Material Classification based on Training Data Synthesized Using a BTF Database

Michael Weinmann, Juergen Gall, Reinhard Klein

Institute of Computer Science, University of Bonn

Abstract. To cope with the richness in appearance variation found in real-world data under natural illumination, we propose to synthesize training data capturing these variations for material classification. Using synthetic training data created from separately acquired material and illumination characteristics allows to overcome the problems of existing material databases which only include a tiny fraction of the possible real-world conditions under controlled laboratory environments. However, it is essential to utilize a representation for material appearance which preserves fine details in the reflectance behavior of the digitized materials. As BRDFs are not sufficient for many materials due to the lack of modeling mesoscopic effects, we present a high-quality BTF database with 22,801 densely measured view-light configurations including surface geometry measurements for each of the 84 measured material samples. This representation is used to generate a database of synthesized images depicting the materials under different view-light conditions with their characteristic surface geometry using image-based lighting to simulate the complexity of real-world scenarios. We demonstrate that our synthesized data allows classifying materials under complex real-world scenarios.

Keywords: material classification, material database, reflectance, texture synthesis

1 Introduction

Image-based scene understanding depends on different aspects such as the detection, localization and classification of objects. For these tasks, it is essential to consider characteristic object properties like shape or appearance. While its shape tells us how to grasp a particular object, its material tells us how fragile, deformable, heavy, etc. it might be and hence, how we have to handle it. The understanding of the recognized surface material thus guides the interaction of humans with the corresponding object in daily life, and it also represents a key component regarding industrial applications. However, image-based classification of materials in real-world environments is a challenging problem due to the huge impact of viewing and illumination conditions on material appearance. Therefore, training an appropriate classifier requires a training set which covers all these conditions as well as the intra-class variance of the materials.

So far, there have been two main approaches to generate suitable training sets. One approach is to capture a single representative per material category under a multitude of different conditions, such as scale, illumination and viewpoint, in a controlled setting [8,13,7,18] (see Table 1). However, the measured viewing and illumination configurations are rather coarse and hence not descriptive enough to capture the mesoscopic effects in material appearance, which consider the light interaction with material surface regions mapped to approximately one pixel, in an accurate way. In addition, the material samples are only measured under controlled illumination or lab environments which does not generalize to material appearance under complex real-world scenarios. As an alternative, the second category of methods uses images acquired under uncontrolled conditions. In [25], images from an internet image database (Flickr) have been used. This has the advantage that both the intra-class variance of materials and the environment conditions are sampled in a representative way. Unfortunately, the images have to be collected manually, and the materials appearing in the images have to be segmented and annotated. The necessary effort again severely limits the number of configurations that can be generated this way (see Table 1).

In this paper, we instead make use of synthesized data which has already been explored for different applications (e.g. [11,21,29,28,27,19,2,3]). In particular, separately acquired material characteristics and illumination conditions offer the possibility to create synthetic training data for material classification that capture the variations of real-world data. This decoupling of the sampling of material from environment conditions allows us to overcome the limitations of existing material databases that contain only a few hundred configurations of viewing and lighting conditions per material category. For these synthetic images, perfect segmentations are directly available without the need for manual segmentation, and a huge number of them can be obtained easily and fully automatically. This approach requires creating realistic renderings, which accurately simulate the appearance of a material in a real-world scenario. In particular, the appearance of many daily life materials like cloth, skin, etc. is determined by effects taking place on surface structures mapped to a size of approximately 1 pixel (e.g. scratches or fibers) such as subsurface scattering, interreflections, self-shadowing and self-occlusion. These effects cannot be modeled by standard Bidirectional Reflectance Distribution Function (BRDF) models, which are suitable especially for locally smooth surfaces like plastic or metal as these fulfill the assumption of a homogeneous surface reflectance behavior. This was pointed out in [35], where the concept of Apparent BRDFs (ABRDFs) has been introduced to take the above-mentioned effects into account. Bidirectional Texture Functions (BTFs) [9] are a data-driven approach to efficiently capture and store ABRDFs and represent these mesoscopic effects. The results in [16], where training data has been synthesized based on BRDFs, support exactly this claim by showing that using BRDF materials for synthetic training data alone is not sufficient and leads to classification results significantly worse than using real-world images. In contrast, our experiments indicate that using an appropriate representation of the reflectance behavior like the BTF opens the possibility for using solely synthesized training data for classification tasks. We demonstrate that the classification of real-world test data can be boosted significantly by using image-based lighting via environment maps [10] instead of simple directional light sources. To achieve this, we generate synthesized training samples under a vast amount of different lighting conditions simulated by arbitrary HDR environment maps which adequately represent the complexity of real-world materials and lighting.

For this purpose, we have acquired a database containing dense BTF measurements of 84 material samples. The samples can be grouped into 7 categories (i.e. 12 samples per class). Per BTF, all combinations of 151 view and 151 light directions have been measured which results in 22,801 images per sample or a total of $7 \cdot 12 \cdot 22{,}801 > 1.9$M images respectively. The data of our measured database with directional illumination is used as input for generating the synthesized data. By acquiring a height map of each material sample via structured light, we also include the complexity of the geometric structure of the different materials in the process of generating synthetic training images. While in fact an arbitrary number of configurations could be easily included in the synthesized database, we so far used 42 different viewpoints and 30 different illumination conditions per material sample.

