A Sparse Texture Representation Using Affine-Invariant Regions

Svetlana Lazebnik
Beckman Institute
University of Illinois, Urbana, USA
slazebni@uiuc.edu

Cordelia Schmid
Inria Rhône-Alpes
Montbonnot, France
Cordelia.Schmid@inrialpes.fr

Jean Ponce
Beckman Institute
University of Illinois, Urbana, USA
ponce@cs.uiuc.edu

Abstract

This paper introduces a texture representation suitable for recognizing images of textured surfaces under a wide range of transformations, including viewpoint changes and non-rigid deformations. At the feature extraction stage, a sparse set of affine-invariant local patches is extracted from the image. This spatial selection process permits the computation of characteristic scale and neighborhood shape for every texture element. The proposed texture representation is evaluated in retrieval and classification tasks using the entire Brodatz database and a collection of photographs of textured surfaces taken from different viewpoints.

1. Introduction

Over the past decade, computer vision literature has reported texture recognition schemes [7, 12, 19] that perform impressively well on such challenging data sets as the Brodatz database [2]. Unfortunately, these schemes rely on restrictive assumptions about the input (e.g. texture must be stationary) and are not generally invariant with respect to 2D similarity and affine transformations, much less to 3D transformations caused by movement of the camera and non-rigid deformations of the textured surface. However, invariance to such transformations is desirable for many applications, including wide-baseline matching [11, 15, 17], texture-based retrieval [14, 16], segmentation of natural scenes [8], and recognition of materials [18].

In this paper, we design a texture representation that is invariant to any geometric transformations that can be locally approximated by an affine model. In practice, local affine invariants are capable of modeling not only global affine transformations of the image, but also perspective distortions and non-rigid deformations that preserve the locally flat structure of the surface (e.g. the bending of paper or cloth). Our goal is to develop a representation that addresses the problems of *spatial selection* (finding the most salient image locations for computing texture descriptors) and and *shape selection* (finding the characteristic size and shape of local texture neighborhoods).

The proposed method performs spatial and shape selection by restricting the set of candidate texture elements to a sparse set of affine-invariant regions in the image. Several types of affine-invariant region detectors have recently been developed for applications of wide-baseline matching, indexing, and retrieval [11, 17]. To us, these detectors represent powerful tools for creating sparse texture representations. In the current paper, we show how such representations may be built and investigate their effectiveness. The proposed method consists of four steps (see Figure 1):

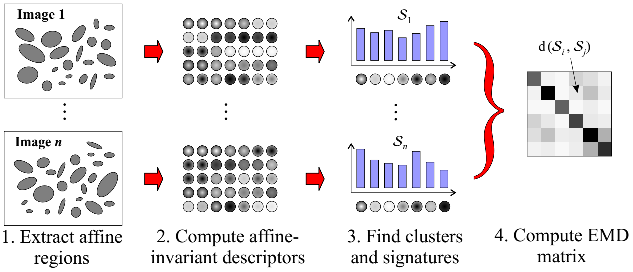


Image 1
Image n
$\mathcal{S}_1$
$\mathcal{S}_n$
$\mathrm{d}(\mathcal{S}_i, \mathcal{S}_j)$
1. Extract affine regions
2. Compute affine-invariant descriptors
3. Find clusters and signatures
4. Compute EMD matrix

Figure 1: Outline of the proposed method.

1. Extract a sparse set of affine regions from a texture image (Section 2.1).
2. For each region, compute an intensity descriptor that is invariant to affine geometric and photometric transformations. In Section 2.2, we discuss a novel descriptor based on *spin images* [3].
3. Perform clustering on the affine-invariant descriptors and summarize the distribution of descriptors in the form of a *signature* consisting of cluster centers and relative weights (Section 2.3).
4. Compare signatures using the *Earth Mover's Distance* (EMD) [13, 14], which is a convenient and effective dissimilarity measure applicable to many types of image information. The output of this stage is an *EMD matrix* whose (i, j)th entry is the distance between the signatures of the ith and jth image in the database. The EMD matrix can be used for retrieval and classification, as described in Section 3.1.

In Section 3, we evaluate the proposed texture representation on two data sets. The first set consists of photographs of textured surfaces taken from different viewpoints and featuring significant scale changes, perspective distortions, and non-rigid transformations. The second set is the Brodatz database, which has significant inter-class variability, but no geometric transformations between members of the same class. Because affine invariance is not required in this case, we modify the basic framework to use neighborhood shape as an additional discriminative feature.

