Evaluation of Texture Segmentation Algorithms

Kyong I. Chang, Kevin W. Bowyer and Munish Sivagurunath
Department of Computer Science & Engineering
University of South Florida
Tampa, FL 33620
chang, kwb or munish@csee.usf.edu

Abstract

This paper presents a method of evaluating unsupervised texture segmentation algorithms. The control scheme of texture segmentation has been conceptualized as two modular processes: (1) feature computation and (2) segmentation of homogeneous regions based on the feature values. Three feature extraction methods are considered: gray level co-occurrence matrix, Laws' texture energy and Gabor multi-channel filtering. Three segmentation algorithms are considered: fuzzy c-means clustering, square-error clustering and split-and-merge. A set of 35 real scene images with manually-specified ground truth was compiled. Performance is measured against ground truth on real images using region-based and pixel-based performance metrics.

1 Introduction

One recent review categorized texture segmentation techniques into feature-based methods, model-based methods and spatial/spatial-frequency methods and structural methods [21]. Feature-based methods characterize a texture as a homogeneous distribution of feature values such as gray level co-occurrence matrix (GLCM) and Laws' texture energy (LAWS). Even though both GLCM and LAWS were originally proposed in the context of texture classification, many researchers have applied them to texture segmentation [14, 23, 6, 5, 15, 4]. Spatial/spatial-frequency methods use a technique to generate a group of features from filtered images computed from frequency information at localized regions, such as Gabor functions or wavelet models [16]. Gabor multi-channel filtering (GABOR) has been selected for this study. Gabor filtering has been applied to the texture segmentation problem by many researchers [9, 2, 17]. Related comparison studies are reviewed in the next section. Few rigorous comparison and evaluation studies have been performed in unsupervised texture segmentation. The purpose of our study is to rigorously compare and evaluate the performance of different unsupervised texture segmentation techniques.

1.1 Review of Texture Segmentations

Previous studies have commonly used "mosaic" images, where region boundaries are artificial. A traditional source of texture samples is Brodatz's album [3]. The reason for using mosaic images is that the texture boundaries are precisely known, which allows for precise quantitative error evaluation [8]. But, such simple boundaries usually do not occur in real scenes.

Texture feature extraction techniques have been compared in several studies. Du Buf *et al.* compared seven different texture feature extraction methods (GLCM, fractal, Michell's, Knutsson's, Laws', Unser's, curvilinear integration) [8]. This is the closest study to ours, in terms of comparing unsupervised segmentation algorithms. It is an important study since they attempted to evaluate issues of image segmentation and boundary accuracy comparison in a quantitative framework. The mean boundary error is used as a criterion. They used several mosaic images with 2 regions. Ojala [19] conducted a comparative study of performance of texture features. This study includes gray level difference, Laws', center-symmetric covariance, local binary patterns, and complementary feature pairs. They measure the performance of each feature by nearest neighbor classification. Experiments with random samples with size 32 × 32 or 16 × 16 showed that local binary patterns performed best. Pichler *et al.* [20] introduced pyramidal and tree-structured wavelet transform and compared with adaptive Gabor filtering. Results are evaluated by comparing the segmented images with those obtained with multichannel Gabor filters. Four different wavelet transformation techniques in texture segmentation are evaluated in [10]. They selected ten wavelet filters to determine how well textures are distinguished. Fuzzy c-means clustering is used to obtain a segmentation based on computed texture features.

0-7695-0149-4/99 $10.00 © 1999 IEEE

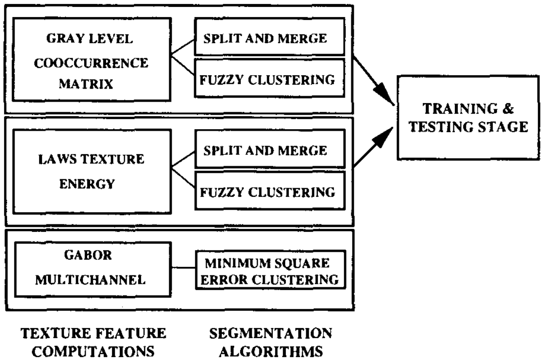


Figure 1: A scheme used in this study

2 Methods and Materials

2.1 Gray level co-occurrence matrix

Gray level co-occurrence matrix (GLCM) was introduced by Haralick [12]. A co-occurrence matrix describes how often one gray level appears in a specified spatial relationship to another gray level. The entry at *(i, j)* of the GLCM indicates the number of occurrences of the pair of gray levels i and j which are a distance d apart along a given direction θ. The values of d and θ are parameters for constructing the GLCM.

