MaterialNet: Multi-scale Texture Hierarchy and Multi-view Surface Reflectance for Material Type Recognition

Dongjin Lee[1]
dj923@khu.ac.kr
Hyun-Cheol Kim[2]
kimhc@etri.re.kr
Jeongil Seo[2]
seoji@etri.re.kr
Seungkyu Lee[1]
seungkyu@khu.ac.kr

[1] Department of Computer Science and Engineering
Kyunghee University

[2] Electronics and Telecommunications Research Institute

Abstract

Material is distinguishing characteristic of real world objects. Recognizing unique texture of certain material type enables improved object detection or semantic segmentation. Incorporating acquired material properties such as surface reflectance of real world objects makes more realistic and richer 3D models in computer graphics. A robot arm essentially requires to recognize the stiffness or roughness of target object for precise and undamaged interaction. Despite the necessities, recognizing material type and its properties from color image is a challenging task. In this work, we propose (1) multi-scale texture hierarchy extraction network (MSTH-Net) encoding view-independent comprehensive multi-scale textures and their hierarchy and (2) multi-view surface reflectance extraction network (MVSR-Net) encoding view-specific features revealing surface reflectance of a material type. Finally, MaterialNet is proposed combining MSTH-Net and MVSR-Net for material type recognition from multi-view color images. Extensive experimental evaluations on six public benchmark datasets show promising performance of proposed method and potential for practical applications.

1 Introduction

Real world objects are composed of multiple distinguishing materials that are essential characteristics in determining object category. Altering material composition of an object changes its sub-category such as wooden vs iron table, fabric vs leather couch. Recognizing material types enables improved object detection and semantic segmentation. In computer graphics, material properties such as surface reflectance enable realistic rendering of a 3D model. Industrial robot arm interacting with many different kinds of objects composed of diverse material types has to react correspondingly for precise grabbing without any damage on

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

them. If the material type of an arbitrary object can be recognized from input images, it simply can be added to corresponding 3D model enabling material-aware realistic rendering or robot arm interaction. Certain material type is well categorized by its surface characteristics such as reflectance, stiffness, friction, roughness, and texture. While haptic properties (stiffness, friction, and roughness) are difficult to be estimated explicitly from visual data, texture of a certain material type could be easily observed from color image and has been widely adopted for semantic segmentation and object detection. Amorphous object classes such as road, grass, and sky are characterized and described well by their local textures. Shape-constrained object classes such as person, cat, and chair also benefit from their unique local textures and corresponding material types in their characterization. Convolutional neural network trained with ImageNet favours texture rather than shape [7] in the description of an object. It is obvious that object description could be optimized when texture information is properly incorporated with shape [12]. Surface reflectance is another distinguishing property of a material type that can be estimated from multiple viewpoint observations (e.g.[19, 22]). Intensity variation observed along the viewpoint changes discloses the surface reflectance of target material type. Texture uniqueness in objects mainly comes from their distinguishing material types. Texture recognition plays an important role in figuring out corresponding material type. We claim that texture features of an object robust to environmental changes, their hierarchy along multiple scales, and surface reflectance obtained from multi-view images are able to characterize material types uniquely and comprehensively. Surface reflectance provides additional discrimination of material types that can be estimated from illumination variation of multiple viewpoint

Based on the observations, we conclude that material types are able to be characterized uniquely and comprehensively by (1) view-independent texture features of an object robust to environmental changes and (2) view-specific visual features observed from multi-view color images revealing surface reflectance. In order to encode both view-independent and view-specific features for material type recognition, we propose multi-scale texture hierarchy extraction network (MSTH-Net) and multi-view surface reflectance extraction network (MVSR-Net). MSTH-Net encodes comprehensive multi-scale textures and their correlations from low-level features of small receptive field to high-level features of large receptive field of backbone convolutional neural networks. MSTH-Net is expected to extract view-independent stationary visual features robust to illumination changes and other variations from multi-view images. On the other hand, MVSR-Net collects visual features of varying illuminations along the viewpoint changes. MVSR-Net is expected to capture view-specific features from multi-view images, which extract the characteristics of object surface reflectance. Finally, MaterialNet is proposed for material type recognition from multi-view images. MaterialNet is composed of a pair of MSTH-Net and MVSR-Net. MaterialNet combines texture and reflectance features for the description of visual characters of material.

