Comparing classification metrics for labeling segmented remote sensing images

Philippe Maillard
Universidade Federal de Minas Gerais
Department of Cartography
Belo Horizonte, Minas Gerais, Brazil
pmaillar@engmail.uwaterloo.ca

David A. Clausi
University of Waterloo
Department of System Design, Engineering
Waterloo, Ontario, Canada
dclausi@engmail.uwaterloo.ca

Abstract

Image segmentation and labelling are the two conceptual operations in image classification. As the remote sensing community uses more powerful segmentation procedures with spatial constraint, new possibilities can be explored for labelling. Instead of assigning a label to a single observation (pixel), whole segments of image are labelled at once implying the use of multivariate samples rather than pixel vectors. This approach to image classification also offers new possibilities for using a priori *information about the classes such as existing maps or object signature libraries. The present paper addresses the two issues. First a labelling scheme is presented that gathers evidence about the classes from incomplete* a priori *information using a "cognitive reasoning" approach. Then, five different metrics are compared for the label assignment and are combined through a voting scheme. The results show that very different results can be obtained depending on the metric chosen. The metric combination through voting, being a suboptimal approach does not necessarily provide the best results but could be a safe alternative to choosing only one metric.*

1. Introduction

In computer vision, the classification of any digital image can be separated into two distinct conceptual operations: segmentation and labelling. Segmentation consists in identifying objects, labelling in naming them. In the remote sensing community, most classification problems are approached using point-dependant algorithms where each pixel is independently compared to feature space clusters (*e.g.* spectral bands and texture features) for its classification. Point-dependant classification can be considered a special case of segmentation with no spatial constraint: each pixel is considered an "object". More powerful segmentation schemes impose some spatial constraint on the identification of objects and create spatial clusters of connected pixels: segments. These segments can then be labelled.

Segment labelling is slightly different from point-dependant classification in the sense that instead of assigning a label to an observation (pixel) by comparing it to known classes (training), entire segments having variable populations of pixels are being considered. This broadens the number of metrics that can be used but also brings up other issues such as distribution parameters. The possibility of a wider range of classification metrics also makes it possible to use a voting scheme that can take advantage of all the classification methods.

The examples used in the present paper are derived from a sea ice mapping application using synthetic aperture radar (SAR) data. For a number of reasons including geometrical, electrical and noise factors [19], [18], [22], the direct classification of SAR images into sea ice types is a very difficult operation that involves not only SAR backscattering values but also information on incident angle, texture, context and even higher visual cues like shape (*e.g.* leads) and size (*e.g.* floe size). In the approach advocated here, a rough map (containing only homogeneous region outlines and the number and type of sea ice classes without their actual location) produced by an ice analyst is fed into a system that takes advantage of this information to specify the number (a classical problem in image segmentation) and types of classes to be found in the "homogeneous" regions drawn by the analysts. This system named MAGSISC (Map-Guided Sea Ice Segmentation and Classification) is being developed with the objective of being inserted in Canadian Ice Service (CIS) operations to provide added value products such as pixel-based ice maps. The complete system is described elsewhere [17] and this paper outlines the labelling solution implemented.

The present paper addresses two main issues:

1. It describes a segment labelling scheme based on evidence gathering and multi-metrics classification by voting.

2. It assesses the performance of five classification metrics: 1) Mahanalobis distance, 2) Fisher criterion, 3) Chi-square test, 4) Kolmogorov-Smirnov test and 5) Student's t-test.