In summary, the key contributions of our paper are:

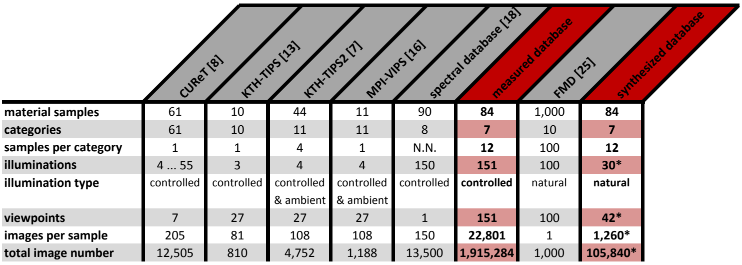
- a technique for decoupling the acquisition of material samples from the environment conditions by generating synthetic training samples
- a publicly available novel BTF database of 7 material categories, each consisting of measurements of 12 different material samples, measured in a darkened lab environment with controlled illumination
- a second, novel database containing data synthesized under natural illumination which is a clear difference to other datasets which only use directional illumination or an additional single ambient illumination
- an evaluation which shows that these synthetic training samples can be used to classify materials in photographs under natural illumination conditions

2 Previous Work

In this section, we briefly review commonly used databases for material recognition and discuss their limitations. Subsequently, we discuss approaches that follow the recent trend of using synthetic training data in various applications.

Databases. Table 1 gives an overview of several different material databases. The CUReT database [8] has been extended in the scope of the KTH-TIPS database [13] in terms of varying the distance of the acquired sample to the camera, i.e. the scale of the considered textures, in addition to changing viewpoint and illumination angle. In both databases, however, only a single material instance is provided per class, and thus the intra-class variation is not represented. Aiming for a generalization to classifying object categories, the KTH-TIPS database was further extended by adding measurements of different samples of the same material category and also considering ambient lighting in the

Table 1. Overview of different databases. Please note that the FMD considers different configurations of viewing and lighting conditions as well as different material samples for each individual image. Our databases are highlighted in red (*: in principle, an arbitrary number of configurations could be considered in the synthesis)



	CUReT [8]	KTH-TIPS [13]	KTH-TIPS2 [7]	MPI-VIPS [16]	spectral database [18]	measured database	FMD [25]	synthesized database
material samples	61	10	44	11	90	**84**	1,000	**84**
categories	61	10	11	11	8	**7**	10	**7**
samples per category	1	1	4	1	N.N.	**12**	100	**12**
illuminations	4 ... 55	3	4	4	150	**151**	100	**30***
illumination type	controlled	controlled	controlled & ambient	controlled & ambient	controlled	**controlled**	natural	**natural**
viewpoints	7	27	27	27	1	**151**	100	**42***
images per sample	205	81	108	108	150	**22,801**	1	**1,260***
total image number	12,505	810	4,752	1,188	13,500	**1,915,284**	1,000	**105,840***

KTH-TIPS2 database [7]. However, taking only four samples per category still limits the representation of the intra-class variance of materials observed in real-world scenarios. More recently, a spectral material database was used in [18]. However, the samples are imaged from only one single viewpoint. A common limitation of all these databases is the rather limited number of measurements, which are furthermore acquired in a lab environment. Hence, the influence of the complexity of real-world environment conditions is not taken into account.

The Flickr Material Database (FMD) [25] is designed to capture the large intra-class variation in appearance of materials in complex real-world scenarios. Images downloaded from Flickr.com show different associated material samples under uncontrolled viewing and illumination conditions and compositions. While manual segmentations are available, these masks are not always accurate, leading to the inclusion of background appearance and problematic artifacts for material classification. Since the manual annotation is time-consuming, the number of images is very small in comparison to the other databases. While standard classification schemes such as [32] reach excellent results on the above-mentioned databases, there is a significant decrease in the performance on the FMD [17]. This is a hint on the fact, that the CUReT and KTH-TIPS databases are not sufficient to represent the complexity of real-world materials.

Synthetic training data. Recombination methods focus on some specific aspects present in real-world examples and recompose them to new examples as done in [11, 21] to enlarge the available training data by recombining shape, appearance and background information for pedestrian detection. In [29], new virtual training images are synthesized via photometric stereo for texture classification. This way, less training images need to be acquired. In contrast, rendering techniques can be used to produce new examples based on an underlying model, e.g. pose estimation was facilitated using synthesized depth maps in [27]. In [28], object detection based on 3D CAD models is investigated using viewpoint-dependent, non-photorealistic renderings of the object contours for learning shape models in 3D, which then can be matched to 2D images showing the corresponding object. Furthermore, an evaluation of the commonly used image descriptors based on a photorealistic virtual world has been carried out in [15]. This virtual scenario represents a well-suited setting for analyzing the effect of illumination and viewpoint changes. The methods in [2] and [3] use a renderer to synthesize shading images based on given depth maps and a spherical harmonic model of illumination for estimating shape, illumination and reflectance from input images. This way, a decoupling of albedo and illumination is reached. The decoupling of measured surface material and environmental lighting has also been addressed in [19], where shape and BRDF of objects have been jointly estimated under known illumination from synthetic data, generated from different combinations of shapes, environment illuminations and BRDFs. In [33], geometric textons are rendered under different view-light configurations for estimating geometric texton labels used in a hybrid model for geometry and reflectance.

