Scale Descriptors : Texture Description by Salient Scales

Timor Kadir, Michael Brady
Robotics Research Laboratory, Department of Engineering Science,
University of Oxford,
Parks Road, Oxford OX1 3PJ, U.K.

October 2, 2002

Abstract

We recently proposed the Scale Saliency method for extracting salient features together with their scales. This paper supports two hypotheses about the Scale Saliency algorithm. First, the technique chooses scales that are relevant to local image characteristics; second, that it is a more general measure of feature saliency and scale compared to conventional methods, such as those based on the wavelet transform, because it is defined independent of any particular feature or basis morphology. To this end, the Scale Saliency algorithm is applied to texture description. We present a novel approach to texture characterisation, called Scale Descriptors, which represents textures using just the PDF of their salient scales. We demonstrate comparable discrimination performance to conventional ‘complete’ feature-based texture descriptors, such as those based on the wavelet transform. Such methods typically use a combination of texture properties (although, in most cases these properties are measured implicitly), for example texton orientation, morphology, density and regularity. A comparable discrimination performance employing just one property, in this case texton scale, indicates a superior measure of that property.

We show that for many textures, salient scales themselves make good descriptors of texture. In some cases, the discrimination performance is far superior to conventional methods. Furthermore, Scale Descriptors are largely invariant to translation, rotation and uniform photometric variations (robust to non-uniform), and handle zoom in a natural way.
Keywords

Texture Description, Saliency, Scale Space, Salient Features, Texton Scale, Scale Saliency, Salient Scales.

1 Introduction

Many types of image content can be described well through the characterisation of their scale-space properties. Textured regions, such as those found in aerial images, are a case in point. Conventional approaches to texture description attempt to capture multi-scale behaviour by extracting descriptions at multiple scales, usually within a pre-set range of scales, for example using linear scale-space (Koenderink, 1984, Lindeberg and ter Haar Romeny, 1994, Witkin, 1983) or wavelet transform methods (Mallat, 1998). Though the resulting description represents the texture at many scales, relatively few provide significantly new information.

The main need for such a multi-scale representation is that existing methods cannot measure accurately the scales of perceptually salient features in the general case. The problem arises from the assumptions about saliency and scale in such methods, namely that salient features are those for which the convolution with a scaled basis function gives a large response. A corollary to this is that a feature is deemed salient only if its morphology resembles that of one of the available basis functions. When this is not the case, that is when there is a mis-match between the basis function and feature morphologies, energy is distributed across a number of scales, resulting in spurious measurements.

In this paper, we argue that in practice, in the vast majority of cases, few significant or salient scales carry sufficient information for texture description and matching, and that feature scale and saliency can be measured better by the Scale Saliency algorithm introduced previously by the authors (Kadir and Brady, 2001). In the case of texture characterisation, the main advantage of this algorithm is that the definitions of scale and saliency are independent of any particular basis morphology. Features are considered salient if they are simultaneously unpredictable in some feature and scale space.

We support our hypothesis by posing the following problem: given that we know that scale plays a significant role in the definition of a texture, can we define useful descriptions of a texture which are comprised solely of its characteristic or salient scales? Conversely, can we validate a given method’s capability for selecting salient features (and their scales), if the extracted scale information alone can be used to describe texture regions? Our approach is to generate descriptors of texture from the local PDF of salient scales. We contend that this provides stronger evidence for our hypothesis.
than if descriptions were extracted at fewer ‘better’ scales, compared to conventional methods for multi-scale texture analysis.

The paper is organised as follows. Following a brief summary of the Scale Saliency algorithm in Section 2, we discuss the roles of scale and saliency in the characterisation of textures in Section 3. We review prior work on texture description and modelling with an emphasis on those methods that capture scale characteristics in Section 4. After outlining our approach to texture characterisation in Section 5, in Section 6 we describe our texture Scale Descriptor method. In Section 7 we illustrate some of the properties of the Scale Descriptors. After discussing various texture description tasks and test conditions in Section 8, we present our results in Section 9. The technique is applied to supervised texture classification and unsupervised segmentation tasks, using a variety of synthetic and real data.

2 Salient Scale Selection

In this section, we briefly describe the Scale Saliency approach to salient feature selection. The benefits are that it offers a general definition of saliency independent of any particular basis morphology, and that it is robust. Moreover, the selected features are largely invariant to translation, rotation, scale and uniform photometric variations (robust to non-uniform). The algorithm is described in detail in (Kadir and Brady, 2001), so is summarised briefly here.

2.1 Saliency as Local complexity

Gilles (1998) investigated the use of salient local image patches or ‘icons’ for matching and registering two images. He defined saliency in terms of local signal complexity or unpredictability. More specifically, he estimated saliency using the Shannon entropy of local attributes. Figure 1 shows local intensity histograms from a number of image segments. Areas corresponding to high signal complexity tend to have flatter distributions, hence higher entropy. More generally, high complexity of any suitable descriptor can be used as a measure of local saliency. Local attributes, such as colour or edge strength, direction or phase, may be used.

Given a point $x$, a local neighbourhood $R_X$, and a descriptor $D$ that takes values from $\{d_1, \ldots, d_r\}$ (e.g. in an 8 bit grey level image $D$ would range from 0 to 255), local entropy is defined as:

$$H_{D,R_X} = - \sum_i p_{D,R_X}(d_i) \log_2 p_{D,R_X}(d_i)$$

where $p_{D,R_X}(d_i)$ is the probability of descriptor $D$ taking the value $d_i$ in the local region $R_X$. 

3
Underlying this definition of saliency is the assumption that complexity is rare in real images. This is generally true, except in the case of pure noise or self-similar images (e.g. fractals) where complexity is independent of scale and position. Although complexity in textured regions is more prevalent than in non-textured regions, the assumption of spatially and scale localised features still holds. Exceptions include so-called noise-like or statistical textures. We have not applied our method to these since such textures typically lack well-defined ‘features’ (and hence scales). Although Gilles was primarily interested in aerial images, he demonstrated the use of saliency on a number of different image types.

However, Gilles’ original method has a number of drawbacks. The most important is that it requires the specification of a local window scale over which the local PDF is estimated; it cannot measure saliency over the scale dimension.

In (Kadir and Brady, 2001), we extended Gilles’ technique to analyse saliency in both feature space and scale dimensions. Our method works as follows: as in the Gilles method, for each pixel location, calculate entropy from the PDF of a descriptor; a circular window is used to sample the image. This process is repeated for increasing scales; that is we grow circular windows. Scales are selected at which the entropy is peaked, then the entropy value is weighted by some measure of the self-dissimilarity in scale-space of that feature.

The reasoning is as follows: we seek local salient features based on what is deemed complex, as measured by entropy (predictability). Features that exist over large ranges of scale exhibit self-similarity in their PDFs, which, in feature space, we regard as non-salient. In extending our saliency measure to scale, we prefer to detect features that exist over a narrow range of scales. Therefore, we weight the peaked entropy with a measure of the statistical self-dissimilarity. In practice, we simply use the sum of absolute difference between the histograms near the entropy peak.

In the continuous case the saliency measure $\mathcal{Y}_D$ as a function of scale $s$ and position $x$ is defined as follows:

$$\mathcal{Y}_D(s_p, x) \triangleq \mathcal{H}_D(s_p, x) \times \mathcal{W}_D(s_p, x)$$

(2)

where entropy $\mathcal{H}_D$ is defined by:

$$\mathcal{H}_D(s, x) \triangleq -\int_{d \in D} p(d, s, x) \log_2 p(d, s, x) \, dd$$

(3)

and where $p(d, s, x)$ is the probability density as a function of scale $s$, position $x$ and descriptor value $d$ which takes on values in $D$, the set of all descriptor values. The “saliency over scale” measure,
\( W_D(s, x) \), is defined by:

\[
W_D(s, x) \triangleq s \times \int_{d \in D} \left| \frac{\partial}{\partial s} p(d, s, x) \right| \, \dd d
\]  

(4)

The vector of scales \( s_p \), at which entropy peaks, is defined by:

\[
s_p \triangleq \left\{ s : \frac{\partial^2 \mathcal{H}_D(s, x)}{\partial s^2} < 0 \right\}
\]  

(5)

In the discrete case, \( \mathcal{H}_D, W_D \) and \( s_p \) become:

\[
\mathcal{H}_D(s, x) \triangleq - \sum_{d \in D} p_{d, s, x} \log_2 p_{d, s, x}
\]  

(6)

\[
W_D(s, x) \triangleq \frac{s^2}{2s - 1} \times \sum_{d \in D} |p_{d, s, x} - p_{d, s-1, x}|
\]  

(7)

\[
s_p \triangleq \{ s : \mathcal{H}_D(s - 1, x) < \mathcal{H}_D(s, x) > \mathcal{H}_D(s + 1, x) \}
\]  

(8)

The algorithm generates a space in \( \mathbb{R}^3 \) (two spatial dimensions and scale) sparsely populated with scalar saliency values. In essence, the method searches for scale-localised features with high entropy, currently subject to the constraint that scale is isotropic. The method therefore favours blob-like features. Alternatively, we can relax the isotropic requirement and use anisotropic regions. This has the drawback of increasing the dimensionality of the saliency space. Moreover, the relatively simple notion of scale as a single parameter is lost. However, for some features, such as those with local linear structure, this may be necessary to correctly characterise local scale behaviour. In this case two scales may be used to analyse the feature; one tangential and one perpendicular to the direction of the feature. This is the subject of ongoing research, and in this paper we concentrate on isotropic features. Such features are useful for matching because they are locally constrained in two directions. Features such as edges or lines only locally constrain matches to one direction (of course depending on their length).

