NORTHWESTERN UNIVERSITY

Visual Texture Analysis for Material Understanding

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Electrical Engineering and Computer Science

By

Dzung Nguyen

EVANSTON, ILLINOIS

March 2018

ABSTRACT

Visual Texture Analysis for Material Understanding

Dzung Nguyen

Texture is an important visual attribute for both human perception and image analysis. It provides useful information for object and scene understanding. However, unlike other computer vision techniques that focus on object shape, texture analysis provides important clues for material understanding. The focus of this thesis is on material identification from texture images. This is a challenging problem that has not received adequate attention. It is important for a variety of applications including surveillance and security, environmental monitoring, agriculture and forestry, health, product design, and sense substitution.

The proposed techniques account for dramatic changes in texture appearance due to variations in illumination and viewing conditions. The key elements of the proposed approach are (1) an adaptation of the Visual Similarity by Progressive Grouping (ViSiProG) procedure for identifying clusters of visually similar textures (2) the characterization of each material with a small set of exemplars; (3) the use of machine learning techniques for

training of Structural Texture Similarity Metrics (STSIMs) to agree with human perception based on clusters from ViSiProG, and as a result, to be able to separate the textures according to the material they correspond to.

Human perception has long been considered as a benchmark for computer vision and image analysis applications. However, current human-based annotation methods mostly deal with application outputs. Instead, we explore the possibilities of capturing human perception at a lower level (visual similarity of images), and utilize a learning framework for metric adaptation.

We demonstrate the effectiveness of the proposed techniques using "CUReT," a fully labeled database of real-world surfaces, viewed under different illuminations and viewing angles at a fixed distance. However, the proposed approaches can be applied to different domains, especially when semantic labeling is not available, for example in an unlabeled database of satellite images, which offers a much wider range of materials, illuminations, and viewing angles, and an unlabeled database of building fronts obtained from "Google Earth Street View."

Acknowledgements

Firstly I would like to express my gratitude and appreciation to professor Thrasyvoulos Pappas for his mentorship, valuable advice and patience throughout my doctoral study. His guidance for research from first principles enables this dissertation. Secondly, I would like to thank my committee members: professor Ying Wu, professor Oliver Cossairt, professor Alan Sahakian, professor Nabil Alshurafa and professor Apostolos Raptis for their feedback, suggestions on the thesis and training at different stages throughout my study. I gratefully acknowledge the funding received towards my PhD from Vietnam Education Foundation Fellowship.

I would also like to thank labmates in professor Pappas's lab and HABits lab who have made research much more collaborative and fruitful: Jana Zujovic, Lulu He, Pubudu Silva, Guoxin Jin, Shengxin Zha, Jing Wang, Jue Lin, Izaiah Wallace, Shibo Zhang, Rawan Alharbi, Runsheng Xu, Zachary King and Lida Zhang. I also want to thank Hoa for being a supportive companion and reminding me of the end goals. Most importantly, no words could describe the unconditional love and support from my family throughout the years; and their encouragement and being role models for pursuing a research career. They have always been there during challenging times for me. This thesis is dedicated to them.

To my parents and sister

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Texture is an important visual attribute for both human perception and scene analysis. Changes in texture provide cues for object boundary detection and localization and foreground/background separation; texture variations can be used to infer object shape; and texture characteristics provide important clues for material identification [**2**]. Even though the importance of all three uses of texture for human perception is obvious, computer vision techniques has mostly focused on object detection and shape, rather than material perception [**3**]. The focus of the thesis is on material identification. Identifying materials from visual texture is important for a variety of applications including surveillance and security, environmental monitoring, forestry and agriculture, construction, health, product quality, virtual reality, as well as sense substitution (visual to acoustic-tactile conversion [**4, 5**], which relies on semantic mapping of visual textures).

While the precise definition of visual texture is highly variable in the literature, several authors (e.g., Portilla and Simoncelli [**6**]) loosely define texture as a spatially homogeneous image that typically contains repeated structures, often with some random variation (e.g., random positions, size, orientations or colors). This definition of texture can apply to materials at different scales. For example, a forest or cityscape can be considered a texture in a satellite image, a concrete or brick wall is a texture in a picture taken from the street, a picture of a cloth at arm's length is a texture, and so is the image of human tissue in a microscope. Texture appearance depends on the intrinsic material properties

(light reflectivity, absorbance, transmittance), the surface geometry (extrinsic material properties), the illumination (direction, spectral composition, polarization, etc.), and the viewing angle. Separating each of these components from the two-dimensional image of a texture, often referred to as "inverse optics," is a difficult, ill-defined, and computationally demanding problem. This is because each of the components can be quite complex, stochastic, and in most cases, unknown or uncertain. An alternative approach, motivated by human perception, is to rely on statistical cues for the determination of material properties. This "ecological optics" approach (see for example [**7, 8**]) is well defined and computationally efficient, which is critical for fast response in a highly dynamic environment (for perception) and for processing large amounts of data (for image analysis). We know from human perception that the ecological approach works most of the time; however, we also know that it can make errors, manifesting themselves as visual illusions, but this only happens in really uncommon and typically contrived situations [**9**].

Since the illumination and viewing conditions can vary dramatically, so does the appearance of the texture that corresponds to a given material. Thus, a simple model or characterization is very difficult to obtain. The approach proposed in this thesis is to characterize each material with a small set of exemplars. A given image can then be classified into a given material if it matches one of the exemplars. Given a large database of materials with different illumination and viewing conditions, we will present techniques for identifying such exemplars. A more challenging problem is to be able to obtain the necessary exemplars without having to collect a large set of images for each material. The goal is to determine the illumination and viewing angles that provide exemplars for robust identification of a new material. Of course, these depend on basic material properties,

such as, whether it is Lambertian, glossy, translucent, smooth, rough, etc. We will discuss how this goal can be attained based on the results of our research.

This thesis research builds on the work of Prof. Pappas's research group on texture similarity [**10**]. Zujovic *et al.* have proposed Structural Texture Similarity Metrics (STSIMs), both grayscale [**11**] and color [**12**]. In contrast to traditional image fidelity metrics, which are sensitive to point-by-point deviations, and thus cannot adequately model the stochastic nature of texture and how it is perceived by humans [**13, 11**], STSIMs allow substantial (visible upon careful observation) point-by-point differences in textured regions that appear virtually the same.

Zujovic *et al.* have also proposed techniques for establishing ground truth for evaluating the performance of similarity metrics in the context of different applications, based on human performance [**11, 1, 14**]. In particular, they have proposed ViSiProG, a Visual Similarity by Progressive Grouping procedure for conducting subjective experiments to organize a texture database into clusters of visually similar images [**1**]. The grouping is based on visual blending and greatly simplifies labeling image pairs as similar or dissimilar. ViSiProG collects subjective data in an efficient and effective manner, so that a relatively large database of textures can be accommodated.

STSIMs have enabled robust and reliable retrieval of textures based on visual similarity [**15, 11, 1**]. However, as has been shown in [**15, 1**], STSIMs can reliably identify textures as similar only if they are similar in every respect. As we discussed, when lighting and viewing conditions change, there can be significant changes in texture appearance. Figure 1.1 shows samples of six materials from the "CUReT" database [**16, 17**], taken under different lighting and viewing conditions. Note that there are significant appear-

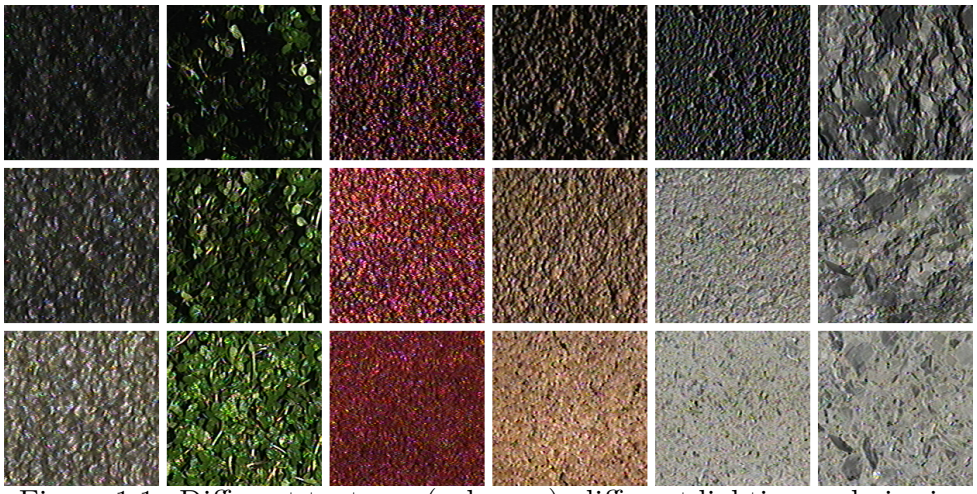

Figure 1.1. Different textures (columns), different lighting and viewing conditions (rows)

ance differences in the samples of each material, which may be comparable to or even greater than the differences between samples of different materials.

To overcome this problem, we need (1) to identify clusters of visually similar textures within each material class; (2) to represent each cluster by an appropriate exemplar; (3) a texture similarity metric that agrees with visual perception. Human vision has been a key guiding principle for our research. This is because it has a remarkable ability for identifying materials based on texture appearance and, as such, provides inspiration, feasibility clues, and performance milestones.

To demonstrate the effectiveness of the proposed techniques, we have primarily relied on the "CUReT" database of real-world surfaces, which offers a range of 61 materials, viewed under different illuminations and viewing angles at a fixed distance [**16, 17**]. However, we have also considered two additional databases. One is a database of satellite images, provided by the Lawrence Livermore National Laboratory, which offers a much wider range of materials, illuminations, and viewing angles. The other is a database of

building fronts obtained from "Google Earth Street View." In the following, we will refer to these databases as CUReT, LLNL, and Street View. The CUReT database is fully semantically labeled, while the other two are not. We will argue that the techniques proposed in this thesis are particularly useful for material extraction from databases that are not fully labeled.

The need for exemplars was demonstrated by Lin and Pappas [unpublished work], who used a simple K-means clustering approach. To characterize each texture, they used a feature vector consisting of statistics computed over a texture image or patch. This is the same feature vector that the STSIMs are based on [**11**], and consists of statistics (mean, variance, horizontal and vertical autocorrelations, and crossband correlations), computed over each subband of a multiscale frequency decomposition, such as the steerable filter decomposition [**18**]. Given a semantically labeled database like CUReT, they used K-means to obtain a number of exemplars for the set of textures that correspond to each material, and showed that there are significant gains in performance as the number of exemplars increase to about 5 or 6, and that the gains diminish as more exemplars are added.

In this thesis we will show that a more elaborate distance metric can do a lot better with fewer exemplars. Given a fully labeled database like CUReT, we rely on metric learning to train STSIMs to separate the textures according to the material they correspond to. Then, we use this metric to obtain K clusters for each material, which we use as the basis for classification.