Recently, this trend has resulted in the development of the virtual MPI-VIPS database introduced in [16] (see Table 1). This database is based on using BRDFs for representing the light exchange on the surface of an object and does not rely on physical measurements but uses a texture map and material shaders of available rendering packages. Bump maps are used to simulate the local mesostructure of the material surface for improving the shading effects. The selection of shaders, viewpoints and illuminations for rendering the materials are closely oriented on the KTH-TIPS2 database. The texture map does not capture intra-class variance and the approximate rendering models result in a loss concerning the realistic depiction of some materials such as aluminum foil, which appear rather artificial, especially in complex light situations. The reason for this is that mesoscopic effects contributing to the appearance of many materials are not modeled. The investigations in [16] therefore indicate that a training set based on the utilized virtual samples alone performs poorly for material classification and a mixture of real and rendered samples is necessary to get acceptable results. In contrast to these studies, we show that the approach for synthesizing virtual samples matters. Our measured database better covers intra-class variances and includes significantly more viewing and lighting configurations than any of the other databases. These dense measurements are required for the realistic depiction of many materials with their characteristic traits in a virtual scene via BTFs to preserve the mesoscopic effects in the synthesized data.

3 Generation of Synthetic Training Data

In this section, we discuss the details of our database of measured BTF material samples and how it is used to produce synthetic training images of these samples under a range of different viewing and illumination conditions.

3.1 BTF Material Database

Since we intend to create synthetic training images, it is necessary to digitize the material samples in such a way that it becomes possible to reproduce the material appearance under nearly arbitrary viewing and lighting conditions. Though a wide range of material descriptions exists, image-based BTFs have proven to be a representation which is suitable for a wide range of materials as already discussed in Section 1. Since their introduction in [8], the technology has advanced considerably, and today devices for the practical acquisition of BTFs at high angular and spatial resolutions are available (e.g. [24]). In contrast to the small number of representative images acquired for the other databases listed in Table 1, these setups allow to acquire tens of thousands of images. Those images are taken in a lab environment and, hence, not directly applicable for typical real-world scenarios. However, this much larger number of viewing and lighting conditions offers the possibility to render high-quality images of the materials under nearly arbitrary viewing and lighting conditions, where material traits are still accurately preserved. For a recent survey on BTFs, we refer to [12].

Our measured database is formed by 7 semantic classes which are relevant for analyzing indoor scenarios (see Fig. 1). To sample the intra-class variances, each of these 7 material categories of our database contains measurements of 12 different material instances. These instances share some common characteristics of the corresponding category but also cover a large variability. With a total of 84 measured material instances, we provide more than CUReT, KTH-TIPS and KTH-TIPS2 (see Table 1). For each of the materials, we have measured a full BTF with 22,801 HDR images (bidirectional sampling of 151 viewing and 151 lighting directions) of a 5cm × 5cm patch with a spatial resolution of 512 × 512 texels. Thus, our database contains more than 1.9 million images. Additionally, for each sample, a height map has been acquired via structured light. This helps to reduce compression artifacts and allows to render realistic silhouettes. We employed a reference geometry to evaluate the RMS error between the reconstruction and the ground truth geometry which was approx. 25μm. The acquisition of both geometry and BTF of a material sample was achieved fully automatically in approximately 3 hours, and up to 4 samples can be acquired simultaneously. In particular, there is no need for manual annotation which is not feasible for large image collections. Our database is available at http://cg.cs.uni-bonn.de/en/projects/btfdbb/download/ubo2014/.

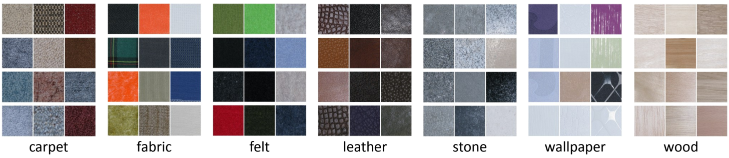


Fig. 1. Representative images for the material samples in the 7 categories

3.2 Synthesizing Novel Training Images

Once the materials have been measured, it is in principle possible to render images showing the materials under nearly arbitrary viewing and lighting con-

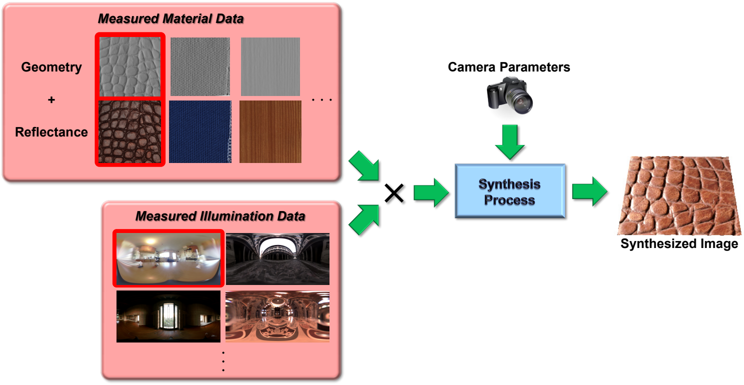


Fig. 2. Synthesis of representative training data: The full Cartesian product of material data (corresponding geometry and reflectance) and environment lighting (environments taken from [1]) can easily be rendered by using a virtual camera with specified extrinsic and intrinsic parameters. The illustrated output image is generated using the material and illumination configuration highlighted in red

