MDPI

Article

Multi-View Learning for Material Classification

Borhan Uddin Sumon, Damien Muselet and Alain Trémeau

Univ Lyon, UJM-Saint-Etienne, CNRS, Institut Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France.

* Correspondence: damien.muselet@univ-st-etienne.fr

Abstract: Material classification is similar to textures classification and consists in predicting the material class of a surface in a color image, such as wood, metal, water, wool or ceramic. It is very challenging because of the intra-class variability. Indeed, the visual appearance of a material is very sensitive to the acquisition conditions such as viewpoint or lighting conditions. Recent studies show that deep convolutional neural networks (CNN) clearly outperform the hand-crafted feature in this context but suffer from a lack of data for training the models. In this paper, we propose two contributions to cope with this problem. First, we provide a new material dataset with large range of acquisition conditions so that CNN trained on this data provide features that can adapt to the appearance diversity of the material samples encountered in real-world. Second, we leverage the recent advances in multi-view learning methods to propose an original architecture designed to extract and combined features from several views of a single sample. We show that such multi-view CNN significantly improves the performance of the classical alternatives for material classification.

Keywords: Material classification; Multi-view learning; Texture analysis; Visual appearance; Material dataset.

1. Introduction

Material classification is a visual recognition task closely related to texture classification and dedicated to classify input texture/material images into categories such as fabrics, wood, steel or cotton [1]. It is of great interest to computer vision because predicting the material of objects in a scene can help for many applications : object manipulation by a robot, automatic waste sorting, predicting the appearance of an object under different lighting conditions, object recognition, ...

However, this is still a challenging problem since material images show a large intra-class variability [1,2]. First, the visual appearance of a material or a texture sample may significantly vary across viewing and lighting conditions. This is illustrated in Fig. 1, where each column represents the same sample, but observed under different lighting conditions and viewpoints. Second, different samples made from the same material can have different visual features, even when observed under similar conditions. This is the case, for example, of the two wool samples displayed in columns 2 and 3 of Fig 1. These two problems are very important for material recognition tasks and make it very challenging to extract relevant features from color images.

Recent studies have shown that deep neural networks clearly outperform many alternatives for material classification, but it is also clear that their performances are highly related to the data on which they are trained and tested [1–3]. For a material dataset showing small variations across acquisition conditions, a deep network can easily learned the specific features of each material and provide a very good recognition accuracy. When high variability exists in the acquisition conditions of the images (as for real world material appearance), we show, in this paper, that the performances can significantly drop. The first contribution of this paper is the constitution and provision of a dataset of material images with large intra-class variability, see Section 3.1. This dataset is called UJM-TIV (UJM is the abbreviation of our university, and TIV stands for Textures under varying Illumination,

Citation: Title. *Journal Not Specified* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

pose and Viewing). In this paper, we leverage this dataset to confirm that current classical neural network solutions do not generalize sufficiently to new data for real-world material observations. We hope that such a new diverse dataset will help to learn better material features in the future.

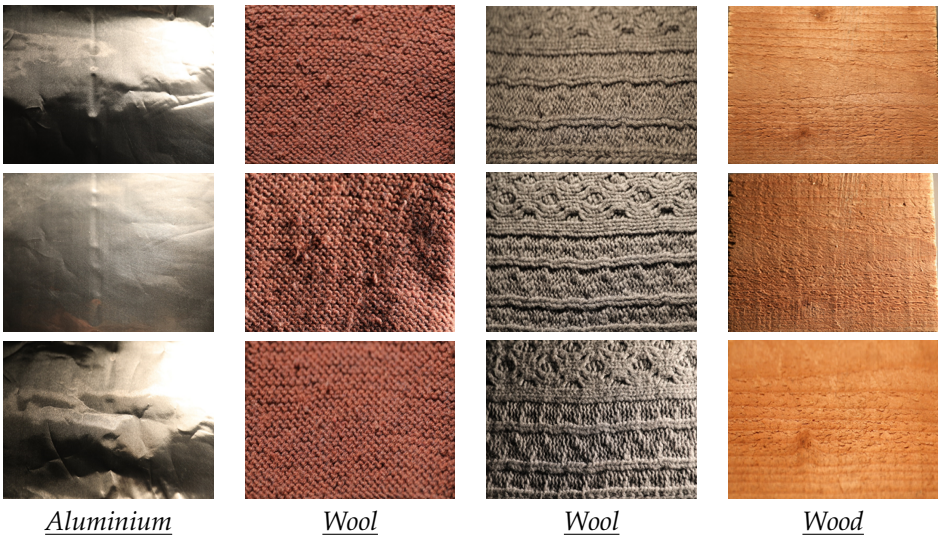


Figure 1. Appearance variation across acquisition conditions. The images of each column contain the same sample under different (lighting or viewpoint) conditions. These images are extracted from our new dataset.

Then, in order to go a step further towards better generalization of deep features for material classification, we propose to exploit a multi-view learning solution. Indeed, since one image provides a single view of a material sample, we claim that the performance could be significantly improved by considering a set of images for each material sample. Indeed, when a human being tries to determine the material that constitutes an object, he often tends to vary his point of view by moving his head or manipulating the object when possible to vary viewpoint and light direction. We propose to mimic this natural behaviour by taking advantage of the recent advances in multi-view learning [4] which makes it possible to extract features from several images and to merge them into a relevant representation. To the best of our knowledge, this is the first time that a multi-view learning approach is applied to material images in order to tackle the problem of appearance variations across viewing conditions.

Our contributions are fourfold:

- we analyze the current material datasets and show that they do not have enough intra-class diversity for material classification tasks,
- we provide a new public material dataset with high variations across acquisition conditions (lighting and viewpoint) in order to better represent the multiple appearances of a single real world material sample,
- we propose to exploit a multi-view learning approach to extract features from a set of images of the same material sample and to merge them into an accurate material representation,
- extensive tests on two material datasets show that exploiting multiple views of the same material sample clearly outperform the single-view alternative.

In Section 2 we present state of the art solutions designed for material classification and multi-view learning and discuss the different public material datasets. Next, Section 3.1 is devoted to the description of our new dataset. We detail the used materials, the lighting conditions, the acquisition device and the viewing conditions. We show why this dataset is more adapted for multi-view learning than the classical KTH-TIPS2 dataset [5] or any other existing datasets. Next, in Section 3.2, we present a deep network architecture designed for

two-view learning and test it on two material datasets, showing that it outperforms the alternative deep single-view classifier. The experimental results are reported in Section 4. Lastly a conclusion is drawn and future research directions are indicated in Section 5.

2. Related work

2.1. Material classification

Several categories of method have been proposed in the state of the art. The first ones were related to pattern recognition based methods, i.e. the computation of image features such as textons [6,7]. Next, filter banks based methods were proposed. They are related to the computation of local texture features [8–12]. Then, local texture features aggregation methods, like bags-of-textons [13], were introduced, they were designed in order to compute global texture features.

