Automatically Discovering Local Visual Material Attributes

Gabriel Schwartz Ko Nishino
Department of Computer Science, Drexel University
{gbs25,kon}@drexel.edu

Abstract

Shape cues play an important role in computer vision, but shape is not the only information available in images. Materials, such as fabric and plastic, are discernible in images even when shapes, such as those of an object, are not. We argue that it would be ideal to recognize materials without relying on object cues such as shape. This would allow us to use materials as a context for other vision tasks, such as object recognition. Humans are intuitively able to find visual cues that describe materials. Previous frameworks attempt to recognize these cues (as visual material traits) using fully-supervised learning. This requirement is not feasible when multiple annotators and large quantities of images are involved. In this paper, we derive a framework that allows us to discover locally-recognizable material attributes from crowdsourced perceptual material distances. We show that the attributes we discover do in fact separate material categories. Our learned attributes exhibit the same desirable properties as material traits, despite the fact that they are discovered using only partial supervision.

1. Introduction

Computer vision has relied heavily on shape cues. Popular image features, including HoG [7] and SIFT [13], essentially encode object shapes for recognition. Images, however, capture much more than merely the shape of objects in a scene. Real-world images capture the appearance of materials, such as fabric, plastic, and wood, even when the complete shape of the object is not discernible. The recognition of these materials in images can provide crucial information for scene understanding. Recognizing that a cup is made of glass and not plastic, for example, can be used to inform a robot how to properly grasp it.

Recognizing real-world materials has proven challenging. Materials exhibit a wide variety of appearances, particularly at a local (image-patch) level. Current methods [5, 12, 18] for material recognition address this challenge by making single image-wide material predictions or by using very large image patches that in fact contain all or most of the object extents. These approaches inevitably use object cues (*i.e.*, their shapes) to recognize materials. It is unclear whether plastic is recognized because it looks like a cup, or because of its characteristic material appearance. This is undesirable, as it will prevent the use of material information for scene understanding tasks including object recognition. Why would knowing that an object is made of plastic help recognize that it is a cup if recognition of plastic relies on knowing that it is a cup?

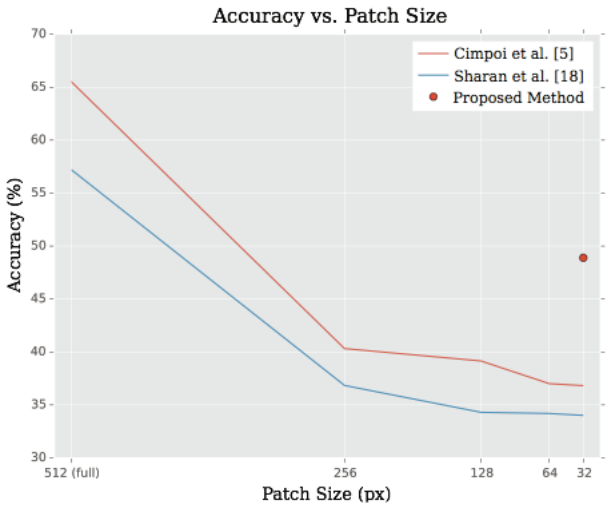


Figure 1. Local material recognition is challenging. When adapted to use aggregated features from local image patches, methods that perform well on full images quickly lose accuracy. Previous work has shown that using material traits as an intermediate representation addresses this issue [17]. We propose a method for weakly-supervised discovery of such traits.

We argue that material recognition should be performed without relying on object cues such as shape. We would instead like to be able to locally recognize characteristic material appearance. If we could do that, we could aggregate local information inside an object region to predict what material the object is made of. This is a challenging task, and all previous methods essentially rely on object cues. Figure 1 clearly shows that when existing methods[1]

[1]We implement the Improved Fisher Vectors of [5] and all features of Sharan *et al*. [18] and aggregate them across local patches within object regions to generate the results in Figure 1.

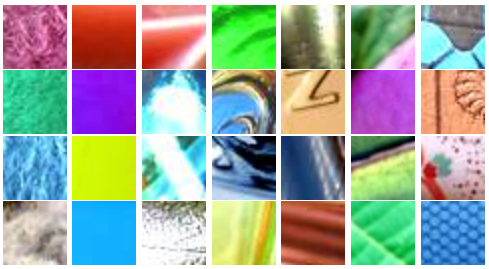


Figure 2. Sample material image patches. Asking annotators to merely "describe" the patches is an ambiguous question. Patches may look similar even though the annotator does not have a concrete word to define the similarity. We instead ask only for binary visual similarity decisions.

are restricted to use only local image patches they perform increasingly poorly.

