Master in Computational Colour and Spectral Imaging (COSI)

Deep Learning Methods for Classification of Photometric Images of Materials

Master Thesis Report

Presented by

Md Borhan Uddin Sumon

and defended at the

Norwegian University of Science and Technology

September 2022

Academic Supervisor(s): Professor Alain Trémeau, and Associate Professor Damien Muselet
Co-Supervisor: Associate Professor Simone Bianco
Jury Committee:

1. Professor Marius Pedersen, NTNU, Norway
2. Dr. Dmitry Semenov, UEF, Finland

Submission of the thesis: 10th August 2022
Day of the oral defense: 1st September 2022

Abstract

A key topic in the field of computer vision is image classification, which involves predicting one class for each input image. Additionally, one of its tasks is the categorization of materials from images, which is difficult for both human and computer systems since materials might appear differently based on their surface characteristics, lighting geometry, viewing geometry, camera settings, etc. The revolutionary image classification architecture, deep convolutional neural networks (CNN) has shown promising results as compared to hand-crafted computer vision methods in recent studies for material classification. However, the number of material datasets that mimic the behavior of the real-world material is limited. To this end, our two contributions are reported. We proposed a new material dataset where images were acquired with larger acquisition settings. The dataset is developed in such a way that convolutional neural networks used to train on this dataset can produce features that can be adjusted with the varying appearance changes found in real-world material images. In order to integrate key features extracted from multiple perspectives of a same material sample, we proposed a distinct architecture that takes advantage of the current developments in multi-view learning techniques. We show that the proposed multi-view network can be used for both feature extraction and classification while significantly outperforming the traditional single-view network for material classification.

Dedication

I would like to dedicate this thesis to my parents and elder brother who always support me and without them it would not be possible for me to complete this long journey.

Acknowledgment

I want to express my gratitude and acknowledgement to my supervisor Alain and Damien, who made this thesis possible. Especially Damien who always helped me with proper guidance to complete the thesis. And thanks to Alain for his insightful corrections and suggestions to enrich the quality of the thesis report.

Acronyms

CNN	Convolutional Neural Network
BRTF	Bi-directional Reflection Transmittance Function
BTF	Bidirectional Texture Function
ILSVRC	The ImageNet Large Scale Visual Recognition Challenge
UJM TIV	University Jean Monnet, Texture under varying Illumination, pose and Viewing
LBP	Local Binary Pattern
SURF	Speeded Up Robust Feature
SIFT	Scale Invariant Feature Transform
LEM	Learnable Encoding Module
DEP	Deep Encoding Pooling Network
EOT	Extent-of-texture
EOS	Extent-of-shape
GAP	Global Average Pooling
GMP	Global Max Pooling
FC	Fully Connected
KNN	K Nearest Neighbors
GPU	Graphics Processing Unit
TF	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
PCA	Principal Component Analysis
BRDF	Bidirectional Reflectance Distribution Function
SVBRDF	Spatially Varying Bidirectional Reflectance Distribution Function

Contents

1 Introduction

Material classification can be defined as a computer vision task that involves categorizing images of a material into class like wool, linen, brown bread etc. based on the available visual information. It has great importance in the field of object tracking, robotics, waste management, automatic sorting of textile samples etc. For instances material recognition can be implemented into robotic visual systems which allow product search, object manipulation or autonomous navigation on a surface made of specific material. Classification of a material using an image is challenging because the visual appearance of the material can be vary depending on a number of factors such as illumination conditions, viewing angle, texture properties of the material etc.

Material classification has gained the attraction of the researchers in 1960's (Julesz, 1962) aiming on describing material with expert defined features. Julesz introduced his pioneering work on texton theory (Julesz, 1981; Julesz and Bergen, 1983) where elementary local features such as edges or corners are used to define the texture or material's descriptors, called textons. After that, a number of researchers began to work on designing efficient filter banks to extract texture features (Bovik et al., 1990; Jain and Farrokhnia, 1991; Malik and Perona, 1990; Turner, 1986; Manjunath and Ma, 1996; Zhu, 2003) . Besides the study of local features extraction, a number of approaches like bags of textons (Leung and Malik, 2001) were proposed to aggregate local features into global representation which can more effectively depicts material images. At physical level materials appearances were collected under control conditions which means the parameters such as lighting color, or direction and viewing condition were strictly set and recorded. With these controlled conditions and appearances some models such as BRTF (Bi-directional Reflection Transmittance Function) and BTF (Bidirectional Texture Function) can be built to characterize the appearance of material instances. These models provide instance level features which are more useful to identify material instances rather than material categories. A key characteristic for material images is that target material occupied the whole region of an image and no clutter background was involved.

Deep learning-based approaches have become more popular as processing power and access to large datasets have increased, making them an attractive solution to solve many problems in society. After the AlexNet (Krizhevsky et al., 2012) a deep convolutional neural network broke the image classification accuracy record in 2012 in the ImageNet ILSVRC (Russakovsky et al., 2015) on a very large dataset, deep learning methods have become the spotlight for the researchers to solve many unsolved problems. These problems were not possible to solve before because of the lack of appropriate computing power and large datasets were not also available to train the deep learning models.Recent research has demonstrated that deep learning-based approaches significantly outperform the various material classification methods(Xu, 2021; Trémeau et al., 2020; Sticlaru, 2017). It is also true that the performances of these methods are highly dependent on the data they are trained and tested. Deep neural networks trained on a material dataset that has small variation across its viewing and illumination directions, in such scenario relevant features from the material dataset can be easily learn by the neural network and achieve high classification accuracy. On the other hand, we demonstrate in this thesis that accuracy can drastically decrease when the networks are trained on a material where there is great variability in the image acquisition settings.

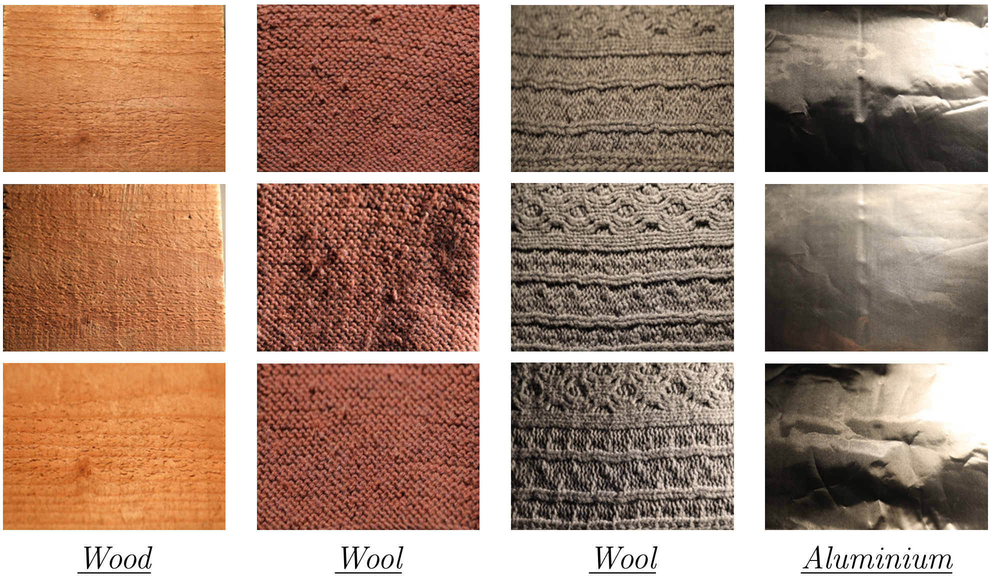


Figure 1.1: *Variation in appearance across different acquisition settings. Each column shows the same images sample captured under various (illumination or viewing) conditions. These images were taken from UJM-TIV dataset. Image from (Sumon et al., 2022)*

1.1 Our Contributions

The goal of this thesis is to improve the material classification performance using deep learning methods and real world datasets. The first contribution is we constructed a new material dataset consisting of 11 different classes such as aluminum foil, lettuce leaf, brown bread, wood, wool etc. The dataset is constructed in such a way that there remains large intra class variability across the images of the samples. We named the dataset as UJM-TIV where (UJM is the abbreviation of University Jean Monnet, where this thesis was conducted and TIV stands for Texture under varying Illumination, pose and Viewing). Figure 1.1 shows the images of samples of four different category from newly created UJM TIV dataset where each column contains the images of same sample under varying viewing and lighting conditions.With the use of this newly developed dataset, we demonstrate that traditional neural networks do not adapt well to datasets including real-world material data with significant intraclass variability. We believe that learning the material characteristics more effectively will be made possible by such a diversified dataset in the future.

We propose a Multiview solution for material classification in order to better generalize the deep learning features. In most of the existing material classification solutions, only one image is used to classify the material, however in this thesis, we demonstrate how classification accuracy could be improved significantly if we provide a collection of images for each material sample. Humans frequently try to modify their head position or manipulate objects in order to compensate for changing lighting conditions and viewpoints to determine the material of an object. In order to imitate this natural behavior, we suggest employing multiview learning to extract image features across a number of images and integrate those features into an appropriate representation. We believe that, this is the first instance of using a multi-view learning technique to for material classification in order to solve the problem of appearance variations caused by viewing conditions. Our contributions are the following:

- We demonstrate that there is insufficient intra-class variability for material classification tasks by analyzing the existing material dataset,
- We propose a new dataset with large variation across different acquisition settings (illumination and viewing) for better representation of various real-world material sample appearances. The first two contributions led to a publication in the journal of imaging in June 2022 (see (Sumon et al., 2022))
- Using a multi-view learning strategy, we propose combining features from several images of the same material sample into an accurate representation of the material,

- Extensive testing on KTH-TIPS2 and our new material datasets demonstrate that using several images of the identical sample outperforms the traditional single view solution by a significant margin.

1.2 Organization of the Thesis

The **second chapter** describes the related works from using handcrafted features based approach to deep learning approaches for material classification. **Chapter 3** introduces existing material datasets used for previous experiments. **Chapter 4** is devoted to the construction of proposed material dataset and methodologies applied in this thesis. Discussion and presentation of the results obtained through various experiments in this thesis is described in **chapter 5**. Finally, conclusion and future research directions are proposed in **chapter 6**.

2 Related Works

In this chapter we explore and present the most remarkable works in the context of material classification, starting from hand-crafted features to deep learning solutions.It will allow us to discover the limitations of the current state of the art approach for material classification. These remarks will be the starting points of our original solutions detailed in the upcoming chapters.

2.1 Handcrafted Features

The earliest work in 60's about material analysis reveals that material or texture can be perceived spontaneously if proximate pixels of uniform brightness form a specific connectivity (Julesz, 1962). To further explain human perception of material, Jules introduced texton theory (Julesz, 1981; Julesz and Bergen, 1983) in 80's. He argued that textures can be perceived if elementary local conspicuous features, called textons, are present, such as crossing, corners, etc. Additionally, he claimed that these texton's first-order statistics are the only ones that are meaningful. In other words, spontaneous perception cannot be triggered if the probability of every texton in one material region is equal.

To produce local features from the input material images Gabor filter (Bovik et al., 1990; Jain and Farrokhnia, 1991; Turner, 1986; Zhu, 2003), Gabor wavelets (Manjunath and Ma, 1996), Differences of Gaussians (Malik and Perona, 1990) serving as sliding wiondows are proposed as expert design filter banks for local conspicuous features extractions. A series of works (Wu et al., 2000; Xie et al., 2015; Zhu et al., 1998, 2000, 2005) based on texton theory tried to mathematically model textons and consequently Bags-of-textons (Leung and Malik, 2001) and Bags-of-words (Csurka et al., 2004) were proposed to accumulate features into a histogram representation over a given texton dictionary.By the end of the last century, researchers concentrate on the extraction of invariant feature representation. Some types of features are more robust than others to certain variations, such as background illumination or object size. The most notable invariant features include

Local Binary Pattern (LBP) (Ojala et al., 2002), Speeded Up Robust Feature (SURF) (Bay et al., 2006) and Scale Invariant Feature Transform (SIFT) (Lowe, 2004). Besides material classification, they dominated visual recognition field before the deep learning era.

2.2 Deep Learning Based Features

Image classification is one of the most fundamental research fields in the computer vision community, and its spur progress always influences greatly not only itself but also other visual recognition tasks, like video classification, image segmentation, medical image analysis. And it even has impact on other domains, such as natural language processing or brain-computer interface. Convolutional Neural Networks (CNN) represent a breakthrough in computer vision, since AlexNet (Krizhevsky et al., 2012) clearly outperformed the state-of-the-art in ImageNet ILSVRC competition (Russakovsky et al., 2015). This achievement is considered as one of the milestones both for deep learning and computer vision.

Indeed, it appeared that knowledge learned from a large image dataset for classification task can be helpful for other related task and dataset. This approach is known as transfer learning and has been widely used in the context of material classification. In (Wieschollek and Lensch, 2016) Wieschollek and Lensch used a network pretrained on Imagenet dataset to extract material features and trained a classifier with those features which outperforms alternatives based on handcrafted features with an evident margin. It clearly indicate that generic deep features are transferable to material classification. A typical convolutional neural network usually applies three operations to the input data. First, A number of convolutions by means of linear filters, second, introducing non-linearity by using activation functions such as sigmoid or Rectified linear unit, Third, pool the local features using different pooling methods (average pooling, Max pooling, etc.). All these operations are quite related to the filter banks used in (Randen and Husoy, 1999) for analyzing the textures materials.

As input goes through the network layer by layer, features extracted from different levels also contain complementary and rich information. For material classification both primitive and semantic information could be combined together in a more discriminative representation.Cerezo et al. show (Bello-Cerezo et al., 2019) how deep neural networks and pre-trained CNN-based features outperform conventional, hand-crafted descriptors, especially when dealing with textures like non-stationary spatial patterns and when the acquisition parameters have been changed repeatedly. The authors used 68 image descriptors both from hand-crafted (35 descriptors) and

CNN-based (33 descriptors) utilizing 23 different textures dataset over 10 different experiments. Cimpoi et al. not only proposed to pool local features of the last convolutional layer of VGG-19 to remove global spatial information, but also found that after combining these representation with penultimate Fully Connected layer's output, there is an obvious increase of the classification accuracy (Cimpoi et al., 2015). The authors explained that the FC layer can be considered as a pooling method which is able to capture the overall shape of the object present in the image. Napoletano (Napoletano, 2017) compared the CNN based features with hand crafted features where he considered five color texture datasets. The datasets contain images under varying lighting directions, viewing conditions and scales where the results achieved by the CNN based features outperforms the results from hand crafted features with a significant margin. Andrearczyk and Whelan designed a network architecture called Texture CNN (T-CNN) where features from different layers are respectively average pooled into a compact feature vector and then all the compact vectors are concatenated into a global one (Andrearczyk and Whelan, 2016).

Inspired by the findings in the work (Cimpoi et al., 2015), Xue et al. extended the Deep Texture module (Zhang et al., 2017) to the Deep Encoding Pooling Network (DEP) that feeds the output of the last convolutional layer of ResNet into two branches: Deep Texture module and global average pooling layer (Xue et al., 2018). The outputs from the two branches are then fused with a bilinear operation. Hu et al. encapsulated the two-branch structure of the work (Xue et al., 2018) into a Learnable Encoding Module (LEM) and plugged it to the end of basic blocks in the ResNet-50 in order to encode multi-level texture representations (Hu et al., 2019). In the bilinear pooling community, Dai et al. combined first-order features computed by average pooling and second order features computed by compact bilinear pooling with a simple concatenation (Dai et al., 2017). They also tried to fuse multi-level features to get a better performance. Herarchical Bilinear Pooling (Yu et al., 2018) runs bilinear pooling on local features across different layers and thus enhances bilinear representation by capturing inter-part feature relations.

In (Ghose et al., 2021) the authors explicitly modeled the extent-of-texture (EOT) and extent-of-shape (EOS) on a local group of feature vectors. According to the EOT (resp. EOS), feature vectors are split into two groups and are encoded separately into a global representation for each group. In the end, with the guide of EOT (resp. EOS), two global representations are combined and finally aggregated into an image-wise representation with bilinear pooling. Unlike these previous methods which concatenate pooled features from several layers, Zhai et al. propose to concatenate multi-layer feature maps (Zhai et al., 2019, 2020).

Then, in (Zhai et al., 2019), they applied a module with a cascade structure, in which global image representation at actual level should guide the next level representation. At the end, a fusion module is introduced to jointly exploit each level's global representation and to make strong classification prediction. In the work (Zhai et al., 2020), they designed a different encoding module that, first generates multiple texture primitives and then encodes texture primitives at one position by its correlations to other local neighbors. Note that, at the end, the output is an orderless pooled representation and it is finally integrated with spatial ordered information.

Most material classification techniques until recently relied solely on **single-view** images or incorporated a limited number of single view image attributes. Using close-range high-resolution texture imaging and near-infrared spectroscopy, for instance, the authors of (Erickson et al., 2020) performed material classification using a multi-modal sensing method. The authors in (Gorpas et al., 2013; Kampouris et al., 2016a) illustrated how the use of **photometric stereo acquisition** may boost the performance of material classification techniques. The authors demonstrated how a surface's microgeometry and reflectance characteristics may be utilized to determine its material.In a similar vein, Maximov et al. (Maximov et al., 2019) and Vrancken et al. (Vrancken et al., 2019) showed that integrating different illumination and viewing conditions might considerably enhance the material classification performance. One would like to be able to anticipate how a material would seem in the best-case scenario, regardless of the direction in which it is being viewed or any other elements that may affect the way it is being captured.This problem is difficult to solve in the general situation (Xu et al., 2019) because it is poorly formulated and has too few constraints.

