Towards texture classification in real scenes

Paul Southam and Richard Harvey
School of Computing Sciences
University of East Anglia
Norwich, UK
{*pauls*, *rwh*}@cmp.uea.ac.uk

Abstract

Two new texture features, based on morphological scale-space processors are introduced. The new methods are shown to have good performance over a variety of tests. We demonstrate that if texture classifies are to be used in real world scenes, then the choice of test is critical and that Brodatz-like tests are unlikely to represent reality.

1 Introduction

There are a number of motivations for studying texture classification. Sometimes there is a real problem such as the classification of marble, sorting of wood, classification of carpet and so on. Other times one is interested to study a stylised classification problem as a convenient benchmark. This latter position is more difficult since there is a nagging doubt that results on stylised experiments will not extrapolate to reality. This paper examines this doubt, using available texture databases. The well known texture databases are Brodatz; VisTex; MeasTex and Outex (see [6] for a summery). CURet [3], although designed for a different task, might also be mentioned.

	Brodatz	**VisTex**	**MeasTex**	**Outex**	**CUReT**
No. Classes	100–112*	19	4	29	61
No. Images per class	1–4*	1–20	4–25	1–47	205†
Evaluation Framework?	×	×	√	√	×
Defined test/training data?	×	×	√	√	×
Unique copy?	×	√	√	√	√

Table 1: Available texture databases (* Depending on implementation, † At differing angle and illuminant.)

With the exception of Brodatz, Table 1 shows that all are available for download which ensures that each experiment is performed with identical texture samples. Brodatz, although the most common, is available as a book and since different methods and

BMVC 2005 doi:10.5244/C.19.25

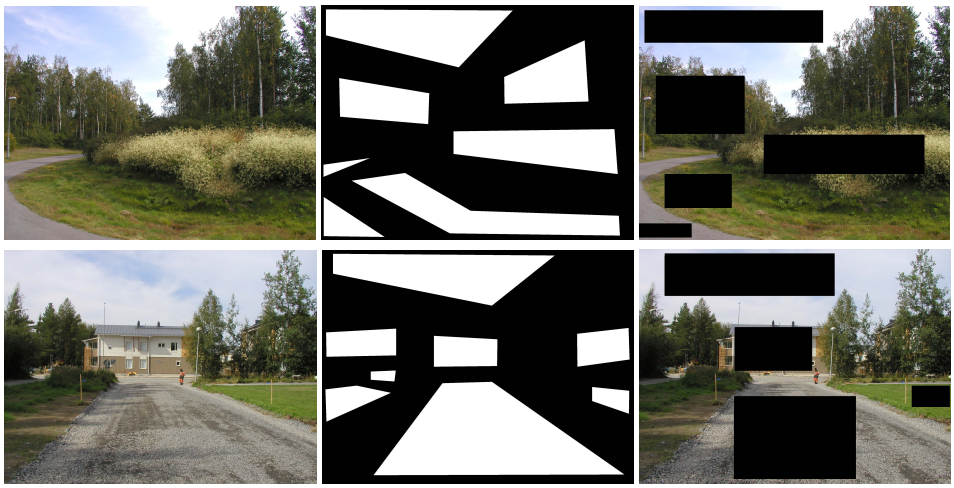

Figure 1: Three Outex natural scene images (left column), hand-segmented ground-truth supplied with the database (middle column) and our new ground-truth (right column). Note that each region has lighting, scale and rotation variations.

equipment were used for digitisation there are several electronic versions at different file, size, format and resolution; all of which affect results[1]. Brodatz, VisTex and CuRET do not have pre-defined test and training data which hinders comparisons of different classification methods. See [10], for example, where different test and training data lead to conflicting results. MeasTex and VisTex have a few samples and classes. VisTex is no longer maintained. In [11] it is claimed that the MesTex classes are non-overlapping for a greater number of methods. Therefore Outex is the best dataset for comparing texture classifiers and should be the database of choice for comparing texture classifiers.

Currently Outex has 319, grey-scale and colour textures spanning 29 classes. The images are organised into *test suites* setup for, classification, segmentation, and retrieval. The test suites are designed to examine illumination, scale, rotation and colour invariance. There is also a natural scenes test suite which contains 20 colour images (2272×1704 pixels) of scenes under varying illumination and orientation. There are five texture classes, *sky*, *trees*, *grass*, *road* and *buildings* which are defined through hand-labelled ground-truth. Examples of the scene and ground truth images are in Figure 1.

