Perceptual Similarity: A Texture Challenge

Alasdair D. F. Clarke[1]
http://www.macs.hw.ac.uk/texturelab
Fraser Halley[1]
http://www.macs.hw.ac.uk/texturelab
Andrew J. Newell[2]
emailhomepage
Lewis D. Griffin[2]
emailhomepage
Mike J. Chantler[1]
http://www.macs.hw.ac.uk/~mjc/

[1] The Texture Lab
School of Mathematical and Computer Sciences
Heriot-Watt University, Edinburgh, UK

[2] Vision & Image Sciences
Computer Science
University College London, UK

Abstract

Over the last thirty years evaluation of texture analysis algorithms has been dominated by two databases: Brodatz has typically been used to provide single images of approximately 100 texture classes, while CUReT consists of multiple images of 61 physical samples captured under a variety of illumination conditions. While many highly successful approaches have been developed for classification, the challenging question of measuring *perceived* inter-class texture similarity has rarely been addressed. In this paper we introduce a new texture database which includes a collection of 334 samples together with perceptual similarity data collected from experiments with 30 human observers. We have tested four of the leading texture algorithms and they provide accurate ($\approx 100\%$) performance in a CUReT style classification task, however, a second experiment shows that resulting inter-class distances do not correlate well with the perceptual data.

1 Introduction

Texture classification and segmentation have been extensively researched over the last thirty years. Early on the Brodatz album[1] quickly became the de facto standard in which a *texture class* comprised a set of non-overlapping sub-images cropped from a *single* photograph. Later, as the focus shifted to investigating illumination- and pose-invariant algorithms, the CUReT database[2] became popular and the *texture class* became the set of photographs of a single physical sample captured under a variety of imaging conditions. While extremely successful algorithms have been developed to address classification problems based on these databases, the challenging problem of measuring perceived *inter-class* texture similarity has rarely been discussed. Rao & Lohse[7] asked participants to sort 56 textures into perceptually similar subsets using any criteria they wished and used multidimensional scaling to derive a 3D perceptual space from these results. They concluded that the axes correlated

BMVC 2011 http://dx.doi.org/10.5244/C.25.120

with the visual properties of repetitiveness, orientation and complexity. These results were questioned by Heaps & Hande [8] who argued that a dimensional model of texture is inappropriate and that texture similarity judgements are context dependent. Better results have been obtained for computational similarity by Long & Leow[9] and Petrou *et al.* [11] however, these again used the limited Brodatz album which provides a relatively sparse sampling of texture space.

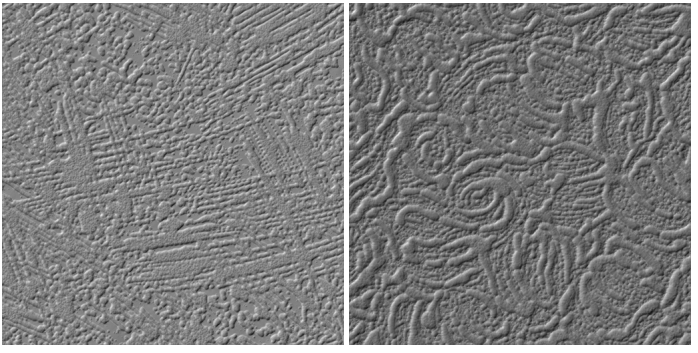
This paper makes use of a new texture collection[7] that was originally developed to investigate navigation and browsing of large image databases[6]. It comprises 334 texture samples and an associated *perceptual* similarity matrix. We used this to perform two experiments testing *classification* (Section 3.1) and *inter-class similarity* (Section 3.2) performance. Experiment 1 was designed to check that our dataset was not taking the classification algorithms outside their normal range of operation, while Experiment 2 tested their ability to mimic human perception.

