# Statistical Approaches to Material Classification

Manik Varma
Robotics Research Group
Dept. of Engineering Science
University of Oxford
Oxford, OX1 3PJ
manik@robots.ox.ac.uk

Andrew Zisserman
Robotics Research Group
Dept. of Engineering Science
University of Oxford
Oxford, OX1 3PJ
az@robots.ox.ac.uk

## Abstract

*The objective of this paper is classification of materials from a single image obtained under unknown viewpoint and illumination conditions. Texture classification under such general conditions is an extremely challenging task. Our methods are based on the statistical distribution of rotationally invariant filter responses in a low dimensional space.*

*There are two points of novelty: first, two representations of filter outputs, textons and binned histograms, are shown to be equivalent; second, two classification methodologies, nearest neighbour matching and Bayesian classification, are compared.*

*In essence, given the equivalence of texton and bin representations, the paper carries out an exact comparison between the texton based* distribution comparison *classifiers of Leung and Malik [IJCV 2001], Cula and Dana [CVPR 2001], and Varma and Zisserman [ECCV 2002], and the Bayesian classification scheme of Konishi and Yuille [CVPR 2000].*

*The comparisons are assessed by classifying images of* all *61 materials present in the Columbia-Utrecht database. Classification rates of over 97% are achieved for both the methods while classifying more than 2800 images in all.*

## 1 Introduction

In this paper, we investigate the problem of classifying materials from their imaged appearance, without imposing any constraints on, or requiring any *a priori* knowledge of, the viewing or illumination conditions under which these images were obtained. Classifying textures from a single image under such general conditions is a very demanding task.

A texture image is primarily a function of the following variables: the texture surface, its albedo, the illumination, the camera and its viewing position. Even if we were to keep the first two parameters fixed, i.e. photograph exactly the same patch of texture every time, minor changes in the other parameters can lead to dramatic changes in the resultant image (see figure 1). This causes a large variability in the imaged appearance of a texture and dealing with it successfully is one of the main tasks of any classification algorithm. Another factor which comes into play is the fact 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 both these factors which makes the texture classification problem so hard.

Weak classification algorithms based on the statistical distribution 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 function (PDF) of filter responses represented by a binned histogram [5], and (b) the nearest neighbour $\chi^2$ distribution comparison classifiers based on a texton frequency representation [2, 6, 9]. In this paper, we draw an equivalence between these two representations. We are then able to compare the performance of Bayesian and distribution comparison classifiers using either representation.

The success of Bayesian classification applied to filter responses was convincingly demonstrated by Konishi and Yuille [5]. Working on the Sowerby and San Francisco outdoor datasets, their aim was to classify image regions into

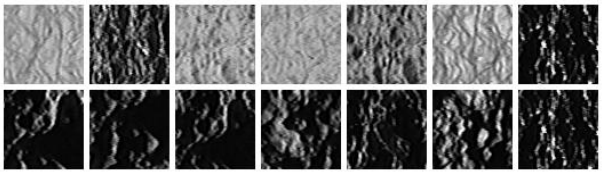

Figure 1: The change in imaged appearance of the same texture (# 30, Plaster B) with variation in imaging conditions. Top row: constant viewing angle and varying illumination. Bottom row: constant illumination and varying viewing angle. There is a considerable difference in the appearance across images.

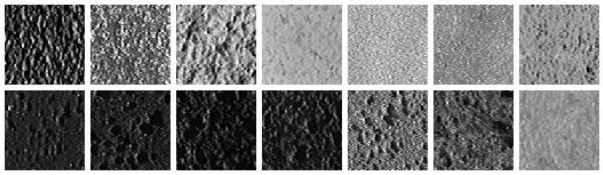

Figure 2: Small inter class variations between textures can make the problem harder still. In the top row, the first and the fourth image are the same texture while all the other images, even though they look similar, belong to different classes. Similarly, in the bottom row, the images appear similar and yet there are three different texture classes present.

one of six texture classes. The joint PDF of the class conditional likelihood 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 classified by computing their filter responses and using Bayes' decision rule. Finally, to classify an entire image or image region, pixel independence was assumed and pixel posterior probabilities multiplied together.

In contrast, distribution comparison classifiers such as [2, 6, 9] 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, Earth Mover's distance, KL divergence, etc., but the $\chi^2$ significance test, in conjunction with a nearest neighbour rule, is often preferred.

Leung and Malik [6] 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 viewpoints 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) [4] database. Later, Cula and Dana [2] and Varma and Zisserman [9] showed that 2D textons could be used to classify single images without any loss of performance.