2. Image segmentation

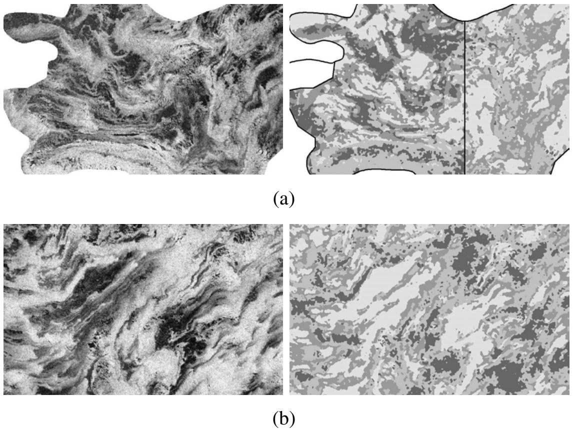
2.1. Pre-processing

Apart from geographical registration and eight-bit quantization, the pre-processing essentially consists in generating texture features to complement SAR intensities in the feature space.

The grey-level co-occurrence matrix (GLCM) [11] is a commonly used texture method in remote sensing applications and has proven very powerful for sea ice using SAR data [1], [24], [3]. More specifically, the GLCM approach can accurately capture the textural characteristics of sea-ice using SAR data [2] and has often proven superior to other popular methods in a classification context [4], [16]. Based on previous results [12], the two preferred GLCM statistics are contrast and entropy. A window size of nine by nine pixels, sixty-four (64) grey level and rotationally invariant features (GLCM statistics were averaged over four directions: 0, 45, 90 and 135 degrees) have proven to generate acceptable results for sea ice classification [3] and were adopted here.

2.2. Segmentation

The segmentation is performed using a Markov random field (MRF) model. MRFs can provide solutions for many image analysis problems such as image restoration, texture description and image segmentation [8], [14]. MRF models inherently describe spatial context: the local spatial interaction among neighboring pixels. This is most appropriate since neighboring pixels are generally not statistically independent but are linked by spatial correlation. Furthermore, MRFs can effectively combine the relative importance of the pixel being considered and its neighborhood. Numerous MRF-based segmentation methods have been developed [5], [26], [10], [20], [6]. Considering SAR images, MRF models have already shown to provide an appropriate representation of SAR images given their variance (due to speckle) and texture [7], [20], [13], [15]. In the MAGSISC system, the "Modified adaptive Markov random field segmentation" (MAMSEG) [6] was adopted because of its good performance with SAR data and sea ice [6]. MAMSEG is innovative because it does not fix *a priori* the relative weight of the central pixel and its neighborhood but rather lets it vary with each iteration in the simulated annealing solution. Examples of segmentation results for a few regions are shown in Fig. 1.



(a)

(b)

Figure 1. Original SAR image samples (left hand side) and segmentation result (right hand side). (a) and (b) have both three ice types and open water. (a) had to be segmented in two parts because of its size; note the consistency of the segmentation across the artificial border. The images are 700 x 430 pixels and were scaled down for display purposes.

2.3. Segment description

The segmentation process is guided in terms of region (large areas defined by the ice analysts) and number of sea ice classes (and could include open water) but yields no clues as to which segment correspond to which class. This is referred to as the labelling process. For labelling to take place, segments must be described in a way that they can be compared between themselves and with known classes. Unlike pixel-based classification which compares a single observation (a m-feature vector) to a population sample, segment-based classification requires the comparison of population samples between themselves. A separate component of MAGSISC extracts a number of statistics to describe each segment and stores this information in a spreadsheet file. The following statistics are stored as vectors or matrices: mean (μ), standard deviation (σ), population (N), covariance matrix (Σ) and histogram (H).

3. Labeling and classification metrics

The image labelling is based on a "learning" approach which seeks to provide high-level semantic concepts from low-level visual features [27]. In our case, the examples from which the system can learn are provided a priori from a rough interpretation stating what ice types (labels) are found in a limited region. Unlike the pure "learning by examples" approach, these training data are not necessarily "ready-to-use" but has to be deduced. The low-level features are based on image texture (GLCM) and SAR intensity.

3.1. The cognitive reasoning approach

The segmentation is guided by pre-defined large regions (on the order of 1000-5000 km^2 or 1-5 Megapixels) through the maps produced by CIS's ice analysts and in some cases these regions might have only one or two classes. In the case of a unique class, these regions can be treated as training areas. In other cases a single class might be common to two of these regions and the segments corresponding to the shared class can be deduced easily through some means of comparison. These possibilities lead to developing an approach that can take advantage of these deductions and has been named "classification by cognitive reasoning". The evidence was classified in three types: first-, second- and third-degree.