Some recent papers demonstrated the efficiency of CNN methods for material recognition (e.g. [14]) and the superiority of deep networks and off-the-shelf CNN-based features (e.g. [15]), particularly with non-stationary spatial patterns, such as textures, and in the presence of multiple changes in the acquisition conditions, against traditional, hand-crafted descriptors [1]. In [3] a selection of CNN architectures were evaluated and compared on various widely used material databases and achieved up to 92.5% mean average precision using transfer learning on MINC 2500 material database. In [1] a selection of state-of-the-art solutions (LFV+FC-CNN [16], Deep Ten [17], FV-CNN [18], B-CNN [19]) designed for material classification were evaluated and compared on various datasets (FMD, KTH-TIPS-2b, 4D-Light). The best classification accuracy obtained with these networks was around 83.% only for the KTH-TIPS-2b dataset.

Until recently most of material classification methods used only a single view image as input, or combined few single view image features as input. For example in [20] the authors used a multi-modal sensing technique, leveraging near-infrared spectroscopy and close-range high resolution texture imaging, to perform material classification.

In [21,22] the authors demonstrated that the concept of photometric stereo acquisition could improve the efficiency of material classification methods. They showed how micro-geometry and reflectance property of a surface could be used to infer its material. Likewise, Maximov et al. [23] and Vrancken et al. [24] demonstrated that combining different lighting and viewing conditions could slightly improve the material classification task.

In the ideal case, one would like to predict what would be the appearance of a material whatever the viewing direction and other factors having an impact on the capturing process. It is a quite challenging, ill-posed and under-constrained problem that remains hard to solve for the general case [2].

2.2. Multi-view learning

The aim of multi-view learning is to extract accurate features from data of different modalities (color image, text, audio, Lidar, ...), or representing different views of the same sample (different languages for texts, different acquisition conditions for images, ...) [4].

Very accurate features can be extracted from images with convolutional neural networks (CNN) and many approaches have integrated multi-view learning in the CNN [4,25–27]. The idea is to aggregate CNN features from different views into a more accurate general representation. Two main approaches based on multi-view CNN exist, as presented in [4]: the so called *one-view-one-net* mechanism uses one network per view and aggregates all the features through a fusion process [25,26] while the *multi-view-one-net* mechanism feeds a single network with all the views to extract features [27]. For the one-view-one-net solutions, the first networks used to extract the features usually share their weights in order to minimize the number of learned weights. The crucial points of such approaches lie in the feature fusion process. The main question with the multi-view-one-net solutions is about the aggregation of the inputs images before feeding the single network. The straightforward approach consists in concatenating these images into a multi-channel image and to apply convolutions on this image. This means that local features are extracted at the same

locations in these images, which requires a coarse registration between the images in order to get consistent features. Second, such a concatenation prevents the use of pre-trained networks that are usually fed with 3-channels images. This is the reason why, in this paper, we have chosen a one-view-one-net approach with a specific architecture.

Even if each element of our designed network has been carefully selected, the contribution of this paper is not in the definition of a new architecture for a general multi-view CNN. The main aim is rather to show that multi-view learning is an appropriate solution to tackle material classification. To the best of our knowledge, this is the first time that a multi-view CNN is used for this task.

2.3. Material datasets

Several categories of texture/material dataset have been introduced over the years. Some image sets were collected in lab settings from cropped stand-alone samples (eg CUReT [28] in 1999, KTH-TIPS [29] in 2005), meanwhile others were collected in the wild (eg FMD [30] in 2009, OpenSurfaces [31] in 2013, MINC [32] in 2015 and LFMD [33] in 2016) with more diverse samples and real-world scene context. The number of classes and the number of samples in each class vary a lot from one dataset to another one (eg 10 classes/810 images in total for KTH-TIPS, 61 classes/5612 images in total for CUReT), likewise the diversity of input parameters vary also significantly (eg small viewpoint changes in KTH-TIPS, higher viewpoint changes in CUReT) [34]. The KTH-TIPS (Textures under varying Illumination, Pose and Scale) image database was created to extend the CUReT database by providing variations in scale [29].

KTH-TIPS2 is an extension of KTH-TIPS [5] database. The KTH-TIPS2 contains 4 physical samples of 11 different materials (same material classes as KTH-TIPS) [35]. As the KTH-TIPS dataset it provides planar images with variations in scale, as well as variations in pose and illumination. From one physical sample to another one there is in some classes some strong (intra-class) variations (eg between wool or cracker samples) meanwhile for some other classes intra-class variations are lower (eg. between wood or cork samples). There is also some similitude between cotton and linen classes (ie. a small inter-class variance). In CUReT, only a single material instance is provided per class, consequently no generalization can be done to classifying material categories, due to a lack of intra-class variation. Changes in KTH-TIPS2 induced by a change of viewing directions or by a change of lighting conditions are respectively illustrated in Fig. 2 and in Fig. 3.

In most of material datasets the viewing and lighting conditions, and the camera settings, are well controlled and image acquisition is performed by a technician (a photograph) who takes care to perform the best acquisition (e.g. to minimize the blur, to minimize specularities) with the available setup system. But for some materials, such as aluminium foil samples, this is very challenging as this kind of material is very reflective.

Our aim was therefore to create a new dataset giving greater flexibility to the user in the images acquisition process. Our main objective was to perform images acquisition under various lighting and viewing directions, rather than under very strict and well controlled (and limited) lighting and viewing conditions. We assume that from one viewing direction to another one the average lightness of the sample may differ, as illustrated in Fig. 4(f) in comparison with 4(h). Lightness/color invariance is one of the invariance properties that a material classifier should have. We also assume that from one viewing direction to another one the contrast of the sample may differ, depending of the roughness and thickness of the materials, as illustrated in Fig. 4(a) in comparison with 4(e). Contrast invariance is one of the invariance properties that a material classifier should also have.

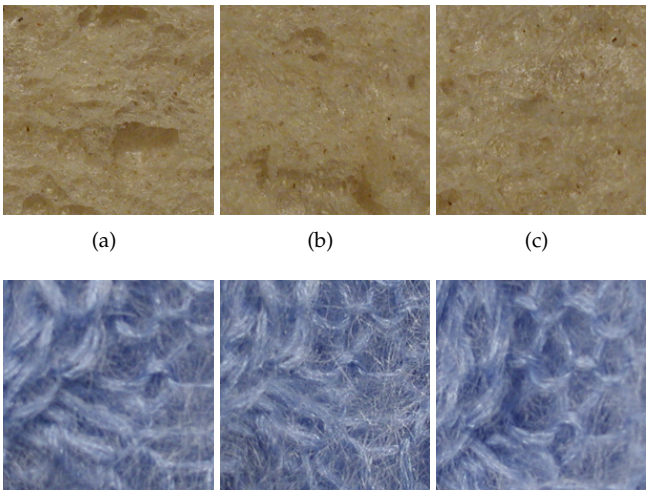


Figure 2. Changes of visual appearance of a white bread and wool sample from KTH-TIPS2 dataset under various lighting and viewing directions. Images (a) to (c) were captured with frontal illumination direction and frontal, 22.5° right and 22.5° left viewing directions, respectively, for a white bread sample. Similarly, images (d) to (f) were captured with frontal illumination direction and frontal, 22.5° right and 22.5° left viewing directions, respectively, for a wool sample.

