On the Significance of Real-World Conditions for Material Classification

Eric Hayman, Barbara Caputo, Mario Fritz, and Jan-Olof Eklundh

Computational Vision and Active Perception Laboratory
Dept. of Numerical Analysis and Computer Science
Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden
{hayman,caputo,mjfritz,joe}@nada.kth.se

Abstract. Classifying materials from their appearance is a challenging problem, especially if illumination and pose conditions are permitted to change: highlights and shadows caused by 3D structure can radically alter a sample's visual texture. Despite these difficulties, researchers have demonstrated impressive results on the CUReT database which contains many images of 61 materials under different conditions. A first contribution of this paper is to further advance the state-of-the-art by applying Support Vector Machines to this problem. To our knowledge, we record the best results to date on the CUReT database.

In our work we additionally investigate the effect of *scale* since robustness to viewing distance and zoom settings is crucial in many real-world situations. Indeed, a material's appearance can vary considerably as fine-level detail becomes visible or disappears as the camera moves towards or away from the subject. We handle scale-variations using a pure-learning approach, incorporating samples imaged at different distances into the training set. An empirical investigation is conducted to show how the classification accuracy decreases as less scale information is made available during training.

Since the CUReT database contains little scale variation, we introduce a new database which images ten CUReT materials at different distances, while also maintaining some change in pose and illumination. The first aim of the database is thus to provide scale variations, but a second and equally important objective is to attempt to recognise *different samples* of the CUReT materials. For instance, does training on the CUReT database enable recognition of *another* piece of sandpaper? The results clearly demonstrate that it is *not* possible to do so with any acceptable degree of accuracy. Thus we conclude that impressive results even on a well-designed database such as CUReT, does not imply that material classification is close to being a solved problem under real-world conditions.

1 Introduction

The recognition of materials from their visual texture has many applications, for instance it facilitates image retrieval and object recognition. As a step towards the use of such techniques in the real world, recent developments have concentrated on being able to recognise materials from a variety of poses and with different illumination conditions [16,9,31]. This is a particularly challenging task when the material has considerable 3-dimensional structure. With such 3D textures, cast shadows and highlights can cause the

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3024, pp. 253–266, 2004.

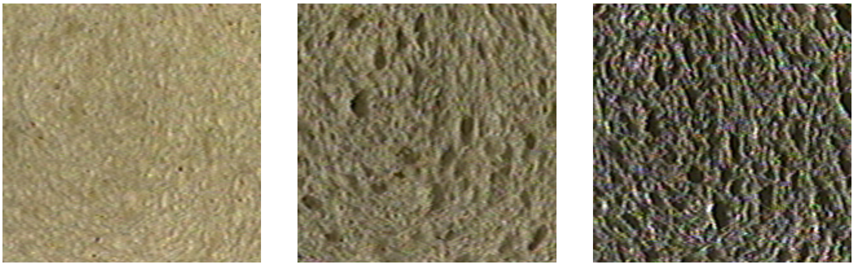


Fig. 1. Three images of white bread taken from the CUReT database demonstrating the variation of appearance of a 3D texture as the pose and illumination conditions change.

appearance to change radically with different viewing angles and illumination conditions. An example from the CUReT database [10] (white bread) is given in Fig. 1.

The overall goal of our work is to bring material recognition algorithms closer still to the stage where they will be useful in real-world applications. Thus a major objective is providing robustness to variations in *scale*. Experiments will show that failure in this regard rapidly leads to a deterioration in classification accuracy. Our solution is a pure-learning approach which accommodates variations in scale in the training samples, similar to how differing illumination and pose are modelled.

A further contribution concerns demonstrating the suitability of Support Vector Machines (SVMs) [8,29] as classifiers in this recognition problem. Experiments show that the SVM classifier systematically outperforms the nearest-neighbour classification scheme adopted by Varma and Zisserman [31] with which we compare our results, and we also demonstrate that we achieve an improvement on their Markov Random Field (MRF) approach [32] which, to our knowledge, previously yielded the best overall classification rate on the CUReT database.

As already alluded to, experiments are conducted on the CUReT image database [10] which captures variations in illumination and pose for 61 different materials, many of which contain significant 3D structure. This database does not, however, contain many scaling effects. Some indication of the performance under varying scale can be achieved by artificially scaling the images by modifying the scales of the filters in the filter bank. However, we also investigate classification results on pictures of materials present in the CUReT database, imaged in our laboratory. The objectives of these experiments are two-fold. First, it permits a systematic study of scale effects while still providing some variations in pose and illumination. Second, we investigate whether it is possible to recognise materials in this new database given models trained on the CUReT database. This indeed proves a stern test, since both the sample of material, the camera and lighting conditions are different to those used during training.

Thus the final contribution of this paper is the construction of a new database, designed to complement the CUReT database with scale variations. This database, called KTH-TIPS (Textures under varying Illumination Pose and Scale) is freely available to other researchers via the web [12].