2. Building the Representation

2.1 Affine-Invariant Regions

Blostein and Ahuja [1] have first introduced a multiscale blob detector based on maxima of the Laplacian of Gaussian. Lindeberg [6] has extended this detector in the framework of automatic scale selection, where a "blob" is defined by a maximum of a normalized Laplacian measure in scale space. Informally, the spatial coordinates of the maximum become the coordinates of the center of the blob, and the scale at which the maximum is achieved becomes the *characteristic scale*. Lindeberg and Gårding [5] have also shown how to design an affine-invariant blob detector using an *affine adaptation* process based on the second moment matrix. Mikolajczyk and Schmid [10, 11] have proposed alternative scale- and affine-invariant detectors that use a multi-scale version of the Harris interest point detector to localize interest points in space while employing Lindeberg's scheme for scale selection and affine adaptation.

Tuytelaars and Van Gool [17] have articulated the goal of building an "opportunistic" system that would combine the output of several region detectors tuned to different kinds of image structure. In this spirit, the texture representation proposed in this paper is designed to support multiple "channels" based on the regions found by different detectors. The four steps listed in Section 1 can be carried out separately for each available channel; at the final stage, the channels are combined as described in Section 2.3.

Our prototype implementation uses two channels: one based on the Harris-Affine detector of Mikolajczyk and Schmid [11], and the other on the affine-adapted Laplacian blob detector after Lindeberg and Gårding [5]. From now on, the respective detectors will be dubbed H and L. Figure 2 shows the output of L and H on two natural images. Intuitively, the two detectors provide complementary kinds of information about the image: H responds to corners and other regions of "high information content" [11], while L responds to blob-like regions of relatively uniform intensity. Note that L tends to produce a denser set of regions than H, though the absolute number of regions extracted from an image is small in any case (for the experiments of Section 3, this number is thresholded at 800 for each detector).

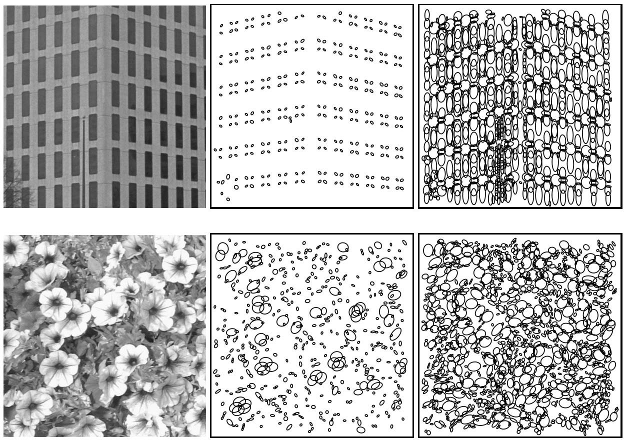


Figure 2: Left: original images, center: output of the H detector, right: output of the L detector.

The regions localized by the H and L detectors can be thought of as ellipses defined by $(\mathbf{x}-\mathbf{x}_0)^T M(\mathbf{x}-\mathbf{x}_0) \leq 1$, where $\mathbf{x}_0$ is the center of the ellipse, and M is a 2×2 symmetric *local shape matrix* (see [5, 11] for details). We can perform *affine normalization* on the patch defined by M by applying to it any transformation that would map the ellipse onto a unit circle. It can be shown that if two image patches are initially related by an affine transformation, then the respective normalized patches are related by an arbitrary orthogonal transformation [5, 11]. To eliminate this remaining one-parameter ambiguity, we can represent each normalized patch by a rotationally invariant descriptor.

2.2. Spin Images as Intensity Descriptors

In this paper, we describe a novel intensity-based rotation-invariant descriptor inspired by the idea of *spin images* introduced by Johnson and Hebert [3] for matching range data. An *intensity domain spin image* is a two-dimensional histogram encoding the distribution of brightness values in an affine-normalized patch. The two dimensions of the histogram are d, the distance from the center or the origin of the normalized coordinate system of the patch, and i, the intensity value. The "slice" of the spin image corresponding to a fixed d is simply the histogram of the intensity values of pixels located at a distance d from the center. Since the d and i parameters are invariant to orthogonal transformations, spin images offer exactly the right degree of invariance for representing affine-normalized patches. To achieve invariance to affine transformations of the intensity (transformations of the form $I \mapsto aI + b$), it is sufficient to normalize the range of the intensity function within the support region of the spin image. We implement the spin image as a "soft histogram" where each pixel within the support region contributes to more than one bin. Namely, the contribution of a pixel having distance d from the center and intensity value i to the bin indexed by (d_0, i_0) is proportional to $\exp[-(d-d_0)^2/(2\alpha^2)-(i-i_0)^2/(2\beta^2)]$, where α and β are "soft width" parameters. In our implementation, the distance between two spin images is given by the sum of squared differences (SSD), with the spin images normal-