Classification performance is evaluated here 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 [9] 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 [1, 2, 3, 6, 10]. 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 [5, 8] may fail in such a situation. The solution adopted here is to represent each texture class by probability distributions (models) conditioned implicitly on viewpoint and illumination. Hence, multiple models are generated from the various training images for each texture class and these models characterise the different appearances of the texture with variation in imaging conditions. Thus, each training image can potentially generate a model, and the choice of which models are used is based on a greedy algorithm which maximises classification performance over the training set [9]. On average, 7-8 models represent each texture class.

The layout 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 [9]. We also describe the two common representations, texton and binned histogram, of the joint PDF of filter responses and comment on their equivalence. Then, in section 3, we present a comparison between the texton and bin represen-

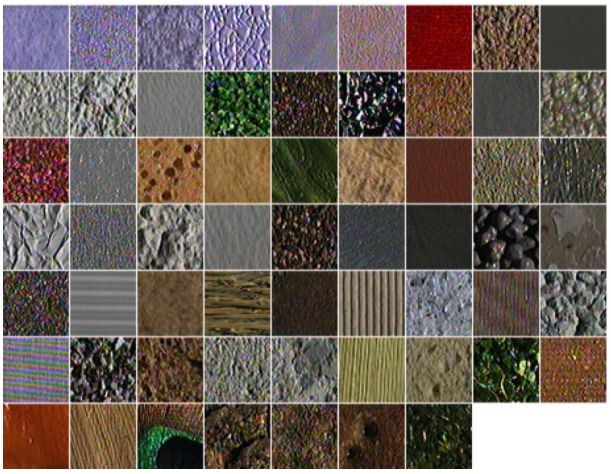

Figure 3: Textures from the Columbia-Utrecht database. All images are converted to monochrome in this work, so colour is not used in discriminating different textures.

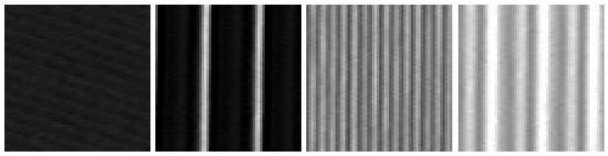

Figure 4: Images of ribbed paper taken under different viewing and lighting conditions. The material has significant surface normal variation and therefore changes its appearance drastically with imaging conditions thereby making single image classification an extremely difficult job.

tations using a distribution comparison classifier. Next, we implement a Bayesian classifier using a texton representation in section 4 and contrast its performance to the distribution comparison classifier. We conclude by discussing some of the advantages and limitations of the CUReT database.

## 2 Filter Responses and their Representation

In this section, we first describe the filters bank that is used here, and then discuss two popular representations of the filter responses, textons and binning, which we will show to be equivalent.

**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 novel filter bank which does this is the Maximum Response (MR8) filter bank which comprises 38 filters but only 8 filter responses (see figure 5). The filters include a Gaussian and a Laplacian of a Gaussian (LOG) filter, an edge (first derivative) filter at 6 orientations and 3 scales and a bar (second derivative) filter also at 6 orientations and 3 scales. The response 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 total.

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 angular information can still be recorded. Furthermore, we expect that more significant textons are generated when clustering in a low dimensional, rotationally invariant space. Further details of the filter bank, as well as pre and post image processing steps, are given in [9].

Both the Bayesian and the distribution comparison classifiers discussed in this paper are divided into two stages - learning and classification. In the learning stage, training images are convolved with the filter bank to generate filter responses. The representation of these filter responses is described next. This representation is the learnt statistical model for a given texture under particular imaging conditions.

**Texton representation of filter responses:** Each training image is convolved with the filter bank to generate a set of filter responses. These filter responses are then aggregated over various images from the texture class 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 here, 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. $K = 40$ textons are learnt from every texture class resulting in a total of $61 \times 40 = 2440$ textons. Hence, a model is a $2440$-vector where each component is the proportion of pixels which are labelled as a particular texton. Implementation details are given in [9].

**Histogram representation by binning:** In this representation, the model corresponding to a given image is the 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 while binning has the advantageous effect of preventing

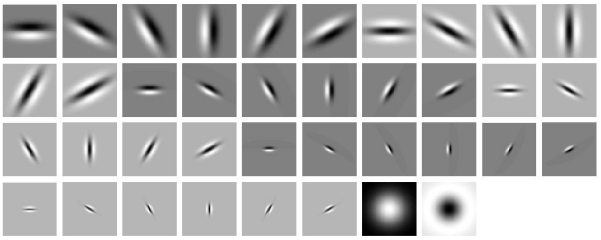

Figure 5: The MR8 filter bank consists of 2 oriented filters (an edge and a bar filter, at 6 orientations and 3 scales), and 2 rotationally symmetric ones (a Gaussian and a Laplacian of Gaussian). However, by taking the maximal response of the oriented filters across orientations and at each scale, only 8 filter responses are recorded.

over-fitting of the data it also changes the underlying distribution. Therefore, the number of bins and their spacing can be important parameters.