The remainder of the paper is organised as follows. Section 2 reviews previous literature in the field. Particular emphasis is placed on the algorithm of Varma and

Zisserman [31] on which we ourselves to a large extent build. Section 3 discusses the application of Support Vector Machines to this problem, and also presents experiments which demonstrate their superior performance relative to the original approach of [31]. Further experiments in the paper also make use of SVMs. Then, Section 4 discusses issues with scale, presents a pure learning approach for tackling the problem, and conducts experiments on the CUReT database. Section 5 introduces the new database designed to supplement the CUReT database for experiments with scale. Conclusions are drawn and potential avenues for future research outlined in Section 6.

2 Previous Work

Most work on texture recognition [21,23,14] has dealt with planar image patches sampled, for instance, from the Brodatz collection [4]. The training and test sets typically consist of non-overlapping patches taken from the same images. More recently, however, researchers have started to combat the problems associated with recognising materials in spite of varying pose and illumination. Leung and Malik [16] modelled 3D materials in terms of *texton histograms*. The notion of textons is familiar from the work of Julesz [13], but it was only recently defined for greyscale images as a cluster centre in a feature space formed by the output of a filter bank. Given a vocabulary of textons, the filter output of each pixel is assigned to its nearest texton, and a histogram of textons is formed over an extended image patch. This procedure was described for 2D textures in [20] and for 3D textures in [16] by stacking geometrically registered images from the training set. Recognition is achieved by gathering multiple images of the material from the same viewpoints and illuminations, performing the geometric registration, computing the texton histogram and classifying it using a nearest-neighbour scheme based on the χ^2 distance between model and query histograms.

Cula and Dana [9] adapted the method of Leung and Malik to form a faster, simpler and more accurate classifier. They realised that the 3D registration was not necessary, and instead described a material by multiple histograms of 2D textons, where each histogram is obtained from a single image in the training set. This also implies that recognition is possible from a single query image.

Varma and Zisserman [31] argued strongly for a rotationally invariant filter bank. First, two images of the same material differing only by an image-plane rotation should be equivalent. Second, removing the orientation information in the filter bank considerably reduced the size of the feature vector. Third, it led to a more compact texton vocabulary since it was no longer necessary for one texton to be a rotated version of another. Rotational invariance was achieved by storing only the maximum response over orientation of a given type of filter at a given scale. As Fig. 2 indicates, the filter bank contains 38 filters, but only 8 responses are stored, yielding the so-called MR8 (Maximum Response 8) descriptor. Not only did the use of this descriptor reduce storage requirements and computation times, an improvement in recognition rate was also achieved. In their experiments [31] they use 92 of the 205 images in the CUReT database, removing samples at severely slanted poses. Splitting these 92 images of each material equally into 46 images for training and 46 images for the test set, they obtain an impressive classification accuracy of up to 97.43% [32]. This is the system that we will be using as a reference in our own experiments.

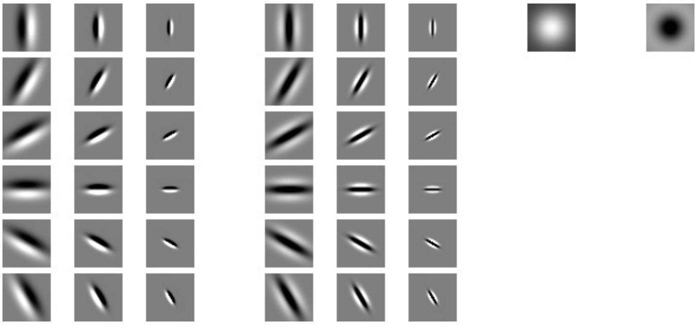


Fig. 2. Following [31] we use a filter bank consisting of edge and bar filters (first and second Gaussian derivatives) at 3 scales and 6 orientations, and also a Gaussian and Laplacian. Only the maximum response is stored for each orientation, yielding the 8-dimensional MR8 descriptor.

Many different descriptors have been proposed for texture discrimination. Filter banks are indeed very popular [21,16,9,31,24], and there is evidence that biological systems process visual stimuli using filters resembling those in Fig. 2. However, non-filter descriptors have recently been regaining popularity [11,32,19,15]. [32] presents state-of-the-art results on the CUReT database using a Markov Random Field (MRF) model. Määenpää and Pietikäinen [19] extend the Local Binary Pattern approach [23] to multiple image resolutions and obtain near-perfect results on a test set from the Outex database. However, this database does not contain any variations in pose or illumination, and the variation in scale is rather small (100dpi images in the training set and 120dpi images in the test set). Recent, impressive work by Lazebnik *et al.* [15] considers simultaneous segmentation and classification of textures under varying scale. Interest points are detected, normalised for scale [18], skew and orientation, and intensity domain spin images computed as descriptors. Each interest point is assigned to a texture class before a relaxation scheme is used to smooth the response. It remains to be seen, however, whether this scheme can handle large variations in illumination, and the number of classes in their experiments is rather small. Scale-invariant recognition using Gabor filters on Brodatz textures was considered by Manthalkar *et al.* [22].