First-degree evidence falls into two categories: 1) the egg code region contains only one class or 2) the egg code region contains several classes and all but one have already been solved and assigned. In either case, the association is straight-forward and no additional information is needed to solve the association.

Second-degree evidence is characterized by the fact that although all or some classes of a region have previously been solved (in other regions), the program still has to find which set of associations is the most likely. For a total of n_c classes, there are $n_c!$ permutations of matching each segment of a particular region to one of the n_c classes. The objective is then to determine which permutation is more likely according to some metric.

In third-degree evidence, reasoning is based on the fact that while comparing two egg code regions, although no association was previously solved, if only one class is common to both egg code regions (intersection), then one can deduce which is more likely by calculating a distance metric between all pairwise possibilities. The optimal result is retained as the correct association.

Each time the evidence leads to the labelling of a segment, that segment becomes part of the class sample and so statistics for that class must be updated. Hence, for the segment i having n_i samples and found to belong to class j (with n_j samples), the mean of j is updated using the following equation:

$$\mu_{ij} = \frac{\mu_i n_i + \mu_j n_j}{n_i + n_j} \tag{1}$$

Similarly, the covariance of the merged samples Σ_{ij} can be updated using the following closed form update equation ([9], p.119):

$$\Sigma_{ij} = \frac{S_{ij}}{(n_i + n_j)} \tag{2}$$

where

$$S_{ij} = S_i + S_j + (\mu_i - \mu_j) \cdot (\mu_i - \mu_j)^t \frac{n_i n_j}{n_i + n_j} \tag{3}$$

is the scatter matrix defined by $\Sigma_i n_i$. The histograms are also merged by simple summation:

$$H_{ij} = H_i + H_j \tag{4}$$

3.2. Classification metrics

Numerous methods have been developed to compare two populations or samples. Five of these methods are tested here. A multi-classifier approach is also proposed as a means to take advantage of all five methods. The five methods are:

1. Mahanalobis distance (MD), [9]
2. Fisher criterion, [9]
3. Chi-square (goodness-of-fit) test ($\chi 2$), [21]
4. Kolmogorov-Smirnov test (KS), [21]
5. Student's t-test (t-test), [21]

The Mahanalobis distance ([9], p. 36) is used to measure the distance between a single observation $\mathbf{x}$ and a class distribution (μ, Σ). Although it uses the mean and covariance matrix, unlike the maximum likelihood it does not assume a Gaussian distribution. The traditional form of MD follows:

$$MD = (\mathbf{x} - \mu)^t \Sigma^{-1} (\mathbf{x} - \mu) \tag{5}$$

To compensate for the fact that only the spread of one sample is considered, the MD has been replaced by the maximum MD (MMD) which is defined as:

$$\begin{aligned} MMD = \quad & max[(\mu_i - \mu_j)^t \Sigma_j^{-1} (\mu_i - \mu_j), \\ & (\mu_j - \mu_i)^t \Sigma_i^{-1} (\mu_j - \mu_i)] \end{aligned}$$

The Fisher criterion (FC, [9], p. 117) projects the feature space onto a line that best separates between two classes:

$$J(\omega) = \frac{\omega^t S_B \omega}{\omega^t S_W \omega} \tag{6}$$

where S_B and S_W are the between- and within-class scatter matrices respectively and $\omega = \Sigma_{12}^{-1}(\mu_1 - \mu_2)$. A Fisher criterion is calculated for each class pair and the segment is assigned to the class with the smallest J. The Fisher criterion offers the advantage of taking into consideration the spread of both distributions being compared.