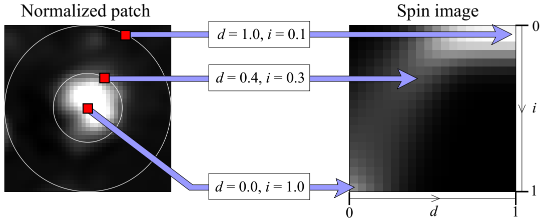


Figure 3: Spin image construction: three sample points in the normalized patch (left) map to three different locations in the spin image (right).

ized to have zero mean and unit Frobenius norm. Figure 3 illustrates the construction of spin images, and Figure 4 shows several examples.

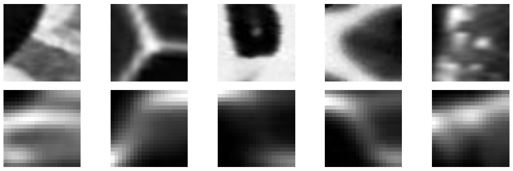


Figure 4: Normalized patches (top) and spin images (bottom).

We have compared the effectiveness of spin images to descriptors based on outputs of rotation-invariant “Gabor-like” linear filters [16, 18]. In a recent evaluation [18], Gabor-like filters had performance levels similar to those of several alternative filter banks. Therefore, they can be seen as fair representatives of currently used filter-based descriptors. Figures 6 (a) and (b) show the performance of spin images and Gabor-like filters on the data set described in Section 3.2. The fact that spin images perform better on our data makes intuitive sense: as a two-dimensional descriptor, the spin image is a richer representation of local appearance than a one-dimensional array of filter outputs.

2.3. Clustering and Signatures

Our method performs clustering to discover a small set of basic primitives in the set of descriptors (spin images) extracted from any single image. We use a standard agglomerative clustering algorithm [4] that successively merges clusters until either the desired target number of clusters is reached (10 to 15 in the implementation), or the distance between clusters exceeds a pre-specified threshold. Agglomerative clustering takes as input not the descriptors themselves, but only a matrix of pairwise distances between descriptors. Thus, the running time of clustering is effectively independent of the dimensionality of the feature space (we use spin images of size 20×20, so technically we are working with 400-dimensional features).

After the clustering stage is completed, we form the final representation for the image: a *signature* $\{(\mathbf{m}_i, u_i)\}$, where $\mathbf{m}_i$ is the *medoid* (the most centrally located element of the ith cluster) and u_i is the *weight* (the size of the cluster divided by the total number of descriptors in the image). Signatures have been introduced by Rubner et al. [13, 14] as representations suitable for matching using the *Earth Mover’s Distance* (EMD). The EMD between two signatures $\{(\mathbf{m}_i, u_i)\}$ and $\{(\mathbf{n}_j, v_j)\}$ has the form

$$\frac{\sum_i \sum_j f_{ij}\, \mathrm{d}(\mathbf{m}_i, \mathbf{n}_j)}{\sum_i \sum_j f_{ij}}$$

where the f_{ij} are *flow values* determined by solving a linear programming problem (see [13] for details), and the $\mathrm{d}(\mathbf{m}_i, \mathbf{n}_j)$ are *ground distances* between medoids. In our case, $\mathbf{m}_i$ and $\mathbf{n}_j$ are spin images and the distance is SSD.

For our application, the signature/EMD framework offers several important advantages. A signature is more descriptive than a histogram, and it does not require global clustering of the descriptors found in all images. EMD can match signatures of different sizes, and is not very sensitive to the number of clusters — that is, if one cluster is split into several, the magnitude of the EMD is not greatly affected [14]. This is an important property, since automatic selection of the number of clusters is generally difficult.

Recall that our texture representation involves multiple channels corresponding to different detectors. Each channel generates its own signature representation for each image in the database, and therefore its own EMD value for any pair of images. We have experimented with several methods of combining these separate values to obtain a cumulative inter-image distance measure. It was empirically determined that simply adding the separate distances with equal weights produces the best results.