It should be noted however, that the method does detect non blob-like features, but these are considered less salient than their isotropic equivalents. In the case of linear structure, this is because there is a degree of self-similarity in the tangential direction. The selected scale is determined predominantly by the spatial extent of the feature in the perpendicular direction. The spatial extent in the tangential direction of such anisotropic regions could be analysed by a post-processing grouping algorithm. In solving the correspondence problem, this ranking is somewhat desirable as it reflects the information gained by matching each type of feature.
3 Scale and Saliency in Textures

The texture analysis and description problem has received a great deal of attention within the vision community (Bigun, 1990, Cross and Jain, 1983, Haralick et al., 1973, Randen and Husøy, 1999, Unser and Eden, 1989). The applications of such techniques are broad and include aerial image analysis (Manjunath and Ma, 1998), medical image analysis (Wu et al., 1992) and image retrieval (Manjunath and Ma, 1996).

There have been several suggestions for a definition of texture. Though none seem to completely encompass the wide variety of images that humans would consider as textures, they do nevertheless provide a useful insight. Jain (1988) proposes the following definition:

The term texture generally refers to repetition of basic texture elements called texels.

The texel contains several pixels, whose placement could be periodic, quasi-periodic or random. Natural textures are generally random, whereas artificial textures are often deterministic or periodic. Texture may be coarse, fine, smooth, granulated, rippled, regular, irregular, or linear.

This idea of a texture being composed of repeated elements called texels or textons, first suggested by Julesz (1981), implies a high degree of redundancy, hence predictability, at some set of scales. Clearly, this is related to, and has implications for, our approach to scale and saliency. We now briefly discuss some of these issues.

The texton idea suggests several notions of characteristic scale. For example, in so-called structural textures, those consisting of well-defined tessellated primitives, the scales of the texture primitives and the scales of their tessellations might be characteristic. These two sets of scales are embodied in the terms ‘granularity’ and ‘density’ often used to describe such textures.

Other types of texture consist of less well-defined primitives and exhibit smooth wave-like variations. In such cases, a better approach might be to measure the period of the wave-like function.

Finally, in so-called noise-like or statistical textures, the primitive is the individual pixel and, at first glance, scale seems almost completely undefined. Such textures are more conveniently considered as samples drawn from probability distributions of pixel values. However, even in this case, a local window must be used to estimate the local PDF and its size must be set appropriately for the texture. This scale can be considered to be somewhat (albeit weakly) characteristic of the texture. A more discriminating set of scales might be those at which particular pixel values are repeated. The co-occurrence matrix (Haralick et al., 1973) attempts to capture this information.
The above ‘taxonomy’ of textures is one of the fundamental problems associated with texture description. Often, proposed methods may perform satisfactorily with only one or two types of texture. The problem is caused essentially by the discrete nature of the imaging process. Image features exist as separate entities over specific ranges of scale (Witkin, 1983) and since the (digital) imaging process is by definition discrete, the available scales of observation can only capture a subset of all the ‘possible’ features within the real world scene. At a specific range of scales, determined by the field of view and the resolution of the camera, certain features that might otherwise be characteristic of a texture might not exist as separate entities within the image.

As an example, consider the image in Figure 3 from the Vicky1 sequence originally used in (Kadir and Brady, 2001) where we applied our Scale Saliency technique to a simple recognition task. Fine scale texture regions, such as the trees in the background and the grass in the foreground, were found to be the relatively non-salient parts of the image. This is because they are more self-similar in feature and scale space (relative to the other regions in the image) at the range of scales available for analysis. These scales are constrained by the focal length and the resolution of the camera. This may not have been the case had the camera been closer to those objects or the resolution much higher.

In this paper, we analyse textured regions at such close focal lengths, where the textons are clearly visible and are the more salient of the features in the image. The more general problem of obtaining a more ‘complete’ image description is the subject of ongoing investigation.

Texture scale is important not only for texture characterisation and modelling but is also a powerful cue for depth perception, and hence useful for shape from texture tasks (Blake and Marinos, 1990, Lindeberg and Garding, 1993). In this paper, we restrict our analysis primarily to fronto-parallel textures and only consider Extended Euclidean or Similarity group of transforms such as translation, rotation and uniform scaling (zoom). Future work will consider more general imaging conditions, for example those that can be described with the affine group of transforms.

4 Measuring Scale in Texture

In this section, we review previous work in texture analysis concentrating on those techniques that attempt to capture scale behaviour. Each method assumes, either implicitly or explicitly, definitions of scale and saliency. These clearly have a significant impact on the resulting descriptors and so we attempt to highlight what these are. Where appropriate, we discuss which image features are used and any assumed generative model.
4.1 Morphological Methods

Morphological operators have a geometric interpretation which makes them a convenient tool for texture analysis. They are of interest to our work because they attempt to measure texton scale directly.

One morphological approach to texture characterisation is that of Volumetric analysis (Vincent and Dougherty, 1994). Here, the image is treated as a surface which is progressively morphologically eroded using increasing sizes or scales of the structuring element. The volume versus scale relationship captures the characteristic scales of the texture primitives. The process is conceptually that of sieving the image with different sized sieves to measure texton granularity.

The method has two main problems. The first is that an assumption has to be made about the foreground and background intensities in order to form the image surface. This means that the definition of the texton depends upon the grey-level value. Two textures sharing the same texton pattern but inverted in intensity will be analysed differently. The second problem is that when applied to grey level images, morphological operators require the setting of the structuring element grey-level. This grey-level must be set appropriately for the texture otherwise an accurate measure cannot be taken. It effectively sets the saliency threshold for grey-levels.

4.2 Filter banks, Transforms and Basis Projection

The characterisation of textures by features extracted by multiple filters has had a long history within the field (Bajcsy, 1973, Laws, 1980, Xie and Brady, 1996), using a variety of filters such as Gabors (Jain and Farrokhnia, 1991) and the closely related Gaussian, Gaussian Derivative and Laplacian pyramids (Bigun, 1990, Lindeberg and Garding, 1993, Unser and Eden, 1989). More recently, wavelet methods have been used and now dominate the field (Mallat, 1998).

All such methods may be considered from one of three perspectives: filterbanks, coordinate transform or projection onto a basis set (Mallat, 1998). The widely-applied Fourier analysis methods may also be placed within this framework. As such they operate in a similar fashion. The image is filtered at multiple scales and orientations at every pixel location. The number of orientations may be reduced through the use of steerable filters (Freeman and Adelson, 1991). The responses to these filtering operations at each pixel are treated as multi-dimensional vectors which are then used to characterise the texture.

Recognition is then simply a matter of choosing an appropriate similarity metric and classification algorithm. Many similarity metrics have been used including Sum Square Difference (SSD), Mahalanobis Distance, $\chi^2$ or Kullback-Leibler. Usually the measurement in the test image is taken.
in a local neighbourhood to enable a statistically significant measure. Similarly, many classification
algorithms have been applied including clustering, Bayesian, Fuzzy, and Neural Networks.

There appears to be no general consensus with regards to which particular filter or wavelet basis
is best. Randen and Husøy (1999) present a detailed evaluation of many texture descriptors based
on feature extraction. Their conclusion was that no particular set of filters consistently outperforms
all others. Largely, it seemed to depend on the particular types of texture under test, reinforcing
the view that the analysis must be adapted to the specific texture. However, there were a number
of methods which consistently under-performed. Gabor filters were outperformed in most cases, and
the over-complete wavelet schemes outperformed those that were non-redundant. More recently,
deRivaz (2000) has reported improved results using the Dual Tree Complex Wavelet Transform
(DT-CWT) applied to the Randen and Husøy test set.

The main problem with the wavelet and filterbank methods is that a large number of filtering
operations must be carried out in order to sufficiently characterise the image. This results not only in
a computationally expensive algorithm but also large amounts of data to be processed. Furthermore,
the resultant feature space is likely to be highly under-sampled — the so-called curse of dimension-
ality problem. This necessitates the application of some form of dimensionality reduction algorithm
such as Principal Component Analysis (PCA). Such methods not only reduce dimensionality of the
data but can also infer a statistical model from the measurements.

One of the causes of this high dimensionality is the fact that a number of the parameters such
as orientation and scale are unnecessarily left free. Appropriate scale selection can significantly
reduce the number of measurements and furthermore can improve the subsequent matching stages
by ‘fixing’ an otherwise free parameter (Kadir and Brady, 2001, Lindeberg, 1993a,b). Orientation
analysis can also be used in a similar way.

A solution to this problem could be to choose a scale or set of scales for each pixel location. For
example, one way to do this could be to choose the largest magnitude wavelet coefficient (Lindeberg,
1993b, Starck and Murtagh, 1999). Since each coefficient is the resultant of a convolution of the
local image with a scaled version of the mother wavelet, large magnitude coefficients would indicate
the presence of a significant texture feature at that scale.

The problem with using the wavelet transform in such a way is that there is an implicit assumption
that saliency and scale are defined with respect to a particular basis set. A technique that determines
feature saliency and scale based on coefficient magnitude tends to select features whose morphology
matches one from the available basis set. Since textures can contain many types of feature, the
resulting saliency and scale information would not be particularly discriminating if only part of the

9
decomposition was kept.