We intuitively expect to find some form of local visual cues that indicate material without relying on a global view of an image. Schwartz and Nishino [17] indeed demonstrate this by introducing locally-recognizable material attributes which they refer to as "material traits." Material traits enable recognition of material categories using features extracted from patches as small as 32×32 pixels. They describe materials based on characteristic appearances like shiny and smooth. Fabric, for example, can appear woven or smooth, but not liquid. When taken as an aggregate across an image region, their traits can be used to identify material categories. Since material traits do not imply the presence of any particular object, they can be used to perform object-independent material recognition.

Material trait recognition, as previously proposed, relies on a set of fully labeled material trait examples. This assumption hinders scaling the method with larger training datasets. We hardly have a mutually agreeable vocabulary for describing materials and their visual characteristics. Considering the images in the first column of Figure 2, for instance, one may call them fuzzy and others may call them fluffy. People may also be inconsistent in annotating material traits. Some may only annotate the patches in the second column as smooth and others may only see them as translucent. Cimpoi *et al.* [5] alleviate these problems for texture recognition by preparing a pre-defined vocabulary. They may do so by focusing on apparent texture patterns like stripes and dots. Materials underlie these texture patterns (*i.e.*, the stripes or dots on a plastic cup are still plastic) and do not follow such a vocabulary.

We propose to automatically discover locally-recognizable material attributes. We achieve this by exploiting human perception of visual material similarity. Humans are able to reliably assess material similarity from local visual information [10]. We hypothesize that people judge material similarity based on characteristics equivalent to visual material traits. For instance, a person would perceive an image patch of wool to be similar to that of sheep fur as both look fuzzy. Humans can look at the images in Figure 2 and see that images in each column share visual properties without necessarily being able to identify them. By analyzing human assessments of the visual similarities of different local material image patches, we should be able to build a classifier to recognize these implicit local visual attributes. In this work we show that such assessments can easily and reliably obtained.

We use crowdsourcing (Amazon Mechanical Turk) to determine the visual similarity of material categories as seen by humans. For this, we show image patches of different materials as references and ask whether other image patches from other materials look similar. These results are aggregated to compute pairwise visual distances between material categories. The idea is to identify a space of material attributes that preserves these pairwise distances while permitting reliable recognition of the attributes on local image patches. For this, we first convert the distance matrix into a category-attribute matrix that realizes desirable characteristics such as sparsity. We then train a joint attribute classifier that predicts, on average for each category, the desired attribute likelihoods. Our formulation requires no supervised labeling of attributes on training data.

There are an infinite number of random local attributes which can be used to recognize materials. In our work, we are specifically discovering those that underlie human perception. We show that humans agree on a common perception of similar materials, and that we can in fact encode this perception in our discovered attributes. The discovered attributes show clear correlations with known visual material traits [17]. Due to the constraints we impose on the learning process, the discovered attributes also exhibit similar spatial sparsity patterns to those that are characteristic of known traits. This is in contrast with random attributes which exhibit no such patterns. Our framework requires only simple annotations that can be quickly and consistently collected. Unlike previous methods, we do not rely on visual properties that are only visible at a global image level.

2. Related Work

The majority of existing methods for attribute-based recognition, whether that be for objects or materials, rely on a pre-defined set of attributes that are relevant to the specific task in question [4, 8, 9]. Such approaches benefit from the fact that we can intuitively understand the attributes used for recognition. All of these approaches are task-specific and would require manual definition of a new set of attributes for each new task. Our approach requires minimal annotation, and the attributes arise entirely from the inherent visual similarity of the relevant categories. While we focus on materials in this paper, our method is not task-specific.

Berg *et al.* [3] describe a framework for learning object

attributes from web data (images and associated text). This approach learns some localized attributes, however the required text annotations are image-wide and do not guarantee locality. Recently, Patterson and Hays [14] proposed a process to discover and recognize scene-wide attributes in natural images. While they are able to discover a large amount of attributes, their learned attributes are not local. Rastegari *et al*. [15] learn a binary attribute representation (binary codes) for images. As with existing methods, however, these attributes are image-wide and not local.

Cimpoi *et al*. [5] demonstrate a method for learning an arbitrary set of describable texture attributes based on terms derived from psychological studies. As noted by Adelson [1], texture is only one component of material appearance, and cannot alone describe our perception of materials. Though their results demonstrate impressive performance on the Flickr Materials Database [19], their learned attributes apply only globally. Our proposed formulation explicitly models the desired properties of the category-attribute space at an image patch level.

Akata *et al*. [2] formulated attribute discovery as a label embedding problem. Yu *et al*. [22] propose a two-step procedure for discovering and classifying attributes based on a similarity matrix. They propose to compute a distance matrix using Euclidean distances in the raw feature space of labeled image patches. Our attribute discovery process can be viewed as a form of label embedding. We cannot, however, simply use raw features to define similarities. We show that raw features do not encode the human perceptual information required to discover intuitive material traits.

We rely on human perception of material similarity to discover meaningful local material attributes. Wills *et al*. [21] have proposed a relevant procedure for measuring human perception of gloss. They measure perceptual gloss via non-metric multidimensional scaling applied to a series of relative comparisons. Their proposed method is, however, limited to measurement of one property (gloss). It does not apply to material recognition and is not a local process.

