humans use more than local image information for material recognition. On the other hand, the fact that the recognition rate in the globally scrambled conditions is greater than chance (10%) indicates that local image information plays a role in material recognition. Man-made materials (e.g., plastic) are harder to recognize than natural materials (e.g., stone) when global surface information is removed, possibly because the texture appearance of man-made surfaces is far more variable than that of natural surfaces.

In the next section, we describe the image features that were used in our system, many of which are influenced by the conclusions of our Mechanical Turk studies.

4 A proposed set of image features for material category recognition

We used a variety of features based on what we know about the human perception of materials, the physics of image formation, and successful recognition systems in computer vision. The results of our Mechanical Turk studies lead to specific guidelines for selecting and designing images features: (a) we should include features based on object shape and object identity; and (b) we should include local as well as non-local features. A different set of guidelines comes from the perspective of image formation. Once the camera and the surface of interest are fixed, the image of the surface is determined by the BRDF of the surface, the geometric surface shape including micro-structures, and the lighting on the surface. These factors can, to some extent, be estimated from images, which suggests the following additional guidelines: (c) we should use estimates of material-relevant factors (e.g., BRDF, micro-structures) as they can be useful for identifying material categories; and (d) we don't need to estimate factors unrelated to material properties (e.g., camera viewpoint, lighting). Ideally, the set of image features we choose should satisfy all of these guidelines. Practically, our strategy has been to try a mix of features that satisfy some of these guidelines, including standard features borrowed from the fields of object and texture recognition and a few new ones developed specifically for material recognition.

We used four groups of image features, each group designed to measure a different aspect of surface appearance. The four groups corresponded to: color and texture, micror-texture, outline shape, and reflectance. These groups consist of one to three features each and consider image regions of varying sizes, as illustrated in Table 1. Note that we did not design a feature group to measure object identity (e.g., chair) because half of the images in FMD are partial or close-up views, which makes it difficult to use object recognition techniques. In addition, FMD contains a large number of objects that are not typical for their category (e.g., a cat-shaped pillow), which makes the task of designing features that cover the range of objects in FMD very challenging.

Our selection of features is by no means exhaustive or final. Rather, our efforts to design features for material category recognition should be viewed as a first attempt at understanding which feature types are useful for recognizing materials. We will now describe the four feature groups that we designed and the reasons for including them.

4.1 Color and texture

Color is an important attribute of material appearance; wooden objects tend to be brown, leaves tend to be green, and plastics tend to have saturated colors. Color properties of surfaces can be diagnostic of the material category, so we used 3×3 RGB pixel patches as a color feature. Similarly, texture properties can be useful for distinguishing materials. Wooden surfaces tend to have characteristic textures that are instantly recognizable and different from those of polished stone or printed fabrics. We used two sets of features to capture texture information. The first set of features comprises the filter responses of an image through a set of multi-scale, multi-orientation Gabor filters, often called filter banks or jets [Koenderink and van Doorn, 1987]. Jet features have been used to recognize 3-D textures [Leung and Malik, 2001, Varma and Zisserman, 2009] by clustering to form "textons" and using the distribution of textons as a feature. We used Gabor filters of both cos and sin phases at 4 spatial scales (0.6, 1.2, 2, and 3) and 8 evenly spaced orientations to form a filter bank to obtain jet features. The second set of features comprises SIFT features [Lowe, 2004] that have been widely used in object and scene recognition to characterize the spatial and orientational distribution of local gradients in an image [Fei-Fei and Perona, 2005]. The SIFT descriptor is computed over a grid of 4×4 cells (8 orientation bins per cell), where a cell is a 4×4 pixel patch. As we do not use the spatial pyramid, the SIFT feature we use functions as a measure of the texture properties rather than the object properties in an image.

4.2 Micro-texture

Two surfaces that have the same BRDF can look very different if their surface micro-structures are not similar (e.g., if one is smooth and the other is rough). Tiny hairs on fabric surfaces, rounded edges of glass objects, and crinkles in leather surfaces add to the distinctive appearances of those surfaces. In order to extract information about surface micro-structure, we followed the idea of [Bae et al., 2006] of smoothing an image by bilateral filtering (σ_s=5, σ_r estimated automatically from image gradients) [Durand and Dorsey, 2002] and then using the residual image for further analysis. The process is illustrated in Figures 7 and 8. In Figure 7, images from three categories - glass, metal, and fabric - are presented along with the base and residual images obtained after bilateral filtering. The residuals of bilateral filtering reveal the variations in pixel intensity at a finer scale. For the fabric and metal examples in Figure 7, the residual contains surface micro-structure information whereas for glass, these variations are related to translucency. Although it is difficult to cleanly separate the contributions of surface micro-structure from those of surface reflectance, the residual often contains useful information about the material category. To characterize the information in these residual images, we applied the same analysis that was used to measure texture properties in the original images. We computed jet and SIFT features of residual images, and for clarity, we refer to them as micro-jet and micro-SIFT features. The parameters for the micro-jet and micro-SIFT features were the same as those for the jet and SIFT features respectively.

4.3 Outline shape

The outline shape of a surface and its material category are often related. For example, fabric and glass surfaces tend to have long, curved edges, while metallic surfaces tend to have straight edges and sharp corners. The outline shape of a surface can be estimated from an edge map. We used the Canny edge detector [Canny, 1986] to obtain edge maps by running the detector on the base images (i.e., the output of bilateral filtering) and trimming out the short edges. We used MATLAB's Canny edge detector with the larger threshold set to a function of 90% of luminance gradients and the smaller threshold set to 40% of the larger one. Examples of edge maps are shown in Figure 7c. To characterize the variations in the edge maps across material categories, we measured the curvature of edges in the edge map as a feature. The curvature feature was computed at every third edge point in the edge map and at three different scales (2, 8, and 16) as shown in Figure 9a.

4.4 Reflectance-based features

Glossiness and transparency are important cues for material recognition. Metals are usually shiny, whereas wooden surfaces are usually dull. Glass and water are translucent, whereas stones are often opaque. These reflectance properties sometimes manifest as distinctive intensity changes at the edges of surfaces [Fleming and Bülthoff, 2005]. To measure such changes, we used histogram of oriented gradients (HOG) features [Dalal and Triggs, 2005] for regions near strong edges in images, as shown in Figures 9b & c. We took slices of pixels normal to and along the edges in the images, computed the gradient at every pixel in those slices, divided the slices into 6 cells of size 3×3 pixels each, and quantized the oriented gradients in each cell into 12 angular bins. We call these composite features edge-slice and edge-ribbon respectively. Both edge-slice and edge-ribbon features were extracted at every edge point in the edge map.

We have described eight sets of features that can be useful for material category recognition: color, SIFT, jet, micro-SIFT, micro-jet, curvature, edge-slice, and edge-ribbon. Of these features, color, jet, and SIFT are *low-level* features that are computed directly on the original images, and they are often used for texture analysis. The remaining features, micro-SIFT, micro-jet, curvature, edge-slice, and edge-ribbon are *mid-level* features that rely on estimates of base images and edge maps. A priori, we did not know which of these features were best suited for material category recognition. To understand which features were useful, we combined our features in various ways and examined the recognition accuracy for those combinations. In the next two sections, we will describe and report the performance of a Bayesian learning framework and an SVM model that utilize the features described in this section.

5 Classifiers for material category recognition

Now that we have a selection of features, we want to combine them to build an effective material category recognition system. In this paper, we examine both generative and discriminative models for recognition. For the generative model, we extend the LDA framework [Blei et al., 2003] to select good features and learn per-class distributions for recognition. For the discriminative model, we use support vector machines (SVMs), which have proven useful for a wide range of applications, including the object detection and object recognition problems in computer vision. It is important to note here that the focus of this work is not designing the best-performing classifiers but exploring features for material category recognition. We will now describe how features are quantized into visual words, how visual words are modeled, and how an optimal combination of features is chosen using a greedy algorithm.

5.1 Feature quantization and concatenation

We used the standard k-means algorithm to cluster the instances of each feature to form visual words and create dictionaries. Suppose there are m features in the feature pool and m corresponding dictionaries, $\{D_i\}_{i=1}^{m}$. Each dictionary contains V_i codewords, i.e., $|D_i| = V_i$. The m features are quantized separately, so the words generated by the ith feature are $\left\{w_1^{(i)}, \cdots, w_{N_i}^{(i)}\right\}$ where $w_j^{(i)} \in \{1, 2, \cdots, V_i\}$ and N_i is the number of words. A document with m sets of words

$$\left\{w_1^{(1)}, \cdots, w_{N_1}^{(1)}\right\}, \left\{w_1^{(2)}, \cdots, w_{N_2}^{(2)}\right\}, \cdots, \left\{w_1^{(m)}, \cdots, w_{N_m}^{(m)}\right\} \quad (1)$$

can be augmented to form one set

$$\left\{w_1^{(1)}, \cdots, w_{N_1}^{(1)}, w_1^{(2)}+V_1, \cdots, w_{N_2}^{(2)}+V_1, \cdots, w_1^{(m)}+\Sigma_{i=1}^{m-1} V_i, \cdots, w_{N_m}^{(m)}+\Sigma_{i=1}^{m-1} V_i\right\} \quad (2)$$

using a joint dictionary $\mathbb{D}=\cup_i D_i$, $|\mathbb{D}|=\Sigma_{i=1}^{m} V_i$. In this way, we can transform multiple dictionaries into a single dictionary.

5.2 Augmented Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) [Blei et al., 2003] was developed to model the hierarchical structures of words. Details of the LDA framework can be found elsewhere [Blei et al., 2003, Fei-Fei and Perona, 2005], however, in order to be self-contained, we will now briefly describe LDA in the context of material recognition. As depicted in Figure 10, we randomly draw the material category $c \sim \mathrm{Mult}(c|\boldsymbol{\pi})$ where $\mathrm{Mult}(\cdot|\boldsymbol{\pi})$ is a multinomial distribution with parameter $\boldsymbol{\pi}$. Based on c, we select hyper-parameter α_c, based on which we draw $\theta \sim \mathrm{Dir}(\cdot|\alpha_c)$ where $\mathrm{Dir}(\cdot|\alpha_c)$ is a Dirichlet distribution with parameter α_c. θ has the following property: $\Sigma_{i=1}^{k} \theta_i=1$ where k is the number of elements in θ. From θ, we can draw a series of topics $z_n \sim \mathrm{Mult}(z|\theta), n = 1, \ldots, N$. The topic $z_n (= 1, \ldots, k)$ selects a multinomial distribution β_{z_n} from which we draw a word $w_n \sim \mathrm{Mult}(w_n|\beta_{z_n})$, which corresponds to a quantization cluster of the features. Unlike [Fei-Fei and Perona, 2005] where β is assumed to be a parameter, we impose a conjugate prior η upon β to account for insufficient data as suggested by [Blei et al., 2003].