The chi-square "goodness-of-fit" test ($\chi 2$, [21], p. 620) is usually employed to compare a sample's distribution to a theoretical distribution such as the Gaussian function. In the present case, it is used to compare two independent sample distributions. The chi-square statistic is defined by:

$$\chi^2 = \sum_{k=1}^{N_b} \frac{(H_{i_k} - H_{j_k})^2}{H_{i_k} + H_{j_k}} \tag{7}$$

where H_i and H_j are the two sample histograms and N_b represent the number of bins in the histograms. The chi-square probability function is an incomplete gamma function defined by χ^2 and ν degrees of freedom and corresponds to the probability that the sum of square differences between H_i and H_j are attributable to the samples' variance (H_0). The largest the probability, the more likely the two samples belong to the same population.

The Kolmogorov-Smirnov (K-S, [21], p. 623) test computes the probability of two distributions belonging to the same population based on the distance between their cumulative distributions:

$$D = \max_{-\infty<x<\infty} |S_{N_i}(x) - S_{N_j}(x)| \tag{8}$$

The probability that D is significant (reject H_0 that the two distributions are the same) is then computed using the following sum:

$$Q_{KS}(\lambda) = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2\lambda^2} \tag{9}$$

where $\lambda = D(\sqrt{N_e}+0.12+0.11/\sqrt{N_e})$ and $N_e = \frac{N_i N_j}{N_i+N_j}$. Thus, the larger Q_{KS}, the less likely the two distributions belong to the same population.

Finally, Student's t-test (t, [21], p. 616) verifies if the difference between two sample's mean is significant to a certain degree of confidence or if the difference can be attributed to the sample's pooled variance. The t-test assumes a Gaussian distribution but this assumption can usually be relaxed if N is large ([23], p. 410). For samples that could have different variances, the sample's pooled variance is defined as:

$$SD = \sqrt{\frac{S_{x_i}^2}{n_i} + \frac{S_{x_j}^2}{n_j}} \tag{10}$$

and the t statistic as:

$$t = \frac{\mu_i - \mu_j}{SD} \tag{11}$$

and the number of degrees of freedom (ν) can be approximated for large samples by:

$$\nu = n_i + n_j - 2 \tag{12}$$

The t statistic is compared with Student's distribution $A(t|v)$ being an incomplete beta function and returns the probability that the samples are drawn from the same population (H_0).

3.3. Combining the classifier: the most likely solution?

With the possibility of using various classifiers arises the possibility of combining them by stating that they are complimentary and no single approach is better in all situations [25]. When all methods are probabilistic, their probabilities can be combined by simple averaging. Since in the present case includes two distance metrics and three probabilistic metrics, a simple "voting" scheme was adopted and the candidate (label combination) that receives the highest number of "votes" wins the election. In this approach, the ambiguous results (no or more than one winner) are not classified.

This is a sub-optimal approach and a candidate with only two very strong votes (very high probability) would lose to one with three "barely" votes (low probability but higher than for the other candidates) even though it should maybe win. It is hoped that the solutions of the different metrics do not differ to the point of provoking such situation but it remains nonetheless a possibility.

3.4. Experimental design for evaluating the results

Evaluating results for most remote sensing application consists of counting hits and misses while comparing the classified image and a ground truth sample. Here however, only the labelling is being evaluated and not the segmentation (the latter has already been evaluated and approved elsewhere [6], [17]). Furthermore, the labels are known for each region but not their exact assignment. So instead of comparing individual segments, whole "label assignment" solutions have to be evaluated and compared. The goal is to evaluate each of the five metrics, compare them and evaluate their combination as well (combining classifiers by voting). To do so, four tests were designed which are summarized below and detailed in section 4.

1. Level of agreement. This first test is intended to find out how the five metrics differ from one another. The test consists in simply counting for each metric, how many times it matches each of the other metrics in a contingency table. A chi-square test was also performed on the contingency table to check if the level of agreement between each metric pair is significant.