To compare texture classifiers that are restricted to rectangular windows, we provide alternative hand-segmented ground truth also shown in Figure 1. We use the same class labels as the original ground-truth.

[1] [9] for example shows images that have an accidental gray-scale inversion compared to the book.

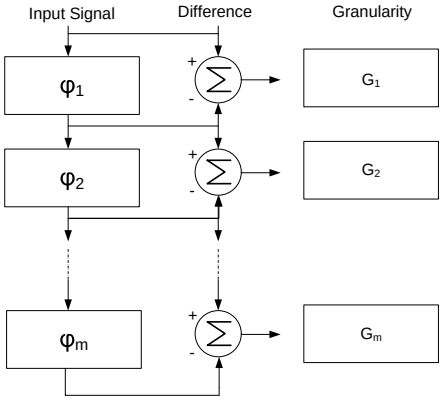

Figure 2: The structure of a sieve decomposition where φ is a filtering operator chosen from a set [1, 2]. Non-zero regions in the output are called *granules* and the set of granules is called the *granularity domain* in an analogy to granulometries.

2 Methods

There are a very large number of proposed methods of classifying texture. Here we attempt to represent performance via three techniques that have been reported to work well in the literature.

The first benchmark is the dual-tree complex wavelet transform (DTCWT) [5], which is known to have better shift invariance and directional selectivity than conventional wavelets. The DTCWT decomposes an image into eight sub-bands: two low-pass and six high-pass at orientations of $\pm 15^\circ, \pm 45^\circ, \pm 75^\circ$. Decomposing to three levels produces $8 + 8^2 + 8^3$ bands. The sample mean and standard deviation of absolute values of these bands are taken to form a feature vector of 336 elements.

The second benchmark is the local binary pattern (LBP) [7] method, which produces features based upon the spatial structure of an image using absolute grey-level intensity differences between neighbouring pixels. It works by passing a 3×3 window over every pixel in the texture. Pixels that have intensities higher than or equal to, the target pixel are masked off. These masked pixels are assigned weightings that are summed to produce a LBP-score within the range $0 \longrightarrow 255$. Hence LBP produces a feature vector of 256 elements.

The third benchmark uses co-occurrence matrices (co-occ) introduced in [4]. Co-occurrence matrices extract second order statistics based upon the vector between two pixels in an image. Once the matrix of frequencies has been created, a number of features can be derived from it. We use 12 features representing energy, contrast, entropy and homogeneity at orientations of $0^\circ, 90^\circ$ and 45°.

We also introduce new methods based upon the morphological operator known as the *sieve*. Sieves were originally defined as one-dimensional systems [1] but were later extended to *n*-dimensional filters [2] that adopt techniques from graph morphology. They are cascades of morphological scale-space operators that remove intensity extrema at a

specific scale via the structure shown in Figure 2.

In the first new method, denoted (2D-sieve), φ is a 2D $\mathscr{M}$-filter [2] which filters the image using a morphological opening followed by a morphological closing in one operation. This method produces a decomposition that, ignoring sampling errors, is invariant to rotation. At small scales, the processor tends to remove noise; then, as the scale increases, texture; then, objects within the scene as in, for example, Figure 3.

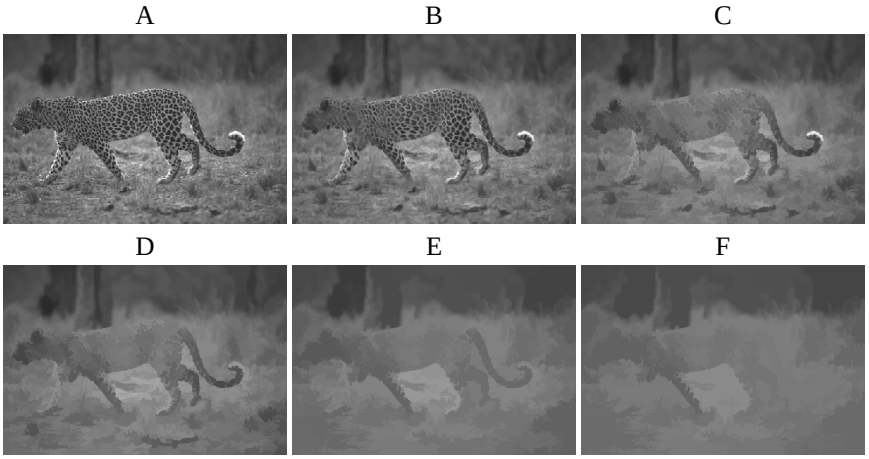

Figure 3: An original image (A) sieved using a 2D-sieve to scales 15(B), 90(C), 251(D), 2000(E) and 5000(F). Each image has fewer intensity extrema than its predecessor. A full decomposition may be summed to re-create the original thus the sieve is a transform of the original image.