2.2 Laws' texture energy

Laws' texture energy (LAWS) combines predetermined one-dimensional kernels into various convolution masks [18]. The output image of the convolution process is considered as an "energy image", followed by a texture energy transformation in which each pixel at the center of a local window $(l(i,j))$ is replaced by the mean of absolute value in the filter window $(f(i,j))$ as follows:
$s(i,j) = \frac{1}{(2\times n+1)^2}\sum_{k=i-n}^{i+n}\sum_{l=j-n}^{j+n}|f(k,l) - l(i,j)|$,
where n is size of mask.

2.3 Gabor multi-channel filtering

In this study, we use multi-channel filtering with Gabor functions (GABOR) as proposed by Jain and Farrokhnia [17]. The set of channels was designed with even-symmetric Gabor filters, cosine part only, as an impulse response, defined as:
$h(x,y) = \exp\{-\frac{1}{2}[\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}]\}\cos(2\pi u_0 x)$,
where u_0 is radial frequency of filter and σ_x,σ_y are the space constants of the Gaussian envelope along the x and y axes, respectively [17]. Feature images are obtained by submitting each selected filtered image to a nonlinear transformation and computing a measure of energy around each pixel. Then, the average absolute deviation from the mean in small overlapping windows is computed.

2.4 Quad-tree split-and-merge

Split-and-merge (SPMG) takes the entire feature image as an initial input and successively divides into four sub-regions based on the degree of homogeneity of feature values in sub-regions. Once a splitting has been accomplished, a merging process is then performed with more restricted threshold values (T_F). Pairs of regions which are spatial neighbors are merged if following test is satisfied:
$|Max(F_{R_1}, F_{R_2}, F_{R_3}, F_{R_4}) - Min(F_{R_1}, F_{R_2}, F_{R_3}, F_{R_4})| < T_F$,
where F_{R_i} indicates feature value of region R_i.

2.5 Clustering

Clustering labels regions in an image by partitioning a given feature set into compact and well-separated clusters in feature space. It does not necessarily use spatial information.

Fuzzy c-means clustering : Fuzzy c-means clustering (FCM) minimizes the objective function J_m with respect to fuzzy membership grade $\mu_{i,j}$ and the center of cluster V_i [1]. The objective function is defined as:
$J_m = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{i,j})^m d^2(X_j, V_i)$,
where $d^2(X_j, V_i) = (X_j, V_i)^T A(X_j, V_i)$. A can be any positive definite $p \times p$ matrix, where p is the dimension of the feature vectors X_i, c is number of clusters and n number of data points.

Square error clustering : This clustering method (CLST) uses the square-error criterion to achieve a set of partitions in the feature space [17]. After initial set-up for K clusters followed by iteration of measurement of minimum square-error between the data points where K is upper bound of clusters, the partition process halts when there is no change in the clusters.

Validity measurement : We incorporate the clustering validity measure for each method to make a comparison of unsupervised texture segmentation. Estimating the true number of clusters is critical since the number of clusters directly affects the segmentation performance, as shown in the Table 3 for the training set. This study uses a validity measure from [22]. Xie *et al.* measure the validity metric using the degree of *separation* and *compactness* of clusters as fuzzy validity criteria. For CLST, this study uses a modified Hubert (MH) index as in [17]. It computes the Euclidean distance between each data point of each cluster. Both validity measurement methods here search for significant change in the range of minimum and maximum number of clusters.

2.6 Performance Metric

Misclassified pixel based comparison: Previous texture segmentation studies have generally evaluated performance in terms of misclassified pixel rate [8, 10, 4, 15, 20, 17].