This is a key part of our argument and deserves further analysis, particularly as there are several different approaches to using wavelet-based methods in feature detection. All make slightly different assumptions about the relationship between the basis set and feature saliency and scale.

At one extreme, it is assumed that all image features of interest will resemble some scaling, translation and rotation of a particular basis function. A salient feature is then one for which a large magnitude wavelet coefficient results and its scale is simply the scaling of the mother wavelet at which this large response occurred.

Alternatively, a more pragmatic assumption is that salient features do not correspond directly to any particular scaling of a single wavelet basis but rather can be synthesised from a number of such scalings; that is, to deliberately design a basis set such that no single scaling of the Mother wavelet corresponds to image features. Rather, some small sub-set of all the wavelet scalings, translations and rotations can succinctly capture local signal behaviour. Most feature detection methods typically make one of these two assumptions, though rarely state explicitly which.

At the other extreme, it can be assumed that feature morphologies do not match with the basis set at all and total mis-match occurs; the convolution of the feature with the basis function scalings causes energy in many, if not all, scales. In such a case, dominant coefficient magnitudes cannot be used as a measure of saliency and some other method must be used. One possibility is to use the phase information, as is used in Phase Congruency (Kovesi, 1999)\(^3\). In fact, if such band-spreading did not occur then a measure of Phase Congruency alone would not detect any features. For example, the application of Phase Congruency to an image grating pattern composed of a single frequency sine wave would not result in any detected features. Since the mismatch of features and basis set is much more likely than perfect matches in the general case, a method that uses phase information to determine saliency is likely to be more general than one using magnitude.

The particular assumptions made by any feature detection method should reflect the prior knowledge available for a given application. For example, if the types of feature of interest can be well modelled by a particular basis morphology then using the first of the above assumptions is perfectly reasonable, and is in fact advantageous for noise resilience. Wavelet basis sets have been found that work well with a wide variety of low-level image features of interest such as edges, lines, Corners and roof-edges.

However, we contend that in order to work well and robustly for a wide class of textures there should be few a-priori assumptions about the nature of the textures. For texture analysis, our goal is to succinctly characterise the textons, their scale or the inter-texton spacing. Given the wide
range of possible textures, hence the wide variety of texton features that we may encounter, it is very unlikely that a particular basis function will match most of these. Therefore we cannot rely on conventional filterbank or wavelet methods to select the perceptually salient features (and hence their scales). We can, however, assume that our basis set can synthesise many of the features of interest, but in this case we must maintain the entire scale decomposition.

Wavelet Packet methods (Coifman and Wickerhauser, 1992, Laine and Fan, 1993) attempt to overcome this problem by trying many different basis functions. However, even these can only try a relatively small number of basis functions and they also necessitate a significant increase in computational load.

It follows then that one should keep all the measured data and use some data reduction technique to maintain a manageable feature volume and to generalise models from it. We mentioned PCA earlier as an example, next we review more general statistical models which can also be used for the same purpose.

4.3 Statistical Models

Statistical texture models consider the texture surface, or at least its appearance, as having been formed by some stochastic process and characterise it by estimating the underlying PDF of the process. The stochastic model may be parametric, for example assuming the PDF to be of a Gaussian form, or non-parametric. One of the benefits of statistical texture models is that they can be generative. Another advantage is that an appropriate model can significantly reduce the image data into a manageable form whilst allowing for (and in some cases explaining) the variations observed in the real image.

Markov Random Fields (MRF) are a class of statistical model that have been applied to a wide range of image processing and computer vision tasks (Besag, 1986, Geman and Geman, 1984, Li, 1995). Their application to texture modelling was first suggested by Hassner and Sklansky (1980) and soon after by Cross and Jain (1983); most subsequent techniques have essentially adopted their main ideas.

The basic assumption is that the value of a pixel (or site in the MRF lexicon) is probabilistically conditional only on the values of its neighbours. This assumption is reasonable for modelling textures because of their main defining property: textures contain some kind of self-similarity at some particular scale or set of scales. The size of the neighbourhood and the possible configurations of pixels within this neighbourhood contributing to this conditional probability are constraints assumed in a particular model. To discriminate between only a few types of texture that are well distinguished,
for example by texton size and directionality, a fairly small neighbourhood system may be used. More complex textures require that larger neighbourhoods and more pixel configurations must be considered resulting making training more difficult and rapidly increasing the computational cost — the main Achilles heal of the technique.

Despite their high computational cost, there is certainly convincing evidence that a sufficiently complex MRF (i.e. one with a sufficiently large neighbourhood and considering all relevant cliques) can model very challenging textures. One example is the impressive texture synthesis work of Efros and Leung (1999) using an algorithm originally proposed by Garber (1981) but dismissed at the time as impractical. Their approach is to consider the source texture patch as an MRF and sample from it directly. Their algorithm builds new textures pixel by pixel by re-calculating the conditional probabilities from the source patch at every stage. The results are very impressive, but reported computation time is in the order of tens of hours. A newer algorithm by Efros and Freeman (2001) called Texture Quilting has provided significant computational speed-ups.

4.4 Hierarchical Markov Models

Multi-resolution Markov Random Fields (MMRF) have been proposed to overcome these computational problems (Jeng, 1992, Lakshmanan and Derin, 1993). The idea is that the MRF is solved at a coarser level (that is, for a sub-sampled version of the image) and then at progressively finer levels up to the original resolution of the image. However, the main problem with such methods is that the spatial Markovian property is lost through the sub-sampling procedure except in a few special cases (Lakshmanan and Derin, 1993).

An alternative approach to this problem is to replace the spatial Markovian requirement with one across scale, hence creating a Markov Tree. This was proposed by Bouman and Shapiro (1994) and Luettgen et al. (1993) who extended previous work of Basseville et al. (1992). They propose a tree structure to represent the image information at multiple resolutions where each node is conditionally dependent only on the values of its parents and independent of the other nodes from the same parent. The Markovian property across scale is reasonable since sub-sampled versions of a signal will be correlated with the original signal.

4.5 Combined Transform and Multi-scale Statistical Models

Recently, methods have been proposed that apply multi-resolution Markov models to wavelet decomposed data (Basseville et al., 1992, De Bonet and Viola, 1998, Luettgen and Willsky, 1995, Portilla and Simoncelli, 2000). Significant computational savings may be achieved if the transform goes some
way to decorrelating the image data — in essence, it presents the image to the Markov model in a more compact form and hence a much simpler statistical model can be used. One such transform is the wavelet transform and as a multi-scale method it fits well with the multi-scale statistical models.

Examples of such approaches vary from the parametric approaches of Portilla and Simoncelli (2000), to the non-parametric methods of De Bonet and Viola (1998). The Portilla and Simoncelli (2000) algorithm tries to model explicitly different properties of a texture by applying a number of different constraints to the texture model at the training stage. The De Bonet and Viola (1998) algorithm uses a non-parametric approach and employs a multi-scale statistical representation which is very similar to those of Basseville et al. (1992), Luettgen and Willsky (1995). The model does not try to fit a parametric model to these probabilities, rather to place them all in what is called a multi-dimensional flexible histogram, a kind of vector quantisation of the cross-scale conditional probabilities.

It is interesting to note that the Markov Tree is used to model cross scale dependencies, despite the decorrelating properties of the wavelet transform. Several authors have commented on this issue (Buccigrossi and Simoncelli, 1999, De Bonet and Viola, 1998). Buccigrossi and Simoncelli (1999) present a rigorous analysis of the joint statistical properties of images in wavelet space, through scale, position and orientation. One of their observations is that although the raw coefficients are generally decorrelated, the magnitudes of the coefficients are highly correlated.

There are two relevant issues here. First, only non-redundant wavelet transforms have this optimal decorrelation property. Most of the texture description methods use over-complete, that is redundant, wavelet decompositions. Second, the correlation only measures second-order statistical dependencies. High-order relations may still be present. The Markov tree models these dependencies. Lastly, the cross-scale dependencies are there precisely because of feature and basis mismatch. The only case where the scales will be probabilistically independent is if the morphology of the image features exactly matches with those of the basis set.

The success of the combination of multi-scale transform methods with multi-scale statistical models is due primarily to the complimentary properties of the two methods. MRF statistical models can model even quite complex textures; the work of Efros and Leung (1999) has shown that. However, they are computationally very expensive. Multi-scale transform methods such as the wavelet transform can capture some of the essential qualities of textures such as scale, directionality and texton shape, and represent this information in a very compact form. However, as discussed in Section 4.2, most transform methods cannot capture such qualities independent of the feature morphology because they implicitly assume definitions of saliency, scale and orientation with respect
to a particular basis morphology. This leads to the ‘spreading’ of responses across different scales or, in general, any measurement class of interest (e.g. orientation). The information across the measurement classes is not independent. This effect causes any texture class feature space to be cluttered with spurious measurements making discrimination more difficult. The application of Markov Tree methods to an efficient representation of the image, as generated by a wavelet transform, significantly reduces the required complexity of statistical model for a given complexity of texture.

A final observation about these methods is that they can be considered as ‘continuously hybrid’ statistical and feature-based texture models. By this we mean they automatically adapt between statistical models and feature-based models according to the texture. In the case where the texture comprises textons which correspond exactly to the morphology of the available basis functions, the hybrid Markov model acts as little more than a probabilistic feature detector. Where the texture is a purely statistical one, the wavelet transform has little effect and the hybrid Markov model operates in a similar to conventional MRF approaches. In between these two extremes the model acts in a hybrid manner.