2.3 Multi-view Learning

With the improvement of imaging technology and popularity of social media the amount of **multiview** data has been increased exponentially in recent decades in the field of social network (Fan et al., 2020), medical imaging (Wei et al., 2019; Xu et al., 2020), video surveillance (Guo et al., 2015; Feichtenhofer et al., 2016; Deepak et al., 2021) and entertainment industry (Srivastava and Salakhutdinov, 2012; Mao et al., 2014; Karpathy and Fei-Fei, 2015). Multiview data refers to the data that have same meaning but captured from different views, modalities and sources. Multiview learning aims at extracting precise characteristics from several data sources or modalities in order to aggregate the common features across various types of data. Because of the increasing amount of Multiview data it has gained the attraction of the researches in computer vision community in recent years. Convolutional neural networks (CNN) are capable of extracting very precise features from images, and several methods have included **multi-view learning** in the CNN (Yan et al., 2021; Feichtenhofer et al., 2016; Andrearczyk and Whelan, 2016; Dou et al., 2016). The objective of multiview CNN is to aggregate features from many views to get more generalize features from the objects. Two main multiview CNN based approaches exists namely **one-view-one-net mechanism** and **multi-view-one-net mechanism**. In one-view-one-net mechanism a dedicated CNN is used for each view to extract distinct features of that view and in the end all the features from different views are fused together using a fusion method (Feichtenhofer et al., 2016; Yang et al., 2018). In contrast, a single network is fed with all of the views in a multi-view-one-net technique in order to extract features (Dou et al., 2016). Su et al. proposed a multiview convolutional neural network for recognition of 3D shape where multiple 2D views are used to aggregate the multiview information (Su et al., 2015).

In one-view-one-net approach to reduce the amount of learnt weights, the network used to extract features often shares its weights. Feature fusion process used is critical in such approaches. The main challenge with a multi-view-one-net strategy is aggregating the input features before feeding it to the network. One way could be applying the convolution after these images are concatenated into a multi-channel image. By doing this, it is ensured that local features are pulled from these images at the same regions. In addition to that, to get consistent features, it requires a coarse registration between the images One disadvantage of such concatenation is that pre-trained networks cannot be used because usually those pre-trained networks are fed with 3 channels images. This is the rationale for our decision to choose a one-view-one-net strategy in this thesis over a multi-view-one-net method. Finally, in some research people used **Siamese network** where they used Multiview data for person re-identification (Varga and Szirányi, 2017). image quality assessments (Liang et al., 2016) for instance, Liang et al.(Liang et al., 2016) also suggest feeding a Siamese network using sub-patches taken from RGB images, whereas Varga et al. (Varga and Szirányi, 2017) suggest to extract a collection of overlapping sub-windows from a single image. Guo et al. proposed a dynamic Siamese network for tracking of moving object from video data with a rapid general transformation learning model that allows for efficient online learning of target appearance variation and background suppression from prior frames.(Guo et al., 2017). Whereas authors in (Zhang et al., 2018) proposed a siamese network based on local structure learning that takes into account both the target's local patterns and its structural relationships for more precise target tracking. By incorporating into the Siamese network in place of pairwise loss for training, an unique triplet loss is presented in (Dong and Shen, 2018) to extract meaningful deep features for object tracking. Khvedchenya et al. employed satellite imagery with siamese network for the purpose of assessing the damage of building before and after the disaster (Khvedchenya and Gabruseva, 2021). With experimental results the authors have shown that concatenation of output features from the branches of siamese network leads to better performance of the model. The authors in (Fatima et al., 2021) developed a system to detect the adulteration of papaya seed through the use of a siamese network. Moreover, in (Wang and Wang, 2019) siamese network is used for classification of plant leaves

The construction of a novel architecture for a multiview CNN is not the main goal of this thesis, even if the design of our network has been thoroughly explored. **The primary goal is instead to show how multi-view learning can be utilized to classify materials**. As far as we know, this is the first attempt for material classification a multi-view CNN has been utilized. It's also true for many other computer vision tasks, multi-view learning is a very recent research area.

3 | Material Datasets

To study and validate material recognition methods described in the previous chapter, material databases are always needed and we investigate them thoroughly in this chapter. In our opinion, they can be roughly grouped into three types and we consider their past, their present and their future. In the next three subsections, the three types will be presented according to this timeline.

3.1 Datasets under Controlled Environment

Before the deep learning era, when deep neural networks weren't used to perform large scale image classification, the first group of material datasets were created with the goal of characterizing the appearance of material instances. BRTF (Bidirectional Reflection Transmittance Function) and BTF (Bidirectional Texture Function) are widely used models to output parameterized visual appearance with lighting and viewing condition inputs. In order to build BRTF/BTF models for real-world material instances, images in these kinds of datasets were collected under controlled conditions in labs, and the parameters of these conditions were provided.Because these datasets focus on the study of material instances, for one instance, images with different visual appearance need to be extremely collected. Hence, the resulting BRTF/BTF model is able to perfectly describe this material instance and enables to produce synthesized images. On the other hand, in each category, the number of instances is rather limited and instances were carefully chosen by the dataset creators. In one words, this type of datasets can be well exploited to build instance-level features, that are invariant to different conditions, but these features may be less transferable to other instances of the same category which are not included in the dataset.In the next subsequent section we will describe the representative datasets.

3.1.1 Columbia-Utrecht Reflectance and Texture (CUReT) Dataset

Researchers at at Columbia University and Utrecht University created this dataset (Dana et al., 1999) which includes 61 material samples from 10 different types of surface materials ranges from natural surface to man-made surfaces. These samples are taken under **205 different lighting and viewing conditions**. Figure 3.1 shows some example images from CUReT database. Only one image per class is provided in this dataset which leads to lack of intra-class variability across the categories of the dataset.

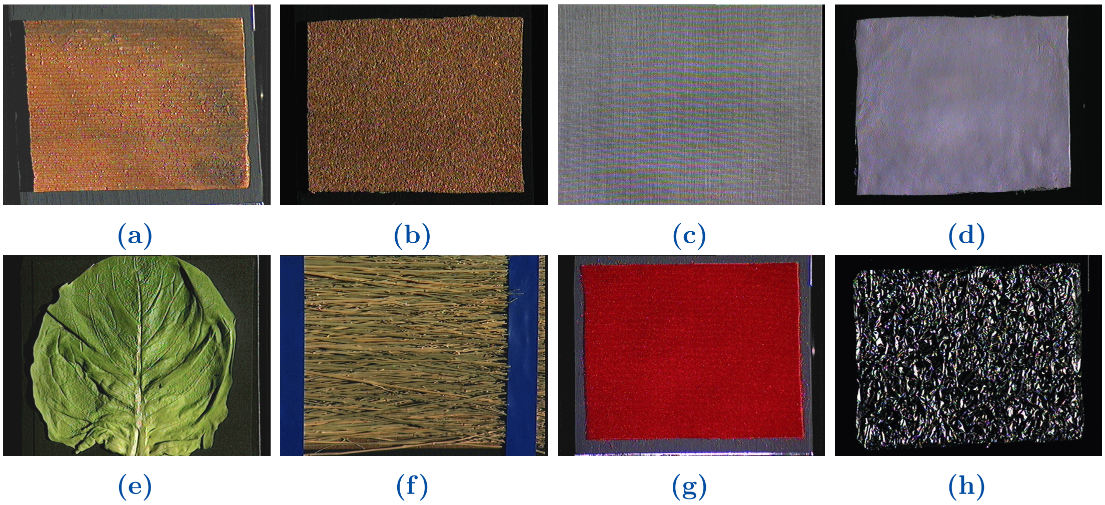


Figure 3.1: *Example images of a sample of h wood, b cork, c cotton, d leather, e lettuce, f straw, g velvet and a aluminium foil from CUReT dataset.*

3.1.2 KTH-TIPS2 Dataset

KTH-TIPS2 dataset (Mallikarjuna et al., 2006) is an extension of KTH-TIPS dataset where all the categories of KTH-TIPS2 dataset is also included into CUReT database.The images of the samples are captured under **3 viewing direction and 4 illumination conditions**. It contains images of 11 material categories with 9 different scales for each image.For each sample there are 108 images are captured under varying illumination, pose and scales. Images of a sample of aluminium foil, brown bread,corduroy, cotton, lettuce leaf, cork, cracker and wool is shown in figure 3.2. Four image samples for each category is provided in this dataset. Because fewer samples are offered, the intra-class variety of materials encountered in real-world situations cannot be as accurately represented. We also found that the variation in the viewing and illumination direction is not significant for some classes (such as cork, brown bread and white bread) that's why the images from this classes suffer from lower intra-class variability.

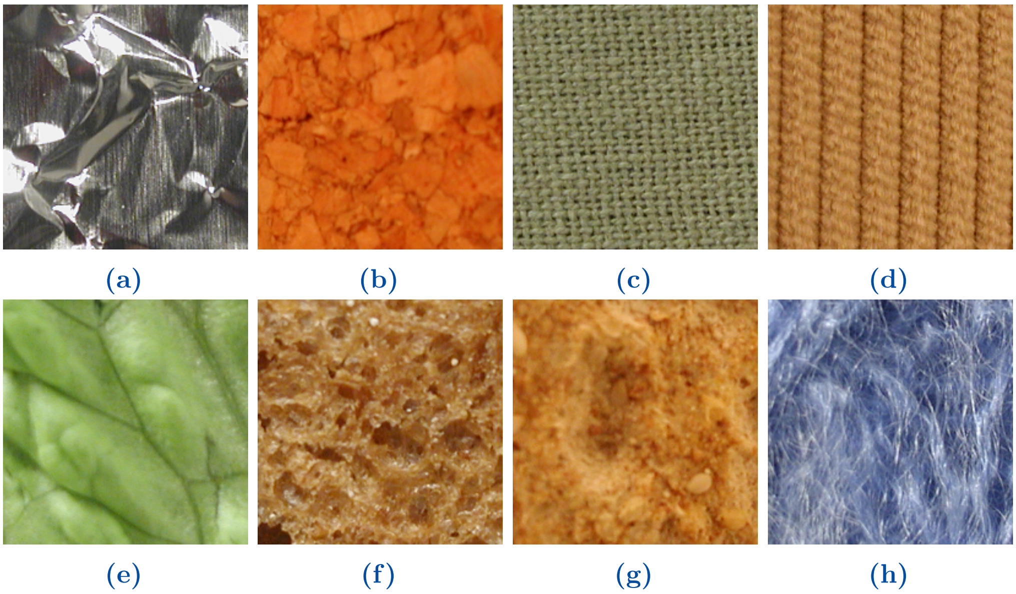


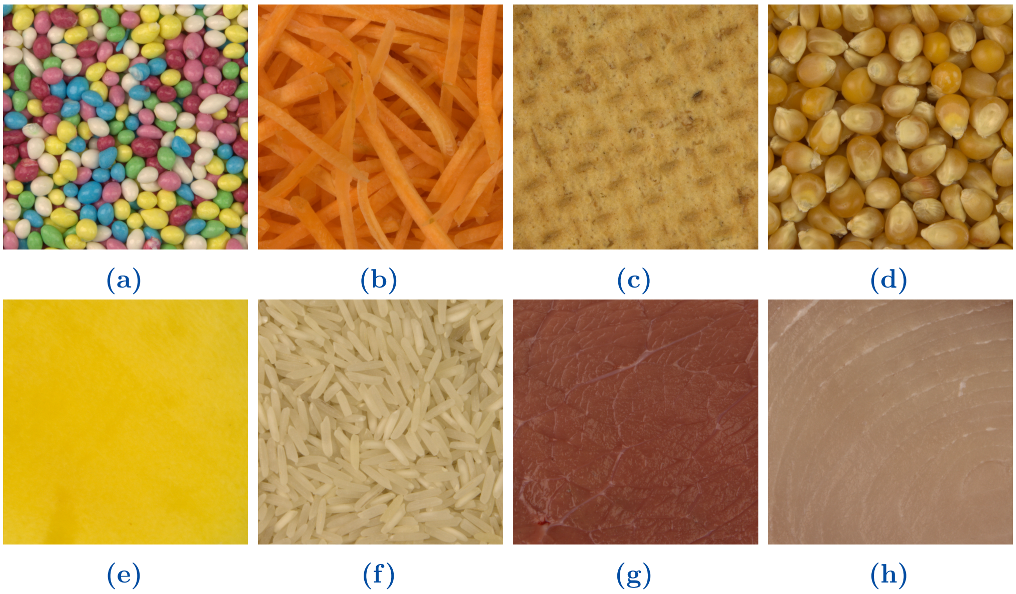
Figure 3.2: *Example images of a aluminium foil, b cork, c cotton, d corduroy, e lettuce leaf, f brown bread, g cracker and h wool from KTH-TIP2 dataset.*

3.1.3 RawFooT Database

RawFooT (Cusano et al., 2016) dataset is constructed by the researchers of imaging and vision laboratory from University of Milan-Bicocca which contains images of 68 raw food samples. Images of these food samples are captured with **46 different illumination conditions**. The authors simulate natural daylight(D65, D45,..,D95) and artificial lights(L27,L30,..,L65) under different color temperatures and captured images under those illuminant. Some example images from the dataset is shown in figure 3.3 The category includes a whole ranges of food sample from meat to rice, cookies, fruits etc.

3.1.4 UBO 2014 Dataset

A larger dataset (Michael et al., 2014) taken under controlled conditions , which consists of seven different material categories (carpet, fabric, felt, leather, stone, wallpaper, and wood), each one has 12 material samples that may be used to



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

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

Figure 3.3: *Example images of a sample of a candy, b carrot, c cookie, d corn, e mango, f basmati rice, g steak and h swordfish from RawFooT dataset.*

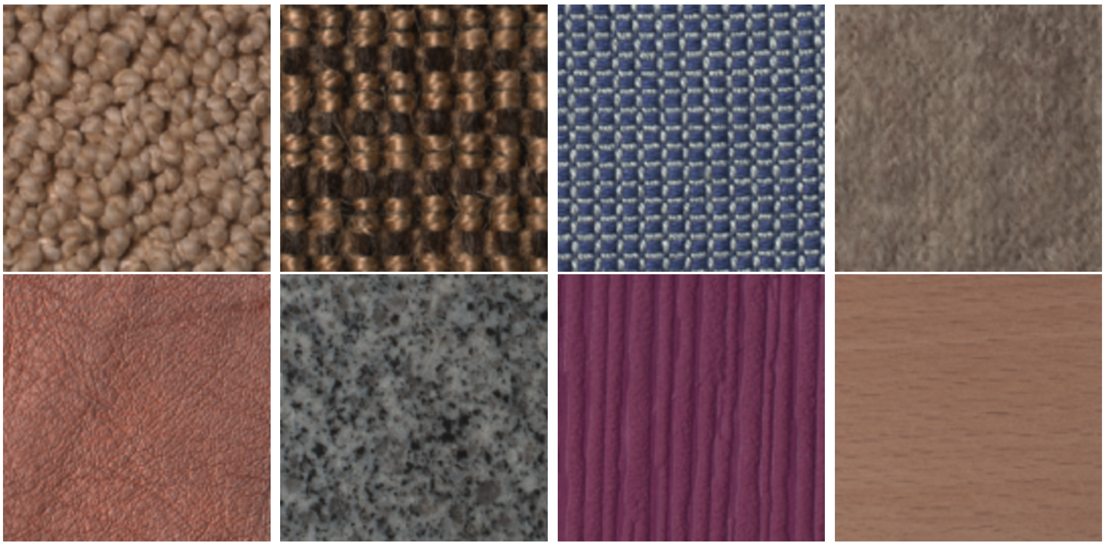


Figure 3.4: *Images of different samples from UBO2014 dataset*

illustrate the associated intra-class variations. The dataset contains total 84 different samples. Figure 3.4 shows some sample images from the dataset. The dataset contains 22,801 images taken under **151 illumination directions and 151 viewing conditions** using a bi-directional sampling method which combines series of images captured under different lighting and viewing conditions.

3.2 Real-world Datasets

In contrast to datasets taken under controlled conditions, in real-world datasets, images are captured in the natural world, which results in a wider range of visual looks. It is based on unobserved material details, sunlight, random pose, etc. Besides, images no longer necessarily show only the material but context information surrounding the target. Based on the fact that material images are collected from multiple online sources and images are **taken under random conditions**, it becomes possible to exploit invariant features of material categories. With these datasets, the majority of current research that uses deep learning networks to retrieve invariant characteristics may produce good classification results.