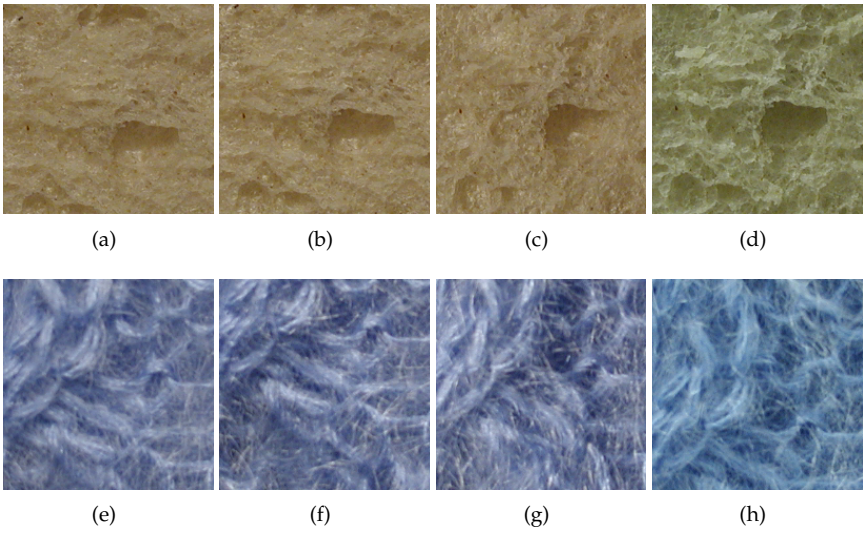


Figure 3. Changes of visual appearance of a white bread and wool sample from KTH-TIPS2 dataset under various lighting and viewing directions. Images (a) to (d) were captured with frontal viewing direction and frontal, 45° from top, 45° from side and ambient illumination condition, resp., for a white bread sample. Similarly images (e) to (h) were captured with frontal viewing direction and frontal, 45° from top, 45° from side and ambient illumination condition, resp., for a wool sample.

The fabric dataset introduced in [22] illustrates another kind of lightness shift due to a lighting field (an array of 12 LEDs) which is not spatially uniform on the sample area. This dataset contains 1266 samples which belong to one of the following fabric classes: cotton, terrycloth, denim, fleece, nylon, polyester, silk, viscose, and wool. The number of samples in each class is very unbalanced (588 in the cotton class, 32 in the terrycloth class). The samples were acquired under near-grazing illumination from a frontal view only. To perform photometric reconstruction the setup was geometrically calibrated.

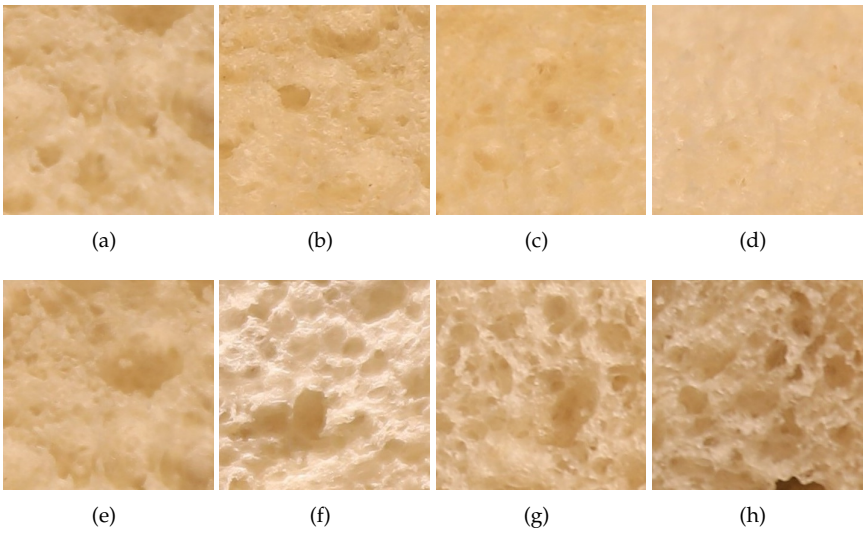


Figure 4. Changes of visual appearance of a white bread sample under various lighting geometries and viewing directions. Images (a) to (d) were acquired under with same lighting direction (90°). Images (e) to (h) were acquired under with same viewing direction (90°). For images (a) to (d) the lighting direction is fixed at 90°and viewing directions are 90°, 60°, 35°, and 10°, respectively. For images from (e) to (h) the viewing direction is fixed at 90°and lighting directions are 90°, 65°, 45°, and 20°, respectively.

Playing with lighting and viewing conditions, we can increase the difference of visual appearance for a material sample. In this paper we claim that the diversity of visual appearances of a material sample over acquisition condition variations should be accounted in the final feature vector to optimize the classification accuracy. For example, image differences observed in Fig. 6 are more significant than those observed in Fig. 5 as higher viewing and lighting angles were considered in the UJM-TIV dataset than in the KTH-TIPS2 dataset (see complementary information provided in Tables 1 and 2).

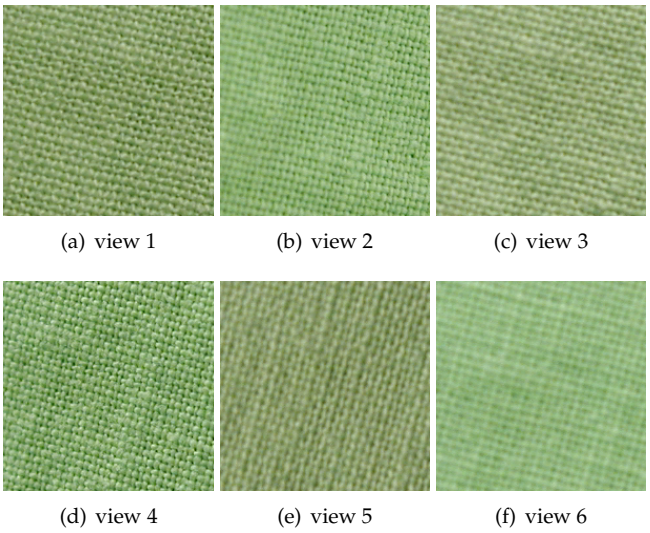
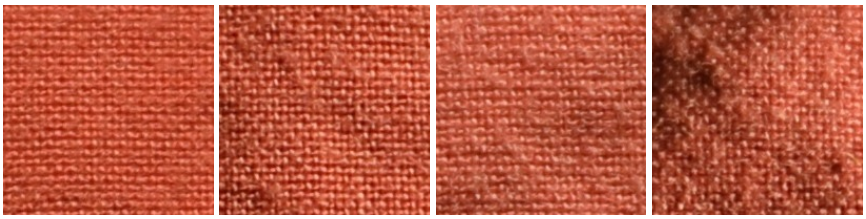


Figure 5. Images used for different views for a sample of cotton from KTH-TIPS2 dataset.