4.6 Miscellaneous methods

Any discussion of scale and texture would not be complete without at least a mention of Fractal methods (Mandelbrot, 1977). The scale self-similarity property of Fractals makes them an promising candidate for texture modelling. For example, Xie and Brady (1992) use an estimate of the Fractal Dimension to drive a segmentation algorithm. However, despite their theoretical attraction and the initial surge of interest after their introduction to signal processing, it appears they have been superseded by other methods. One reason for the lack of success of Fractal-based texture descriptors is the difficulty associated with measuring their properties. Perhaps another reason is that the belief that many textures can be represented in the Fractal model is unfounded.

An approach similar in spirit to Volumetric Analysis and not dissimilar the one presented in this paper, is that of Sporring and Weickert (1999). Their idea is to capture the characteristic scales of a texture (more generally any image) by measuring its information content using the Generalised Renyi entropy (Renyi, 1961), at different levels of scale in linear (Lindeberg and ter Haar Romeny, 1994, Witkin, 1983) and non-linear (Perona and Malik, 1988, Weickert, 1997) scale-spaces.

The authors demonstrate the use of their method for global scale selection and size estimation of features and propose a ‘fingerprint’ description for textures in which the variation in entropy over scale-space is used to characterise the texture. However, the method is global and requires that a foreground and background intensity be defined through appropriate setting of a parameter.
This latter problem derives from the interpretation of pixel values as probabilities. This could be avoided if the probabilities were obtained from a PDF of local intensities. However, based on this assumption, the authors prove some useful properties, such as decreasing entropy with scale which, in the general case, is not true for entropy calculated from the PDF of local intensities.

5 Salient Scale Descriptors

The role of scale in texture modelling and description is widely acknowledged to be significant. However, the review in the previous section suggests that conventional methods of measuring scale, such as those based on transform techniques, are unsatisfactory because they are implicitly tied to a particular basis set. The crux of the problem is that both scale and saliency are intrinsically linked to the morphology of the available bases of the particular transform. Features are only salient if their form matches that of the basis set and their scale is only measured accurately in this case.

The Scale Saliency algorithm, described in Section 2, defines scale and saliency in an information theoretic framework which is independent of a particular feature morphology. Saliency is defined as unpredictability of a particular descriptor measured in feature space and scale. In essence, scale is a constraint placed on the spatial configurations of pixels. For example, in most of our saliency experiments, we have constrained scale to be isotropic, with one parameter controlling the radius of a circular sampling kernel which we have referred to as scale. This one-parameter model may be generalised by a N-parameter one. For example, using such a definition of scale, the algorithm will generate an anisotropic scale space parameterised by the major and minor axes of an ellipse plus a rotation.

The feature morphology independence properties of our Scale Saliency method make it a promising candidate for scale-based texture description. Our hypothesis is that since our method can measure feature scale and saliency in a more general way, it should result in more accurate characterisation of the texton scales than is achievable with conventional multi-scale transform methods alone.

We may adopt a number of different strategies to test our hypothesis. The most obvious is to use our algorithm to select salient features and their scales and use descriptors of these to model the texture. This is simply an extension of the recognition experiments we carried out in (Kadir and Brady, 2001). However, the problem with this is that we cannot explicitly test the accuracy of the scale information we extract. It would be unclear how much discrimination power comes from the scales and how much from the salient texture patches.
Instead, we pose a slightly different question. Given that we know that scale information can be characteristic of a texture and that our method can capture this information more accurately than conventional methods: can the salient scales that our method selects be used to characterise a texture?

Our approach is to apply our algorithm to a texture image, then use the most salient scales in the saliency space as a description of the texture. Obviously, such an approach will only work on those textures that can be discriminated by scale; texton scale is not the only defining quality of a texture. For more complete texture characterisation texton density, orientation and regularity and morphology must all be measured. However, as demonstrated in our experiments, a surprising number of textures can be discriminated by such a method.

6 The Algorithm

In this section, we describe the Scale Descriptor algorithm for characterising textures. The basic idea is to generate a PDF of salient scales for each texture class. In the implementation described in this paper, the PDF is approximated using a histogram. An alternative approach could be to use a Parzen window algorithm. Recognition is then simply a matter of finding the best matching PDF; we have used a simple SSD measure for this purpose. Placing the salient scales in a probabilistic representation provides for the possibility of using Bayesian techniques for matching (though we have not yet investigated this).

The main steps in the algorithm are:

1. Generate \( Y_D(s, x) \) by applying the Scale Saliency algorithm as described in Section 2 to the image.

2. Apply local thresholding to \( Y_D(s, x) \).

3. For each point in \( x \), approximate a PDF of salient scales using a local window of size \( w \).

For the training stage, the last step should be modified to generate one PDF for the whole texture. This may be done by averaging all the local PDFs, better still by applying a PCA type method to the ‘stack’ of local PDFs. The latter approach creates a more flexible model by capturing the principal moments of variation as well as the average set of salient scales in the PDFs across the texture. Such an approach may be used to characterise ‘allowed’ intra-texture class scale variations or those due to projective distortion. In our experiments, we have used an average PDF generated from the global set of salient scales of training textures.
For the test image, each point has an approximation to the local PDF of salient scales, which will be used to describe the local texture. Of course, in common with other methods, we require that the size of the local window should be large enough to capture a representative sample of the texture. However, in our case, the setting of this is less sensitive because we are compiling a local PDF of scales which themselves may be much bigger than the local window used. The size of this window is determined predominantly by the inter-texton spacing, something which our basic algorithm does not measure. In the experiments presented in this paper, we assume that this local window size is fixed and chosen manually. The selection of this window is clearly related to the saliency scales and characteristics of the particular texture under test. This is the subject of current investigation and techniques to automatically select this ‘extrinsic’ or semi-local scale parameter will be reported later.

As an enhancement to the basic technique, we note that some textures may have similar PDFs of scales, but that their respective saliencies can be quite different. Therefore, we replace the one dimensional PDF described above with a two dimensional one which captures the joint distribution of characteristic scales and their saliencies.

In a further enhancement to the basic technique, the counts in the histograms are weighted by the saliency values. This reduces the effect of the hard cut-off of the local thresholding stage. The result is a smoother histogram of salient scales.

A pseudo-code description detailing the generation of all four types of Scale Descriptors (Hist1D, Hist1DW, Hist2D, Hist2DW) is provided below (i.e. step 3 in above):

```
01 HIST1D[1..N]=Initialise histogram accumulator array of size N
02 HIST1DW[1..N]=Initialise histogram accumulator array of size N
03 HIST2D[1..N, 1..N]=Initialise histogram accumulator array of size NxN
04 HIST2DW[1..N, 1..N]=Initialise histogram accumulator array of size NxN
05 For each point, (x,y) in the a local window, W, in the image do:
06     For each scale, s between SMIN and SMAX do:
07         {                                 
08             HIST1D[s x N/SMAX]+=1
09             HISTSUM+=1
10             HIST1DW[s x N/SMAX]+=YD(s,x,y)
11             HISTWSUM+=YD(s,x,y)
12             HIST2D[s x N/SMAX, YD(s,x,y) x N/MAXYD ]+=1
13             HIST2DW[s x N/SMAX, YD(s,x,y) x N/MAXYD]+=YD(s,x,y)
14         }
15     For each bin, (i) in HIST1D do: HIST1D[i]=HIST1D[i]/HISTSUM
16     For each bin, (i) in HIST1DW do: HIST1DW[i]=HIST1DW[i]/HISTWSUM
17     For each bin, (i,j) in HIST2D do: HIST2D[i,j]=HIST2D[i,j]/HISTSUM
18     For each bin, (i,j) in HIST2DW do: HIST2DW[i,j]=HIST2DW[i,j]/HISTWSUM
```
where $\mathcal{YD}(s, x, y)$ is the input saliency space, and $\text{MAX}YD$ is the maximum saliency value in $\mathcal{YD}(s, x, y)$.

Figure 4 shows an example of the Scale Descriptor method applied to the D101 and D102 Wicker patterns from the Brodatz (1966) set of reference textures. The salient scales are highlighted in the images of the textures. We have applied the clustering algorithm described in (Kadir and Brady, 2001) to the saliency space in order to provide a clear demonstration of the method, though this step is not necessary for the algorithm proper. There are two main texton scales in the patterns which can be seen clearly both in the 1D and 2D salient scale histograms (brightness represents probability in the latter).

Figure 5 shows the 1D salient scale histograms for pairs of samples taken from a number of simple textures. The numbers at the bottom of the figure give the best matches between all twelve samples.

Figure 6 shows examples of the 2D joint scale and saliency histograms generated from a selection of Brodatz textures. These Scale Descriptors can discriminate between texture with similar sets of salient scales but at different saliencies.

### 7 Properties of Scale Descriptors

In this section, we discuss some of the main properties of the Scale Descriptors, namely, the robustness of the technique to variations in intensity and rotation, and also the effect of zoom. Properties such as these are desirable for texture descriptors, indeed for image descriptors in general, because in most applications, hypotheses should be unaffected by changes in the imaging conditions such as illumination intensity or camera position. In practice though, descriptors cannot be made totally invariant to all imaging conditions, however robustness to a reasonable set of variations is achievable and useful.

The results for these experiments are shown in Figure 7. In all the experiments reported in this paper, the Scale Descriptors are built from salient scales obtained by application of the Scale Saliency algorithm operating on the intensity PDF. Two specific implementations of the Scale Saliency algorithm were tested. The first used a 16 bin histogram to estimate the local PDF of intensities; the second, a Parzen window technique with a Gaussian kernel of $\sigma = 25^4$. Three experiments are presented: a 50% scaling in intensity, a $45^\circ$ rotation (clockwise) and a scaling of 50% (sub-sampling by two). For clarity, only the 1D Scale Descriptors are shown. We discuss each case below.