3. Perceptual Distance between Materials

Our goal is to discover a set of attributes that exhibit the desirable properties of material traits. We want to achieve this without relying on fully-supervised learning. Known material traits, such as "smooth" or "rough," represent visual properties shared between similar materials. We expect that attributes that preserve this similarity will satisfy our goal. We propose to define a set of attributes based on the perceived distances between material categories. By working with distances rather than similarities, we avoid any need to assume a particular similarity function. For this, we obtain a measurement of these distances from human annotations.

From a high-level perspective, our attribute discovery consists of three steps:

1. Measure perceptual distances between materials
2. Define an attribute space based on perceptual distances
3. Train classifiers to reproduce this space from image patches

Defining perceptual distance between material categories poses a challenge. If each material had a single typical appearance (*e.g.,* if metal was always shiny and gray), we could simply compute the difference between these typical appearances. This is not the case. Materials may exhibit a wide variety of appearances, even sharing appearances between categories (what we refer to as material appearance variability). An image patch from a leaf, for example, may appear similar to certain fabrics or plastics.

Directly measuring distances via human annotation would be ideal, as we have an intuitive understanding of the differences between materials. As Sharan *et al*. [18] showed, this understanding persists even in the absence of object cues. It is, however, also a difficult task to obtain these distance. Given two query image patches, annotators would have to decide how different the patches look on a consistent quantitative scale. We would instead like to ask simple questions that can be reliably answered.

We propose that instead of asking how different patches look, we reduce the question to a binary one: "Do these patches look different or not?" We assume that this will give us sufficient information to obtain consistent and sensible perceptual information. Our underlying assumption for this claim is that if a pair of image patches look similar, they do so as a result of at least one shared visual material trait.

To transform a set of binary similarity annotations into pairwise distances, we represent each material as a point defined by the average probabilities of similarity to each material category. The pairwise distances between these points define the material perceptual distance matrix. This process treats each material category as a point in a space of typical (but not necessarily realizable) material appearances. The resulting distance between a pair of materials depends on joint similarity with all material categories, including the pair in question, and is thus robust to material appearance variability.

Formally, given a set of N reference images with material category $c_n \in \{1 \dots K\}$, we obtain binary similarity decisions $\mathbf{s}_n \in \{0,1\}^K$ for each reference image against a set of sample images from each category. We represent each material category in the space of typical material category appearances as K-dimensional vectors $\mathbf{p}_k$:

$$\mathbf{p}_k = \frac{1}{N_k} \sum_{n|c_n=k} \mathbf{s}_n, \tag{1}$$

where $N_k = |\{c_n | c_n = k\}|$. Entries $d_{kk'}$ in the $K \times K$ pairwise distance matrix D are then defined as:

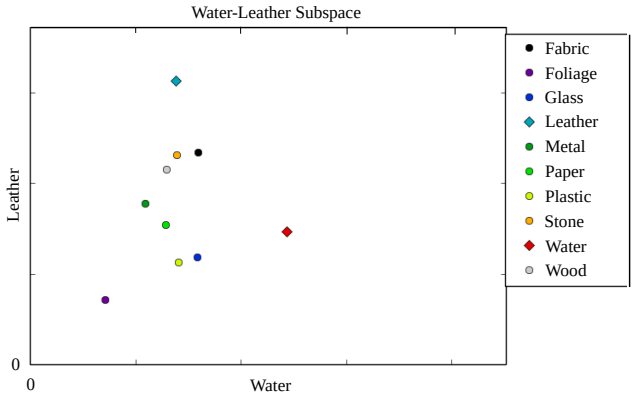


Figure 3. Example projections of materials into a 2D similarity subspace. The locations of the two material categories corresponding to the axes are marked. We would expect that, in this case, water would lie furthest along the "water" axis and likewise with leather. Materials with common visual properties, such as the smoothness of plastic and glass, lie close to each other. Materials with distinct visual properties, such as woven fabric and shiny metal, do not.

$$d_{kk'} = \|\mathbf{p}_k - \mathbf{p}_{k'}\|_2 . \tag{2}$$

We obtain the required set of binary similarity annotations through Amazon Mechanical Turk (AMT). Each task presents annotators with a reference image patch of a given material category (unknown to the annotator) and a row of random image patches, one from each material category. We use patches from images of the 10 material categories from the Flickr Materials Database of Sharan *et al*. [19]. Annotators are directed to select image patches that look similar to the reference. Examples of suggested similar image patches are given based on known material traits. Each set of patches is shown to 10 annotators, and final results are obtained from a vote where at least 5 annotators must agree that the patches look similar. We collect similarity decisions for 10,000 reference image patches.