(a) view 1 (b) view 2 (c) view 3 (d) view 4

Figure 6. Images of a sample of cotton as different views from UJM-TIV dataset

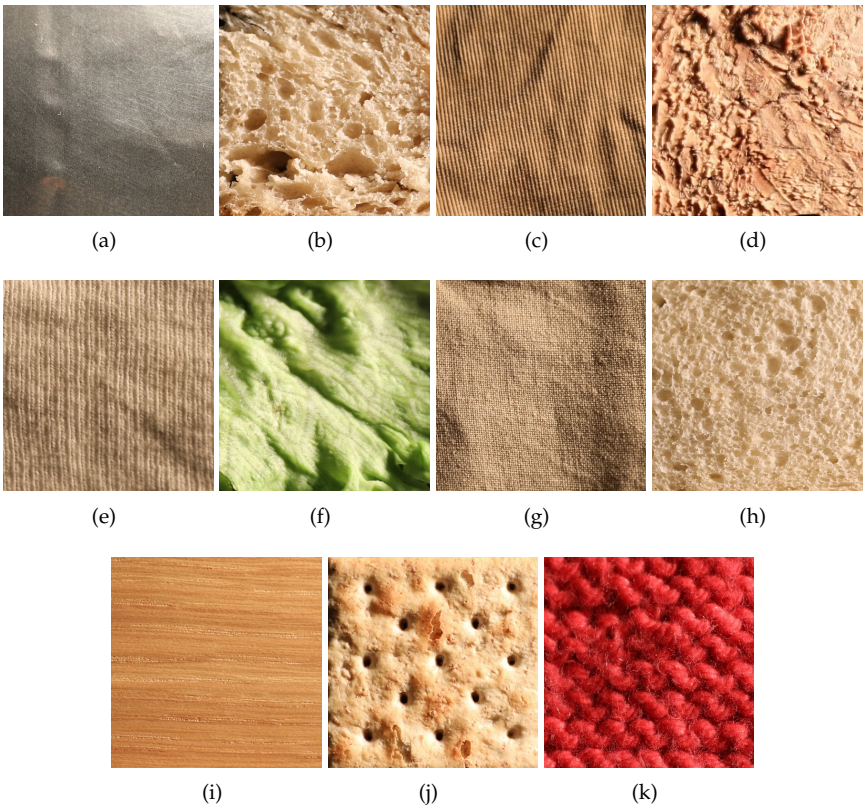
In next section, we present the details of our new datasets and the way we propose to exploit multiple views of a single material in order to boost the classification performance.

3. Materials and methods

3.1. Our new material dataset : UJM-TIV

3.1.1. General comments

The UJM-TIV material dataset consists of images from 11 distinct classes, namely aluminium foil, brown bread, corduroy, cork, cotton, lettuce leaf, linen, white bread, wood, cracker and wool (see Figure 7).



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k)

Figure 7. Images of a sample of : (a) aluminium foil, (b) brown bread, (c) corduroy, (d) cork, (e) cotton, (f) lettuce leaf, (g) linen, (h) white bread, (i) wood, (j) cracker, and (k) wool category from UJM-TIV dataset taken under illumination condition 65°and viewing condition 90°.

These images were acquired under controlled viewing and lighting conditions. These 11 classes are also included in the KTH-TIPS2 [35] dataset. Due to the diversity of samples in each material category, the visual appearance of UJM-TIV samples is not similar to the one of KTH-TIPS2 samples. Stronger appearance differences appeared at lower viewing angles or lower illumination angles (see Figure 8).

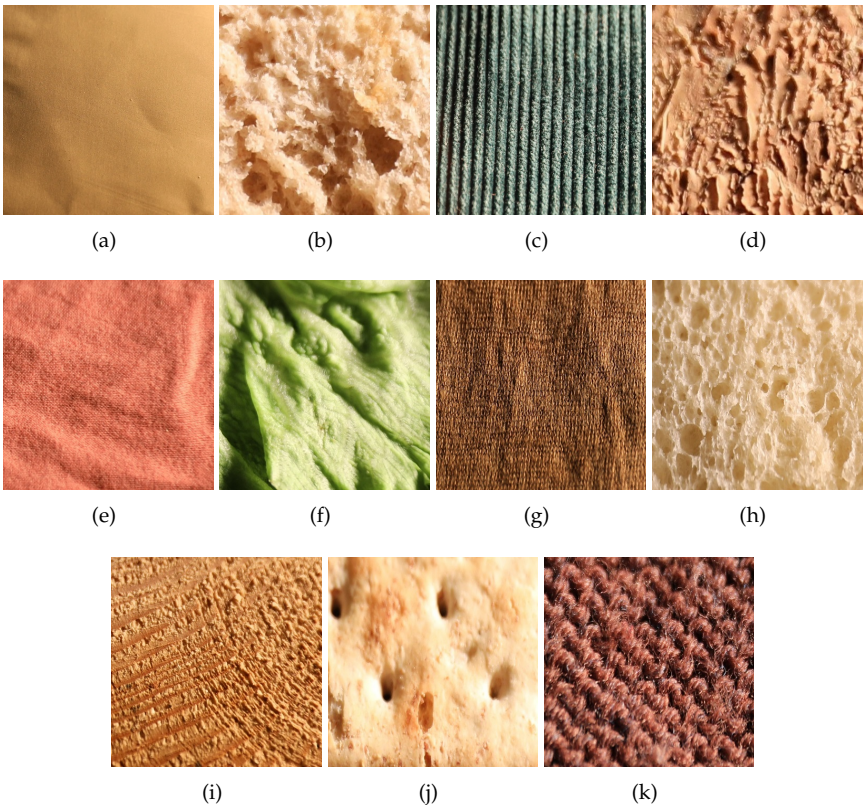


Figure 8. Images of a sample of (a) aluminium foil, (b) brown bread, (c) corduroy, (d) cork, (e) cotton, (f) lettuce leaf, (g) linen, (h) white bread, (i) wood, (j) cracker, and (k) wool category from UJM-TIV dataset taken under illumination direction 65° and viewing condition 35°.

In the UJM-TIV dataset the variation in appearance between samples is clearly larger for some categories (eg. wood and wool) than in KTH-TIPS2. Furthermore, among UJM-TIV, wool and cotton have the highest appearance variations, while cork, brown bread, and white bread have the lowest intra-class variations. As illustration, see changes of appearance shown in Figures 1 and in 9.