ditions. For training a material classifier, we have to decide for which conditions we synthesize the training images, and we need a technique to synthesize a sufficiently large number of images efficiently. Additionally, the material representation used for producing the renderings needs to be capable of accurately depicting the traits in material appearance. In the synthesis process (see Fig. 2), the measured geometry and BTF of a considered material sample are rendered under different illumination conditions simulated by environment maps which is a standard in computer graphics (e.g. [10]). Furthermore, utilizing the measured geometry allows compensating parallax effects. The latter would otherwise be induced by surface regions which significantly protrude from the modeled reference surface and result in a blurring of the surface details. We followed the technique in [23] which is based on the reprojection of the BTF onto the geometry. The result remains a BTF parameterized over the respective (non-planar) reference geometry (and not a Spatially Varying BRDF), as the reflectance functions still remain data-driven ABRDFs. Hence, effects like interreflections, self-shadowing, etc. can still be reproduced. For geometric details not contained in the reference geometry, the major parallaxes are removed by the reprojection and the remaining disparities do not significantly influence the appearance of the synthesized material.

In the rendering process, the exitant radiance $L_r(\mathbf{x}, \omega_o)$ is calculated for each surface point $\mathbf{x}$ via the image-based relighting equation

$$L_r(\mathbf{x}, \omega_o) = \int_{\Omega} \mathrm{BTF}(\mathbf{x}, \omega_i, \omega_o)\, L_i(\omega_i)\, V(\mathbf{x}, \omega_i)\, d\omega_i, \tag{1}$$

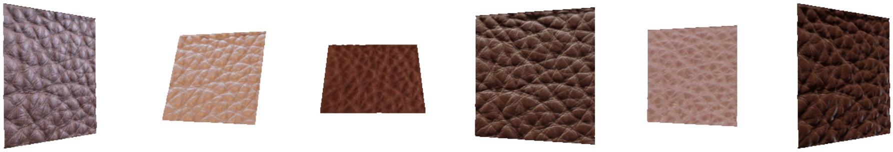


Fig. 3. Examples for synthesized images of the same material sample demonstrating the large variation under different viewing and illumination conditions

where ω_i and ω_o represent the incoming and outgoing light direction. $L_i(\omega_i)$ denotes the radiance distribution in the environment map over the spherical domain Ω. The visibility function $V(\mathbf{x}, \omega_i)$ represents a binary indicator function considering if the environment map is visible from surface point $\mathbf{x}$ in the direction ω_i. To solve the integral, the Mitsuba pathtracer [34] can be used. Due to the enormous number of images we want to synthesize, the use of an efficient rendering technique is mandatory. Therefore, we decided to additionally use an OpenGL-based renderer for generating our database. To simulate the HDR environment in this renderer, we approximated it in a similar way to the work in [4] with 128 directional light sources, distributed representatively over the environment via a relaxation algorithm. In this case, the equation for evaluating the exitant radiance $L_r(\mathbf{x}, \omega_o)$ reduces to

$$L_r(\mathbf{x}, \omega_o) = \sum_{\omega_i \in \mathcal{L}} \mathrm{BTF}(\mathbf{x}, \omega_i, \omega_o)\, L_i(\mathbf{x}, \omega_i)\, V(\mathbf{x}, \omega_i) \qquad (2)$$

where $V(\mathbf{x}, \omega_i)$ represents a shadowing term computed via shadow mapping [22] and $\mathcal{L}$ denotes the set of light source directions, i.e. the ω_i represent the directions to the utilized directional light sources. That way, it becomes possible to render the images with a double resolution full-scene anti-aliasing at a resolution of 1,280 × 960 pixels in about 2s on a GPU, including the computation of the 128 shadow-maps necessary to compute $V(\mathbf{x}, \omega_i)$. Fig. 3 illustrates the considerable variations in material appearance captured in the synthesized data due to changes in the illumination and viewing conditions.

For every combination of material sample and environment map, we then generated training images, depicting a planar material sample under a range of 21 different rotations of the material sample ($\theta \in \{0.0°, 22.5°, 45.0°\} \times \varphi \in \{-67.5°, -45.0°, -22.5°, 0.0°, 22.5°, 45.0°, 67.5°\}$) and in two different distances to also consider the scale-induced variations in appearance of the materials. To further increase the variance captured by our dataset, we also use 6 rotated versions of each of the 5 environment maps available at [1]. As a consequence, we obtain 1,260 images per material sample (see Table 1). Though we only used planar samples for this paper, the BTFs could in principle also be rendered on arbitrary geometry to further increase the space of sampled conditions.