As it is intractable to compute the log likelihood $p(w|\alpha_c, \eta)$, we instead maximize the lower bound $\mathscr{L}(\alpha_c, \eta)$ estimated through the variational distributions over θ, $\{z_d\}$, β. Please refer to [Blei et al., 2003] for the derivation of the variational lower bound and parameter learning for α and η. Once α_c and η are learned, we use the Bayesian MAP criterion to choose the material category

$$c^* = \underset{c}{\operatorname{argmax}} \mathscr{L}(\alpha_c, \eta) + \lambda_c. \quad (3)$$

where $\lambda_c = \log \pi_c$.

5.2.1 Prior learning

5.2.1 Prior learning—A uniform distribution is often assumed for the prior $p(c)$, i.e., each material category is assumed to occur equally often. As we learn the LDA model for each category independently (only sharing the same β), our learning procedure is not guaranteed to converge in finite iterations. Therefore, the probability density functions (pdfs) have to be grounded for a fair comparison. We designed the following greedy algorithm to learn λ by maximizing the recognition rate (or minimizing the error).

Suppose $\{\lambda_i\}_{i \neq c}$ is fixed and we want to optimize λ_c to maximize the rate. Let y_d be the ground truth label for document d. Let $q_{d,i} = \mathscr{L}_d(\alpha_i, \eta) + \lambda_i$ be the "log posterior" for document d to belong to category i. Let $f_d = \max_i q_{d,i}$ be the maximum posterior for document d. We define two sets:

$$\begin{aligned} \Omega_c = & \ \{d | y_d = c, f_d > q_{d,c}\}, \\ \Phi_c = & \ \left\{d | y_d \neq c, f_d = q_{d,y_d}\right\}. \end{aligned} \quad (4)$$

Set Ω_c includes the documents that are labeled as c and misclassified. Set Φ_c includes the documents that are not labeled as c and correctly classified. Our goal is to choose λ_c to make $|\Omega_c|$ as small as possible and $|\Phi_c|$ as large as possible. Note that if we increase λ_c, then $|\Omega_c|$ and $|\Phi_c| \downarrow$, therefore an optimal λ_c exists. We define the set of correctly classified documents using $\lambda_c' > \lambda_c$:

$$\begin{aligned} \Psi_c = & \left\{d | d \in \Omega_c, f_d < q_{d,c} + \lambda_c' - \lambda_c\right\} \cup \\ & \left\{d | d \in \Phi_c, f_d > q_{d,c} + \lambda_c' - \lambda_c\right\}, \end{aligned} \quad (5)$$

and choose the new λ_c that maximizes the size of Ψ_c:

$$\lambda_c \leftarrow \underset{\lambda_c'}{\operatorname{argmax}} |\Psi_c|. \quad (6)$$

We iterate this procedure for each c repeatedly until each λ_c does not change much.

5.2.2 Greedy algorithm for combining features—Should we use all the features we selected in Section 4? Do more features imply better performance? Unfortunately, we have limited training data, and the more features we use the more likely it is that the LDA model will overfit to the training data. To avoid overfitting, we designed a greedy algorithm, shown in Figure 11, to select an optimal subset from our feature pool. The key idea is to construct the best set of features, one feature at a time, so that the recognition rate on an evaluation set is maximized. The greedy algorithm stops when adding more features decreases the recognition rate. Note that we randomly split the training set H into L, for parameter learning, and E, for cross evaluation. After $\mathbb{D}$ is learned, we use the entire training set H to relearn the LDA parameters for $\mathbb{D}$.

5.3 Support Vector Machines (SVMs)

Support vector machines (SVMs) have become standard tools for learning maximum margin classifiers. Details about the learning theory and algorithm for SVM can be found in Burges [1998]. For our problem, we first form a histogram of words $h_d \in \mathbb{R}^{|\mathbb{D}|}$ from document d, and then we use following binary SVM

$$c(h) = \sum_i a_i k(h, h_i) + b \quad (7)$$

to train a one-vs-all classifier for each category. The kernel function $k(\cdot,\cdot)$ is defined as histogram intersection:

$$k(h, h_i) = \sum_{j=1}^{|\mathbb{D}|} \min(h(j), h_i(j)). \quad (8)$$

In general, more features lead to better performance with SVMs. However, we can also apply the feature pursuit algorithm in Figure 11 to learn the optimal feature set for our SVM model. In the next section, we report the results of running our LDA-based and SVM-based systems on the Flickr Materials Database (FMD).

6 Experimental results

We will now describe the performance of our aLDA and SVM models on the Flickr Materials Database (FMD) [Sharan et al., 2009]. There is a binary, human-labeled mask associated with each image in FMD that marks the spatial extent of the material of interest. The average mask of each category is shown in Figure 12. For most of our experiments, we only consider the pixels inside these binary masks and disregard the background pixels. In one experiment, we will test the importance of using these masks and show that using masks to isolate regions of interest has negligible impact on recognition performance.

We started by computing features for each image according to Figure 8. Mindful of computational costs, we sampled *color, jet, SIFT, micro-jet*, and *micro-SIFT* features on a coarse grid (every 5^{th} pixel in both horizontal and vertical directions). Because there are far fewer pixels in edge maps than in the original images, we sampled every other edge pixel for *curvature, edge-slice*, and *edge-ribbon* features. Once all eight features were extracted, they were clustered separately using the k-means algorithm to form dictionaries. We chose the number of clusters for each feature after considering both the dimensionality and the number of instances per feature, as shown in Table 2.

For each of the ten FMD categories, we randomly chose 50 images for training and 50 images for test. In the next four subsections, we report the results that are based on: (i) *one* particular split into training and test sets; and (ii) binary masking of FMD images. In subsection 6.5, we report results that are based on several random splits and that do not rely on binary masking of FMD images.

6.1 aLDA

Before running the feature selection algorithm with the aLDA model, we split the 50 training images per category (set H) randomly into a set of 30 images for estimating parameters (set L) and a set of 20 images for evaluation (set E). After an optimal set of features had been learned for L, we re-learned the parameters using all 50 training images per category and measured the recognition rates on the training and test sets. In the aLDA learning step, we varied the number of topics N from 50 to 250 with a step size of 50, and we selected the best value. The feature selection procedure is shown in Figure 13. First, the our greedy algorithm tries every single feature and discovers that amongst all features, *SIFT* leads to the best performance on the evaluation set, E. In the next iteration, the system picks up *color* from the remaining features and then *edge-slice*. Including more features causes the performance to drop, so the selection procedure of Figure 11 terminates. For this optimal feature set, "*color* + *SIFT* + *edge-slice*", the training rate is 49.4% and the test rate is 44.6%. The recognition rate for random guesses is 10%. It is important to clarify that the recognition rates shown in Figure 13 pertain to the training and test sets even though the feature selection procedure utilized the learning and evaluation sets as described in Figure 11.

The difference in the performance of the best individual feature (SIFT, 35.4%) and the best set of features (color + SIFT + edge-slice, 44.6%) can be attributed to the aLDA model. Interestingly, when all eight features are combined by the aLDA model, the test rate (38.8%) is lower than when fewer features are combined. Using more features can cause overfitting, especially for a database as small as FMD. The fact that SIFT is the best-performing single feature indicates the importance of texture information for material recognition. Edge-slice, which measures reflectance features, is also useful.

6.2 SVM

We used an aggregated one-vs-all SVM classifier, using MATLAB's built-in package, for multi-category material recognition. We tried both the histogram intersection kernel and the linear kernel, and we found that the histogram intersection kernel performed better (linear: 50.9%, histogram intersection: 57.1% [3]). In general, combining more features leads to higher performance for discriminative classifiers. However, we still use the feature selection process of Figure 14 to understand the relative importance of features. Because a multiclass SVM with a histogram intersection kernel always produces 100% training rate, we do not show the training rate in Figure 14.

In the scan line order, the first eight plots in Figure 14 show the test rate of each individual feature. The SVM model performs much better than the aLDA model for *SIFT, micro-jet*, and *micro-SIFT*, slightly better than the aLDA model for *color, jet*, and *curvature*, and slightly worse than the aLDA model for *edge-ribbon* and *edge-slice*. When the features are combined, SVM performs much better than aLDA. The next seven plots show the feature selection procedure for SVM. Because the feature set grows as features get added, we use the term, "preset", to denote the feature set used in the previous step of the feature selection process. The first two features selected by SVM are the same as aLDA, namely, *SIFT* and *color*, but the test rate for this combination (50.2%) is much higher than for aLDA (43.6%). The remaining features are selected in following order: curvature, edge-ribbon, micro-SIFT, jet, edge-slice, and *micro-jet*. This order of feature selection illustrates the importance of the edge-based features.

We also explored the importance of features by subtraction in the last three plots of Figure 14. With only a 2.6% drop in performance, *SIFT* is not as important in the presence of other features. Meanwhile *color* is more important because a 8.6% drop is obtained by excluding color. Excluding edge-based features (*curvature, edge-slice*, and *edge-ribbon*) leads to a 6.2% drop in performance, which reinforces the importance of these features.

6.3 Nearest neighbor classifier

We implemented and tested a former state-of-the-art system for 3-D texture recognition [Varma and Zisserman, 2009]. The performance of this system on FMD serves as a baseline for our results. The Varma-Zisserman (VZ) system uses 5×5 pixel gray-scale patches as features, clusters features into codewords, obtains a histogram of the codewords for each image, and employs a nearest neighbor classifier. First, we ran our implementation of the

[3] Kernel comparison results were obtained by averaging over 14 different splits of FMD into training and tests sets. All other results in Sections 6.1 - 6.4 pertain to a single split.

VZ system on the CURET database [Dana et al., 1999] and reproduced Varma and Zisserman's original results (our implementation: 96.1% test rate, VZ: 95 ~ 98% test rate). Next, we ran the same VZ system that we tested on CURET on FMD. The VZ test rate on FMD was 23.8%. This result supports the conclusions from Figure 4 that the FMD is a much more challenging database than CURET for recognition purposes.