We then compare the performance of this approach to human perception. Based on our experience, we conjecture, and demonstrate using ViSiProG, that humans are very good

at identifying differences in textures that correspond to different materials; that is, *texture metamerism* (same appearance, different material) is rare. On the other hand, humans are not necessarily good at associating textures with substantial differences in appearance with the same material. Thus, in order to exploit human perception, we rely on ViSiProG to obtain clusters of visually similar textures. As we will see, applying ViSiProG to a large database like CUReT and LLNL requires a number of adaptations. Our results indicate that, with a few exceptions (texture metamerism), these clusters correspond to a given material. We then use metric learning to train STSIMs, and use the trained metrics for finding the cluster exemplars, as well as for matching a query texture to the right exemplar. Of course, if we have a fully labeled database like CUReT, it contains more information than the clusters of similar textures provide, and as expected, metric learning based on the labeled database outperforms ViSiProG-based learning. However, we show that the performance difference is small. More importantly, the ViSiProG-based approach does not require a fully labeled database. For metric training all we need is the clusters of visually similar textures. For classification, we only need to label the cluster centers, which we find using the trained metric. If we assume that each cluster contains textures that belong to the same material, then the cluster centers should be good representatives (exemplars) of that material. For material identification, we match any given texture sample to one of these exemplars.

The problem of material identification under varying viewing and illumination conditions has been studied in the literature. The most elaborate approaches have been proposed by Leung and Malik [**19**], Cula and Dana [**20**], and Varma and Zisserman [**21, 22**]. They all rely on a dictionary of filter response vectors, which they call *textons*

as the basic building blocks, and use a histogram representation for each material. Their results in the context of the "CUReT" database are quite impressive. Our goal is not to compete with these results, but instead, to obtain a more fundamental understanding of the problem and to propose computationally efficient techniques that are easily adaptable to new application domains without the need for extensive training.

The focus of this thesis research is on visual textures. However, it can also be applied to other modalities and wavelengths, such as polarized light, near infrared (NIR), and short-wavelength infrared (SWIR) images. While human perception relies primarily on texture images in the visible range, the analysis of other modalities can be based on the same perceptual principles as the analysis of data in the visible range.

1.1. Main contributions

The main contributions of this thesis are summarized as follows.

- Development of techniques for visual-texture-based material identification that rely on the characterization of each material by a small set of exemplars.
- Modification of ViSiProG to allow faster convergence and higher coverage for large texture datasets.
- Use of ViSiProG for forming clusters of visually similar textures.
- Development of machine learning approaches for training STSIM metrics, based on ViSiProG clusters.
- Use of ViSiProG clusters and trained STSIMs to obtain exemplars for each cluster, and use of those to characterize each material.
- Implementation of ViSiProG as a crowd-sourcing tool for efficient data gathering.

- Contribute to fundamental understanding of human visual perception.

The techniques proposed in this thesis can be applied to different application domains, such as:

- Real-word texture surface identification for product design and quality.
- Geospatial data analysis for national security, environmental monitoring, agriculture, and forestry.

They also contribute to better fundamental understanding of human visual perception.

The thesis is organized as follows. Chapter 2 reviews structural texture similarity metrics. Chapter 3 discusses ViSiProG and adaptations to large scale dataset. Chapter 4 discusses our human perception based approach where a similarity metric is trained from ViSiProG groups, then exemplars are found using k-means clustering. Chapter 5 shows results of our approach, and comparison with original STSIM-M and semantic approaches.

CHAPTER 2

Review of Texture Similarity Metrics

2.1. Texture appearance of materials

Appearance of real-life materials, especially rough surfaces, are known to vary depending on illumination and viewing angles. This is in contrast to ideal Lambertian surface [**23**], where reflected light is constant independent of viewing angles. Texture appearance depends on the intrinsic material properties (light reflectivity, absorbance, transmittance), the surface geometry (extrinsic material properties), the illumination (direction, spectral composition, polarization, etc.), and the viewing angle. For example, a single pixel of a rough surface would contain ridges and facets [**24**]. There would be masking, shadows and interreflection of light within a single pixel.

Bidirectional Reflection Distribution Function (BRDF) is defined as a 4-dimensional function that capture this phenomenon. It is the ratio of reflected radiance divided by incoming irradiance, parameterized by illumination angle (θ_i, ϕ_i) and viewing angle (θ_v, ϕ_v). Here θ and ϕ are polar and azimuth angles with respect to surface normal.

$$BRFD(\theta_i, \phi_i, \theta_v, \phi_v)$$

BRDF could be extended to account for all points on the material instead of a single point. Then it is termed as Bidirection Texture Function (BTF) which is a 6-dimensional function, parameterized not only by illumination and viewing angles but also planar texture coordinates (x, y). BTF is usually represented as a set of images parameterized

by illumination and viewing angles. For example CuReT dataset has 205 images for one material.

$$BTF(x, y, \theta_i, \phi_i, \theta_v, \phi_v)$$

In computer graphics literature, there has been many studies on explicitly modeling BRDF and BTF which could capture large variance of material appearance. One particular model is Oren-Nayar model [**24**] where a surface is modeled as combination of V-cavities with random slopes. This model accounts for masking, shadowing, interreflection and could generate appearance that look realistic, for example a rough vase.

However, these models are usually constrained to a specific type of materials. There is a second, implicit approach to modeling BTF using bag-of-words model [**20**] [**21**] [**22**]. Firstly, a multiscale filterband is applied to all BTF textures. Secondly, k-means clustering is applied on the filter responses to build a dictionary of filter responses, named as textons. This clustering step finds the characteristic textons, and remove redundant ones. After this, each texture is represented as a histogram of texton, based on the filter responses of that texture. Utilizing these histogram representation along with chi-square distance, material of a new unseen texture could be determined with a high accuracy.

2.2. Texture similarity metrics

The main idea behind the development of texture similarity metrics is to give high similarity scores to pairs of textures that have relatively large point-by-point deviations, yet according to human judgment are visually similar or essentially identical. This can be accomplished by replacing point-by-point comparisons with comparisons of region statistics.

Zujovic *et al.* [**15, 12**] have argued that the color composition and the spatial pattern (structure) of a texture are quite separate attributes that should be considered separately, and have shown that this leads to more effective metrics. They also argued that texture structure can be reasonably approximated with its grayscale component.

The basic elements of the grayscale STSIMs [**15, 11**] are the following:

- A real or complex subband decomposition, typically a steerable filter decomposition.
- A set of statistics computed for each image, each subband or pair of subbands, and each window in that subband. Either a local sliding window or a global window (the entire subband) can be used. The statistics typically include the mean, variance, horizontal and vertical autocorrelations, and crossband correlations, and can be computed on the complex subband coefficients or their magnitudes.
- Formulas for computing similarity scores for each pair of corresponding statistics, one from each image. The form that each formula takes depends on the range of values of the particular statistic and may also include a normalization factor.
- A pooling strategy for combining the similarity scores, over statistics, subbands, and window positions, to produce an overall STSIM score.

The two main variations, STSIM-2 and STSIM-M, are presented in detail in [**11**]. STSIM-2 computes the statistics (on a local or global window) and compares images in a similar fashion as CW-SSIM [**25**]. In STSIM-M, each image is represented with a vector of its statistics and the metric computes the dissimilarity between images as the distance between their respective feature vectors normalized by the standard deviation of each component. Zujovic *et al.* [**15, 11, 1**] have shown that these metrics offer significant

Figure 2.1. Operating domains for evaluating performance of texture similarity metrics

improvement over existing methods. We should also point out that the metrics in [**11**] are not scale or rotation invariant. However, they can easily be modified to account for such invariance.

The color composition metrics are based on the dominant colors of the textures and their percentages [**12, 15**]. The dominant color representation was introduced by Ma *et al.* [**26, 27**] and adopted by Mojsilović *et al.* [**28**]. The need for compact color representations was also emphasized in [**29**]. The dominant colors are obtained using adaptive clustering [**30**] to segment the image and then local averaging within each segment class to obtain the colors. The similarity of two textures is then based on the optimal color composition distance (OCCD) [**31**], which is closely related to the earth mover's distance [**29**].

For the systematic performance evaluation of texture similarity metrics, in [**15, 1**], we identified three operating domains: the ability to retrieve "identical" textures; the top of the similarity scale, where a monotonic relationship between metric values and subjective scores is desired; and the ability to distinguish between perceptually similar and dissimilar textures. Figure 2.1 shows examples of identical textures, textures in the monotonic range, similar, and dissimilar textures.

CHAPTER 3

ViSiProG

3.1. Original procedure

Visual Similarity by Progressive Grouping (ViSiProG) [**1**] is a subjective procedure based on human annotation that can be used to form clusters of visually similar images from a texture database. ViSiProG builds similarity groups one at a time, in a step by step fashion. Each user builds multiple groups of typically 9 similar textures. The texture must be similar in every respect: color, brightness, pattern, orientation, and scale. Then, the groups from multiple users are pooled together to form a similarity matrix, which is analyzed using spectral clustering [**32**] to form the final clusters, each of which can have any number of textures.

Figure 3.1 shows the interface of ViSiProG. There are two boxes. The box on the right, the *batch box,* contains a set of textures randomly drawn from the database. The box on the left, the *group box,* is where the user collects a group of similar textures. Initially, the batch box is full and the group box is empty. The user drags a group of images that are most similar to each other from the batch box to the group box, and then presses the *shuffle* button to obtain another batch of textures from the database. The user then refines the group of similar images in the group box by dragging images back and forth between the two boxes. With each iteration, the similarity of the images in the group box should increase. This probability with which the images are drawn from the database

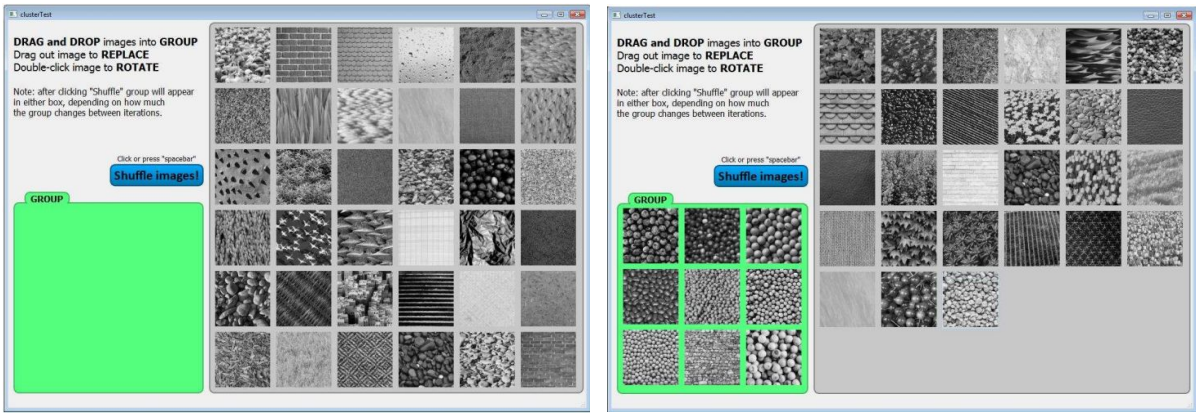

Figure 3.1. Snapshots of the Original ViSiProG interface [**1**]. *Left:* Initially the group box is empty. The user is asked to drag images that appears similar to each other from the batch box into the group box. *Right:* After several iterations, a visually similar group starts to be formed in group box; the user continues to refine the group, until all the images have been presented and the user is satisfied with the similarity of the group.

favor images that have not yet been presented, so that the user can see all the images in the database in resonable time. The system then gives the user the option to stop. However, the user has the option to continue until she/he is satisfied with the similarity of the group.