2 Related Work

Prior efforts on material type recognition are categorized according to the material properties primarily used for the classification task: texture-based approach, reflectance-based approach. Recently, texture recognition has been conducted employing convolutional neural networks. Cimpoi et al.[6] propose Fisher Vector CNN(FV-CNN) for texture recognition with orderless pooling, because in general texture instances from same class do not maintain common shape as if an object class does. Song et al.[18] propose locally-transferred Fisher vector (LFV) by combining the output of FV-CNN with a learnable locally connected layer to keep the benefits of both FV encoding and neural network. Andrearczyk et al.[1] propose Texture-CNN (T-CNN) using a simple Global Average Pooling (GAP) layer that averages the features of each channel obtaining spatially orderless description while reducing the memory usage and computation complexity. There has been a bunch of work that analyze various visual properties of material such as visual texture for material type recognition. Hu et al.[10] and Sharan et al.[17] experimentally analyze the properties of materials and suggest new set of features such as micro-SIFT[17] and variance of gradient orientation[10]. Bell et al.[2] demonstrate patch-wise material classification and semantic segmentation using large-scale annotated training database named Materials in Context (MINC). They show that surrounding context information such as object and scene is crucial for real-world material classification. Schwartz et al.[15] improve material semantic segmentation performance by explicitly integrating local appearance of materials and global context such as objects and scenes. To integrate the global context information into material semantic segmentation model, they propose to combine a pre-trained network for object recognition task and trainable material recognition network. Zhang et al.[28] propose DeepTEN, an end-to-end framework, using residual dictionary learning to learn inherent visual properties of texture. DEP[23] integrates orderless texture features and local spatial features by using DeepTEN as a texture encoding backbone claiming that local spatial information is important cue for the recognition of real world material types. Zhai et al.[25] propose MAPNet progressively learning visual texture properties by mutually reinforcing manner. They also propose DSRNet[26] utilizing spatial dependencies among texture primitives to capture the structural information of textures. The most recent texture recognition studies are based on fractal analysis. Zhile Chen et al. propose CLASSNet[4] to model CNN feature maps across multiple layers to take advantage of statistical self-similarity (SSS), one of the main properties of textures. Xu et al. propese FENet[21] characterizing spatial layout via a local-global hierarchical fractal analysis. CLASSNet and FENet integrated with ResNet backbone show state-of-the-art performance in texture recognition.

Recently, some material recognition studies are introduced extracting surface reflectance which is one of the core visual properties of materials. Zhang et al.[27] propose reflectance hashing that captures reflectance disk of material surface using a unique optic camera for material type recognition. Georgoulis et al.[8] perform material classification by taking the advantages of the 3D shape of an object and the reflectance map. lee et al.[11] propose two-stream deep neural networks using both color and Infrared (IR) surface reflectance estimation using Time-of-Flight depth camera. There are several material classification methods [9, 19, 20] estimating reflectance using 4D light-field camera. Light field camera takes multi-view images of narrow angle changes using single camera equipped with lenslet array extracting partial reflectance information. Purri et al. [14] propose reflectance residual encoding, which captures both multi-angle and multi-spectral information to improve material semantic segmentation performance on satellite image. These methods work under limited experimental conditions or employ customized or special type of cameras. On the other hands, Xue et al.[22] use differential angular images encoding angular gradient reflectance, and multi-view images to improve the performance of ground terrain material classification. Additionally, they propose TEAN[24] to improve classification performance by adding a reflectance branch that takes a differential angular image as input to their DEP structure[23].

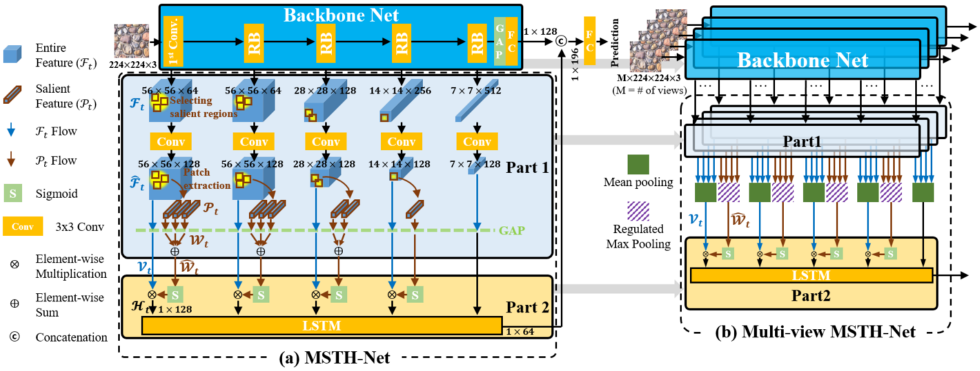


Figure 1: Multi-scale texture hierarchy network (MSTH-Net) and its multi-view extension

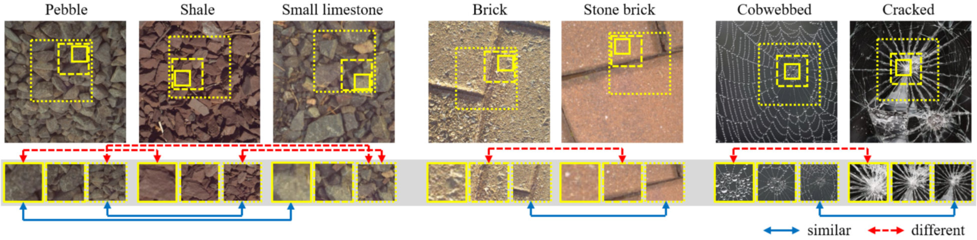


Figure 2: Sample multi-scale texture hierarchy of similar material groups from GTOS[22] and DTD[5]. The yellow boxes in an image are salient regions of different scales.

