Unifying Statistical Texture Classification Frameworks

Manik Varma and Andrew Zisserman

Dept. of Engineering Science
University of Oxford
Oxford, OX1 3PJ, UK
{manik,az}@robots.ox.ac.uk

Abstract

The objective of this paper is to examine statistical approaches to the classification of textured materials from a single image obtained under unknown viewpoint and illumination. The approaches investigated here are based on the joint probability distribution of filter responses.

We review previous work based on this formulation and make two observations. First, we show that there is a correspondence between the two common representations of filter outputs - textons and binned histograms. Second, we show that two classification methodologies, nearest neighbour matching and Bayesian classification, are equivalent for particular choices of the distance measure. We describe the pros and cons of these alternative representations and distance measures, and illustrate the discussion by classifying all the materials in the Columbia-Utrecht (CUReT) texture database.

These equivalences allow us to perform direct comparisons between the texton frequency matching framework, best exemplified by the classifiers of Leung and Malik [IJCV 2001], Cula and Dana [CVPR 2001], and Varma and Zisserman [ECCV 2002], and the Bayesian framework most closely represented by the work of Konishi and Yuille [CVPR 2000].

Key words: Texture, Classification, PDF representation, Textons

1 Introduction

In this paper, we examine two of the most successful frameworks in which the problem of texture classification has been attempted – the texton frequency comparison framework, as exemplified by the classifiers of [3,8,14], and the Bayesian framework, most closely represented by [7]. While the frameworks appear seemingly unrelated, we draw out the similarities between them, and show that the two can be made equivalent under certain choices of representation and distance measure.

The classification problem being tackled is the following: given a single image of a textured material obtained under unknown viewing and illumination conditions, classify it into one of a set of pre-learnt classes. Classifying textures from a single image under such general conditions and without any *a priori* information is a very demanding task.

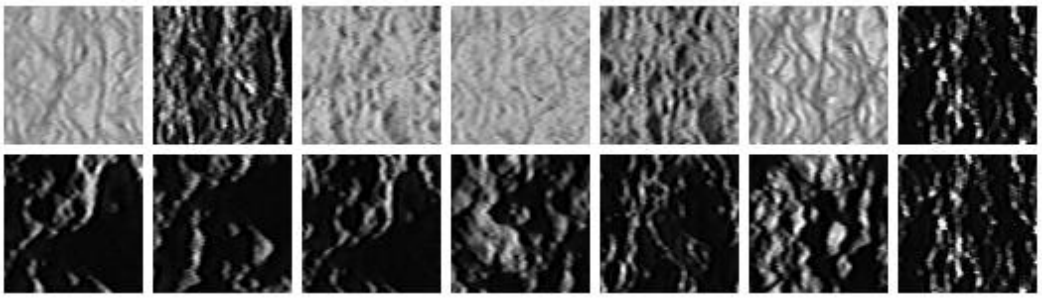

Fig. 1. The change in imaged appearance of the same texture (Plaster B, texture # 30 in the Columbia-Utrecht database [5]) with variation in imaging conditions. Top row: constant viewing angle and varying illumination. Bottom row: constant illumination and varying viewing angle. There is a considerable difference in the appearance across images.

What makes the problem so hard is that unlike other forms of classification, where the objects being categorised have a definite structure which can be captured and represented, most textures have large stochastic variations which make them difficult to model. Furthermore, textured materials often undergo a sea change in their imaged appearance with variations in illumination and camera pose (see figure 1). Dealing with this successfully is one of the main tasks of any classification algorithm. Another factor which comes into play is that, quite often, two materials when photographed under very different imaging conditions can appear to be quite similar, as is illustrated by figure 2. It is a combination of all these factors which makes the texture classification problem so challenging.

Nevertheless, weak classification algorithms which exploit the statistical nature of textures by characterising them as distributions of filter responses have shown much promise of late. The two types of algorithm in this category that have been particular successful are (a) the Bayesian classifier based on the joint probability distribution of filter responses represented by a binned histogram [7], and (b) the nearest neighbour χ^2 distribution comparison classifier based on a texton frequency representation [3,8,14]. In this paper, we draw an equivalence between these two schemes which then allows a direct comparison of the performance of the two classification methodologies.

The success of Bayesian classification applied to filter responses was convincingly demonstrated by Konishi and Yuille [7]. Working on the Sowerby and San Francisco outdoor datasets, their aim was to classify image pixels into one of six texture classes. The joint PDF of the empirical class conditional probability of six filter responses for each texture was learnt from training images. It was represented as a histogram by quantising the filter responses into bins. Novel image pixels were then