The 2D projection in Figure 3 shows that the similarity values obtained from the AMT annotations agree with our own intuitive understanding of material appearance. The plot shows the locations of material categories projected into one 2D subspace of the 10-dimensional space of typical material appearances. We would expect that the two materials corresponding to the typical materials in each subspace will lie close to their respective axes. In this case, water is most similar with itself, but is also similar to glass. Leather is likewise most similar with itself, but also similar to fabric.

To show that we do in fact obtain a consistent distance matrix, we compute the difference between the distance matrix computed with all annotations versus that from only n of the N total annotations. The difference drops quickly (within the first few hundred samples of 10,000), showing that annotators agree on a single common set of perceptual distances.

4. Defining the Material Attribute Space

Discovering attributes given only a desired distance matrix poses a challenge. A straightforward approach would be to directly train classifiers to predict attributes that encode the distance matrix. This would be a particularly under-constrained problem as we do not even know which attributes to associate with which categories.

We instead propose to separate attribute association and classifier learning into two steps. First, we discover attributes in an abstract form by discovering a mapping between categories and attribute probabilities. We ensure that the mapping preserves the pairwise perceptual material distances, and then train classifiers to predict the presence of these attributes on image patches.

As described in Section 3, we obtain a distance matrix $\mathbf{D}$ from crowdsourced similarity answers for K material categories $C = \{1 \dots K\}$. Using $\mathbf{D}$, we find a mapping that indicates which attributes are associated with which categories. The number of attributes we discover is arbitrary, and we refer to it as M. The mapping is encoded in the $K \times M$ category-attribute matrix $\mathbf{A}$. We restrict values in $\mathbf{A}$ to lie in the interval $[0, 1]$ so that we may treat them as conditional probabilities.

We impose two constraints on the category attribute mapping. $\mathbf{A}$ should map categories to attributes in a way that preserves the measured distances in $\mathbf{D}$, and the mapping should contain realizable values. If the values in $\mathbf{A}$ are not plausible, we will not be able to recognize the attributes on image patches. For example, one potential attribute mapping would be to assign each attribute to a single category. Attribute recognition then becomes the same as the intractable probelm of material category recognition on single image patches.

We formulate the attribute discovery process as a minimization problem over category-attribute matrices $\mathbf{A}$:

$$\mathbf{A}^* = \arg\min_{\mathbf{A}} d\left(\mathbf{D}; \mathbf{A}\right) + w_A \kappa_A\left(\mathbf{A}\right) \tag{3}$$

with hyperparameter w_A. d describes how well the current estimate of $\mathbf{A}$ encodes the pairwise perceptual differences between material categories, and κ_A is a constraint that makes the discovered attribute associations exhibit a realizable distribution.

The category-attribute matrix that best encodes the desired pairwise distances will minimize the following term defined over rows $\mathbf{a}_k$ of the matrix $\mathbf{A}$:

$$d\left(\mathbf{D}; \mathbf{A}\right) = \sum_{k,k' \in C} \left(\|\mathbf{a}_k - \mathbf{a}_{k'}\|_2 - \mathbf{D}_{kk\prime}\right)^2 . \tag{4}$$

To discover realizable attributes, we encode our own prior knowledge that recognizable attributes exhibit a particular distribution and sparsity pattern. We observe that

semantic attributes, specifically visual material traits, have a Beta-distributed association with material categories [17]. Generally, a material category will either strongly exhibit a trait or it will not exhibit it at all. Intermediate cases occur when a material category exhibits a particularly wide variation in appearance. Fabric, for example, sometimes has a clear "woven" pattern but, in the case of silk or other smooth fabrics, does not. We would like the values in $\mathbf{A}$ to be Beta-distributed to match the distribution of known material trait associations.

The canonical method for matching two distributions is to minimize a divergence measure between them. To incorporate this into a minimization formulation, we need a differentiable measurement for the unknown empirical distribution of values in $\mathbf{A}$. We choose the KL-divergence and Gaussian kernel density estimator. The Gaussian kernel density estimate at point p is:

$$q\left(p;\mathbf{A}\right)=\frac{1}{KM}\sum_{k,m}\left(2\pi h^{2}\right)^{-\frac{1}{2}}\exp\left\{-\frac{\left(a_{km}-p\right)^{2}}{2h^{2}}\right\} \tag{5}$$

The KL-divergence between the distribution of the values in the category-attribute matrix $\mathbf{A}$ and the target Beta distribution $\beta\left(p;a,b\right)$ with $a=b=0.5$ can then be written as:

$$\kappa_{A}\left(\mathbf{A}\right)=\sum_{p\in P}\beta\left(p;a,b\right)\ln\left(\frac{\beta\left(p;a,b\right)}{q\left(p;\mathbf{A}\right)}\right). \tag{6}$$

5. Training a Material Attribute Classifier

We now must derive classifiers that recognize the attributes defined by the category-attribute mapping. As attributes are not defined semantically, we cannot ask for further annotation to label training patches with attributes. Instead, we propose a model and a set of constraints that will enable us to predict our discovered attributes on material image patches.

