Combining Local and Global Features for Image Segmentation Using Iterative Classification and Region Merging

Qiyao Yu, David A. Clausi
Systems Design Engineering, University of Waterloo
Waterloo, ON, Canada
{*q2yu,dclausi*}*@engmail.uwaterloo.ca*

Abstract

In MRF based unsupervised segmentation, the MRF model parameters are typically estimated globally. Those global statistics sometimes are far from accurate for local areas if the image is highly non-stationary, and hence will generate false boundaries. The problem cannot be solved if local statistics are not considered. This work incorporates the local feature of edge strength in the MRF energy function, and segmentation is obtained by reducing the energy function using iterative classification and region merging.

Keywords: Markov random field (MRF), region merging, segmentation, classification, edge detection

1. Introduction

Markov random field (MRF) allows the modelling of the joint probability distribution of image pixels in terms of the local spatial interactions, and hence has found great applications in image segmentation. MRF can not only be used to extract texture features [4], but can also model spatial context constraint under the Bayesian framework [2][6][17][11]. The latter is referred as the region model, and combined with the feature distribution model forms a 2-layer model for the segmentation problem.

For unsupervised applications, the 2-layer MRF model parameters are not known in advance, and hence need to be estimated from the same image, mostly using expectation-maximization (E-M) methods [5]. A main drawback of this kind of approach is that the model parameters estimated from global statistics may not fit for local regions, resulting in incorrect identification of boundaries during segmentation. To overcome this drawback, here the edge strength is introduced as a term into the MRF energy.

Finding the global minimum is expected to give the correct solution, and optimization techniques such as iterated conditional mode (ICM) [2] and simulated annealing [7] are typically used to solve the MRF formulation. It has been found, however, that the solution can become trapped into the local minima due to the impossibility of implementing an ideal optimization method. The non-optimal result is intolerable if the configuration corresponding to the local minima is far from the desired. This situation is likely to occur for highly non-stationary images, of which the corresponding MRF energy function is highly non-convex with the introduction of the edge strength term. Annealing performed on single pixels is not efficient for such situations and methods having larger perturbation units instead of single pixels are needed.

The region merging method proposed in this paper, called ICRM (iterative classification and region merging), is one such approach. It keeps merging the segments if the merging result gives a reduced MRF energy. Classification is then performed on those regions instead of single pixels. Therefore, the perturbation unit can be much larger.

There has been a significant amount of work on hybrid region and edge segmentation [14][12][15][10] and region merging [9][13][15]. For example, in [10] a deformable model (balloon) is coupled with the regional MRF model, but the method is not designed for simultaneous segmentation of multiple regions. Tu and Song [15] use an edge map to guide the region splitting and merging step in finding minima of the model energy. Their model has not integrated the edge strength directly into the energy function and is quite complex. Sakar et al. [13] also merge similar regions to reduce MRF energy. But the edge process in their work is based on neighboring regional statistics not on edge strength. The most similar to the proposed work is [3], where the optimization is performed by graph theory technique which actually is also region merging.

The organization of the paper is as follows. Section 2 gives an introduction to the traditional 2-layer MRF segmentation model, from which the new model is derived and presented in section 3. Section 4 introduces the ICRM method. Section 5 gives the experiments and discussions.

Proceedings of the Second Canadian Conference on Computer and Robot Vision (CRV'05)
0-7695-2319-6/05 $ 20.00 **IEEE**

Summary and future works comprise section 6.

2. Segmentation Model

2.1. Markov Random Field (MRF)

An MRF [8][11] assumes that each site is independent of sites outside its defined neighborhood given the neighborhood sites. The neighborhood, with appropriate size or order, makes computation of the dependence among sites manageable, especially for images which often have hundreds of thousands of pixels. The Hammersley-Clifford theorem [1] shows that a random field X is an *MRF* if and only if X is a Gibbs distribution. That is, for any configuration $\mathbf{x}$ of the state space T of random field X, the joint probability is

$$P(X=\mathbf{x}) = \frac{1}{Z}\exp\{-E(\mathbf{x})\} = \frac{1}{Z}\exp\left\{-\sum_{c\in C} V_c(\mathbf{x})\right\} \tag{1}$$

where C is the set of cliques which are defined as the sets of mutually neighboring sites, $V_c(\mathbf{x})$ is the energy of configuration $\mathbf{x}$ on clique c, $E(\mathbf{x})$ is the total energy of configuration $\mathbf{x}$, and the normalizing constant $Z = \sum_{\mathbf{y}\in T}\exp\{-E(\mathbf{y})\}$ is called the partition function.