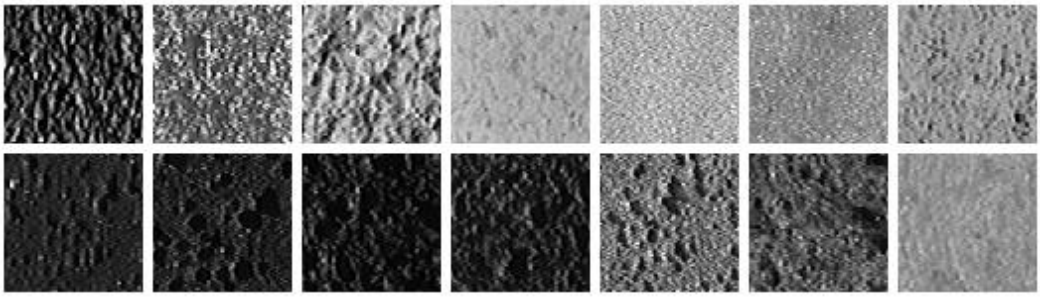

Fig. 2. Small inter class variations between textures present in the Columbia-Utrecht database. In the top row, the first and the fourth image are of the same texture while all the other images, even though they look similar, belong to different classes. Similarly, in the bottom row, the images appear similar and yet there are three different texture classes present.

classified by computing their filter responses and using Bayes' decision rule. However, for image regions, Konishi and Yuille only reported the Chernoff information (which is a measure of the asymptotic classification error rate) but didn't perform any actual classification. In this paper, we take the final step and use Bayes' decision rule to classify entire images by making the same assumption as Schmid [11], i.e. that a region is a collection of statistically independent pixels, and whose probability is therefore the product of the individual pixel probabilities.

In contrast, frequency comparison classifiers such as [3,8,14] learn distributions of texton frequencies from training images and then classify novel images by comparing their texton distribution to the learnt models. Different comparison methods may be used, such as the Bhattacharya metric [12], Earth Mover's distance [10], KL divergence and other entropy based measures [16], but the χ^2 significance test [9], in conjunction with a nearest neighbour rule, has become the default choice.

Leung and Malik [8] were amongst the first to seriously tackle the problem of classifying 3D textures and, in doing so, made an important innovation by giving an operational definition of a texton. They defined a 2D texton to be a cluster centre in filter response space. This not only enabled textons to be generated automatically from an image, but also opened up the possibility of a *universal* set of textons for all images. To compensate for 3D effects, they proposed 3D textons which were cluster centres of filter responses over a stack of 20 training images with representative viewpoint and lighting. They developed an algorithm capable of classifying a stack of 20 registered, novel images using 3D textons and applied it very successfully to the Columbia-Utrecht (CUReT) [5] database. Later, Cula and Dana [3] and Varma and Zisserman [14] showed that 2D textons could be used to classify single images without any loss of performance.

Classification performance is evaluated here also on image sets taken from the CUReT texture database. All 61 materials present in the database are included, and 92 images of each material are used with only the most extreme viewpoints

being excluded (see [14] for details). The variety of textures in this database is illustrated in figure 3. The 92 images present for each texture class are partitioned into two, disjoint sets. Images in the first (training) set are used for model learning, and classification accuracy is assessed on the 46 images for each texture class in the second (test) set.

The materials in the CUReT database are examples of 3D textures and exhibit a marked variation in appearance with changes in viewing and illumination conditions [2–4,8,14,17]. The difficulty of single image classification is highlighted by figure 4 which illustrates how drastically the appearance of a texture can change with varying imaging conditions.

Modelling such textures by a single probability distribution of filter responses may fail in these situations. The solution adopted in this paper is to represent each texture class by multiple models which characterise the different appearances of the texture with variation in imaging conditions. These models are generated from the various training images for each texture class, and essentially this means that each texture class is represented by a set of probability distributions implicitly conditioned on viewpoint and illumination. In our experiments, each training image is used to generate a model and thus there are 46 models per texture class. However, the number of models can be substantially reduced to, on average, 7-8 per class by