4 Classification Scheme

Fig. 4 illustrates our classification scheme. For capturing different aspects of material appearance, we use densely sampled 3×3 color patches and SIFT features which represent standard descriptors (e.g. [17]). Although the color of a material varies depending on the environmental conditions and the viewpoint, it still contains valuable information as the variance of the color of a certain material sample under natural illumination is typically limited. Furthermore, we extract dense SIFT features which has become a popular choice in scene, object and material recognition [5, 36, 17, 16]. These features capture the local spatial and directional distribution of image gradients and provide robustness to variations in illumination and viewpoint. In our system, these features are extracted on multiple scales ($s \in \{1, 2, 4, 6, 8\}$). Both descriptor types are extracted on a regular grid with a spacing of 5 pixels as in [17].

Once features have been extracted, an appropriate representation for the content of the masked image regions has to be computed for each type of descriptor. For this purpose, we first generate a dictionary of visual words for the individual feature types by k-means clustering of the respective descriptors extracted from the images in the training set. This allows us to represent the single images either by histograms as used in standard bag-of-words (BOW) approaches or by more sophisticated representations such as Fisher vectors [20] or vectors of locally aggregated descriptors (VLADs) [14] which have shown to yield superior performance when compared to standard BOW. Hence, we choose VLADs for describing the content of the masked regions. This means that all the local descriptors $\mathbf{x}_i$ in an image are first assigned to their nearest neighbor $\mathbf{c}_j$ with $j = 1, \ldots, k$ in the corresponding dictionary with k visual words for each feature type. Subsequently, the entries in the VLAD descriptor are formed by accumulating the differences $\mathbf{x}_i - \mathbf{c}_j$ of the local descriptors and their assigned visual words according to

$$\mathbf{v}_j = \sum_{\{\mathbf{x}_i | \mathrm{NN}(\mathbf{x}_i) = \mathbf{c}_j\}} \mathbf{x}_i - \mathbf{c}_j . \tag{3}$$

The final descriptor is built via the concatenation $\mathbf{v} = \left[\mathbf{v}_1^T, \ldots, \mathbf{v}_k^T\right]^T$. However, the dimensionality of this representation is rather high-dimensional $(d \cdot k)$. Here, d represents the dimensionality of the local descriptors (e.g. $d = 128$ for SIFT) and k the number of words in the dictionary. We utilize PCA and take the 250 most relevant components of the PCA space per descriptor type for the training data. The VLAD representations of the test set are projected into this space.

The final classification task can be performed using standard classifiers such as the nearest neighbor classifier, random forests [6] or support vector machines [31]. The latter have already been successfully applied in the domain of material recognition [32, 7, 26]. Since an SVM with RBF kernel outperformed the nearest neighbor classifier or random forests in our experiments, we only report the numbers for the SVM, where the regularization parameter and the kernel parameter are estimated based on the training data using grid search.

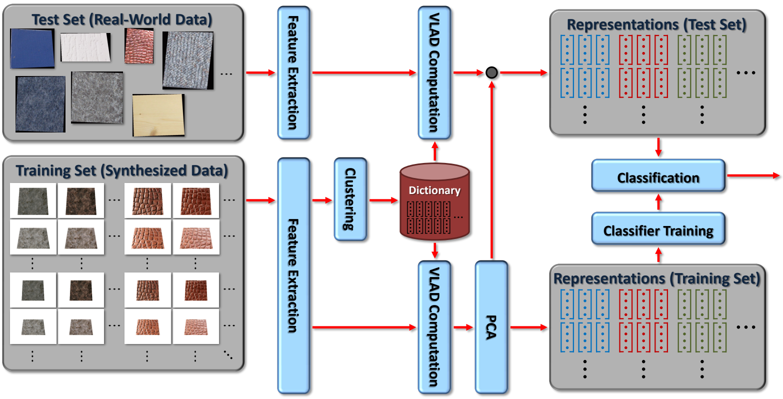


Fig. 4. Classification scheme: Based on the descriptors extracted from the synthetic training data (where the masks for the presence of materials are automatically given) we calculate a dictionary via k-means clustering. This dictionary is used to encode the descriptors per masked region via VLADs. Then, a dimensionality reduction of these VLADs is performed via PCA which is followed by an SVM-based classification

5 Experimental Results

In the scope of our experiments, we focus on whether real-world materials can be classified using synthesized training data. For this purpose, we first validate our classification scheme on standard material databases. In the next step, we perform a detailed evaluation of using our synthesized training data for material classification which is followed by a comparison to using other datasets. After this, we analyze the potential of our synthesized training data for classifying materials in internet photos. For obtaining the VLAD representation of the individual feature types, we use dictionaries with 150 visual words for the color descriptor and 250 visual words for the SIFT descriptor in our experiments.

Validation of classification scheme on commonly used material databases. With accuracies of 99.11% and 99.25% on the CUReT database and the KTH-TIPS database respectively, our system is on par with recent state-of-the-art approaches as listed in [30] which achieve accuracies of around 99%.

Analysis of using synthetic training data. Our main experiments target material classification under everyday illumination conditions. For this reason, we acquired photographs of the samples of the 7 classes considering arbitrarily chosen poses of the camera w.r.t. the material samples for the test set $\mathcal{T}_{te,1}$. Different illumination conditions are taken into account by placing the material samples into different environments: a room with natural illumination, a room with a mix of natural illumination and neon lamps, a room with neon lamps and two darkened room scenarios with a rather directional illumination. In each of the 5 scenarios, each material sample is photographed twice using different viewpoints which results in a test set of 840 images. Based on this test set, we evaluate if our synthesized training data (both pathtraced and OpenGL-based) can be used for training a robust classifier. Additionally, we perform an evaluation of considering natural illumination vs. considering directional illumination as present in the measurement data. This will indicate if and how much can be gained from the training data synthesized under natural illumination. The results are summarized in Table 2.