The relatively small set of images (27) in each batch is dictated by the display resolution and size and by the attention span of the user. The separation of the visually similar images in the group box is a critial part of ViSiProG. Placing images adjacent to each other makes it easy to check the visual similarity. The borders separating the images are also important, as they facilitate visual blending (window effect).

3.1.1. Detailed ViSiProG procedure

In the original ViSiProG setup, the number of images in the batch box is $N_b = 36$, with 27 new images presented each time. The number of images in the group box is $N_g = 9$. The original setup was designed and tested with a database of about $N = 500$ images.

The subset of images shown in the batch box is determined by random sampling with replacement. The initial probability of selecting each image is $1/N$. After the user presses the shuffle button to get another batch, the probability of the images that remained in the batch box after the prior selection is reduced by a factor of 4. Then, a new batch is drawn with the updated probabilities, and the procedure is repeated.

At any point, the user is free to change the group completely by dragging images between the group box and the batch box. Thus, the user is may choose to refine the current group or to start forming another entirely different group. To facilitate the latter case, the original procedure includes a reset feature: if during an iteration the group changes by more than 50%, then all images in group box are moved to the batch box after the shuffle button is pressed.

When all images from the dataset have been displayed at least once, a *Submit* button appears. If the user is content with the similarity of the group, she/he can press the Submit button to finalize the group and exit the test. Otherwise the user can continue refining the group. We use the term *trial* to refer to one group formation, from the initial presentation to the time that the submit button is pressed. Each user performs multiple trials, and there is no limit on duration of each ViSiProG trial.

3.1.2. ViSiProG algorithms

Data from multiple subjects are pooled together into a similarity matrix D. It has dimension N by N, and each entry D_{ij} is the number of ViSiProG groups that image i and image j belong together. Then spectral clustering algorithm [**32**] is applied on this matrix to obtain clusters of visually similar images. We notated these cluster as ViSiProG clusters. While ViSiProG groups has a fixed number of images ($N_g = 9$), ViSiProG clusters has variable number of images depending on the clustering algorithm.

3.2. ViSiProg adaptations for the material identification problem

In this section, we describe the limitations of the original ViSiProG procedure, and adaptations that are needed for material identification in a large database, e.g., with $N > 5000$.

3.2.1. Saturation

The most important limitation that we found is that the percentage of images in the database that have been placed in at lease one group, which we define as ***coverage***, does not increase linearly with the number of trials and saturates at some point. Figure 3.2 shows that saturation is reached at approximately 25% after 200 trials. Thus, the number of trials to reach a desired coverage grows exponentialy, deeming it infeasible for a larger scale database.

A possible explanation for this observation is that the user tends to select the most salient groups of similar images, ignoring similarity groups that are more difficult to find.

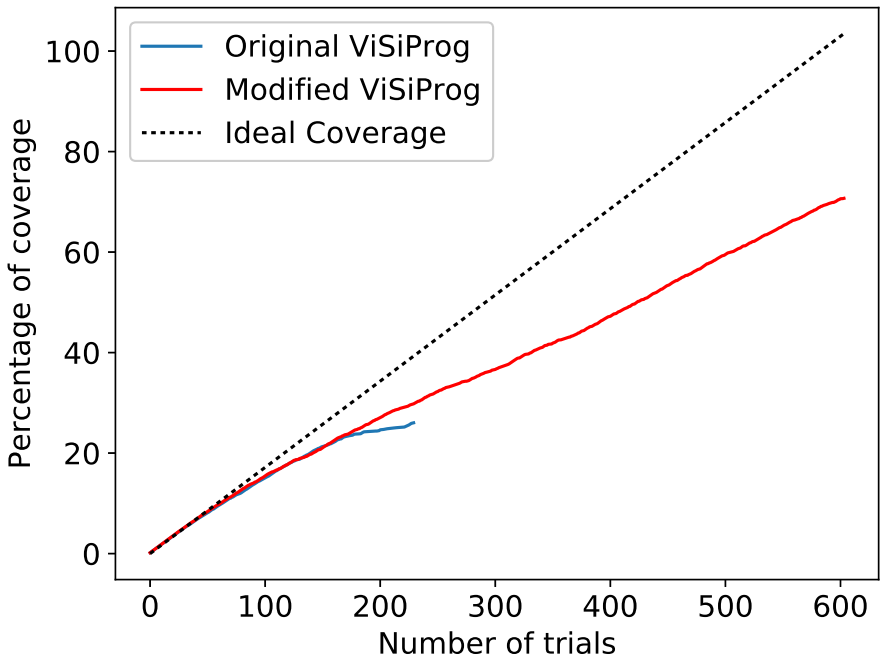

Figure 3.2. Coverage rate: Original ViSiProG saturates when reaching coverage of 25%. Modified ViSiProG increases linearly, reaching coverage over 70% after 600 trials.

Once some groups are formed, there are no constraints to force the user to avoid selecting images from these groups as the core for forming groups in subsequent trials.

To mitigate this problem, we modified the initial selection of the probability of each image. Instead of equalizing the initial probabilities, we make them proportional to $1/n$ where n is the number of ViSiProG groups that the image belongs to. Thus, at the first iteration, the not yet labeled images are more likely to be shown to the user, forcing her/him away from groups that have already been formed. The number n should be calculated for each user, but as we will discuss below, in order to expedite things, in the tests we conducted with three experienced users, we calculated n over all users. Figure 3.2 shows the linear coverage curve of the modified ViSiProG procedure.

3.2.2. Interface optimization

To further expedite the group formation, we increased the number of images in the batch from $N_b = 36$ is increased from 36 (6 rows by 6 columns) to $N_b = 54$ (6 rows by 9 columns). We found that the users have no difficulty spotting similar textures out of the larger batch. More significantly, we allowed the user to terminate the trial after only 50% of the images were presented (instead of 100%). That is, the Submit button is shown when the user has seen 50% of the database in a given trial. This is because in a large database, it is very likely that a group of similar textures will be found long before the entire database has been scanned. Figure 3.3 shows the optimized interface.

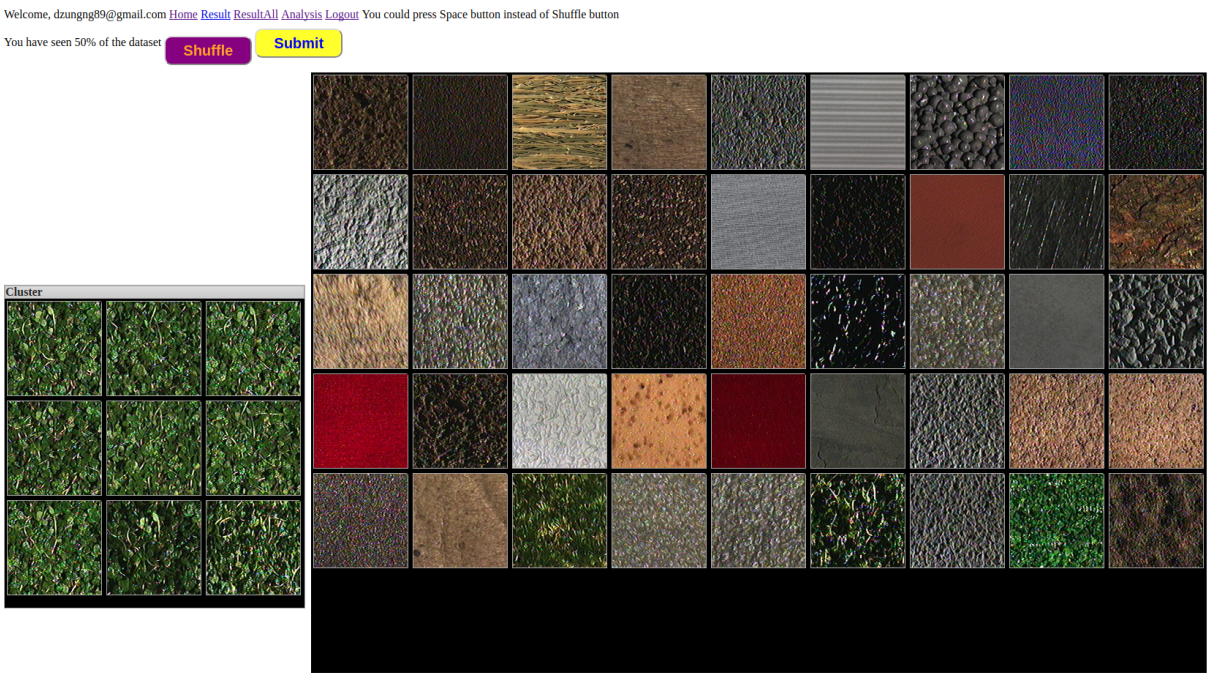

Figure 3.3. Modified ViSiProg procedure which has larger number of images in the batch N_b, earlier stopping criterion at 50% of N, as well as different initial probabilities

Finally, ViSiProG was implemented as a web-based interface rather than a desktop application. The advantages are two-fold: multiple users can label data concurrently from different locations, and initial probabilities can be calculated automatically in real-time across users.

CHAPTER 4

Learning image similarity metric from ViSiProG

In this section, we discuss the training of image similarity metrics based on similarity clusters obtained via ViSiProG. The key advantage of this approach is that it bypasses the need for a fully semantically labeled database. All that is needed is to associate semantic labels with each similarity cluster, provided that texture metamerism is rare in the image dataset. First, we discuss how training based on ViSiProg clusters differs from traditional techniques in Section 4.1, and then we will present a detailed training framework in Section 4.2. The trained metric can be used to obtain exemplars, then to map a querry texture into one of the exemplars for material identification (Section 4.3). In the rest of this thesis, we will differentiate between *semantic labels* (material classification) and *ViSiProG labels* (allocation to a particular cluster or group obtained via ViSiProG).

4.1. Properties of training from ViSiProG clusters

Figure 4.1 illustrates several ViSiProG clusters or groups along with the corresponding semantic labels.

The difference between metric learning from ViSiProG labels versus lerning from semantic labels consists of the following:

a) Semantic labels are *absolute* as each texture is assigned to a specific material. On the other hand, ViSiProG labels are *relative* since they represents the visual similarity between textures.