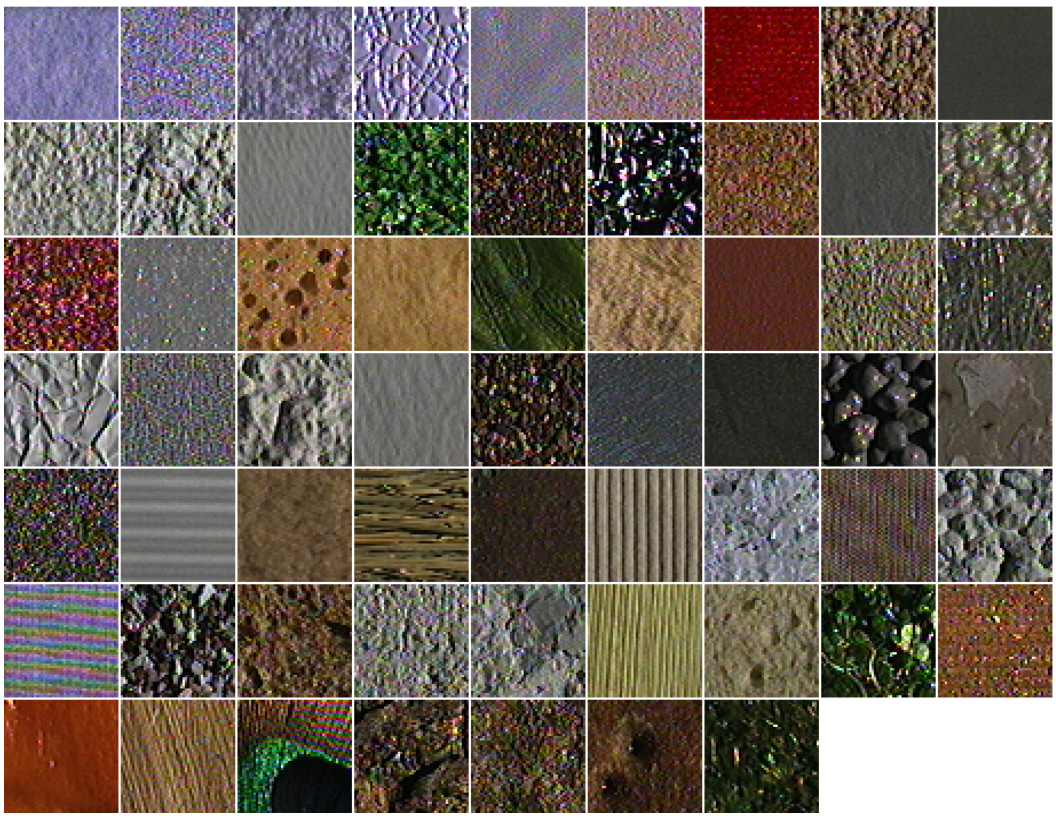

Fig. 3. Textures from the Columbia-Utrecht database. All images are converted to monochrome in this work, so colour is not used in discriminating different textures.

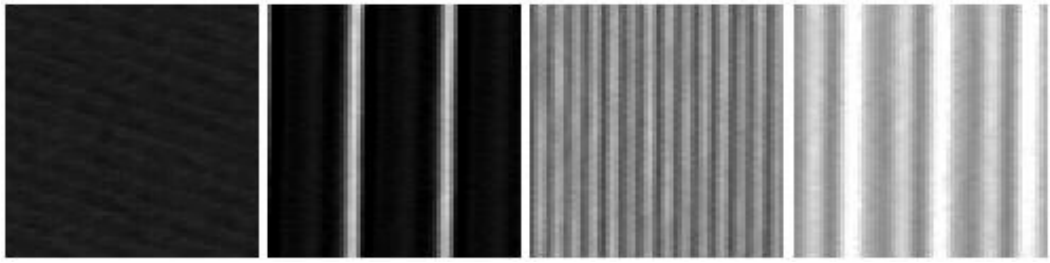

Fig. 4. Images of ribbed paper taken under different viewing and lighting conditions. The material has significant surface normal variation and consequently changes its appearance drastically with imaging conditions. Characterising such a texture by only one model, i.e. a single probability distribution of filter responses, will lead to poor classification results. Instead, it is better to represent a texture by multiple models generated under different viewing and illumination conditions.

the use of *K-Medoid* and *Greedy* algorithms as described in [14].

The layout of the rest of the paper is as follows: in section 2, we outline a low dimensional representation of rotationally invariant filter responses which was first introduced in [14]. We also describe the two common semi-parametric representations, texton and binned histogram, of the joint PDF of filter responses and show how it is possible to convert one to the other. Then, in section 3, we present a comparison between the texton and bin representations using a distribution comparison classifier. Finally, we implement a Bayesian classifier using a texton representation in section 4 and contrast its performance to the distribution comparison classifier.

2 Filter Responses and their Representation

In this section, we first describe the filter bank that will be used to generate features, and then discuss two popular representations of filter responses, textons and binned histograms, between which we will draw a correspondence.

2.1 Filter bank

Traditionally, the filter banks employed for texture analysis have included a large number of filters in keeping with the philosophy that many diverse features at multiple orientations and scales need to be extracted accurately to successfully classify textures. However, constructing and storing PDFs of filter responses in a high dimensional filter response space is computationally infeasible and therefore it is necessary to limit the dimensionality of the filter response vector. Both these requirements can be achieved if multiple oriented filters are used but their outputs combined to form a low dimensional, rotationally invariant response vector. A filter bank which does this is the Maximum Response (MR8) filter bank which com-

prises 38 filters but only 8 filter responses (see figure 5). The filters consists of a Gaussian and a Laplacian of Gaussian (LOG) both with $\sigma = 10$ pixels (these filters have rotational symmetry), an edge filter at 3 scales (σ_x,σ_y)=$\{$(1,3), (2,6), (4,12)$\}$ pixels and a bar filter at the same 3 scales. The latter two filters are oriented and occur at 6 orientations at each scale. The responses of the isotropic filters (Gaussian and LOG) are used directly, but the responses of the oriented filters (bar and edge) are, at each scale, "collapsed" by using only the maximum filter responses across all orientations – thereby giving 8 rotationally invariant filter responses in all.