3.1.2. Acquisition settings and image processing

For our dataset, a Canon EOS 5D Mark IV digital camera was used to capture the images of the samples with a resolution of 6720 x 4480 pixels. The background surrounding each sample was removed using a post processing step. For each object sample, two object poses were taken into consideration, with a variation of 90° rotation around the surface normal (**N** in Fig. 11). The example shown in Fig. 10 illustrates how such a change can modify the material appearance for a given material sample.

The image acquisition setup used to capture the images under controlled viewing and lighting conditions is illustrated in Figure 11. In this figure **S** is the material sample, **I** the illumination source and **V** the viewpoint direction. Four standard light sources (60W tungsten light bulb) were used, one for each lighting direction $\{\theta_i, \varphi_i\}$ used (frontal, roughly 20°, roughly 45°, and roughly 65°). Four viewing directions $\{\theta_v, \varphi_v\}$ (frontal, roughly 60°, roughly 30°, and roughly 10°) were used for each object pose. Therefore, there is a total of 16 (4 illumination directions x 4 viewing directions) images per sample position captured for each material sample. For two poses, a total of 32 images were captured for each sample. The acquisition were performed in a dark room without any ambient illumination.

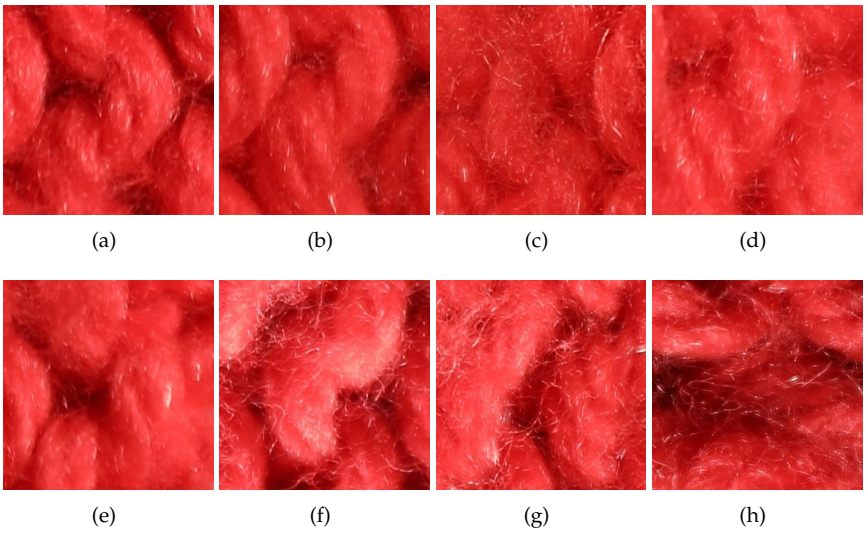


Figure 9. Changes of visual appearance of a wool sample under various lighting geometries and viewing directions. Images (a) to (d) were acquired under with same lighting direction (90°). Images (e) to (h) were acquired under with same viewing direction (90°). For images (a) to (d) the lighting direction is fixed at 90° and viewing directions are 90°, 60°, 35°, and 10°, resp.. For images from (e) to (h) the viewing direction is fixed at 90° and lighting directions are 90°, 65°, 45°, and 20°, resp.

Figure 10. Images of a cotton sample from UJM-TIV: (a) when viewing condition is frontal and lighting condition is at 20°. (b) with same viewing and lighting condition when sample orientation is perpendicular.

The Patchify [36] library was used to extract 200 x 200 pixel image patches from the samples. Patches with background and too blurry images were removed manually from all extracted patches. The number of patches extracted from each sample varies from sample to sample. The dataset contains around 75 thousand image patches after removing the blurred and patches with out of focus from the all extracted patches.

3.1.3. Comparison with previous datasets

The viewing directions used in UJM-TIV are different from those used in KTH-TIPS2 (frontal, rotated 22.5° left and 22.5° right) and with a larger range. The lighting directions used in UJM-TIV are also different from those used in KTH-TIPS2 (frontal, 45° from the top and 45° from the side, all taken with a desk-lamp with a Tungsten light bulb).

All samples captured in the KTH-TIPS2 were acquired under a combination of three viewing directions (frontal, rotated 22.5° left and rotated 22.5° right) and four illumination directions (from the front, from the side at roughly 45° and from the top at roughly 45°, and using ambient lighting), different from the ones used in UJM-TIV. They were also captured at different scales oppositely to UJM-TIV.

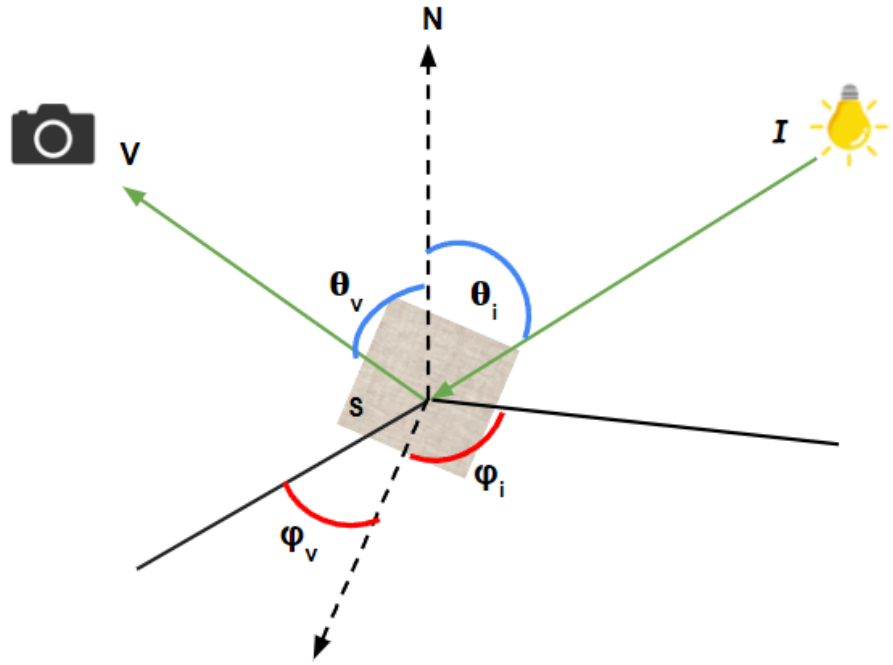


Figure 11. Schematic diagram of the image acquisition setup. In our experiments the plane defined by vectors **N** and **I** was set perpendicular to the plane defined by vectors **N** and **V**.

As with KTH-TIPS2, in UJM-TIV few images of fine-structured materials appear out of focus at working distances due to perspective effects and roughness of materials, see Figure 12 where all the images shown are captured under viewing direction around 10° and illumination direction 20°.

Oppositely to some other setups, such as the ones described in [22] or [37], in this study our aim was not to tailor a lighting system which optimizes the light source positions depending of the various materials to acquired.