3. Performance Analysis

3.1 Methodology

We have exercised the proposed texture representation in retrieval and classification tasks. For retrieval, we follow the procedure standardized in several previous studies [7, 12, 19]. Given a *query image*, we select images from the database in increasing order of EMD. Each image in the database is used once as a query image. The performance is summarized in a plot of average recognition rate (number of images from the class retrieved so far over the total number of images in the class) vs. the number of closest images retrieved. Perfect performance would correspond to 100% recognition rate after $n-1$ retrievals, where n is the number of images of the given class in the database.

In the above retrieval framework, we effectively treat each image as a model for its class, hoping that the rest of the images from the class will be very similar to this model. However, this reasoning does not work for databases with significant intra-class variability, where a given texture cannot be adequately modeled by any individual sample. To avoid this problem, we suggest classification as an addi-

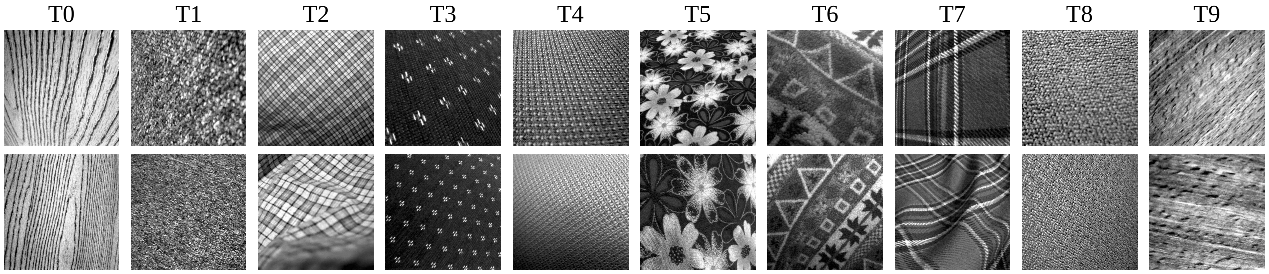


Figure 5: Samples of the ten texture classes used in the experiments of Section 3.2.

tional method for performance evaluation. In the classification framework, a model for a class is created not from a single (possibly atypical) image, but from several training images. As long as the training set adequately reflects the range of variation within the class, classification rate is not unduly affected by inhomogeneity.

In our implementation, we use nearest-neighbor classification with EMD. The training set is selected as a fixed-size random subset of the class, and all remaining images are added to the test set. To eliminate the dependence of classification rates on the particular training images used, this procedure is repeated with different random training sets (we use 50 iterations).

3.2. Viewpoint-Invariant Texture Recognition

We have acquired a data set consisting of 20 samples each of ten different textured surfaces, for a total of 200 images (Figure 5). Significant viewpoint and scale changes are featured within each class. To push the limits of our system, we have allowed additional sources of variability: inhomogeneities in the texture pattern, non-rigid transformations, illumination changes, and unmodeled viewpoint-dependent appearance changes (these are the most severe for class T4, whose "textured" look is due largely to the pattern of light and shadow on a bumpy surface).