2. Nearest segment. Since the previous test was performed on "solutions" that includes a combination of labels and not each segment individually, a test was designed to measure the degree of consistency between the metrics. This test consists in finding, for each segment taken as a sample, the nearest segment in the rest of the data. Since all ice types are found more than once, all segments should find another segment of the

same class. The number of agreements was also compared in a contingency table.

3. Level of agreement with “winning label”. This test is similar to the first one but this time the individual “solutions” are compared with the “winning” solution in order to determine if any metric can act as a gauge metric that usually agrees with the most voted solution.

4. Level of agreement with ground truth. This last test is meant to evaluate the metrics individually and the combined metrics as a successful approach to the labelling problem.

These tests were performed on two sets of data from two different SAR images from RADARSAT-1 totalling over fifty regions segmented in two to four classes. The images cover the Western (10 March 2002) and Eastern (13 March 2000) part of the Gulf of Saint-Lawrence (including the Canadian East coast) respectively.

4. Results and discussion

4.1. Level of agreement

The level of agreement between the five metrics is calculated by summing the number of times that one “solution” (for a set of labels in a single region) of a metric exactly matches the solution of another. The regions had to be separated by the number of classes they contain because the probability of a match (p_m) decreases dramatically with increasing number of classes (n_c): $p_m = 1/n_c!$. Table 1 shows for each region set the number of matches in a contingency table. For each metric pair, a contingency table was constructed for observed and expected matches (according to chance alone) in the fashion presented in Table 2. A chi-square test was applied to see if the observed matches were larger than the expected by-chance matches: $H_0 : P_{observed} = P_{expected}$ and $H_1 : P_{observed} > P_{expected}$.

The results of the chi-square tests are presented in Table 3. They clearly show that, although most metrics have a significant level of agreement, no metric pair match perfectly. Based on these results, three rough groups can be formed: 1) MMD and FC have a high level of agreement (35 in 62) and form a first group; 2) $\chi 2$ and KS are both distribution comparison methods and generally agree (35 in 62); and 3) the t-test only marginally agrees with the other methods and can form a group of its own. The $\chi 2$ metric is the only one for which H_0 is always rejected: its level of agreement with all the other metrics cannot be attributed to chance alone (at a 95% level of confidence).

2 classes (2 possibilities)	MMD	FC	$\chi 2$	KS	t-test
MMD	24*	22	19	13	19
FC		24*	19	16	16
$\chi 2$			24*	17	20
KS				24*	17
t-test					24*

3 classes (6 possibilities)	MMD	FC	$\chi 2$	KS	t-test
MMD	20*	8	7	6	5
FC		20*	6	7	1
$\chi 2$			20*	18	3
KS				20*	3
t-test					20*

4 classes (24 possibilities)	MMD	FC	$\chi 2$	KS	t-test
MMD	18*	5	1	1	1
FC		18*	0	2	0
$\chi 2$			18*	0	1
KS				18*	2
t-test					18*

Table 1. Number of pairwise agreements between the five metrics for the two- (top), three- (middle) and four-classes (bottom) regions (* represents the total number of regions).

	Observed			Expected		
	=	≠	Σ	=	≠	Σ
2 classes	22	2	24	12	12	24
3 classes	8	12	20	3.33	16.66	20
4 classes	5	13	18	0.75	17.25	18
Σ	35	27	62	16.08	45.92	62

Table 2. Example of contingency tables for observed (left) and expected by-chance (right) matches between the Mahanalobis distance (MMD) and Fisher criterion (FC) metrics.

4.2. Nearest segment

The “nearest segment” test was performed in order to test the level of consistency of the metrics between themselves in an independent fashion. Here, it is not the “solutions” that are tested but each segment independently of the others. The two test images together contain a total of 776 (340 and 446 for the first and second image respectively). Table 4 shows the compiled results for this test.