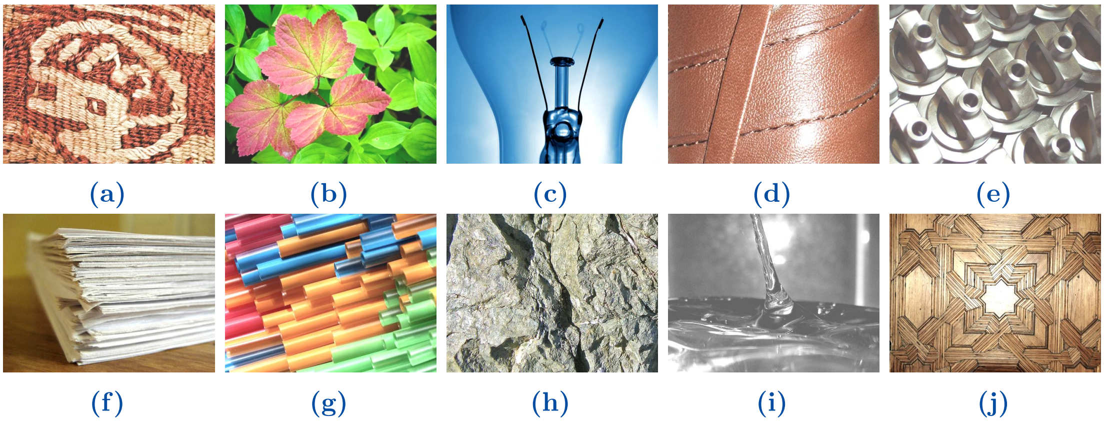


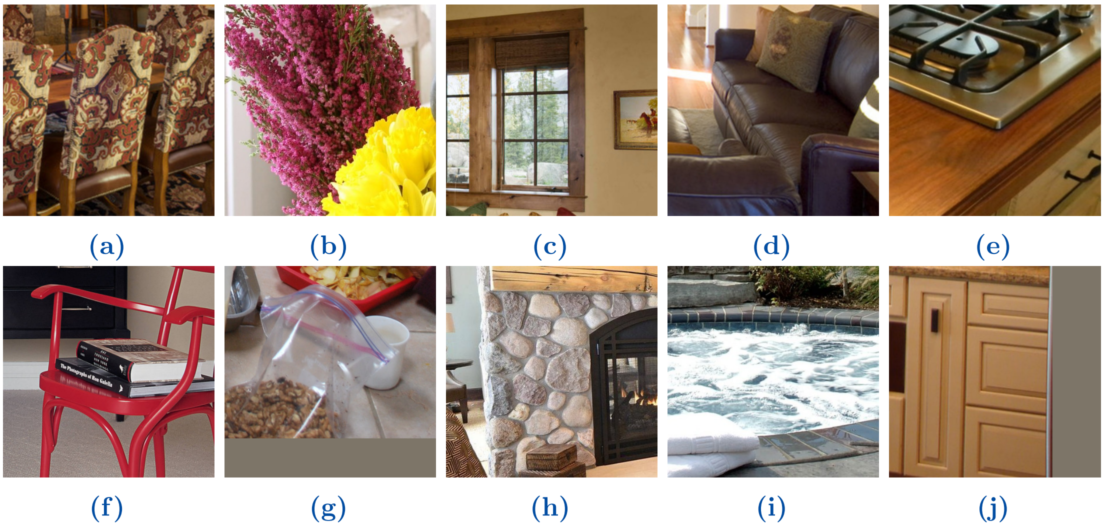
Figure 3.5: *Representative images of a sample of a fabric, b foliage, c glass, d leather, e metal, f paper, g plasctic, h stone, i water, and j wood from Flickr material database.*

3.2.1 Flickr Material Database (FMD)

Flickr Material Dataset (FMD) (Sharan et al., 2014) is a small but popular real-world material dataset, containing 10 categories. Each category contains 100 images. Representative images of FMD database is shown in figure 3.5. Images were **downloaded from flickr.com** and they were carefully chosen to cover a wide range of visual appearance in one category. Masks, locating material region, are also provided for every image. They are helpful for studies where masking out clutter background is needed (e.g. when the influence of background context impacts the classification performance).

3.2.2 MINC 2500 (Materials in Context Database)

MINC 2500 (Bell et al., 2015) is a subset of MINC database. Its large size makes it very suitable for training a deep CNN material classification task. Image patches were manually cropped from the material samples. Abundant background context appearing around target material makes this dataset quite challenging, see Fig 3.6. It contains 23 commonly-seen material categories and 2500 images per category.



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

(f) (g) (h) (i) (j)

Figure 3.6: *Representative images of a sample of a fabric, b foliage, c glass, d leather, e metal, f paper, g plasctic, h stone, i water, and j wood from MINC 2500 database.*

3.2.3 Ground Terrain in Outdoor Scenes (GTOS)

This dataset contains images of ground terrain captured under **19 viewing directions** and varying weather and lighting conditions(different time of the day) (Xue et al., 2017).It can be used to study of the ground terrain recognition, which can be implemented into autonomous driving systems to detect current ground terrain's condition. This dataset is challenging because some inter-class boundaries

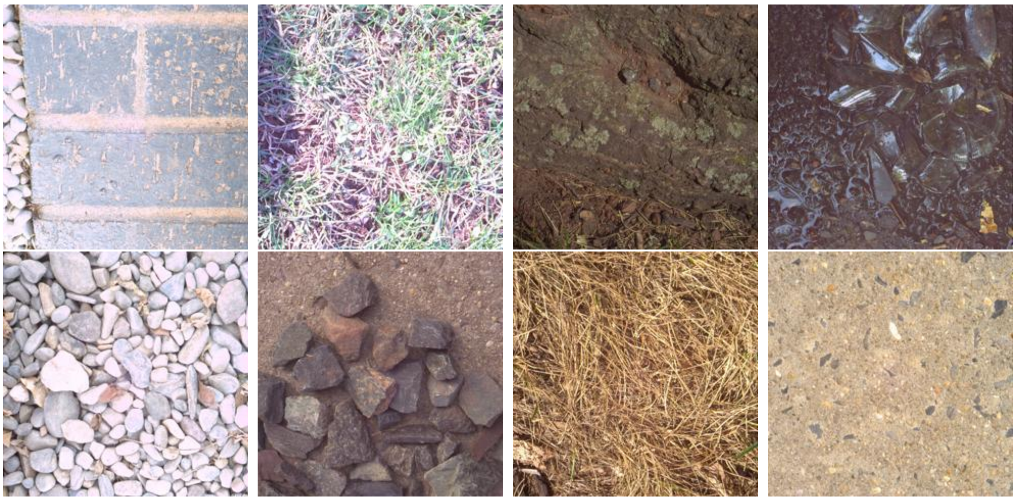


Figure 3.7: *Images of different samples from GTOS dataset*

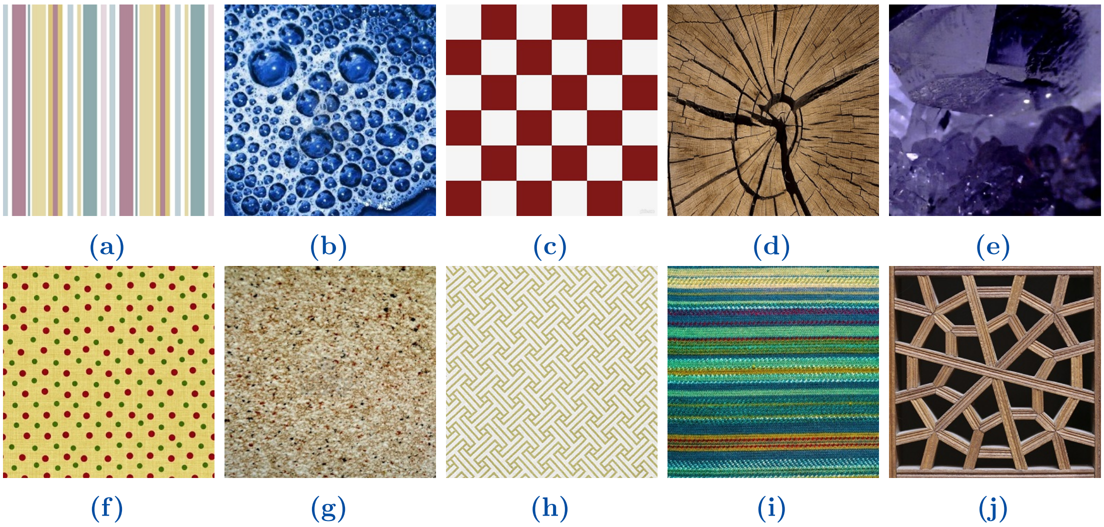
are ambiguous. Some images from GTOS dataset is shown in 3.7. The dataset comprises of 30,000 pictures from 40 classes of outdoor scenes.

3.2.4 4D Light Dataset

4D-Light dataset (Wang et al., 2016) consists of 12 categories where each category contain 100 images. It is the first medium-size dataset for light-field images. In addition to 2D projections of the scene, light field images additionally includes the angle of the light beams that reach the projection as a third dimension. A special camera known as plennoptic camera is used to capture the amount of light present in a scene as well as the precise direction in which the light rays are moving across space. Different from RGB images, light-field images are taken with plenoptic camera, which not only captures light intensity and color in a scene, as a conventional camera does, but also records **light directions with multi-view points**. Light-field images can be seen as an alternative way to determine materials when it is difficult to determine a material with its surface reflectance or BTF. As in our study case we limited our investigations to RGB input images, light-field information was not investigated in our experiments.

3.2.5 Describable Textures Dataset (DTD)

Describable Textures Dataset (Cimpoi et al., 2014) is constructed by **collection of images from Google and Flickr**. It consists of 47 categories where each category contains 120 images.Instead of defining categories by material name, like wood,



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

(f) (g) (h) (i) (j)

Figure 3.8: *Representative images of a sample of a banded, b bubbly, c chequered, d cracked, e crystalline, f dotted, g flecked, h grid, i lined, and j meshed category from DTD database.*

water, a collection of images having the same texture attributes (e.g: dotted) is viewed as a category inspired by the human perception. See fig. 3.8 for illustration. The dataset contains a total of 5,640 images, most of them have limited surrounding background.

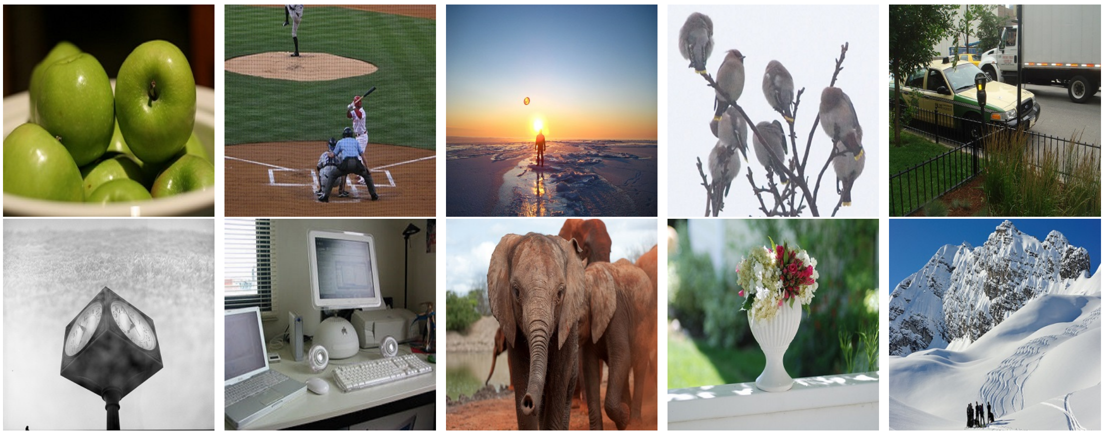


Figure 3.9: *Example images from COCO dataset*

3.2.6 Common Objects in Context (COCO) dataset

A large sacale dataset (Lin et al., 2014)which contains more than 200 thousands labeled images composed of 80 objects categories and 91 stuff categories. Figure 3.9 shows some example images from COCO dataset. It has a variety of material categories, many of which are subdivided into more precise classes, such as water, whose examples exist in words like "see" and "river," etc. The dataset is partioned into train (118K image), test (41K images) and validation(5K images) along with annotations for 91 stuff classes.

3.3 Synthesized dataset

In computer graphics, material or texture rendering is the process of generating a photorealistic image of an object. Since there are not plenty of material datasets available, one way could be to generate synthesized images of the material objects and use those synthesized databases for the experiment. An advantage of the synthesized dataset is one can generate as many images as one wants very quickly and without setting up any physical arrangement for the acquisition of the images. It is an alternative way to enrich the existing dataset.

Targhi et al. used photometric stereo to rendered synthesized images (Targhi et al., 2008) and showed that by including these synthesized images in the existing

Table 3.1: *Summary of different material datasets.*

Dataset	**Multi-lighting condition**	**Multi-viewing conditions**	**Multi-pose**	**Acquisition environment**
CUReT	205		-	controlled
KTH-TIPS2	4	3	3	controlled
RawFooT	46	-	-	controlled
UBO2014	151	151	-	controlled
UJM-TIV	4	4	2	controlled
FMD	-	-	-	uncontrolled
MINC 2500	-	-	-	uncontrolled
GTOS	-	19	-	uncontrolled
4D Light	-	-	-	uncontrolled
DTD	-	-	-	uncontrolled
COCO	-	-	-	uncontrolled

training data can improve the classification accuracy. In (Weinmann et al., 2014) Weinmann et al. used synthesized images to conducted some experiments . A classifier was trained and applied to a test dataset containing also real world images. Thanks to its large scale, a synthesized training dataset can achieve a comparable performance to a small dataset containing real-world images only. Combining these two dataset together can consistently boost the classification accuracy. Synthesized images can be considered as a good complementary training data if the size of the real-world images training dataset is too small.

As we discussed in subsection 3.1.4 UBO 2014 dataset also contains BTF measurements for all the 84 samples (7 category x 12 samples per category). Combined with 30 environment lighting maps (**5 directions x 6 natural lights**) from (Debevec, 2008), a virtual camera can take 1260 synthesized images while considering **42 viewing points**. Therefore, the total number of generated images is 105,840 (7 category x 12 samples per category x 1260 images per sample).

This chapter was devoted to the different material datasets, which we have proposed to classify according to their characteristics that meet different demands in each epoch of material classification's history. Table 3.1 summarises the different acquisition settings used in the datasets discussed in this chapter as well as acquisition settings of our newly created UJM-TIV dataset. Because some datasets (for instance MINC-2500, FMD, DTD, COCO) were constructed by gathering images from internet source e.g. google search, Flickr.com acquisition conditions are not available for these kind of datasets. We propose a new material dataset named UJM TIV where images are taken under controlled conditions. Then, we decided to use the dataset for experiments, considering its suitability for CNN-based approaches.

4 Methodology

This chapter is devoted to the materials and methods used in this thesis. In the first section of the chapter we describe about the newly created UJM TIV dataset, the acquisition setup, as well as comparison with other existing datasets. In the next section we will discuss about the use of siamese network for material classification.

4.1 UJM TIV: Our newly created material dataset

The newly created UJM-TIV material dataset consists of 11 different classes: images of aluminium foil, brown bread, corduroy, cork, cotton, lettuce leaf, linen, white bread, wood, crackers and wool illustrated in figure 4.1. **Controlled viewing and illumination conditions** were used to acquire these images. KTH-TIPS2 (Mallikarjuna et al., 2006) dataset also includes each of these 11 categories of images. In KTH-TIPS2 dataset for each category four physical samples are provided where images were taken under varying pose, illumination and viewing conditions. In some classes (such as wool and cracker samples), intra-class variation is stronger from one physical sample to another. On the other hand, in some other classes, the variation within the class is small (for example, in a wood or cork sample). Because of small inter-class variance some similarities is also observed between linen and cotton classes. Changes in visual appearances in KTH-TIPS2 dataset due to changes in illumination and viewing directions are illustrated in figure 4.2 and 4.4. Given the variety of acquisition conditions, the samples in the UJM-TIV dataset do not have the same visual look as the KTH-TIPS2 samples. At lower viewing angles or lower lighting angles, significant visual variations from figure 4.1 are visible in figure 4.3.