As the VZ system uses features tailored for the CURET database (5×5 pixel patches), we ran their nearest neighbor classifier with our features that were developed for FMD. The results are shown in Figure 15. The training rate for nearest neighbor classifiers is always 100%, so we only report the test rate. Many of our features outperform fixed-size grayscale patches on FMD (i.e., test rate > 23.8%), which is not surprising given the diversity of appearances in FMD. The nearest neighbor classifier with *SIFT* features alone has a test rate of 31.8%, which is similar to our aLDA model with *SIFT* features (35.2%). However, combining features in the nearest neighbor classifier leads to only a modest increase in the test rate, 37.4%. Clearly, both the SVM and aLDA models are better at combining features for material category recognition.

6.4 Confusion matrices and misclassification examples

In Figure 16, we present the confusion matrices and examples of misclassification for the aLDA and SVM models. The confusion matrix for the LDA-based system (*color* + *SIFT* + *edge-slice* with average recall 44.6%) tells us how often a given category is misclassified as another. For example, fabric is often misclassified as stone, leather is misclassified as fabric, and plastic misclassified as paper. The category metal is more likely to be classified as glass than itself. The misclassification examples shown in Figure 16 make sense as there are certain commonalities in the appearance of leather and fabric, plastic and paper, and metal and glass.

The confusion matrix for the SVM-based system (all features with average recall 60.6%) is cleaner than that of the LDA-based system. Using the labeling scheme of Figure 16 for clarity, the most visible improvements occurs for the following pairs: fabric: *stone*, leather: *fabric*, and metal: *glass*. As one might expect, SVM outperforms aLDA on samples that reside close to decision boundaries in feature space.

One can compare the errors made by our systems to those made by humans by examining Figures 6 and 16. There are some similarities in the misclassifications (leather: *fabric*, water: *glass*, and wood: *stone*) even though humans are much better at recognizing material categories than our systems. We will return to this point in Section 7.

6.5 Variations

We will now report the results averaged over 14 random splits of FMD categories into training and test sets while retaining the 50% training ratio. For aLDA, the feature selection procedure yields different feature sets for the 14 splits. For all splits, we limit the maximum number of features to be three as one needs significantly more computation time and memory for large vocabularies in LDA modeling. *SIFT* and *color* were always selected as the first and second features. However, the third feature that was selected varied based on the split. For the 14 splits, *edge-ribbon* was selected six times, *edge-slice* was selected four times, *curvature* was selected three times, and *micro-SIFT* was selected once. This result confirms the importance of edge-based features introduced in the paper. The average recognition rate for 14 splits was 42.0% with standard deviation 1.82%. For SVM, all eight features were selected for all 14 splits, and the average recognition rate was 57.1% with standard deviation 2.33%.

Next, we ran both systems on the same 14 splits without the binary masking step. Averaged over all 14 splits, the test rate was 39.4% for aLDA and 55.6% for SVM. The standard deviations were similar to those in the masking condition. The drop in performance (aLDA: 2.6%, SVM: 1.5%) that results from skipping the masking step is quite minor, and it is comparable to the standard deviation of the recognition performance. Therefore, it is fair to conclude that using masks has little effect on system performance.

7 Comparison to human performance

As one compares Figures 6a and 16, it is clear that humans are much better (84.9%) at recognizing material categories in FMD images than our computer vision system (SVM: 57.1%, aLDA: 42%). To understand this gap in performance better, we conducted the same analysis for our computer vision system that we had conducted for human observers in Section 3. We ran our SVM-based system on the four sets of modified images (bilateral filtered, high-pass filtered, and texture synthesized 15×15 and 30×30) shown in Figure 5. By comparing the performance of our system to that of humans on these modified images, we can gain insights into why our system falls short of human performance. We chose the SVM-based system over the LDA-based system for these comparisons because the SVM classifier led to the best performance. The results of running our system on the modified images are presented in Figures 17 through 20. The same split of FMD images into training and test sets was used in each case.

The bilateral filtered images were created to emphasize outline shape and object identity while suppressing color, texture, and reflectance information. On examining Figure 17, one notices that using only the *color* feature yields chance performance (10%), which makes sense because bilateral filtering removes color information. The best individual feature is *SIFT* (20.6%), and the best set of features (26.2%) comprises *SIFT, edge-ribbon, curvature, color*, and *edge-slice* features. When compared to human performance, the best performance of our system falls short (65.3% vs. 26.2%), which shows that humans are much better at extracting shape and object-based features than our system.

The high-pass filtered images were also created to emphasize outline shape and object identity, although the texture information is suppressed to a lesser extent than in the bilateral filtered images. In Figure 18, we notice, once again, that *color* by itself is not very useful. The best individual feature is *micro-SIFT*, and the best set of features (35.6%) comprises *micro-SIFT, jet, color, edge-ribbon*, and *micro-jet*, which suggests that texture-based features are more important for these images than the bilateral filtered images. While human performance on bilateral filtered (65.3%) and high-pass-filtered (64.8%) images is comparable, our system performs better on high-pass filtered images (35.6%) than bilateral filtered images (26.2%). The improvement in performance can be attributed to the fact that subtle texture details can be utilized more effectively than shape or object-based information by our system.

The texture synthesized images were created with the goal of emphasizing texture information while suppressing shape and object-based information. In Figures 19 and 20, we see that the best feature sets tend to include texture features (e.g., *SIFT, jet*) and edge-based features (e.g., *edge-ribbon, edge-slice*). The utility of edge-based features might seem surprising given that the texture synthesized images were spatially scrambled, however this result can be explained as follows: locally preserved regions in texture synthesized images have the same pixel relationships across and along edges as the original images, so edge-based features are still useful. The most interesting observation about Figures 19 and 20 is that humans and computer vision systems perform similarly on locally preserved but globally scrambled images (humans: 38.7% and 46.9%, our system: 33.8% and 42.6%). This result leads to two conclusions: (i) local image information is of limited value to humans for material category recognition; and (ii) our SVM-based system can utilize local image information nearly as effectively as humans.

To summarize, the comparisons with human performance show that it is important to model non-local image information in order to succeed at material category recognition. This non-local image information includes many aspects of surface appearance – 3-D shape, surface reflectance, illumination, and object identity. Under normal circumstances, humans are able to untangle these variables and recognize material properties. However, when given only local patches, humans fail to untangle these variables, and their performance at material recognition is poor. Therefore, computer vision systems should not rely only on features based on local image patches.

8 Discussion and conclusion

We have presented a recognition system that can categorize single, uncalibrated images of real-world materials. We designed a set of image features based on studies of human material recognition, and we combined them in an SVM classifier. Our system achieves a recognition rate of 57.1%, and it outperforms a state-of-the-art method (23.8%, [Varma and Zisserman, 2009]) on a challenging database that we introduce to the computer vision community, the Flickr Materials Database (FMD) [Sharan et al., 2009]. The sheer diversity of material appearances in FMD, as illustrated in Figures 2 and 4, makes the human performance we measured on FMD (84.9%) an ambitious benchmark for material recognition.

Readers may have noted that the recognition performance of our system varies with the material category. For example, performance is highest for ‘Foliage’ images and lowest for ‘Metal’ images in Figure 14 (all features). This trend makes sense; images of metal surfaces tend to be more varied than images of green leaves. Color information, by itself, allows ‘Foliage’ images to be categorized with >70% accuracy, as shown in Figure 14 (color). The confusions between categories, shown in Figure 16, are also reasonable. Glass and metal surfaces are often confused as are leather and fabric. These material categories share certain reflectance and texture properties, which leads to similar image features and eventually, confusions. Based on these observations, we suggest that material categories be organized according to shared properties, similar to the hierarchies that have been proposed for objects and scenes [Rosch and Lloyd, 1978, WordNet, 1998].

We evaluated two models for material recognition, a generative LDA model [Blei et al., 2003] and a discriminative SVM classifier. The SVM classifier (57.1%) was better at combining our features than our augmented LDA model (42%), and that is why we chose the SVM classifier for our system. We also evaluated different combinations of features in our experiments and found that *color* and edge-based features (i.e., *curvature* and the new features that we have proposed, *edge-slice* and *edge-ribbon*) are important. Although *SIFT* was the best individual feature, it was not as necessary for ensuring good performance as *color* and edge-based features (Figure 14, all: 60.6%, all except SIFT: 58%, all except color: 52%, all except edge-based: 54.4%). In fact, edge-based features achieve slightly higher accuracies than SIFT by themselves (edge-based: 42.8%, SIFT: 41.2%) and in combination with color (edge-based + color: 54.4%, SIFT + color: 50.2%).

Beyond specific image features, we have shown that local image information (i.e., color, texture, and local shape), in itself, is not sufficient for material recognition (Figures 6d, 6e, 19, and 20). Humans struggle to identify material categories when they are presented globally scrambled but locally preserved images, and their performance is comparable to that of our system for such images. Natural categories are somewhat easier to recognize than man-made categories in these conditions, both for humans and our computer vision system (Figures 6f, 19, and 20). Based on these results, we suggest that the future progress will come from modeling the non-local aspects of surface appearance (e.g., extended highlights, object identity) that correlate with the material category.

One might wonder why local surface properties are not sufficient for material recognition. Consider Figure 3. Local surface information such as color or texture is not always helpful; the surfaces in Figure 3a could be made of (top row) shiny plastic or metal and (bottom row) human skin or wood. It is only when we consider non-local surface features such as the elongated highlights on the grill and the hood or the edge of the table in Figure 3b that we can identify the material category (metal and wood, respectively). When objects are fully visible in an image (e.g., Figure 3c), shape-based object identity, another non-local surface feature, further constrains the material category (e.g., tables are usually made of wood not skin). Identifying and modeling these non-local aspects of surface appearance is, we believe, the key to successful material recognition.

To conclude, material recognition is an important problem in image understanding, and it is distinct from 3-D texture recognition ([Varma and Zisserman, 2009]’s 3-D texture classifier does poorly on FMD) and shape-based object recognition (outline shape information is, on average, not useful; see Figure 12 and Section 6.5). In this paper, we are merely taking one of the first steps towards solving it. Our approach has been to use lessons from perception to develop the components of our recognition system. This approach differs significantly from recent work where perceptual studies have been used to evaluate components of well-established computer vision workflows [Parikh and Zitnick, 2010]. Material recognition is a topic of current study both in the human and computer vision communities, and our work constitutes the first attempt at automatically recognizing high-level material categories “in the wild”.

References

Adelson EH. On seeing stuff: The perception of materials by humans and machines. SPIE, Human Vision and Electronic Imaging VI. 2001; 4299:1–12.

Bae, S.; Paris, S.; Durand, F. ACM SIGGRAPH. 2006. Two-scale tone management for photographic look.

Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts. TPAMI. 2002; 24(4):509–522.