2 Texture Database and Perceptual Data

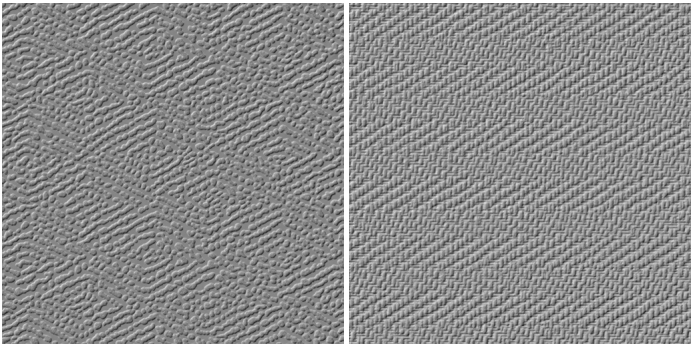
Our database comprises of 334 texture samples and we assume each sample to represent a separate texture class. It includes examples of embossed vinyl, woven wall coverings, carpets, rugs, window blinds, soft fabrics, building materials, product packaging, etc. As our studies focus on the perception of surface relief we calculated the heightmaps of all textures and then relit them in order to remove variations due to reflectance function.

The height-maps were captured using four light photometric stereo[15]. However, only the image data were used in this paper and these were generated assuming constant albedo, Lambertian surfaces. For the first experiment a variety of illumination elevations, $\{50^\circ, 60^\circ, 70^\circ\}$, and azimuths $\{n45^\circ | n = 1, \ldots, 8\}$, were used providing 24 images per class. For Experiment 2 the illumination and viewing conditions were kept constant (elevation 45° and azimuth 135°) to provide one image per sample. These images were printed out onto photographic card ($280g/m^2$), dimensions 10.16×10.16cm, and a grouping experiment, similar to those used by Rao & Lohse[12] and Petrou *et al.* [11], was carried out. Thirty participants were asked to sort the images into similar subsets; although due to the large size of the database, only six participants carried out a grouping experiment on the whole set. This was a time consuming task with each participant taking between $2\frac{1}{2}$ and 4 hours. The results of the 6-participant pilot experiment were used to partition the dataset into three subsets of approximately equal size, which were grouped individually by a further 8 people per subset. The participants were asked simply to group the images into perceptually similar subsets and were given no explicit directions on what visual features to use or ignore. They were allowed to make as many groups as they wished and were allowed to split or merge groups as they carried out the experiment. Some examples of the results are shown in Figure 1.

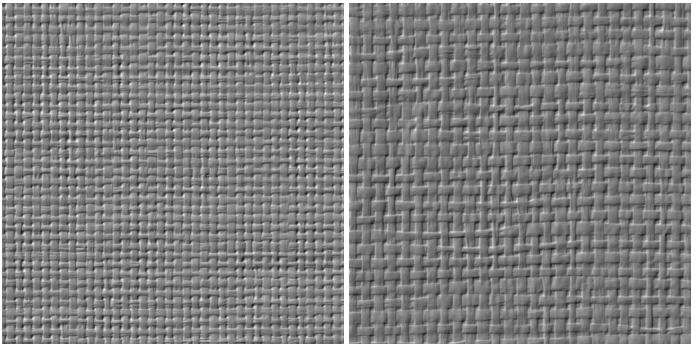
The similarity scores, $S(I_i, I_j)$, for each texture pair were calculated simply by dividing the number of observers that grouped the pair into the same sub-set by the number of observers that had the opportunity do so. A *dissimilarity matrix* was then defined as $d_{sim}(I_i, I_j) = 1 - S(I_i, I_j)$. Hence $d_{sim}(I_i, I_i) = 0$ for all images I_i, and $d_{sim}(I_i, I_j) = 1$ if none of the participants grouped images I_i together with I_j. Further experimental details are given in [6], where the similarity matrix has been used to construct fast texture database navigation environments.



(a) A pair of images with: $d_{sim}(27,131) = 1$. None of the human observers grouped these textures together



(b) A pair of images which approximately half of human observers grouped together: $d_{sim}(5,86) = 0.57$.



(c) A pair of images which all but one of human observers grouped together: $d_{sim}(168,176) = 0.07$.

Figure 1: Some examples of textures from the dataset with their similarity matrix results

Feature	Type	Dictionary	Dimensions
MRV8	Filter based	Training stage	$8 \rightarrow 3340$
MRF	Pixel based	Training stage	$49 \rightarrow 3340$
LBP	Pixel based	Predefined	26
BIF	Filter based	Predefined	1296

Table 1: Summary of texture features. Dictionaries for the MRV8 and MRF methods are created in an initial training stage where k-means is used to find 10 textons per image. These are then aggregated across texture classes, giving a 3340-dimensional representation for textures.