Region based comparison: We adapted the methodology used in Hoover *et.al* to evaluate range segmentation algorithms [13]. It can also be applied to evaluate performance of texture segmentation algorithms. When measuring segmented regions against ground truth regions, the primary criteria of the method in [13] is the degree of overlap between machine segmented and ground-truthed regions. Results fall into one of five categories: *correctly detected*, *over-segmented*, *under-segmented*, *missed* and *noise* depending on degree of overlap regions between machine segmentation (MS) and ground truth (GT) regions. Details can be found in [13]. Performance metrics are computed as a $X_{region}Rate = \frac{n(X_{region})}{T_{region}}$, where X can be *correctly detected*, *over-segmented*, *under-segmented* or *missed*. $n(X_{region})$ is total number of such instances and T_{region} is total number of regions in an image.

2.7 Image Data Set

Evaluation with ground truth information is an important aspect in texture segmentation. In order to achieve a more rigorous comparison, we have constructed a set of images with a large number and variety of textures. Images may contain outdoor or indoor scenes including a variety of textured objects and backgrounds, as well as non-textured areas. Ground truth is more subjective for real scene images than for mosaic images. However, we feel that there is no other option for evaluating the accuracy of vision algorithms which must ultimately deal with real images. There are two stages in the ground truth process. First, homogeneous textured boundaries in all images have been selected by 9 observers in order to decide which images give good agreement on the ground truth.

Second, the agreed-upon GT regions in the 35 images were carefully traced out at a pixel level. After the ground truth was completed, we selected 10 images to measure variations due to the drawing of the GT outlines. The person (A) who generated pixel-level ground truth repeated the same process at different times to measure within-subject (A:A) variation and another person (B) performed the process for the same images to obtain across-subject (A:B) variation. The average percent overlap in region area between within-subject GTs is 97.01%, and the percent overlap between across-subject GTs is 95.06%. Thus 95 % may represent an ideal agreement between MS and GT.

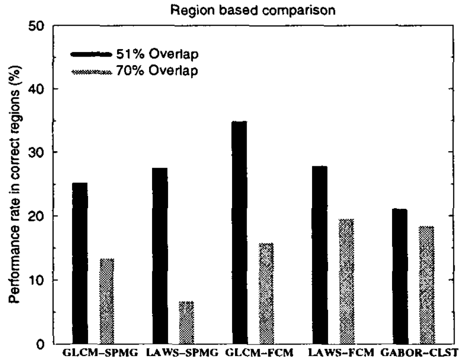


Figure 2: Training results on 10 real scene images

3 Experimental Results

3.1 Training

Each texture feature extraction technique requires a set of parameters to characterize textures. Often parameters in different texture feature algorithms are not clearly comparable. The list of the parameters required in GLCM, LAWS and GABOR is shown in Table 1. Both GLCM and LAWS require a set of parameters to train whereas GABOR does not (Figure 1). *Spatial resolution* and *sampling window* in GLCM and LAWS were trained. The selection of features for feature-based methods in this study has been made based on previous literature. The popularly used features in GLCM [6, 7] are energy, entropy, homogeneity, contrast and correlation and in LAWS are edge-edge, edge-level, edge-shape, shape-level and ripple-ripple [11, 4]. Twenty filters formed with four orientations and five radial frequencies were selected for GABOR-CLST [17]. There are two important aspects that should be considered in order to achieve fair comparison between texture feature computation algorithms. First, the level of achieved information or the level of importance of required parameters *to* each algorithm should be similar. Second, the number of features or usefulness of features trained should be similar across algorithms.

Segmentation results of five algorithms on two training images are presented in Figure 3. The texture segmentation performance in training was not close to the ideal, as shown in Table 3 (pixel classification rate) and Figure 2 (correct region rate). In the evaluation of pixel classification rate, GABOR-CLST performed best, then LAWS-FCM, GLCM-FCM, LAWS-SPMG and GLCM-SPMG (Table 3). In the correct region rate evaluation, LAWS-FCM performed best, GABOR-CLST performed second, then GLCM-FCM, GLCM-SPMG followed by LAWS-SPMG when 70% overlap is used as a correct criterion (Figure 2).