Berzhanskaya J, Swaminathan G, Beck J, Mingolla E. Remote effects of highlights on gloss perception. Perception. 2005; 34(5):565–575. [PubMed: 15991693]

Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003; 3:993–1022.

Bloj M, Kersten D, Hurlbert AC. Perception of three-dimensional shape influences color perception through mutual illumination. Nature. 1999; 402:877–879. [PubMed: 10622251]

Boivin, S.; Gagalowicz, A. ACM SIGGRAPH. 2001. Image-based rendering of diffuse, specular and glossy surfaces from a single image; p. 107-116.

Boyaci H, Maloney LT, Hersh S. The effect of perceived surface orientation on perceived surface albedo in binocularly viewed scenes. Journal of Vision. 2003; 3:541–553. [PubMed: 14632606]

Brainard, DH.; Kraft, JM.; Longere, P. Color Perception: From Light to Object, chapter Color constancy: developing empirical tests of computational models. Oxford University Press; 2003. p. 307-334.

Burges C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 1998; 2(2):121–167.

Canny J. A computational approach to edge detection. TPAMI. Nov; 1986 8(6):679–698.

Caputo B, Hayman E, Mallikarjuna P. Class-specific material categorization. ICCV. 2005; 2:1597–1604.

Barbara, Caputo; Eric, Hayman; Mario, Fritz; Jan-Olof, Eklhund. IDIAP-RR 69. IDIAP; Martigny, Switzerland: 2007. Classifying Materials in the Real World.

Cula OG, Dana KJ. 3d texture recognition using bidirectional feature histograms. IJCV. 2004a; 59(1): 33–60.

Cula OG, Dana KJ, Murphy FP, Rao BK. Bidirectional imaging and modeling of skin texture. IEEE Transactions on Biomedical Engineering. 2004; 51(12):2148–2159. [PubMed: 15605862]

Cula OJ, Dana KJ. 3d texture recognition using bidirectional feature histograms. International Journal of Computer Vision. 2004b; 59(1):33–60.

Dalal N, Triggs B. Histograms of oriented gradients for human detection. CVPR. 2005; 2:886–893.

Dana, KJ.; Nayar, S. CVPR. 1998. Histogram model for 3d textures; p. 618-624.

Dana KJ, Van-Ginneken B, Nayar SK, Koenderink JJ. Reflectance and texture of real world surfaces. ACM Transactions on Graphics. 1999; 18(1):1–34.

Debevec, P.; Hawkins, T.; Tchou, C.; Duiker, HP.; Sarokin, W.; Sagar, M. ACM SIGGRAPH. 2000. Acquiring the reflectance field of a human face; p. 145-156.

Debevec, P.; Tchou, C.; Gardner, A.; Hawkins, T.; Poullis, C.; Stumpfel, J.; Jones, A.; Yun, N.; Einarsson, P.; Lundgren, T.; Fajardo, M.; Martinez, P. Estimating surface reflectance properties of a complex scene under captured natural illumination. University of Southern California; 2004. ICT-TR-06

Dror, R.; Adelson, EH.; Willsky, AS. Recognition of surface reflectance properties from a single image under unknown real-world illumination; IEEE Workshop on identifying objects across variation in lighting; 2001;

Durand, F.; Dorsey, J. ACM SIGGRAPH. 2002. Fast bilateral filtering for the display of high-dynamic-range images.

Efros, AA.; Freeman, WT. ACM SIGGRAPH. 2001. Image quilting for texture synthesis and transfer.

Fei-Fei L, Perona P. A bayesian hierarchical model for learning natural scene categories. CVPR. 2005; 2:524–531.

Fleming RW, Bülthoff H. Low-level image cues in the perception of translucent materials. ACM Transactions on Applied Perception. 2005; 2(3):346–382.

Fleming RW, Dror R, Adelson EH. Real world illumination and the perception of surface reflectance properties. Journal of Vision. 2003; 3(5):347–368. [PubMed: 12875632]

Fleming RW, Torralba A, Adelson EH. Specular reflections and the perception of shape. Journal of Vision. 2004; 4(9):798–820. [PubMed: 15493971]

Forsyth D, Fleck MM. Automatic detection of human nudes. IJCV. 1999; 32(1):63–77.

Fritz, M.; Black, M.; Bradski, G.; Darrell, T. NIPS. 2009. An additive latent feature model for transparent object recognition.

Gilchrist A, Kossyfidis C, Bonato F, Agostini T, Cataliotti J, Li X, Spehar B, Annan V, Economou E. An anchoring theory of lightness perception. Psychological Review. 1999; 106:795–834. [PubMed: 10560329]

He, XD.; Torrance, KE.; Sillion, FS.; Greenberg, DP. A comprehensive physical model for light reflection; 18th Annual Conference on Computer Graphics and Interactive Techniques; ACM; 1991. p. 175-186.

Ho YX, Landy MS, Maloney LT. Conjoint measurement of gloss and surface texture. Psychological Science. 2008; 19(2):196–204. [PubMed: 18271869]

Hu, D.; Bo, L.; Ren, X. BMVC. 2011. Towards robust material recognition for everyday objects.

Jensen, HW.; Marschner, S.; Levoy, M.; Hanrahan, P. ACM SIGGRAPH. 2001. A practical model for subsurface light transport; p. 511-518.

Khan EA, Reinhard E, Fleming RW, Bülthoff HH. Image-based material editing. ACM SIGGRAPH. 2006:654–663.

Koenderink JJ, Van Doorn AJ, Dana KJ, Nayar S. Bidirectional reflectance distribution function of thoroughly pitted surfaces. International Journal of Computer Vision. 1999; 31:129–144.

Koenderink JJ, van Doorn AJ. Representation of local geometry in the visual system. Biological Cybernetics. 1987; 545:367–375. [PubMed: 3567240]

Leung T, Malik J. Representing and recognizing the visual appearance of materials using three-dimensional textons. IJCV. 2001; 43(1):29–44.

Liu, C.; Sharan, L.; Rosenholtz, R.; Adelson, EH. CVPR. 2010. Exploring features in a bayesian framework for material recognition.

Liu, C.; Yuen, J.; Torralba, A. CVPR. 2009. Nonparametric scene parsing: Label transfer via dense scene alignment.

Lowe DG. Distinctive image-features from scale-invariant keypoints. IJCV. 2004; 60(2):91–110.

Maloney, LT.; Yang, JN. Color Perception: From Light to Object. Oxford University Press; 2003. The illumination estimation hypothesis and surface color perception; p. 335-358.

Marschner, S.; Westin, SH.; Arbree, A.; Moon, JT. ACM SIGGRAPH. 2005. Measuring and modeling the appearance of finished wood; p. 727-734.

Marschner, S.; Westin, SH.; LaFortune, EPF.; Torrance, KE.; Greenberg, DP. Image-based brdf measurement including human skin; 10th Eurographics Workshop on Rendering; 1999. p. 139-152.

Matusik, W.; Pfister, H.; Brand, M.; McMillan, L. ACM SIGGRAPH. 2000. A data-driven reflectance model; p. 759-769.

McHenry K, Ponce J. A geodesic active contour framework for finding glass. CVPR. 2005; 1:1038–1044.

McHenry K, Ponce J, Forsyth D. Finding glass. CVPR. 2005; 2:973–979.

Motoyoshi I, Nishida S, Sharan L, Adelson EH. Image statistics and the perception of surface reflectance. Nature. 2007; 447:206–209. [PubMed: 17443193]

Nicodemus F. Directional reflectance and emissivity of an opaque surface. Applied Optics. 1965; 4(7): 767–775.

Nillius P, Eklundh J-O. Classifying materials from their reflectance properties. ECCV. 2004; 4:366–376.

Nishida S, Shinya M. Use of image-based information in judgments of surface reflectance properties. Journal of the Optical Society of America A. 1998; 15:2951–2965.

Nishino, K.; Zhang, Z.; Ikeuchi, K. ICCV. 2001. Determining reflectance parameters and illumination distributions from a sparse set of images for view-dependent image synthesis; p. 599-601.

Oren M, Nayar SK. Generalization of the lambertian model and implications for machine vision. International Journal of Computer Vision. 1995; 14(3):227–251.

Parikh, D.; Zitnick, L. CVPR. 2010. The role of features, algorithms and data in visual recognition.

Pellacini, F.; Ferwerda, JA.; Greenberg, DP. Towards a psychophysically-based light reflection model for image synthesis; 27th Annual Conference on Computer Graphics and Interactive Techniques; ACM; 2000. p. 55-64.

Phong B-T. Illumination for computer generated pictures. Communications of ACM. 1975; 18:311–317.

Pont SC, Koenderink JJ. Bidirectional texture contrast function. IJCV. 2005; 62(1/2):17–34.

Ramamoorthi, R.; Hanrahan, P. ACM SIGGRAPH. 2001. A signal processing framework for inverse rendering; p. 117-128.

Robilotto R, Zaidi Q. Limits of lightness identification of real objects under natural viewing conditions. Journal of Vision. 2004; 4(9):779–797. [PubMed: 15493970]

Romeiro F, Vasilyev Y, Zickler TE. Passive reflectometry. ECCV. 2008; 4:859–872.

Romeiro F, Zickler TE. Blind reflectometry. ECCV. 2010; 1:45–58.

Rosch, E.; Lloyd, BB., editors. Cognition and categorization, chapter Principles of categorization. Erlbaum; 1978.

Sato, Y.; Wheeler, M.; Ikeuchi, K. ACM SIGGRAPH. 1997. Object shape and reflectance modeling from observation; p. 379-387.

Savarese, S.; Criminisi, A. Classification of folded textiles. Aug. 2004 URL: http://research.microsoft.com/vision/cambridge/recognition/MSRC_MaterialsImageDatabase.zip

Sharan L, Li Y, Motoyoshi I, Nishida S, Adelson EH. Image statistics for surface reflectance perception. Journal of the Optical Society of America A. 2008; 25(4):846–865.

Sharan L, Rosenholtz R, Adelson E. Material perception: What can you see in a brief glance? Journal of Vision. 2009; 9(8):784, 784a. Abstract.

Todd JT, Norman JF, Mingolla E. Lightness constancy in the presence of specular highlights. Psychological Science. 2004; 15:33–39. [PubMed: 14717829]

Tominaga S, Tanaka N. Estimating reflection parameters from a single color image. IEEE Computer Graphics and Applications. 2000; 20(5):58–66.

Varma M, Zisserman A. A statistical approach to texture classification from single images. IJCV. 2005; 62(1-2):61–81.