3 Using Support Vector Machines for Texture Classification

The first contribution of this paper is to demonstrate that recent advances in machine learning prove fruitful in material classification. Support Vector Machines are state--of-the-art large margin classifiers which have gained popularity within visual pattern recognition, particularly for object recognition. Pontil and Verri [26] demonstrated the robustness of SVMs to noise, bias in the registration and moderate amounts of occlusion while Roobaert *et al.* [27] examined their generalisation capabilities when trained on only a few views per object. Barla *et al.* [2] proposed a new class of kernel inspired by similarity measures successful in vision applications. Other notable work includes [17,5,1]. Although SVMs have previously been used on planar textures [14], they have not, to our knowledge, been applied to 3D material classification under varying imaging conditions.

Before demonstrating in experiments the improvements that can be achieved with SVMs, we provide a brief review of the theory behind this type of algorithm. For a more detailed treatment, we refer to [8,29].

3.1 Support Vector Machines: A Review

Consider the problem of separating a set of training data $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) ... (\boldsymbol{x}_m, y_m)$, where $\boldsymbol{x}_i \in \Re^N$ is a feature vector and $y_i \in \{-1, +1\}$ its class label. If we assume that the two classes can be separated by a hyperplane $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, and that we have no prior knowledge about the data distribution, then the optimal hyperplane (the one with the lowest bound on the expected generalisation error) is that which has maximum distance to the closest points in the training set. The optimal values for $\boldsymbol{w}$ and b can be found by solving the following constrained minimisation problem:

$$\underset{\boldsymbol{w},b}{\text{minimise}} \ \frac{1}{2}\|\boldsymbol{w}\|^2 \qquad \text{subject to} \qquad y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, \forall i = 1, \dots m \tag{1}$$

Introducing Lagrange multipliers $\alpha_i (i = 1, \dots m)$ results in a classification function

$$f(x) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{w} \cdot \boldsymbol{x} + b\right). \tag{2}$$

where α_i and b are found by Sequential Minimal Optimisation (SMO, [8,29]). Most of the α_i's take the value of zero; those $\boldsymbol{x}_i$ with nonzero α_i are the "support vectors". In cases where the two classes are non-separable, Lagrange multipliers are introduced, $0 \leq \alpha_i \leq C, i = 1, \dots m$, where C determines the trade-off between margin maximisation and training error minimisation. To obtain a nonlinear classifier, one maps the data from the input space $\Re^N$ to a high dimensional feature space $\mathcal{H}$ by $\boldsymbol{x} \to \Phi(\boldsymbol{x}) \in \mathcal{H}$, such that the mapped data points of the two classes are linearly separable in the feature space. Assuming there exists a kernel function K such that $K(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{y})$, a nonlinear SVM can be constructed by replacing the inner product $\boldsymbol{w} \cdot \boldsymbol{x}$ by the kernel function $K(\boldsymbol{x}, \boldsymbol{y})$ in eqn. (2). This corresponds to constructing an optimal separating hyperplane in the feature space. Kernels commonly used include polynomials $K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} \cdot \boldsymbol{y})^d$, and the Gaussian Radial Basis Function (RBF) kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \exp\{-\gamma||\boldsymbol{x} - \boldsymbol{y}||^2\}$. The Gaussian RBF has been found to perform better for histogram-like features [7,5], thus unless specified otherwise, this is the kernel we will use in the present paper.

The extension of SVM from 2-class to M-class problems can be achieved following two basic strategies: In a *one-vs-others* approach, M SVMs are trained, each separating a single class from all remaining classes. Although the most popular scheme for extending to multi-class problems (see for instance [8,5,7]), there is no bound on its generalisation error, and the training time of the standard method scales linearly with M [8]. In the second strategy, the *pairwise approach*, $M(M-1)/2$ two-class machines are trained. The pairwise classifiers are arranged in trees, where each tree node represents an SVM. Decisions can be made using a bottom-up tree similar to the elimination tree used in tennis tournaments [8], or a Directed Acyclic Graph (DAG, [25]).

3.2 Results

Platt and others [25] presented an analysis of the generalisation error for DAG, indicating that building large margin DAGs in a high dimensional feature space can yield good generalisation performance. On the basis of this result and of several empirical studies, we used a pairwise approach with DAG in this paper, using the *LibSVM* library [6]. C was fixed at 100 whereas γ in the RBF was obtained automatically by cross-validation. The histograms were treated as feature vectors and normalised to unit length.

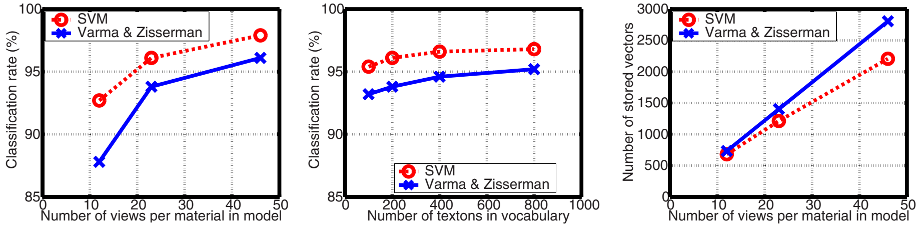
We compared the SVM classifier with our own implementation of the algorithm of Varma and Zisserman [31], which from now on will be denoted the VZ algorithm, and we use the same 200×200 pixels greyscale image patches as they do. The patches are selected such that only foreground is present.