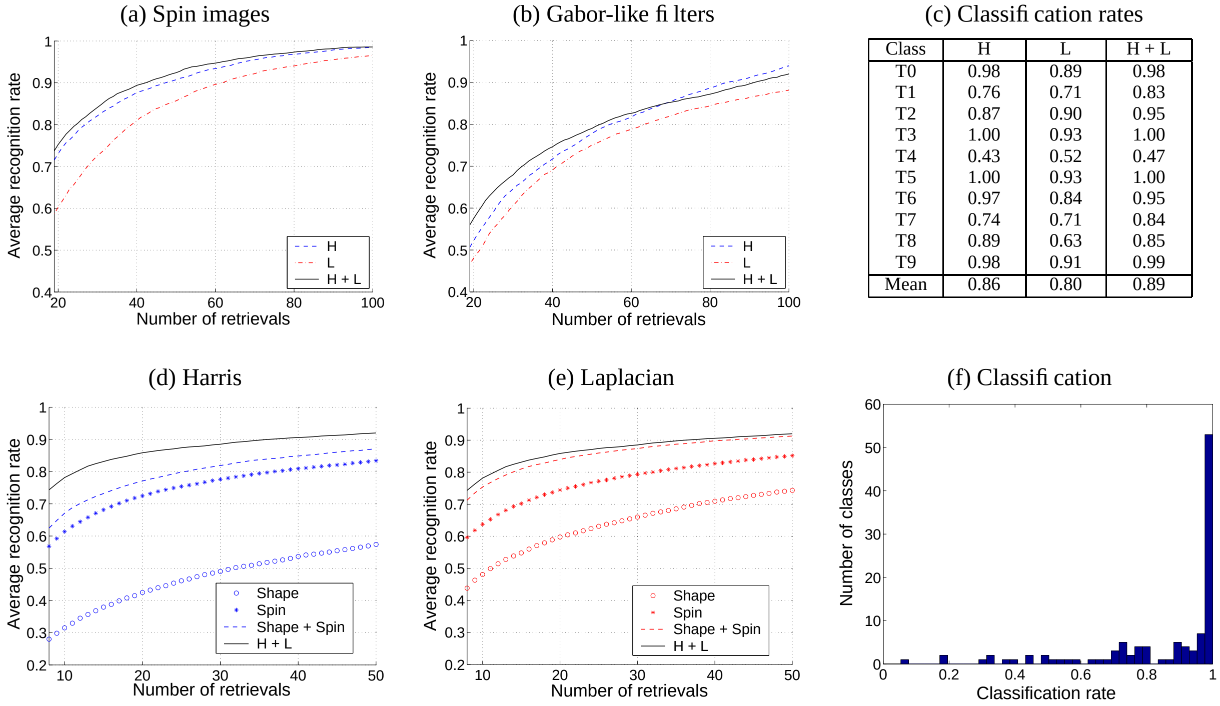
Figure 6 (a) shows retrieval results using spin images as intensity-based descriptors. Notice that for this data set, the H channel is more discriminative than the L channel. Nevertheless, adding the two EMD estimates results in improved performance. This "hyperacuity effect" is the strongest argument for combining the output of different feature extractors. Figure 6 (b) shows the results using Gabor-like filters as descriptors instead of spin images (see Section 2.2 for discussion). Figure 6 (c) summarizes the classification results obtained by using 5 samples from each class as training images. The classification rate for each class provides an indication of the "difficulty" of this class for our representation. The mean rate is 0.89, with two classes achieving 1.0, showing the robustness of our system against a large amount of intra-class variability. In particular, performance is very good for relatively inhomogeneous textures T5 and T6. Class T4 has the lowest recognition rate, which is probably due to excessive variability in local appearance caused by the complex interaction between viewpoint, lighting, and fine-scale 3D structure. Indeed, about half of the signatures for this class are dominated by (roughly speaking) light center/dark surround blobs, while the remaining signatures are dominated by dark center/light surround blobs. The EMD between these two types of signatures is quite high, explaining the relatively low recognition rate.

3.3. Brodatz Database Evaluation

In the experiments shown in the previous section, we discard the information contained in the shape of the patches computed using affine adaptation (recall Section 2.1). However, shape can be a distinctive feature when affine invariance is not required. In this section, we present a modification of our system that takes advantage of affine shape for recognition, and evaluate it on the Brodatz database.

Let E_1 and E_2 be two regions defined by local shape matrices M_1 and M_2. We eliminate the translation between E_1 and E_2 by aligning their centers, and then compute the dissimilarity between the regions as $\mathrm{d}(E_1, E_2) = 1 - \mathrm{Area}(E_1 \cap E_2) / \mathrm{Area}(E_1 \cup E_2)$. Notice that this measure takes into account the relative rotations of the two ellipses. We can achieve rotation invariance simply by aligning the major and minor axes of the ellipses before comparing their areas. In the experiments of this section, we use local shape to obtain one additional channel for each region detector. This involves a separate clustering stage for each set of regions based on the "ellipse distance" $\mathrm{d}(E_1, E_2)$, as well as the computation of shape-based signatures and EMD's between all pairs of images. At the end, the shape-based EMD matrices are combined with the spin image-based EMD matrices through addition.

The Brodatz database is a widely used benchmark for texture recognition. Following the same procedure as previous evaluations [7, 12, 19], we form the classes by partitioning each of the 111 images into nine non-overlapping fragments, for a total of 999 images. As many authors have noted before, the wide variety of Brodatz textures, combined with a certain degree of subjectivity in assigned texture classes (some perceptually similar textures have different labels, while a few textures are too inhomogeneous to permit successful recognition) makes the Brodatz database a challenging platform for performance analysis.