The first experiment was concerned with the resilience of the method to variations in intensity. First, we discuss shifts in intensity. In the continuous case, shifts in intensity do not cause a change in entropy because it is intensity invariant. However, since a histogram is only a discrete approximation to a PDF, a shift will have a small effect on the entropy value. The size of this effect is governed
by the histogram bin size. While the absolute saliency value will change a little, the salient scales themselves will remain relatively unaffected by shifts since they are determined by peaks in entropy (with respect to scale), even in the histogram case. PDF estimators such as the Parzen window method do not suffer from the problems associated with histograms, but are computationally more expensive. If the shift is large then the pixels might saturate. In such a case, there might be a loss of features which will be reflected in the distribution of salient scales. However, we would expect that the less salient features would be the first of the features to disappear, thus leaving the majority of the salient scales intact.

For scalings in intensity, there would be a change because of the effective decrease in the spread of the PDF due to quantisation of the intensity values. However, it could be expected that the distribution of salient scales would remain stable, since these are generated from where entropy attains a peak with respect to scale. Therefore, the 1D Scale Descriptors should be robust to scalings in intensity. For the 2D Scale Descriptors, that is those capturing both scale and saliency, we would expect some variation due to the change in saliency values, however a Parzen window PDF estimator would minimise this effect.

The ‘Intensity Scaling’ plot in Figure 7 shows the effect on the 1D Scale Descriptor of a 50% scaling of the image intensity values. As expected, the distribution of scales remains largely unaffected for both the Parzen window and histogram cases.

The second experiment tested robustness to rotation. Here, we would expect that the method is robust because the saliency algorithm uses circular local windows to sample the image. This is confirmed in the ‘Rotation’ plot in Figure 7, which shows the effect of a 90° clockwise rotation of the image on the salient scale histogram.

Such robustness raises an interesting point. In some cases of intensity change, such as inversion or severe scaling, it is debatable whether the texture should be classed as the same. Certainly if an image were made up of patches of the same texture with various intensity transforms applied to it, the human visual system would probably class each patch separately. Probably, this is because humans are able to use multiple cues to cluster features and segment regions, and, of course, they also apply contextual information. In the case of automated recognition and classification of real images with textured regions, such as aerial images, it may be beneficial to separate out the variations due to scale, from those of other attributes. We can, of course, add extra dimensions to the scale descriptor to measure other cues such as average intensity and orientation.

Finally, we discuss the effect of zoom. Since the Scale Descriptors are essentially a representation of the texton scales, we would expect that a change in scale, for example due to a change in focal length of a camera, would cause a shift of the scale distribution along the scale axis of the descriptor.
In the 2D case, a similar effect would be observed. However, it should be noted that for a large change in the focal length, effects due to the imaging process might cause problems. For example, what might have been one feature at scale \( x \), could become two or more features at scale \( 2x \) or vice-versa. This effect would cause a change in the descriptors. In practice though, as long as the camera has sufficient resolution and the change in zoom not too great, these problems can be minimised.

The ‘Zoom’ plot in Figure 7, shows the effect of zoom on the salient scale histogram. In this case the texture was sub-sampled by two, but the size of the sample window was kept constant for a fair comparison. Both sets of 1D Scale Descriptors show a shift in the dominant scale, from 5 to 3, as expected. As with the previous experiment, there is some spreading of the histogram.

This zoom property could be used in a number of ways. For example, a classification algorithm could use it to identify textures observed at scales quite different from those at which the database was trained. Here, the shape of the distribution is the important discriminant. Normally, texture databases must be trained over a range of expected scales of observation, hence this method could provide substantial computational savings. A logical progression of this idea is to develop ‘families’ of textures, where the distribution of scales (amongst other properties) defines membership. For example, all Polka-dot patterns would be members of the same family of textures. The zoom property could also be used to improve tracking of textures in sequences by analysing the effects on the Scale Descriptor. For example, if a textured object were to move towards the camera, a shift to the right in the corresponding Scale Descriptor could be expected.

8 Testing Texture Descriptors

In this section, we briefly discuss commonly-adopted practices for testing texture descriptors and outline our approach. There are two main objectives for our experiments. First, we wish to demonstrate that texton scale can be a key visual cue in the characterisation of textures. Second, that our Scale Saliency algorithm is a more appropriate measure of texton scale and saliency than other conventional approaches. For these two objectives the results will need to show that discrimination using salient scales alone is at least as good as conventional methods.

There are several texture analysis tasks and conditions that we might consider in testing the Scale Descriptors. For example, we might choose a classification task, where pixels are classified as one of a number of known texture classes. In some cases there is only one texture per test image, while in others multiple textures are combined to make up a texture mosaic. MeasTex (1998) is a texture test suite package widely used to test and compare classification performance. Alternatively a segmentation task might be more appropriate. This may be considered as a classification task
with additional constraints, such as piecewise smoothness, imposed on the solution. In supervised segmentation knowledge about the problem, for example the expected number of classes, is supplied to the algorithm under test. In the unsupervised case this must be determined by the algorithm. Texture synthesis is applicable only to methods which incorporate a generative model. A model is deemed accurate if a human cannot distinguish the synthesised example from the source texture — “a type of texture Turing test” (Cross and Jain, 1983).

We must also specify an appropriate set of test conditions for the experiments. For example, we must decide on the particular texture test set to use, their number, and type. Variations in imaging conditions are also an often overlooked factor; texture test sets are commonly created using fronto-parallel views. A more general set of test conditions could use the Affine group where rotations, anisotropic scalings and skewings are allowed. For example, Chetverikov (2000) extracts an Affine invariant texture descriptor through characterising the global autocorrelation of the intensity function and in a related approach Schaffalitzky and Zisserman (2001) use the second moment matrix in a two-step procedure, first skew-normalising a local image patch and then applying scale and rotation invariant descriptors.

Another important factor is the position of the texture with respect to sources of light. For example, Chantler (1995) demonstrated that typical classification methods based on filter banks such as Gabors can fail catastrophically in such conditions.

We have used two main tasks in our experiments: a classification task using two and five texture mosaics, and an unsupervised segmentation task using a variety of mosaics and real images. For the former we have used the test set used by Randen and Husøy (1999) who carried out a detailed study of feature-based texture classifiers. By using this test set we are able to compare the performance of our method with that of a wide range of published algorithms. The second task is in some ways more difficult because the texture classes must be learnt at run-time. However, this test is less demanding of the discriminatory power of the descriptor because only neighbouring regions are compared.

9 Results

9.1 Classification of Randen and Husøy Test Set

In this section, we report results of applying the Scale Descriptor method to the Randen and Husøy (1999) test-set. The aim here, is to demonstrate that the Scale Descriptors can classify a range of textures with a performance comparable to those of standard methods. Since the technique uses only scale information, it cannot be expected to work on all textures. However, where texton scale provides a powerful perceptual discriminator, good results should be achieved.
The test images consist of two-, five-, ten- and sixteen-texture mosaics. We have used only the two- and five-texture ones. The textures are from the Brodatz (1966), MeasTex (1998) and VisTex (1998) databases. Figures 8 and 9 show the set of test mosaics.

The experimental conditions were as follows. Scale Saliency was used with 8-bin histogram PDF estimation and run over scales $1 \rightarrow 30$ (circles of $3 \rightarrow 61$ pixels in diameter). The local threshold was 75% (by saliency value) applied using a running 8x8 window. 32-bin Scale Descriptors were used ($32 \times 32$ for 2D case) and were generated using a 45 pixel circular window. Four Scale Descriptor variants were tested: one dimensional (Hist1D), one dimensional with weighting (Hist1DW), two dimensional (Hist2D), and two dimensional with weighting (Hist2DW).

The classification test was SSD (sum-square-difference), and for efficiency, only one in five pixels were classified. The implementation of Scale Saliency we used does not operate within $s_{max}^2$ pixels of the image edges (in these experiments this corresponds to 30 pixels). Therefore, the classification was similarly limited.

### 9.1.1 Histogram Equalisation

The original Randen and Husøy (1999) test images were histogram equalised. This was to ensure that the algorithms under test could not utilise the local mean or variance of the pixel values. However, this is problematic for the Scale Descriptors, particularly the 2D variants. The reason is that the 2D Scale Descriptors utilise the saliency information to increase discrimination. Histogram equalisation spreads out the global histogram of grey-values, thereby increasing the global entropy. The result is that the entropy of the textures is made more similar. This effectively nullifies the discrimination power gained through the use of the saliency distribution information in the 2D descriptors. For the two-texture test cases, this makes little difference, because sufficient discrimination is available with the 1D Scale Descriptors. However, in the five-texture cases this is problematic, especially where the textron scales are not clearly distinct. Therefore, where possible, we have used the textures without histogram equalisation. For mosaics 8a, 8b, 8c, and 9a, the histogram equalised mosaics have been used as the original images could not be obtained. For all other cases, non-equalised images have been used.