A first experiment ascertains the maximum performance that can be achieved on the CUReT database by using a very large texton vocabulary. 40 textons were found from each of the 61 materials, giving a total dictionary of $40 \times 61 = 2440$ textons. The 92 images per sample were split equally into training and test sets. Varma and Zisserman [32] previously reported a 97.43% success rate, while our own implementation of their algorithm gave an average of 97.66% with a standard deviation of 0.11% over 10 runs[1]. In contrast, the SVM classifier gave $98.36 \pm 0.10\%$ using an RBF kernel and $98.46 \pm 0.09\%$ using the χ^2-kernel $K = \exp\{-\gamma\chi^2\}$. We implemented this Mercer kernel [3] within *LibSVM*. This performs better even than the very best result obtained in [32] using an MRF model (98.03%) which, to our knowledge, previously represented the best overall classification rate on the CUReT database.

Another natural extension to the Varma and Zisserman algorithm is to replace the Nearest Neighbour classifier with a k-Nearest Neighbour scheme. Several variants of k-NN were tried with different strategies to resolve conflicts [28]. Of these, Method 2 from [28] proved best in our scenario, but no variant yielded an improved recognition rate for any choice of $k > 1$. This is probably due to a relatively sparse sampling of the pose and illumination conditions in the training set.

Further experiments examine the dependency on the size of the training set (Fig. 3a) and the texton vocabulary (Fig. 3b). Both plots clearly demonstrate that the SVM classifier reduces the error rate by 30 – 50% in comparison with the method of [31]. In both experiments, textons were found from the 20 materials specified in [16] rather than all 61 materials. In Fig. 3a, 10 textons per material are used, giving a dictionary of $20 \times 10 = 200$ textons. In Fig. 3b, the training set consists of 23 images per material, and the remaining 69 images per material are placed in the test set.

[1] The variability within experiments is due to slightly different texton vocabularies; images are selected at random when generating the dictionary with K-means clustering. The difference of 0.23% between our results and the figure of 97.43% reported in [32] is caused by our use of more truncated filter kernels (41×41 compared to 49×49 [30]) although the scales used to compute the kernels were identical. For a texton to be assigned to a pixel, the entire support region of the filter kernel is required to lie inside the 200×200 image patch. Thus the texton histograms contain more entries when a smaller filter kernel is used. It may be noted that the MRF algorithm of [32] computes descriptors from significantly smaller regions, for instance 7×7.



(a) (b) (c)

Fig. 3. Experiments comparing our SVM scheme with the VZ [31] approach. (a) plots the reliance on the number of views in the training set, (b) the dependency on the size of the texton vocabulary, and (c) the size of the stored model. In (c) the model reduction schemes of [31,32] were not implemented.

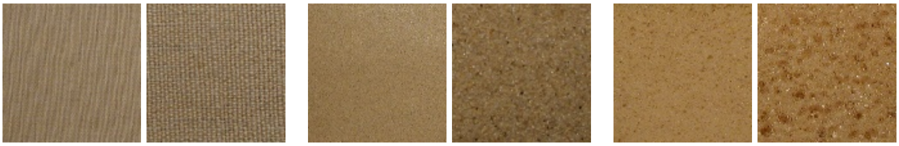
Training times for SVM vary from about 20 seconds (with a vocabulary of 100 textons, 12 views per material in the training set) up to roughly 50 minutes (for 2440 textons, 46 views per material). Finding γ by cross-validation, if required, typically incurs a further cost of 3–7 times the figures reported above.

The size of the resulting model is illustrated in Fig. 3c. Recalling that only the support vectors need be stored, and noting that storing the coefficients α_i incurs little overhead, SVM reduces the size of the model by $10 - 20\%$. This is significantly less than the reduction by almost 80% obtained using the greedy algorithms described in [31] and [32]. However, the scheme in [31] used the test set for validating the model, which is unreasonable in a recognition task, while the method in [32] was extremely expensive in training, in fact by a few orders of magnitude [30] in comparison with the more expensive times listed for SVM above. Moreover, their procedure for selecting a validation set from the training set is largely heuristic and at a high risk of over-fitting, in which case the performance on the test set would drop very significantly [30].

4 Material Classification under Variations in Scale

The results presented so far on the CUReT database were obtained without significant scale variation in the images [2]. In the real world, scale undoubtedly plays an important role, and it seems unlikely that the classifiers described so far will perform well. First, the individual filters are tuned to certain frequencies, and zooming in or out on a texture changes the characteristic frequencies of its visual appearance. Second, zooming in on a texture can make visible fine-level details which could not be recorded at coarser scales due to the finite resolution of the imaging device. Examples are given in Fig. 4. With cotton, for instance, at a coarse scale a vertical line structure is just about visible, whereas at a fine scale the woven grid can be seen clearly, including horizontal fibres.

[2] Four samples are zoomed in images of other materials. In the experiments reported in this paper, classifying one material as the zoomed in version of that same material is labelled an incorrect match. In practise such confusions are fairly common for those four materials, but this does not have a very large effect on classification rates when averaged over all materials.