3 Experiments

Much of the work on texture analysis over the last decade has concerned the problem of classifying images under varying illumination conditions. While some model-based approaches [2] have been developed most of the work in this area has been data-driven (appearance-based). These data driven approaches typically involve extracting feature vectors derived from filter-banks or local neighbourhood statistics for each pixel in the image and then applying vector quantisation to some pre-defined texton dictionary. In the two experiments, we evaluated four different texture features (LBP[10], MRV8[13], MRF[14], BIF[3] (see Table 1). We also tested a multi-scale implementation of the BIF algorithm (MS-BIF). These feature sets were chosen as they have all demonstrated excellent results on previous databases and are available to run unchanged as 'off the shelf algorithms.'

The aim of Experiment 1 (classification) was to test the texture classification algorithms on the new dataset, mimicking the structure of the CUReT database[4] and the protocol commonly used on it (for example, see [3]). In the second experiment, we tested the same algorithms on a set of images consisting of one render per texture (all with constant viewing and lighting conditions) and compared the computed histogram distances with the perceptual dissimilarity data.

3.1 Experiment 1: Classification

In order to mimic the classification assessments made using the CUReT database a subset of 60 height maps were randomly selected from the database. Each heightmap was then rendered under a variety of illumination conditions (eight azimuths and three elevations) to give a set of 24 images per class. Images were 256×256 pixels in dimension. This smaller size was chosen so the set of images were comparable with the CUReT database (200×200 pixels). We also carried out the classification protocol with the fullsize, 1024×1024, images (256×256 crops were taken in the dictionary classification stage for the MRFs and MRV8s).

Half the images from each texture class were randomly selected as the training set, with the remaining half used as the test set. For the MRV8 and MRF methods, were a texton dictionary needs to be constructed, the feature vectors from all 12 training images were aggregated together and k-means was used to quantise the feature vectors into 10 clusters or "textons" per texture class. For local binary patterns, we use $\mathrm{LBP}^{riu2}_{24,3}$. The training phase involves creating a histogram for each image in the training set. To classify an image, its texton histogram is created and compared to the histograms from all the training images using $1 - \sqrt{h} \cdot \sqrt{g}$, (a simplified form of the Bhattacharyya distance, as used in [3]).

Feature	LBP	MRV8	MRF	BIF
accuracy for 256×256 images	99.4%	92.8%	94.7%	98.1%
accuracy for 1024×1024 images	96.7%	96.0%	84.0%	99.2%

Table 2: Classification performance. Accuracy rates for all features are similar to those reported in previous studies. There does not appear to be a big difference in performance between sample size.

3.1.1 Classification Results

Our results are in agreement with previous work[3, 10, 13, 14], with each of the feature sets achieving a very high classification accuracy (see Table 2). This is especially impressive as all the computational features were developed, trained and tested on a completely different dataset. In particular, the CUReT texture samples are much smaller than the samples in our new dataset, and hence the important information is on a different scale. Surprisingly, this had no effect on classification accuracy.

3.2 Experiment 2: Similarity

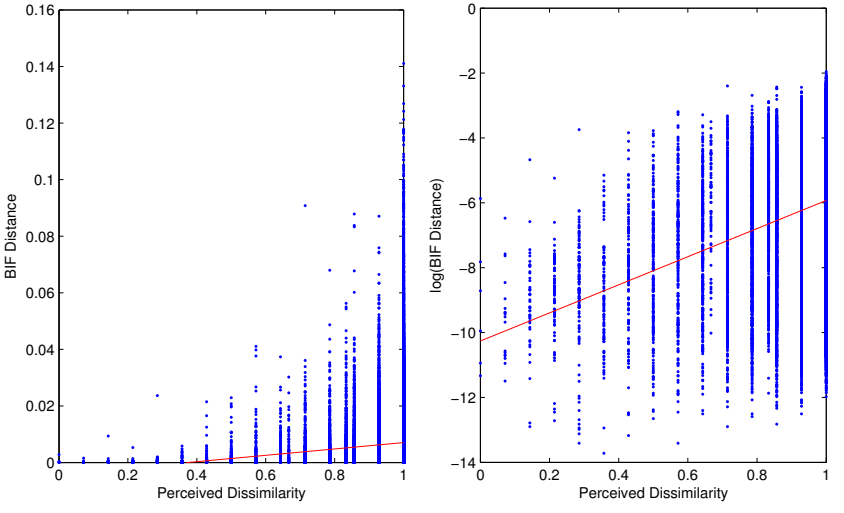
In the second experiment we tested the ability of the four texture algorithms to predict the inter-class similarities derived from our 30 observers. All evaluations were conducted over two scales (256×256 and 1024×1024). To test the LBPs, we computed the histograms for all 334 images in the database, and then compared the distances between histograms using $ln(1-\sqrt{h}\cdot\sqrt{g})$. In order to test the MRV8 and MRF methods, we need to first construct a texton dictionary. This was done by extracting 10 clusters per image using k-means which were there then aggregated to give 3340 textons. We also experimented with running the MRF and MRV8 algorithms using reduced dictionaries (by a further application of k-means with $k = 50, 500$), however, there was little difference in performance.