We contend that since the Scale Descriptor method does not use local mean or variance information, the results are comparable to the published results of Randen and Husøy (1999). We did not re-implement their algorithms, therefore we could not ascertain whether histogram equalisation improved or degraded the performance of the methods under test. For some methods, such as the filterbanks, the increased contrast after histogram equalisation might improve classification. However, other methods, such as the co-occurrence matrix, might benefit from local mean and variance...
Randen and Husøy generated their training and test images from non-overlapping parts of the source texture. However, they did not specify exactly which parts were used for which. In re-generating the test-set, we have used the top-left $300 \times 300$ square of the source image for the test mosaics and the largest remaining square for the training images. 30 pixels of overlap were allowed since this area is not used in the classification stage (due to the border left by the Scale Saliency algorithm).

9.1.2 Classification Results

The classification results for four Scale Descriptor implementations are given in Table 1. The labels `Hist1D', `Hist1DW', `Hist2D' and `Hist2DW' refer to the one dimensional, one dimensional with weighting, two dimensional, and two dimensional with weighting Scale Descriptors respectively. The reported results for the co-occurrence and QMF (f16b) methods from Randen and Husøy (1999) and DT-CWT from deRivaz (2000) are also given in the table. The QMF (f16b) results are representative of the best of all the results reported by Randen. The full classification maps for the Scale Descriptors are provided in Figures 10 and 11.

For completeness, we have also run the algorithms on the histogram equalised images. These are not intended for the evaluation of the Scale Descriptors, but have been included because they provide a useful insight into the operation of the 2D descriptors. These results are provided separately in Table 2.

Mosaic-based tests such as these include the misclassifications caused by texture region boundaries. While the inclusion of such effects represents a more realistic test scenario, the results do not measure the absolute classification performance. Another problem is that the magnitude of these inter-regional misclassification errors is affected by the particular textures used and their arrangement in the mosaic. Therefore, in order to get an idea of the baseline classification performance of the Scale Descriptors, the five textures that make up each mosaic were classified separately. The average classification performance without boundary effects is provided in Table 3. Neither Randen nor deRivaz report such results, therefore we do not use these for direct comparison.

9.1.3 Discussion

The main results are given in Table 1. Overall, they show a comparable performance to the best of the Randen and Husøy (1999) and deRivaz (2000) algorithms and in some cases they significantly outperform all methods.
Table 1: Main classification results using the Randen 2 and 5 mosaic test-set. Numbers are percentage errors in classification, i.e. a low value means that few pixels were incorrectly classified. (HE) indicates that mosaic is histogram equalised.

<table>
<thead>
<tr>
<th>Mosaic</th>
<th>Co-occurrence</th>
<th>QMF-f16b</th>
<th>DT-CWT</th>
<th>Hist1D</th>
<th>Hist1DW</th>
<th>Hist2D</th>
<th>Hist2DW</th>
</tr>
</thead>
<tbody>
<tr>
<td>8a(HE)</td>
<td>1.9</td>
<td>8.1</td>
<td>0.6</td>
<td>19.6</td>
<td>17.2</td>
<td>17.1</td>
<td>14.5</td>
</tr>
<tr>
<td>8b(HE)</td>
<td>4.8</td>
<td>0.8</td>
<td>1.1</td>
<td>1.3</td>
<td>0.9</td>
<td>4.5</td>
<td>4.4</td>
</tr>
<tr>
<td>8c(HE)</td>
<td>3.3</td>
<td>8.2</td>
<td>9.3</td>
<td>1.3</td>
<td>1.6</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>9a(HE)</td>
<td>9.9</td>
<td>8.7</td>
<td>10.9</td>
<td>24.6</td>
<td>21.6</td>
<td>22.3</td>
<td>17.2</td>
</tr>
<tr>
<td>9b</td>
<td>27.0</td>
<td>18.9</td>
<td>21.8</td>
<td>38.7</td>
<td>34.7</td>
<td>25.7</td>
<td>16.1</td>
</tr>
<tr>
<td>9c</td>
<td>26.1</td>
<td>23.3</td>
<td>16.2</td>
<td>34.6</td>
<td>32.5</td>
<td>20.1</td>
<td>19.0</td>
</tr>
<tr>
<td>9d</td>
<td>51.1</td>
<td>18.4</td>
<td>16.6</td>
<td>49.5</td>
<td>52.6</td>
<td>25.3</td>
<td>21.5</td>
</tr>
<tr>
<td>9e</td>
<td>35.7</td>
<td>17.2</td>
<td>17.3</td>
<td>37.5</td>
<td>53.0</td>
<td>19.5</td>
<td>38.1</td>
</tr>
</tbody>
</table>

In general, the weighted descriptors outperform their non-weighted counterparts and the 2D descriptors outperform the 1D ones. In the case of the latter (2D vs. 1D), this difference is primarily due to the extra discrimination power provided by the saliency information. It is interesting to compare the results in Table 1 to those of the histogram equalised experiments, shown in Table 2. In the latter, the performance of the 2D descriptors about the same as the 1D versions, in some cases worse. This seems to validate the discussion in the previous section on the effects of histogram equalisation. That is, equalisation acts to maximise entropy in images, thereby causing all the textures to exhibit similar saliency. This effectively nullifies the extra discrimination power of the 2D descriptors.

We now discuss each case in detail.

First, we analyse the two-texture tests. These mosaics were histogram equalised. For mosaic 8a, the results for all the Scale Descriptors are significantly worse than those for the benchmark methods. However, this is not an altogether surprising result given the very similar scales of the textons in the component textures. The primary discriminating property of these two textures is directionality. Since the Scale Descriptors are invariant to rotation and the texton scales are very similar, a relatively poor classification performance could be expected. This result could be improved by replacing the isotropic Scale Saliency implementation, with an anisotropic version.

In test cases 8b and 8c the performance of the Scale Descriptors is very good. In fact, in the latter case, the Scale Descriptors significantly outperform all other methods. This is very interesting especially considering the comment made by Randen and Husøy:
<table>
<thead>
<tr>
<th>Mosaic</th>
<th>Hist1D</th>
<th>Hist1DW</th>
<th>Hist2D</th>
<th>Hist2DW</th>
</tr>
</thead>
<tbody>
<tr>
<td>8a</td>
<td>19.6</td>
<td>17.2</td>
<td>17.1</td>
<td>14.5</td>
</tr>
<tr>
<td>8b</td>
<td>1.3</td>
<td>0.9</td>
<td>4.5</td>
<td>4.4</td>
</tr>
<tr>
<td>8c</td>
<td>1.3</td>
<td>1.6</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>9a</td>
<td>24.6</td>
<td>21.6</td>
<td>22.3</td>
<td>17.2</td>
</tr>
<tr>
<td>9b</td>
<td>36.0</td>
<td>34.9</td>
<td>30.5</td>
<td>33.7</td>
</tr>
<tr>
<td>9c</td>
<td>35.9</td>
<td>35.1</td>
<td>34.6</td>
<td>30.3</td>
</tr>
<tr>
<td>9d</td>
<td>42.9</td>
<td>40.2</td>
<td>38.7</td>
<td>37.8</td>
</tr>
<tr>
<td>9e</td>
<td>35.9</td>
<td>39.5</td>
<td>43.9</td>
<td>49.7</td>
</tr>
</tbody>
</table>

Table 2: Classification results using the histogram equalised Randen 2 and 5 mosaic test-set. Number are percentage errors in classification.

<table>
<thead>
<tr>
<th>Mosaic</th>
<th>Hist1D</th>
<th>Hist1DW</th>
<th>Hist2D</th>
<th>Hist2DW</th>
</tr>
</thead>
<tbody>
<tr>
<td>9a</td>
<td>15.7</td>
<td>14.6</td>
<td>9.2</td>
<td>7.8</td>
</tr>
<tr>
<td>9b</td>
<td>33.9</td>
<td>31.0</td>
<td>23.7</td>
<td>18.9</td>
</tr>
<tr>
<td>9c</td>
<td>21.3</td>
<td>22.1</td>
<td>11.2</td>
<td>9.8</td>
</tr>
<tr>
<td>9d</td>
<td>33.0</td>
<td>38.3</td>
<td>9.5</td>
<td>13.4</td>
</tr>
<tr>
<td>9e</td>
<td>13.3</td>
<td>23.8</td>
<td>2.0</td>
<td>20.0</td>
</tr>
</tbody>
</table>

Table 3: Average classification results for Randen 5 mosaics without boundary errors. Number are percentage errors in classification.

We note that although Fig 12c [8c] looks simple, it was very difficult to discriminate for several of the approaches.

For these particular test cases, the QMF and co-occurrence results represent the best of all the methods tested by the authors. The two textures in this case, D5 and D92 from the Brodatz test set, can be clearly discriminated by their texton scales. Yet both filterbank methods, the QMF-f16b (full rate Quadrature Mirror Filter) and the DT-CWT (Dual Tree Complex Wavelet Transform) perform rather poorly compared to the Scale Descriptors. We contend that the difference is due primarily to basis-related definitions of saliency and scale used in such methods. The Scale Saliency method, though biased towards isotropic features by the scale constraint, captures the perceptually salient textons and their scales.

Initially then, these results seem to support our hypotheses. The relatively poor performance for
the 8a mosaic in fact provides weight to the argument that the Scale Saliency algorithm really does measure texture scale. In other words, since there is not a distinct perceptual difference between the two textures in terms of their texton scales, a good classification result would have been indicative of discrimination power coming from a cue other than scale. Cases 8b and 8c complement this result by demonstrating good classification where the texton scales are clearly distinct.