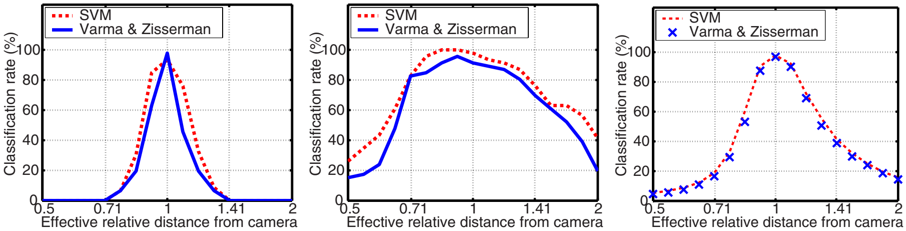


(i) Distant (ii) Close
(a) Cotton

(i) Distant (ii) Close
(b) Sandpaper

(i) Distant (ii) Close
(c) Sponge

Fig. 4. The appearance of materials can change dramatically with distance to the camera.



(a) Sandpaper (b) Sponge (c) Average

Fig. 5. Variations in scale can have a disastrous effect. In this experiment the training set contains images only at the default scale whereas the test set contains images rescaled by amounts up to a factor of two both up and down. For sandpaper (a) the recognition rate drops dramatically, whereas for sponge (b) they are more stable, probably since the salient features are repeated over a wide range of scales. Results averaged over the entire CUReT database are shown in (c).

4.1 A Motivational Experiment

Experimental confirmation of the scale-dependence of the texton-histogram based schemes was obtained by supplementing the CUReT database with artificially scaled versions of its images. Rather than rescaling the images, which raises various issues with respect to smoothing and aliasing, the *filters* were rescaled. For instance, reducing the size of the image (zooming out) by a factor of two is equivalent to doubling the standard deviations in the filters. This procedure was repeated at eight logarithmically spaced intervals per octave, scaling both up and down one octave. This resulted in $2 \times 8 = 16$ scaled images in addition to the unscaled original, giving a total of 17 images. Only the unscaled images were placed in the training set, whereas recognition was attempted at all 17 scales [3]. The 92 images per sample were split evenly into training and test sets, and a texton vocabulary of 400 textons was used.

Fig. 5 illustrates this dependency on scale for two materials. Sandpaper (Fig. 5a), shows almost no robustness to changes in scale, whereas sponge (Fig. 5b) is much more resilient. These effects can be attributed to two main factors. The first concerns *intra-class* properties: materials with a highly regular pattern have a clear characteristic scale, whereas others, such as sponge, exhibit similar features over a range of scales. The

[3] We acknowledge that this method is no true replacement for real images since (i) it is not possible to increase the resolution while artificially zooming in, and (ii) the information content is reduced somewhat when artificially zooming out since the size of the 200×200 pixels patch is effectively reduced.

Table 1. The recognition rate (in %) on the artificially rescaled CUReT database as the richness of the model is varied both with respect to the sampling density in the scale direction and in how many of the original 92 images are incorporated in the training set (per scale). With 3 scales present, the training set includes the original image and also samples at scales one octave up and one octave down. With five scales, half-octave positions are made available during training, and with 9 scales, quarter-octave positions are also used.

		No. of original images per sample		
		46	**23**	**12**
No. of scales	**9**	97.58	94.59	91.60
	5	95.89	92.67	89.89
	3	81.75	78.04	76.80
	1	36.85	36.12	34.08

(a) SVM

		No. of original images per sample		
		46	**23**	**12**
No. of scales	**9**	92.14	89.23	83.65
	5	81.19	77.91	71.95
	3	58.00	55.69	51.57
	1	34.47	33.16	30.90

(b) Varma and Zisserman [31]

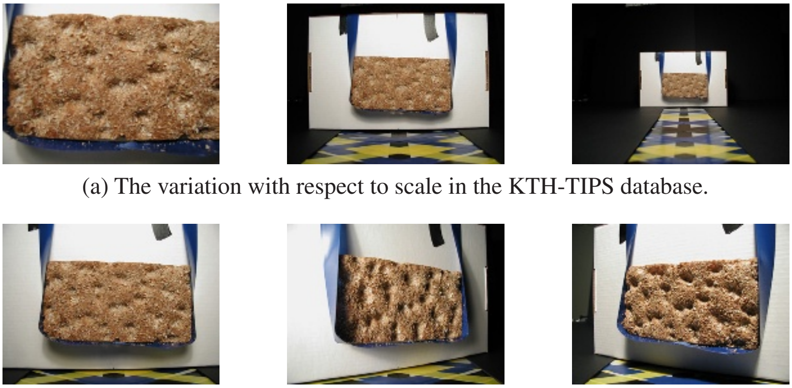
feature vector for the former material could be severely mutated, whereas we expect the descriptor of the latter to be more robust to changes in scale. The second factor depends on the *inter-class* variation in the database: the recognition rate depends on the degree of distraction caused by other materials. It is feasible that a material imaged at a certain scale closely resembles another material at the default scale. Fig. 5c shows corresponding plots for an average over all 61 materials in the CUReT database.