Table 1: Parameters in GLCM, LAWS and GABOR

Texture Algorithm	Orientation	Spatial resolution	Sampling window size	Radial frequency size
GLCM	Average of $0°, 45°, 90°, 135°$	1,3,5	8,16,32	not applic.
LAWS	not applic.	3,5,7	8,16,32	not applic.
GABOR	$0°, 45°, 90°, 135°$	not applic.	not applic.	$4N$, $8N$, $16N$ $32N$,$64N$
	$N = \sqrt{2} \times cycles/image\_width.$			

Table 2: Performance on five mosaic images

Texture Algorithm	Classification pixel rate (%)	Correct region rate (%)
GLCM-FCM	67.10 % (62.98 %)	39.80 % (15.79 %)
LAWS-FCM	73.82 % (56.78 %)	46.41 % (19.59 %)
GABOR-CLST	85.04 % (57.96 %)	96.00 % (18.42 %)
Rates in parentheses are ones in using real scene images.		
Correct region rate was measured at 70% overlap.		

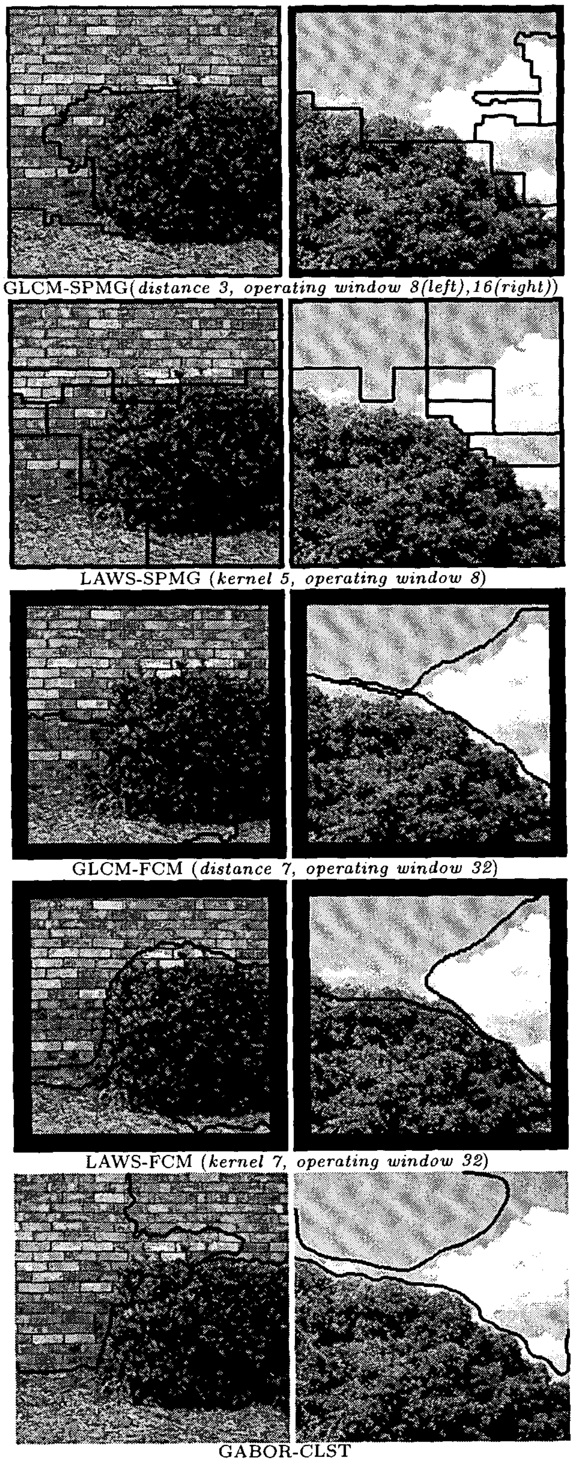
LAWS-FCM has a greater number of 70% or higher overlapped regions than GABOR-CLST. Texture features with SPMG showed relatively poor performance in both evaluation criteria.