In the five-texture experiments, it is interesting to note that while the 2D joint scale and saliency Scale Descriptors perform reasonably well, the 1D ones do not. This is mainly due to the limited discriminatory power of the 1D descriptors. If they were tested in pairs, many of the textures would be quite easy to discriminate using the scales alone. Several options are available to further increase discrimination power. For example, the anisotropic version of Scale Saliency could be used, or a measure of texton density developed. Possibly the most powerful would be to use the salient features themselves.

Another general point is that it is a little unfair to use a plain classification algorithm in a task that would clearly benefit from the application of a few spatial constraints. Many of the misclassifications in these results are due to the inter-region boundaries. A multi-scale segmentation algorithm would provide significant gains here.

The first case, 9a, is quite interesting because the Scale Descriptor method is quite significantly outperformed by the other methods. Examination of the classification map, shown in Figure 11, indicates that most of the errors are arising in the inter-region boundaries. The baseline classification performance given in Table 3 is 7.8%, which confirms this hypothesis. Reducing the size of the local window used to generate the Scale Descriptors should alleviate this problem in this case, but would degrade the performance of the other test cases. An alternative solution would be to use a segmentation algorithm that competes at region boundaries, for example Region Competition (Zhu and Yuille, 1996). Perhaps better performance could be achieved with the non-histogram equalised versions of the textures.

In test case 9b the Scale Descriptors outperform all other methods. This can be explained by the quite disparate scales of the textons in the textures. Examining the four image maps (see Figure 11), we see that the 1D histograms do quite well for the textures in the left, top and centre segments (Fabric.0000, Fabric.0017 and Leaves.0013 from (VisTex, 1998)). Perceptually, it is clear that they have quite distinct texton scales. However, the remaining two segments in the right and bottom (Flowers.0002 and Leaves.0006) cause problems for the 1D descriptors. Neither texture has a dominant set of scales associated with it. The 2D descriptors can discriminate between them, because they utilise the difference in the distribution of saliency values.

The performance in cases 9c and 9d are approximately on par with the other methods, with
the latter performing slightly worse. In Figure 11 we can see that for the Hist2DW descriptors the majority of the errors come from the boundaries near the left segment (Fabric.0007). It is likely that this texture shares similar texton scales to those of its neighbours and hence gets misclassified at the segment edges. The non-boundary performance for this case, Hist2DW = 13.4% confirms this.

The final test case, mosaic 9e, is a very interesting result because unlike most of the other cases, the weighted descriptors perform significantly worse than the unweighted ones. In most cases, the weighting improves the results because it reduces the effect of lower-saliency features on the distribution of salient scales. However, in this case, the weighting has the opposite effect. The classification maps for the 2D and 2DW descriptors for this case reveal that the primary cause of the high misclassification is the texture in the centre segment (Rock.0005), whose main distinguishing features are the faint vertical lines or columns. These elongated structures have a relatively low scale-based saliency. The coarse binning of the histograms in the version of Scale Saliency used in these experiments is likely to further compound the problem. The resultant low saliency values for these features means that they do not contribute very much to the final description in the weighted histograms. In the non-weighted versions, all scales contribute equally therefore the Hist2D results are significantly better.

9.2 Unsupervised Segmentation

In this section, we present results for two sets of experiments using the Scale Descriptors to drive the unsupervised segmentation algorithm described in (Kadir and Brady, 2002). In the first, the algorithm has been applied to multi-texture mosaics, similar to those used in Section 9.1. In the second, ‘real’ images have been used. The aim of these experiments is to demonstrate that the Scale Descriptors can induce and generalise texture classes from small patches, or exemplars, of a texture. These exemplars correspond to the seed regions defined by the segmentation initialisation.

For the purposes of the discussion here the algorithm described in (Kadir and Brady, 2002) can be considered to be a standard implementation of Region Competition (Zhu and Yuille, 1996) with the exception that a boundary term is not included in the cost functions. Furthermore, two regions models are presented in (Kadir and Brady, 2002) — a standard Gaussian parametric model and a novel non-parametric model. Our initial experiments indicated that the Gaussian model is too simplistic for all but the most basic texture images. Therefore, the non-parametric region model has been used exclusively. Both the 1D and 2D Scale Descriptors are used, but only the weighted versions since these generally outperform the non-weighted ones. The Level Sets code was suitably modified to handle both the 1D and 2D region models.

Local Thresholding for all of these experiments is 70% by saliency value in a $8 \times 8$ window. All
images are 256 × 256 pixels, unless otherwise stated. The segmentation initialisation shown in Figure 12a has been used throughout the experiments.

9.2.1 Texture Mosaics

Figure 12b shows the segmentation results on a simple two-texture Brodatz mosaic. In this case, the 1D Scale Descriptor was used. The segmentation is good, but the boundaries are not well-localised everywhere along the segment borders. This is a direct artifact of the feature-space driving the segmentation, namely the Scale Saliency features. If a texture has sparse textons, as is the case with the wood grain texture in Figure 12, then the resulting saliency space is also sparse. In such cases, large windows are often necessary in order to generalise the texture class. In fact, in this experiment a window size of 35 pixels was used.

One solution to this problem is to adopt a coarse-to-fine strategy. That is, start the segmentation with a large window, then progress to smaller windows once convergence has been achieved. However, in some cases, other problems can occur at the smaller window sizes. For example, the sparse nature of the saliency space can cause uneven forces to act on the contour, resulting in rough region boundaries, even with a coarse-to-fine approach. Figure 12c shows such an effect, where the local scale has been reduced to 15 after convergence at local scale= 35. One solution to this problem could be to include a boundary smoothness term (i.e. boundary regularisation) in the cost function. Alternatively, additional, non-sparse, features could be used to drive the segmentation, for example, local intensity information. The latter could be used in the form a special boundary detection term, such as suggested by (Paragios and Deriche, 2000).

The 1D Scale Descriptor is useful for images which comprise textures with ‘strong’ and distinct texton behaviour. By ‘strong’, we mean that the textures can be considered as texton-based, or structural, as opposed to statistical or wave-like. Weakly-structural textures, that is those with poorly defined features (and scales), tend to give rise to scale histograms in which many scales are likely. The resulting discrimination performance is poor. However, the joint saliency and scale distribution, as represented in the 2D Scale Descriptors, can distinguish between such textures. This is because weakly-structural textures typically exhibit low saliency features, whereas those that are strongly-structural result in high saliency features.

This point is demonstrated in the Figure 13, where a three-texture mosaic has been segmented using the 1D and 2D descriptors. In this example, the segment at the top (water) is the weakly-structural texture. In the 1D descriptor result, Figure 13a, the algorithm has failed to correctly partition the image into the desired regions, whereas for the 2D descriptor, Figure 13b, the correct result has been achieved.
Figure 13c illustrates another interesting phenomenon. Here, the same three-texture mosaic has been segmented with the same parameters as those used for (b), but with local scale= 35. As was the case in Figure 12, the large window has caused poor region boundary localisation. However, in this case, it appears that the error is systematic because the boundary has ‘leaked’ evenly into the weakly-structural texture region at the top. The reason for this, once again, is the sparse nature of the saliency space. The problem is that the Region Competition cost function assumes a uniformly sampled feature-space, and in our implementation, this is not true. The saliency space is, in general, sparse and non-uniformly populated. The linen texture (bottom right) has many more features across the Scale Saliency space than the water texture (top), therefore its features dominate in the Scale Descriptors formed at the region boundaries. The solution is to correct for this in the Region Competition cost functions.

Finally, in Figure 14, we show the results of the algorithm applied to a five-texture Brodatz mosaic similar to those used for the classification experiments in Section 9.1. The images here were of size 320 $\times$ 320 pixels. Figure 14a shows the result of the segmentation using a local scale of 35 pixels. The overall segmentation is good, but once again, the boundary localisation is poor. This time, the effect has been to extend the region of one texture, the grass in left segment, to areas that do not correspond to that texture. This could be due to the strong/weak ‘leak’ problem discussed above, however all of these textures seem equally strongly-structural. A more likely explanation is that a sort of feature ‘mixing’ occurs at the boundaries. For example, in this case the scales and saliencies from the textures in the top, right and bottom segments ‘mix’ with those from the centre segment at their respective boundaries. The resultant Scale Descriptors at the region boundaries are closer to that of the left segment, than those for any of the others. This is not an unreasonable supposition, given that the grass texture contains many scales at many saliency values. The problem can be alleviated by reducing the local scale once convergence has been achieved. The result of reducing the local scale to 25, 15, then 5 is shown in Figure 14b. The region overhead cost, $\lambda$, (a larger value encourages fewer regions) was also reduced inline with the local scale to ensure the regions did not merge. An alternative solution is to use an additional set of features to drive the segmentation, such as local intensity.

9.2.2 Real Images

In this section, the segmentation algorithm has been applied to two ‘real’ images; the first, is of a zebra, and the second, of a cheetah. The characteristic hide-patterns of such animals makes them a good choice for the scale-based texture descriptor. The results are shown in Figures 15 and 16.

For the Zebra image, results are presented for increasing region cost: $\lambda = 500, 1000, 2000$. Overall,
the segmentation is good, with the ‘best’ partition achieved with $\lambda = 2000$. However, one problem is that the region boundary localisation is not very good. It is worse for the lower $\lambda$ results. This is another instance a strongly-structural texture dominating a neighbouring weakly-structural one, and requires the correction of the Region Competition cost terms. It is interesting to note that the problem is worse at the back or saddle point$^6$ of the zebra, than at the bottom (i.e. near its stomach). This is because the neighbouring texture at the bottom is more strongly-structural than the one at the top.