We do not know *a priori* any particular semantics or structure associated with the attributes, thus we model our attributes using a general two-layer non-linear model [6]. We constrain the predictions such that they reproduce the desired values in the attribute matrix (in expectation) while also separating material categories when possible.

Formally, given a training set of N image patches represented by D-dimensional raw feature vectors $\mathbf{x}_n$ with corresponding material categories $c_n \in C$, we train a model f with parameters $\mathbf{\Theta}$ that maps an image patch to M attribute probabilities: $f\left(\mathbf{x}_n;\mathbf{\Theta}\right):\mathbb{R}^{D}\rightarrow\left[0,1\right]^{M}$. Given an intermediate layer with dimensionality H and parameters $\mathbf{W}_1\in\mathbb{R}^{H\times D}$, $\mathbf{W}_2\in\mathbb{R}^{M\times H}$, $\mathbf{b}_1\in\mathbb{R}^{H}$, $\mathbf{b}_2\in\mathbb{R}^{M}$ the prediction for an instance $\mathbf{x}_n$ is defined as:

$$\begin{aligned} f\left(\mathbf{x}_n;\mathbf{\Theta}\right) &= h\left(\mathbf{W}_2 h\left(\mathbf{W}_1\mathbf{x}_n+\mathbf{b}_1\right)+\mathbf{b}_2\right) \\ h\left(x\right) &= \min\left(\max\left(x,0\right),1\right). \end{aligned} \tag{7}$$

As additional regularization, used only during training, we mask out a random fraction of the weights used in the model to discourage overfitting (akin to dropout [11]).

We formulate the full classifier training process as a minimization problem:

$$\mathbf{\Theta}^{\star}=\arg\min_{\mathbf{\Theta}} r\left(\mathbf{X};\mathbf{A},\mathbf{\Theta}\right)+w_1\kappa\left(\mathbf{X};\mathbf{\Theta}\right)-w_2\pi\left(\mathbf{X};\mathbf{A},\mathbf{\Theta}\right), \tag{8}$$

with hyperparameters w_1 and w_2. r (Equation 9) is a data term indicating the difference between predicted and expected attribute probabilities. κ and π (Equations 10 and 11) are, respectively, constraints on the the distribution of attribute predictions and on the pairwise separation of material categories.

The category-attribute matrix encodes the probabilities that each category will exhibit each attribute. We represent this in our classifier training by matching the mean predicted probability for each attribute to the given entry in the category-attribute matrix:

$$r\left(\mathbf{X};\mathbf{A},\mathbf{\Theta}\right)=\sum_{k\in C}\left\|\mathbf{a}_k-\frac{1}{N_k}\sum_{i|c_i=k}f\left(\mathbf{x}_i;\mathbf{\Theta}\right)\right\|_{2}^{2}. \tag{9}$$

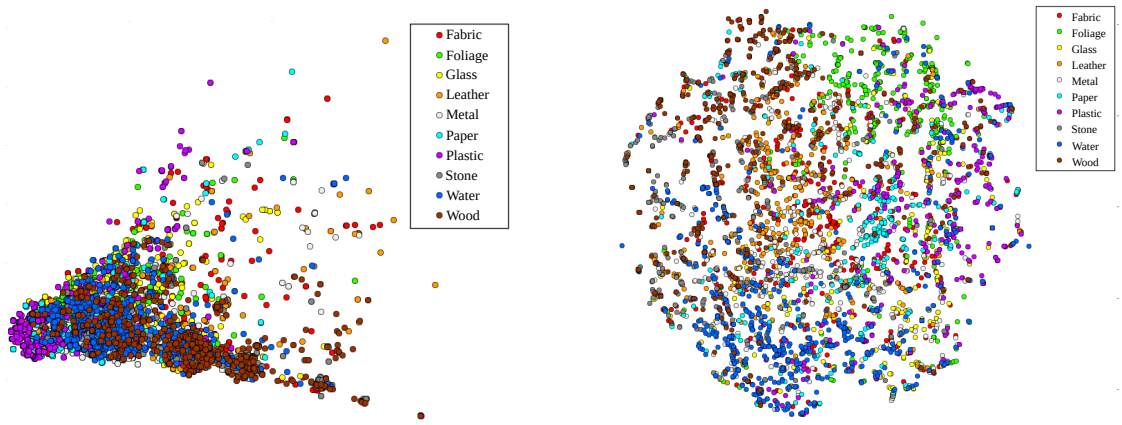
Equation 9 directly encodes the desired behavior of the classifier, but it alone is under-constrained. Each prediction for each instance may take on any value so long as their mean matches the target value.