2.2. Two-Layer MRF Model

Let Y denote the observed data (i.e., the original image) and X the desired segmentation. A 2-layer model is constructed from the Bayes' rule [6][11]:

$$P(X=\mathbf{x}|Y=\mathbf{y}) = \frac{p(Y=\mathbf{y}|X=\mathbf{x})P(X=\mathbf{x})}{p(Y=\mathbf{y})} \tag{2}$$

where $P(X=\mathbf{x}|Y=\mathbf{y})$ is the posterior conditional probability of the segmentation given the observed data, $p(Y=\mathbf{y}|X=\mathbf{x})$ is the conditional probability distribution of the observation conditioned on the segmentation, $P(X=\mathbf{x})$ is the prior probability of the segmentation, and $p(Y=\mathbf{y})$ is the probability distribution of the observation.

Maximization of the posterior $P(X=\mathbf{x}|Y=\mathbf{y})$ gives a segmentation solution, and since $p(Y=\mathbf{y})$ is a constant it is equivalent to maximizing $p(Y=\mathbf{y}|X=\mathbf{x})P(X=\mathbf{x})$. $p(Y=\mathbf{y}|X=\mathbf{x})$ describes the gray level, texture or any other features while $P(X=\mathbf{x})$ can incorporate spatial context constraint. Two separate models, a feature distribution model and a region model, are needed to obtain the analytical expressions for $p(Y=\mathbf{y}|X=\mathbf{x})$ and $P(X=\mathbf{x})$ respectively.

For the region model, a multi-level logistic model (MLL) [6] is typically used, whose clique potential is defined as:

$$V_c(X) = \begin{cases} \beta_c & \text{if all sites on } c \text{ equal} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The term $p(Y=\mathbf{y}|X=\mathbf{x})$ is typically assumed to be Gaussian at each site. Maximizing the posterior is equivalent to minimizing the objective function in (4)

$$\arg\min_{\Omega_1,\ldots,\Omega_n} \sum_{i=1}^{n}\sum_{s\in\Omega_i}\left\{\log(\sigma_i) + \frac{[Y(s)-\mu_i]^2}{2\sigma_i^2} - \beta U_i(s)\right\} \tag{4}$$

where n is the number of classes, $\Omega_1,\ldots,\Omega_n$ are the obtained classes, μ_i is the mean gray level of class Ω_i, σ_i is the standard deviation of class Ω_i, and $U_i(s)$ is the number of neighbors of s that belongs to class Ω_i.

3. Proposed Model

3.1. Objective Function

To estimate the mean and standard deviation for Equation (4), most approaches assume that the number of classes is known and solve the Gaussian mixture by E-M based method [5]. The estimation uses global statistics and does not describe local regions well if the image is highly non-stationary. Incorporating local statistics is thus necessary.

Consider Equation (4) from another point-of-view. Minimization of (4) is equivalent to the minimization of:

$$\arg\min_{\Omega_1,\ldots,\Omega_n} \sum_{i=1}^{n}\sum_{s\in\Omega_i}\left\{\log(\sigma_i) + \frac{[Y(s)-\mu_i]^2}{2\sigma_i^2} + \beta\bar{U}_i(s)\right\} \tag{5}$$

where $\bar{U}_i(s)$ is the number of neighbors of s that does not belong to class Ω_i. That means the term $\beta\bar{U}_i(s)$ is not zero only when the site s is at the boundary. Therefore, Equation (4) becomes

$$\arg\min_{\Omega_1,\ldots,\Omega_n} \left\{\sum_{i=1}^{n}\sum_{s\in\Omega_i}\left\{\log(\sigma_i) + \frac{[Y(s)-\mu_i]^2}{2\sigma_i^2}\right\} + \beta\sum_{i=1}^{n}\sum_{s\in\partial\Omega_i}\bar{U}_i(s)\right\} \tag{6}$$

$\sum_{i=1}^{n}\sum_{s\in\partial\Omega_i}\bar{U}_i(s)$ is approximately proportional to the length of the obtained boundary. Therefore, the role of the region model term is actually penalizing the existence of boundary by its population.