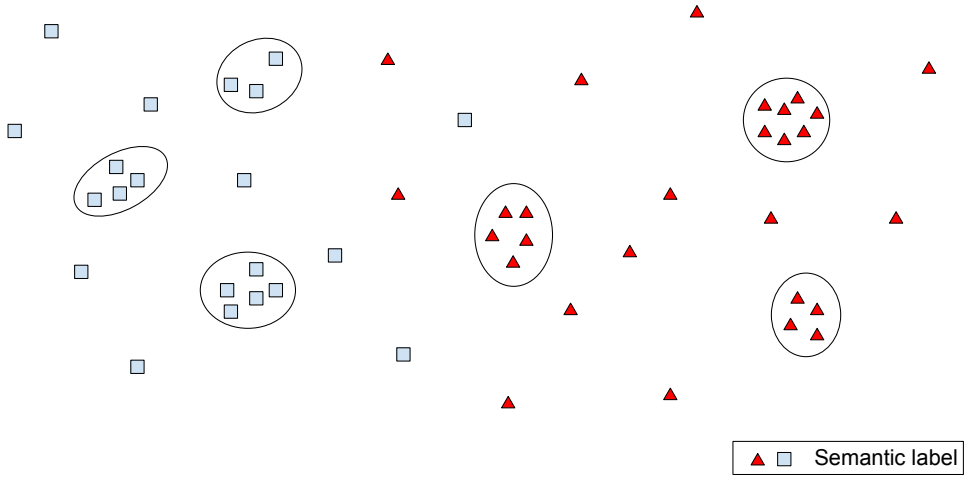

Figure 4.1. Training from ViSiProG groups: *a)* Semantic labels are absolute, while ViSiProG labels are relative *b)* Different ViSiProG clusters/groups may be similar to each other *c)* ViSiProG labels require semi-supervised training *d)* Variance of ViSiProG clusters/groups is much smaller than the variance of semantic classes.

b) Textures with different semantic labels are different. On the other hand, textures in different ViSiProG clusters or groups may be similar to each other, as one material typically contains multiple ViSiProG clusters. ViSiProG clusters, in particular may be very similar to each other, as a cluster is typically constructed (via spectral clustering) from several groups.

c) Semantic labels are provided for the whole dataset, while ViSiProG labels only contains part of the dataset where images which are visually similar. The remaining images in the dataset are tagged as ***unlabeled***. A metric trained from ViSiProG labels only (without utilizing unlabeled data) would not generalize well into the similar/dissimilar operating domain.

d) The variance of a ViSiProg group is guaranteed to be small since the textures are visually similar. The within variance of ViSiProG clusters is typically small, but

larger than the variance of the groups. On the other hand, a semantic category contains all the textures that correpspond to a material, and given the variations in viewing conditions, it is expected to have much higher intra-class variability.

One particular method that can address all of the above attributes is the Local Fisher Discriminant Analysis (LFDA) [**33**]. First, LFDA minimizes the within class variance of each ViSiProG group, thus guarantees that trained metric will produce small variance for visually similar groups. Second, instead of discriminating between every pair of different ViSiProG groups, LFDA maximizes the total difference between clusters and thus allowing some groups to be close to each other. Last but not least, the modified LFDA approach utilizes unlabeled data, thus it learns from all image pairs in the dataset.

4.2. Metric learning

4.2.1. Model

As we saw in Section 2.2, each texture can be characterized by a feature vector that consists of statistics computed over a texture image or patch. The statistics include the mean, variance, horizontal and vertical autocorrelations, and crossband correlations, computed over each subband of a multiscale frequency decomposition, using steerable filter decomposition [**18**]. Here we use a 3-scale, 4-orientation steerable filter decomposition for a total of 82 statistics [**11**]. We also use Mahalanobis distance between the 82-dimensional feature vectors $f(x)$ and $f(y)$ of two images x, y:

$$d(x, y) = f(x)^T M f(y)$$

where M is a positive semidefinite matrix. An equivalent formulation is a linear mapping of the feature vector $f(x)$ into a new space $Lf(x)$, where L is the Cholesky

decomposition of M: $M = L^T L$. This linear representation is widely used in the metric learning literature [**34**].

4.2.2. Baseline metric: STSIM-M

We will use the STSIM-M [**11**] as a baseline method that requires minimal training. As we saw in Section 2.2, in STSIM-M, the matrix M is diagonal that contains the variance of each statistic:

$$M = dig(\sigma_1, \sigma_2, \ldots, \sigma_N)$$

where σ_i is variance of the i-th component of the feature vector $f()$ over entire database (training set).

4.2.3. Local Fisher Discriminant Analysis

Scatter matrix of whole dataset is computed as follows:

$$S_T = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^T (x_i - \mu)$$

where x_i the feature vector of the ith element of the dataset, and μ is average of entire dataset. The within class scatter matrix is computed across all clusters as follows:

$$S_W = \frac{1}{N_C} sum_{c \in C} \sum_{i:l_i=c} (x_i - \mu_c)^T (x_i - \mu_c)$$

where x_i the feature vector of the ith element of the dataset, l_i is its ViSiProg cluster/group label, c is the index of ViSiProG cluster/group, μ_c is the average of the feature vectors in group c, and C is the collection of all the ViSiProG clusters/groups. The between class scatter matrix is computed as:

$$S_B = S_T - S_W$$

An equivalent formulation computed pairwise is:

$$S_W = \sum_{i,j=1}^{N} A_{i,j}^{(W)} (x_i - x_j)^T (x_i - x_j)$$

$$S_T = \frac{1}{N} \sum_{i,j=1}^{N} (x_i - x_j)^T (x_i - x_j)$$

$$S_B = S_T - S_W$$

where $A_{i,j}^{(W)} = 1/n_c$ if x_i and x_j belongs to the same group g, and 0 otherwise.

To preserve locality, Sugiyama *et al.* [**33**] weighted coefficients $A_{i,j}^{(W)}$ by an affinity matrix $\bar{A}_{i,j}^{(W)}$ to reduce the influence of far apart data points on variance. $\bar{A}_{i,j}^{(W)}$ is 1 if x_j is nearest neighbor of x_i, and 0 otherwise. This enables learning the metric from multimodal clusters. This is particularly important when we train the metric on ViSiProG clusters, since clusters usually consists of multiple groups.

The optimization objective function is:

$$\arg\max_L tr((L^T S_W L)^{-1}(L^T S_B L))$$

The optimal solution to this problem can be found in closed form as general eigenvalues of the following equation:

$$\sigma_B L = \lambda \sigma_W L$$

4.2.4. Inclusion of unlabeled data

When applying LFDA for metric learning based on the ViSiProG clusters/groups, we use the same calculation of within class variance σ_W. However, σ_T is calculated on the

whole dataset (combining data from ViSiProG labeled and unlabeled). This new model accounts for all of the requirements stated at section 4.1.

4.3. Finding exemplars

Given the trained metric, the material of an unseen texture can be found using nearest neighbor approach. We compare then feature vector of the unseen texture with the exemplars obtained from the training dataset, and return the material of the one with smallest distance.

To find these characteristic exemplars, we use k-means clustering algorithm in the trained metric space. Since the trained metric can discriminate between different materials, we would need just a few exemplars to represent one material.

CHAPTER 5

Results

5.1. Databases

In the development and testing of the proposed techniques, we have primarily relied on the CUReT database, which is fully semantically labeled. However, the techniques proposed in this thesis are particularly useful for material extraction from databases that are not fully labeled, such as the LLNL and the Street View Databases, the analysis of which we discuss briefly at the end of this chapter.

5.1.1. CUReT

The CUReT database contains images of 61 samples of materials that include plaster, concrete, pebbles, rugs, cloth, wood, etc. [**16, 17**]. The texture images have been captured with different combinations of illumination and viewing directions, at a distance of 2 meters from the sample. We should point out that CUReT includes several images that contain severe image aliasing. We were able to restore the textures of three materials (44, 33, 46), in which the aliasing was primarily in the chrominance component. To restore the textures, we extracted the grayscale component, and interpolated the chrominance from adjacent texture samples. However, we had to eliminate several samples of texture 05, as well as a few textures that could not be restored. Textures 02, 06, and 29 appear to be aliased but look like natural textures, so we included them in our experiments. Finally,

we also eliminated three classes of textures that were not uniform (peacock, orange peel, leaf). In addition, they are very easy for humans to identify, which makes it trivial to group in ViSiProG.

5.1.2. LLNL

LLNL is a database of satellite images, provided by the Lawrence Livermore National Laboratory, which offers a much wider range of materials. Here, the illumination conditions (angle, time of day, weather conditions) and viewing angles are variable but not under direct control. The only controlled parameter is scale.

5.1.3. Street View

Street View is a database of building fronts obtained from "Google Earth Street View." Here, again, the illumination and viewing conditions vary and are not under direct control, while the scale has limited variation that depends on the distance of the buildings or other structures from the street.

5.2. Data collection with adapted ViSiProG

We collected data using a web-based ViSiProG interface. We relied primarily on three experienced subjects (DTN, TNP, TNL) but also collected some data from 7 additional users, both local and remote. Using the adapted ViSiProG procedure, we were able to achieve 70.7% coverage with a total of 606 groups.

Figures 5.1 and 5.2 show sample groups from materials 03, 08, 18, and 50. Note that we selected groups with significantly different appearance, however, there are multiple groups that are similar to each other. Figure 5.3 shows examples of metamerism. The

first column includes two scales of the same material (rough paper), the second includes pebbles and stones, and the third, rough tile and rough paper, the latter with illumination and viewing azimuthal angles. It is thus understandable that these could be confused.

Figure 5.4 shows the distribution of number of ViSiProG groups per material. The red color bar indicates groups that contain images from different materials, that is, they are examples of texture metamerism. Note that there are no groups for materials 23, 55 and 57, which we removed as we discussed above.

We then pooled the groups from by the different users together to form a similarity matrix, and applied spectral clustering [**32**] to form the final clusters. Note that each of the clusters can have any number of textures. After obtaining lower dimension space through spectral clustering, DBScan algorithm [**35**] is applied to form clusters.

Figures 5.5 and 5.6 show three clusters that correspond to material 18. Figures 5.7 and 5.8 show three clusters that correspond to material 50. Note that the clusters are quit cohesive and distinct from each other.

Figure 5.9 shows the distribution of number of ViSiProG clusters per material. Many textures from material 14 belongs to metamerism groups, thus spectral clustering algorithm shifted clusters from material 14 to metamerism case (material 0).

5.3. ViSiProG cluster and BTF

In this section, we analyze ViSiProG clusters with respect to bidirectional texture function. Concretely, utilizing ViSiProG clusters as perceptual ground truth, we find whether visually similarity could be predicted from illumination or viewing angles. Our hypothesis is that given a pair of illumination-viewing angles, a second pair with either

nearest illumination **or** viewing angles would be visually similar to the first one. One possible edges cases could be convex-concave visual illusion [**36**], where if either illumination or viewing angles is flipped around surface normal texture would appear still the same.