Most of the material datasets have well-controlled lighting, viewing, and camera settings. A technician (photographer) performs image capture, paying attention to take the best images while attempting to reduce blur and specularity by utilizing the setup system at their disposal.It is very difficult to capture images without

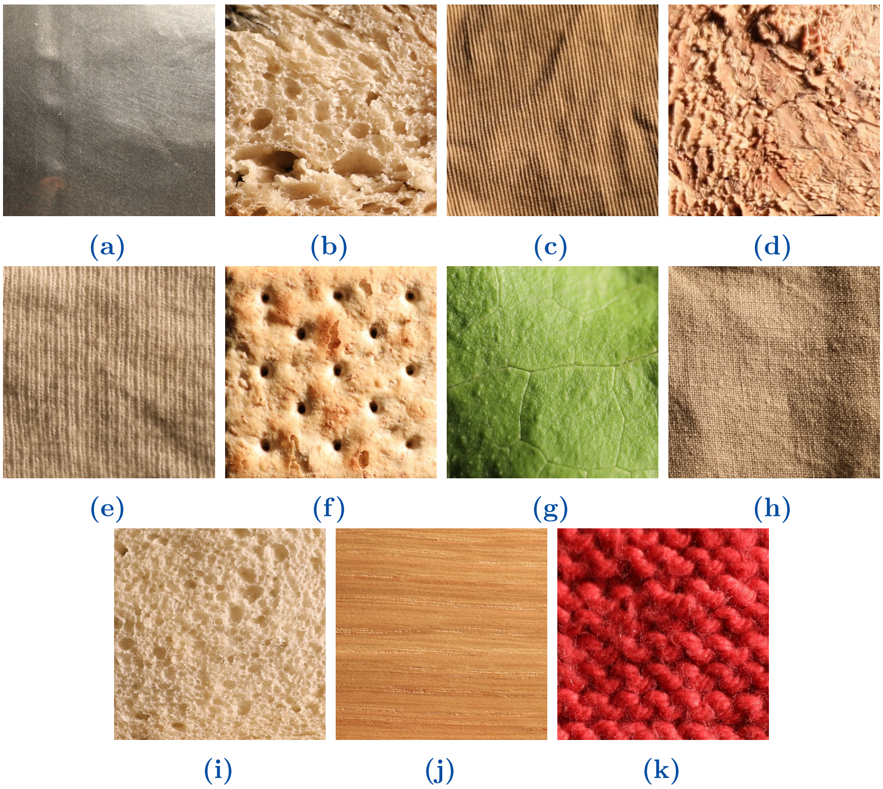


Figure 4.1: *Image samples from UJM TIV dataset observed under viewing direction 90°and lighting condition 65°. Sample a aluminium foil, b brown bread, c corduroy, d cork, e cotton, f cracker, g lettuce leaf, h linen, i white bread, j, k wool respectively. Images from (Sumon et al., 2022).*

specularity for some reflective materials such as aluminum foil. Therefore, our objective was to create a new dataset that would enable users to obtain images in a flexible ways, instead of relatively strict and carefully monitored viewing and lighting conditions our primary aim was to accomplish image capturing under a variety of illuminations and viewing directions. Depending on the direction of viewing we presume that , the sample's average brightness may change as shown in figure 4.5f in comparison with 4.5h. One of the invariant features that a material classifier should have is lightness/color invariance. We also suppose that depending on the material's thickness and degree of roughness, the sample's contrast can change depending on the viewing angle as shown in figure 4.5a in comparison with 4.5e. One of the invariant features that material classifiers also require is contrast invariance.

The fabric dataset presented in (Kampouris et al., 2016b) shows different types of shift in illumination due to a spatially non-uniform lighting field (consist of 12

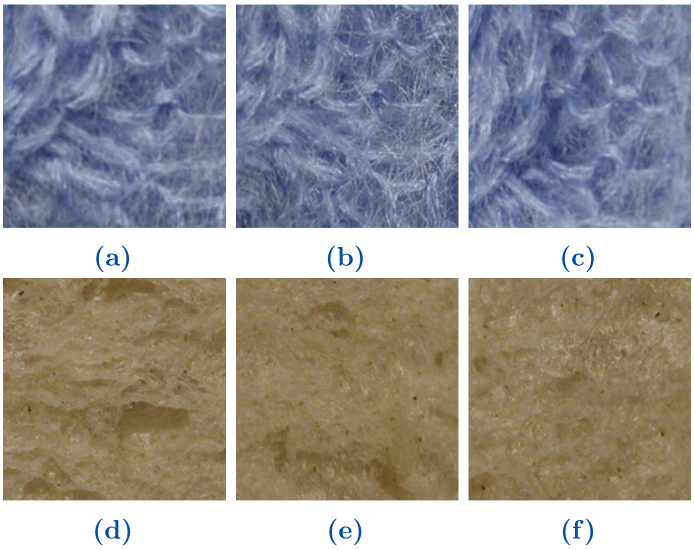


Figure 4.2: *Changes in the appearance of a wool and white bread sample from the KTH-TIPS2 dataset observed with different illumination and viewing conditions.Images a to c were taken with frontal lighting condition and frontal, 22.5° right and 22.5° left viewing conditions respectively for a wool sample. Likewise, for a white bread sample mages d to f were taken with frontal lighting condition and frontal, 22.5° right and 22.5° left viewing conditions, respectively. Images from (Sumon et al., 2022).*

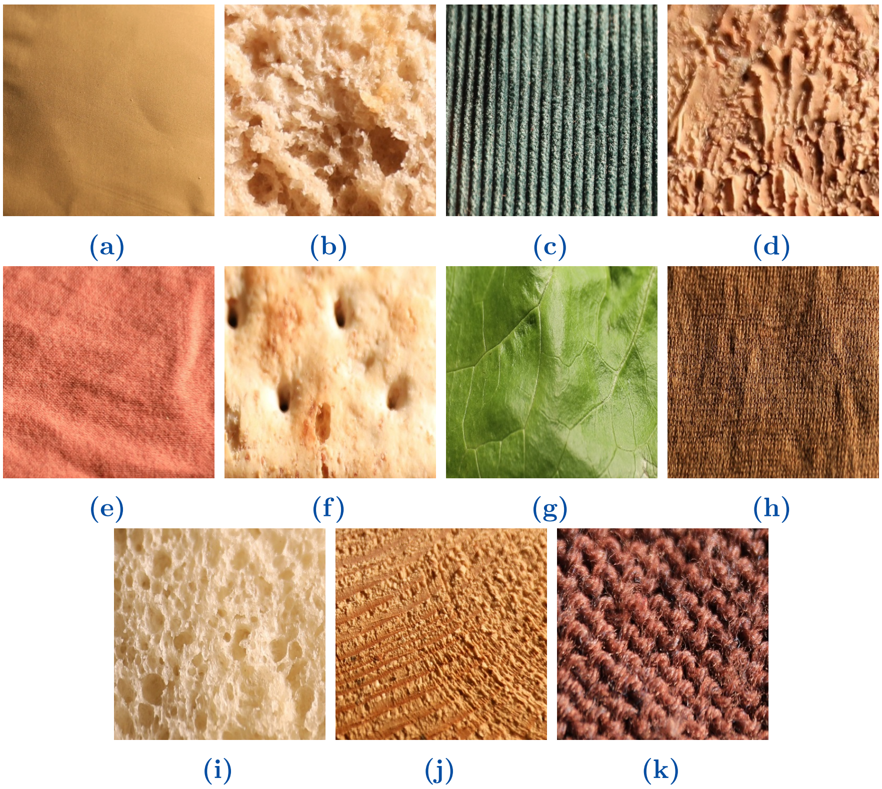
LEDs forming an array) on the sample surface. The 1266 samples in this dataset are made up of different fabric class such as cotton, denim, corduroy, wool, silk ,polyster terrycloth etc. The amount of samples for each class is very imbalanced for example there are only 32 samples in terrycloth class whereas 588 samples are available in cotton class. Samples were taken only under grazing light from the front. Photometric reconstruction was carried out using a geometrically calibrated system. We can increase the variation in a material sample's visual appearance by experimenting with illumination and viewing directions. In this research, we argue that to maximize classifier performance, **the final feature vector should take into consideration the variety of visual appearances of a material sample across different acquisition settings**.Because larger viewing and illumination angles were taken into account in UJM TIV dataset as compared to KTH-TIPS2 dataset, for instance, the differences between the images seen in Figure 4.6 are more substantial than those seen in Figure 4.7. The UJM-TIV dataset's sample diversity is noticeably greater than KTH-for TIPS2 for several categories (such as wood and wool). Additionally, in UJM-TIV, white bread, cork, and brown bread show the least intra-class variances whereas cotton and wool have the largest appearance changes . See, for instance, the visual differences in figures 1.1 and 4.8.

Table 4.1: *As shown in figure 4.7 viewing and lighting directions of all the selected views from KTH-TIPS2 dataset. Table from (Sumon et al., 2022).*

View	Viewing condition	Illumination condition
View1	Frontal	Frontal
View2	22.5° left	Ambient
View3	Frontal	45° from top
View4	22.5° right	Ambient
View5	Frontal	45° from side
View6	Frontal	Ambient
View7	22.5° right	Frontal
View8	22.5° left	45° from side
View9	22.5° right	45° from top
View10	22.5° left	45° from top
view11	22.5° right	45° from side
view12	22.5° left	Frontal

Table 4.2: *As shown in Figure 4.6, Viewing and illumination directions for selected views from UJM-TIV dataset. Table from (Sumon et al., 2022).*

View	Viewing condition	Illumination condition
View1	90°	90°
View2	90°	45°
View3	90°	20°
View4	60°	65°
View5	60°	20°
View6	30°	90°
View7	90°	65°
View8	60°	45°
View9	60°	90°
View10	30°	20°
View11	30°	45°
View12	30°	65°
View13	10°	90°
View14	10°	20°
View15	10°	45°
View16	10°	65°



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

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

(i) (j) (k)

Figure 4.3: *Image samples from UJM TIV dataset observed under viewing direction 35°and lighting condition 65°. Sample a aluminium foil, b brown bread, c corduroy, d cork, e cotton, f cracker, g lettuce leaf, h linen, i white bread, j wood, k wool respectively. Images from (Sumon et al., 2022).*

4.1.1 Configurations for Acquisition and Image Processing

A Canon EOS 5D Mark IV digital camera was used to acquire the images for UJM-TIV dataset. The camera resolution is 6720x4480 pixels. Most of the images contain background along with the sample, which were removed in the post processing step. **Two object poses** were taken into account for each object sample, with each object pose being rotated by 90 degrees along the surface normal N of the angle shown θ_S in fig 4.9. Figure 4.10's example demonstrates how such a modification might affect the material's looks for a particular material sample. The acquisition settings utilized to acquire the photos under controlled lighting and viewing conditions are shown in Figure 4.9. In the illustration, the sample is denoted by S, light source by I, and viewing direction by V. The vectors N and V define a plane that is perpendicular to the plane formed by the vectors surface normal N and light

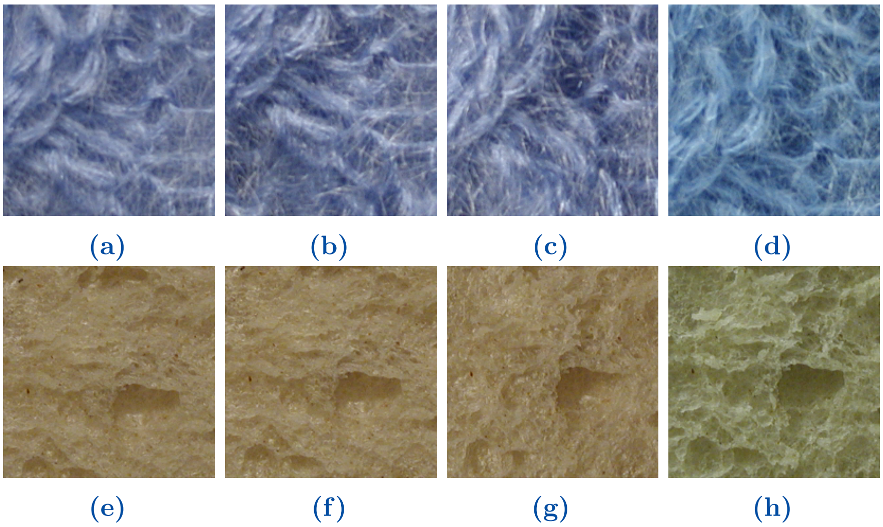


Figure 4.4: *Changes in the appearance of a wool and white bread sample from KTH-TIPS2 dataset observed with different illumination and viewing conditions.Images a to d were taken with frontal viewing condition and frontal, 45° from top, 45° from side and ambient illumination directions, respectively for a wool sample. Likewise, for a white bread sample images e to h were taken with frontal viewing condition and frontal, 45° from top, 45° from side and ambient illumination directions, respectively. Images from (Sumon et al., 2022).*

source I. Each of the **4 illumination directions** θ_I (frontal, approximately 20°, approximately 45°, and approximately 65°) was illuminated by a single conventional light source (a tungsten light bulb of 60 W). For each object orientation, **four viewing angles** (frontal, approximately 60°, approximately 30°, and approximately 10°) were employed. As a result, for each material sample total 16 images were obtained (4 lighting conditions x 4 viewing conditions). Total 32 images were taken for each sample in two positions. Image acquisition was carried out in a room that with no ambient light. 200 x 200 pixel image patches were extracted from the samples using the Patchify (Weiyuan Wu, 2020) library. All extracted patches had backgrounds and portions of the images that were too blurry removed manually. From one sample to another, different numbers of patches were extracted. Following the removal of all blurry and out-of-focus regions from the collected patches, the **dataset comprises about 75,000 image patches**.

4.1.2 Comparison with KTH-TIPS2 dataset

Compared to KTH-TIPS2, which employed frontal, rotated 22.5°left and 22.5°right viewing conditions, UJM-TIV uses viewing directions that are more broad and

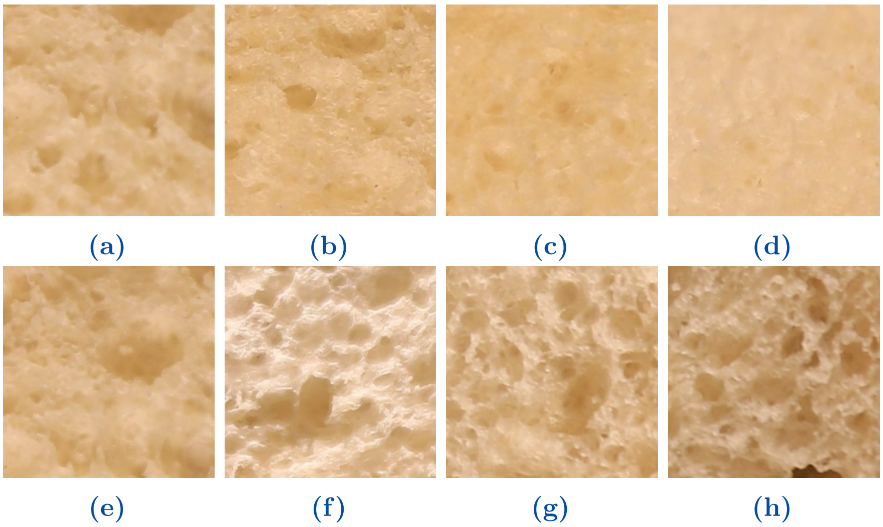


Figure 4.5: *Changes in the appearance of a sample of white bread from UJM TIV dataset observed under various illumination and viewing conditions. The illumination direction is same (90°) for images a to d where viewing conditions are 90°, 60°, 35°, and 10°, respectively. Viewing condition is fixed at 90°for images e to h where illumination conditions are 90°, 65°, 45°, and 20°, respectively. Images from (Sumon et al., 2022).*

have a wider range.Additionally, UJM-TIV's illumination directions are differ from KTH-TIPS2 .Unlike the UJM-TIV dataset, the KTH-TIPS2 dataset's samples were all obtained with a combination of 4 lighting conditions and 3 viewing conditions. Moreover, in KTH-TIPS2 dataset images were also acquired with 9 different scales, on the other hand for UJM-TIV scaling was considered while capturing the images.

Similar to KTH-TIPS2, UJM-TIV suffers from perspective effects and material roughness, which causes a small number of images of fine structured materials to seem blurry at working distances.As illustrated in figure 4.11 where all the images were taken at a viewing angle of roughly 10°and 20°illumination angle. Unlike those setting described in (Kampouris et al., 2016b) and (Kapeller et al., 2020) our objective wasn't to develop an optimal illumination system that modifies the light source settings in accordance with specific materials.