The final example is shown in Figure 16, where an image of a cheetah has been segmented. All parameters were identical to the previous example and, once again, results for increasing $\lambda$ have been shown. Here, the boundary localisation is much better. The most likely reason for this is that the background texture is more strongly-structural. It is interesting to note that in the first set, for which $\lambda = 500$, two regions are found in the body of the cheetah which seem to approximately correspond to two scales of spots.

The high region costs used in these two examples are necessary in order to encourage the merging of the desired regions. This is because they do not correspond well to the piecewise constant model assumed in the segmentation algorithm. In other words, they are not homogeneous regions. Fortunately, in these examples, the textures in the component regions are distinctive enough to allow this. For other images, this may not be the case and the algorithm would require a careful choice of parameters in order to converge to the desired result. A piecewise smooth model would alleviate this problem.

10 Conclusion

In this paper, we have applied our Scale Saliency algorithm, introduced in (Kadir and Brady, 2001), to the task of texture description. We have developed novel descriptors of texture, called Scale Descriptors, which are based solely on the PDF of salient scales. These descriptors essentially measure the scales of salient textons. Although not every texture can be discriminated by texton scale alone, the results presented in this paper demonstrate that the method can work for a surprising number of naturally occurring textures. Furthermore, the Scale Descriptors are essentially invariant to translation, rotation, and uniform shifts and scalings in intensity.

We have presented texture classification results that demonstrate that the method performs comparably to other published methods and in some cases significantly outperforms them. We have also applied the technique to unsupervised segmentation using the Level Sets based algorithm described in (Kadir and Brady, 2002).
References


VisTex. Vision texture webpage, 1998. URL


Figure 1: High saliency regions, such as the eye, exhibit unpredictable local intensity hence high entropy. Image from NIST Special Database 18, Mugshot Identification Database.
Figure 2: Salient scale region selection and salient icons (10% most salient shown) are robust to rotation (45° clockwise) and scaling (60% of original size).
Figure 3: A frame from the Vicky1 sequence used in the recognition experiments in Kadir and Brady (2001). At this particular focal length the textons of the various textures in the image are not particularly salient.
Figure 4: The Scale Descriptors of textures D101 and D102 from the Brodatz test set. Two variants are shown for each texture — 1D and 2D. The upper plots (1D histograms) show the probability (vertical axis) of a feature occurring in the texture patch with a certain scale (horizontal axis). In the lower plots (2D histograms), the brightness represents the joint probability of a feature occurring in the texture patch with a certain saliency value (vertical axis) and scale (horizontal axis). There are two main texton scales in the images which can be clearly seen in the Scale Descriptors.
Figure 5: Histograms of salient scales for a pair of samples from each of six textures. Numbers at the bottom show the best matches. Images from (Bigun and du Buf, 1994)
Figure 6: Joint saliency and scale histograms for Brodatz textures (Brodatz, 1966). These 2D histograms can distinguish between textures that have similar ranges of scale present, but at different saliencies. Images from USC-SIPI image database.
Figure 7: Robustness of the Scale Descriptors. Only the 1D versions are shown for clarity. Two implementations of the Scale Saliency algorithm were used, the first with a histogram (16 bins) PDF estimator, and the second with a Parzen window (Gaussian kernel with $\sigma = 25$). Four cases are shown: (1) Original Texture. (2) Intensity scaling - all intensities multiplied by 0.5. (3) Rotation $-45^\circ$. (4) Scaling (zoom out) - texture subsampled by 2.
Figure 8: The two-texture Randen mosaics.
Figure 9: The five-texture Randen mosaics.

Figure 10: Classification maps for mosaics in Figures 8a, 8b and 8c. Correct classification colours are: Left — Red. Right — Green.
Figure 11: Classification maps for mosaics in Figures 9a, 9b, 9c, 9d and 9e. Correct classification colours are: Left — Red, Top — Green, Right — White, Bottom — Yellow, Centre — Blue.
Figure 12: Unsupervised segmentation of a simple two-texture Brodatz mosaic using 1D parametric model (Hist16). Image (a) is the initialisation. Image (b) Local scale=35. Image (c) Local scale=35,15 (coarse-to-fine). $\lambda = 300$ for both cases.

Figure 13: Comparing 1D and 2D Scale Descriptor region models for segmentation of a three-texture Brodatz mosaic. Image (a) 1D : Hist16, Local scale=15, $\lambda = 100$. Image (b) 2D : (Hist8 x 8), Local scale=15. $\lambda = 1000$. Image (c) 2D : (Hist8 x 8), Local scale=35. $\lambda = 1000$. 

46
Figure 14: Segmentation of five-texture Brodatz mosaic. Image (a) shows the result of coarse-scale stage (local scale= 35), $\lambda = 400$. Image (b) shows the result after the fine-scale stage (local scale=5) $\lambda = 100$. 2D (Hist8) Scale Descriptors were used.
Figure 15: Unsupervised segmentation of a Zebra with varying $\lambda$. Region model — Hist2DW (8 bins). Local Scale=15.
Figure 16: Unsupervised segmentation of a Cheetah with varying $\lambda$. Region model — Hist2DW (8 bins). Local Scale=15.
## List of Figures

1. High saliency regions, such as the eye, exhibit unpredictable local intensity hence high entropy. Image from NIST Special Database 18, Mugshot Identification Database. 36
2. Salient scale region selection and salient icons (10% most salient shown) are robust to rotation (45° clockwise) and scaling (60% of original size). 37
3. A frame from the Vicky1 sequence used in the recognition experiments in Kadir and Brady (2001). At this particular focal length the textons of the various textures in the image are not particularly salient. 38
4. The Scale Descriptors of textures D101 and D102 from the Brodatz test set. Two variants are shown for each texture — 1D and 2D. The upper plots (1D histograms) show the probability (vertical axis) of a feature occurring in the texture patch with a certain scale (horizontal axis). In the lower plots (2D histograms), the brightness represents the joint probability of a feature occurring in the texture patch with a certain saliency value (vertical axis) and scale (horizontal axis). There are two main texton scales in the images which can be clearly seen in the Scale Descriptors. 39
5. Histograms of salient scales for a pair of samples from each of six textures. Numbers at the bottom show the best matches. Images from (Bigun and du Buf, 1994) 40
6. Joint saliency and scale histograms for Brodatz textures (Brodatz, 1966). These 2D histograms can distinguish between textures that have similar ranges of scale present, but at different saliencies. Images from USC-SIPI image database. 41
7. Robustness of the Scale Descriptors. Only the 1D versions are shown for clarity. Two implementations of the Scale Saliency algorithm were used, the first with a histogram (16 bins) PDF estimator, and the second with a Parzen window (Gaussian kernel with $\sigma = 25$). Four cases are shown: (1) Original Texture. (2) Intensity scaling - all intensities multiplied by 0.5. (3) Rotation - 45°. (4) Scaling (zoom out) - texture subsampled by 2. 42
8. The two-texture Randen mosaics. 43
9. The five-texture Randen mosaics. 44
10. Classification maps for mosaics in Figures 8a, 8b and 8c. Correct classification colours are: Left — Red. Right — Green. 44
11. Classification maps for mosaics in Figures 9a, 9b, 9c, 9d and 9e. Correct classification colours are: Left — Red, Top — Green, Right — White, Bottom — Yellow, Centre — Blue. 45
Unsupervised segmentation of a simple two-texture Brodatz mosaic using 1D parametric model (Hist16). Image (a) is the initialisation. Image (b) Local scale=35. Image (c) Local scale=35,15 (coarse-to-fine). $\lambda = 300$ for both cases.

Comparing 1D and 2D Scale Descriptor region models for segmentation of a three-texture Brodatz mosaic. Image (a) 1D : Hist16, Local scale=15, $\lambda = 100$. Image (b) 2D : (Hist8 $\times$ 8), Local scale=15. $\lambda = 1000$. Image (c) 2D : (Hist8 $\times$ 8), Local scale=35. $\lambda = 1000$.

Segmentation of five-texture Brodatz mosaic. Image (a) shows the result of coarse-scale stage (local scale= 35), $\lambda = 400$. Image (b) shows the result after the fine-scale stage (local scale=5) $\lambda = 100$. 2D (Hist8) Scale Descriptors were used.

Unsupervised segmentation of a Zebra with varying $\lambda$. Region model — Hist2DW (8 bins). Local Scale=15.

Unsupervised segmentation of a Cheetah with varying $\lambda$. Region model — Hist2DW (8 bins). Local Scale=15.
Footnotes

0 Affiliation of Authors - Robotics Research Laboratory, Department of Engineering Science, University of Oxford, Parks Road, Oxford. OX1 3PJ, U.K.

1 Timor Kadir is the recipient of a Motorola University Partners in Research grant.

2 This is true for the real wavelet transform. For the complex wavelet transform the mother wavelet can be scaled and also its phase shifted in the Fourier sense, thus resulting in a richer basis set.

3 Note that in the Kovesi formulation, Phase Congruency is defined with Fourier phase and defines salient features as discontinuities. We could redefine phase with reference to a particular wavelet basis.

4 Note that here, we are referring to the PDF estimators used in the Scale Saliency algorithm, not the Scale Descriptors themselves, which are represented as histograms.

5 Available from http://www.ux.his.no/~tranden/data.html

6 in both senses of the word.
Address for correspondence

Dr. Timor Kadir
Department of Engineering Science
University of Oxford
Ewert House, Ewert Place,
Summertown, Oxford.
OX2 7BZ.