3 Methods

Proposed MaterialNet is composed of multi-scale texture hierarchy network (MSTH-Net) and multi-view surface reflectance network (MVSR-Net). Multi-scale texture hierarchy network (MSTH-Net) attached to existing CNN based backbone encoder is designed to collect texture features from local regions of multiple scales as illustrated in Figure 1 (a). MSTH-Net takes both entire and salient features from each layer in multi-scale texture extractor (part 1 in Figure 1) that are fed to texture attention and texture hierarchy builder (part 2 in Figure 1). Multi-scale texture hierarchy features are extracted and concatenated with the features extracted from the backbone encoder. Backbone encoder in Figure 1 (a) is, but not limited to, ResNet18 that consists of sequentially connected residual blocks (RBs).

3.1 Multi-Scale Texture Hierarchy Network

Multi-scale texture extractor (part 1 in Figure 1 (a)) collects low to high level textures of salient regions according to the size of the receptive field. Each texture delivers different aspects of visual characteristics observed from difference scales of input image. Figure 2 shows several examples of multi-scale textures from different datasets used in our experiments. Lower scale texture (of smaller salient patch) captures intrinsic surface texture of a material type regardless of corresponding object shape. Mid-scale texture shows the shape of a piece or an object of the material type. Higher scale texture (of bigger salient patch) shows the texture of multiple pieces of the material type. In general, higher scale texture reveals contextual aspects such as the shape of a pile of multiple pieces or unique arrangement of multiple small objects of the material type.

Let T is total number of residual blocks (RBs) plus 1(1st Conv layer) of ResNet18 and $Z_t(t=1,...,T)$ is the number of channels at each RB. First, entire-features $\mathcal{F}_t(t=1,...T) \in \mathbb{R}^{H_t \times W_t \times Z_t}$ are extracted from each RB. And then, we select N_t number of the most activated regions as salient regions within the window of size $H' \times W'$. (Our implementation settings are $N_1=3, N_2=3, N_3=2, N_4=1, N_5=0, H'=W'=10$). In the entire-feature of the last RB $\mathcal{F}_T$, we do not select salient region because spatial size of $\mathcal{F}_T$ is smaller than the window size. The $\mathcal{F}_t$ from each RB are fed to a convolutional layer, and the output feature map is $\hat{\mathcal{F}}_t \in \mathbb{R}^{H_t \times W_t \times C}$. (ResNet18: C=128, ResNet50: C=192). Then we extract salient-feature patches $\mathcal{P}_t = [P_t^1;...;P_t^{N_t}] \in \mathbb{R}^{N_t \times H' \times W' \times C}$ from $\hat{\mathcal{F}}_t$ based on previously found salient regions. From the residual block of lower to higher level, receptive field of the patch increases. $\hat{\mathcal{F}}_t$ contains entire textures of input image captured with different scales of local regions. On the other hands, $\mathcal{P}_t$ contains only selected local salient textures out of $\hat{\mathcal{F}}_t$. Obtained $\hat{\mathcal{F}}_t$ and $\mathcal{P}_t$ are fed to global average pooling (GAP) layer. Output of GAP layers of $\hat{\mathcal{F}}_t$ and $\mathcal{P}_t$ are single entire vector $\mathcal{V}_t \in \mathbb{R}^C$ and N_t number of salient vectors $\mathcal{W}_t = [W_t^1;...;W_t^{N_t}] \in \mathbb{R}^{N_t \times C}$, respectively. Finally with the salient vectors, all W_t are element-wise summed to each other $(W_t^1 + ... + W_t^{N_t})$ making unified salient vector $\hat{\mathcal{W}}_t \in \mathbb{R}^C$. If N_t is 1, $\hat{\mathcal{W}}_t$ equals to $\mathcal{W}_t$. Details of MSTH-Net shown in Figure 1 are based on input image size of 224×224.