4.2 Robustness to Scale Variations: A Pure Learning Approach

The experiment described above indicated that providing robustness to changes in image scale can be crucial if material recognition is to function in the real world. A natural strategy for providing such robustness is to extend the training set to cover not just variations in *pose* and *illumination* conditions, but also *scale*. An alternative, left unexplored here, would be to include only images at one scale during training, but then artificially rescale the query image to a number of candidate scales by rescaling the filter bank.

An open question is how densely it is necessary to sample in the scale direction, particularly since the size of the training set has obvious implications for algorithm speed and memory requirements. Clearly there will be some dependence on the bandwidth of the filters, but the amount of inter-class variation will also be of consequence.

This dependence on sampling in the scale dimension was ascertained empirically on the rescaled CUReT database, and our findings are summarised in Tables 1a and b for the SVM and VZ classifiers respectively with a vocabulary of 400 textons. The most noteworthy aspect of these results is that impoverishing the model in the *scale* dimension appears to have a more severe effect than reducing the size of the training set with respect to the proportion of the original 92 images which were placed in the training set. Both SVM and the VZ schemes exhibit such behaviour. A further point worth emphasising is that SVM systematically outperforms the VZ classifier, as was also seen in Section 3. Again, we attempted replacing the Nearest Neighbour classifier in the Varma and Zisserman approach with k-Nearest Neighbour schemes, but without observing any improvement for $k > 1$.



(a) The variation with respect to scale in the KTH-TIPS database.

(b) The variation of pose and illumination present in the KTH-TIPS database.

Fig. 6. The variations contained in the new KTH-TIPS (Textures under varying Illumination Pose and Scale) database. In (a) the middle image, depicting the central scale, was selected to correspond roughly to the scale used in the CUReT database. The left and right images are captured with the sample at half and twice that distance, respectively. 3 further images per octave (not shown) are present in the database. (b) shows 3 out of 9 images per scale, showing the variation of pose and illumination. Prior to use, images were cropped so only foreground was present.

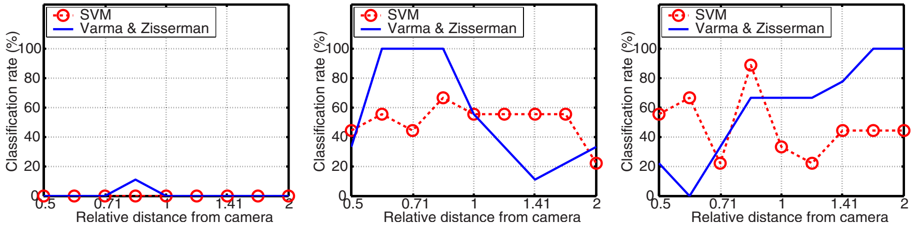
5 The KTH-TIPS Database of Materials under Varying Scale

Although the results presented above gave some indication as to the deterioration in performance under changes in scale, the artificial rescaling is no perfect replacement for real images. Therefore we created a new database to supplement CUReT by providing variations in *scale* in addition to pose and illumination. Thus we named it the KTH-TIPS (Textures under varying Illumination Pose and Scale) database. A second objective with the database was to evaluate whether models trained on the CUReT database could be used to recognise materials from pictures taken in other settings. This could indeed prove challenging since not only the camera, poses and illuminant differ, but also the actual samples: can *another* sponge be recognised using the CUReT sponge?

To date, our database contains ten materials also present in the CUReT database. These are sandpaper, crumpled aluminium foil, styrofoam, sponge, corduroy, linen, cotton, brown bread, orange peel and cracker B. These are imaged at nine distances from the camera to give equidistant log-scales over two octaves, as illustrated in Fig. 6a for the cracker. The central scale was selected, by visual inspection, to correspond roughly to the scale used in the CUReT database. At each distance images were captured using three different directions of illumination (front, side and top) and three different poses (central, 22.5° turned left, 22.5° turned right) giving a total of $3 \times 3 = 9$ images per scale, and $9 \times 9 = 81$ images per material. A subset of these is shown in Fig. 6b. For each image we selected a 200×200 pixels region to remove the background.

The database is freely available on the web [12].

We now present three sets of experiments on the KTH-TIPS database, differing in how the model was obtained. The first uses the CUReT database for training, the second a combination of both databases, and the third only KTH-TIPS.



(a) Sandpaper (b) Sponge (c) Corduroy

Fig. 7. Experiments attempting to recognise images from the new KTH-TIPS database using a model trained on all 61 materials of the CUReT database. The recognition rate is plotted against scale for three materials.

Using the CUReT database for training. We attempted to recognise the materials in KTH-TIPS using a model obtained by training on the 61 materials of the CUReT database. 46 out of 92 images per material were placed in the training set. To cope with variations in scale, the procedure described in Section 4.2 is used: the model is acquired by rescaling each training sample from the CUReT database by adapting the Gaussian derivative filters. For this experiment the training set contained data from 9 scales, equidistantly spaced along the log-scale dimension over two octaves.

Results for sandpaper, sponge and corduroy can be seen in Fig. 7a, b and c respectively. Performance on sandpaper is very poor. This failure could be due to differences between our sample of sandpaper and the CUReT sample of sandpaper, despite our efforts to provide similar samples. We did, however, note that sandpaper was a very difficult material to recognise also in experiments using the CUReT database as the test set. This indicates that many of the other materials can be confused with sandpaper.