3.2.1 Similarity Results

As can be seen in Figure 2 the computed distances between histograms do not correlate well with human judgements. Interestingly though, for pairs of images that were judged very similar by human observers, the computer also judged the images to be similar. However, for pairs that human observers judged to be dissimilar the computational algorithms gave a wide range of responses. This is born out in the correlation coefficients, with the best performance giving $R^2 = 0.04$ (see Table 3). Applying a post-hoc log transform to the data slightly improves matters ($R^2 \approx 0.07$). We also examined Spearman's rank correlation coefficient, which showed that there was only a weak relationship between perceptual and computational dissimilarity ($\rho = 0.21$). Overall the MRFs appear to correlate best with human judgement, although none of the relationships are particularly strong.

Another way of analysing the data is to look at database retrievals. For each image I_i we use the perceptual similarity matrix to find a subset of images, $S_p(i)$, that have $d_{sim} < k$ for some constant k. Let n_i be the number of elements in $S_p(i)$. We now use the computational features to retrieve the n_i most similar images, denoted $S_f(i)$. The performance of the feature is then given by:

$$R(k) = \Sigma_{i=1}^{334} \frac{|S_p(i) \cap S_f(i)|}{n_i} \tag{1}$$



(a) (b)

Figure 2: MS-BIF performance. The scatter plots for the other features all look very similar.

Feature	R^2	R^2 - log(Feature)	ρ
LBP	0.031	0.025	0.131
MRV8	0.042	0.077	0.180
MRF	0.031	0.071	0.206
BIF	0.009	0.063	0.166
MS-BIF	0.011	0.058	0.176

Table 3: Similarity performance. R^2 gives the proportion of the variation in y explained by x (where variation is measured in a sum of squares sense). So we can see that the best computational feature explains $\approx 5\%$ of the variation in human responses. ρ is Spearman's correlation coefficient. Each feature was also run on 256×256 down sampled images, but the results were very similar. Furthermore, the MRF and MRV8 features were also computed using reduced dictionaries (obtained through a further application of k-means) with 50 and 500 textons. Again, this made little difference to the results.

Computational Feature	V. SimRet. <0.25	M. SimRet <0.5	L SimRet <1.0
Number of Pairs:	147	482	9061
LBP	19.3%	18.0%	25.1%
MRV8	37.2%	32.9%	31.1%
MRF	31.5%	31.2%	33.1%
BIF	45.4%	38.1%	32.9%
MS-BIF	45.0%	37.2%	35.4%

Table 4: Results from the retrieval test. The 2nd row, "Number of Pairs" indicated the number of texture pairs that have a perceptual dissimilarity of less than 0.25, 0.5 and 1.0.

The results for $k \in \{0.25, 0.5, 1.0\}$ are shown in Table 4. The BIF algorithm gave the best performance in terms of retrieving the most similar textures for any given query texture with a retrieval rate of 45%. Figure 3 illustrates this with an example. The top three retrievals, according to the perceptual similarity matrix (Figure 3b) all appear a closer fit for the query texture (Figure 3a) than the three images with the closest BIF-distance (Figure 3c). Furthermore, as we increased k (and hence increase the number of retrievals) performance decreased and when $k = 1$ (which corresponds to retrieving all textures that at least one human participant grouped with the query texture), there was little difference between the algorithms, with precision $\approx 33\%$.