We have observed that, similar to category-attribute associations, predicted probabilities for known material traits are also Beta-distributed. Local image regions exhibiting a trait will have uniformly high probability for that trait, only decreasing around the trait region edges. We constrain the predicted probabilities such that they are Beta-distributed. Using the formulation discussed in Section 4, we again minimize a KL-divergence of a kernel density estimate:

$$\kappa\left(\mathbf{X};\mathbf{\Theta}\right)=\sum_{p\in P}\beta\left(p;a,b\right)\ln\left(\frac{\beta\left(p;a,b\right)}{q\left(p;f\left(\mathbf{X};\mathbf{\Theta}\right)\right)}\right), \tag{10}$$

where $f\left(\mathbf{X};\mathbf{\Theta}\right)$ represents the $N\times M$ matrix of attribute probability predictions for the training dataset, and q,a,b are defined as in Equation 6.

One of the goals for our attribute representation is to discover attributes that allow for material classification. If this were our only goal, we could simply maximize the distance between the predicted attributes for all pairs of different material categories. This would conflict with our goal of preserving human perception, as material categories do not always exhibit different appearances. We instead modify this separation by weighting each component of the distance based on the values in the category-attribute matrix:



(a) Raw Features (b) Discovered Attribute Predictions

Figure 4. t-SNE [20] embedding of materials from the raw feature [17] space (a) and from our discovered attributes (b). We embed a set of material image patches into 2D space via t-SNE using raw features and predicted attribute probabilities as the input space for the embeddings. Though t-SNE has been shown to perform well in high-dimensional input spaces, it fails to separate material categories from the raw feature space. Material categories are, however, clearly more separable with our attribute space.

$$\pi\left(\mathbf{X};\mathbf{A},\boldsymbol{\Theta}\right) = \sum_{i,j\in N|c_i\neq c_j} \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{p}_{ij} \tag{11}$$

$$\mathbf{p}_{ij} = \left(2\left|\mathbf{a}_{c_i} - \mathbf{a}_{c_j}\right| - 1\right)\left(f\left(\mathbf{x}_i;\boldsymbol{\Theta}\right) - f\left(\mathbf{x}_j;\boldsymbol{\Theta}\right)\right).$$

This separates the material categories in attribute space only when the attributes dictate that there is a perceptual difference.

6. Analysis of Discovered Attributes

To analyze the properties of attributes discovered by our framework, we follow the procedures outlined above to collect annotations and discover a set of attributes. Since both learning steps involve minimization of a non-linear, non-convex function, we rely on existing optimization tools[2] to find suitable estimates. As a raw feature set, we use the local features of Schwartz and Nishino [17].

If our attributes described a space that successfully separates material categories, we would expect categories to form clusters in the attribute space. To verify this, we compute a 2D embedding of a set of labeled image patches. For the embedding, we use the t-SNE method of van der Maaten and Hinton [20]. t-SNE attempts to generate an embedding that matches the distributions of neighboring points in the high- and low-dimensional spaces. In Figure 4, we represent image patches by their raw feature vectors (a) and predicted attribute probability vectors (b), and compare the 2D embeddings resulting from each. Material categories are separated much more clearly in our attribute space than in the raw feature space.

Part of the usefulness of visual material traits, as shown in [17], is derived from the fact that they each represent a particular intuitive visual material property. This is evident in the spatial sparsity pattern of the traits, specifically the fact that they appear in regions and not randomly within an image. Traits such as "shiny" are highly localized, while others such as "woven" or "smooth" exist as coherent regions within a particular material instance. Figure 5 shows examples of per-pixel attribute probabilities predicted from our discovered attribute classifiers. The attributes exhibit both sparse and dense spatial patterns that are consistent within local regions. Dense attributes generally correspond with smooth image regions. Sparse attributes often indicate localized surface features such as specific texture patterns.

For comparison, in Figure 6 we visualize per-pixel predictions for an attribute classifier trained on a random attribute matrix $\mathbf{A}$. Unlike attributes based on human perception, these random attributes do not exhibit the same meaningful spatial consistency.

We aimed to discover attributes similar to the visual material traits that underlie human perception. We thus expect that the discovered attributes exhibit a correlation with known traits. Figure 7 shows the correlation between 13 discovered attributes and 13 known material traits using attributes predicted on labeled material trait image patches. Collectively, we can indeed describe material traits using the discovered attributes. Visually similar traits, such as rough and woven, show similar correlations with the attributes. Discovered attributes are also consistent with the semantic properties of material traits. Rough and smooth are mutually exclusive traits, and we see that discovered attributes that positively correlate with smooth do not generally correlate with rough.

We quantitatively evaluate the discovered attributes using logic regression [16]. Given a set of image patches with known traits, we predict our discovered attributes as binary values for use as input variables in a logic regression model

[2]Specifically, L-BFGS with box constraints for $\mathbf{A}$ and stochastic gradient descent for $\boldsymbol{\Theta}$.