Class	H	L	H + L
T0	0.98	0.89	0.98
T1	0.76	0.71	0.83
T2	0.87	0.90	0.95
T3	1.00	0.93	1.00
T4	0.43	0.52	0.47
T5	1.00	0.93	1.00
T6	0.97	0.84	0.95
T7	0.74	0.71	0.84
T8	0.89	0.63	0.85
T9	0.98	0.91	0.99
Mean	0.86	0.80	0.89

Figure 6: (a), (b), (c) Retrieval and classification performance for the data set shown in Figure 5. (d), (e) Retrieval using the Brodatz database. The solid black curve is the same in both plots. This curve represents the best performance of our system (see text). (f) Histogram of classification rates for all 111 classes.

Parts (d) and (e) of Figure 6 show retrieval results using the shape channel alone, the spin image channel alone, and the two channels combined. By comparing the two plots, we can see that H and L spin image channels produce very similar results, but the shape information for H is much less descriptive than for L. As a result, the combined L channel outperforms the combined H channel, which is the reverse of the situation seen in Section 3.2. The final performance of the system, shown as the solid curve in Figures 6 (d) and (e), is obtained by combining the shape and spin image channels from both detectors.

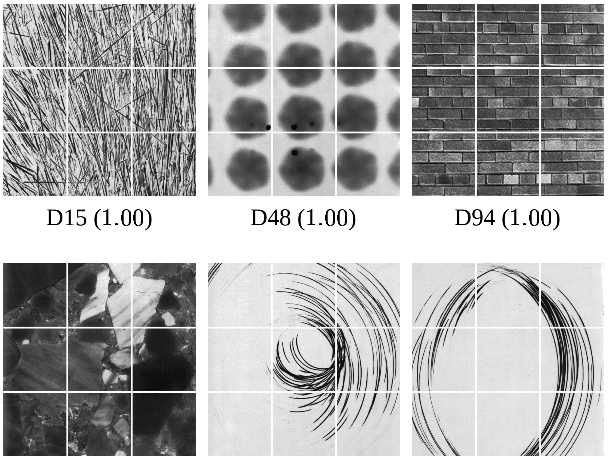
Figure 6 (f) gives us a more detailed means of performance analysis. It shows a histogram of classification rates for all 111 classes obtained by putting three images from each class into the training set. The histogram reveals that the majority of textures are highly distinguishable, with only a few stragglers at the low end. In fact, 41 classes have 1.0 classification rate, and the mean rate is 0.85. Three of these classes are shown in the top part of Figure 7, with the three bottom textures below. Not surprisingly, the bottom three textures are highly inhomogeneous. While the tests of Section 3.2 show that our method provides some robustness against inhomogeneity, textures with large amounts of spatial variation remain difficult to recognize.

The best performance curve of our system has about 0.75 recognition rate after 8 retrievals. This result is on par with Picard et al. [7, 12], but somewhat below Xu et al. [19], who report 0.84 recognition rate using the multiresolution simultaneous autoregressive (MRSAR) model [9]. MRSAR is completely different from our method: it models texture as a stationary random field, uses a dense representation with fixed neighborhood shape and size, and has no scale, rotation, or affine-invariance. Traditional texture models such as MRSAR have been studied and perfected for at least a decade, while our method is built on new techniques that have not previously been applied to texture analysis. We believe that "mature" methods such as MRSAR have been pushed close to the intrinsic limit of their performance, while novel methods such as ours have a much greater potential for improvement in the future.

4. Discussion and Future Work

In this paper, we have introduced a non-parametric texture representation that provides several kinds of geometric invariance, does not make any statistical assumptions about the input texture, and applies spatial and shape selection to automatically determine the locations and support regions of salient texture neighborhoods.

In the future, we will pursue several avenues for improvement of our method. One important research direction is acquiring a better understanding of the relative expressiveness of the H and L detectors. In our experiments, H worked better for the data set of Section 3.2, while L gave



D15 (1.00) D48 (1.00) D94 (1.00)

D59 (0.05) D43 (0.18) D44 (0.19)

Figure 7: Top: three of the 41 textures having classification rate of 1.0. Interestingly, the representation for D48 has an average of only 9 L and 5 H regions per sample. Bottom: three textures with the worst classification rates (shown in parentheses).

better performance on the Brodatz database. We plan to conduct a more systematic study of the on larger databases, and to develop a quantitative measure of the discriminative power of different detectors on different types of texture. This study should also involve more detectors, e.g., the ones proposed by Tuytelaars and Van Gool [17].