Instead of penalizing equally for all boundary pixels, greater penalty can be applied to weak edge pixels and a lesser penalty to strong edge pixels, so that local statistic such as edge strength can be incorporated. Therefore, we can replace the penalty term with some monotonically decreasing function of edge strength $g(|\nabla(\cdot)|)$, where $|\nabla(\cdot)|$ is the gradient magnitude. The proposed objective function is therefore

$$\arg\min_{\Omega_1,\ldots,\Omega_n} \left\{\sum_{i=1}^{n}\sum_{s\in\Omega_i}\left\{\log(\sigma_i) + \frac{[Y(s)-\mu_i]^2}{2\sigma_i^2}\right\} + \beta\sum_{i=1}^{n}\sum_{s\in\partial\Omega_i} g(|\nabla(Y(s))|)\right\} \tag{7}$$

Our method of penalizing differently on the edge strength has the shortcoming of bias on certain classes since boundaries between some classes can be generally weaker than those of others. However, including edge strength is advantageous in describing local behaviors, and we will also show next that the bias problem can be alleviated by properly manipulating the edge penalty function $g(\cdot)$.

3.2. Choosing Edge Penalty Function

The edge penalty function $g(\cdot)$ can be any monotonically decreasing function, so that the greater the edge strength, the smaller the penalty. Suppose the gradient $|\nabla(Y)|$ has been normalized. Then, the penalty function can be formulated as:

$$g(|\nabla(Y)|) = e^{-(|\nabla(Y)|/K)^2}$$

The parameter K (Fig. 1) defines how fast the edge penalty decays with the increase of edge strength. As K increases, the penalty difference between weak and strong edges decreases. When K approaches infinity, all edge penalties are equally 1.

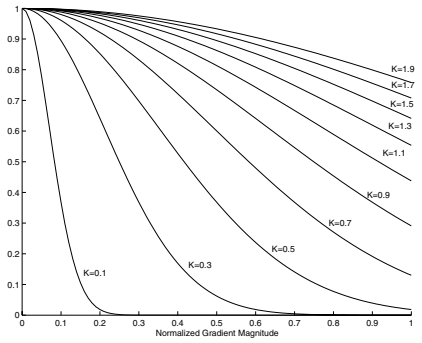


Figure 1. Edge penalty functions

Therefore, the parameter K of the edge penalty function $g(\cdot)$ is allowed to increase during the segmentation, details of which are described in Section 4.2. The local feature of edge strength is hence utilized, but the bias is significantly reduced.

4. Segmentation Algorithm

Finding the global minimum of (7) is expected to give the desired segmentation. Optimization techniques such as simulated annealing can be used. However, due to the impossibility of implementing an ideal optimization method, the solution searching may be trapped into some local minima. If the image is highly non-stationary, the local minima are easily located far away from the desired solution and convergence to such local minima is intolerable.

Consider the following example. Fig. 2(a) is an image that has two distinct regions, one relatively bright and one relatively dark. Both clusters have Gaussian distributed intensities. However, they are highly non-stationary. Fig. 2(b) gives the segmentation result using Gaussian mixture model whose parameters are estimated by E-M. It is quite obvious that the globally estimated statistics for the clusters are not sufficient for the segmentation due to the non-stationarity of the image. Therefore, edge information is incorporated as in Equation (7), and simulated annealing is used to give the segmentation result in Fig. 2(c). Interestingly however, there is no visible improvement. The region of the global minimum of (7) is too small and surrounded by energetically unfavorable configurations (a phenomena referred to as a "golf course" [7]), thus making the algorithm easily trapped into the local minima.

(a) An example image.

(b) By Gaussian mixture model.

(c) By proposed model, simulated annealing used.

(d) By proposed model, ICRM used.

Figure 2. An example highly non-stationary image.