Figure 5.10 shows illumination angles of CUReT dataset. Since the angles are not sampled uniformly (denser sampling at the top of the sphere), absolute angle difference would not be a good measure of adjacency between illumination angles. Instead, we apply spherical Voronoi partition on all illumination points, and consider number of adjacent Voronoi subset as distance between two illumination angles. For example, two illumination angles that belong to two adjacent Voronoi partitions will have distance 1. This distance is termed as ***adjacency distance***. Figure 5.11 shows the Voronoi partition. Similar representation is applied to viewing angles.

Our hypothesis states that within a ViSiProG cluster, there exists a path between textures where adjacent textures exhibit only gradual changes in either illumination or viewing angles between textures. We represented this path as a minimum spanning tree of adjacency distances. We found the average distance over all ViSiProG clusters is 1.33. This shows that our hypothesis is correct, and there is no convex-concave illusion in ViSiProG clusters.

5.4. Material identification

The performance of the proposed approaches has been evaluated in two scenarios: (1) Within material validation, where we split the textures of each material for training and testing, and (2) across material validation where we split the materials for training and testing. We compare the performance of four grayscale metrics:

- STSIM-M: This is the STSIM-M metric proposed in [**11**], where the matrix M is diagonal with values equal to the variance of the corresponding statistic over the entire dataset (no labels).
- Semantic: M is trained using the semantic labels with the original LFDA approach.
- ViSiProG with LFDA: M is trained using only the ViSiProG labeled textures (groups or clusters) with the original LFDA approach. That is, the unlabeled data is not utilized.
- ViSiProG with modified LFDA: M is trained using the the ViSiProG labeled textures (groups or clusters) as well as the unlabeled data with the modifed LFDA approach.

5.4.1. Within material validation

For within material validations, for training, we randomly selected 80% of the textures that correspond to each material in the CUReT database, and for testing, the remaining 20%. We ran this procedure 10 times with different random selections and averaged the results (5-fold cross-validation).

First, the metric is learned from the training data. Then, we apply K-means clustering with the trained metric within each material (all the textures in a class for the semantic approach, and all the labeled data for the ViSiProG-based approaches) to obtain a fixed number of K examplars for each material. Finally, the trained metric and the exemplars are used to classify each texture in the test set. We used *precision@1* [**37**] as the performance measure.

Across Material Splitting – KNN			
	Semantic	ViSiProG	
STSIM-M	LFDA	LFDA	mod. LFDA
0.914	0.992	0.975	0.993

Table 5.1. Classification performance across material

The performance results are shown in Figures 5.12 and 5.13 as a function of the number of exemplars per material. Figure 5.12 is based on the ViSiProG group labels, while Figure 5.13 is based on the ViSiProG cluster labels. As expected, in both cases, the best performance is obtained with the semantically trained metric, and the worst performance is obtained when the covariance matrix is used without any labels or similarity clusters. However, the performance of the ViSiProG with modified LFDA approach is close to the semantic approach, and better than the ViSiProG with the original LFDA approach. In addition, when color similarity is included, performance will improve further.

5.4.2. Across material validation

In across material validation, we evaluate whether a metric trained from a subset of materials can work well on new, unseen materials. We randomly selected half of the materials for training and used the other half for testing. Since the labels, and hence the exemplars for the new materials are not known, we used the 1-nearest-neighbor approach to test metric performance. That is, for each texture in the test set, we compared it with the remaining textures, and checked if the nearest one belongs to the same material. We run this evaluation over 10 different random splits of the data and averages the results, which are shown in Table 5.1. Note that the performance of the semantically trained metric and the ViSiProG-based modified learning approach are about the same.

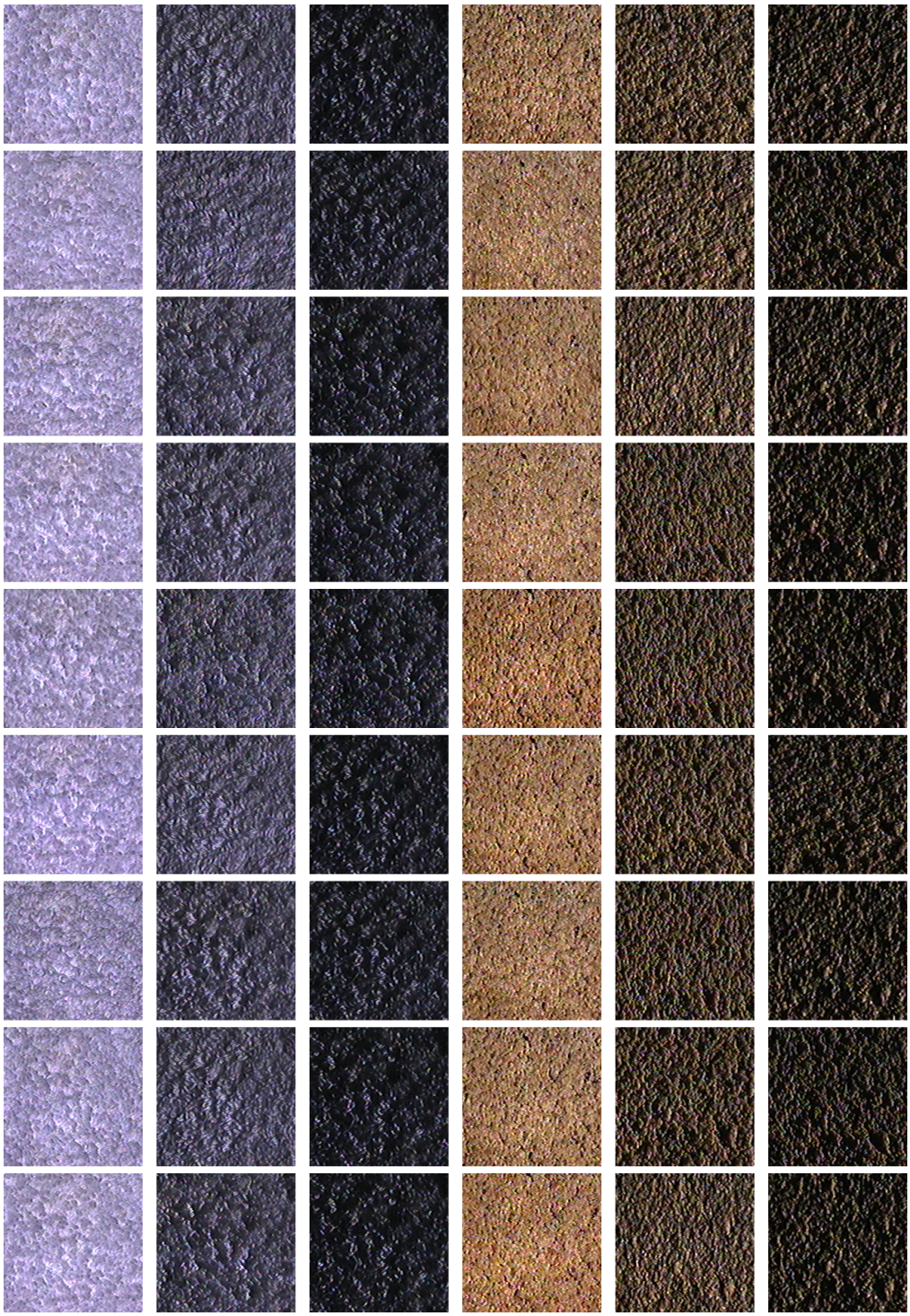

Figure 5.1. ViSiProG groups that correspond to materials 03 (3 left columns) and 08 (3 right columns)

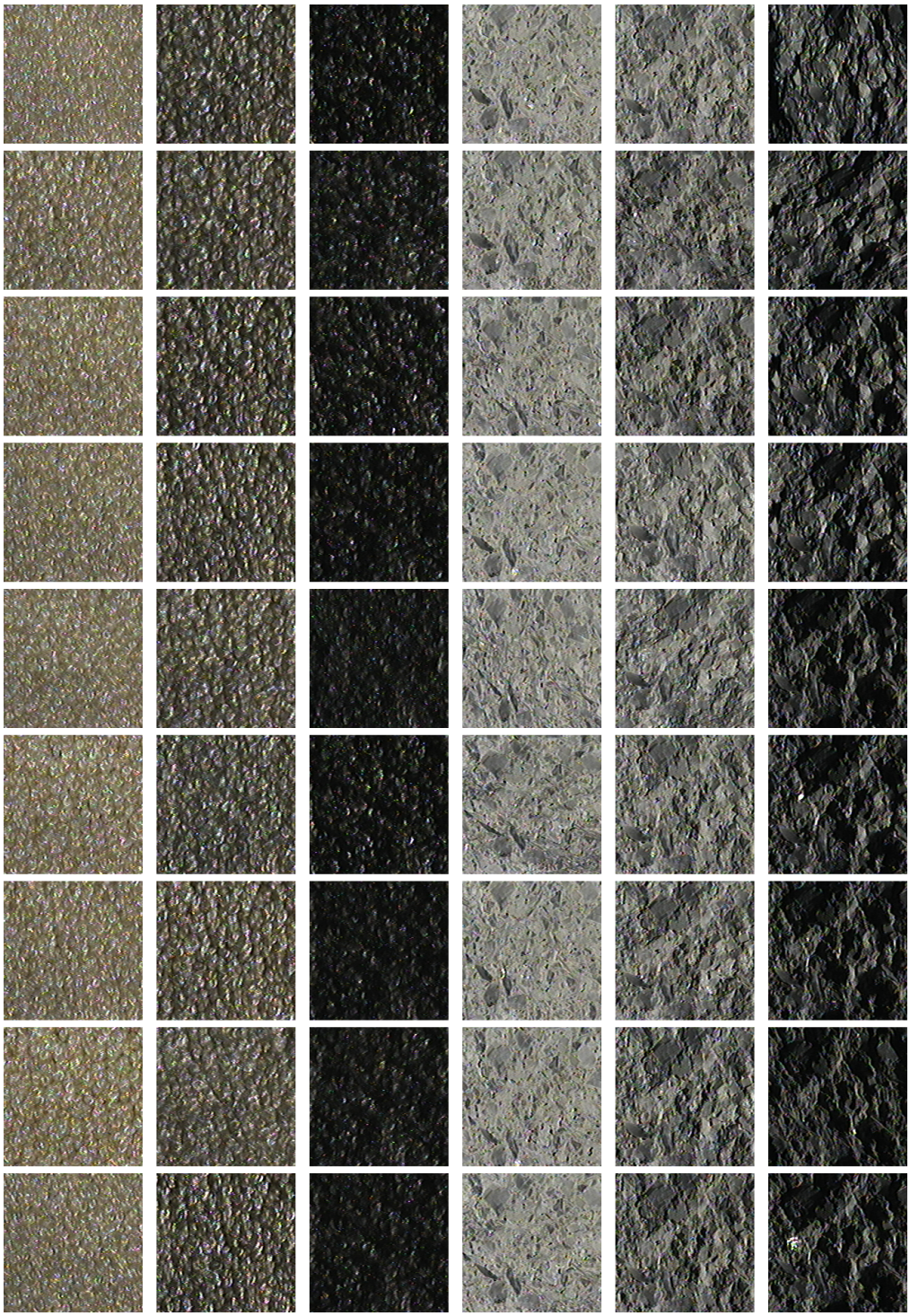

Figure 5.2. ViSiProG groups that correspond to materials 18 (3 left columns) and 50 (3 right columns)