3.2 Experiment with Mosaic Images

One of the major motivations for this study is to investigate the importance of using real images rather than mosaic images. In order to investigate whether mosaic images are a sufficient way to evaluate the performance of texture segmentation, we applied three higher-ranking texture segmentation algorithms (GABOR-CLST, GLCM-FCM, LAWS-FCM) to 5 popularly-used mosaic images. One of the images with results by the three algorithms is shown in Figure 4. Testing on mosaic images showed substantially higher performance than for real scene images (Table 2). Therefore, use of mosaic images may lead one to expect unrealistically high performance on real scene images.

3.3 Testing

Testing is performed with the parameters selected from the 70% overlap measure in the training. In a 25-image test set with trained parameters, GABOR-CLST performed best, LAWS-FCM second, and LAWS-SPMG worst in both pixel classification rate and correct region rate as shown in Table 3. Graphs for the region-based performance metrics appear in Figure 5. The highest correct pixel classification rate, 71.15 % was achieved with the GABOR-CLST approach with the number of clusters selected manually. Performance dropped substantially when the number of clusters was selected automatically using a validity metric (71.15% and 57.96%). Choice of segmentation algorithm plays an important role



GLCM-SPMG(*distance 3, operating window 8(left),16(right)*)

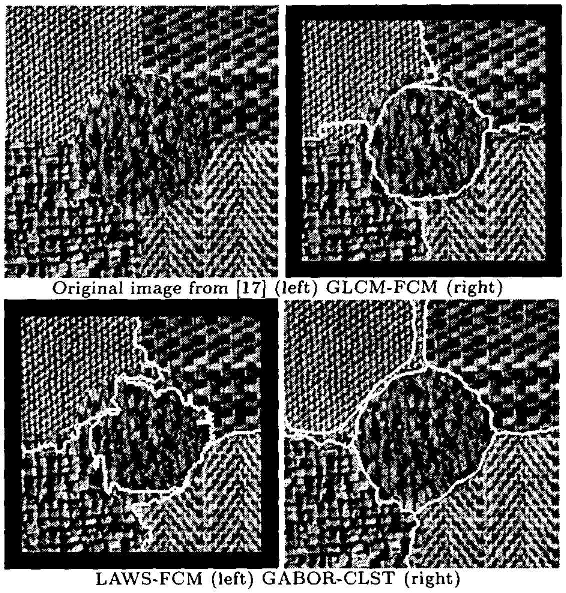
LAWS-SPMG (*kernel 5, operating window 8*)

GLCM-FCM (*distance 7, operating window 32*)

LAWS-FCM (*kernel 7, operating window 32*)

GABOR-CLST

Figure 3: Examples of machine segmented images



Original image from [17] (left) GLCM-FCM (right)

LAWS-FCM (left) GABOR-CLST (right)

Figure 4: Machine segmentations on a mosaic image

Table 3: Performance results on training / testing set

Algorithms	Pixel classification rate(%)		Correct region based rate(%)		Average runtime (min.)
	Train	Test	Train	Test	
GLCM-SPMG	38.8	38.9	13.5	8.2	28.5
LAWS-SPMG	41.2	32.1	6.7	3.7	29.1
GLCM-FCM	63.0	36.0	15.8	14.1	25.5
GLCM-FCM 1	69.1				
LAWS-FCM	56.8	48.8	19.6	20.0	31.4
LAWS-FCM 1	67.6				
GABOR-CLST	58.0	67.3	18.4	38.8	54.4
GABOR-CLST 1	71.2				
1. is methods without using validity measure.					
Correct region rate measured at 70% overlap.					
Average time is measured in Sun Ultra Sparc.					

in texture segmentation. LAWS interacts very sensitively with segmentation algorithms whereas GLCM does not show much difference with the choice of segmentation algorithm. FCM works better with LAWS than with GLCM. SPMG works better with GLCM than with LAWS. Texture segmentation with SPMG performed worse than with clustering based methods. Therefore, we might infer that the choice of segmentation algorithm indeed makes a difference in a texture segmentation. The misclassified pixel rate evaluation measures the average overlapping pixels with highest overlapped region to its ground truth region without counting over-segmentation or under-segmentation.