Texture attention and texture hierarchy builder (part 2 in Figure 1 (a)) get entire vector $\mathcal{V}_t$ and unified salient vector $\hat{\mathcal{W}}_t$ from multi-scale texture extractor (part1 in Figure 1 (a)). Then the $\hat{\mathcal{W}}_t$ fed to the sigmoid function is multiplied to $\mathcal{V}_t$ providing attention of strongly activated features in salient regions to entire vector. This feature-wise attention acts as salient texture boosting in the following texture hierarchy builder. While delivering global information extracted from entire image, added texture attention from the salient features enriches the texture representation in the feature space. Output of texture attention $\mathcal{H}_t \in \mathbb{R}^C, (t=1,...,T)$ is sequentially fed to Long Short-Term Memory (LSTM) to build multi-scale texture hierarchy. If $N_t = 0$, $\mathcal{H}_t = \mathcal{V}_t$.

In our MaterialNet, MSTH-Net is combined with MVSR-Net explained in the next section that accepts M multi-view input images. For the multi-view environment, multi-view MSTH-Net(Figure 1 (b)) is constructed by collecting as many texture extractors (part 1) as the number of views MaterialNet accepts. Multiple texture extractors are followed by pooling layers, single texture attention and hierarchy builder, thereby view-independent common multi-scale texture hierarchy features from multi-view input images are extracted. $[\mathcal{V}_{t,1},...,\mathcal{V}_{t,M}]$ and $[\hat{\mathcal{W}}_{t,1},...,\hat{\mathcal{W}}_{t,M}]$ obtained from M views are fed to the channel-wise mean and regulated max pooling to extract view-independent features.

$$\mathcal{V}_t = mean(\mathcal{V}_{t,1},...,\mathcal{V}_{t,M}) \tag{1}$$

$$\mathcal{W}_t = \frac{max(\hat{\mathcal{W}}_{t,1},...,\hat{\mathcal{W}}_{t,M})}{max(\hat{\mathcal{W}}_{t,1},...,\hat{\mathcal{W}}_{t,M}) - mean(\hat{\mathcal{W}}_{t,1},...,\hat{\mathcal{W}}_{t,M})}, \tag{2}$$

where $\mathcal{V}_t, \mathcal{W}_t \in \mathbb{R}^C$ are respective pooling result. We use mean pooling for entire-features and regulated max pooling for salient-features. Since entire-features describe global characteristics of an image and multi-view images give similar features of sharing target scene, entire-features extracted from multi-view images are highly redundant. Therefore, mean pooling is applied to entire-features. On the other hand, salient-features represent different local regions. In order to secure outstanding local features, regulated max pooling is adopted.

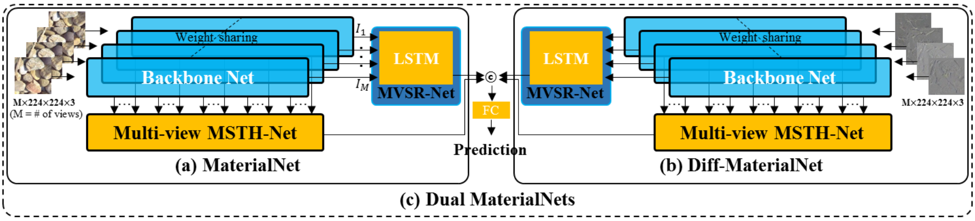


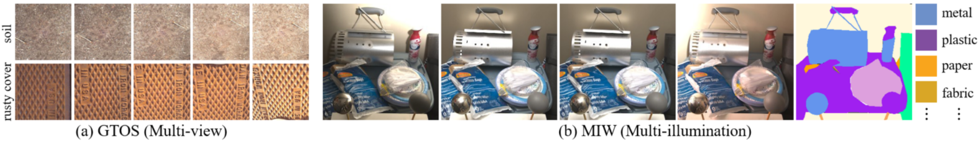
Figure 3: Dual MaterialNets with MSTH-Net and MVSR-Net: (a) MaterialNet using color multi-view images as inputs (b) Diff-MaterialNet using differential angular images as inputs (c) Dual MaterialNets

3.2 Multi-View Surface Reflectance Network

Multi-view surface reflectance network (MVSR-Net) works on multi-view images fed to M number of weight sharing backbone encoders. MVSR-Net adds another LSTM to the outputs of the backbone encoders to predict single reflectance type out of M multi-view images as illustrated in Figure 3 (a). Each backbone encoder gives one-dimensional vector building $\mathcal{I} = [I_1;\ldots;I_M] \in \mathbb{R}^{M\times D}$ where D is channel size. We expect that each encoder finds view-specific features from each viewpoint image and LSTM encodes feature variation over viewpoints. The aspect of illumination changes over viewpoints reveals distinguishing reflectance characteristic of surface material. Since the feature vector $\mathcal{I}$ of multiple views are sequentially fed to the LSTM, the order of input images requires certain criterion for the extraction of surface reflectance. In order to observe physically meaningful illumination variation of reflectance in the sequence of input images, we constrain that input images are sorted in order of camera locations so that observed illumination increases or decreases monotonically. In real situation, however, camera location of each image is missing. Instead, images are sorted by increasing order of brightness. Note that multiple images could be taken within narrow camera view-angle changes observing only partial illumination variation. Even worse, the partial observations could come from only side views with respect to material surface or could be uneven in the angle step size.