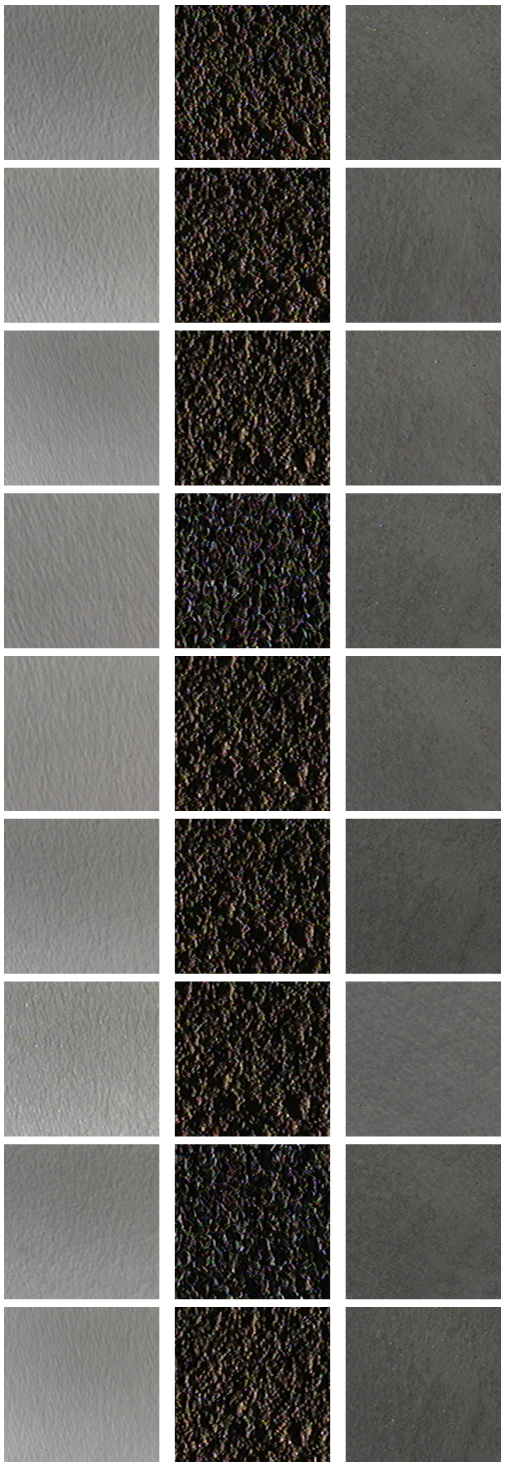

Figure 5.3. ViSiProG groups that correspond to metamerism

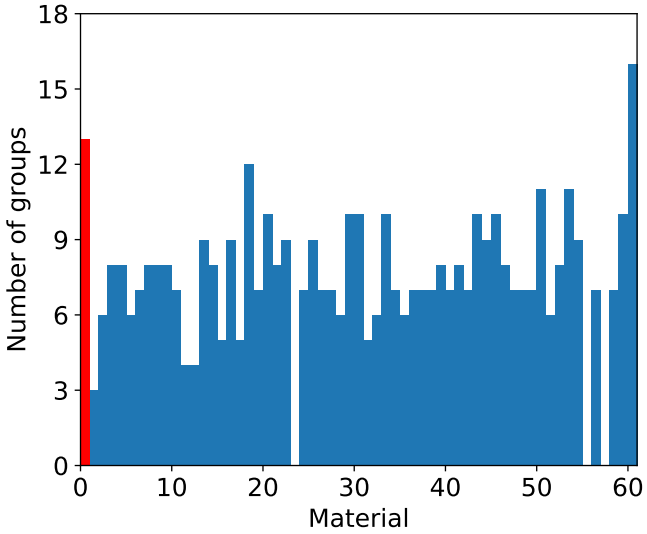

Figure 5.4. Histogram of number of groups in each material. The red bar indicates metamerism.