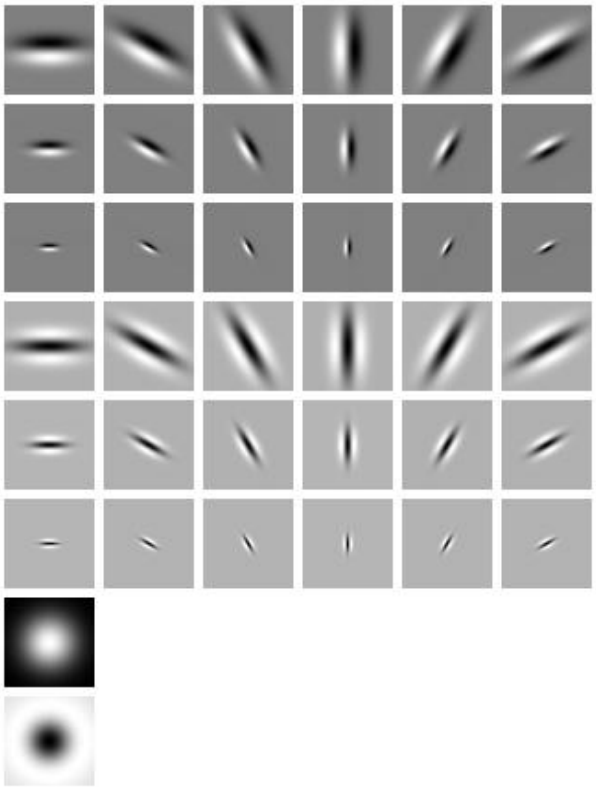

Fig. 5. The rotationally invariant MR8 filter bank consists of (in order, from top to bottom) an oriented edge filter at 6 orientations and 3 scales, a bar filter at the same set of orientations and scales, an isotropic Gaussian filter and a Laplacian of Gaussian filter. The responses of the isotropic filters are recorded directly. However, taking the maximal response of the anisotropic filters across all orientations results in 2 filter responses being recorded per scale, giving 8 filter responses in total. Essentially this is the maximum response of each row of the filter bank above.

Rotation invariance is desirable in that it leads to the correct classification of rotated versions of textures present in the training set. Another motivation for using the MR8 filter bank is that despite being rotationally invariant, its oriented filters should accurately pick up anisotropic features whose relative angular information (as determined by the angle of maximum response) can still be recorded. Furthermore, we expect that more significant textons are generated when clustering in a low dimensional, rotationally invariant space.

Both the Bayesian and the texton frequency comparison classifiers discussed in this paper are divided into two stages – learning and classification. In the learning stage, features are extracted from the training images by convolution with the MR8 filter bank. The choice of representation of the resultant filter response features determines the learnt statistical model for a given texture under particular imaging conditions and is described next.

2.2 Texton representation of filter responses

Each training image is convolved with the MR8 filter bank to generate a set of filter responses. These filter responses are then aggregated over various training images from the same texture class (as described in the following paragraph) and clustered. The resultant cluster centres form a dictionary of exemplar filter responses and are called textons. Given a texton dictionary, the first step in learning a model from a particular training image is labelling each of the image pixels with the texton that lies closest to it in filter response space. The (normalised) frequency histogram of pixel texton labellings then defines the model for the training image.

In the implementation in this paper, the filter responses of 13 randomly selected images per texture class (taken from the training set) are aggregated and clustered via the *K-Means* algorithm. The textons (cluster centres) learnt from each class are then combined into a single dictionary of size S. Various values of K are tried ($K = 10, 20, \ldots, 50$) and results are reported for each case. Thus, in each experiment, K textons are learnt from every texture class resulting in a dictionary comprising a total of $S = 61 \times K$ textons. Hence, a model of a training image is a S-vector where each component is the proportion of pixels which are labelled as a particular texton.

2.3 Histogram representation by binning

In this representation, the model corresponding to a given image is the joint probability distribution of the image's filter responses – obtained by quantising the responses into bins and normalising so that the sum over all bins is unity. It should be noted that the number of bins and their placement can be important parameters as they determine how crudely, or how well, the underlying probability distribution is approximated and whether the data is over-fitted or not.

As an implementation detail, the histogram is stored as a sparse matrix and the space it occupies is given by: number of non-empty bins $\times$ number of bytes required to store a bin value and its corresponding index. This is bounded above by the number of data points and compares favourably to a naive implementation which stores the full matrix in $\mathcal{O}(\text{total number of bins})$ bytes, but where most of the

bins are empty. For example, using this implementation for MR8 with 20 bins per dimension, we were able to store the PDF of all the training images in less than a hundred megabytes whereas the naive implementation would have taken over five hundred gigabytes. Also, it is efficient to store the histogram as a sparse matrix as the χ^2 statistic can be evaluated in $\mathcal{O}(\text{number of non-empty bins})$ flops.

2.4 Moving between representations

The two representations of filter responses can be made identical by a suitable choice of bins or textons. For example, an equally spaced bin representation can be converted into an identical texton representation by placing a texton at the centre of every bin (see figure 6). In the algorithm implemented here, the textons are generated by clustering and do not coincide with the bin centres. Hence, the two representations are not identical in this case.

It is possible to go the other way round as well. Every texton representation can be converted into an identical bin representation. In this case, the bins will be irregularly shaped and will correspond to the Voronoi polytopes obtained by forming the Voronoi diagram of the texton sites. Thus, clustering to get textons can be thought of as an adaptive binning method and a histogram of texton frequencies can be equated to a bin count of filter responses. In essence, the comparisons made in section 3 can be thought of as a comparison between two different texton dictionaries.