Results were much improved for sponge and corduroy where recognition results of around 50% were achieved. It is interesting to note that the VZ classifier outperformed SVM in these experiments. The success rate of the VZ approach varies considerably with scale. It would seem that there is not perfect overlap between the two octaves in scale in the two datasets. Another explanation for a drop-off in performance at fine scales is that the rescaling of the CUReT database cannot improve the resolution: rescaling the filters does not permit sub-pixel structure to appear. A third reason is that the images closest to the camera were poorly focused in some cases. The SVM classifier provided much more consistent results over varying scales, as could perhaps be expected from the experiment reported in Table 1. However, the recognition rate was consistently fairly low over *all* scales. By supplying a test set too different to the samples provided during training, we are asking the SVM to perform a task for which it was not optimised; SVMs are designed for *discrimination* rather than *generalisation*.

The recognition rates for all 10 materials, averaged over all scales, is provided in Table 2a. Results are, on the whole, well below 50%, clearly demonstrating that material recognition cannot be performed reliably in the real world merely using the CUReT database to form the model. We have, however, confirmed that many of the confusions are reasonable. For instance, cotton was frequently confused with linen.

Table 2. Attempting to recognise samples from the KTH-TIPS database. Results are averaged over all scales.

Material	Recognition rate (%)	
	SVM	VZ
sandpaper	0.00	1.23
aluminium foil	11.35	12.35
styrofoam	34.72	38.27
sponge	50.62	54.32
corduroy	46.91	59.26
linen	30.41	25.93
cotton	11.11	20.99
brown bread	5.11	7.41
orange peel	11.11	11.11
cracker B	3.70	7.41
AVERAGE	**20.50**	**23.83**

(a) Training only on CUReT

Material	Recognition rate (%)	
	SVM	VZ
sandpaper	77.78	66.67
aluminium foil	91.67	88.89
styrofoam	100.00	91.67
sponge	100.00	100.00
corduroy	80.56	80.56
linen	61.11	41.67
cotton	61.11	47.22
brown bread	77.78	80.56
orange peel	100.00	63.89
cracker B	91.67	80.56
AVERAGE	**84.17**	**74.17**

(b) Training on *both* CUReT and KTH-TIPS

Using a combination of databases for training. In a second experiment we combined the CUReT and KTH-TIPS databases for training. Thus we no longer needed to worry about training and tests being performed on different samples, but now some classes in the model contained a wider variety, thus increasing the risk of classes overlapping in the feature space. We report experimental results for training with 5 equidistant scales in the log-scale dimension, spanning two octaves. For KTH-TIPS materials, at each scale 3 out of 9 images in the KTH-TIPS database were used for training, as were 43 images from the CUReT database. This same total number of 46 training images per scale was also used for the 51 materials only found in CUReT; these were included as distractors in the experiment. Results are summarised in Table 2b. As expected, including the KTH-TIPS samples in the training set yielded much better results; the average over all materials increased to 84.17% for SVM and 79.17% for VZ .

Training on KTH-TIPS. We also performed similar experiments using only the KTH-TIPS database for training, implying that the model contained only 10 classes rather than 61. Thus there are fewer distractions, and the overall recognition rate increased to 90.56% for SVM and 84.44% for VZ with 5 scales. Using only the central scale resulted in classification rates of 64.03% and 59.70% for SVM and VZ respectively. We will not report results from these experiments further.

6 Discussion and Conclusions

This paper attempted to bring material classification a step closer to real-world applications by extending work on 3D textures under varying *pose* and *illumination* to also accommodate changes in *scale*. We showed in experiments that it is crucial to model scale in some manner, and we demonstrated a scale-robust classifier which incorporates the variations in scale directly into the training set. Experiments were conducted both on an artificially rescaled version of the CUReT database, and on a new database designed to supplement the CUReT database by imaging a subset (currently 10 out of 61) of the materials at a range of distances, while still maintaining some variation in pose and illumination. This database represents the second contribution of this paper, and is available to other researchers via the web [12].

A third contribution was to demonstrate the superiority of Support Vector Machines (SVMs) in this application. We obtained a recognition rate of 98.46% on the CUReT database at constant scale which, to our knowledge, represents the highest rate to date.

However, a more sobering conclusion, and perhaps the most important message from this paper, is that such success on the CUReT database does *not* necessarily imply that it is possible to recognise those materials in the real world, even when scale is modelled. The main reason is probably that the samples imaged in our laboratory were not identical to those in CUReT. Naturally it is possible to include multiple samples of the same material in a database, but with increased intra-class variability, the risk of inter-class confusion increases. As this risk depends on the number of classes in the database, keeping this number low (e.g. in production line applications) should make it feasible to separate the classes, but with a large number it might only be possible to classify into broader *groups* of materials. The performance will again depend on scale since most materials appear more homogeneous with increased imaging distance.

In other work we are currently investigating mechanisms for scale selection as a pre-processing step [18]. Although it might still be necessary to store models at multiple characteristic scales, this number should still be smaller than with the pure-learning approach. This would reduce storage requirements, and also the recognition time.