Aiming at suppressing the unfavorable configurations surrounding the global minimum and increasing the relative area of the global minimum region, the perturbation unit is changed from a single pixel to a homogenous region. The idea is intuitively simple. Pixels or regions that have similar local statistics are grouped together first, and then labelling is performed on the obtained regions using global statistics. Back to the example of Fig. 2, we are expecting

that we can group pixels belonging to the left part into one region and those belonging to the right part into another region before we label them. This leads to the combined region merging and classification approach.

4.1. Merging and Classification

4.1.1. Region Merging For the optimization purpose, the region merging process in our approach aims at reducing the overall energy function. However, we need to modify the energy function (7) a little for the convenience of the merging process. The formula is still the same, but the notation of class is replaced by segment (closed region). That is, n is the number of segments, $\Omega_1, \ldots, \Omega_n$ are the segments, μ_i is the mean gray level of segment Ω_i, σ_i is the standard deviation of segment Ω_i, and $U_i(s)$ is the number of neighbors of s that belongs to segment Ω_i. During each merging operation, the overall energy for the resulting configuration is computed and compared with that of the unmerged ones. If there is a decrease, the merging is justified.

The overall energy need not be calculated in order to make a decision for merging a pair of segments. There is only a need to compute the difference between the sum energy of the two neighboring segments tested and the energy of the obtained merged one. Suppose we are investigating segment Ω_i and segment Ω_j, and let $\Omega_k = \Omega_i \bigcup \Omega_j$. The energy difference δE is:

$$\sum_{s\in\Omega_k} \log(\sigma_k) - \sum_{s\in\Omega_i} \log(\sigma_i) - \sum_{s\in\Omega_j} \log(\sigma_j) - 2\beta \sum_{s\in\partial\Omega_i \bigcap \partial\Omega_j} g(|\nabla(Y(s))|) \quad (8)$$

If (8) gives a negative value, Ω_i and Ω_j can be merged. If (8) gives a positive value, Ω_i and Ω_j are not merged.

A sequential merging order needs to be defined since the merging cannot be performed simultaneously. A natural method is to find the two segments whose energy decrease most if merged, and merge them first.

The merging begins on an initial configuration obtained by a watershed algorithm [16], and is iterated until the energy cannot be reduced any more. This will usually generate an over-segmentation result, which is further labelled (classified).

4.1.2. Classification Aiming at finding the minimum of (7), we define for each segment Ω_i its individual energy to be

$$E(Y(\Omega_i)|k) = \sum_{s\in\Omega_i} \left\{ \log(\sigma_k) + \frac{[Y(s)-\mu_k]^2}{2\sigma_k^2} \right\} + \beta \sum_{s\in\partial\Omega_i \bigcap \partial\Omega_j} g(|\nabla(Y(s))|) \quad (9)$$

where k is the class label, Ω_j is a neighboring segment to Ω_i and does not belong to class k. Simulated annealing is then used for the classification.

4.2. Iterative Classification and Merging (ICRM)

As we have mentioned in section 3.2, using a single penalty function will produce bias. Different classes that have fuzzy boundaries between could be merged, and class that has high intra-class variation or noise could be split. Both could result in poor performance of the system, even though in the later situation the split parts could probably be assigned the same class label in the subsequent classification.

Therefore, the combined merging and classification process is iterative. The diagram for the overall algorithm is shown in Fig. 3. At each iteration, there is a classification process followed by merging. The edge penalty function parameter K increases while the iteration continues until there is no merging occurred or a maximum number of iterations have been reached.