Varma M, Zisserman A. A statistical approach to material classification using image patch exemplars. TPAMI. 2009; 31(11):2032–2047.

Ward, G. Measuring and modeling anisotropic reflection; 19th Annual Conference on Computer Graphics and Interactive Techniques; ACM; 1992. p. 265-272.

WordNet. WordNet: An electronic lexical database. MIT Press; Cambridge, MA: 1998.

Xiao B, Brainard DH. Surface gloss and color perception of 3d objects. Visual Neuroscience. 2008; 25:371–385. [PubMed: 18598406]

Yu, Y.; Debevec, P.; Malik, J.; Hawkins, T. ACM SIGGRAPH. 1999. Inverse global illumination: recovering reflectance models of real scenes from photographs; p. 215-224.

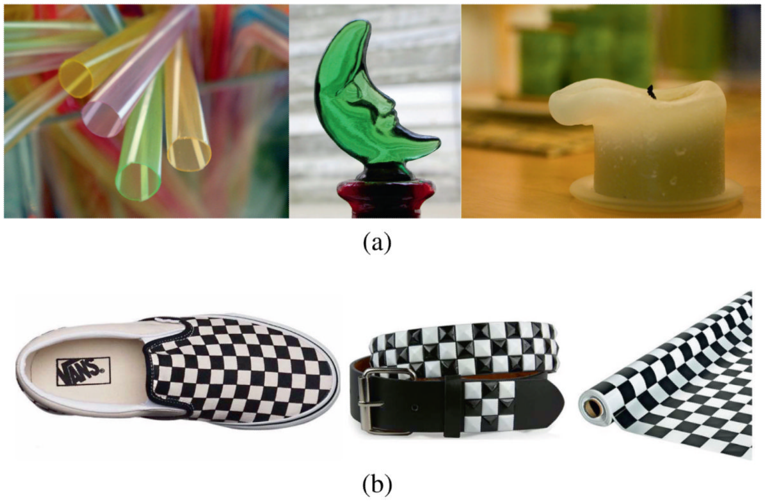


Fig. 1. Material categorization vs. reflectance estimation and vs. texture recognition
(a) Surfaces that are made of different materials can exhibit similar reflectance properties. These translucent surfaces are made of (from left to right): *plastic, glass*, and *wax*. (b) Surfaces with similar texture patterns can be made of different materials. These objects share the same checkerboard patterns, but they are made of (from left to right): *fabric, plastic*, and *paper*.

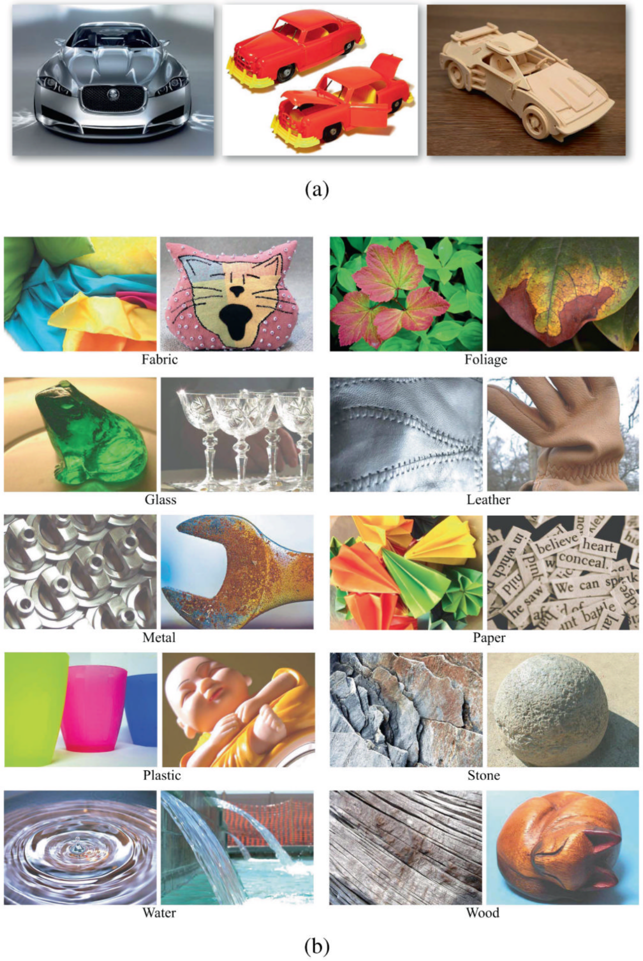


Fig. 2. Material categorization vs. object categorization

(a) Objects that belong to the same category can be made of different materials. These vehicles are made of (from left to right): *metal, plastic*, and *wood*. (b) On the other hand, objects that belong to different categories can be made of the same material. Consider the *fabric* and *glass* examples. The scarves and the cat-shaped pillow are both made of fabric. Similarly, the frog and the crystal glasses are both made of glass. These images belong to the Flickr Materials Database (FMD) [Sharan et al., 2009], which consists of a diverse selection of surfaces in ten common material categories. We will use FMD in this paper to design image features and to evaluate material recognition systems.

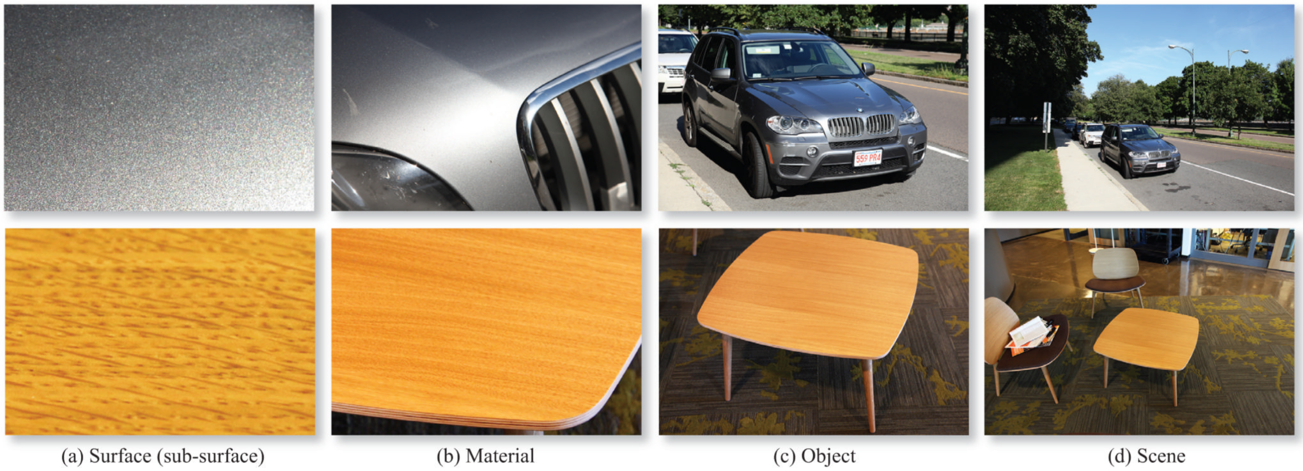


Fig. 3. Visual recognition as a function of spatial scale

(a) Extreme close-up views, (b) close-up views, (c) regular views, and (d) zoomed-out views of (top row) a car and (bottom row) a table. Visual recognition is driven by (a) 2-D and 3-D textures properties, (b) material properties, (c) 3-D shape and object properties, and (d) scene properties. In this work, we will use images from the Flickr Materials Database (FMD) [Sharan et al., 2009] that depict spatial scales in the range (b)-(c).

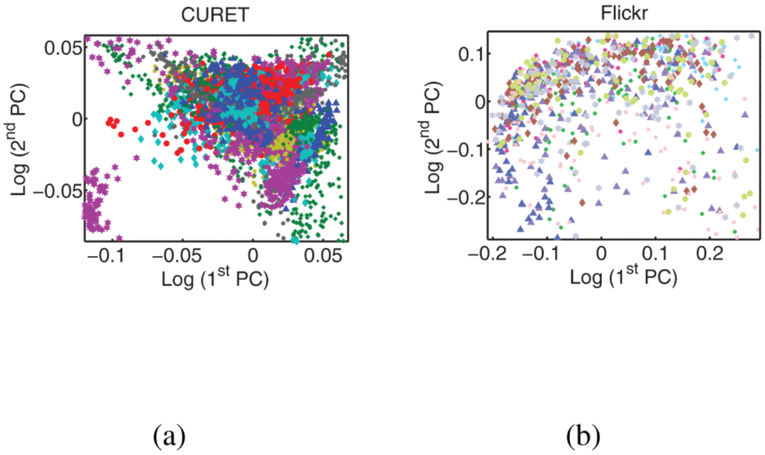


Fig. 4. CURET vs. Flickr Materials Database

The projection of the first two principal components (PCs) of texton histograms are shown for all images in (left) the 61-class CURET database [Dana and Nayar, 1998] and (right) the 10-class Flickr Materials Database [Sharan et al., 2009]. The textons were derived from 5×5 pixel patches as described in [Varma and Zisserman, 2009]. The colors indicate the various texture/material categories. These plots demonstrate that CURET samples are more separable than Flickr.