The second new method (1D-sieve) is similar, but now there is an additional parameter: the orientation at which the filter is applied. Here φ is a 1D recursive median filter, as in [13] for example, at orientations of $0^\circ, \pm 30^\circ, \pm 45^\circ, +90^\circ$. The 2D-sieve decomposes an image by scale equal to the area of intensity extrema, where as the 1D-sieve decomposes by length. Figure 4 shows an example 1D-sieve decomposition from which we can see that the decomposition is anisotropic.

The features from both new methods are based on *granules*. In [2], granule images are defined as the difference between successive sieve outputs, $G_n = \varphi_n - \varphi_{n-1}$ where φ_n is the nth stage in the serial structure shown in Figure 2. There are thus a great number of granule images. The G_n are a transform of the texture which can be reconstructed through a simple summation. Here, each texture image is sieved to a few scales, $[s_1 \ldots s_N]$ where $\log_{10} s_n$ are equispaced between 0 and $\log_{10} P$, where $P = 30$ is chosen to remove all textural information from all images. The difference between these images are termed *channels*, $C_n = \varphi_{s_n} - \varphi_{s_{n-1}}$. We choose $N = 5$ to give 5 channels. The magnitude of the channel images as a function of scale is an indicator of the scale-distribution of the texture features. The sample mean, standard deviation and skewness of the magnitude of the granule images may be used as features. Thus the 2D-sieve has 15 features and the 1D-sieve has 90 features.

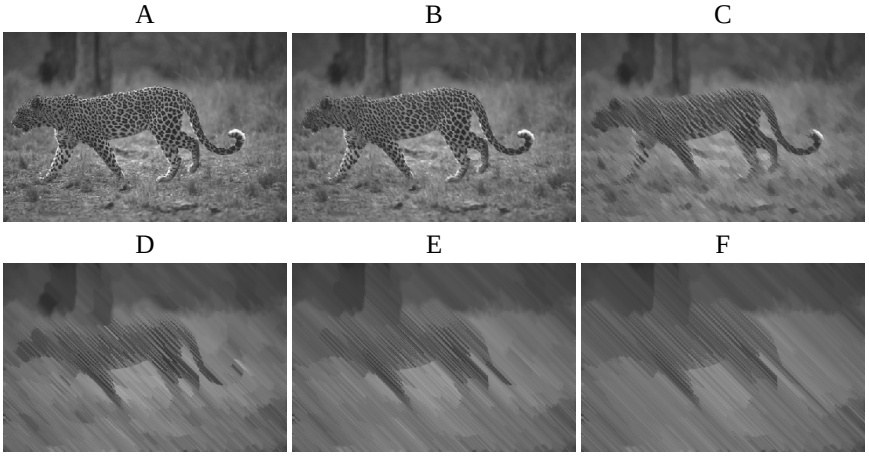

Figure 4: An original image (A) sieved using a 1D-sieve to scales 1(B), 5(C), 16(D), 29(E) and 50(F) at an angle of 45°.

Figure 5 (left) hows the sample mean for three texture classes using a 2D-sieve.The standard deviation and skewness are not show here but have similar variations. Finer scale textures have peak responses at lower scales and the coarser textures have peaks at higher scales.

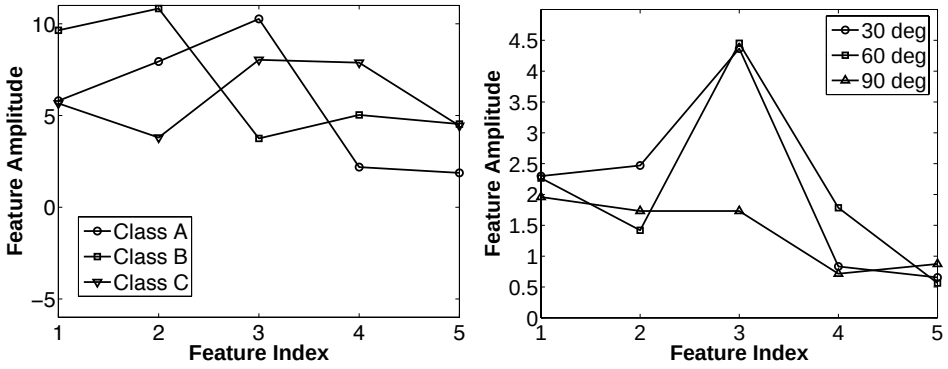

Figure 5: Left: Sample mean against feature index of three texture classes A,B and C shown in Figure 7. Right: The sample mean of a texture class D shown in Figure 7 at 30°, 60° and 90°. Feature index is correlated with scale: 1 is scale (0-1), 2 is scale (1-2), 3 is scale (2-5), 4 is scale (5-13) and 5 is scale (13-30).