3.3 MaterialNets

Single MaterialNet(Figure 3 (a)) encodes both texture hierarchy and surface reflectance from multi-view images. MaterialNet is composed of weight sharing backbone encoders, multi-view MSTH-Nets, and single MVSR-Net. The output of multi-view MSTH-Nets and MVSR-Nets are concatenated and fed to final fully connect layers for material type classification. As shown in Figure 3 (c), we build dual MaterialNets complementarily extracting 2D color and 3D relief based features. MaterialNet (Figure 3 (a)) accepts multi-view color images. Diff-MaterialNet (Figure 3 (b)) accepts difference images (differential angular images [22]) of every two consecutive color images aligned by affine transformation before subtraction [22]. Color difference images observe the gradients of reflectance and 3D relief textures[22]. In Diff-MaterialNet, MVSR-Nets encode the gradients characterizing material reflectance and MSTH-Nets encodes 3D relief textures especially in the mid or high scales of texture hierarchy.



(a) GTOS (Multi-view)

(b) MIW (Multi-illumination)

Figure 4: Multi-view/illumination datasets: (a) GTOS [22] sampled from 19 view directions with instance label and sorted increasing order of brightness (b) MIW [13] sampled from 25 illuminations with per-pixel label

4 Experimental Evaluation

Experimental evaluation is conducted in two ways. First, MSTH-Net is evaluated on single-view texture or material datasets and compared with state of the art prior methods. We use six benchmark datasets for the evaluation of our Multi-Scale Texture Hierarchy network (MSTH-Net). (a) Describable Texture Dataset (DTD)[6] contains 47 categories of wild textures with 120 images per category. (b) KTH-TIPS-2b[3] is surface-level dataset composed of 11 material categories with total 4752 images. (c) Flickr Material Dataset (FMD)[16] is object-level dataset composed of 10 material categories with 100 images for each category. (d) Materials in Context 2500 (MINC-2500)[2] is scene-level dataset of 23 material categories each of which contains 2500 images. (e) Ground Terrain in Outdoor Scenes (GTOS)[22] is a dataset of outdoor ground materials with 40 categories observed from multiple viewpoints. When GTOS dataset is used for single-view input test, one viewpoint image is treated as one instance. (f) GTOS-Mobile[28] is a dataset collected from GTOS via mobile phone, which consists of 31 material categories. Training data of GTOS-mobile consist of small, medium, and large scales, but we only use small scale training data. In our evaluation on DTD, MINC-2500 and GTOS, we use provided splits of each dataset. KTH-TIPS-2b and FMD are randomly divided into 10 splits of training and test images with recommended split size, and the mean accuracy across splits are reported [4].

Secondly, MaterialNets are evaluated on existing two material datasets. (a) Multi-view GTOS dataset has 19 viewpoint images per sample. Proposed method is compared with prior material type recognition methods under multi-view conditions [22, 24] on GTOS dataset. (b) Multi-Illumination Images in the Wild (MIW) [13] consists of multiple illumination images originally designed for vision tasks such as single-image illumination estimation, image relighting, and mixed illuminant white balance. Multi-illumination in MIW means varying location of light source. We ignore Fresnel effect and assume stationaly surface reflectance over incidence angle. Under such conditions, multi-illumination properly simulates multi-view observations. ResNet18 and ResNet50 pre-trained on ImageNet are used as backbone. SGD optimizer with momentum of 0.9 is used and batch size is set to 64 for GTOS, GTOS-mobile and MINC-2500, and 32 for others. Learning rate with cosine decay is initialized to 0.01. Training is finished after 50 epochs. All training and test images are resized to 256 × 256 and then cropped to 224 × 224. Horizontal flipping with probability 0.5 is applied to input images for data augmentation.