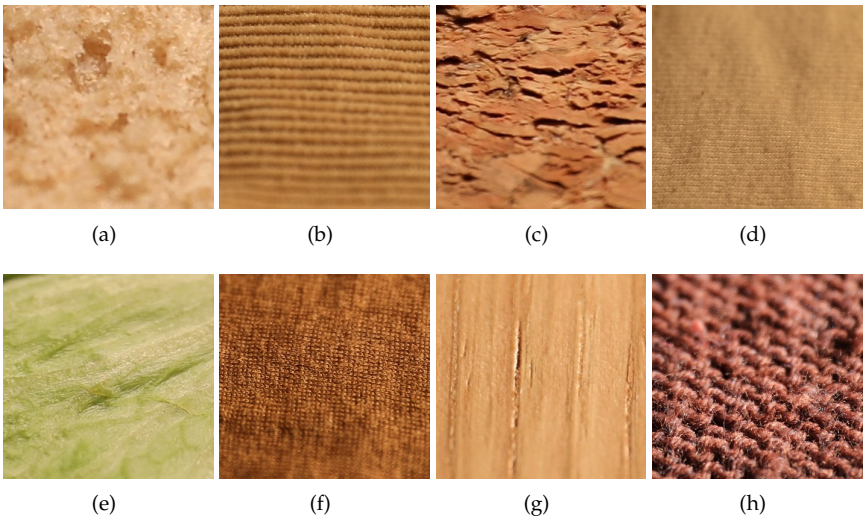


Figure 12. Image samples appeared as out of focus for category (a) brown bread, (b) corduroy, (c) cork, (d) cotton, (e) lettuce leaf, (f) linen, (g) wood, and (h) wool from UJM-TIV dataset.

3.2. Multi-view learning with Siamese networks

Multi-view learning is attracted many researchers today [4] since it allows to extract features from multiple views and to merge them into an accurate global representation. As explained earlier, a *one-view-one-net* mechanism is more adapted to material classification. In this case, each image (view) is feed to a deep backbone to extract features, then the features of each view are merged and used as input to a classification network that predicts the class of the considered sample. Once again, our contribution, here, is not in the definition of the best architecture for this task but rather to leverage the multi-view learning area to show that it can significantly improve the performance for material classification.

Hence, we have selected a simple *one-view-one-net* architecture with a pre-trained network, leaving for future works any improvements related to the architecture choice.

Since each view is feeding a backbone, we propose to share the weights between these backbones in order to minimize the number of learned weights and prevent overfitting. Furthermore, sharing the weights between backbones also helps to improve the generalization power of the model, since the same backbone has to extract accurate features from different views (different appearances). A single architecture merging the outputs of two identical branches is a Siamese network [38–40].

The architecture of the proposed network is shown in Figure 13. The Siamese network takes a pair of images as input from two different views and feeds it to one backbone. In our case, a pre-trained ResNet50 [41] is used as backbone. Each branch learns the features from each input view. Then the learned features are concatenated together and the result feeds the fully connected layers for classification. It is worth mentioning that all the blocks are differentiable so that this architecture can be trained end-to-end (feature extraction and classification) with a single classification loss.

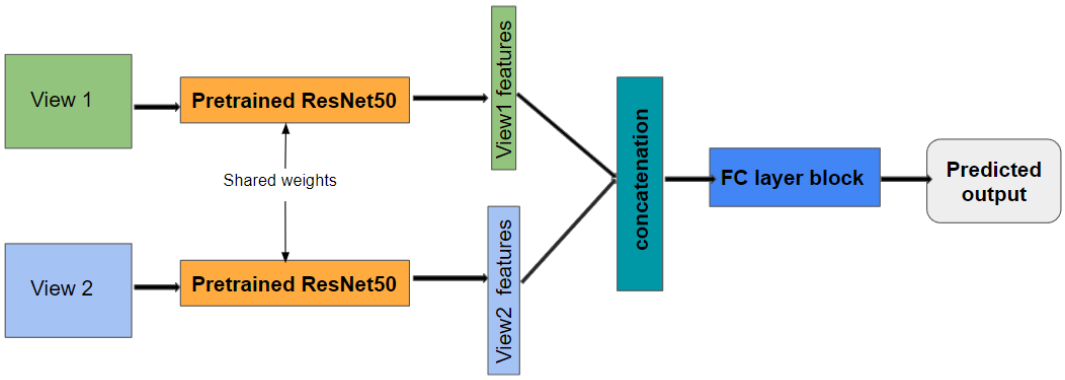


Figure 13. The proposed Siamese architecture for multi-view learning.

Before concatenating the features of each view, a global average pooling layer is used in order to reduce the number of parameters to learn in the first fully-connected (FC) layer of the architecture. This block also helps to prevent overfitting problems. Furthermore, in order to regularize the classifier, dropout is applied in the FC layers.

The advantage of such an architecture is that it can be easily adapted to more views than two. Indeed, the pre-trained backbone can be used to extract features from any new views and only the FC layer has to be adapted and retrained to perform classification. In this paper, we have just trained and tested a two-branch architecture.

4. Results and discussion

In order to assess the quality of our new dataset and the performance of the propose multi-view CNN, we have conducted many tests on two datasets. The idea was to compare the advantages of our dataset over the KTH-TIPS2 dataset and to compare the performance of our two-views CNN with a single view alternative.

4.1. Experimental settings

We have created two architectures for our tests. One classical single-branch architecture with a convolutional backbone to extract features and FC layers for classification. The accuracy provided by this network is called Single-view accuracy. Then, we have used our Siamese architecture with two backbones with shared weights that extract features from two views and FC layers for classification. This architecture provides the so called Multi-view accuracy. As backbone for these architectures we have selected a residual network ResNet50 [41] pretrained on ImageNet dataset. The last convolutional layer of this network is fine-tuned on the considered data, while the other layers are frozen. For each architecture, the number of FC layers and the number of neurons in each layer are cross-validated for fair comparison. Finally, the number of learned parameters is equivalent between each architecture (7.1 millions for the single-view and 7.7 millions for the multi-view).

Likewise, the hyperparameters and optimization algorithms are the same for both networks. We use Adam optimizer with initial learning rate of 0.001. For each experiment, the learning rate automatically decreases with a factor of 0.2 when the loss meets a *plateau*. The maximum number of epochs is fixed to 350. Input images were resized to 224 x 224 before feeding the network with a batch size of 16.

Keras framework with TensorFlow 2.8.0 backend and Python version 3.9.5 was used to implement the both single branch and Siamese network. The models were trained on a high-performance GPU with an NVIDIA RTX 8000 8GB graphics card, CUDA version 11.2, and RAM of size 16 GB.

4.2. Data

We run experimental tests with two different configurations. The first configuration consists in training and testing on the whole considered dataset. Each dataset is randomly splitted in training and test sets with respectively 70% and 30% of the data, providing the sets called *KTH-TIPS2 Train*, *KTH-TIPS2 Test*, *UJM-TIV Train* and *UJM-TIV Test*.