However, it should be noted that in general, not every bin representation can be converted to an equivalent texton representation in which there is a one to one mapping between textons and bins. Though it might be possible to find a similar

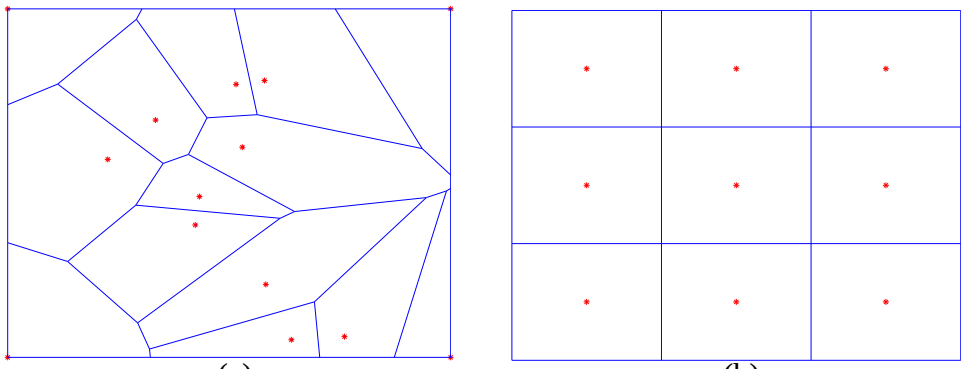

Fig. 6. Texton and bin equivalence in two dimensions: (a) Every texton representation can be converted into an equivalent bin representation where the bins are the Voronoi polygons. (b) Conversely, an equally spaced bin representation can be converted into an identical texton representation by placing a texton at the centre of each bin (though not every bin representation has an equivalent bijective texton representation). A similar equivalence applies in $\Re^N$.

representation if there are more textons than bins with certain textons being grouped together to form a particular bin.

3 Classification by Distribution Comparison

In this section the effect of the representation on classification performance is investigated. Given a set of models characterising the 61 material classes, the task is to classify a novel (test) image as one of these textures. This proceeds as follows: the filter response distribution is computed for the test image, and both types of representation (texton and bin) are then determined. In either case, the closest model image, in terms of the χ^2 statistic, is found and the novel image declared to belong to model's texture class.

3.1 Experimental setup and results

Classification is carried out on all 61 texture classes for both the representations. We consider the case where every image in the training set, for each texture class, is used to generate a model. Thus, there are 46 models per texture, for each of the 61 texture classes. The classification performance is measured by the proportion of test images which are correctly classified as the right texture.

For the texton based representation, figure 7 plots the classification accuracy versus the number of textons in the dictionary when classifying all 2806 test images. The best results are obtained when $K = 40$ textons are learnt per texture (resulting in a dictionary of size $S = 2440$ textons) and an accuracy rate of 97.43% is achieved.

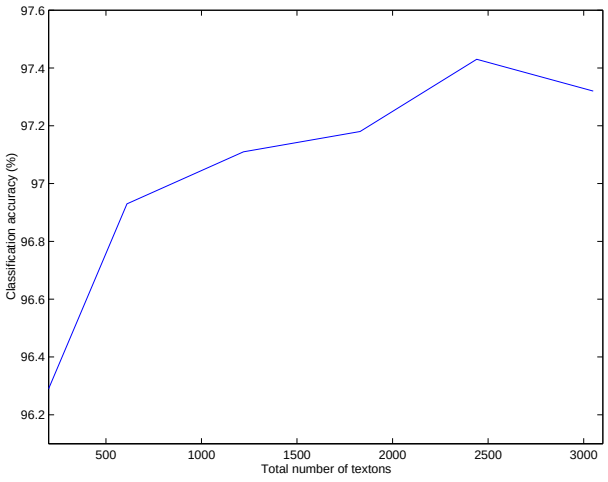

Fig. 7. The variation in classification performance with the number of of textons. The best classification result obtained is 97.43% using a dictionary of size $S = 2440$ textons.

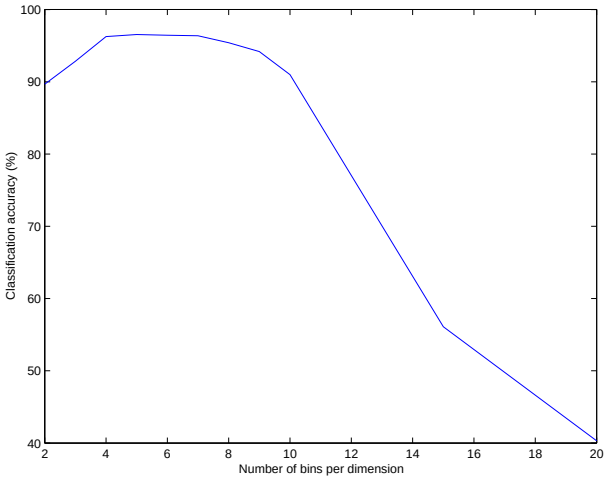

Fig. 8. The variation in classification performance with the number of bins used during quantization: The best classification results are obtained when the filter responses are quantized into 5 bins per dimension.