4.2 Multiview learning with Siamese network

Since multiview learning enables the extraction of features from several views and their fusion into appropriate global representations, multi-view learning is

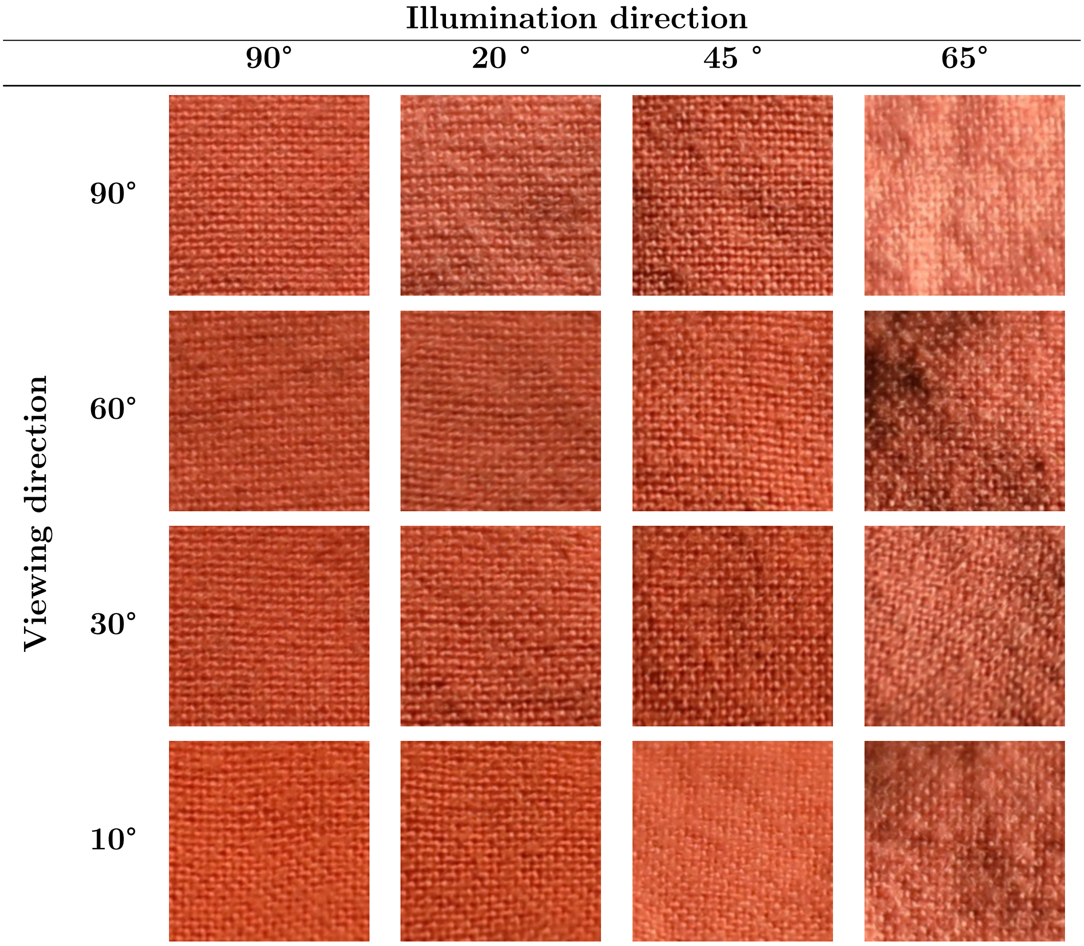


Figure 4.6: *A cotton sample's images from the UJM TIV dataset used for different views. Images from (Sumon et al., 2022).*

attracting the attention of many researchers nowadays (Xiaoqiang et al., 2021).A one-view-one-net technique is more suited for classification of material, as we stated previously. Here,images from each view are used in this case to feed a deep network that serves as the branch for feature extraction. The aggregated features from each view are then utilized as inputs to a classification model to predict the class of the sample under consideration. Once more, **our intention here is not to define the optimum architecture for the job, but rather take advantages of multiview learning to demonstrate how it may considerably enhance material classification performance**. As a consequence, we picked a straightforward one-view-one-net design employing a pre-trained neural network, allowing any improvements related to the architectural selection for future studies. Sharing the parameters between the branches that each view receives would help to

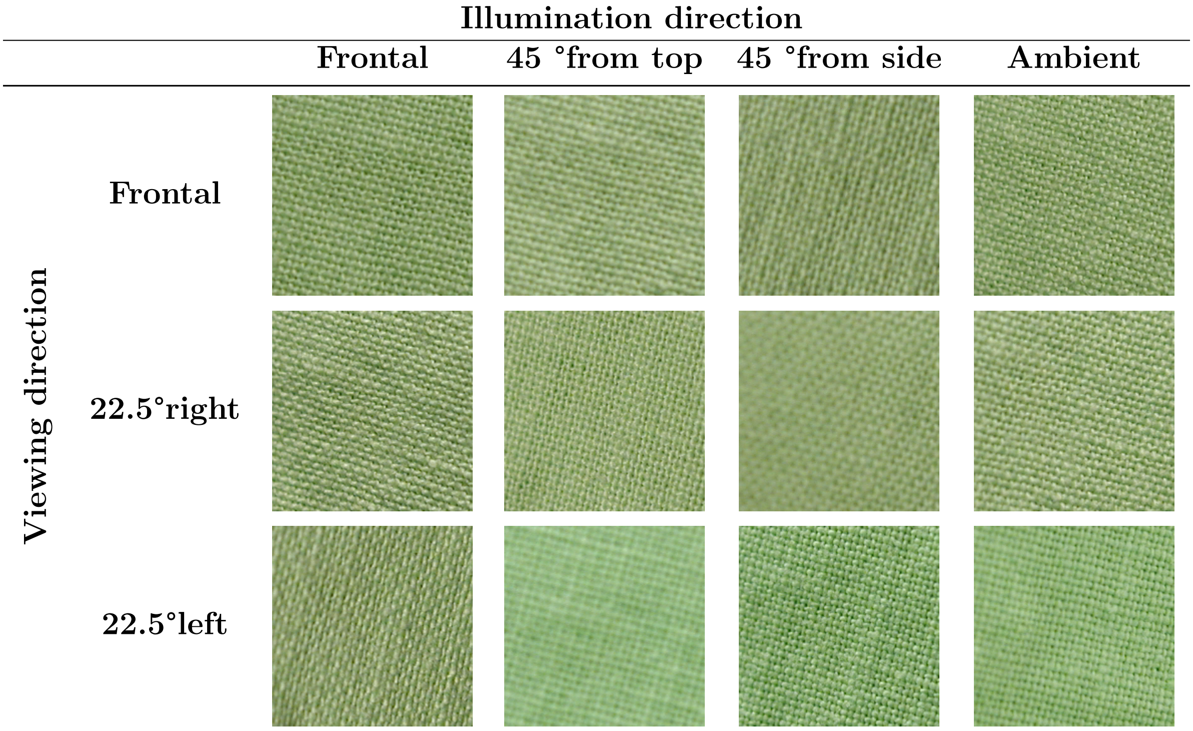


Figure 4.7: *A cotton sample's images from the KTH-TIPS2 dataset used for different views. Images from (Sumon et al., 2022).*

reduce the set of learned weights and thus help prevent overfitting. Because each branches must extract precise information from a variety of views that represent images of varying looks, sharing the parameters between the branches could also assist to boost the model's generalization ability. As we want the backbone to share weights, siamese network is an appropriate choice for us. A siamese network is a kind of neural network composed of two or more branches where each branch is basically share the same network architecture, parameters and weights between themselves (Koch et al., 2015; Wiggers et al., 2019; Melekhov et al., 2016). In other words these two branches are not different network instead two copies of same network. Each branch of the network is a convolutional neural network. Lets say two input images x1 and x2 denotes two different views passed through the branches of siamese network, each branch generates a fixed length feature vectors for each input view. These feature vectors are then compared to each other to determine whether the images belongs to same class or not assuming the model is trained properly.The similarity is computed by using the difference between the feature vectors. If two images belongs to the same class the feature vectors difference will be small, on the other hand if they belong to different classes the

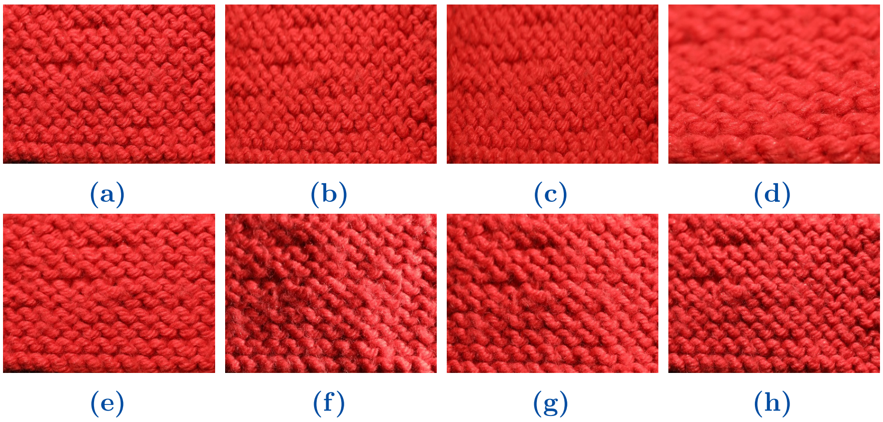


Figure 4.8: *Differences in a wool sample's visual appearance caused by varied lighting conditions and viewing angles. The same lighting angle (90°) was used to capture the images (a – d). Images (e - h) were taken at a 90°angle from the viewing direction. For images (a - d), the viewing angles are 90°, 60°, 35°, and 10°, respectively, with the illumination direction set at 90°. The viewing angle is set at 90 degrees for images from (e to h), while the illumination angles are 90 degrees, 65 degrees, 45 degrees, and 20 degrees, respectively. Images from (Sumon et al., 2022).*

difference will be large. Figure 4.12 depicts the proposed network's architecture. The branches of the Siamese network receives a couple of images from two different viewpoints. We used a ResNet50 pretrained on imagenet dataset as the backbone of the network. The architecture details of resnet50 network is illustrated in figure 4.13. Resnet is the winner of ILSVRC in 2015 classification competition.Resnet introduced residual block. The first difference we notice with traditional CNN is that there is a direct connection, skipping certain levels (which may change in different models), in between. The center of residual blocks is a connection known as the skip connection. The output of the layer has changed as a result of this skip connection. The input matrix is multiplied by the layer weights and then a bias is added if skip connection is not used. The idea of skip connection is to skip a number of layers and connect the output of layer directly to its input.There are 4 stages in the ResNet-50 network as shown in diagram 4.13, each having a convolution and an identity block. Each of the identity block and convolution block contains three convolutional layers. The number of trainable parameters of resnet-50 is around 23 million. As mentioned before preatrined resnet50 is used as a branch of the proposed siamese network. Every branch of the siamese network learns the distinct features from the images of distinct views passed to that branch as input. The fully connected layer is then provided with the combined features from both views for classification. In fact this design can be learned end-to-end since all of

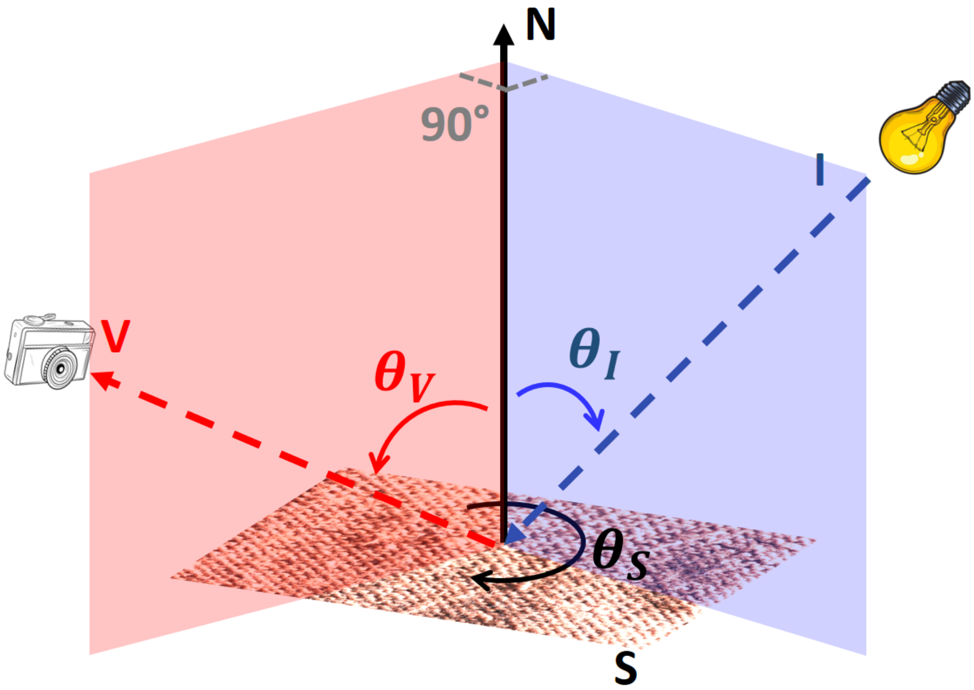


Figure 4.9: *The configuration for image acquisition is shown schematically. For the sake of our studies, the plane bounded by vectors **N** and **I** was positioned perpendicular to the plane bounded by vectors **N** and **V**.*

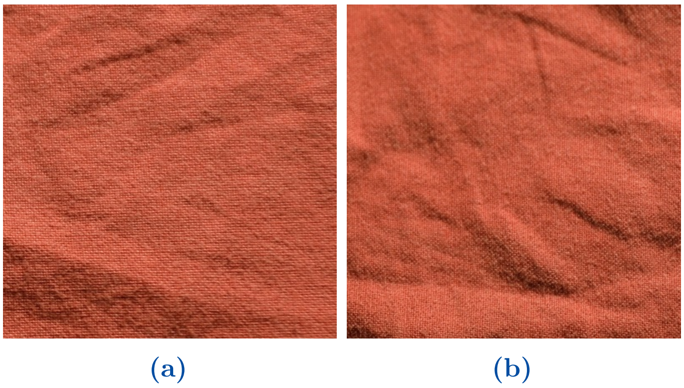

(a) (b)

Figure 4.10: *Images from UJM TIV cotton sample (a) where the illumination direction is at 20°angle and the viewing direction is frontal (b) when the sample orientation is perpendicular and under the identical viewing and illumination conditions. Images from (Sumon et al., 2022).*

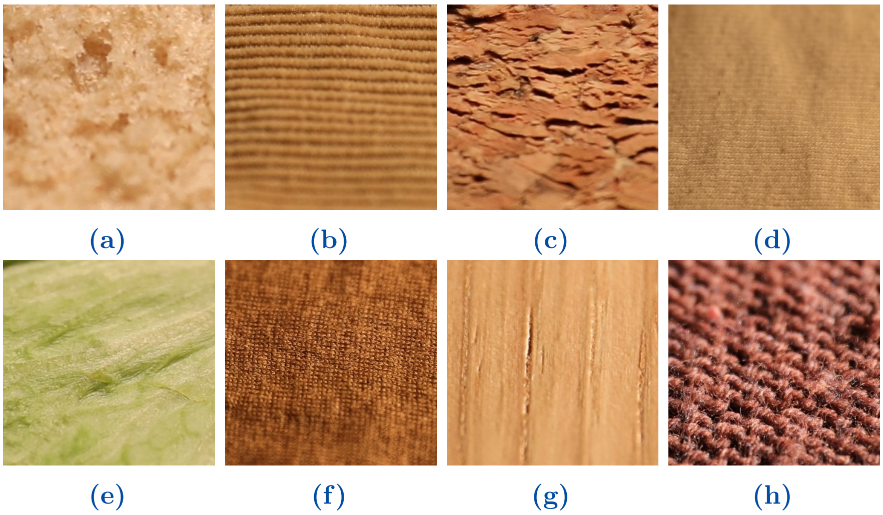


Figure 4.11: *Out of focus image samples from UJM TIV dataset for class a brown bread, b corduroy, c cork, d cotton, e lettuce leaf, f linen, g wood, and h wool. Images from (Sumon et al., 2022).*

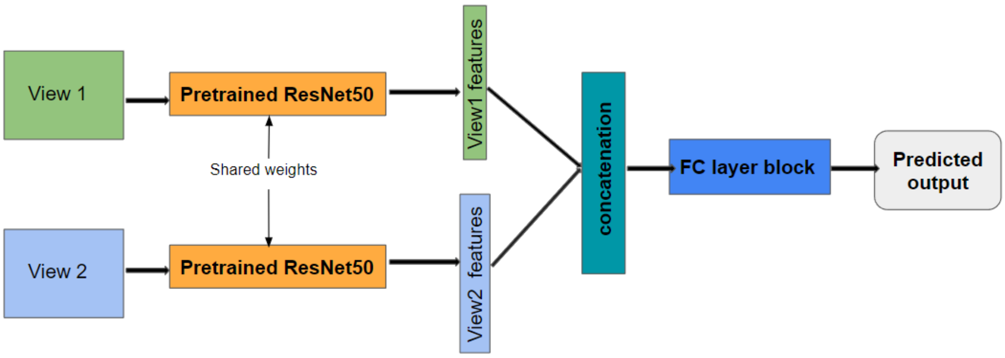


Figure 4.12: *Schemetic of proposed siamese network architecture. Image from (Sumon et al., 2022).*

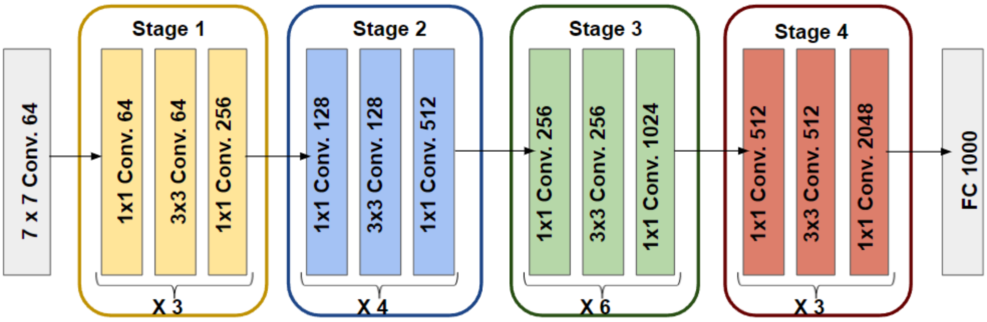
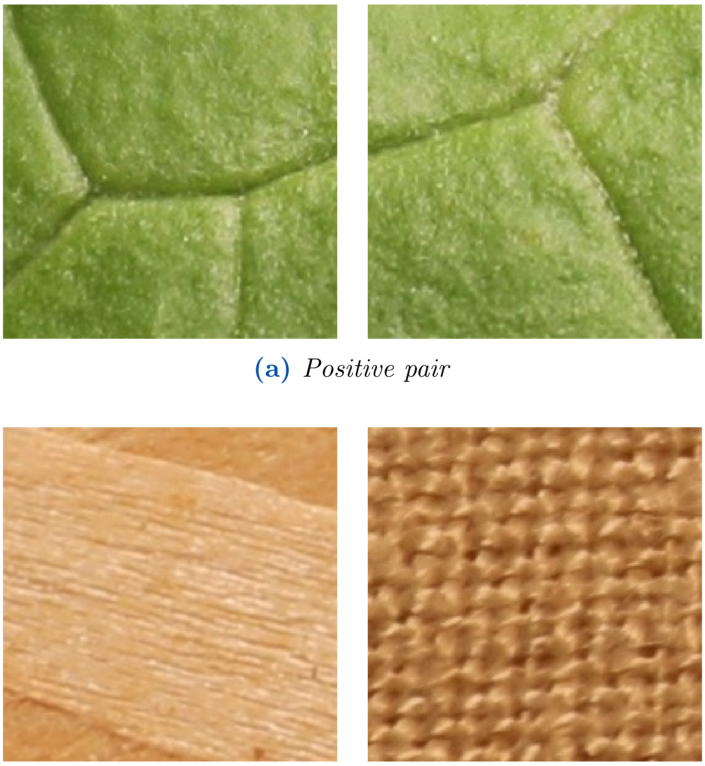


Figure 4.13: *Architecture of ResNet50 (He et al., 2016)*

the building components are trainable with just one classification loss. In order to go a step further, positive and negative pairs of images is created where positive pairs consist of images from same material category and same view. Similarly for creating negative pairs images from different category and different views is chosen. Figure 4.14 shows an example of positive and negative pair of images, where in 4.14a both samples of lettuce leaf is taken from same view. In 4.14b the images on left is an image of wood sample and image of right is taken from category linen but from different views.