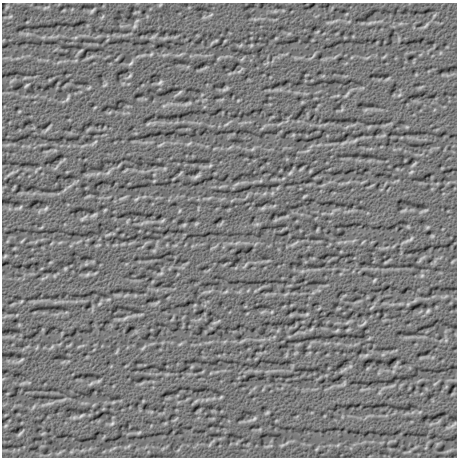
3.3 Discussion

The question that the above results raise is: why do algorithms which achieve near perfect ($> 90\%$) performance on classification tasks only account for a relatively small proportion ($< 8\%$) of judgements made by human observers? The authors believe that this may be due to two main reasons.

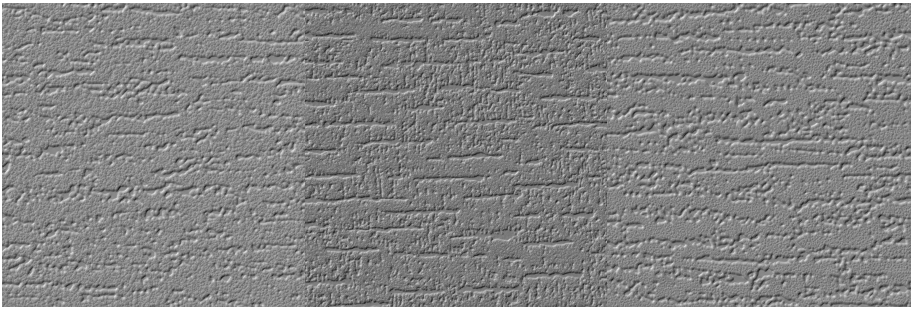
One possibility is that the grouping data obtained from our thirty observers are not representative of that which would be obtained from the wider population. However, visual inspection using a dendrograms[1] suggest that the data are reasonable. Furthermore, the perceptual similarity matrix has been extensively used to design intuitive database and browsing environments and allows for statistically significantly faster navigation[6] when compared to environments based on the trace transform[11]. However, the authors recognise that this possibility can by no means be dismissed and what is really required is to significantly increase the numbers of observers in the grouping experiments and we have planned a series of further observer sessions that we will analyse later this year.

The second possibility is due to the fact that the two tasks are fundamentally different. The test *classification* task used in the majority of the literature (and followed here in Experiment 1) is essentially to match an unseen (test) image to other (training) images obtained from the *same physical sample*. This is because no human judgements are involved and therefore the only information that is available is that images were either obtained from a particular physical sample (or not). For the Brodatz album, within-class test and training images are obtained from different (generally non-overlapping) areas of the same texture sample. In CUReT, test and training images are typically obtained from the same physical area of the sample, but captured under different illumination and/or viewing conditions (although there are some exceptions e.g.[5]). In both the Brodatz and CUReT cases therefore,

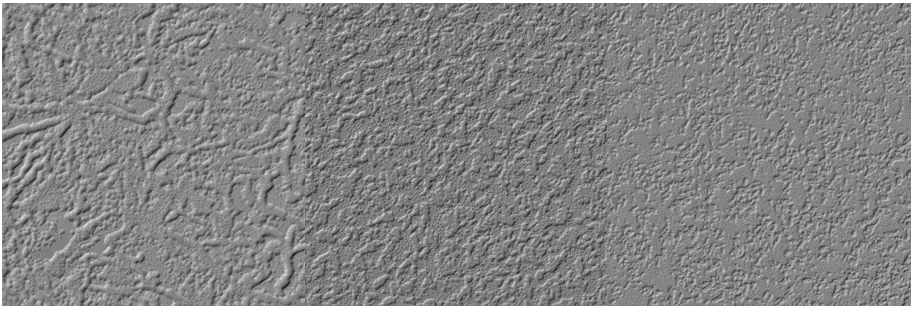
[1] Please see supplementary materials.