Then, in order to test the multi-view learning, we select some views in both datasets: 6 views in KTH-TIPS2 and 4 views in UJM-TIV. All the images of each selected view are also randomly split with ratio 70% and 30% for training and testing, respectively.

Table 1 details the viewing and illumination conditions of the selected views from the KTH-TIPS2 dataset. As observed in Fig. 5, changes of viewing and illumination directions have an impact on the overall appearance of the observed cotton sample (more blur, less contrast, etc.), but these changes are not significant (lower than changes of appearance between samples belonging to the same category, i.e. changes induced by intra-class variation).

Table 1. Viewing and illumination conditions of selected views from KTH-TIPS2 [35] dataset

View	Viewing direction	Illumination direction
View1	Frontal	Frontal
View2	22.5° left	Ambient
View3	Frontal	45° from top
View4	22.5° right	Ambient
View5	Frontal	45° from side
View6	22.5° right	Ambient

Table 2 details the viewing and illumination conditions of the selected views from the UJM-TIV dataset used for the experiment. Similarly, Fig. 4 shows the images of four different views of a white bread sample from our new dataset used in multi-view experiment.

Table 2. Viewing and illumination condition for selected views from UJM-TIV dataset showed in Figure 6

View	Viewing direction	Illumination direction
View1	90°	90°
View2	90°	45°
View3	90°	20°
View4	60°	65°

4.3. Results

The results are organized in two sections, depending on which data the networks have been trained and tested. First, we show results for test on the whole datasets and then, results on selected views.

4.3.1. Appearance diversity of the datasets

First, the idea is to analyze the results of a single-branch network on the whole datasets. The results are provided in Table 3 for both datasets. First, we can notice that the obtained accuracy for KTH-TIPS2 (80%) is similar to the ones obtained by classical deep networks in [42]. Second, we notice that the accuracy obtained on our UJM-TIV dataset with the same settings as the ones used on KTH-TIPS2 is much lower. This means that a single branch network performs better on KTH-TIPS2 than on our dataset. We think that it is directly related to the higher intra-class variability of our dataset.

Table 3. Model accuracy of single branch network with KTH-TIPS2 and UJM-TIV when considering all the views.

Train data	Test data	Val. accuracy
KTH-TIPS2 Train	KTH-TIPS2 Test	80.00
UJM-TIV Train	UJM-TIV Test	55.26

4.3.2. Multi-view learning

In this section, we provide results on both datasets when the networks are trained and tested on selected views. We consider the views by pairs in order to test our two-views deep architecture. Thus, we have trained a network (single- or two- views) with the images of the two considered views (training set) and tested on the same views (test set).

The results are provided in Table 4 for the KTH-TIPS2 dataset and in Table 5 for our UJM-TIV dataset. First, we can notice that considering only two views for training overall reduce the accuracy compared when training on the whole dataset (which was 80% for KTH-TIPS2 and 55% for UJM-TIV). This is not surprising since here, the network has been trained on less data than when the whole dataset was used. Second, we observe that the multi-view network significantly outperforms the single-view network for all selected view pairs. This clearly shows that multi-view learning is a relevant solution for material classification. And we can notice that the improvement provided by the multi-view training over the single-view is much higher when the two views present very different appearance. This is the case for the dataset KTH-TIPS2 between view5 and view6 (from 56% to 71%) where view6 has a very different lighting conditions than view5. For our dataset, the improvement from single-view to two-views is important for all the considered pairs of views. This is due to the high variation in appearance between the views for our dataset.

Table 4. Model accuracy of single-view and multi-view learning on KTH-TIPS2.

Train data	Test data	Single-view accuracy	Multi-view accuracy
KTH-TIPS2 view1,view2 Train	KTH-TIPS2 view1, view2 Test	56.90	68.53
KTH-TIPS2 view3,view4 Train	KTH-TIPS2 view3, view4 Test	60.34	67.24
KTH-TIPS2 view5,view6 Train	KTH-TIPS2 view5, view6 Test	56.91	71.98

Table 5. Model accuracy of single-view and multi-view learning on our UJM-TIV dataset.

Train data	Test data	Single-view accuracy	Multi-view accuracy
UJM-TIV view1,view2 Train	UJM-TIV view1, view2 Test	50.28	79.52
UJM-TIV view3,view4 Train	UJM-TIV view3, view4 Test	60.00	75.29

These results clearly show that our dataset is well designed to train networks for material classification and that the proposed Siamese architecture is a relevant solution for two-views learning.

5. Conclusion

In this paper, we have proposed several contributions for material classification. We have introduced a new dataset with large intra-class variability. The appearance variations within each class are due to large range of acquisition conditions and selection of diverse material samples. We have shown that classical deep networks cannot generalize easily on such data, demonstrating the need of alternative solutions for this task. In order to exploit the appearance variations across viewing conditions, we have proposed to leverage the strengths of recent solutions in multi-view learning. We have shown that a Siamese architecture significantly outperforms the single-branch alternative by merging features from two views. Obviously, increasing the number of views at the input of the network is a solution that will be investigated in our future works. The challenge here, will be to extract features from uncontrolled views and to merge them into a general representation of the considered sample. Next, we plan to demonstrate that multi-view learning could also contribute to better reconstruct (photometrically) complex spatially varying BRDF and to improve the efficiency of single image SVBRDF-based rendering methods (see [43]).

Author Contributions:

Supervision/project administration: A.T. and D.M.; Conceptualization, D.M., B.U.S. and A.T.; Methodology/formal analysis, D.M. and B.U.S.; Software/investigation, D.M. and B.U.S.; Validation, D.M., B.U.S. and A.T.; Writing—original draft preparation, A.T., B.U.S. and D.M.; Writing—review and editing, A.T. and D.M.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement:

The UJM-TIV dataset will be publicly available on an open access repository. The UJM-TIV dataset is already available upon request.

Conflicts of Interest:

The authors have no conflict of interest to declare.

References

1. Xu, S. Transfer learning for material classification based on material appearance correspondances. PhD thesis, University Jean Monnet, 2021.