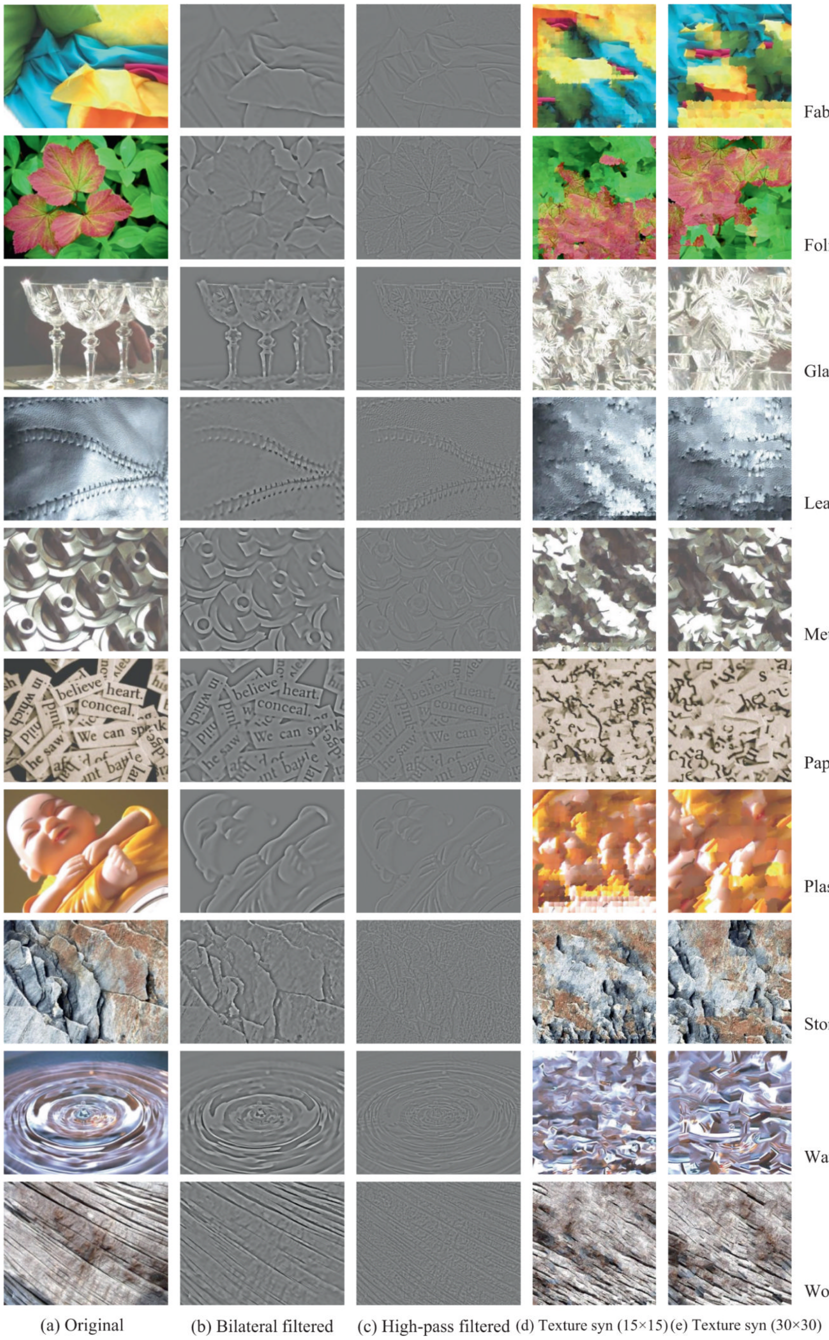
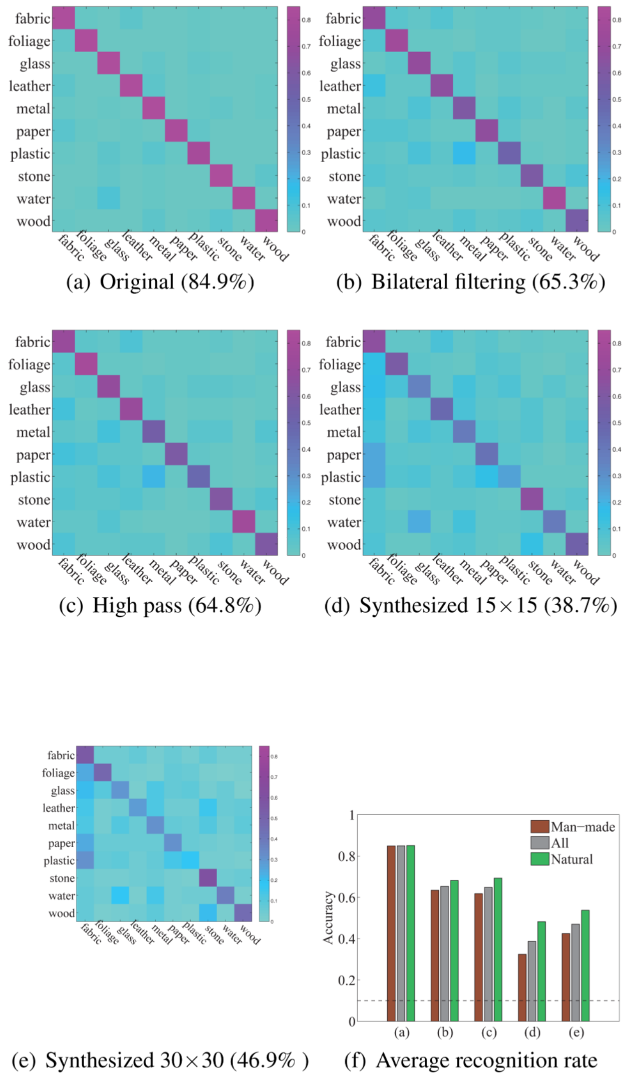


Fig. 5. Images used in the Mechanical Turk experiments

Images from FMD were presented either in (a) their original form or in their (b) bilateral filtered, (c) high-pass filtered, or (d,e) texture synthesized forms. Texture synthesized images were created for two different patch sizes. The images shown here were derived from the following FMD categories (from top to bottom): fabric, foliage, glass, leather, metal, paper, plastic, stone, water, and *wood*.



(a) Original (84.9%)

(b) Bilateral filtering (65.3%)

(c) High pass (64.8%)

(d) Synthesized 15×15 (38.7%)

(e) Synthesized 30×30 (46.9%)

(f) Average recognition rate

Fig. 6. Results of the Mechanical Turk experiments

The accuracy of categorization is highest for (a) the original images, and it decreases when (b),(c) filtered images and (d),(e) texture synthesized images are presented. (f) The average recognition rate is plotted as function of experimental condition for all FMD categories including natural and man-made categories.

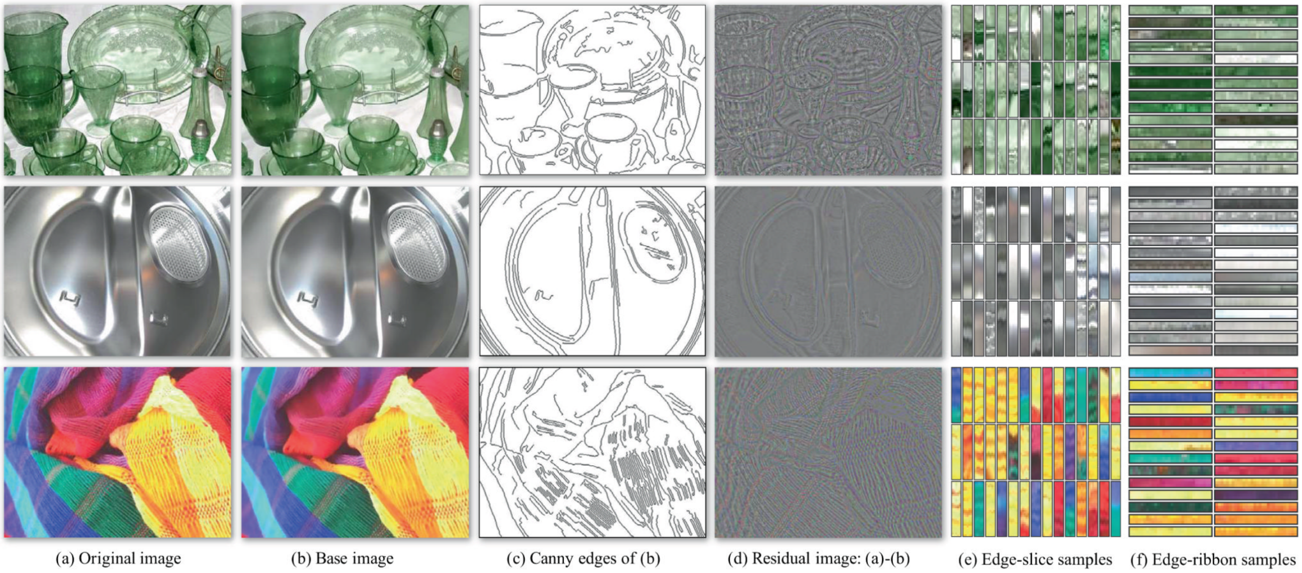


Fig. 7. Candidate features for material recognition

(a) These images belong to the (top row) Glass, (middle row) Metal, and (bottom row) Fabric categories of the Flickr Materials Database. We perform bilateral filtering [Bae et al., 2006] on the original images to obtain (b) the base images. We run the Canny edge detector [Canny, 1986] on the base images to obtain (c) edge maps, which are then used to compute *curvature* features. Subtracting (b) from (a) results in (d) residual images that contain 'micro-texture' information. We use *micro-jet* and *micro-SIFT* features, as described in the text, to characterize this information. Finally, we extract slices of pixels (e) normal to and (f) along the tangential direction of the Canny edges in (c). Arrays of HOGs are used to compute *edge-slice* and *edge-ribbon* features from these pixel slices.

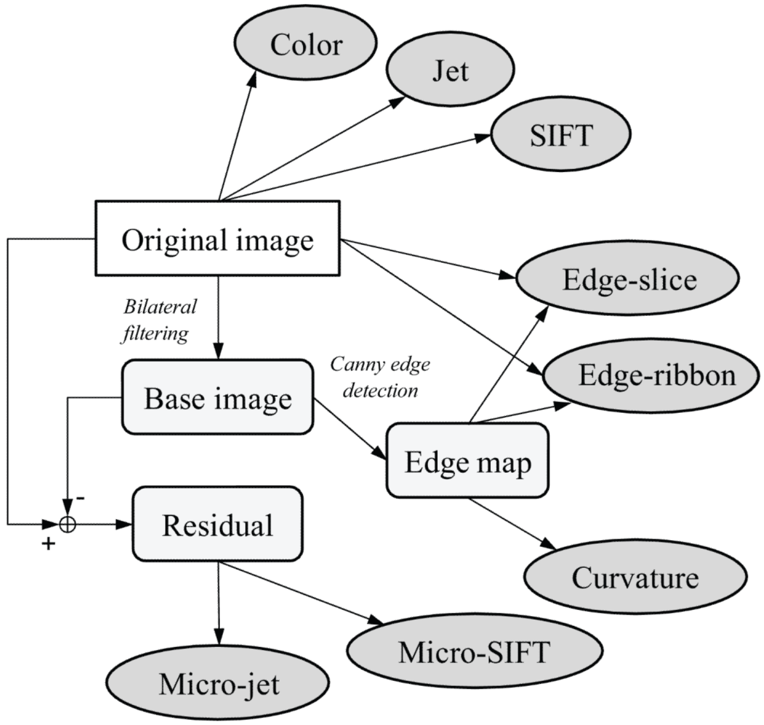


Fig. 8.
An illustration of how our features are generated. We tested eight sets of features, shown here in ellipses: color, jet, SIFT, micro-jet, micro-SIFT, curvature, edge-slice, and *edge-ribbon*.