Figure 5. Per-pixel discovered attribute probabilities for four attributes (one per column). These images show that the discovered attributes exhibit patterns similar to those of known material traits. The first attribute, for example, appears consistently within the woven hat and the koala; the second attribute tends to indicate smooth regions. The last two columns show we are discovering attributes that can appear both sparsely and densely in an image, depending on the context. These are all properties shared with visual material traits.

Figure 6. Typical per-pixel attribute probabilities based on a random attribute matrix. Unlike the predictions for attributes derived from human perception, these attributes appear randomly within a region and do not reflect any local visual properties.

for material traits. Logic regression from 30 attributes alone (no other features) achieves comparable accuracy (75% vs. 77%) to the method of [17] using a complex feature set. These results show that the discovered attributes do collectively encode intuitive visual material properties.

7. From Discovered Attributes to Materials

Seeing that discovered attributes encode visual material properties, we would expect them to also serve as an intermediate representation for material category recognition.

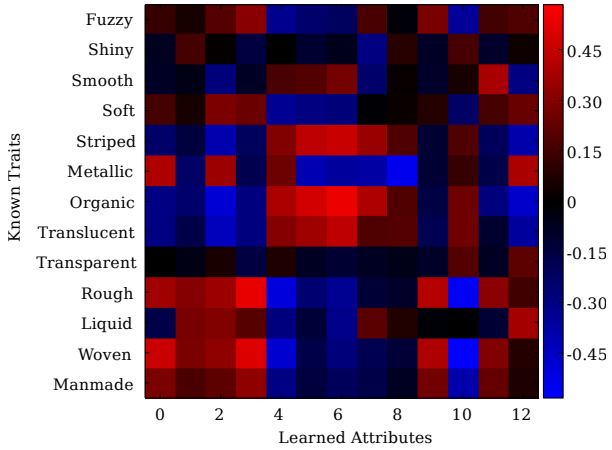


Figure 7. Correlation between discovered attribute predictions and material traits. Groups of attributes can collectively indicate the presence of a material trait. Metallic, for example, correlates positively with attribute 0 and negatively with attribute 8.

To test this, we follow the material recognition procedure described by Schwartz and Nishino [17], substituting our discovered attributes for their labeled material traits. We

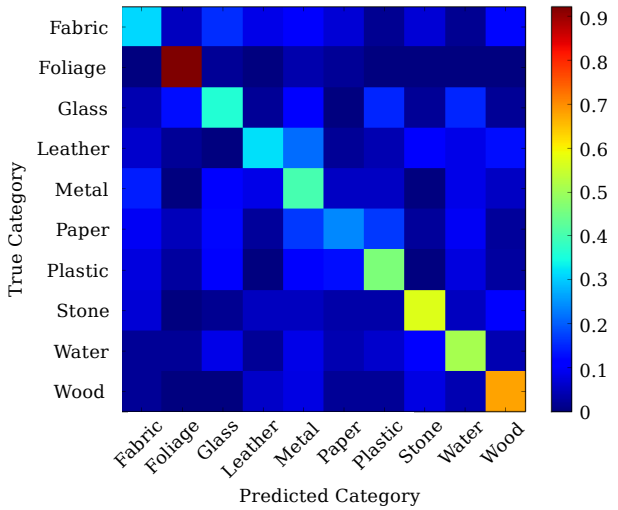


Figure 8. Confusion matrix for material recognition on FMD images. Well-recognized categories, such as foliage, correspond with categories that appeared distinct in human annotations for perceptual distance. Annotators regularly selected foliage patches as appearing different from all other categories.

compute the histograms of these predicted probabilities across the material region and use them as input for a histogram kernel SVM. As we focus on local attributes, these previous local results (and those of Sharan *et al*. [18] on scrambled images) serve as the correct baseline.

For comparison with Schwartz and Nishino [17], we compute average material recognition accuracy on the Flickr Materials Database (FMD). All results are computed using $M = 30$ discovered attributes and 5-fold cross-validation unless otherwise specified.

Our attributes achieve an average accuracy of 48.9% ($\sigma = 1.2\%$) on FMD images using only local information. This is comparable to results reported by Schwartz and Nishino [17] and Sharan *et al*. [18] (using only local information) even though we are discovering attributes using only weak supervision.

Figure 8 shows a confusion matrix for FMD images. In agreement with previous work, metal is the most challenging category to identify. Foliage is very well-recognized. This follows from the results of our measurements of human perception, as annotators consistently found that foliage image patches looked different from all other material categories. Fabric was previously somewhat challenging to recognize locally [17], and we see that paper is also challenging in this case. It is possible that subtle cues separating paper and plastic were not visible to the annotators.

Figure 9 shows that accuracy reaches a plateau as the training dataset size increases. We also compute accuracy for varying values of M and find that past $M = 30$, there is little (<0.1%) gain in accuracy from additional attributes. These plateaus indicate that we are in fact extracting as much perceptual material information as we can from the available data.