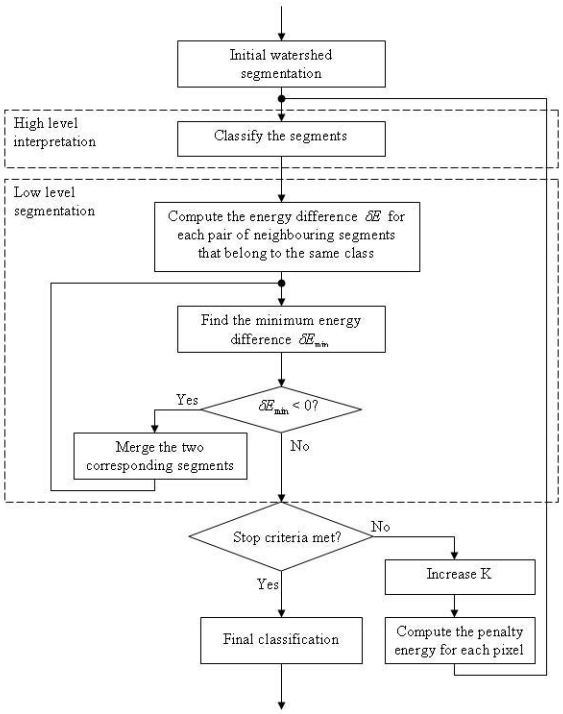


Figure 3. Edge penalty functions

Different from the merging-classification approach in section 4.1, classification is placed ahead of merging during each iteration. Two segments are not mergable if they

belong to different classes, the purpose of which is to suppress the merging between segments of different classes that have weak boundaries in between.

It is noteworthy that the high level interpretation part is not necessarily to be restricted to the classification in section 4.1.2. Nor does it have to be aiming at minimizing the objective function of (7). We can replace it with any reasonable interpretation incorporating any features such as texture and shape. Therefore this iterative classification and merging method provides a way of bilateral communication between low level segmentation and high level interpretation.

4.3. Computational Complexity

The computational complexity of ICRM is reasonably low compared to that of the traditional model, even though this algorithm has the classification and merging process that is iterative. In this method the classification is performed on segments rather than over image pixels, which considerably reduced the computation time. And moreover, even if the initial number of segments is large, it often reduced dramatically after the first round of merging. As to the merging process, the corresponding computation is trivial since efficient way of estimating segment parameters (mean and variance) can be used, and the pixels are scanned only once at the initial stage.

The algorithm is implemented in C++. Running on a Pentium 4 1.8G PC with memory 512M, it takes about 5-7 minutes to segment images of 1252×873 pixels.

5. Experiments and Discussions

Before beginning the experiment, the parameter β needs to be determined. Images that have a large boundary population (i.e. having many small irregular segments) should be given a small β, while those having few boundary pixels (i.e. consisting of only a few large regular segments) should be given a large β. Table 1 gives the βs we set experimentally, which is determined based on the percentage of boundary pixels in the image. Obviously, the percentage of boundary pixels are not known in advance. But as the iterative classification and merging progresses, this percentage number will gradually stabilize.

The first experiment is on the example image of Fig. 2. Fig. 2(d) shows an accurate segmentation using the proposed edge weighted model and ICRM optimization method. Parameters of the Gaussian mixture are globally estimated by the E-M method. Class distributions are all assumed to be Gaussian for these experiments.

The next three experiments are performed on three SAR images acquired by Radarsat ScanSAR C-band mode with the resolution of $100m \times 100m$. The traditional model of (4) and the edge weighted model of (7) are compared. Simulate annealing is used for the traditional model and ICRM is used for the edge weighted model.