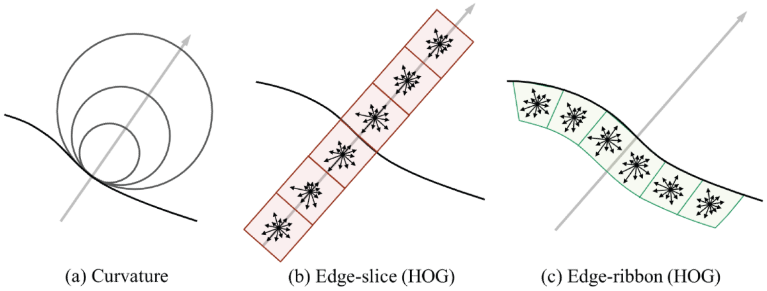


Fig. 9.
(a) We computed the curvature of edges at three spatial scales to measure outline shape information. (b,c) We computed HOGs over 6 cells defined near the edges in images to measure reflectance information.

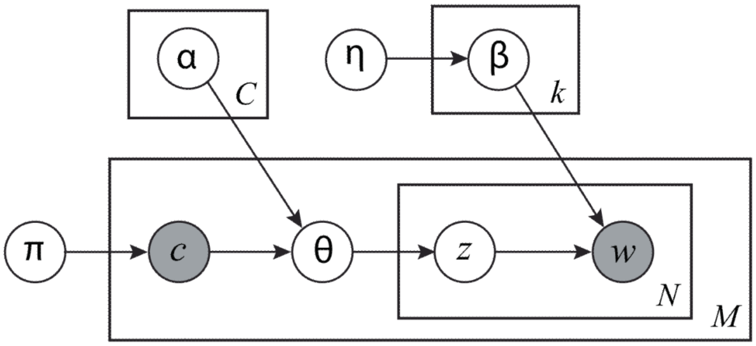


Fig. 10.

The graphical model of LDA [Blei et al., 2003]. Note that our categorization shares both the topics and codewords. Unlike [Fei-Fei and Perona, 2005], we impose a prior on β to account for insufficient data.

```
Input: dictionary pool $\{D_1, \cdots, D_m\}$, training set $H$

• Initialize: $\mathbb{D} = \emptyset$, recognition rate $r = 0$
• Randomly split $H = L \cup E$
  for $l = 1$ to $m$
      for $D_i \notin \mathbb{D}$
          • Augment dictionary $\mathbb{D}' = \mathbb{D} \cup \{D_i\}$
          • Concatenate words according to $\mathbb{D}'$ using Eqn. (2)
          • Train LDA on $L$ for each category (sharing $\beta$)
          • Learn prior $\lambda$ using Eqn. (5) and (6)
          • $r_i =$ recognition rate on $E$ using Eqn. (3)
      end
      if $\max r_i > r$
          • $j = \arg\max_i r_i$, $\mathbb{D} = \mathbb{D} \cup \{D_j\}$, $r = r_j$
      else
          break
      end
  end
• Train LDA and learn prior $\lambda$ on $H$
• $r =$ recognition rate on $H$

Output: $\mathbb{D}$, $r$
```

Fig. 11. The augmented LDA (aLDA) algorithm to learn an optimal feature set in a greedy manner

The LDA model in this flow-chart can be substituted by an SVM, although in general, adding more features improves the performance of SVMs.

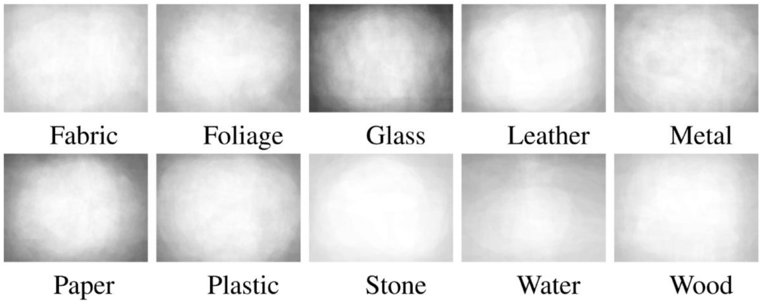


Fig. 12. Average mask for each category in FMD

Although binary masks are provided for each image in the Flickr Materials Database, we see here that, on average, the masks are not very informative. In Section 6.5, we show that using only the pixels inside the masks vs. using all the pixels in the images has minor impact on the recognition performance.

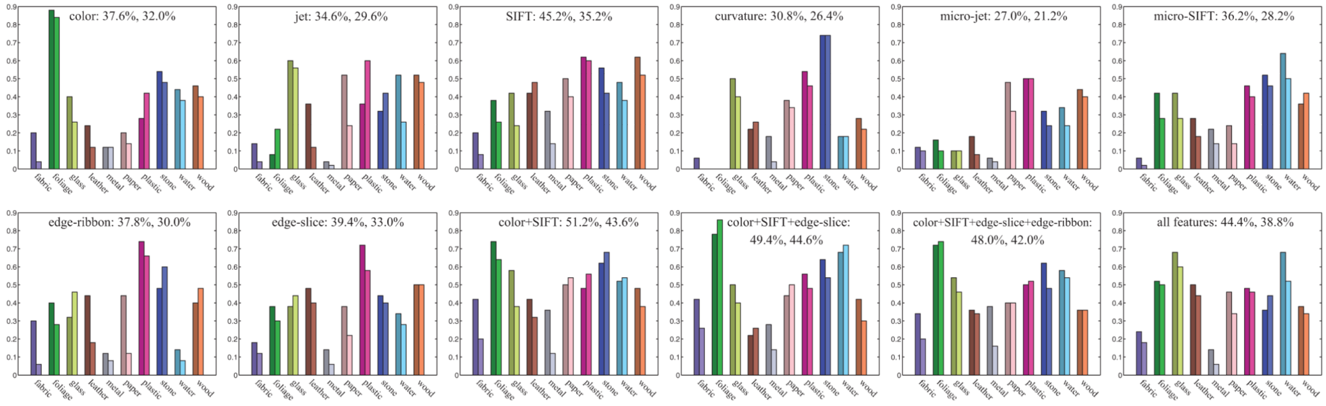


Fig. 13. Feature selection with aLDA

Each plot shows the per-category recognition rate for different combinations of features. The training rate is shown on the left with a darker bar. The test set is shown on the right with a lighter bar. The two numbers displayed right after the feature combination label are the training and test recognition rates averaged across material categories. Our feature selection algorithm finds “*color* + *SIFT* + *edge-slice*” to be the optimal feature set for the aLDA model on FMD images.

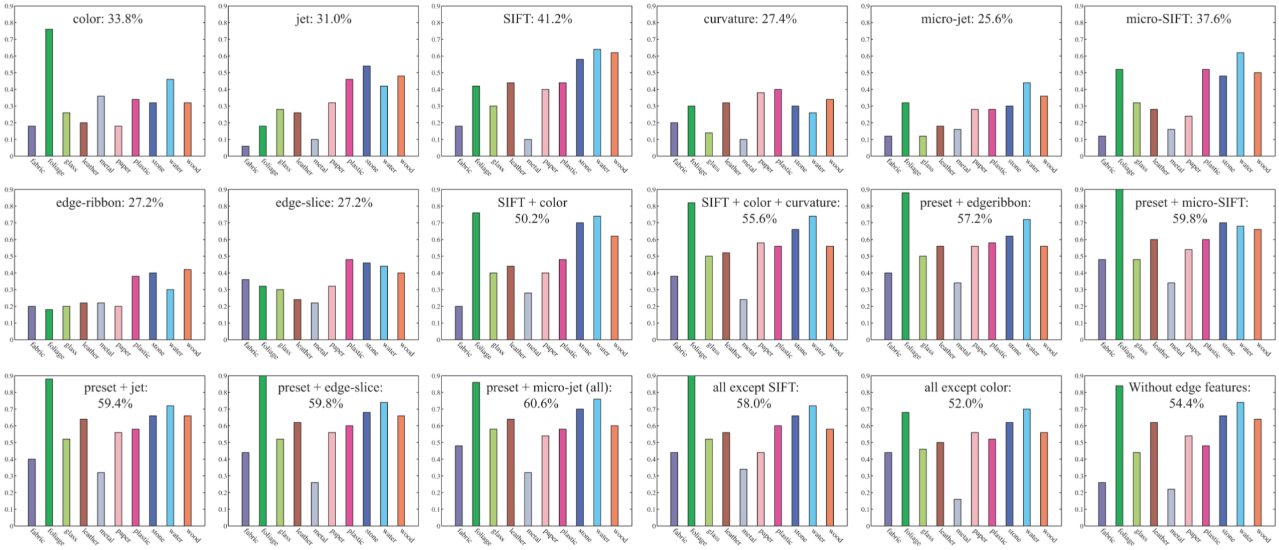


Fig. 14. Feature selection with SVM

These plots are arranged similar to Figure 13. The SVM with the histogram intersection kernel always produces 100% training rate, so we do not show the training rate in these plots. In contrast to the aLDA model, adding more features always seem to improve the performance of the SVM model. The last three plots show features sets where one feature has been removed. *Color* and edge related features, namely, *curvature, edge-ribbon*, and *edge-slice*, turn out to be more useful than *SIFT* for FMD images.

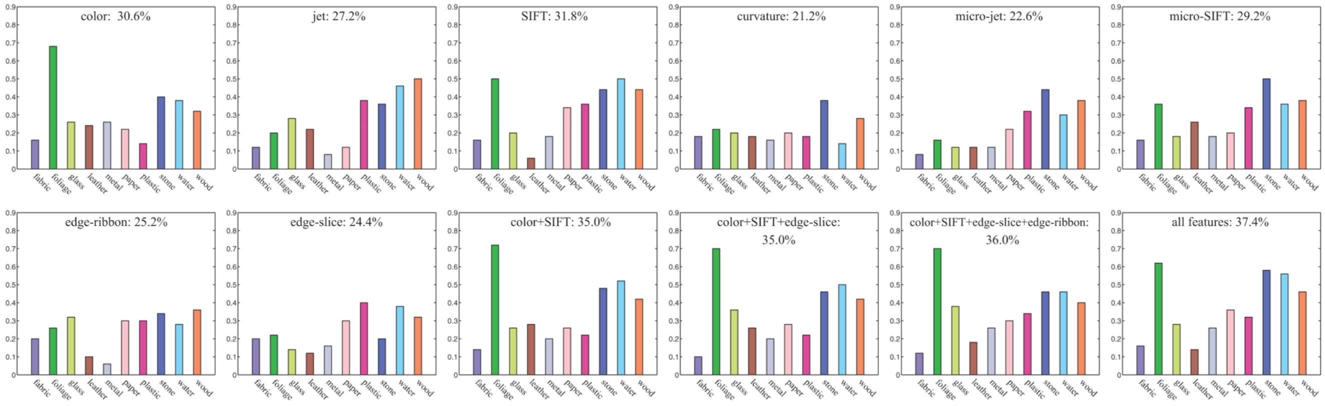


Fig. 15. Comparison to Varma-Zisserman

We ran Varma-Zisserman's system [Varma and Zisserman, 2009], which uses a nearest neighbor classifier, on our feature sets. The training rate for a nearest neighbor classifier is always 100%, so we do not show the training rate in these plots. These results demonstrate the power of using our features vs. fixed-size grayscale patch features (23.8% accuracy) for FMD images.