4.1 Texture/Material Recognition with MSTH-Net

We compare proposed MSTH-Net with state-of-the-art texture recognition approaches including DeepTEN[28], MAPNet[25], DSRNet[26], CLASSNet[4], and FENet[21] with single input image on six benchmark datasets: DTD[6], KTH-TIPS-2b[3], FMD[16], MINC-2500[2], GTOS[22] and GTOS-mobile[23] as summarized in Table 1. In Table 1, pro-

Method	Backbone	Texture	Material (Single-color)				
		DTD[5]	KTH[3]	FMD[16]	MINC[2]	GTOS[22]	GTOS-mobile[23]
MAPNet[25]	VGGVD	74.10±0.6	82.70±1.5	82.90±0.9	NA	80.80±2.5	82.00±1.6
DSRNet[26]		74.90±0.7	83.50±1.5	84.00±0.8	NA	81.80±2.2	82.94±1.6
DeepTEN[28]	ResNet18	NA	NA	NA	NA	NA	76.12±x.x
DEPNet[23]		NA	NA	NA	NA	NA	82.18±x.x
MAPNet[25]		69.50±0.8	80.90±1.8	80.80±1.0	NA	80.30±2.6	82.98±1.6
DSRNet[26]		**71.20**±0.7	81.80±1.6	81.30±0.8	NA	81.00±2.1	83.65±1.5
CLASSNet[4]		**71.50**±0.4	85.40±1.1	**82.50**±0.7	**80.50**±0.6	**84.30**±2.2	**85.25**±1.3
FENet[21]		69.59±0.1	**86.62**±0.1	82.26±0.3	**80.57**±0.1	83.10±0.2	**85.10**±0.4
MSTH-Net		**69.33**±0.9	**86.69**±1.4	**83.17**±1.5	**79.10**±0.5	**84.95**±2.2	**85.10**±0.3
DeepTEN[28]	ResNet50	69.60±x.x	82.00±3.3	80.20±0.9	81.30±x.x	84.50±2.9	NA
DEPNet[23]		73.20±x.x	NA	NA	82.00±x.x	NA	NA
MAPNet[25]		**76.10**±0.6	84.50±1.3	85.20±0.7	NA	84.70±2.2	86.64±1.5
DSRNet[26]		**77.60**±0.6	85.90±1.3	86.00±0.8	NA	85.30±2.0	**87.03**±1.5
CLASSNet[4]		74.00±0.5	87.70±1.3	**86.20**±0.9	**84.00**±0.6	85.60±2.2	85.69±1.4
FENet[21]		74.20±0.1	**88.24**±0.2	**86.74**±0.2	**83.98**±0.1	**85.71**±0.1	85.20±0.4
MSTH-Net		**71.45**±0.6	**87.72**±1.0	**85.65**±1.4	**81.47**±0.6	**85.73**±2.6	**87.45**±0.8

Table 1: **Single-color Material Recognition (MSTH-Net)**: We mark the **best** performance in blue and the **second best** performance in red. Accuracy in "Mean±SD%" is reported. Standard deviation (±SD) marked x.x is not available from prior works.

Input	Method	GTOS[22]	Input	Method	GTOS[22]
Multi-Color	CNN[24]	82.50±2.8	Multi-Color + Diff.	DAIN[22, 24]	86.20±2.5
	DEP[23, 24]	85.80±1.9		TEAN[24]	87.60±2.0
	MVSR-Net (9 views)	**85.54**±2.7		**MVSR-Net (9 views)**	**86.65**±2.3
	MaterialNet (4 views)	**86.20**±2.5		**Dual MaterialNets (4 views)**	**87.84**±2.1
	MaterialNet (9 views)	**86.71**±2.1		**Dual MaterialNets (9 views)**	**88.41**±2.1

Table 2: **Multi-color Material Recognition (MSVR-Net and MaterialNet)** on GTOS dataset[22]. Backbone is ResNet18.

posed MSTH-Net with ResNet18 achieves best material recognition accuracy on KTH-TIPS-2b, FMD and GTOS datasets and second best on GTOS-mobile dataset. MSTH-Net with ResNet50 achieves best material recognition accuracy on GTOS and GTOS-mobile datasets and second best on KTH-TIPS-2b.

Varying performance of MSTH-Net over datasets comes from varying image style and characteristics. Since GTOS and GTOS-mobile images show clean and clear lower, mid, and higher scale textures, performance improvement of our proposed MSTH-Net is outstanding. FMD is an object-level dataset and both local surface and global shape of an object are well observed in an instance. In other words, lower to higher scale textures are distinct enough to achieve gain by applying multi-scale texture hierarchy. Therefore, MSTH-Net with ResNet18 shows best accuracy. Although KTH-TIPS-2b dataset consists of images taken very closely, MSTH-Net shows outstanding performance because the images show unique local textures. MINC-2500 is a scene-level material dataset. Context information such as other surrounding objects or background scene help material recognition. MSTH-Net, however, does not show significant performance improvement on it, since lower scale image does not reveal surface texture of a material type and higher scale texture contains too much surrounding context information and our feature hierarchy does not work effectively. DTD is a texture dataset that does not share a similar global shape within a class and images of the same class show limited correlations.