Comparison of measured vs. pathtraced training data (directional lighting). In a first step, we considered learning the classifier using training data with illumination via point light sources. We randomly selected 50 images per material sample from the measured data resulting in a training set $\mathcal{T}_{tr,m}$ of 4,200 images. Using this training data, we obtain a classification accuracy of 58.92% on $\mathcal{T}_{te,1}$. To support our assumption that virtual images are of a similar quality as their real-world counterparts, we generated a virtual duplicate of the utilized measurement device using the pathtracer implementation in [34]. Using this virtual setup, we produce synthetic training data following Fig. 2 for exactly the same viewing and illumination configurations as present in $\mathcal{T}_{tr,m}$. In this case, we use illumination by point light sources as present in the real device instead of environment maps. The resulting classification accuracy of 60.48% closely matches to the accuracy obtained for the real-world measurement data.

Comparison of measured vs. pathtraced training data (natural lighting). Here, we analyzed the effect of considering more complex illumination as encountered in typical real-world scenarios for the training. We captured all the 84 material samples under two representative room environments and an outside environment in a courtyard, and from two different viewpoints which results in a training set of 504 images. Based on this training set, where we expect the camera settings (viewpoints w.r.t. material samples, white-balancing, ...) to be close to the ones used for the test set, we obtain a classification accuracy of 75.83%. For synthetically simulating this scenario, we captured light probes of the three environments and used them for generating training data under more typical real-world lighting but under the same viewpoints as present in $\mathcal{T}_{tr,m}$. This results in a training set of 12,600 images for which we obtain an accuracy of 68.21%.

Obviously, there is a clear benefit of using representative environments in the generation of the easy-to-produce synthetic training data in comparison to the illumination via point light sources as present in $\mathcal{T}_{tr,m}$. In addition, the characteristic material traits seem to be preserved sufficiently within our synthesized data. However, we recognize a difference in performance between using the training set of 12,600 images synthesized under environment lighting and the use of the 504 photos taken in the respective environments. This might be due to the noise introduced by the pathtracer with only $32spp$ (samples per pixel) which influences the descriptors as well as not perfectly matching the assumptions of far field illumination and the neglection of emitting surfaces. The reason for only taking

32*spp* is that data generation using a pathtracer takes a lot of time, especially if different environmental lighting and different scales are desired. Rendering the 4,200 images for the virtual measurement device (under one single environment map) for instance already takes about two days using 32*spp* with our implementation based on Mitsuba on a Intel Xeon CPU E5 − 2690v2 workstation (32 cores, 3 GHz). We also did not perform a white balancing of the data under the environmental lighting which might influence both descriptor types. Furthermore, the acquisition conditions (view conditions, camera characteristics) of both $\mathcal{T}_{te,1}$ and the 504 real-world training images were similar.

Comparison of measured vs. rasterized training data (natural lighting). As a consequence of the slow rendering via a pathtracer, we used our OpenGL-based synthesis procedure for generating the huge amount of images in our synthesized database. As training set, we consider a random subset of 600 different viewing and illumination conditions from this synthesized data for each of the classes resulting in 4,200 images. In this scenario, our classifier yields a classification accuracy of 72.74% which again significantly outperforms the accuracy of 58.92% obtained when using 4,200 photos acquired during the measurement of the samples in a lab with controlled illumination for the training. It even almost reaches the accuracy of 75.83% from the experiment mentioned before. This might be due to the fact, that we do not encounter the problem of noise induced by the pathtracing approach when using the OpenGL-based synthesis as well as better matching the viewpoint conditions in $\mathcal{T}_{te,1}$ by accounting for multiple scales.

Furthermore, we analyzed the impact of using different numbers of the OpenGL-synthesized images for the training. The accuracy increases with an increasing size of the training data, which is to be expected, as larger training sets cover a larger variance of the utilized viewing and illumination conditions (Table 2).

Comparison of per-class accuracies. There seems to be a trend that in particular the samples of the categories fabric, felt, leather and stone can be categorized more reliably when using the synthesized training data (OpenGL-based) with natural illumination in comparison to measurement data with directional illumination (improvements of around 22% (fabric), 10% (felt), 30% (leather), 35% (stone) and less overfitting to the remaining categories). This agrees with our motivation for this study as we expect the samples of these classes to have more variance in appearance under the different illumination conditions due to their deeper meso-structure and their surface reflectance behavior.