The first FC layer of the architecture's design uses a global average pooling (GAP) layer to minimize the amount of inputs before the concatenation of features from each view. Such pooling is believed to aid in preventing overfitting issues as described in (He et al., 2016). Each channel receives an average of all local neuron features from this GAP layer. A global max pooling layer keeps the largest values only from associated feature maps. GMP could be used as an alternate strategy to GAP in the FC layer. In contrast to GMP, which is intended to gather the most crucial information from each feature map, GAP is logically constructed to work with recurring patterns, in which the mean of resulted features bears a significance while eliminating the noise. We think that GAP is better suitable than GMP in this situation for texture images because most of the texture contains repetitive patterns. Additionally, dropout is used in the FC layers to regularize the classifier. Being easily adaptable to more than two views is one of the benefits of such a network design. In fact, any new views may be analyzed for features using the pre-trained backbone; just the fully connected layers need to be modified and retrained to accomplish classification. Only a two-branch architecture has been taught and tested in this study.



(a) *Positive pair*

(b) *Negative pair*

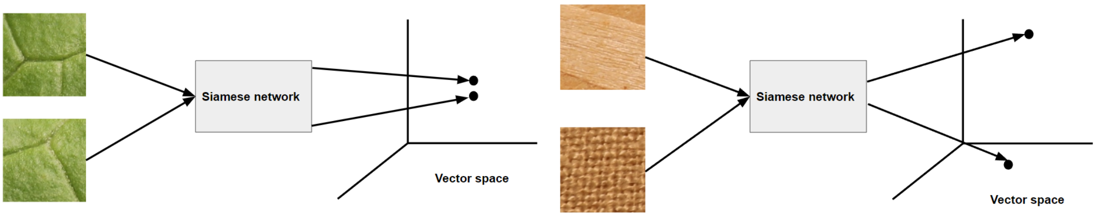
Figure 4.14: *Example of positive and negative pairs of image from UJM TIV.*

4.3 Training with Constructive Loss

As mentioned before siamese network accepts two images as input and each branch generates a fixed length features from the images. These output features are then compared to determine the similarity of the input images. These comparison can be performed in a number of ways such as triplet loss(Schroff et al., 2015), quadruplet loss(Chen et al., 2017), and constructive loss(Hadsell et al., 2006). If triplet loss is used as the loss function in siamese network, at each time step it will take three input samples to compare. The first sample is known as anchor object which is used as a point of comparison with two other data samples. Second one is positive object that is similar to the anchor object. Finally, the third object is negative object which is dissimilar to the anchor object. As the name suggest quadruplet loss consider four data objects at each time step for comparison. In addition to anchor, positive and negative object it requires another negative object which is dissimilar to every other of the 3 data objects (anchor,positive, and negative). On the other hand constructive loss consider two data samples at each step to compare the distance between the feature vectors. This data samples could be either similar(positive) or dissimilar(negative). If the two samples are drawn from the same class the calculated distance will be lower, similarly if they belongs to negative pair the distance will be higher. This scenario is illustrated in figure 4.15. The constructive loss is defined by the following equation

$$loss(d, Y) = \frac{1}{2} * Y * d^2 + (1 - Y) * \frac{1}{2} * max(0, m - d)^2 \tag{4.1}$$

In the equation 4.1 **d** is the distance between the out embedding vectors which could be Euclidean distance or Manhattan distance or any other distance metrics. **Y** is the label of provided input to the model which could be either 1 (if similar) or 0(if dissimilar). Finally, **m** is the margin which usually set to 1. Margin ensures that the observations were properly spaced, i.e. their contribution to error is zero if the distance is greater than margin. As a result, the optimization algorithm may focus on separating challenging Data Point. Hence, margin helps optimization algorithm to separate difficult embedded features in the vector space.



(a) *Positive pair distance*

(b) *Negative pair distance*

Figure 4.15: *Distance measurements on positive and negative image pairs into a vector space.*

4.4 Classification of features using KNN

K Nearest Neighbour is one of the most simple yet useful supervised machine learning algorithm for classification. Based on the characteristics of nearby data points in the training dataset, it produces predictions.By comparing the input sample and the k training instances that are closest to it, the algorithm determines the class to which the test data is most likely to be assigned. It is presumed that similar data points exists close to each other. KNN is effective when is little or no information available about the data distribution. Furthermore, because the method is non-parametric in design, it makes no assumption about the general distribution of the data points. Hence, the KNN algorithm doesn't need any training time. However, finding the appropriate value for K is critical which refers to the number of data points to be considered while making the decision for new data points. This is the key factor since the class to which the majority of these neighboring points belongs determines the classifier output. To show that the proposed model can be trained end to end we take the branch from trained siamese network and used that trained branch to extract the features from the training data. These extracted features are then used with KNN for the classification. KNN calculate the distance between the nearest features with the test features and assign the test features to that class that has the majority votes.

5 Experimental Results and Discussion

We have performed several experiments on two datasets to evaluate the benefits of proposed UJM-TIV dataset and the effectiveness of the suggested multi-view model. This chapter is devoted to the discussion and presentation of results obtained from all these experiments. It is important to note that results shown in subsections 5.3.1, 5.3.2 and 5.3.3 come from paper (Sumon et al., 2022) (officially published in the end of June 2022).

5.1 Experimental Settings

Two architectural designs have been used for the experiments. The first one is a traditional convolutional neural network which extracts the features from the input images, following a FC layer block for the classification of the images. We called this a single branch architecture and accuracy achieved by this network is single view accuracy. The second one is a siamese network with two identical branches where weights are being shared between the branches. As mentioned before pretrained ResNet-50 on imagenet dataset is used as the branch of the siamese network.While the other layers of this network are frozen, the final convolutional layer is tweaked based on the data under consideration. The total number of parameters and fully connected layers are cross validated to ensure fair comparison between two architecture design. We employ the Adam optimization algorithm with a 0.001 learning rate in the beginning for both single-branch and multi-view CNN. If the loss reduction does not take place for several successive epochs, the learning rate dynamically reduced by a factor of 0.2. Input images were resized to 224 x 224 pixels before passing to the network. For all the tests we used batch size 16 and maximum numbers of epochs is 300. Python programming language with version 3.9.5 and Keras deep learning framework with TensorFlow 2.8.0 as backend are used to developed the models. All the experiments were performed on an NVIDIA RTX 8000 high performance GPU with CUDA toolkit version 11.2.

Two configurations were used for the experiments, the first one was using the whole dataset for training and testing. 70% of the data was randomly selected as train set and the rest 30% was used as test set for both KTH-TIPS2 and UJM-TIV dataset. In the second configuration, we evaluate the multi-view learning strategy by choosing a number of views from KTH-TIPS2 and UJM-TIV datasets. Note that there are 16 views in UJM TIV dataset and KTH-TIPS2 has 12 different views (see table 4.2 and 4.1 for detailts). Images from all the views are also splitted into 70-30 for training and testing for both dataset. Figure 4.7 illustrates how viewing and lighting conditions changes effect the general appearance of a sample of cotton that is being observed (blurry, low-contrast), although the variations are not particularly remarkable..

5.2 Evaluation Matrices

5.2.1 Accuracy

Accuracy is a commonly used metric to evaluate the performance of a deep learning model. When all classes are equally important, it is helpful. The number of accurate prediction divided by the total number of predictions is used to compute it. Mathematically this can be written as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

A model can make four types of prediction namely True Positive(TP), True negative(TN), False Positive(FP) and False negative(FN). True positive means positive samples are correctly predicted as positive class. Similarly true negative means negative samples are correctly predicted as negative class. On the other hand false positive means the model predicted test sample as positive class but actually the samples belongs to the negative class. Likewise, false negative denoted as actual test samples belongs to positive class while the model predicted as negative. TP and TN predictions are desired to get a good classification accuracy.

5.2.2 Confusion Matrix

If the dataset has more than two classes or if each class has an uneven amount of observations, classification accuracy alone may be confusing. Confusion matrix is a popular way of summarizing and visualizing the performance of a neural network. For each class, count values are utilized to express the proportion of accurate and inaccurate predictions. One can obtain a better understanding of classification results and helps to identify where the model is getting right and wrong. Confusion matrix not only reveal the mistakes of the classifier is making, but more crucially, it reveals the specific mistakes that are being done.

5.3 Results

Depending on the type of data used to train and test the networks, the findings are divided into two sections. The results of the test on the entire dataset are presented first, followed by the results on the chosen views.

5.3.1 Results from single-view network

In this section performance of a single-branch network is discussed across all data. Table 5.1 lists the outcomes for both datasets.The accuracy achieved by the single view network on KTH-TIPS2 dataset is 80% which is similar to the accuracy of traditional deep neural network used in (Sixiang et al., 2020).On the other hand as we observe from the table, the accuracy on UJM-TIV dataset while using the same network configuration as on KTH-TIPS2, is substantially worse. We believe that because of higher intra class variability in UJM TIV dataset as compared to KTH-TIPS2, the single-view network obtained significant accuracy improvement on KTH-TIPS2.

Table 5.1: *Single branch model's accuracy on KTH-TIP2 and UJM TIV dataset when all the views are considered. Table from (Sumon et al., 2022)*

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

5.3.2 Results from multi-view network

This section is devoted to the discussion about the results obtained from multi-view network. Selected distinct views are used to train and test the multi-view network. From the chosen views we train the model using the training set and evaluate the model using the test set from same selected views. Performance of the multi-view CNN on KTH-TIPS2 dataset is reported in table 5.2. Similarly obtained results on UJM-TIV dataset is presented in 5.3.

Table 5.2: *Model performance of single-view and multi-view network on KTH-TIPS2 dataset.Table from (Sumon et al., 2022)*

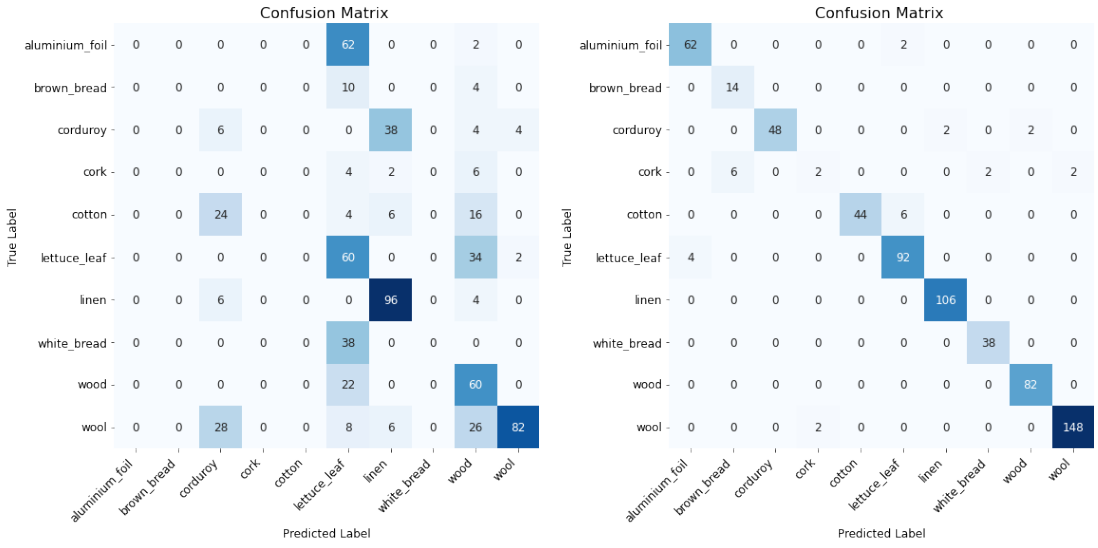
Train data	Test data	Single-view accuracy	Multi-view accuracy	Improvement (%)
view1,view2	view1, view2	56.90	68.53	+29.76
view3,view4	view3, view4	60.34	67.24	+10.26
view5,view6	view5, view6	56.91	71.98	+20.94
view7,view8	view7, view8	39.66	47.41	+16.35
view9,view10	view9, view10	34.48	64.22	+46.31
view11,view12	view11,view12	37.93	67.24	+43.59

From these tables we can see that accuracy of single view model for both dataset is decreased when the model is trained on only two views as compared to the results when it was trained with all the views. This is expected because the single view network is trained on less data when considering only two views. Another observation is for all the pair of selected views multi-view learning approach outperform the single-view network with a significant margin. These results clearly demonstrate the superiority of multi-view CNN over the traditional single view CNN for material classification. Additionally, when the two views are noticeably different, multi-view learning works better than the single-view strategy. It is evident from the result (46% improvement) of view9 and view10 pair from KTH-TIP2 where there is larger viewing differences (45°) between the samples of view9 and view10. Because there exists higher variation in appearances accross the samples of different views in our proposed UJM-TIV dataset, multi-view learning approach achieved

Table 5.3: *Model performance of single-view and multi-view network on UJM-TIV dataset. Table from (Sumon et al., 2022)*

Train data	Test data	Single-view accuracy	Multi-view accuracy	Improvement (%)
view1,view2	view1, view2	50.28	79.52	+36.77
view3,view4	view3, view4	60.00	75.29	+20.31
view5,view6	view5, view6	44.48	95.71	+53.52
view7,view8	view7, view8	51.32	96.52	+46.83
view9,view10	view9, view10	65.59	95.29	+31.17
view11,view12	view11,view12	66.63	94.56	+29.54
view13,view14	view13,view14	80.33	89.34	+10.08
view15,view16	view15,view16	53.91	83.78	+35.65

better results for all the views over the single view CNN as shown in table 5.3. The performance of multi-view learning on UJM-TIV dataset strongly demonstrates the usefulness of our dataset for classification of material as well as the suitability of our proposed siamese network for two-view learning. To summarize and visualize the performance confusion matrix of both single and multi-view network on view5, and view6 pair which achieved the highest relative improvement (+53.5%) on multi-view approach over single view network is depicted in figure 5.1. The confusion matrix clearly shows the reason why single view network perform worse than the multi-view network in view5, and view6. As shown in figure 5.1a except for linen, the single view network fails to correctly classify the test samples from most of the categories. On the other hand, multi-view network is able to correctly predict the almost all the test samples to their corresponding categories (see figure 5.1b).



	aluminium_foil	brown_bread	corduroy	cork	cotton	lettuce_leaf	linen	white_bread	wood	wool
aluminium_foil	0	0	0	0	0	62	0	0	2	0
brown_bread	0	0	0	0	0	10	0	0	4	0
corduroy	0	0	6	0	0	0	38	0	4	4
cork	0	0	0	0	0	4	2	0	6	0
cotton	0	0	24	0	0	4	6	0	16	0
lettuce_leaf	0	0	0	0	0	60	0	0	34	2
linen	0	0	6	0	0	0	96	0	4	0
white_bread	0	0	0	0	0	38	0	0	0	0
wood	0	0	0	0	0	22	0	0	60	0
wool	0	0	28	0	0	8	6	0	26	82

(a) *Single view*

	aluminium_foil	brown_bread	corduroy	cork	cotton	lettuce_leaf	linen	white_bread	wood	wool
aluminium_foil	62	0	0	0	0	2	0	0	0	0
brown_bread	0	14	0	0	0	0	0	0	0	0
corduroy	0	0	48	0	0	0	2	0	2	0
cork	0	6	0	2	0	0	0	2	0	2
cotton	0	0	0	0	44	6	0	0	0	0
lettuce_leaf	4	0	0	0	0	92	0	0	0	0
linen	0	0	0	0	0	0	106	0	0	0
white_bread	0	0	0	0	0	0	0	38	0	0
wood	0	0	0	0	0	0	0	0	82	0
wool	0	0	0	2	0	0	0	0	0	148

(b) *Multi-view*

Figure 5.1: *Confusion matrix of a Single-view CNN and b Multi-view CNN when considering the view5 and view6 pairs from UJM-TIV dataset. Image from (Sumon et al., 2022)*

5.3.3 Experiments with State-of-the-Art Solution

The results on previous sections illustrated that by employing multi-view approach on a simple deep neural network can boost the classification accuracy even if the network is not specially designed for the purpose of material classification.We also wanted to check if our multi-view approach can be applied to the state-of-the-art solution specifically designed for material classification task based on Fisher score

(Xu et al., 2021). This technique involves a training phase made up of three sequential phases and makes use of orderless pooling and sparse coding. The network is also trained and tested for all the selected views from both KTH-TIPS2 and UJM-TIV datasets. The results on KTH-TIPS2 and UJM-TIV dataset are reported in table 5.4 and 5.5 respectively.