4.2 Material Recognition with MaterialNets

We compare MaterialNets with multiview CNN[24], multiview DEP[23, 24] using multi-view color images and multiview DAIN[22], multiview TEAN[24] using both multi-view color and multi-view differential angular images. DEP employs DeepTEN[28] for texture encoding using residual dictionary learning to capture both orderless texture and local spatial features. TEAN is dual network that takes differential angular images by adding a reflectance branch to DEP. DAIN[22] is two-stream convolutional neural networks that use single color and single differential angular images as inputs and combine feature maps. In multiview CNN, DEP, TEAN and DAIN, images are fed to weight sharing network(CNN or DEP or TEAN or DAIN) and material type is inferred by voting in softmax step or pooling features across viewpoints in the network.

Experimental results on GTOS[22] are summarized in Table 2. Multiview CNN, DEP, TEAN, and DAIN accept respective optimal number of viewpoints randomly chosen out of 19 views of GTOS. Since our networks is able to extract partial reflectance by sequentially encoding view-specific features from multiple-views, it outperforms previous studies that simply pool or vote features among views. In general, accuracy increases and standard deviation decreases as the number of training and test views increase. Furthermore, when the number of training and test views are same, accuracy increases. However, when it is trained with 4 views, test accuracy does not change much along the changes of the number of test views. Compared to other voting methods in softmax or pooling methods that take the maximum value of the features extracted at each viewpoint, our networks extract the correlation of features from multi-view images, exhibiting improving performance as the number of viewpoints increases.

We use MIW dataset[13] for further validation of MVSR-Net and MaterialNet in Table 3. Examples of MIW can be seen in Figure 4. There are 41 classes in MIW dataset and we group the classes of similar material types into 12 super-classes. The class names are paper(paper/tissue + cardboard + wallpaper), ceramic, stone(stone + concrete + brick), wood(wood + cork/corkboard + wicker), fabric(fabric/cloth + fur + carper/rug), glass(glass + mirror), granite/marble, metal, painted, plastic-clear, plastic-opaque and tile. We randomly extract 30 patch(31×31) set from each scene(375×250) for patch-wise classification. Single patch set consists of 25 patches with varying illumination conditions. We set the label of the center pixel of a patch as patch label. 25 patches are sorted by increasing order of brightness. We test our MSTH-Net, MVSR-Net, MaterialNet and Dual MaterialNets with ResNet18 as backbone. MSTH-Net are using single-illumination patch, and the rest of the networks are using multi-illumination patches. Due to smaller size of the input patches, MSTH-Net collects only entire-feature $\mathcal{F}_t$ to encode texture hierarchy features. Table 3 shows evaluation results on MIW dataset. Dual MaterialNets with color and difference images show better accuracy than single MaterialNet and MSTH-Net results. In smaller size of images where objects

Input	Method	Accuracy(%)
Single-view Color	MSTH-Net	74.24±2.3
Multi-view Color	MVSR-Net	82.60±2.1
	MaterialNet	84.43±2.1
Multi-view Color + Diff.	Dual MVSR-Net	83.17±2.2
	Dual MaterialNets	**86.21**±2.0

Table 3: MaterialNets Material Recognition Results on MIW dataset[13]: 'Diff' means difference images.

can not clearly be observed, intrinsic surface texture of materials are hardly observed. In such case, material type recognition is very challenging with single-view or single-illumination of single image. On the other hand, with multi-view condition, our MVSR-Net and MaterialNet encodes reflectance of the surface enabling material type recognition even with small images of little texture clue. Difference images obtained from multi-illumination images extract texture similar to 3D relief texture based on the difference in the shadow of different illumination. This is well known characteristic of multi-illumination images that has been exploited by popular photometric stereo method.

4.3 Ablation Studies

MSTH-Net		GTOS	FMD	KTH
Attention	Hierarchy			
✓		83.82±2.3	82.34±1.5	85.31±1.7
	✓	84.64±2.1	82.78±1.6	84.93±2.1
✓	✓	**84.95**±2.2	**83.17**±1.5	**86.69**±1.4

Table 4: Ablation studies on the effectiveness of texture attention and texture hierarchy in MSTH-Net

We conduct ablation studies analyzing the effectiveness of texture attention using salient features and texture hierarchy constructed with LSTM in MSTH-Net as summarized in Table 4. Texture attention is removed from MSTH-Net by discarding salient features and employing only entire features. With only entire features without attention from salient features that represent distinguishing local textures from multiple scales, MSTH-Net only encodes global texture from different scales. Texture hierarchy is removed from MSTH-Net by replacing LSTM by fully connected (FC) layers concatenating $\mathcal{H}_t, t = 1, ..., T$ and feeding them to the FC layers. Although FC layers extract the correlation of multi-scale textures, LSTM encodes the sequential correlation from lower scale texture to higher scale texture feature better. Consequentially, in Table 4, MSTH-Net with both texture attention and texture hierarchy shows best accuracy.