Classifier generalization to unseen material samples in different environments based on synthesized data. For each of the classes, we draw a random subset of 600 images with different viewing and illumination conditions from the complete synthetic training data. We split the material samples of the 7 classes into disjoint training and test sets by using 8 material samples observed under 4 different environments for the training set and the remaining 4 samples rendered under the fifth environment map as the test set. The resulting accuracy of 62.29% indicates the ability of our classifier to generalize to unseen material samples

Table 2. Classification on the manually acquired photos in $\mathcal{T}_{te,1}$ using different training sets (*: pathtraced using Mitsuba renderer [34]; **: OpenGL-based synthesis using 5 environment maps available from [1])

training set	**illumination type**	**type of training data**	**performance on $\mathcal{T}_{te,1}$**
4,200 images from measurement	directional	real-world	58.92%
4,200 synthesized images (pathtraced*) using the same viewing and lighting conditions as present in measurement	directional	synthetic	60.48%
12,600 synthesized images (pathtraced*) using the same viewing conditions as present in the measurement data but under 3 measured environments	natural	synthetic	68.21%
504 photos acquired in 3 measured environments	natural	real-world	75.83%
525 synthesized images (OpenGL-based**)	natural	synthetic	62.74%
1,050 synthesized images (OpenGL-based**)	natural	synthetic	65.71%
2,100 synthesized images (OpenGL-based**)	natural	synthetic	68.69%
4,200 synthesized images (OpenGL-based**)	natural	synthetic	72.74%

and illumination conditions. Using more material samples per category and more environment maps would probably lead to an increasing accuracy.

Using our synthesized database vs. using previous synthesized training data for classifier training. A comparison to other approaches using synthesized data, such as [16], is not directly possible. We focus on different material categories that might be more relevant for analyzing offices, buildings or streets and our synthesized data differs significantly from the data in [16] as we utilize natural lighting conditions which allows material classification outside a controlled lab environment. We show the benefit of using more realistic data for the overlapping material category wood in the supplementary material.

Classifying materials in internet photos. We downloaded for each of our 7 material categories 20 images and performed a manual segmentation on each image. Then, the masked material regions form $\mathcal{T}_{te,20}$. Taking a subset of 15 images per class from $\mathcal{T}_{te,20}$ gives another test set $\mathcal{T}_{te,15}$. Using our aforementioned training data of 4,200 images synthesized using OpenGL and under consideration of environmental illumination gives accuracies of 65.71% ($\mathcal{T}_{te,15}$) and 62.86% ($\mathcal{T}_{te,20}$). In comparison, using $\mathcal{T}_{tr,m}$ for the training results in an accuracy of only 56.19% for $\mathcal{T}_{te,15}$ and 56.43% for $\mathcal{T}_{te,20}$.

In addition, training the classifier on 5 of the images per class not included in $\mathcal{T}_{te,15}$ gives an accuracy of 41.90% on $\mathcal{T}_{te,15}$. The influence of adding synthesized data to this training set on the accuracy obtained for $\mathcal{T}_{te,15}$ as well as a summary of the other results in this paragraph are shown in Table 3. Taking more training data with a larger variance of the utilized illumination conditions and the utilized viewpoints leads to an increasing performance. This clearly demonstrates the power of using synthesized materials for practical applications.

Except for the category leather, we also used the samples present in the CUReT database to represent the categories. For each category, we selected

Table 3. Classification of internet images ($\mathcal{T}_{te,15}$ and $\mathcal{T}_{te,20}$) using different training sets (*: OpenGL based synthesis using 5 environment maps available from [1]; †: category leather is not covered in the CUReT database)

training set	**illumination type**	**type of training data**	$\mathcal{T}_{te,15}$ 15 internet images	$\mathcal{T}_{te,20}$ 20 internet images
CUReT images †	directional	real-world	41.11%	36.67%
4,200 images from measurement	directional	real-world	56.19%	56.43%
4,200 synthesized images*	natural	synthetic	65.71%	62.86%
internet images	natural	real-world	41.90%	–
internet images+ 4,200 synthesized images*	natural	mixed	66.67%	–
internet images+ 16,800 synthesized images*	natural	mixed	72.38%	–

92 images equally distributed over the material samples contributing to the classes (carpet: samples 18,19; fabric: samples 2,3,7,22,29,42,44,46; felt: sample 1; stone: samples 10,11,17,30,33,34,36,37,41,49,50; wallpaper: samples 12,31,38; wood: samples 54,56). In this experiment, we obtained accuracies of 41.11% ($\mathcal{T}_{te,15}$) and 36.67% ($\mathcal{T}_{te,20}$) hinting on a bad generalization of the CUReT database to natural illumination, varying viewing conditions and intra-class variances. Furthermore, the image quality is rather low for the CUReT database.

6 Conclusion

In this paper, we have presented an approach for creating synthetic training samples for material classification. This way, it is possible to decouple the acquisition of the material samples from the acquisition of the illumination conditions under which the material is observed. In addition, using synthesized data overcomes the need for time-consuming manual acquisition, annotation and segmentation of images. To evaluate our approach, we acquired a database of BTFs, containing 7 classes with 12 samples each, from which the training data is generated. Our evaluation demonstrates that our approach represents a significant step towards classifying materials in everyday environments and clearly outperforms the alternative of taking images from the measurement of the material samples under controlled illumination conditions as training data. This makes our approach valuable for many applications in the area of texture classification and automatic segmentation of images. We intend to extend the database by additional material classes. In addition, the number of viewing and lighting conditions could also be increased for the synthesized database.

Acknowledgements The research leading to these results was funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 323567 (Harvest4D), 2013-2016 and by the DFG Emmy Noether program (GA 1927/1-1). The authors also thank Welf Warnecke for helping with the measurements.