Figure 5 (right) also shows how the sample mean varies for a stripy texture such as class A (shown in Figure 7) using a 1D-sieve applied at several angles. At 90° where the

scan-line runs along the direction of the texture, the intensity variation is less resulting in a flatter response.

Fast algorithms exist for sieves, and in practice a full decomposition of a domestic video image takes a couple of hundred ms. In all methods we compute the covariance matrix over the training data and apply principal component analysis (PCA) retaining all components: since we have found that this always improves performance on all methods and all databases.

3 Results

The first test, **Test-1**, uses the Outex_TC_00000 test suite. It contains 100 leave-out-half cross validation, classification experiments (different permutations of 240 testing and 240 training data) of 480 images. There are 20 samples of 24 texture classes. Classification is via a k-nearest neighbour classifier (elsewhere [12] it has been shown that k = 1, Euclidean distance is the best choice for this test). Table 2 shows the results. The success rate of the LBP method differs from [7] (0.995) because in our experiments we are not using a histogram distance measure. The results for Gabor wavelets and Gaussian Markov Random Fields are taken directly from [8] and it is not clear what classifier or distance measure is being used.

	1D-Sieve	**2D-Sieve**	**DTCWT**	**LBP**	**Co-occ**	**Gabor**	**GMRF**
$\bar{x}$	0.998	0.962	0.999	0.986	0.946	0.995	0.961
max	1	0.992	1	0.996	0.983	1	0.992
min	0.988	0.933	0.988	0.967	0.900	0.983	0.925
σ	0.0034	0.0103	0.0019	0.0070	0.0154	0.5	1.3
σ'	0.00034	0.00103	0.00019	0.00070	0.00154	0.05	0.13
f	90	15	336	256	12	??	??

Table 2: Mean success rate, $\bar{x}$, over 100 trials, max and min success rate, standard deviation σ, standard error of the mean $\sigma' = \sigma \setminus \sqrt{N}$ and number of features f for Outex test suite Outex_TC_00000. The results for the Gabor and GMRF methods are taken from [8] which doesn't specify the number of features used.

Table 3 shows the result of using McNemars's test on each trial at a significance of $\alpha = 0.05$. The 1D-sieve and DTCWT are indistinguishable at a significance for all 100 tests with Outex_TC_00000.

Using PCA to reduce feature dimensionality improves the 1D-sieve success rate to 0.999 which is similar to the DTCWT but with only 40 features. The affect of applying PCA to the DTCWT does not improve performance but maintains a 0.999 success rate using 77 features. Figure 6 shows the affect on success rate of reducing the number of features using PCA. The 1D-sieve has the highest success rate for the fewest features.

To test for noise-invariance we repeat the experiments using the Outex_TC_00000 test suite. Multiplicative noise (n) is added to each test image (I) resulting in a noisy image $J = I + n * I$, where n is uniformly distributed random noise with mean 0 and variance

	1D-Sieve	2D-Sieve	DTCWT	LBP	Co-occ
1D-sieve	x	90	0	10	96
2D-sieve	-	x	93	22	10
DTCWT	-	-	x	12	98
LBP	-	-	-	x	67

Table 3: The number of times out of the Outex_TC_00000 100 trials that we can confidently ($\alpha = 0.05$) reject the null hypothesis that the two data distributions are drawn from the same source.

$[V_1 \cdots V_k]$ where $log_{10}V_k$ are equispaced between 0 and 0.0115. Figure 6 show that the 1D- and 2D-sieve are the most invariant to multiplicative noise [2].