Table 5.4: *Accuracy of single-view and multi-view learning models when using State-of-the-art Xu et al. (2021) approach on KTH-TIPS2 dataset. Table from (Sumon et al., 2022)*

Train data	Test data	Single-view accuracy	Multi-view accuracy	Improvement (%)
view1,view2	view1, view2	94.7	97.5	+3.0
view3,view4	view3, view4	90.0	96.67	+6.90
view5,view6	view5, view6	90.83	95.83	+5.22
view7,view8	view7, view8	92.50	98.33	+5.93
view9,view10	view9, view10	92.50	95.83	+3.47
view11,view12	view11,view12	90.00	94.17	+4.40

Table 5.5: *Accuracy of single-view and multi-view learning models when using State-of-the-art Xu et al. (2021) approach on UJM-TIV dataset. Table from (Sumon et al., 2022).*

Train data	Test data	Single-view accuracy	Multi-view accuracy	Improvement (%)
view1,view2	view1, view2	100	98.99	-1.02
view3,view4	view3, view4	99.44	100	+0.56
view5,view6	view5, view6	99.85	100	+0.15
view7,view8	view7, view8	99.88	100	+0.12
view9,view10	view9, view10	99.56	100	+0.44
view11,view12	view11,view12	99.88	100	+0.12
view13,view14	view13,view14	99.80	100	+0.20
view15,view16	view15,view16	98.31	99.58	+1.28

Given that the tested network surpasses every result from Tables 5.2 and 5.3, the above findings show the network's applicability to material classification. Results show that despite having such a good results on single view network, multi-view learning approach helped to improve the accuracy for almost all the views. The relative improvement of multi-view approach over single-view is not very high as compared to the results from previous tests since the single-view network already achieved quite good classification accuracy on the both material datasets. Meanwhile the mean accuracy improvement is about 4.8% on KTH-TIPS2 dataset. The results on UJM-TIV with the tested network is nearly perfect when multi-view learning technique combined with the state-of-the-art solution (Xu et al., 2021). Undoubtedly, the most recent experiments demonstrated how our contributions may be utilized to enhance the performance of any state of the art system for classifying materials.

5.3.4 Experiments with KNN

In this section the experiment results obtained with KNN classifier is reported. Our proposed siamese network is end to end trainable which means that the branch from siamese network can be used for feature extraction and consequently these extracted features can be used for classification purpose. Several experiments are performed on UJM-TIV dataset which proved the above statement. The siamese network generate d-dimensional features or embedding from the input images. The KNN algorithm take those features and classify them by calculating the distances between the test features and considered nearest neighbors features. Figure 5.2 shows the location of generated embedding before and after a the model trained with 300 epochs. It can be observed from the plot that how the model is already producing similar embeddings for images belonging to the same class after only

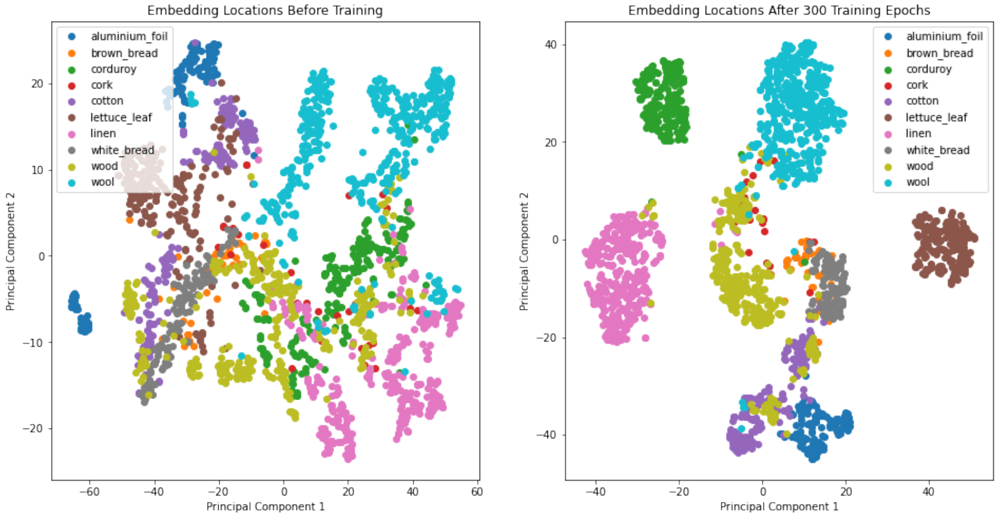
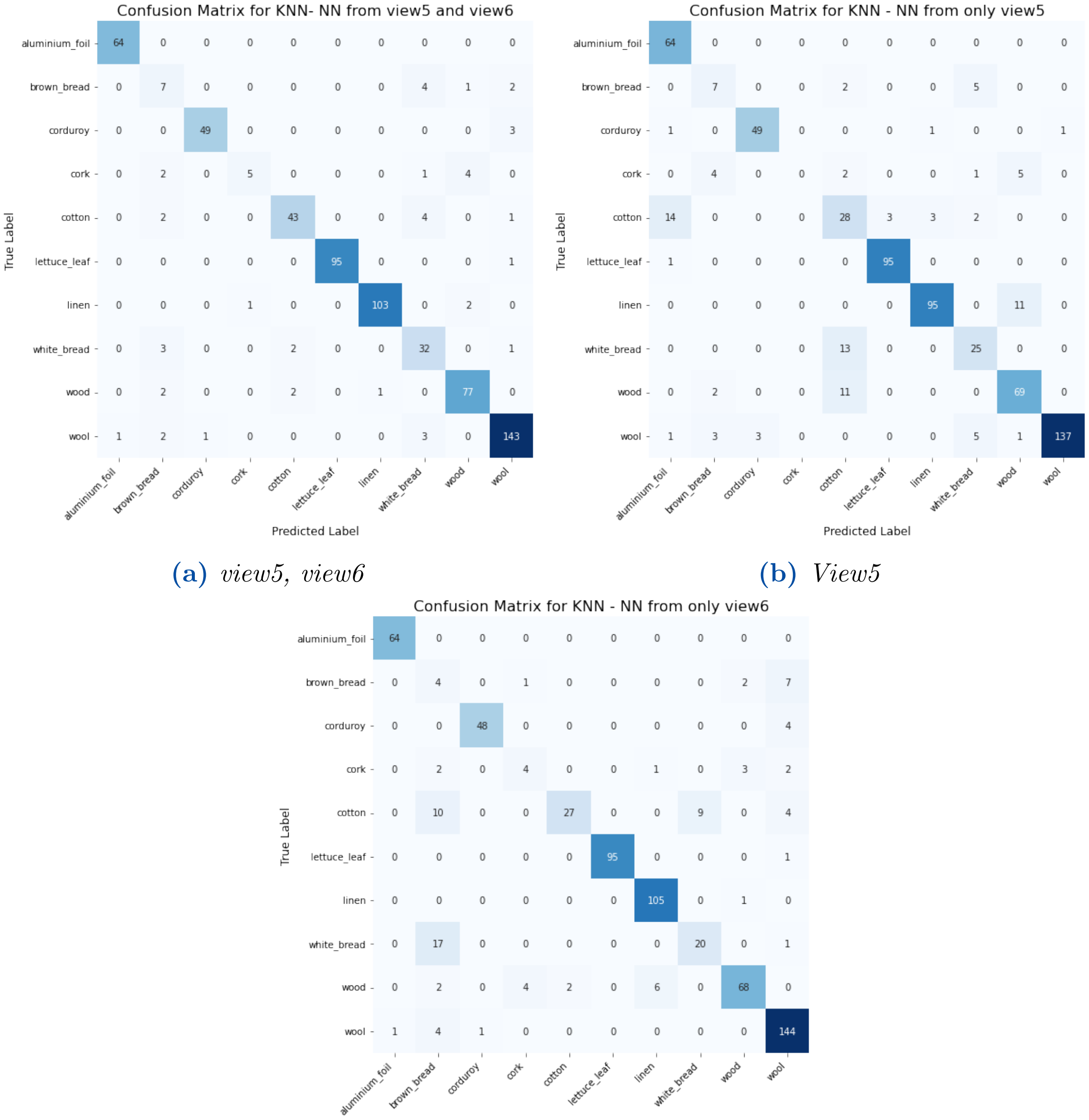


Figure 5.2: *Embedding locations before (left) and after (right) 300 training epochs. Using PCA, projected down to two dimensions.*



(a) *view5, view6*

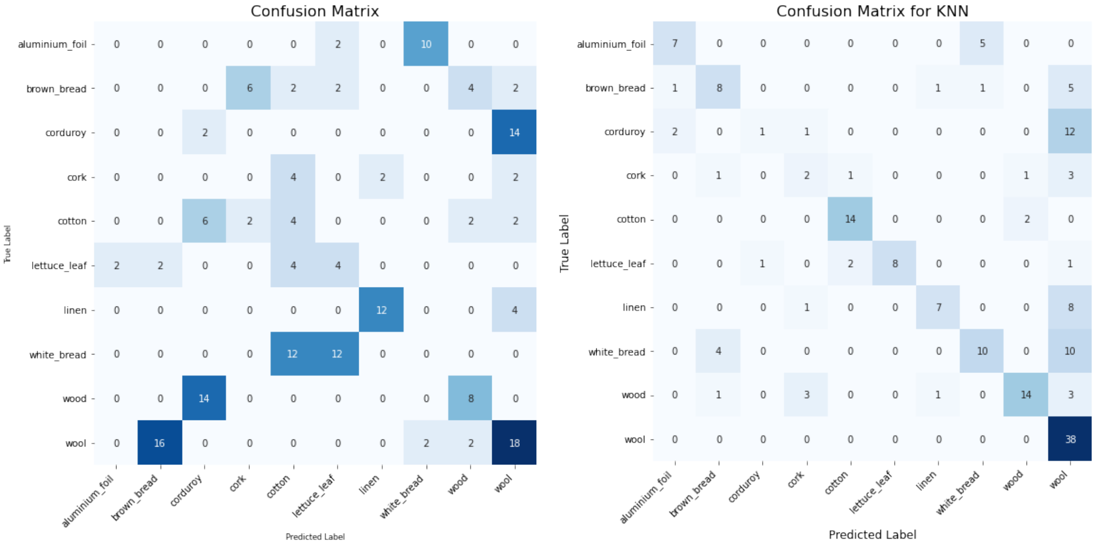
(b) *View5*

(c) *View6*

Figure 5.3: *Confusion matrix obtained using KNN when considering neighbors samples from a view5 and view6, b view5, and c view6 from UJM-TIV dataset.*

300 training epochs. The clusters of identically colored dots in the graph of figures 5.2 demonstrate this; some clusters are seen stacked on top of one another in the plot as a result of the reduction of hyperspace to 2-D by the PCA. This embedding clustering is what gives siamese network their strength. If we want to test the model on new class data we don't need to retrain the model with that new class again. If the new class data is plotted it should be far from the current clusters, but if more samples of the new class are added, the samples from new class should begin to cluster with one another. With just a small amount of data, we may start to get reasonable classifications results for both seen and unseen classes by exploiting this embedding similarity. Results from different experiments with different views is reported on table 5.6. We performed two types of tests using KNN algorithm. First ,we train the KNN with features from a view pairs and tested on the test set belong to same view pairs and compared with the results obtained using single view model. The KNN algorithm checks nearest neighbors sample from both views in this case. The accuracy obtained by KNN on UJM-TIV dataset significantly outperform the single-view model accuracy with a large margin for all the considered view pairs. The maximum accuracy improvement (48%) achieved for view5, view6 pair which is also true when multi-view model (+53% improvement, see table 5.3) is used for this view pair. These results shows that our model is well suited for classification of such kind of material features where multiple view plays an important role.

As a second experiment we wanted to check how the KNN algorithm performs on the test data when samples from only one view is checked while determining the class for the test data. Samples are considered from both alternative samples with separate experiments. Even if when KNN consider only one sample for determining class for the test data points, the accuracy is still higher than the single-view model for the same pair of views. From the table 5.6 it is clear that for all the view pairs single-view model performed worse than the accuracy obtained by KNN when considering samples from only one view as the neighbors of the test data points. Because of there are higher changes in viewing (30°) and illumination (70°) direction between view5 and view6 pair the model trained on these view pair achieved maximum improvement on accuracy. Confusion matrices for the best performing KNN classifier on view5 and view6 pair is shown in figure 5.3. 5.3a shows the performance of KNN when nearest neighbors samples are considered from both view5 and view6 pair while assigning the category for the test data points. Similarly, figures 5.3b and 5.3c show the results when only images from view5 and view6 are considered as neighboring samples respectively. All of the findings point to multi-view learning as a promising material classification technique which includes images captured under different viewing or lighting geometry. As the final experiment we performed cross view test i.e train on one view pair and test on another view pairs. We compared results from both multi-view learning approach with siamese network and KNN algorithm for classification of material images from UJM-TIV dataset. Cross views experiments results are shown in table 5.7. Confusion matrices from the cross view tests for best performing models are displayed in 5.4. It is clear from the results on all the view pairs, KNN classifier achieved greater or at least similar accuracy as multi-view network. Results from confusion matrices clearly show that KNN classifier got better accurate predictions (see figure 5.4b) while multi-view model was struggling to correctly classify almost all the test samples to their corresponding class (see figure 5.4a). Note that in this case both the models were trained on images from view5 and view6 but tested on view3, view4 images.



(a) *Multi-view*

(b) *KNN*

Figure 5.4: *Confusion matrix obtained using a Multi-view and b KNN from cross view experiments when the model trained on view5 view6 and tested on view3, view4 pair.*

These results from KNN is expected because as mentioned before we take the CNN branch of the trained siamese network to extract the d-dimensional features from the selected train and test set. A KNN classifier is then trained using the extracted train features and evaluated the results on the test features extracted by the CNN siamese branch. As seen before the highest improvement in the classification accuracy is observed for trained on view5, view6 and tested on view3, view4 pairs. Siamese model achieved very low accuracy (27%) on test set (view3 and view4) on the other hand KNN classifier gained about 60% accuracy on the same test set. All of these experimental results indicate that our proposed network can be easily adapted for classification of material samples images from unseen

Table 5.6: *Accuracy of single-view network and KNN on different views from UJM-TIV dataset.*

Train data	Test data	Single-view accuracy	KNN accuracy(k=3)	Nearest Neighbors from
view1,view2	view1,view2	50.28	86.36 86.00 81.00	view1,view2 view1 only view2 only
view3,view4	view3,view4	60.00	83.89 80.00 77.00	view3,view4 view3 only view4 only
view5,view6	view5,view6	44.48	93.07 85.69 87.20	view5,view6 view5 only view6 only
view7,view8	view7,view8	51.32	91.5 83.26 88.34	view7,view8 view7 only view8 only
view9,view10	view9,view10	65.59	94.48 90.7 90.26	view9,view10 view9 only view10 only
view11,view12	view11,view12	66.63	90.95 87.78 87.04	view11,view12 view11 only view12 only
view13,view14	view13,view14	80.33	95.16 91.53 90.52	view13,view14 view13 only view14 only
view15,view16	view15,view16	53.91	84.75 81.78 79.03	view15,view16 view15 only view16 only

Table 5.7: *Accuracy of multi-view model and KNN classifier for cross view experiments from UJM-TIV dataset.*

Train data	Test data	Multi-view accuracy	KNN accuracy
view1,view2	view3, view4	52.94	73.89
view3,view4	view1, view2	74.30	74.24
view5,view6	view3, view4	27.65	60.56
view7,view8	view5, view6	57.36	74.55
view9,view10	view7, view8	40.71	57.04
view11,view12	view9,view10	70.00	69.19
view13,view14	view11,view12	56.49	59.54
view15,view16	view13,view14	60.66	74.40

views.

6 Conclusion and Future Work

Material image classification is one crucial task in computer vision because it is involved in many real applications such as robotics or automatic waste sorting, and because it can help in many other problems such as fine-grained image classification. It consists in correctly classifying images with target material from one given category. In the beginning of 2010s, thanks to their superior performances, the deep convolutional neural networks (CNN) arise and become a promising tool to solve many computer vision problems, including image classification. Deep networks have also been introduced into material classification. By simply transferring a network pretrained on a large-scale image classification task, better accuracy is achieved than former state-of-the-arts. However, unlike object recognition, classifying materials requires some specific processing.