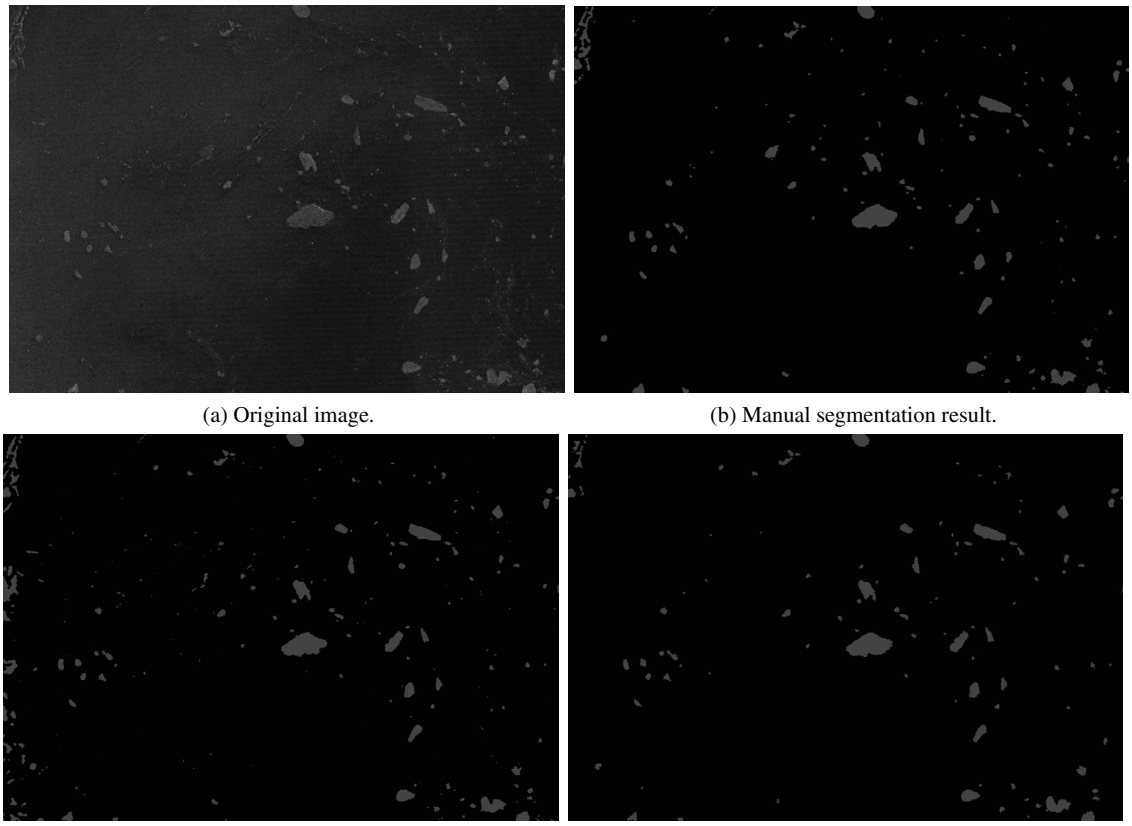
Boundary pixels (%)	β
≤ 0.02	23
≤ 0.03	13
≤ 0.04	8
≤ 0.05	5
≤ 0.06	3
≤ 0.07	2
≤ 0.08	1
else	0.5

Table 1. β corresponding to different percentage of boundary pixels

Fig. 4(a) was extracted from a SAR image of Baffin Bay on June 24, 1998. The extracted region is highly non-stationary. The gray level of the water increases gradually from the center part of the image to the left, reaching some values close to those of ice floes. This is a manifestation of the change of backscatter as a function of antenna's incidence angle with respect to the earth's surface. The traditional model result, given by Fig. 4(c), incorrectly classified some water regions in the left part of the image to ice. The result of the proposed method shown in (d), however, is much better in this aspect. On the other hand, the proposed method has failed in preserving some tiny ice segments, which is probably caused by the smoothing operation in computing the gradient for the initial watershed segmentation. Overall, the quality of the segmentation produced by the ICRM algorithm is an improvement over the traditional model.

Experiments on two other images, Fig. 5(a) and 6(a), provide similar results. Fig. 5(a) was extracted from the same image of Fig. 4(a), and Fig. 6(a) was captured over the Gulf of St. Lawrence on Feb 19, 1997. On the right side of Fig. 5(a) there is a body of rough water. Some of the rough water pixels (indicated by the white rectangular) are classified as ice at the center in the right of Fig. 5(c), while the corresponding part in (d) is quite clean. Similar results are noted in Fig. 6.

To show the advantage of ICRM over simulated annealing in finding the global minimum for the edge weighted model, we compute the energy values corresponding to the segmentation results of the two methods. For edge weighted model with simulated annealing, the energy values for the three SAR images are 3.14285×10^6, 4.00532×10^6, and 4.71885×10^6 respectively. For the same model with ICRM result, the corresponding values are 3.06702×10^6, 3.82866×10^6, and 4.64495×10^6 respectively. Clearly,



(a) Original image. (b) Manual segmentation result.

(c) By traditional model and simulated annealing (d) By edge weighted model and ICRM.

Figure 4. Segmentation of a SAR image of Baffin Bay

ICRM provides a better solution.