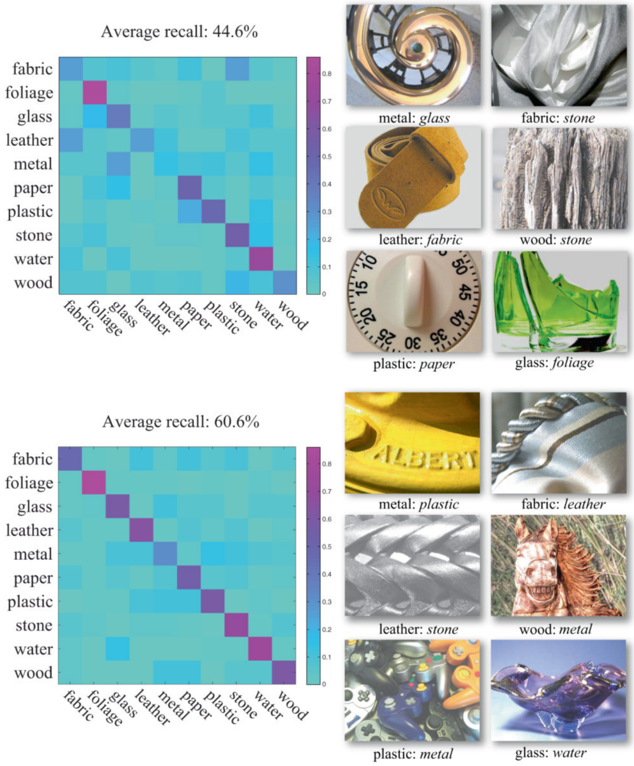


Fig. 16. Confusions made by our LDA-based and SVM-based systems

We present the confusion matrices corresponding to (top panel) our LDA-based and (bottom panel) SVM-based material category recognition systems. The LDA-based system uses *color, SIFT*, and *edge-slice* as features. The SVM-based system uses all eight features. For each confusion matrix, cell (*i, j*) denotes the probability of category *i* being classified as category *j*. On the right, we shows examples of misclassification. Labels of the type "X: *Y*" mean that a surface made of X was misclassified as being made of Y.

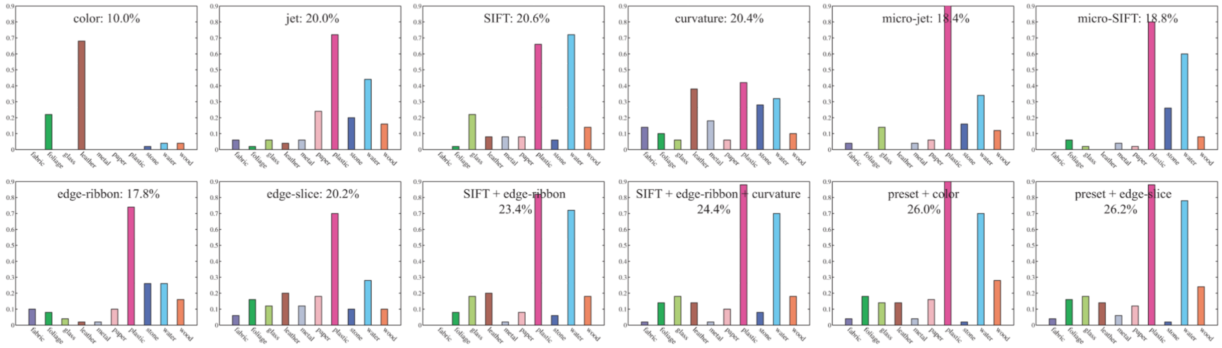


Fig. 17. SVM results on bilateral filtered images

We trained SVM classifiers on the original images (Figure 5a), and we tested them on bilateral filtered images (Figure 5b). These plots are arranged in a similar way to those in Figure 14. The best performance on the bilateral filtered images (26.2%) lags human performance (65.3%) by a large margin.

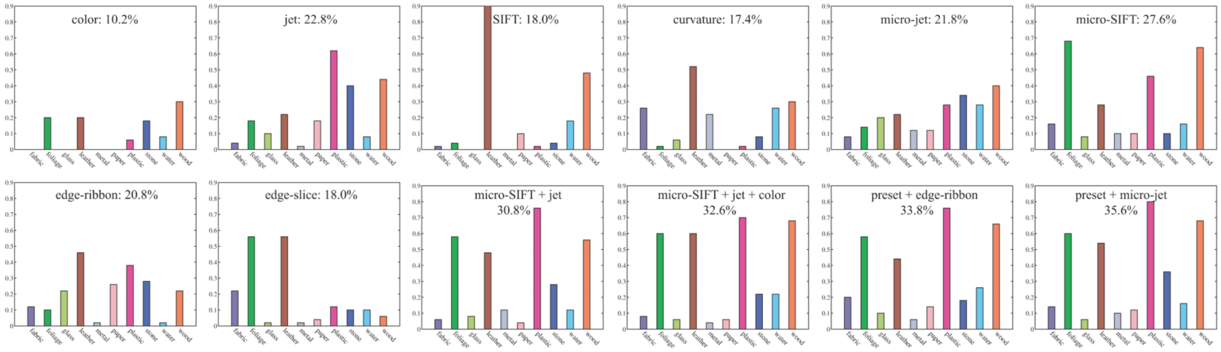


Fig. 18. SVM results on high-pass filtered images

We trained SVM classifiers on the original images (Figure 5a), and we tested them on high-pass filtered images (Figure 5c). The best performance on the high-pass filtered images (35.6%) lags human performance (64.8%) by a large margin.

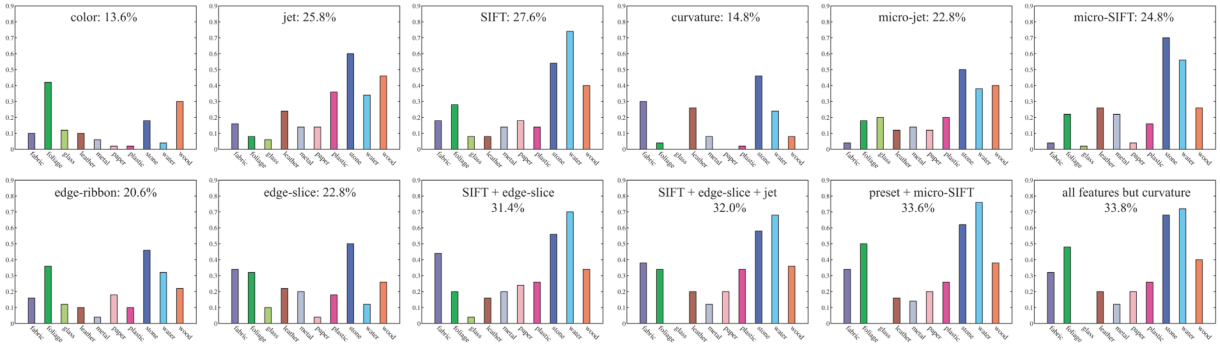


Fig. 19. SVM results on texture synthesized images (15×15)

We trained SVM classifiers on the original images (Figure 5a), and we tested them on texture synthesized images (Figure 5d) with patch size set to 15×15 pixels. The best performance on the texture synthesized images (33.8%) is comparable to human performance (38.7%).

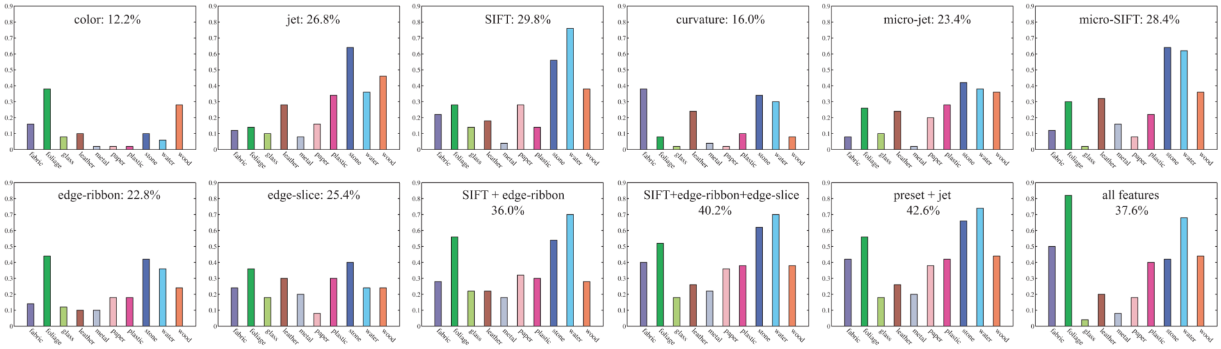


Fig. 20. SVM results on texture synthesized images (30×30)

We trained SVM classifiers on the original images (Figure 5a), and we tested them on texture synthesized images (Figure 5e) with patch size set to 30×30 pixels. The best performance on the texture synthesized images (42.6%) is comparable to human performance (46.9%).

Table 1

For each feature, we list the surface property measured by that feature and the size of image region (in pixels) that the feature is computed over.

Feature name	Surface property	Size of image region
Color	Color	3×3
Jet	Texture	25×25
SIFT	Texture	16×16
Micro-jet	Micro-texture	25×25
Micro-SIFT	Micro-texture	16×16
Curvature	Local shape	16×16
Edge-slice	Reflectance	18×3
Edge-ribbon	Reflectance	18×3

Table 2

The dimension, average number of occurrences per image, and the number of clusters is listed for each feature.

Feature name	Dim	Average # per image	# of clusters
Color	27	6326.0	150
Jet	64	6370.0	200
SIFT	128	6033.4	250
Micro-jet	64	6370.0	200
Micro-SIFT	128	6033.4	250
Curvature	3	3759.8	100
Edge-slice	72	2461.3	200
Edge-ribbon	72	3068.6	200