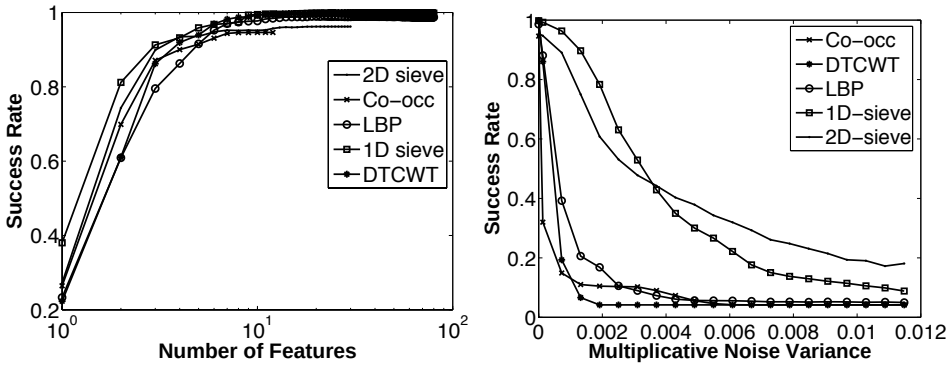

Figure 6: Success rate against log number of features (left) and Success rate against multiplicative noise variance (right)

In the real world the orientation of a texture class may be unknown. So in **Test-2** we use Outex_TC_00010 to test for rotation invariance. This test suite contains textures under constant illuminant and scale but with rotations of $00^\circ, 05^\circ, 10^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ$. There are 480 training images of 24 classes and 3840 test images. Table 4 shows the results. It appears that the performance of some methods degrades badly under unknown rotation. The 2D-sieve is now the best performing.

	1D-Sieve	2D-Sieve	DTCWT	LBP	Co-occ
$\bar{x}$	71.80	94.30	55.63	50.91	69.17
f	90	30	336	256	12

Table 4: Mean success rate, $\bar{x}$, and number of features f for rotation-variant Outex test suite Outex_TC_00010.

[2] We have similar results, not reported here, for additive Gaussian noise.

Test-3 concerns natural scenes. Rectangular regions (right column of Figure 1) are cropped from images of natural scenes to give texture images of the type shown in the centre row of Figure 3. There are 80 images which is too few for hold-out so they are used in 80 leave-out-one cross-validation experiments. The top half of Table 5 shows the results with a knn classifier ($k = 3$, Euclidean distance).

Test-4 uses the Outex natural scenes database but now the ground-truth is polygonal. There are a total of 91 labelled regions so again we use leave-out-one cross-validation. Table 5 shows the success rate across all classes, with a knn classifier ($k = 3$, Euclidean distance). Also shown is the number of samples per class. Not all texture methods are easily applicable to this set because the regions are non-rectangular. We therefore restrict the comparison to the LBP and sieve methods because, for these, we can generate the filtered images, apply the hand-segmented region as a mask, and generate a feature for each region. In Test-3 and Test-4 the best performing method on a per-class-winner or overall mean basis is the 2D-sieve.

	sky	tree	bush	grass	road	building	mean
Test-3							
No. Samples	14	15	10	19	14	8	
1D-sieve	1	0.933	0.4	0.84	0.71	0.38	0.71
2D-sieve	1	1	0.5	0.63	0.71	0.5	0.72
DTCWT	1	0.8	0.3	0.84	0.64	0.63	0.70
LBP	0.79	0.6	0.5	0.53	0.57	0.13	0.52
co-occ	0.93	0.93	0.3	0.37	0.57	0.5	0.60
Test-4							
No. Samples	14	17	15	20	16	9	
1D-sieve	0.79	0.59	0.2	0.6	0.38	0	0.42
2D-sieve	1	0.82	0.53	0.65	0.88	0.33	0.7
LBP	0.64	0.76	0.6	0.4	0.81	0.22	0.57

Table 5: Table shows number of samples per class, mean success rate per class and overall mean success rate for the Outex natural scene database.

4 Conclusions

This paper presents new texture classifiers and compares their performance to some well known benchmarks. The first test, Test-1, was quite conventional and is representative of a large number of texture evaluations in the literature. Had we stopped at this point then, the conclusion would be that one of the new methods, 1D-sieve, was as good as one of the benchmarks (DTCWT) but used fewer features.

Test-2 represents the situation where the orientation is unknown. Now, the previously best performing methods degrade. The DTCWT is now the second worst method and the 1D-sieve is out-performed by the 2D-sieve which we were confident was not the best method previously.