One important aspect in region segmentation evaluation is to examine how a region is segmented by a certain segmentation algorithm instead of measuring only misclassified pixels or correct regions. Over-segmentation rate indicates presence of regions falsely detected along the correct boundary or/and regions confused in homogeneous regions. Graph 2 in Figure 7 shows that GABOR-CLST and GLCM-FCM have a higher over-segmentation rate than others. The fixed size of *operating window* would not perform robustly for a common texture with varying resolution. Multiple sizes of *operating window* might be a partial solution for this problem. Most algorithms in this study show low under-segmentation rate (Graph 3 in Figure 5). GABOR-CLST under-segmentation represents a rate of merged neighbor regions due to the lack of discriminating power.

4 Conclusions

Evaluation of texture segmentation performance is not an easy task since each technique has its own way to characterize textures. GABOR-CLST approach with validity metric offers the best performance in this study. GABOR more readily incorporates multiresolution information than GLCM and LAWS. Since GABOR computes more filters to capture texture characteristics than GLCM and LAWS, it segments better than the other algorithms considered.

Testing on mosaic images showed unrealistically high performance. Two possible reasons can be given for this. First, in a mosaic image, region boundaries are smooth so that performance of region segmentation gets better. Second, very distinctive textures may be positioned as neighbor regions, which would lead to high performance. Clustering methods require a range for the number of clusters or regions. In order to develop an unsupervised texture segmentation with clustering methods, one should carefully test for a validity metric. SPMG seems suitable only as a preliminary or coarse-level segmentation [4].

Our evaluation method is fully automated and easily applicable to other proposed texture segmentation algorithms. It gives more relevant and detailed result information than simple metrics such as percent correct pixel classification. The relevant tools will be available to the public.

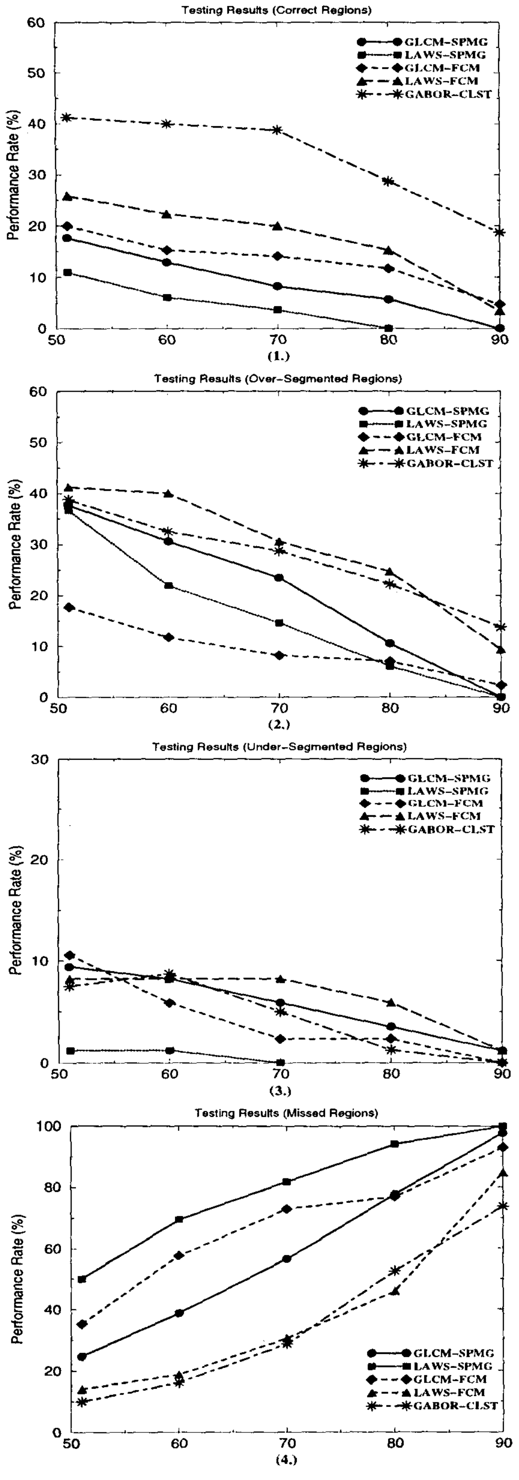
References

[1] J.C. Bezdek. Pattern recognition with fuzzy objective function algorithms. Plenum, New York, 1981.