2. Trémeau, A.; Xu, S.; Muselet, D. Deep Learning for Material recognition: most recent advances and open challenges. *arXiv preprint arXiv:2012.07495* **2020**.
3. Sticlaru, A. Material Classification using Neural Networks. *arXiv preprint arXiv:1710.06854* **2017**.
4. Xiaoqiang, Y.; Shizhe, H.; Yiqiao, M.; Yangdong, Y.; Hui, Y. Deep multi-view learning methods: A review. *Neurocomputing* **2021**, pp. 106–129.
5. Fritz, M.; Hayman, E.; Caputo, B.; Eklundh, J.O. The kth-tips database **2004**.
6. Julesz, B. Textons, the elements of texture perception, and their interactions. *Nature* **1981**, *290*, 91–97.
7. Julesz, B.; Bergen, J.R. Human factors and behavioral science: Textons, the fundamental elements in preattentive vision and perception of textures. *Bell System Technical Journal* **1983**, *62*, 1619–1645.
8. Bovik, A.C.; Clark, M.; Geisler, W.S. Multichannel texture analysis using localized spatial filters. *IEEE transactions on pattern analysis and machine intelligence* **1990**, *12*, 55–73.
9. Jain, A.K.; Farrokhnia, F. Unsupervised texture segmentation using Gabor filters. *Pattern recognition* **1991**, *24*, 1167–1186.
10. Turner, M.R. Texture discrimination by Gabor functions. *Biological cybernetics* **1986**, *55*, 71–82.
11. Zhu, S.C. Statistical modeling and conceptualization of visual patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2003**, *25*, 691–712.
12. Manjunath, B.S.; Ma, W.Y. Texture features for browsing and retrieval of image data. *IEEE Transactions on pattern analysis and machine intelligence* **1996**, *18*, 837–842.
13. Leung, T.; Malik, J. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International journal of computer vision* **2001**, *43*, 29–44.
14. Liu, L.; Chen, J.; Fieguth, P.; Zhao, G.; Chellappa, R.; Pietikäinen, M. From BoW to CNN: Two decades of texture representation for texture classification. *International Journal of Computer Vision* **2019**, *127*, 74–109.
15. Bello-Cerezo, R.; Bianconi, F.; Di Maria, F.; Napoletano, P.; Smeraldi, F. Comparative evaluation of hand-crafted image descriptors vs. off-the-shelf CNN-based features for colour texture classification under ideal and realistic conditions. *Applied Sciences* **2019**, *9*, 738.
16. Song, Y.; Zhang, F.; Li, Q.; Huang, H.; O'Donnell, L.J.; Cai, W. Locally-transferred fisher vectors for texture classification. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4912–4920.
17. Zhang, H.; Xue, J.; Dana, K. Deep ten: Texture encoding network. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 708–717.
18. Cimpoi, M.; Maji, S.; Vedaldi, A. Deep filter banks for texture recognition and segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3828–3836.
19. Lin, T.Y.; RoyChowdhury, A.; Maji, S. Bilinear convolutional neural networks for fine-grained visual recognition. *IEEE transactions on pattern analysis and machine intelligence* **2017**, *40*, 1309–1322.
20. Erickson, Z.; Xing, E.; Srirangam, B.; Chernova, S.; Kemp, C.C. Multimodal material classification for robots using spectroscopy and high resolution texture imaging. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 10452–10459.
21. Gorpas, D.; Kampouris, C.; Malassiotis, S. Miniature photometric stereo system for textile surface structure reconstruction. In Proceedings of the Videometrics, Range Imaging, and Applications XII; and Automated Visual Inspection. International Society for Optics and Photonics, 2013, Vol. 8791, p. 879117.
22. Kampouris, C.; Zafeiriou, S.; Ghosh, A.; Malassiotis, S. Fine-grained material classification using micro-geometry and reflectance. In Proceedings of the European Conference on Computer Vision. Springer, 2016, pp. 778–792.
23. Maximov, M.; Leal-Taixé, L.; Fritz, M.; Ritschel, T. Deep appearance maps. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8729–8738.
24. Vrancken, C.; Longhurst, P.; Wagland, S. Deep learning in material recovery: Development of method to create training database. *Expert Systems with Applications* **2019**, *125*, 268–280.
25. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional Two-Stream Network Fusion for Video Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.
26. Yang, Z.; Tang, L.; Zhang, K.; Wong, P.K. Multi-View CNN Feature Aggregation with ELM Auto-Encoder for 3D Shape Recognition. *Cognitive Computation* **2018**, *10*, 908–921.
27. Dou, Q.; Chen, H.; Yu, L.; Qin, J.; Heng, P.A. Multilevel Contextual 3-D CNNs for False Positive Reduction in Pulmonary Nodule Detection. *IEEE Transactions on Biomedical Engineering* **2017**, *64*, 1558–1567.
28. Dana, K.J.; Van Ginneken, B.; Nayar, S.K.; Koenderink, J.J. Reflectance and texture of real-world surfaces. *ACM Transactions On Graphics (TOG)* **1999**, *18*, 1–34.
29. Caputo, B.; Hayman, E.; Mallikarjuna, P. Class-specific material categorisation. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. IEEE, 2005, Vol. 2, pp. 1597–1604.
30. Sharan, L.; Rosenholtz, R.; Adelson, E. Material perception: What can you see in a brief glance? *Journal of Vision* **2009**, *9*, 784–784.
31. Bell, S.; Upchurch, P.; Snavely, N.; Bala, K. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on graphics (TOG)* **2013**, *32*, 1–17.
32. Bell, S.; Upchurch, P.; Snavely, N.; Bala, K. Material recognition in the wild with the materials in context database. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3479–3487.

33. Wang, T.C.; Zhu, J.Y.; Hiroaki, E.; Chandraker, M.; Efros, A.A.; Ramamoorthi, R. A 4D light-field dataset and CNN architectures for material recognition. In Proceedings of the European conference on computer vision. Springer, 2016, pp. 121–138.
34. Hu, Y.; Long, Z.; Sundaresan, A.; Alfarraj, M.; AlRegib, G.; Park, S.; Jayaraman, S. Fabric surface characterization: assessment of deep learning-based texture representations using a challenging dataset. *The Journal of The Textile Institute* **2021**, *112*, 293–305.
35. Mallikarjuna, P.; Targhi, A.T.; Fritz, M.; Hayman, E.; Caputo, B.; Eklundh, J.O. The kth-tips2 database. *Computational Vision and Active Perception Laboratory, Stockholm, Sweden* **2006**, *11*.
36. Python patchify library. https://pypi.org/project/patchify/. Accessed: 10-05-2022.
37. Kapeller, C.; Antensteiner, D.; Štolc, S. Tailored photometric stereo: Optimization of light source positions for various materials. *Electronic Imaging* **2020**, *2020*, 71–1.
38. Koch, G.; Zemel, R.; Salakhutdinov, R.; et al. Siamese neural networks for one-shot image recognition. In Proceedings of the ICML deep learning workshop. Lille, 2015, Vol. 2, p. 0.
39. Wiggers, K.L.; Britto, A.S.; Heutte, L.; Koerich, A.L.; Oliveira, L.S. Image retrieval and pattern spotting using siamese neural network. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019, pp. 1–8.
40. Melekhov, I.; Kannala, J.; Rahtu, E. Siamese network features for image matching. In Proceedings of the 2016 23rd international conference on pattern recognition (ICPR). IEEE, 2016, pp. 378–383.
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
42. Sixiang, X.; Damien, M.; Alain, T.; Robert, L. Confidence-based Local Feature Selection for Material Classification. In Proceedings of the 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), 2020.
43. Deschaintre, V.; Aittala, M.; Durand, F.; Drettakis, G.; Bousseau, A. Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (ToG)* **2018**, *37*, 1–15.