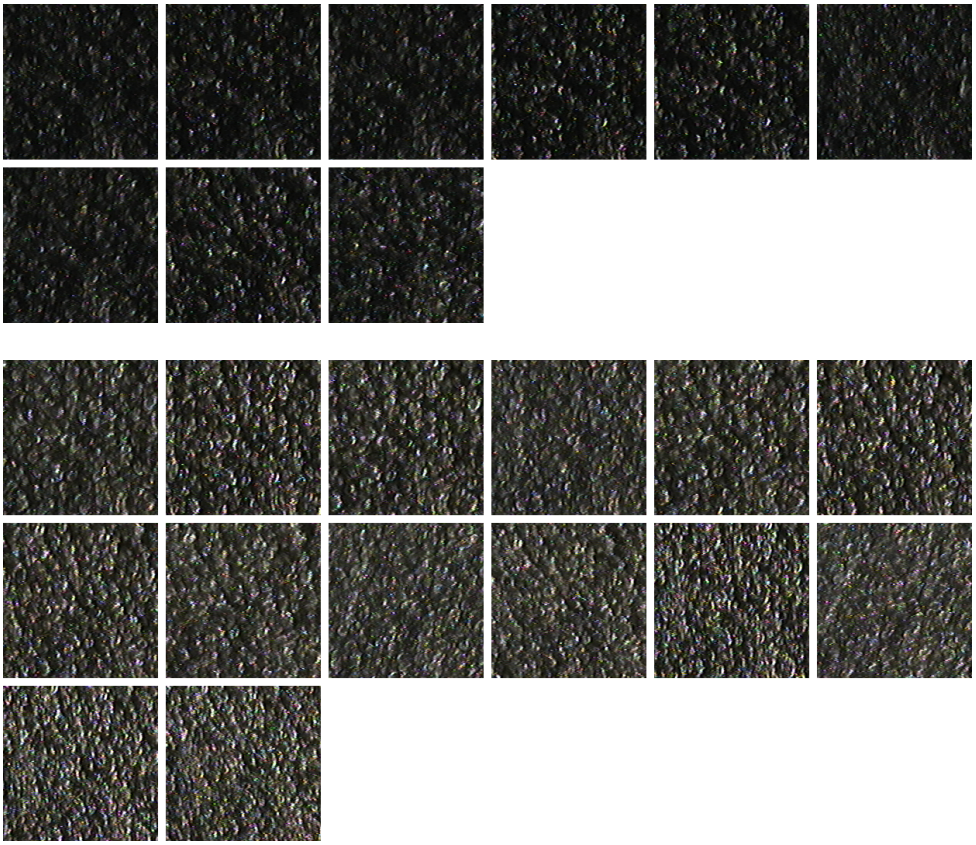

Figure 5.5. ViSiProG cluster 1 and 2 (out of 3) that correspond to material 18

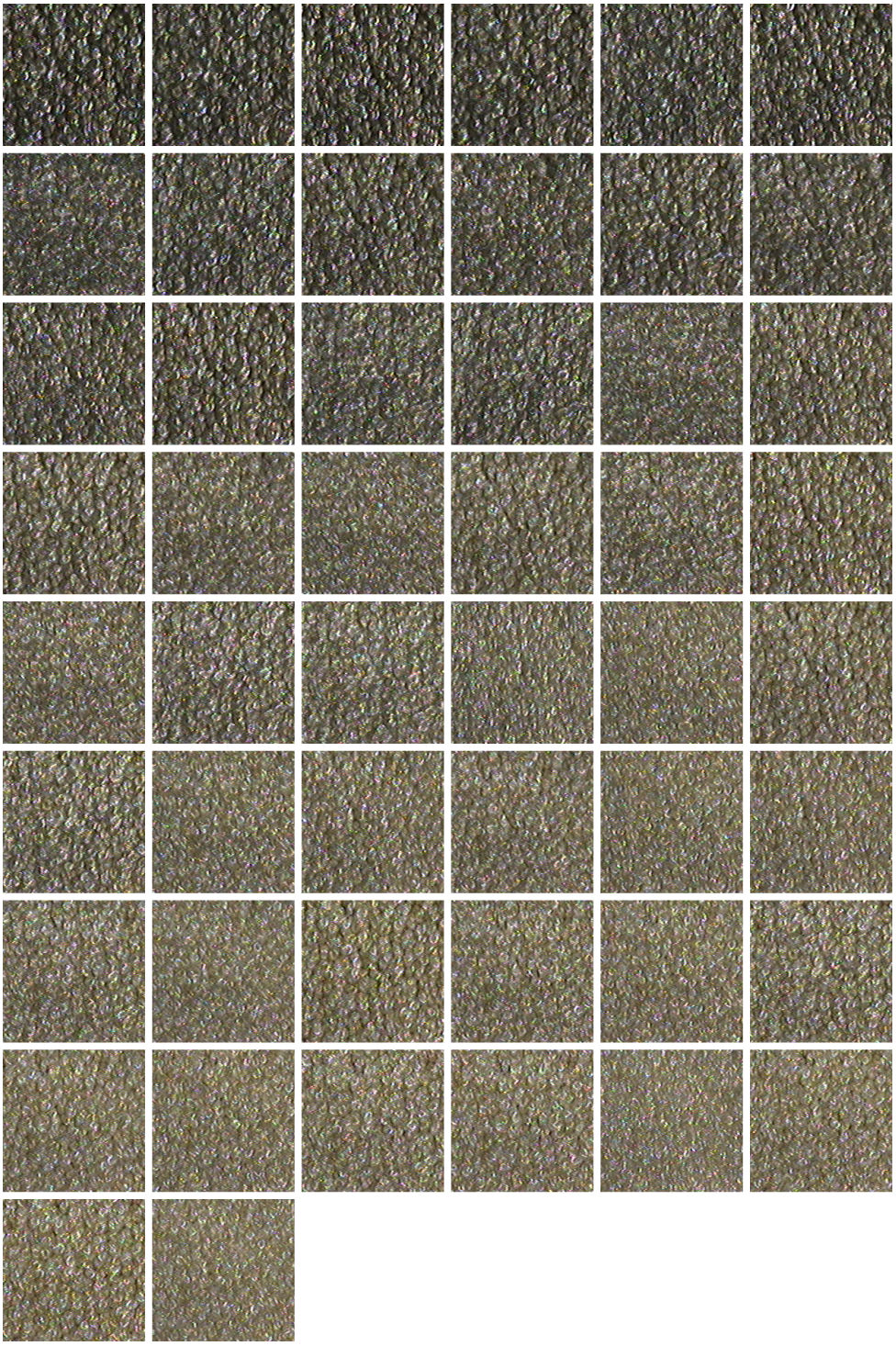

Figure 5.6. ViSiProG cluster 3 (out of 3) that corresponds to material 18

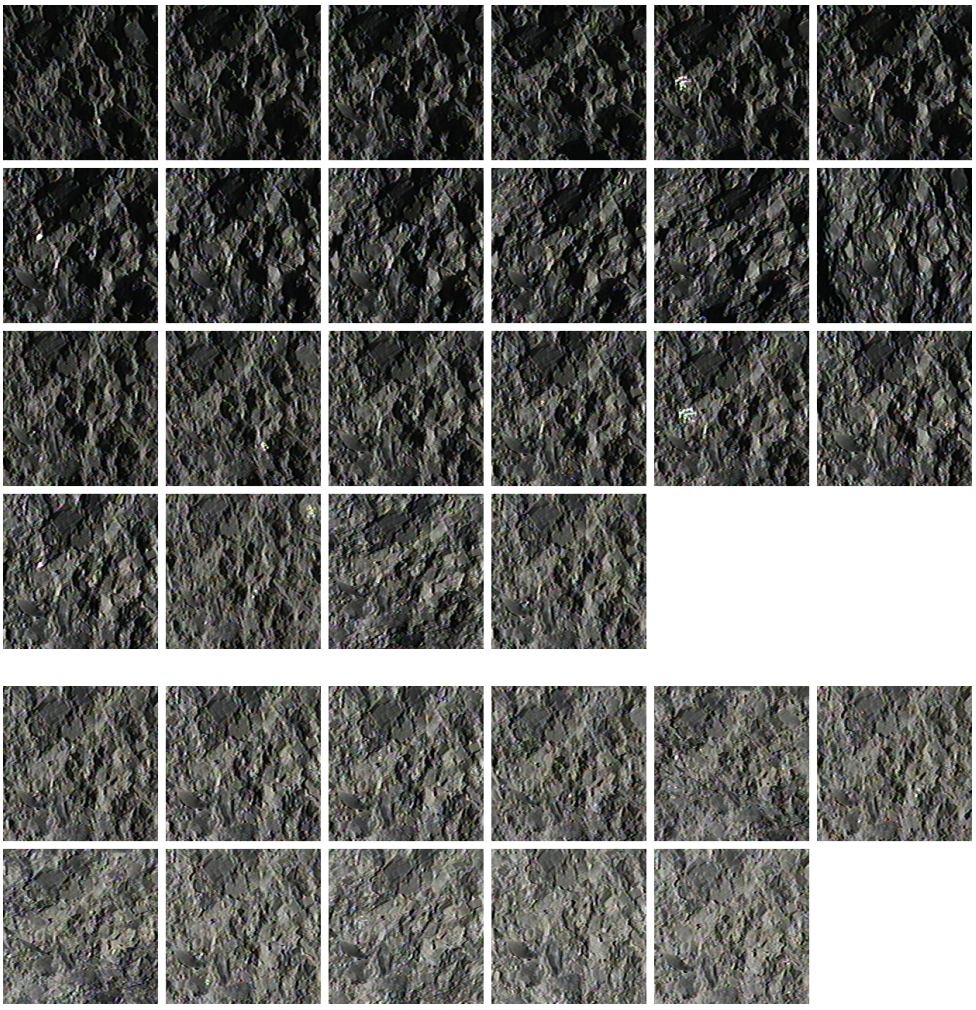

Figure 5.7. ViSiProG cluster 1 and 2 (out of 3) that correspond to material 50

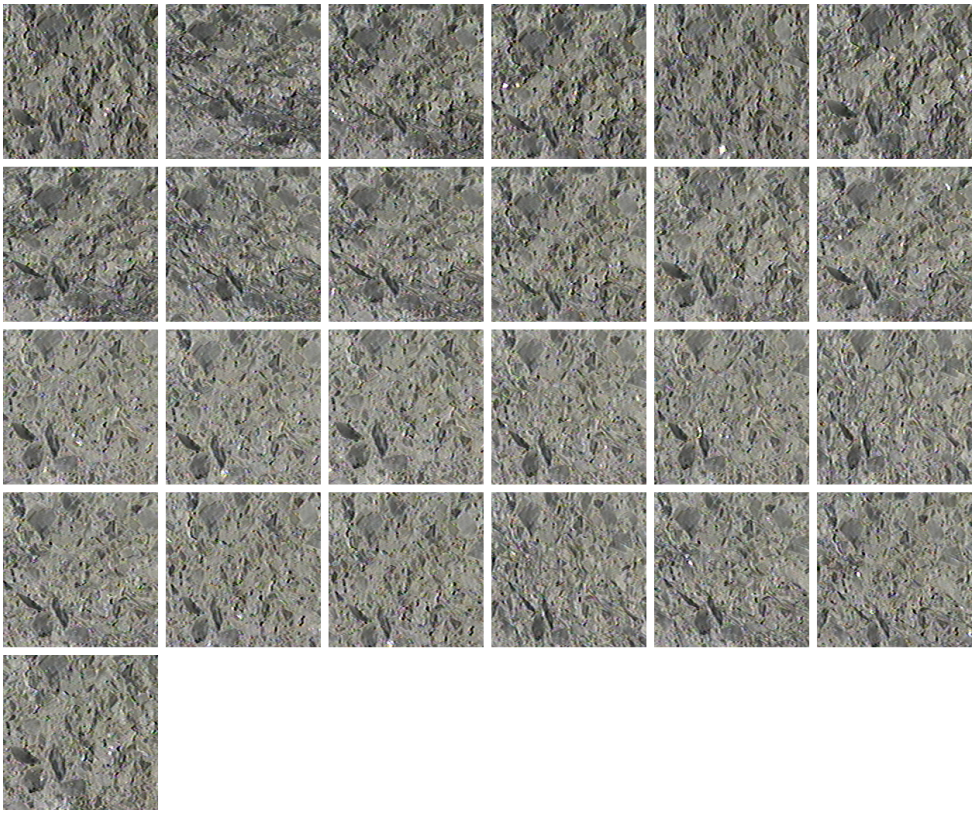

Figure 5.8. ViSiProG cluster 3 (out of 3) that corresponds to material 50

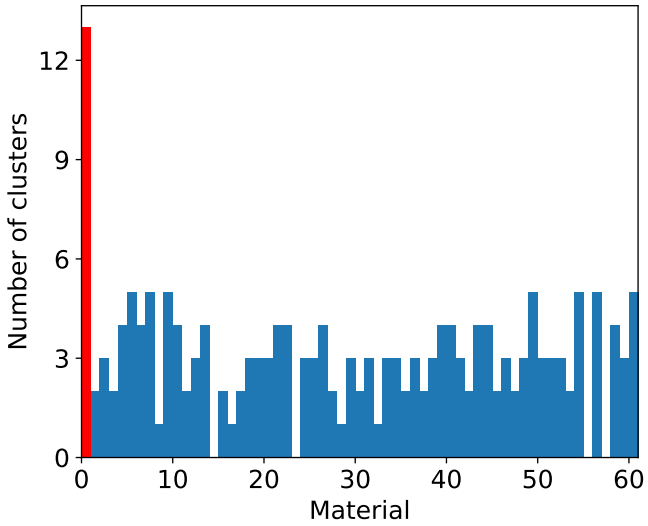

Figure 5.9. Histogram of number of clusters in each material. The red bar indicates metamerism.

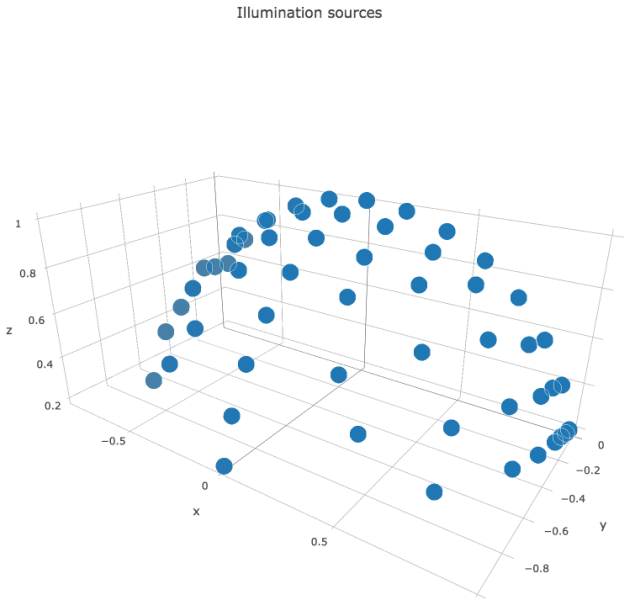

Figure 5.10. All Illumination angles in CUReT dataset for all materials

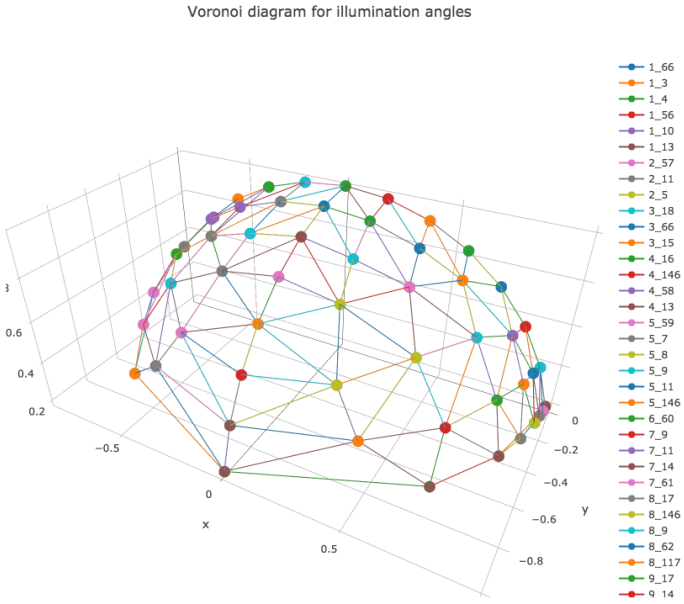

Figure 5.11. Voronoi partition to represent adjacency between angles. Distance between two angles are defined as shortest number of Voronoi subset to reach between them

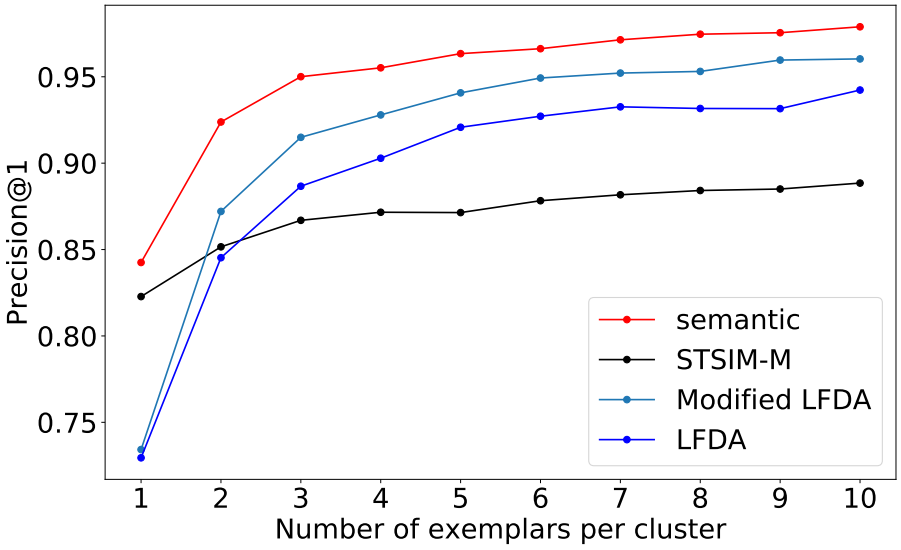

Figure 5.12. Precision@1 of different metrics (LFDA and Modified LFDA are trained from ViSiProG groups)