For the bin representation, the number and location of the bins are, in general, important parameters. However, it turns out that in this case excellent results are obtained using equally spaced bins. Figure 8 plots the classification accuracy for the test set versus the number of bins used in the quantization process. The classifier achieves a maximum accuracy of 96.54% when the filter responses are quantised into 5 bins per dimension. Increasing the number of bins decreases the performance, indicating that the distribution is being over-fitted and that noise is being learnt as well. The classification accuracy also decreases with decrease in the number of bins as the binning is now crude.

Both the representations give very similar classification results. Of course, this is not surprising in the light of the fact that the two can be made identical. In this particular instance, however, the texton representation slightly outperforms the bin representation as the bins are always equally sized while the textons are learnt adaptively from the given data.

4 Bayesian Classification

Given that texton frequencies and histogram binning are equivalent ways of representing the PDF of filter responses, it is now possible to calculate the class conditional probability of obtaining a particular filter response using a texton representation. This setting of a texton representation in a Bayesian paradigm effectively lets us compare, in this section, the classification scheme of Konishi and Yuille [7] with the texton based distribution comparison classifiers of Cula and Dana [3], Leung and Malik [8] and Varma and Zisserman [14].

The Bayesian classifier of [7] is also divided into a learning stage and a classifi-

cation stage. In the learning stage, class priors and empirical filter response probabilities are learnt from the training data. Once again, we emphasise that to take into account the variation due to changing viewpoint and illumination a number of models will be used for each texture class, rather than just learning a single model per texture class as is done by [7,11]. In the classification stage, Bayes' theorem is invoked to calculate the posterior probability of a given filter response from a novel image belonging to a particular class.

4.1 A Bayesian classifier using the texton representation

The class conditional joint PDF of filter responses is obtained directly from the histogram of texton frequencies for the various images in the training set (of section 2). It is straight forward to implement the Bayesian classifier given this information. For a particular model we want to estimate the posterior

$$P(M_{ij}|I) = P(M_{ij}|\{\mathbf{F}(\mathbf{x})\}) \tag{1}$$

where $\{\mathbf{F}(\mathbf{x})\}$ is the collection of filter responses generated from the novel image I to be classified, M_{ij} is a particular model corresponding to training image number i taken from texture class number j, and the equality sign arises from the assumption that all the available information in the image has been extracted by the filtering process. The image I is classified as the texture j for which $P(M_{ij}|\{\mathbf{F}(\mathbf{x})\})$ is maximised over all models (i.e. all ij). Using Bayes' rule,

$$P(M_{ij}|\{\mathbf{F}(\mathbf{x})\}) \propto P(\{\mathbf{F}(\mathbf{x})\}|M_{ij})P(M_{ij}) \tag{2}$$

where $P(\{\mathbf{F}(\mathbf{x})\}|M_{ij})$ is the likelihood of the model M_{ij}, and $P(M_{ij})$ the prior on model M_{ij}. Since all models are equally likely in our case, the MAP class selection reduces to a Maximum Likelihood estimate, i.e.

$$\hat{M} = \underset{M_{ij}}{\textit{argmax}}\, P(\{\mathbf{F}(\mathbf{x})\}|M_{ij}) \tag{3}$$

If the filter responses are assumed spatially independent, then the probability of all the filter responses from the novel image belonging to the model M_{ij} is obtained by taking the product of the probabilities of the individual filter responses, i.e.

$$P(\{\mathbf{F}(\mathbf{x})\}|M_{ij}) = \prod_{\mathbf{x}} P(\mathbf{F}(\mathbf{x})|M_{ij}) \tag{4}$$

At this point, we take logs and focus on the log-likelihood to clarify the subsequent discussion,

$$\hat{M} = \underset{M_{ij}}{\textit{argmax}} \prod_{\mathbf{x}} P(\mathbf{F}(\mathbf{x})|M_{ij}) = \underset{M_{ij}}{\textit{argmax}} \sum_{\mathbf{x}} \log P(\mathbf{F}(\mathbf{x})|M_{ij}) \quad (5)$$

$$= \underset{M_{ij}}{\textit{argmax}} \sum_{k=1}^{S} N_k \log P(T_k|M_{ij}) \quad (6)$$

where N_k is the number of times the k^{th} texton occurs in the novel image labelling and $P(T_k|M_{ij})$ is the probability of occurrence of the k^{th} texton in the model M_{ij}. This equality follows because the log likelihood essentially amounts to counting the number of times each filter response falls in a particular bin – but this is exactly what is recorded in the texton frequency histogram.

4.2 Equivalence with Minimum Cross Entropy and KL Divergence

It will now be shown that Bayesian classification in this form can be viewed as nearest neighbour distribution comparison classification where the distance between two distributions is measured using Cross Entropy or the KL divergence. Cross Entropy is an information theoretic measure of the average number of bits required to encode symbols from a given alphabet using another alphabet. It is minimised if the same alphabet is used throughout. Based on this observation, we can use Cross Entropy to determine how similar a given distribution is to another distribution. The Cross Entropy between two discrete distributions, p and q, is given by $H(p,q) = -\sum p_k \log q_k$ and the smaller this value the better the match between the two distributions. The KL divergence is a related measure of the similarity between two distributions and is defined as $D(p\|q) = \sum p_k \log (p_k/q_k)$.