The results are similar for both images and show that the FC and MMD tend to agree more on their solutions than the other three metrics. Still, their number of agreements

Probability of H_0				
	FC	$\chi2$	KS	t-test
MMD	0.000	0.001	0.245	0.010
FC		0.003	0.008	0.067
$\chi2$			0.000	0.005
KS				0.041

Accepted hypothesis (95%)				
	FC	$\chi2$	KS	t-test
MMD	H_1	H_1	H_0	H_1
FC		H_1	H_1	H_0
$\chi2$			H_1	H_1
KS				H_1

Table 3. Results of the chi-square tests for pairwise agreement between the five metrics for all 62 regions.

Image	MMD	FC	$\chi2$	KS	t-test	Maximum
1	152	156	63	132	77	340
2	164	174	42	127	97	436
Σ	316	330	105	259	174	776

Table 4. Number of times each metric found the same nearest segment as other metrics.

are well below the maximum possible attainable (less than half) which shows that no single metric is universally acceptable and that any single metric likely to yield different results from the other. The $\chi2$ metric has the lowest level of agreement which can seem contradictory with the results in 4 where it was found to be the sole metric with significant level of agreement with all the other metrics. A plausible explanation is that when considering that the $\chi2$ metric never really disagrees with all other metrics, it is possible that it also never quite agrees since a strong level of agreement with one particular metric would signify stronger disagreement with others. However, insufficient evidence is available to fully support such explanation.

4.3. Level of agreement with "winning label"

This test is meant to assess if one of the metrics generally agrees with the label that won the majority of votes. Table 5 is a compilation of the percentage of times that each metric agrees with the winning label. Again, the results have been divided into three to match the number of classes in the regions. The regions with no winner (each of the five votes were given to a different solution) or with ambiguous results (equality of votes between two solutions) were left out of the calculations.

The results appear to indicate that with a total score of 84% the $\chi2$ tends to agree better with the winning combinations of labels than any other single metric. The $\chi2$ is followed by the KS (72.7%), the MMD and FC with both 68.2% leaving the t-test with the worse level of agreement. This ranking is also consistent for the two- and three-class regions. The four-class regions however, show much lower levels of agreement and a different ranking. It should be noted that in their case, the change of a perfect match is only $\frac{1}{24}$ and that there are only eight valid cases (unambiguous). One conclusion that can be drawn from these results is that if one should choose only one metric, the $\chi2$ is apparently the preferred choice.

N-class	MMD	FC	$\chi2$	KS	t-test
(N-cases)	%	%	%	%	%
2 cl.(24)	83.3	83.3	95.8	75	87.5
3 cl.(12)	50.0	50.0	100	100	25.0
4 cl.(8)	50.0	50.0	25.0	25.0	12.5
Total (44)	68.2	68.2	84.1	72.7	56.8

Table 5. Number of times (in percentage) each metric agrees with the solution having the majority of votes. Note that null or ambiguous results were left out.

4.4. Level of agreement with ground truth

The best possible test for any classification scheme is to see how successfully it performs when compared against ground truth. Unfortunately, ground truth is seldom available and is substituted by a sample. In the case of sea ice, ground truth is not practically accessible and here it was substituted by an interpreted version. It should be noted that each segment could not be interpreted individually but in combination with the other labels in the region. Through this interpretation process, various flaws and doubtful solutions were detected. Human operators only approximate the exact answer; statistical validation requires exact answers but these are not available. In some regions, the CIS ice map appeared to have "missed" a class which has caused problems in the segmentation and consequently, in the classification as well. These "doubtful" regions were left out of the compilation.

Table 6 shows the results (in percentage) of comparing the solutions of each individual metric and the winning solution with the interpreted ground truth. Because of the uncertainty that surrounds the interpretation, many regions were left out of the validation data and only 39 of the 62 regions were used. The four-classes regions were particularly affected by the selection process and only five of the 18 were considered. The scores are roughly between 62%

and 80% but the ground truth cannot be considered completely reliable and these scores might vary with better validation data.

Unlike the previous tests, these results seem to indicate that the distance metrics, MMD and FC are more reliable than the probabilistic methods. However, if the three type of regions are considered separately, the results are not as clear. When considering only the three- and four-class regions (because the results for the two-class regions can easily be attributed to chance with a probability of 0.5), all scores fall below the 60% mark but the MMD still performs better followed by the "winner" of the voting approach.