(a) Query texture 011



(b) SimMatrix retrievals: $d_{Sim}(011,062) = d_{Sim}(011,119) = d_{Sim}(011,123) = 0.0714$. i.e. only one participant did not group all these textures together.



(c) MS-BIF retrievals: $d_{BIF}(011,028) = d_{BIF}(011,071) = d_{BIF}(011,162) = 1.42 \times 10^{-}5$

Figure 3: Retrieval results for texture #011. The MS-BIF feature ranks textures 062, 119 and 123 as the 8th, 11th and 13th similar to texture 011. Computational similarity scores: $d_{BIF}(011,028) = 2.05 \times 10^{-}5$, $d_{BIF}(011,119) = 2.28 \times 10^{-}5$ and $d_{BIF}(011,123) = 2.44 \times 10^{-}5$.

it is likely that there are *many* spatial statistics that are capable of discriminating between classes and indeed the literature has reported considerable success in developing and identifying many different sets of texture features. However, there is no reason to believe that these features encapsulate the same set of spatial statistics that human observers use to perform texture similarity judgements. Furthermore, the Brodatz/CUReT databases have necessarily focused attention on near identical textures (i.e. images drawn from the same physical samples) however, Experiment 2 included information right across the similarity range. Thus it is unfair to expect algorithms that have been designed and developed over many years on the Brodatz/CUReT classification task to be good at mimicking human behaviour over a much wider range of similarity judgements.

A caveat to the above is that we took the algorithms as given and there was no training stage (beyond dictionary generation) in the similarity experiment (Section 3.2).

4 Conclusion

We believe that this set of 334 textures is currently the largest texture database that has been captured under controlled illumination conditions and, perhaps more importantly, is accompanied by an associated perceptual similarity matrix. It is also contains height data allowing illumination-independent generation of features and relighting under arbitrary illumination conditions.

In Experiment 1 we investigated the performance of four state-of-the-art classification schemes and showed that they provide near-ceiling texture classification performance when tested on this new texture database using a protocol similar to that commonly used with the CUReT image set. However, Experiment 2 showed that the perceptual similarity matrix obtained using 30 human observers does not correlate well with machine performance. This is likely to be due to either (a) the perceptual data not being representative of the population or (b) the algorithms not exploiting all of the texture features that are used by human observers. If the latter is the case then it may be that it is the salient, longer range spatial interactions that are not detected by the relatively small spatial neighbourhoods that machine vision features use.

References

[1] P. Brodatz. *Textures: A photographic album for artists and designers*. Dover Publications, 1966.

[2] M. J. Chantler, M. Petrou, A.Penirschke, M. Schmidt, and G. McGunnigle. Classifying surface texture while simultaneously estimating illumination. *International Journal of Computer Vision (VISI)*, 62:83–96, 2005.

[3] M. Crosier and L. D. Griffin. Using basic image features for texture classification. *Internation Journal of Computer Vision*, 88:447–460, 2010.

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

[5] O. Drbohlav and A. Leonardis. Towards correct and informative evaluation methodology for texture classification under varying viewpoint and illumination. *Computer Vision and Image Understanding*, 114:439–449, 2010.

[6] F. Halley. *Perceptually Relevant Browsing Environments for Large Texture Databases (submitted)*. Heriot-Watt University, 2011.

[7] F. Halley. Pertex v1.0, 2011. URL http://www.macs.hw.ac.uk/texturelab/resources/databases/pertex/.

[8] C. Heaps and S. Handel. Similarity and features of natural textures. *Journal of Experimental Psycholog.: Human Perception and Performance*, 25:299–320, 1999.

[9] H. Long and W. K. Leow. Perceptual texture space improves perceptual consistency of computational features. In *IJCAI'01 Proceedings of the 17th international joint conference on Artificial intelligence*, 2001.

[10] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:971–987, 2002.

[11] M. Petrou, A. Talebpour, and A. Kadyrov. Reverse engineering the way humans rank texture. *Pattern Analysis Applications*, 10:101–114, 2007.

[12] A. R. Rao and G. L. Lohse. Identifying high level features of texture perception. *CVGIP: Graph. Models Image Process.*, 55:218–233, May 1993.

[13] M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 691–698, 2003.

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

[15] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19:139–144, 1980.