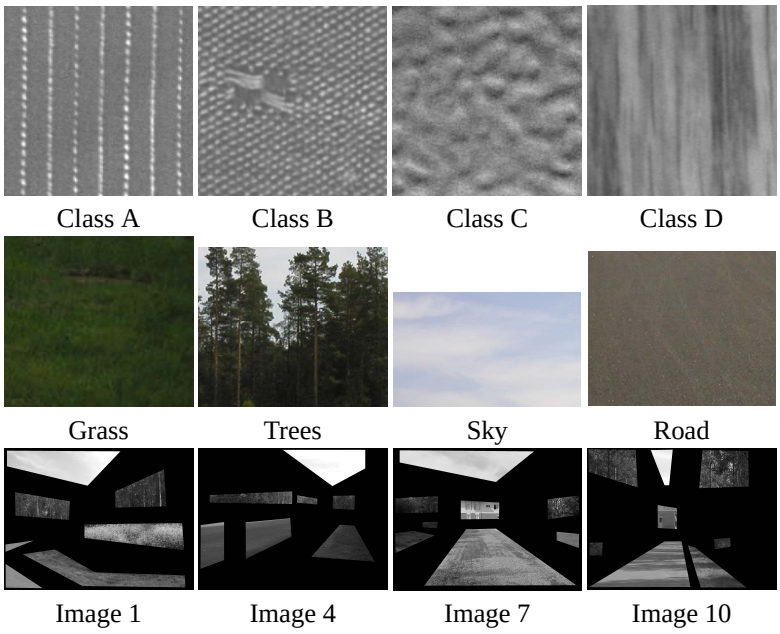

Figure 7: Sample data from: Test-1 (row 1), Test-3 (row 2), Test-4 (row 3).

This conclusion is reinforced by Test-3 and Test-4 with real images which have scale variation, perspective, class-overlap and so on. In Test-3 the 2D-sieve is the best performing method on four of the six classes. In Test-3 it's worth noting that the training and test data were sub-images formed by rectangular regions from the Outex natural scene test suite. This is very favourable to 1D-sieve, DTCWT and co-ooc because these methods have a fixed-shape support region which is unlikely to match the shape of natural texture regions (note for example the decline in performance of the 1D-sieve in Test-4). Therefore, there is a double problem with some conventional texture classifiers: not only can they have poor performance, they may also be impractical.

This paper therefore, firstly, demonstrates that caution is needed when interpreting texture classifications of the Brodatz-like type measured in Test-1: such tests will not represent performance on many real problems. Secondly, we have presented a technique that works well with both stylised situations and in reality. For the future we wish to encourage texture classification in the context of real world scenes since it is here that it is most useful to the general vision community.

References

[1] J.A. Bangham, P. Chardaire, C.J. Pye, and P.D. Ling. Multiscale nonlinear decomposition: The sieve decomposition theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(5):529–539, May 1996.

[2] J.A. Bangham, R. Harvey, P.D. Ling, and R.V. Aldridge. Morphological scale-space preserving transforms in many dimensions. *Journal of Electronic Imaging*, 5:283–299, 1996.

[3] CUReT. http://www1.cs.columbia.edu/cave/curet/.

[4] R. Haralick. Statistical and structural approaches to texture. In *Proccedings of IEEE*, volume 67, pages 786–804, 1979.

[5] N. Kingsbury. Image processing with complex wavelets. In *Phil. Trans. Royal Society London A, September 1999, on a Discussion Meeting on Wavelets: the key to intermittent information?, London, February 24-25*, 1999.

[6] T. Ojala, T. Maenpaa, M. Pietikainen, J. Viertola, J. Kyllonen, and S. Huovinen. Outex - new framework for empirical evaluation of texture analysis algorithms. In *Proc. 16th International Conference on Pattern Recognition, Quebec, Canada*, volume 1, pages 701–706, 2002.

[7] T. Ojala, M. Pietikanien, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29:51–59, 1996.

[8] Outex. http://www.outex.oulu.fi.

[9] C.-M. Pun and M.-C. Lee. Log-polar wavelet energy signatures for rotation and scale invariant texture classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):590–603, 2003.

[10] T. Randen and J. Husoy. Filtering for texture classification: A comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):291–310, 1999.

[11] S. Singh and M. Sharma. Texture analysis experiments with meastex and vistex benchmarks. *Lecture Notes in Computer Science*, 2013:417–424, 2001.

[12] Paul Southam and Richard Harvey. Compact rotation-invariant texture classification. In *International Conference on Image Processing (ICIP 2004)*, pages 3033–3036. IEEE, 24–27 Oct 2004.

[13] R. Zwiggelaar, T.C. Parr, J.E. Schumm, I.W. Hutt, S.M. Astley, C.J. Taylor, and C.R.M. Boggis. Model-based detection of spiculated lesions in mammograms. *Medical Image Analysis*, 3(1):39–62, 1999.