5 Conclusion

We propose MSTH-Net and MVSR-Net extracting view-independent comprehensive multi-scale texture hierarchy and view-specific surface reflectance features. MaterialNet is proposed combining MSTH-Net and MVSR-Net for material type recognition. Extensive evaluations on six public benchmarks have shown promising performance of our proposed method. Proposed material recognition method is able to be adopted in various practical applications where multi-view images can be obtained.

6 Acknowledgements

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [22ZH1200, The research of the fundamental media·contents technologies for hyper-realistic media space].

References

[1] Vincent Andrearczyk and Paul F Whelan. Using filter banks in convolutional neural networks for texture classification. *Pattern Recognition Letters*, 84:63–69, 2016.

[2] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015.

[3] Barbara Caputo, Eric Hayman, and P Mallikarjuna. Class-specific material categorisation. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1597–1604. IEEE, 2005.

[4] Zhile Chen, Feng Li, Yuhui Quan, Yong Xu, and Hui Ji. Deep texture recognition via exploiting cross-layer statistical self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5231–5240, 2021.

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

[6] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3828–3836, 2015.

[7] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[8] Stamatios Georgoulis, Vincent Vanweddingen, Marc Proesmans, and Luc Van Gool. Material classification under natural illumination using reflectance maps. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 244–253. IEEE, 2017.

[9] Bichuan Guo, Jiangtao Wen, and Yuxing Han. Deep material recognition in light-fields via disentanglement of spatial and angular information. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 664–679. Springer, 2020.

[10] Diane Hu, Liefeng Bo, and Xiaofeng Ren. Toward robust material recognition for everyday objects. In *BMVC*, volume 2, page 6. Citeseer, 2011.

[11] SeokYeong Lee and SeungKyu Lee. Surface ir reflectance estimation and material recognition using tof camera. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6463–6470. IEEE, 2021.

[12] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and Cihang Xie. Shape-texture debiased neural network training. *arXiv preprint arXiv:2010.05981*, 2020.

[13] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A dataset of multi-illumination images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4080–4089, 2019.

[14] Matthew Purri, Jia Xue, Kristin Dana, Matthew Leotta, Dan Lipsa, Zhixin Li, Bo Xu, and Jie Shan. Material segmentation of multi-view satellite imagery. *arXiv preprint arXiv:1904.08537*, 2019.

[15] Gabriel Schwartz and Ko Nishino. Material recognition from local appearance in global context. *arXiv preprint arXiv:1611.09394*, 2016.

[16] Lavanya Sharan, Ruth Rosenholtz, and Edward Adelson. Material perception: What can you see in a brief glance? *Journal of Vision*, 9(8):784–784, 2009.

[17] Lavanya Sharan, Ce Liu, Ruth Rosenholtz, and Edward H Adelson. Recognizing materials using perceptually inspired features. *International journal of computer vision*, 103(3):348–371, 2013.

[18] Yang Song, Fan Zhang, Qing Li, Heng Huang, Lauren J O'Donnell, and Weidong Cai. Locally-transferred fisher vectors for texture classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4912–4920, 2017.

[19] Ting-Chun Wang, Jun-Yan Zhu, Ebi Hiroaki, Manmohan Chandraker, Alexei A Efros, and Ravi Ramamoorthi. A 4d light-field dataset and cnn architectures for material recognition. In *European Conference on Computer Vision*, pages 121–138. Springer, 2016.

[20] Yunlong Wang, Kunbo Zhang, and Zhenan Sun. A novel deep-learning pipeline for light field image based material recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2422–2429. IEEE, 2021.

[21] Yong Xu, Feng Li, Zhile Chen, Jinxiu Liang, and Yuhui Quan. Encoding spatial distribution of convolutional features for texture representation. *Advances in Neural Information Processing Systems*, 34, 2021.

[22] Jia Xue, Hang Zhang, Kristin Dana, and Ko Nishino. Differential angular imaging for material recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 764–773, 2017.

[23] Jia Xue, Hang Zhang, and Kristin Dana. Deep texture manifold for ground terrain recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2018.

[24] Jia Xue, Hang Zhang, Ko Nishino, and Kristin Dana. Differential viewpoints for ground terrain material recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[25] Wei Zhai, Yang Cao, Jing Zhang, and Zheng-Jun Zha. Deep multiple-attribute-perceived network for real-world texture recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3613–3622, 2019.

[26] Wei Zhai, Yang Cao, Zheng-Jun Zha, HaiYong Xie, and Feng Wu. Deep structure-revealed network for texture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11010–11019, 2020.

[27] Hang Zhang, Kristin Dana, and Ko Nishino. Reflectance hashing for material recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3071–3080, 2015.

[28] Hang Zhang, Jia Xue, and Kristin Dana. Deep ten: Texture encoding network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 708–717, 2017.