Another issue worthy of further research is the method for combining channels. For our two data sets, the simple method of adding the individual EMD matrices has worked surprisingly well to improve average performance. However, as can be seen from Figure 6 (b), it is occasionally possible for the combined recognition rate to be lower than the single-channel rates. In the future, we plan to study more sophisticated methods for combining channels that would not suffer from similar detrimental effects.

Finally, we believe that a significant increase in discriminative power will come from using features based on spatial relationships between neighoring regions. A few recent representations [8, 16] have used a two-level scheme, with intensity-based textons at the first level and histograms of texton distributions over local neighborhoods at the second level. For many natural textures, the arrangement of affine regions captures perceptually significant information about the global geometric structure. Augmenting our representation with such information is likely to increase its ability to distinguish textures that have similar local neighborhoods but different spatial layouts.

Acknowledgments. This research was partially supported by the UIUC Campus Research Board, by the National Science Foundation under grant IRI-990709, and by the European project LAVA (IST-2001-34405).

References

[1] D. Blostein and N. Ahuja, "A Multiscale Region Detector", *CVGIP* 45, 1989, pp. 22-41.

[2] P. Brodatz, *Textures: A Photographic Album for Artists and Designers*, Dover, New York, 1966.

[3] A. Johnson and M. Hebert, "Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes", *IEEE Trans. PAMI* 21(5), 1999, pp. 433-449.

[4] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York, 1990.

[5] T. Lindeberg and J. Gårding, "Shape-Adapted Smoothing in Estimation of 3-D Depth Cues from Affine Distortions of Local 2-D Brightness Structure", *Image and Vision Computing* 15, 1997, pp. 415-434.

[6] T. Lindeberg, "Feature Detection with Automatic Scale Selection", *IJCV* 30(2), 1998, pp. 77-116.

[7] F. Liu and R. W. Picard, "Periodicity, Directionality, and Randomness: Wold Features for Image Modeling and Retrieval", *IEEE Trans. PAMI* 18(7), 1996, pp. 722-733.

[8] J. Malik, S. Belongie, T. Leung and J. Shi, "Contour and Texture Analysis for Image Segmentation", *IJCV* 43(1), 2001, pp. 7-27.

[9] J. Mao and A. Jain, "Texture Classification and Segmentation Using Multiresolution Simultaneous Autoregressive Models", *Pattern Recognition* 25, 1992, pp. 173-188.

[10] K. Mikolajczyk and C. Schmid, "Indexing Based on Scale Invariant Interest Points", *Proc. ICCV* 2001, vol. 1, pp. 525-531.

[11] K. Mikolajczyk and C. Schmid, "An Affine Invariant Interest Point Detector", *Proc. ECCV* 2002, vol. 1, pp. 128-142.

[12] R. Picard, T. Kabir and F. Liu, "Real-time Recognition with the Entire Brodatz Texture Database", *Proc. CVPR* 1993, pp. 638-639.

[13] Y. Rubner, C. Tomasi and L. Guibas, "A Metric for Distributions with Applications to Image Databases", *Proc. ICCV* 1998, pp. 59-66.

[14] Y. Rubner and C. Tomasi, "Texture-Based Image Retrieval Without Segmentation", *Proc. ICCV* 1999, pp. 1018-1024.

[15] F. Schaffalitzky and A. Zisserman, "Viewpoint Invariant Texture Matching and Wide Baseline Stereo", *Proc. ICCV* 2001, vol. 2, pp. 636-643.

[16] C. Schmid, "Constructing Models for Content-Based Image Retrieval", *Proc. CVPR* 2001, vol. 2, pp. 39-45.

[17] T. Tuytelaars and L. Van Gool, "Matching Widely Separated Views based on Affinely Invariant Neighbourhoods", submitted to *IJCV*, 2001.

[18] M. Varma and A. Zisserman, "Classifying Images of Materials: Achieving Viewpoint and Illumination Independence", *Proc. ECCV* 2002, vol. 3, pp. 255-271.

[19] K. Xu, B. Georgescu, D. Comaniciu and P. Meer, "Performance Analysis in Content-based Retrieval with Textures", *Proc. ICPR* 2000, vol. 4, pp. 275-278.