References

1. http://www.pauldebevec.com/ (November 2013)
2. Barron, J.T., Malik, J.: Shape, albedo, and illumination from a single image of an unknown object. In: CVPR. pp. 334–341 (2012)
3. Barron, J., Malik, J.: Intrinsic scene properties from a single rgb-d image. In: CVPR. pp. 17–24 (2013)
4. Ben-Artzi, A., Ramamoorthi, R., Agrawala, M.: Efficient shadows for sampled environment maps. J. Graphics Tools 11(1), 13–36 (2006)
5. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via pLSA. In: ECCV. pp. 517–530 (2006)
6. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (2001)
7. Caputo, B., Hayman, E., Mallikarjuna, P.: Class-specific material categorisation. In: ICCV. vol. 2, pp. 1597–1604 (2005)
8. Dana, K.J., van Ginneken, B., Nayar, S.K., Koenderink, J.J.: Reflectance and texture of real world surfaces. Tech. rep. (1996)
9. Dana, K.J., Nayar, S.K., Ginneken, B.V., Koenderink, J.J.: Reflectance and texture of real-world surfaces. In: CVPR. pp. 151–157 (1997)
10. Debevec, P.: Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In: SIGGRAPH. pp. 189–198 (1998)
11. Enzweiler, M., Gavrila, D.M.: A mixed generative-discriminative framework for pedestrian classification. In: CVPR. pp. 1–8 (2008)
12. Filip, J., Haindl, M.: Bidirectional texture function modeling: A state of the art survey. IEEE Trans. Pattern Anal. Mach. Intell. 31, 1921–1940 (2009)
13. Hayman, E., Caputo, B., Fritz, M., Eklundh, J.O.: On the significance of real-world conditions for material classification. In: ECCV. pp. 253–266 (2004)
14. Jegou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR. pp. 3304–3311 (2010)
15. Kaneva, B., Torralba, A., Freeman, W.T.: Evaluation of image features using a photorealistic virtual world. In: ICCV. pp. 2282–2289 (2011)
16. Li, W., Fritz, M.: Recognizing materials from virtual examples. In: ECCV. vol. 4, pp. 345–358 (2012)
17. Liu, C., Sharan, L., Adelson, E.H., Rosenholtz, R.: Exploring features in a bayesian framework for material recognition. In: CVPR. pp. 239–246 (2010)
18. Liu, C., Yang, G., Gu, J.: Learning discriminative illumination and filters for raw material classification with optimal projections of bidirectional texture functions. In: CVPR. pp. 1430–1437 (2013)
19. Oxholm, G., Nishino, K.: Shape and reflectance from natural illumination. In: ECCV. pp. 528–541 (2012)
20. Perronnin, F., Dance, C.R.: Fisher kernels on visual vocabularies for image categorization. In: CVPR (2007)
21. Pishchulin, L., Jain, A., Wojek, C., Andriluka, M., Thormählen, T., Schiele, B.: Learning people detection models from few training samples. In: CVPR. pp. 1473–1480 (2011)
22. Reeves, W.T., Salesin, D.H., Cook, R.L.: Rendering antialiased shadows with depth maps. SIGGRAPH Comput. Graph. 21(4), 283–291 (1987)
23. Ruiters, R., Schwartz, C., Klein, R.: Example-based interpolation and synthesis of bidirectional texture functions. Computer Graphics Forum (Proceedings of the Eurographics 2013) 32(2), 361–370 (2013)

24. Schwartz, C., Weinmann, M., Ruiters, R., Klein, R.: Integrated high-quality acquisition of geometry and appearance for cultural heritage. In: The 12th International Symposium on Virtual Reality, Archeology and Cultural Heritage VAST 2011. pp. 25–32 (2011)
25. Sharan, L., Rosenholtz, R., Adelson, E.H.: Material perception: What can you see in a brief glance? Journal of Vision 8 (2009)
26. Sharan, L., Liu, C., Rosenholtz, R., Adelson, E.H.: Recognizing materials using perceptually inspired features. IJCV 103(3), 348–371 (2013)
27. Shotton, J., Fitzgibbon, A.W., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR. pp. 1297–1304 (2011)
28. Stark, M., Goesele, M., Schiele, B.: Back to the future: Learning shape models from 3d cad data. In: BMVC. pp. 106.1–106.11 (2010)
29. Targhi, A.T., Geusebroek, J.M., Zisserman, A.: Texture classification with minimal training images. In: International Conference on Pattern Recognition. pp. 1–4 (2008)
30. Timofte, R., Van Gool, L.: A training-free classification framework for textures, writers, and materials. In: BMVC. pp. 1–12 (2012)
31. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA (1995)
32. Varma, M., Zisserman, A.: A statistical approach to material classification using image patch exemplars. IEEE Trans. Pattern Anal. Mach. Intell. 31(11), 2032–2047 (2009)
33. Wang, J., Dana, K.J.: Hybrid textons: Modeling surfaces with reflectance and geometry. In: CVPR. vol. 1, pp. 372–378 (2004)
34. Wenzel, J.: Mitsuba renderer (2010), http://www.mitsuba-renderer.org
35. Wong, T.T., Heng, P.A., Or, S.H., Ng, W.Y.: Image-based rendering with controllable illumination. In: EGWR. pp. 13–22 (1997)
36. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. IJCV 73(2), 213–238 (2007)