Although the data is too sparse to provide a definite explanation as to the ranking of the metrics, three facts can shed some light some of the reasons.

1. Stationarity: because of the variations in the incident angle in the SAR images, the various classes of ice and open water cannot be expected to have a stationary behavior and can have shifts in their distributions. Frequency comparison methods like $\chi 2$ and KS are particularly sensitive to such shifts and can become "unstable".
2. Feature dependency: the three probabilistic metrics consider the features (SAR intensity and texture features) to be independent which is not a very reliable assumption. Only the MMD and FC metrics truly create a feature space.
3. Variance: neither the MMD and FC metrics are affected by differences in variance between the samples whereas some normalization is necessary for the three probabilistic methods.

Additional testing is required to obtain a better explanation as to the reasons behind the ranking of the different metrics and, more importantly, explain the mechanisms that make them provide different answers to a single problem.

5. Conclusions

This paper presents an innovative method for labelling segmented images using evidence gathering from *a priori* information supplied by an analyst in the form of large, partially interpreted regions containing a pre-determined number of classes. Instead of using training data, the algorithm accumulates evidence about one or two classes and then builds the statistics about the other classes by deduction while continuously updating these statistics as the segments are gradually assigned a label. Five different metrics are used and compared for assigning a label to the segments: two non-parametric distances and three probabilistic methods. The five metrics were also combined in a voting approach to produce a sixth possible solution.

Although the five metrics were found to have a significant degree of agreement between themselves, they yielded very different results. The $\chi 2$ metric was found to generally agree better with the most common ("winning") solution but when compared against interpreted ground truth, the Mahanalobis distance and the Fisher criterion were found to perform better. The voting scheme performed well and could be an alternative to selecting a single metric.

For other applications, the *a priori* information can be replaced by training data or class signature libraries. Work is currently underway to test this methodology on other remote sensing data and with ground truth data for training and validation.

Acknowledgements: The authors are thankful to GEOIDE (http://www.geoide.ulaval.ca), CRYSYS (http://www.crysys.ca) and CIS (http://www.cis.ec.gc.ca) for their their financial support. CIS also provided the SAR sea ice images (all RADARSAT images ©CSA - http://www.space/gc/ca).

References

[1] D. G. Barber and E. F. LeDrew. SAR sea ice discrimination using texture statistics: A multivariate approach. *Photogrammetric Engineering and Remote Sensing*, 57(4):385–395, 1991.

[2] D. A. Clausi. Comparison and fusion of co-occurrence, Gabor and MRF texture features for classification of SAR sea-ice imagery. *Atmosphere-Ocean*, 39(3):183–194, 2000.

[3] D. A. Clausi. An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian J. Remote Sensing*, 28(1):1–18, 2002.

[4] D. A. Clausi and B. Yue. Comparing co-occurrence probabilities and Markov random fields for texture analysis. *IEEE Trans. on Geoscience and Remote Sensing*, 42(1):215–228, 2004.

[5] F. S. Cohen and D. B. Cooper. Simple parallel hierarchical and relaxation algorithms for segmenting noncausal Markovian random fields. *IEEE Trans. Pattern Anal. Machine Intell.*, 9(2):195–219, 1987.

[6] H. Deng and D. A. Clausi. Unsupervised segmentation of synthetic aperture radar sea ice imagery using a novel Markov random field models. *IEEE Trans. on Geoscience and Remote Sensing*, accepted for publication, 2004.

[7] Y. Dong, B. Forester, and A. Milne. Comparison of radar image segmentation by gaussian- and gamma-markov random fields models. *International Journal of Remote Sensing*, 24(4):711–722, 2004.

[8] R. Dubes and A. Jain. Random field models in image analysis. *Journal of Applied Statistics*, 16:131–164, 1993.