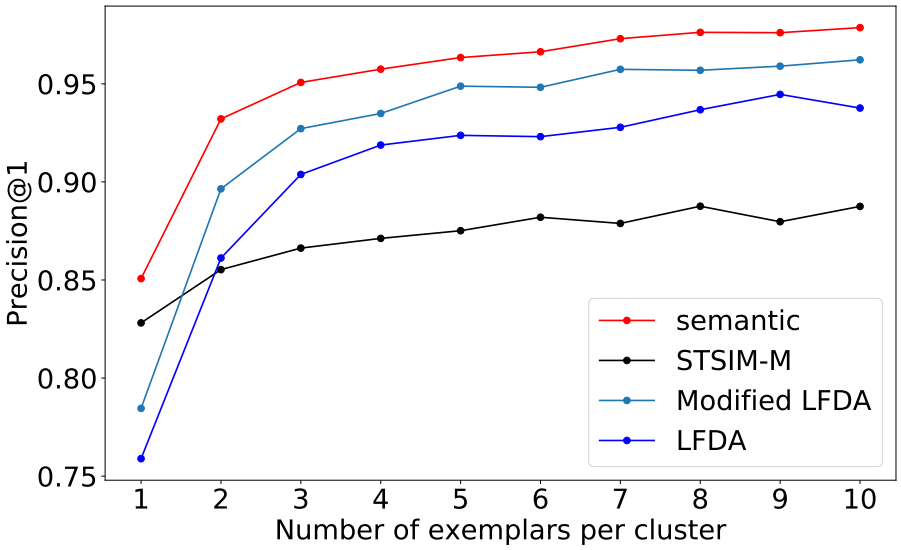

Figure 5.13. Precision@1 of different metrics (LFDA and Modified LFDA are trained from ViSiProG clusters)

CHAPTER 6

Conclusions and future work

6.1. Conclusions

We proposed a new framework for identifying materials from textures, which consists of (1) finding visually similar clusters using ViSiProG, a Visual Similarity by Progressive Grouping procedure; (2) using Local Fisher Discriminant Analysis (LFDA) to train a discriminative texture similarity metric; and (3) representing each material with a few exemplars in the trained metric space.

We first identified limitations of the original ViSiProG procedure when being applied to a large scale dataset. The main issues were the large number of textures that the users have to search through, and the large number of groups that is needed to characterize the materials in the database. In the original ViSiProG procedure users tend to form groups with a small subset of the most salient textures, leading to coverage saturation. These limitations were addressed through optimization of user interface and modification of initial probabilities based on the number of ViSiProG groups each image has been placed in by a particular user or all the users.

We then proposed a metric learning framework based on LFDA for training structural texture similarity metrics (STSIMs) based on the ViSiProG clusters. Experimental results show that the trained metrics consistently outperform the original untrained metrics. In addition, we incorporated data not labeled by ViSiProG in the training procedure, via

the calculation of the total variance. The experiment results demostrate the performance advantages of utilizing the unlabeled data.

The combination of adapted ViSiProG subjective test and metric learning framework provides a complete solution for applications where semantic labels are limited or completely unavailable.

There are numerous applications of this work in geospatial image analysis, environmental monitoring, surveillance and security, forestry and agriculture, construction, health, product quality, and virtual reality.

6.2. Future work

There are many extension of this work that can be pursued in the future. Firstly, the ViSiProG test can be implemented in a crowdsourcing platform. This will enable rapid data collection, resulting in a wider variety of ViSiProG groups. In turn, this would enable the use of multidimensional scaling (rather than spectral clustering), which will provide a visualization of the structure of the texture space. However, crowdsourcing wwil require an independent review mechanism to guarantee the quality of collected groups. Secondly, this framework can be applied to applications where limited data labeling is available, such as geospatial image analysis and national security, environmental monitoring, agriculture, and forestry. Finally, an important application is to "Street View" images, for automatic identification of the materials of each building in a city, thus providing useful information for city planning, fire prevention, and security.

References

[1] Jana Zujovic, Thrasyvoulos N. Pappas, David L. Neuhoff, Rene van Egmond, and Huib de Ridder. Effective and efficient subjective testing of texture similarity metrics. *Journal of the Optical Society of America A*, 32(2):329–342, February 2015.

[2] Michael S. Landy and Norma Graham. Visual perception of texture. In L. M. Chalupa and J. S. Werner, editors, *The Visual Neurosciences*, pages 1106–1118. MIT Press, Cambridge, MA, 2004.

[3] Edward H. Adelson. On seeing stuff: The perception of materials by humans and machines. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors, *Human Vision and Electronic Imaging VI*, volume 4299 of *Proc. SPIE*, pages 1–12, San Jose, CA, January 2001.

[4] Thrasyvoulos N. Pappas, Vivien Tartter, Anrew G. Seward, Boris Genzer, Karen Gourgey, and Ilona Kretzschmar. Perceptual dimensions for a dynamic tactile display. In B. E. Rogowitz and T. N. Pappas, editors, *Human Vision and Electronic Imaging XIV*, volume 7240 of *Proc. SPIE*, pages 72400K–1–12, San Jose, CA, January 2009.

[5] Pubudu Madhawa Silva, Thrasyvoulos N. Pappas, Joshua Atkins, and James E. West. Perceiving graphical and pictorial information via touch and hearing. *IEEE Trans. Multimedia*, 18(12):2432–2445, December 2016.

[6] Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statictics of complex wavelet coefficients. *Int. J. Computer Vision*, 40(1):49–71, October 2000.

[7] Lavanya Sharan, Yuanzhen Li, Isamu Motoyoshi, Shin'ya Nishida, and Edward H. Adelson. Image statistics for surface reflectance perception. *Journal of the Optical Society of America A*, 25(4):846–865, 2008.

[8] Maarten W. A. Wijntjes and Sylvia C. Pont. Illusory gloss on Lambertian surfaces. *Journal of Vision*, 10(9):1–12, 2010.

[9] Stephern E. Palmer. *Vision Science: Photons to Phenomenology*. MIT Press, 1999.

[10] Thrasyvoulos N. Pappas, David L. Neuhoff, Huib de Ridder, and Jana Zujovic. Image analysis: Focus on texture similarity. *Proc. IEEE*, 101(9):2044–2057, September 2013.

[11] Jana Zujovic, Thrasyvoulos N. Pappas, and David L. Neuhoff. Structural texture similarity metrics for image analysis and retrieval. *IEEE Trans. Image Processing*, 22(7):2545–2558, July 2013.

[12] Jana Zujovic, Thrasyvoulos N. Pappas, and David L. Neuhoff. Structural similarity metrics for texture analysis and retrieval. In *Proc. Int. Conf. Image Processing*, pages 2225–2228, Cairo, Egypt, November 2009.

[13] Alan C. Brooks, Xiaonan Zhao, and Thrasyvoulos N. Pappas. Structural similarity quality metrics in a coding context: Exploring the space of realistic distortions. 17(8):1261–1273, August 2008.

[14] Jana Zujovic, Thrasyvoulos N. Pappas, David L. Neuhoff, Rene van Egmond, and Huib de Ridder. Subjective and objective texture similarity for image compression. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1369–1372, Kyoto, Japan, March 2012.

[15] Jana Zujovic. *Perceptual Texture Similarity Metrics*. PhD thesis, Northwestern Univ., Evanston, IL, August 2011.

[16] CUReT: Columbia-Utrecht Reflectance and Texture Database. www1.cs.columbia.edu/CAVE/software/curet/.

[17] Kristin J. Dana, Bram van Ginneken, Shree K. Nayar, and Jan J. Koenderink. Reflectance and texture of real-world surfaces. *ACM Trans. Graphics*, 18(1):1–34, January 1999.

[18] Eero P. Simoncelli and William T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proc. ICIP-95, vol. III*, pages 444–447, Washington, DC, October 1995.

[19] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using thee-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, 2001.

[20] Oana G. Cula and Kristin J. Dana. 3D texture recognition using bidirectional feature histograms. *International Journal of Computer Vision*, 59(1):33–60, 2004.

[21] Manik Varma and Andrew Zisserman. A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1/2):61–81, 2005.

[22] Manik Varma and Andrew Zisserman. A statistical approach to material classification using image patch exemplars. 31(11):2032–2047, 2009.

[23] Shree K Nayar, Katsushi Ikeuchi, and Takeo Kanade. Surface reflection: physical and geometrical perspectives. Technical report, Carnegie Mellon University, 1989.

[24] Michael Oren and Shree K Nayar. Generalization of lambert's reflectance model. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 239–246. ACM, 1994.

[25] Zhou Wang and Eero P. Simoncelli. Translation insensitive image similarity in complex wavelet domain. In *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, volume II, pages 573–576, Philadelphia, PA, 2005.

[26] W. Y. Ma, Yining Deng, and B. S. Manjunath. Tools for texture/color based search of images. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors, *Human Vision and Electronic Imaging II*, volume Proc. SPIE, Vol. 3016, pages 496–507, San Jose, CA, February 1997.

[27] Y. Deng, B. S. Manjunath, C. Kenney, M. S. Moore, and H. Shin. An efficient color representation for image retrieval. *IEEE Trans. Image Processing*, 10(1):140–147, January 2001.

[28] Aleksandra Mojsilović, Jelena Kovačević, Jianying Hu, Robert J. Safranek, and S. Kicha Ganapathy. Matching and retrieval based on the vocabulary and grammar of color patterns. *IEEE Trans. Image Processing*, 1(1):38–54, January 2000.

[29] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. Journal of Computer Vision*, 40(2):99–121, 2000.

[30] Thrasyvoulos N Pappas. An adaptive clustering algorithm for image segmentation. *IEEE Transactions on signal processing*, 40(4):901–914, 1992.

[31] Aleksandra Mojsilović, Jianying Hu, and Emina Soljanin. Extraction of perceptually important colors and similarity measurement for image matching, retrieval, and analysis. *IEEE Trans. Image Processing*, 11(11):1238–1248, November 2002.

[32] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

[33] Masashi Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd international conference on Machine learning*, pages 905–912. ACM, 2006.

[34] Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

[35] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[36] G Brelstaff, MW Greenlee, and P Thompson. Reviews: Shape from shading, the perceptual world: Readings from scientific american magazine, visual allusionspictures of perception, 1992.

[37] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.