[2] A.C. Bovik, M. Clark, and W.S. Geisler. Multichannel texture analysis using localized spatial filters. *PAMI*, 12:55–73, 1990.

[3] P. Brodatz. *Textures – A Photographic Album for Artist and Designers*. New York: Dover, 1965.

[4] J.L. Chen and A.Kundu. Unsupervised texture segmentation using multichannel decomposition and hidden markov models. *Transactions on Image Processing*, 4(5):603–619, May 1995.


Testing Results (Correct Regions)

Testing Results (Over-Segmented Regions)

Testing Results (Under-Segmented Regions)

Testing Results (Missed Regions)

Figure 5: Testing results in 25 real scene images

[5] P.C. Chen and T. Pavlidis. Segmentation by texture using a co-occurrence matrix and split-and-merge algorithm. *CGIP*, 10:172-182, 1979.

[6] R.W. Conners, M.M. Trivedi, and C.A. Harlow. "Segmentation of a high-resolution urban scene using texture operators". In *CVGIP*, volume 25, pages 273-310, 1984.

[7] L.S. Davis, M. Clearman, and J.K. Aggarwal. An empirical evaluation of generalized co-occurrence matrices. *PAMI*, 3:214-221, 1981.

[8] J.M.H. du Buf, M. Kardan, and M. Spann. Texture feature performance for image segmentation. *Pattern Recognition*, 23(3/4):291-309, 1990.

[9] D. Dunn, W.E. Higgins, and J. Wakeley. Texture segmentation using 2d gabor elementary functions. *PAMI*, 16:130-149, 1994.

[10] Fatemi-Ghomi, P.L. Palmer, and M. Petrou. Performance evaluation of texture segmentation algorithms based on wavelets. *Workshop on Performance Characterization of Vision Algorithms in Cambridge, England.*, April 1996.

[11] X. Gong and N.K. Huang. Texture segmentation using iterative estimate of energy states. *9th ICPR, Rome Italy*, pages 51-55, 1988.

[12] R.M. Haralick, K. Shanmugam, and I.H. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 237-247, 1973.

[13] A. Hoover, G. Baptiste, X. Jiang, P. Flynn, H. Bunke, D.B. Goldgof, K. Bowyer, D.W. Eggert, A. Fitzgibbon, and R.B. Fisher. An experimental comparison of range image segmentation algorithms. *PAMI*, 18(7):673 - 689, 1996.

[14] J.Y. Hsiao and A.A. Sawchuk. Unsupervised textured image segmentation using feature smoothing and probabilistic relaxation techniques. *CVGIP*, 48:1-21, 1989.

[15] C. Jacquelin, A. Aurengo, and G. Hejblum. Evolving descriptors for texture segmentation. *Pattern Recognition*, 30(7):1069-1079, 1997.

[16] A.K. Jain. Texture analysis. *Handbook of Pattern Recognition and Computer Vision*, pages 235-276, July. 1989.

[17] A.K. Jain and F. Farrokhnia. Unsupervised texture segmentation using gabor filters. *Pattern Recognition*, 24(12):1167-1186, 1991.

[18] K.I. Laws. *Textured Image Segmentation*. PhD thesis, Univ. Southern California, 1980.

[19] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51-59, 1996.

[20] O. Pichler, A. Teuner, and B.J. Hosticka. A comparison of texture feature extraction using adaptive gabor filtering, pyramidal and tree structured wavelet transforms. *Pattern Recognition*, 29(5):733-742, 1996.

[21] T.R. Reed and J.M.H. du Buf. A review of recent texture segmentation and feature extraction techniques. *CVGIP:Image Understanding*, 57(3):359-372, May 1993.

[22] L.X. Xie and G. Beni. A validity measure for fuzzy clustering. *PAMI*, 13:841-847, Aug. 1991.

[23] J. You and H.A. Cohen. Classification and segmentation of rotated and scaled textured images using texture "tuned" masks. *Pattern Recognition*, 26(2):245-258, 1993.