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 $\times$ (space required to store a bin value + space required to store a bin index). This is bounded above by $\mathcal{O}$(number of data points) and compares favourably to a naive implementation which stores the full matrix in $\mathcal{O}$(total number of bins) space, 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 $\chi^2$ distance can be evaluated in $\mathcal{O}$(number of non-empty bins) flops.

**Equivalence of the representations:** The two representations of filter responses can be made identical by a suitable choice of bins or textons. For example, for equally spaced bins, a 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. In essence, the comparison carried out in section 3 can be thought of as a comparison between two different texton dictionaries.

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 placed in accordance with the hyper plane boundaries demarcated by the various textons in filter response space (as determined by the Voronoi diagram). 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, which facilitates the comparisons made in section 4.

# 3 Classification by Distribution Comparison

**Classification:** 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 $\chi^2$ metric, is found and the novel image declared to belong to the texture class of the closest model.

In this section the effect of the representation on classification performance is investigated.

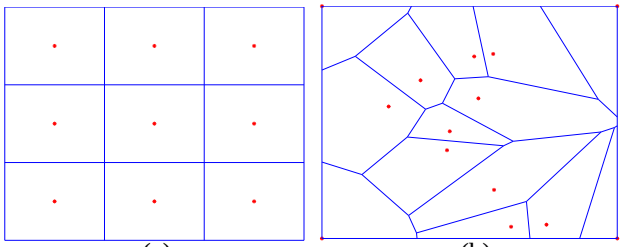


Figure 6: Texton and bin equivalence: (a) A bin representation can be converted into an identical texton representation by placing a texton at the centre of each bin. (b) Conversely, a texton representation can be converted into a bin representation where the bins are the Voronoi polygons.

**Experimental setup and results:** Classification is carried out on all 61 texture classes for both the representations. We first 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. Then, the number of models used is reduced by the Greedy algorithm while maintaining classification accuracy.

Using 46 models per texture for the texton based representation, the classifier achieves an accuracy rate of $97.43\%$ while classifying the $2806$ test images. Upon running the greedy algorithm, the number of models is substantially reduced to, on average, 7.14 per texture.

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 7 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. The greedy algorithm reduced the number of models used to, on average, 7.91 models per texture.

Both the representations give very similar classification results though the texton representation slightly outperforms the bin representation. The situation remains much the same even after the greedy algorithm is used to reduce the number of models. Of course, this is not surprising in the light of the fact that the two representations can be made identical (though they are not here).

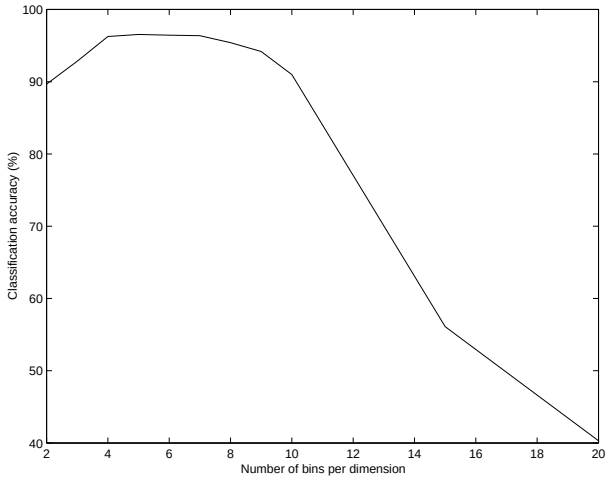


Figure 7: The variation in classification performance with the number of bins used during quantization: In each case, there are $46 \times 61 = 2806$ models and 2806 test images. The best classification results are obtained when the filter responses are quantized into 5 bins per dimension.

# 4 Bayesian Classification

Given that texton frequencies and histogram binning are an equivalent way of representing the PDF of filter responses, it is now possible to calculate the class conditional likelihood 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 [5] (and to some extent that of Schmid [8]) with the texton based distribution comparison classifiers of Varma and Zisserman [9], Cula and Dana [2] and Leung and Malik [6].

The Bayesian classifiers of [5, 8] are also divided into a learning stage and a classification stage. In the learning stage, class priors and filter response likelihoods are learnt from the training data. Once again, we emphasise that to take into account the variation due to changing viewpoint and illumination it is necessary to condition the likelihood on the imaging conditions rather than learning a single likelihood per texture class as is done by [5, 8], i.e. a number of likelihood models will be used for each texture class.

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. A naive Bayesian classifier which assumes the class conditional independence of filter responses is then employed to determine the posterior probability of the novel image as a whole belonging to a particular class.