[9] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2nd edition, 2001.

[10] D. Geman, S. Geman, C. Graffigne, and P. Dong. Boundary detection by constrained optimization. *IEEE Trans. Pattern Anal. Machine Intell.*, 12(7):609–628, 1990.

number of classes (number of regions)	MMD %	FC %	$\chi 2$ %	KS %	t-test %	win %
2 classes(24)	83.3	91.7	87.5	75.0	79.2	83.3
3 classes(10)	80.0	40.0	40.0	50.0	30.0	40.0
4 classes(5)	60.0	40.0	20.0	20.0	60.0	80.0
Total (39)	**79.5**	**71.8**	**66.7**	**61.5**	**64.1**	**71.8**
3 and 4 classes	59.0	46.2	43.6	46.2	46.2	51.3

Table 6. Number of times (in percentage) each metric agrees with the (interpreted) ground truth.

[11] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Trans. Sys. Man Cybern*, 3:610–621, 1973.

[12] R. Jobanputra and D. Clausi. Texture analysis using Gaussian weighted grey level co-occurrence probabilities. In *Proceedings of 1st Canadian Conference on Computer and Robot Vision, May, London, Ontario, Canada*, May 17-19 2004.

[13] S. Krishnamachari and R. Chellappa. Multiresolution Gauss-Markov random field models for texture segmentation. *IEEE Trans. Image Processing*, 6(2):251–267, 1997.

[14] S. Z. Li. *Markov Random Field Modeling in Computer Vision*. New York: Springer-Verlag, 2001.

[15] G. Liu, H. Xiong, and S. Huang. Study on segmentation and interpretation of multilook polarimetric sar images. *International Journal of Remote Sensing*, 21(8):1675–1691, 2004.

[16] P. Maillard. Comparing texture analysis methods through classification. *Photogrammetric Engineering and Remote Sensing*, 69(4):357–367, 2003.

[17] P. Maillard and D. A. Clausi. Map-guided sea ice segmentation and classification using SAR imagery and a MRF segmentation scheme. *Submitted to IEEE Trans. on Geoscience and Remote Sensing*.

[18] M. Mäkynen, A. Manninen, M. Similä, J. Karvonen, and M. Hallikainen. Incidence angle dependence of the statistical properties of C-band HH-polarization backscattering signatures of the baltic eea ice. *IEEE Trans. on Geoscience and Remote Sensing*, 40(12):2593–2605, 2002.

[19] R. Massom. *Satellite Remote Sensing of Polar Regions*. Lewis Publishers, Boca Raton, FLA, 1991.

[20] D. K. Panjwani and G. Healey. Markov random field models for unsupervised segmentation of textured color images. *IEEE Trans. Pattern Anal. Machine Intell.*, 17(10):939–954, 1995.

[21] W. Press, S. Teukolsky, W. Vettering, and B. Flannery. *Numerical Recipes in C: the Art of Scientific Computing*. Cambridge University Press, 2nd edition, 1992.

[22] R. Raney. *Manual of Remote Sensing*, volume 2, chapter Radar fundamentals: technical perspective, pages 9–130. John Wiley and Sons, New York, NY, 3rd edition, 1998.

[23] B. Scherrer. *Biostatistique*. Gaëtan Morin Éditeur, 1st edition, 1984.

[24] L. K. Soh and C. Tsatsoulis. Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. *IEEE Trans. Geoscience and Remote Sensing*, 37(2):780–795, 1999.

[25] B. Tso and P. Mather. *Classification Methods for Remotely Sensed Data*. Taylor and Francis, London, 2001.

[26] C. S. Won and H. Derin. Unsupervised segmentation of noisy and textured images using Markov random fields. *CVGIP: Graphical Models and Image Processing*, 54(4):308–328, 1992.

[27] Y. Xu, E. Saber, and A. Tekalp. Dynamic learning from multiple examples for semantic object segmentation and search. *Computer Vision and Image Understanding*, pages 334–353, 2004.