Nearest neighbour matching using Cross Entropy or KL divergence can be shown to be equivalent to the Bayesian formulation [1,15,16] by noting that from (6),

$$\begin{aligned}
\hat{M} &= \underset{M_{ij}}{\textit{argmax}} \sum_{k=1}^{S} N_k \log P(T_k|M_{ij}) \\
&= \underset{M_{ij}}{\textit{argmax}} \sum_{k=1}^{S} \frac{N_k}{\sum N_k} \log P(T_k|M_{ij}) \quad (7) \\
&= \underset{M_{ij}}{\textit{argmin}} -\sum_{k=1}^{S} P(T_k|M_I) \log P(T_k|M_{ij}) \quad (8) \\
&= \underset{M_{ij}}{\textit{argmin}}\ H(p,q)
\end{aligned}$$

where $p_k = P(T_k|M_I) = N_k/\sum N_k$ is the probability of occurrence of the k^{th} texton in the novel image labelling M_I and $q_k = P(T_k|M_{ij})$, i.e. the normalised texton frequency of the novel image and model image respectively.

Therefore, a Bayesian classifier which assumes uniform priors and the spatial independence of filter responses will give equivalent results to a nearest neighbour classifier based on the Cross Entropy between the texton distributions of the novel and model images.

The result can be straight forwardly extended to KL divergence by adding the constant $\sum_{k=1}^{S} p_k \log p_k$ to (8) and taking it inside the *argmin* operation as it does not depend on M_{ij}. Thus,

$$\begin{aligned} \hat{M} &= \underset{M_{ij}}{\textit{argmin}} \sum_{k=1}^{S} p_k \log p_k - \sum_{k=1}^{S} p_k \log q_k \qquad (9) \\ &= \underset{M_{ij}}{\textit{argmin}} \sum_{k=1}^{S} p_k \log \frac{p_k}{q_k} \qquad (10) \\ &= \underset{M_{ij}}{\textit{argmin}}\, D(p\|q) \end{aligned}$$

4.3 Relationship with χ^2

The equivalence can be taken further by noting that KL divergence and its extensions are bounded above by the χ^2 statistic and that the bound is attained when the two probability distributions being compared are very similar [6,13,15].

In more detail since the KL divergence is not a metric, it is often extended [13] to what is called a *capacitory discriminant* given by

$$C(p,q) = D\left(p\|\tfrac{1}{2}(p+q)\right) + D\left(q\|\tfrac{1}{2}(p+q)\right) \qquad (11)$$

where $\sqrt{C(p,q)}$ is now a metric [6]. Furthermore, Topsoe [13] has shown that C is bounded quite tightly according to the relation

$$\tfrac{1}{2}\chi^2(p,q) \leq C(p,q) \leq \ln 2 \cdot \chi^2(p,q) \qquad (12)$$

where

$$\chi^2(p,q) = \sum_{k=1}^{S} \frac{(p_k - q_k)^2}{p_k + q_k} \qquad (13)$$

Also, by taking the Taylor series expansion of C, it is straight forward to show that

$$\lim_{p \to q} C(p,q) = \frac{1}{2}\chi^2(p,q) \qquad (14)$$

up to second order in p and q.

Most of these results are by now standard. The significance for us is that by combining these equivalence results with the texton-bin correspondence shown in section 2, it becomes immediately clear that the Bayesian method of [7] is equivalent to the texton distribution comparison schemes of [3,8,14].

4.4 Bayesian Classification Experiments

The equivalence results from the previous subsections allow us to cast a Bayesian classifier as a nearest neighbour classifier with KL divergence as the distance measure. Thus, the experimental setup remains exactly the same as in section 3 except now the distance measure being used is KL divergence rather than the χ^2 statistic. Figure 9 plots the classification accuracy versus the size of the texton dictionary for the Bayesian classifier. The best results are for a dictionary of size S=1830 textons (i.e. $K = 30$ textons learnt from each texture class). A classification accuracy of 97.46% is achieved.

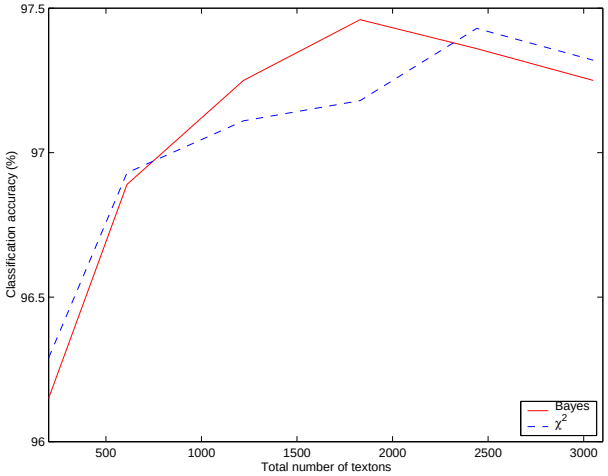

Fig. 9. The variation in classification performance with the size of of the texton dictionary using a Bayesian classifier: In each case, there are 2806 models and 2806 test images. The best classification result obtained is 97.46% using a dictionary of 1830 textons. The results of the texton based χ^2 classifier (i.e. figure 7) are also plotted for the sake of comparison.