A possible reason for sandpaper proving so hard to recognise in the experiments reported in Fig. 5a, is that the representation in terms of filters blurs the information too much with this kind of salt-and-pepper structure. Indeed, the role of filter banks has recently been questioned, and other representations have proved effective [11,32,19]. Thus we intend to explore such descriptors in our future work.

Acknowledgements. The authors are very grateful to Manik Varma and Andrew Zisserman for discussions regarding their algorithms, and for providing appropriately cropped images from the CUReT database.

This work was funded by the EU-IST projects IST-2000-29688 *Insight2+* (EH,MF) and IST-2000-29375 *Cogvis* (BC), and the *VISCOS* project funded by the Swedish Foundation for Strategic Research (EH).

References

1. S. Avidan. Support vector tracking. In *Proc. CVPR, Kauai, Hawaii*, pages I:184–191, 2001.
2. A. Barla, F. Odone, and A. Verri. Hausdorff kernel for 3D object acquisition and detection. In *Proc. ECCV, Copenhagen*, page IV: 20 ff., 2002.
3. S. Belongie, C. Fowlkes, F. Chung, and J. Malik. Spectral partitioning with indefinite kernels using the Nyström extension. In *Proc. ECCV, Copenhagen*, page III: 531 ff., 2002.
4. P. Brodatz. *Textures*. Dover, 1966.
5. B. Caputo and Gy Dorko. How to combine color and shape information for 3D object recognition: kernels do the trick. In *Proc. NIPS, Vancouver*, 2002.
6. Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

7. O. Chapelle, P. Haffner, and V. Vapnik. SVMs for histogram-based image classification. IEEE Trans. on Neural Networks, 10(5), 1999.
8. N. Cristianini and J. S. Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.
9. O.G. Cula and K.J. Dana. Compact representation of bidirectional texture functions. In *Proc. CVPR, Kauai, Hawaii*, pages I:1041–1047, 2001.
10. K.J. Dana, B. van Ginneken, S.K. Nayar, and J.J. Koenderink. Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics*, 18(1):1–34, January 1999.
11. A. Efros and T. Leung. Texture synthesis by non-parametric sampling. In *Proc. ICCV, Kerkyra, Greece*, pages 1033–1038, 1999.
12. M. Fritz, E. Hayman, B. Caputo, and J.-O. Eklundh. The KTH-TIPS database. Available at `http://www.nada.kth.se/cvap/databases/kth-tips`.
13. B. Julesz and R. Bergen. Textons, the elements of texture perception, and their interactions. *Nature*, 290:91–97, 1981.
14. K.I. Kim, K. Jung, S.H. Park, and H.J. Kim. Support vector machines for texture classification. *PAMI*, 24(11):1542–1550, November 2002.
15. S. Lazebnik, C. Schmid, and J. Ponce. Affine-invariant local descriptors and neighbourhood statistics for texture recognition. In *Proc. ICCV, Nice*, pages 649–655, 2003.
16. T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, June 2001.
17. S.Z. Li, Q.D. Fu, L. Gu, B. Scholkopf, Y. Cheng, and H.J. Zhang. Kernel machine based learning for multi-view face detection and pose estimation. In *Proc. ICCV, Vancouver*, pages II: 674–679, 2001.
18. T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):79–116, 1998.
19. T. Mäenpää and M. Pietikäinen. Multi-scale binary patterns for texture analysis. In *Proc. SCIA, Gothenberg, Sweden*, pages 885–892, 2003.
20. J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *Proc. ICCV, Kerkyra, Greece*, pages 918–925, 1999.
21. B.S. Manjunath and W.Y. Ma. Texture features for browsing and retrieval of image data. *PAMI*, 18(8):837–842, Aug 1996.
22. R. Manthalkar, P.K. Biswas, and B.N. Chatterji. Rotation and scale invariant texture classification using gabor wavelets. In *Texture Workshop*, pages 87–90, 2002.
23. T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *PR*, 29(1):51–59, Jan 1996.
24. A. Penirschke, M.J. Chantler, and M. Petrou. Illuminant rotation invariant classification of 3D surface textures using Lissajous‘s ellipses. In *Texture Workshop*, pages 103–108, 2002.
25. J. C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. In *Proc. NIPS 2000, Denver, Colorado*, 2000.
26. M. Pontil and A. Verri. Support vector machines for 3D object recognition. *PAMI*, 20(6):637–646, June 1998.
27. D. Roobaert, M. Zillich, and J.O. Eklundh. A pure learning approach to background-invariant object recognition using pedagogical support vector learning. In *Proc. CVPR, Kauai, Hawaii*, pages II:351–357, 2001.
28. S. Singh, J. Haddon, and M. Markou. Nearest-neighbour classifiers in natural scene analysis. *PR*, 34(8):1601–1612, August 2001.
29. V. Vapnik. *Statistical learning theory*. Wiley and Son, New York, 1998.
30. M. Varma. Private communication, 2003.
31. M. Varma and A. Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. In *Proc. ECCV, Copenhagen*, page III: 255 ff., 2002.
32. M. Varma and A. Zisserman. Texture classification: are filter banks necessary? In *Proc. CVPR, Madison, Wisconsin*, pages II: 691–698, 2003.