In this thesis, a new material dataset is proposed with significant intra-class variability across different material classes. The wide variety of illumination and viewing conditions settings and the choice of varied material samples are responsible for the variances in appearance throughout each class. With the help of a number of experiments we have demonstrated that traditional deep learning networks are not well capable of classification of material samples contain large variation in visual appearances. This leads to an alternative proposed multi-view learning solution which takes the advantages of discriminative features from images of different views. For this we proposed a siamese network consisting of two branches to extract features from different input views and combining these information to achieve better performance. Experiment results show that the proposed multi-view approach through siamese network surpasses the traditional single view solution. One limitation of the proposed approach it can works with only a pair of views. How the network could perform when more that two views are employed could be a future research direction. Multi-view learning for material classification involves tweaking a large number of acquisition parameters such as object pose, viewing and illumination geometry which is a major challenge and it may not be possible to consider all the parameters involved. In the proposed multi-view learning we utilized **one-view-one-net** strategy. Another existing strategy namely

multi-view-one-net could be used in future research to see how the changes in architectural design affect the performance on the material classification task. We performed cross view experiments i.e. train on a view pair and test on different view pairs. This idea, and the corresponding results, will be submitted to a second publication in September 2022. However cross dataset test i.e. train on one dataset and test on another dataset could be an interesting future investing which may helps to reveal the strength and limitations of the proposed multi-view solution. In the future we plan to use the proposed multi-view learning approach to photomatrically reconstruction of complex scene object having spatially varying surface reflectances and which may helps to increase the effectiveness of SVBRDF-based single-image rendering techniques(Deschaintre et al., 2018). In this situation, adding synthetic data to the datasets may be used to modify the input BRDF, would be of interest (Krishna et al., 2021; Zhang et al., 2022; Brochard et al., 2022).

Bibliography

Andrearczyk, V. and Whelan, P. F. (2016). Using filter banks in convolutional neural networks for texture classification. *Pattern Recognition Letters*, 84:63–69. (cited on pages 7 and 9)

Bay, H., Tuytelaars, T., and Gool, L. V. (2006). Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer. (cited on page 6)

Bell, S., Upchurch, P., Snavely, N., and Bala, K. (2015). Material recognition in the wild with the materials in context database. *Computer Vision and Pattern Recognition (CVPR)*. (cited on page 16)

Bello-Cerezo, R., Bianconi, F., Di Maria, F., Napoletano, P., and Smeraldi, F. (2019). 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*, 9(4):738. (cited on page 6)

Bovik, A. C., Clark, M., and Geisler, W. S. (1990). Multichannel texture analysis using localized spatial filters. *IEEE transactions on pattern analysis and machine intelligence*, 12(1):55–73. (cited on pages 1 and 5)

Brochard, A., Zhang, S., and Mallat, S. (2022). Generalized rectifier wavelet covariance models for texture synthesis. *arXiv preprint arXiv:2203.07902*. (cited on page 50)

Chen, W., Chen, X., Zhang, J., and Huang, K. (2017). Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412. (cited on page 34)

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , and Vedaldi, A. (2014). Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. (cited on page 17)

Cimpoi, M., Maji, S., and Vedaldi, A. (2015). Deep filter banks for texture recognition and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3828–3836. (cited on page 7)

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague. (cited on page 5)

Cusano, C., Napoletano, P., and Schettini, R. (2016). Evaluating color texture descriptors under large variations of controlled lighting conditions. *JOSA A*, 33(1):17–30. (cited on page 13)

Dai, X., Yue-Hei Ng, J., and Davis, L. S. (2017). Fason: First and second order information fusion network for texture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7352–7360. (cited on page 7)

Dana, K. J., Van Ginneken, B., Nayar, S. K., and Koenderink, J. J. (1999). Reflectance and texture of real-world surfaces. *ACM Transactions On Graphics (TOG)*, 18(1):1–34. (cited on page 12)

Debevec, P. (2008). Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Acm siggraph 2008 classes*, pages 1–10. (cited on page 20)

Deepak, K., Srivathsan, G., Roshan, S., and Chandrakala, S. (2021). Deep multiview representation learning for video anomaly detection using spatiotemporal autoencoders. *Circuits, Systems, and Signal Processing*, 40(3):1333–1349. (cited on page 8)

Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., and Bousseau, A. (2018). Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (ToG)*, 37(4):1–15. (cited on page 50)

Dong, X. and Shen, J. (2018). Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 459–474. (cited on page 9)

Dou, Q., Chen, H., Yu, L., Qin, J., and Heng, P.-A. (2016). Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection. *IEEE Transactions on Biomedical Engineering*, 64(7):1558–1567. (cited on page 9)

Erickson, Z., Xing, E., Srirangam, B., Chernova, S., and Kemp, C. C. (2020). Multimodal material classification for robots using spectroscopy and high resolution texture imaging. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10452–10459. IEEE. (cited on page 8)

Fan, W., Ma, Y., Xu, H., Liu, X., Wang, J., Li, Q., and Tang, J. (2020). Deep adversarial canonical correlation analysis. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 352–360. SIAM. (cited on page 8)

Fatima, N., Areeb, Q. M., Khan, I. M., and Khan, M. M. (2021). Siamese network-based computer vision approach to detect papaya seed adulteration in black peppercorns. *Journal of Food Processing and Preservation*, page e16043. (cited on page 10)

Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941. (cited on pages 8 and 9)

Ghose, S., Chowdhury, P. N., Roy, P. P., and Pal, U. (2021). Modeling extent-of-texture information for ground terrain recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4766–4773. IEEE. (cited on page 7)

Gorpas, D., Kampouris, C., and Malassiotis, S. (2013). Miniature photometric stereo system for textile surface structure reconstruction. In *Videometrics, Range Imaging, and Applications XII; and Automated Visual Inspection*, volume 8791, pages 271–282. SPIE. (cited on page 8)

Guo, H., Wang, J., Xu, M., Zha, Z.-J., and Lu, H. (2015). Learning multi-view deep features for small object retrieval in surveillance scenarios. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 859–862. (cited on page 8)

Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L., and Wang, S. (2017). Learning dynamic siamese network for visual object tracking. In *Proceedings of the IEEE international conference on computer vision*, pages 1763–1771. (cited on page 9)

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE. (cited on page 34)

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. (cited on pages 33 and 63)

Hu, Y., Long, Z., and AlRegib, G. (2019). Multi-level texture encoding and representation (multer) based on deep neural networks. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4410–4414. IEEE. (cited on page 7)

Jain, A. K. and Farrokhnia, F. (1991). Unsupervised texture segmentation using gabor filters. *Pattern recognition*, 24(12):1167–1186. (cited on pages 1 and 5)

Julesz, B. (1962). Visual pattern discrimination. *IRE transactions on Information Theory*, 8(2):84–92. (cited on pages 1 and 5)

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97. (cited on pages 1 and 5)

Julesz, B. and Bergen, J. R. (1983). Human factors and behavioral science: Textons, the fundamental elements in preattentive vision and perception of textures. *Bell system technical journal*, 62(6):1619–1645. (cited on pages 1 and 5)

Kampouris, C., Zafeiriou, S., Ghosh, A., and Malassiotis, S. (2016a). Fine-grained material classification using micro-geometry and reflectance. In *European Conference on Computer Vision*, pages 778–792. Springer. (cited on page 8)

Kampouris, C., Zafeiriou, S., Ghosh, A., and Malassiotis, S. (2016b). Fine-grained material classification using micro-geometry and reflectance. In *European Conference on Computer Vision*, pages 778–792. Springer. (cited on pages 22 and 27)

Kapeller, C., Antensteiner, D., and Štolc, S. (2020). Tailored photometric stereo: Optimization of light source positions for various materials. *Electronic Imaging*, 2020(6):71–1. (cited on page 27)

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137. (cited on page 8)

Khvedchenya, E. and Gabruseva, T. (2021). Fully convolutional siamese neural networks for buildings damage assessment from satellite images. *arXiv preprint arXiv:2111.00508*. (cited on page 10)

Koch, G., Zemel, R., Salakhutdinov, R., et al. (2015). Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille. (cited on page 29)

Krishna, A., Bartake, K., Niu, C., Wang, G., Lai, Y., Jia, X., and Mueller, K. (2021). Image synthesis for data augmentation in medical ct using deep reinforcement learning. *arXiv preprint arXiv:2103.10493*. (cited on page 50)

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25. (cited on pages 2 and 6)

Leung, T. and Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. *International journal of computer vision*, 43(1):29–44. (cited on pages 1 and 5)

Liang, Y., Wang, J., Wan, X., Gong, Y., and Zheng, N. (2016). Image quality assessment using similar scene as reference. In *European Conference on Computer Vision*, pages 3–18. Springer. (cited on page 9)

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer. (cited on page 19)

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110. (cited on page 6)

Malik, J. and Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *JOSA A*, 7(5):923–932. (cited on pages 1 and 5)

Mallikarjuna, P., Targhi, A. T., Fritz, M., Hayman, E., Caputo, B., and Eklundh, J.-O. (2006). The kth-tips2 database. *Computational Vision and Active Perception Laboratory, Stockholm, Sweden*, 11. (cited on pages 12 and 21)

Manjunath, B. S. and Ma, W.-Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on pattern analysis and machine intelligence*, 18(8):837–842. (cited on pages 1 and 5)

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., and Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*. (cited on page 8)

Maximov, M., Leal-Taixé, L., Fritz, M., and Ritschel, T. (2019). Deep appearance maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8729–8738. (cited on page 8)

Melekhov, I., Kannala, J., and Rahtu, E. (2016). Siamese network features for image matching. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 378–383. IEEE. (cited on page 29)

Michael, W., Juergen, G., and Reinhard, K. (2014). Material Classification Based on Training Data Synthesized Using a BTF Database. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III*, pages 156–171. Springer International Publishing. (cited on page 13)

Napoletano, P. (2017). Hand-crafted vs learned descriptors for color texture classification. In *International Workshop on Computational Color Imaging*, pages 259–271. Springer. (cited on page 7)

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987. (cited on page 6)

Randen, T. and Husoy, J. H. (1999). Filtering for texture classification: A comparative study. *IEEE Transactions on pattern analysis and machine intelligence*, 21(4):291–310. (cited on page 6)

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252. (cited on pages 2 and 6)

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823. (cited on page 34)

Sharan, L., Rosenholtz, R., and Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of vision*, 14(9):12–12. (cited on page 15)

Sixiang, X., Damien, M., Alain, T., and Robert, L. (2020). Confidence-based local feature selection for material classification. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE. (cited on page 39)

Srivastava, N. and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25. (cited on page 8)

Sticlaru, A. (2017). Material classification using neural networks. *arXiv preprint arXiv:1710.06854*. (cited on page 2)

Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E. (2015). Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953. (cited on page 9)

Sumon, B. U., Muselet, D., Xu, S., and Trémeau, A. (2022). Multi-view learning for material classification. *Journal of Imaging*, 8(7):186. (cited on pages 2, 3, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 37, 39, 40, 41, 42, 61, 62, 63, and 65)

Targhi, A. T., Geusebroek, J.-M., and Zisserman, A. (2008). Texture classification with minimal training images. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. (cited on page 19)

Trémeau, A., Xu, S., and Muselet, D. (2020). Deep learning for material recognition: most recent advances and open challenges. *arXiv preprint arXiv:2012.07495*. (cited on page 2)

Turner, M. R. (1986). Texture discrimination by gabor functions. *Biological cybernetics*, 55(2):71–82. (cited on pages 1 and 5)

Varga, D. and Szirányi, T. (2017). Person re-identification based on deep multi-instance learning. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1559–1563. IEEE. (cited on page 9)

Vrancken, C., Longhurst, P., and Wagland, S. (2019). Deep learning in material recovery: Development of method to create training database. *Expert Systems with Applications*, 125:268–280. (cited on page 8)

Wang, B. and Wang, D. (2019). Plant leaves classification: A few-shot learning method based on siamese network. *IEEE Access*, 7:151754–151763. (cited on page 10)

Wang, T.-C., Zhu, J.-Y., Hiroaki, E., Chandraker, M., Efros, A. A., and Ramamoorthi, R. (2016). A 4d light-field dataset and cnn architectures for material recognition. In *European conference on computer vision*, pages 121–138. Springer. (cited on page 17)

Wei, J., Xia, Y., and Zhang, Y. (2019). M3net: A multi-model, multi-size, and multi-view deep neural network for brain magnetic resonance image segmentation. *Pattern Recognition*, 91:366–378. (cited on page 8)

Weinmann, M., Gall, J., and Klein, R. (2014). Material classification based on training data synthesized using a btf database. In *European Conference on Computer Vision*, pages 156–171. Springer. (cited on page 20)

Weiyuan Wu, Divakar Verma, W. Y. (2020). *Python patchify library*. `https://pypi.org/project/patchify/` [Accessed: **12-07-2022**]. (cited on page 26)

Wieschollek, P. and Lensch, H. (2016). Transfer learning for material classification using convolutional networks. *arXiv preprint arXiv:1609.06188*. (cited on page 6)

Wiggers, K. L., Britto, A. S., Heutte, L., Koerich, A. L., and Oliveira, L. S. (2019). Image retrieval and pattern spotting using siamese neural network. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE. (cited on page 29)

Wu, Y. N., Zhu, S. C., and Liu, X. (2000). Equivalence of julesz ensembles and frame models. *International Journal of Computer Vision*, 38(3):247–265. (cited on page 5)

Xiaoqiang, Y., Shizhe, H., Yiqiao, M., Yangdong, Y., and Hui, Y. (2021). Deep multi-view learning methods: A review. *Neurocomputing*, (448):106–129. (cited on page 28)

Xie, J., Hu, W., Zhu, S.-C., and Wu, Y. N. (2015). Learning sparse frame models for natural image patterns. *International Journal of Computer Vision*, 114(2):91–112. (cited on page 5)

Xu, J., Zheng, H., Wang, J., Li, D., and Fang, X. (2020). Recognition of eeg signal motor imagery intention based on deep multi-view feature learning. *Sensors*, 20(12):3496. (cited on page 8)

Xu, S. (2021). *Transfer learning for material classification based on material appearance correspondances*. PhD thesis, University Jean Monnet. (cited on page 2)

Xu, S., Muselet, D., and Alain, T. (2019). Deep learning for material recognition: most recent advances and open challenges. In *International Conference on Big Data, Machine Learning and Applications (BIGDML)*. (cited on page 8)

Xu, S., Muselet, D., and Trémeau, A. (2021). Deep fisher score representation via sparse coding. In *International Conference on Computer Analysis of Images and Patterns*, pages 412–421. Springer. (cited on pages 42, 43, and 65)

Xue, J., Zhang, H., and Dana, K. (2018). Deep texture manifold for ground terrain recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 558–567. (cited on page 7)

Xue, J., Zhang, H., Dana, K., and Nishino, K. (2017). Differential angular imaging for material recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 764–773. (cited on page 16)

Yan, X., Hu, S., Mao, Y., Ye, Y., and Yu, H. (2021). Deep multi-view learning methods: a review. *Neurocomputing*, 448:106–129. (cited on page 9)

Yang, Z.-X., Tang, L., Zhang, K., and Wong, P. K. (2018). Multi-view cnn feature aggregation with elm auto-encoder for 3d shape recognition. *Cognitive Computation*, 10(6):908–921. (cited on page 9)

Yu, C., Zhao, X., Zheng, Q., Zhang, P., and You, X. (2018). Hierarchical bilinear pooling for fine-grained visual recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 574–589. (cited on page 7)

Zhai, W., Cao, Y., Zha, Z.-J., Xie, H., and Wu, F. (2020). Deep structure-revealed network for texture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11010–11019. (cited on pages 7 and 8)

Zhai, W., Cao, Y., Zhang, J., and Zha, Z.-J. (2019). Deep multiple-attribute-perceived network for real-world texture recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3613–3622. (cited on pages 7 and 8)

Zhang, H., Xue, J., and Dana, K. (2017). Deep ten: Texture encoding network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 708–717. (cited on page 7)

Zhang, Y., Wang, L., Qi, J., Wang, D., Feng, M., and Lu, H. (2018). Structured siamese network for real-time visual tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 351–366. (cited on page 9)

Zhang, Y., Wang, Q., and Hu, B. (2022). Minimalgan: diverse medical image synthesis for data augmentation using minimal training data. *Applied Intelligence*, pages 1–18. (cited on page 50)

Zhu, S.-C. (2003). Statistical modeling and conceptualization of visual patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6):691–712. (cited on pages 1 and 5)

Zhu, S.-C., Guo, C.-E., Wang, Y., and Xu, Z. (2005). What are textons? *International Journal of Computer Vision*, 62(1):121–143. (cited on page 5)

Zhu, S. C., Liu, X. W., and Wu, Y. N. (2000). Exploring texture ensembles by efficient markov chain monte carlo-toward a" trichromacy" theory of texture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):554–569. (cited on page 5)

Zhu, S. C., Wu, Y., and Mumford, D. (1998). Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126. (cited on page 5)

List of Figures

List of Tables