A technical point about implementing probability products is that if the model histograms are determined directly from the textons frequencies of the training images then the classification accuracy of the Bayesian classifier is an astonishingly low 1.06%, i.e. almost all the test images are classified incorrectly. This is because most novel images contain a certain percentage of pixels (filter responses) which do not occur in the correct class models in the training set. This may be a result of an inadequate amount of training data, or due to outliers or noise. As a consequence, the posterior probability of these pixels is zero and hence when all the pixel probabilities are multiplied together the image posterior probability also turns out to be

zero.

This is a standard pitfall in histogram based density estimation and three solutions are generally proposed: (a) smoothing the histogram, (b) assigning small nonzero values to each of the empty bins, and (c) discarding a certain percentage of the least occurring filter responses in the belief that they are primarily noise and outliers. A combination of (b) and (c) is used here: instead of starting the bin occupancy count from 0, it is started from 1 to ensure that no bin is ever empty. Some of the least frequently occurring bins are also discarded. These modifications lead to the classification performance plotted in figure 9.

4.5 Comparisons

On the basis of experimental results, there is very little to choose between the Bayesian and distribution comparison classifiers using the texton representation. This is to be expected as, in essence, the test being performed is a comparison of χ^2 and KL divergence as distance measures. Nevertheless, while both classifiers attain rates of over 97%, there are different theoretical pros and cons associated with the two approaches.

There can be no doubt that theoretically, when the underlying distributions are known perfectly, the Bayesian classifier minimises the classification error. However, when we don't have enough data to accurately determine the true distribution or only have noisy approximations which suffer from the inherent quantisation effects of either clustering or histogram binning then the superiority of the Bayesian classifier is much less clear. This is evident from the capability of χ^2 to practically cope with empty bins (even though it is theoretically incapable of doing so) and noisy measurements while the Bayesian classifier completely collapses unless the probability distribution is modified. Furthermore, as can be seen from figure 9, the Bayesian classifier is often marginally surpassed by the nearest neighbour χ^2 classifier.

There is also the question about the Naive Bayesian assumption that the observed data is independent. However, in our case, this can not be considered a major drawback of the Bayesian classifier as compared to χ^2 because (a) χ^2 also makes the very same assumption in its derivation, (b) the experimental results indicate that extremely good classification results are obtained even when the assumption is violated (Schmid [11] notes that this holds true even for other texture datasets) and (c) if violating the assumption was leading to large errors then this could be tackled by randomly sampling filter responses from disjoint regions of the novel image in a bid to decrease their dependence.

Yet, despite their theoretical limitations, both classifiers appear to work extremely well in practise as is evidenced by the classification results.

5 Conclusions

In conclusion, we have shown that the texton representation of the PDF of filter responses is equivalent to an adaptive bin representation and, conversely, that every regularly partitioned bin representation can be converted into an equivalent texton representation. This has enabled us to use texton densities for texture classification in the Bayesian framework which itself, under certain circumstances, can be viewed as another measure of distance in a distribution comparison classification scheme. In doing so, we have brought together two seemingly unrelated schools of thought in texture classification – one based on the Bayesian paradigm and the other on textons and their first order statistics.

Acknowledgements

We would like to thank Alan Yuille for discussions on Bayesian methods and Phil Torr for discussions on density estimation. Financial support was provided by a University of Oxford Graduate Scholarship in Engineering, an ORS award and the EC project CogViSys.

References

[1] F. R. Bach and M. I. Jordan. Tree-dependent component analysis. In *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)*, 2002.

[2] M. J. Chantler, G. McGunnigle, and J. Wu. Surface rotation invariant texture classification using photometric stereo and surface magnitude spectra. In *Proceedings of the 11th British Machine Vision Conference, Bristol*, pages 486–495, 2000.

[3] O. G. Cula and K. J. Dana. Compact representation of bidirectional texture functions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1041–1047, December 2001.

[4] K. J. Dana and S. Nayar. Histogram model for 3d textures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 618–624, 1998.

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

[6] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1857–1860, July 2003.

[7] S. Konishi and A. L. Yuille. Statistical cues for domain specific image segmentation with performance analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 125–132, 2000.

[8] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, June 2001.

[9] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1988.

[10] Y. Rubner, C. Tomasi, and L.J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[11] C. Schmid. Constructing models for content-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 39–45, 2001.

[12] N. A. Thacker, F. Ahearne, and P. I. Rockett. The bhattacharya metric as an absolute similarity measure for frequency coded data. *Kybernetica*, 34(4):363–368, 1997.

[13] F. Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609, July 2000.

[14] M. Varma and A. Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, volume 3, pages 255–271. Springer-Verlag, May 2002.

[15] N. Vasconcelos and A. Lippman. A unifying view of image similarity. In *Proceedings of the International Conference on Pattern Recognition*, pages 1038–1041, 2000.

[16] P. Viola. *Alignment by Maximization of Mutual Information*. Aitr 1548, MIT, AI Lab, June 1995.

[17] A. Zalesny and L. Van Gool. A compact model for viewpoint dependent texture synthesis. In *Proceedings of the European Conference on Computer Vision*, LNCS 2018/5, pages 124–143. Springer-Verlag, 2000.