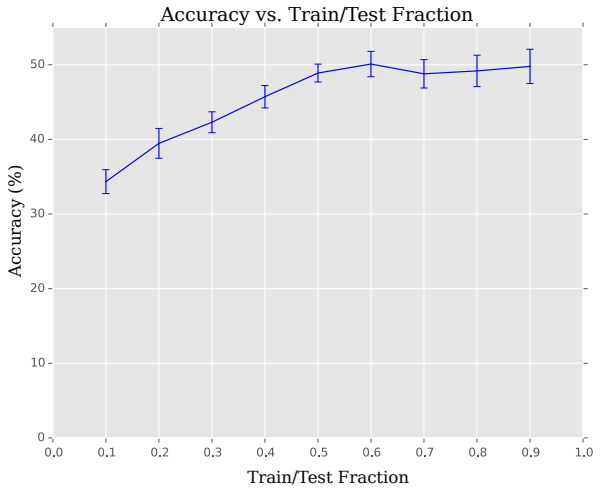


Figure 9. Accuracy vs. training set size. Accuracy does not continue to increase as we use larger training datasets. This shows that we have successfully extracted as much local information as possible from human perception.

8. Conclusion

We introduced a local attribute discovery and recognition framework for visual material attributes. We measure distances between material categories as perceived by humans, and use these distances to discover and recognize a set of attributes. We accomplish this using only simple annotations that can be quickly and consistently collected in large amounts. Our framework results in automatically discovered local material attributes that encode human material perception while still proving useful for material recognition from local information. This is in contrast to random attributes which do not exhibit such properties.

By embedding material image patches into a 2D space using our discovered attributes, we see that they do in fact separate materials into distinct clusters in a way that raw features cannot. Per-pixel visualizations show that the discovered attributes exhibit the same desirable properties as those of visual material traits. We have shown that it is possible to recognize materials using only local information by exploiting human perception. Using only basic annotations, we achieve the same accuracy as existing methods which rely on exhaustive per-patch material trait annotations. We expect our perceptually-discovered attributes to prove useful in further scene understanding tasks.

Acknowledgements

This work was supported by the Office of Naval Research grant N00014-14-1-0316 and the National Science Foundation awards IIS-0964420, IIS-1353235 and IIS-1421094.

References

[1] E. H. Adelson. On Seeing Stuff: The Perception of Materials by Humans and Machines. In *SPIE*, pages 1–12, 2001.

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-Embedding for Attribute-Based Classification. pages 819–826, 2013.

[3] T. L. Berg, A. C. Berg, and J. Shih. Automatic Attribute Discovery and Characterization from Noisy Web Data. In *ECCV*, pages 1–14, 2010.

[4] H. Chen, A. Gallagher, and B. Girod. Describing Clothing by Semantic Attributes. In *Proceedings of the European Conference on Computer Vision*, 2012.

[5] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. *arXiv*, abs/1311.3618, 2013.

[6] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control Signals and Systems MCSS*, 2(4):303–314, 1989.

[7] N. Dalal and W. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, pages 886–893, 2005.

[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing Objects by their Attributes. In *CVPR*, pages 1778–1785, 2009.

[9] V. Ferrari and A. Zisserman. Learning Visual Attributes. In *NIPS*, pages 433–440, 2007.

[10] R. W. Fleming, R. O. Dror, and E. H. Adelson. Real-world Illumination and the Perception of Surface Reflectance Properties. *Journal of Vision*, 3(5):347–368, 2003.

[11] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving Neural Networks by Preventing Co-adaptation of Feature Detectors, 2012.

[12] D. Hu, L. Bo, and X. Ren. Toward Robust Material Recognition for Everyday Objects. In *BMVC*, pages 48.1–48.11, 2011.

[13] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[14] G. Patterson and J. Hays. SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. In *CVPR*, 2012.

[15] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute Discovery via Predictable Discriminative Binary Codes. In *ECCV*, pages 876–889, 2012.

[16] I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic Regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, 2003.

[17] G. Schwartz and K. Nishino. Visual Material Traits: Recognizing Per-Pixel Material Context. In *Color and Photometry in Computer Vision (Workshop held in conjunction with ICCV'13)*, 2013.

[18] L. Sharan, C. Liu, R. Rosenholtz, and E. H. Adelson. Recognizing Materials Using Perceptually Inspired Features. *International Journal of Computer Vision*, 2013.

[19] L. Sharan, R. Rosenholtz, and E. Adelson. Material Perception: What Can You See in a Brief Glance? *Journal of Vision*, 9(8):784, 2009.

[20] L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. *JMLR*, 9:2579–2605, 2008.

[21] J. Wills, S. Agarwal, D. Kriegman, and S. Belongie. Toward a Perceptual Space for Gloss. *ACM Transactions on Graphics*, 28(103):1–15, 2009.

[22] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang. Designing Category-Level Attributes for Discriminative Visual Recognition. pages 771–778, 2013.