**A Bayesian classifier using the texton representation:** Since we have already computed the histogram of texton frequencies for the various images in the training set, we already have all the class conditional likelihoods of filter responses and it is straight forward to implement the Bayesian classifier. First, class priors are learnt by counting the number of pixels present in each texture class and normalising this distribution (which turns out to be uniform in our case). Next, each texton frequency histogram model defines a texture subclass (the texture class is the set of all models for that class) and yields the likelihood of a particular filter response belonging to that subclass. Finally, given the filter responses of a novel image, Bayes' theorem is used to determine the posterior probability of a particular filter response belonging to a specific subclass. If the filter responses are assumed independent, the posterior probability of all the filter responses from the novel image belonging to the same subclass can be obtained by taking the product of the posterior probabilities of the individual filter responses. The novel image is classified as belonging to the subclass (and therefore texture class) with the maximum posterior probability.

**Experiments:** The experimental setup was kept exactly the same as in the previous section. Using the same training, test and textons sets, the classification accuracy of the Bayesian classifier when using 46 models per texture was an astonishingly low 1.06%. Almost all the test images were classified incorrectly. The reason for this was that most novel images contained a certain percentage of pixels (filter responses) which did not occur in the correct class models in the training set. This could be as a result of inadequate amount of training data as compared to the number of textons in the representation, outliers or noise. As a consequence, the posterior probability of these pixels was zero and hence when all the pixel probabilities were multiplied together the image posterior probability also turned out to be zero.

This is a standard pitfall in non-parametric 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) improved the classification performance dramatically. Instead of starting the bin occupancy count from 0, it is started from 1 to ensure that no bin was every empty. The 1% of the least frequently occurring bins are also discarded. Under these conditions the Bayesian classifier achieved an accuracy rate of 97.36% while using 46 models per texture. The greedy algorithm reduced the number of models to, on average, 7 per texture.

**Comparisons:** There is very little to choose between the Bayesian and distribution comparison classifiers using the texton representation. Both attain classification rates over 97% while using roughly 7 models per texture. And yet, there are different theoretical pros and cons associated

with the two approaches.

The biggest theoretical drawback of the Bayesian paradigm is the assumption that the filter responses are independent. However, this can be tackled by randomly sampling filter responses from disjoint regions of the novel image in a bid to decrease their dependence. Another problem, albeit one that was successfully addressed here, is of the non-parametric representation of empty bins.

The $\chi^2$ classifier appears to be a lot more forgiving towards this problem and no modification had to be made to cope with zero texton frequencies. However, $\chi^2$ has its own theoretical limitations [7] and the various criteria for using the $\chi^2$ probability function must be fulfilled. Also, $\chi^2$ is not sensitive to shuffling the bin order, something which is important when textons are interrelated. Of course, a different comparison method could be used but its conditions too would need to be fulfilled.

However, 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 the bin representation and that every texton representation can be converted into an identical bin representation. This lets us use non-parametric texton densities for texture classification in the Bayesian framework.

The Columbia-Utrecht database used here has many advantages over Brodatz textures but also some limitations. As compared to Brodatz, it is by far the superior database, as it has many real world textures photographed under varying image conditions. One can actually see the effects of specularities, shadowing and other surface normal variations, unlike in Brodatz where there is only a single viewpoint and illumination available for each image.

The limitations of the database are mainly in the way the images have been photographed and the choice of textures. For the former, there is no significant scale change for most of the textures and limited in plane rotation. Also, because the photographs were taken under controlled conditions, the illumination is somewhat contrived as there is very little change in illuminant intensity or number of illuminants. As regards choice of texture, the most serious drawback is that multiple instances of the same texture are present for only a very few of the materials, so intra-class variation cannot be investigated. Hence, it is difficult to make generalisations. Finally, almost all the textures can be classified on the basis of their first order statistics. There are almost no instances of textures having the same first order statistic but different higher order statistics.


## Acknowledgements

We are grateful to Phil Torr for discussions on non-parametric 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] M. J. Chantler, G. McGunnigle, and J. Wu. Surface rotation invariant texture classification using photometric stereo and surface magnitude spectra. In *Proc. 11th British Machine Vision Conference, Bristol*, pages 486–495, 2000.

[2] O. G. Cula and K. J. Dana. Compact representation of bidirectional texture functions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

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

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

[5] S. Konishi and A. L. Yuille. Statistical cues for domain specific image segmentation with performance analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[6] 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.

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

[8] C. Schmid. Constructing models for content-based image retrieval. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

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

[10] A. Zalesny and L. Van Gool. A compact model for viewpoint dependent texture synthesis. volume 2018 of *Lecture Notes in Computer Science*, pages 124–143. Springer, July 2000.