6. Summary and Future work

In this paper, an advanced image segmentation algorithm called ICRM (iterative classification and region merge) has been designed, implemented, and tested successfully. The essence of the proposed method lies in two aspects: edge information is incorporated, and neighboring similar segments are constrained to be of the same class. It should be mentioned, however, that the underlying MAP criterion of the ICRM algorithm does not consider the size of the discrepancy between the obtained configuration and the truth configuration. Therefore, merging first and then classifying could possibly cause more pixels to be labelled incorrectly and result in a more discrepant configuration than that of the traditional model which is pixel-based. However, the utilization of edge information has greatly decreased this probability, and more information such as texture and shape can easily be introduced.

Those information can be utilized in the high level interpretation part of the system. Multiple separate interpretations can be performed using different kinds of features, the result of each determines which two segments are allowed to merge and which are not. Incorporating those information will be the future work.

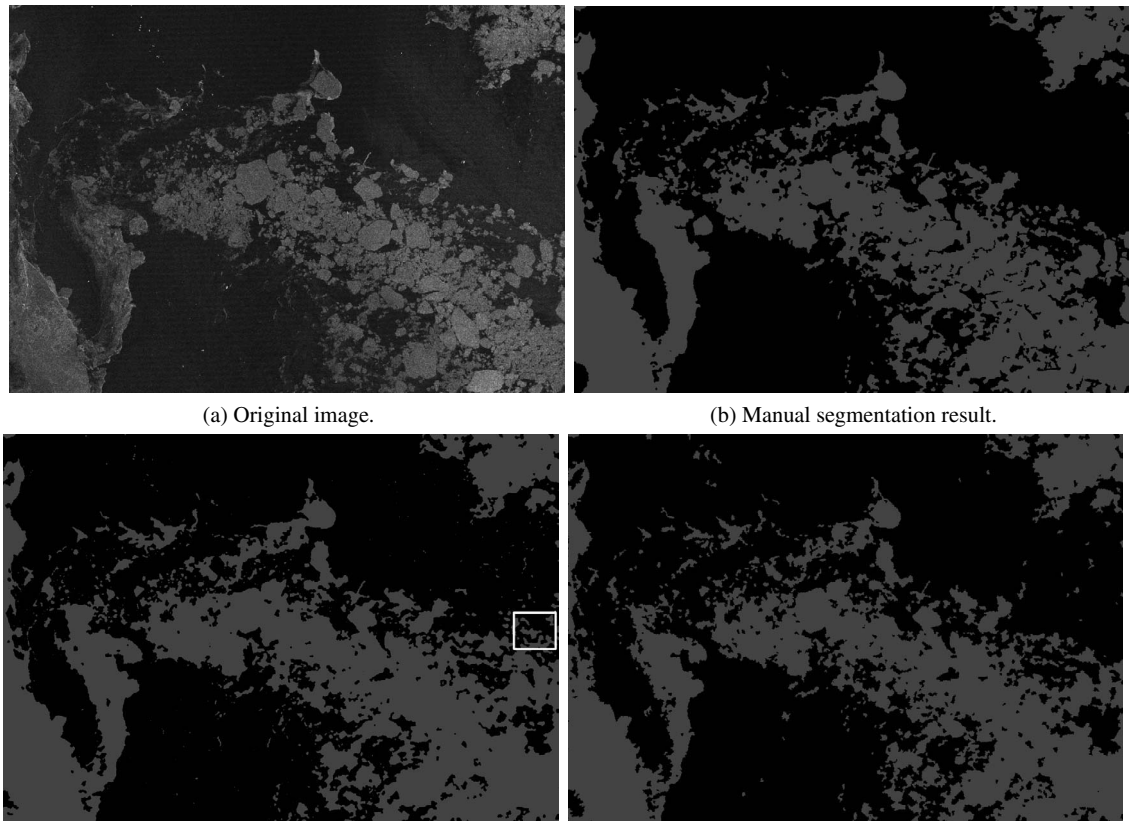
Acknowledgements

RADARSAT images are copyright the Canadian Space Agency (CSA). This work is supported by the NSERC Networks of Centres of Excellence (NCE) called GEOIDE (Geomatics for Informed Decisions) (http://www.geoide.ulaval.ca/) as well as CRYSYS (CRYospheric SYStem in Canada) (http://www.crysys.ca).

References

[1] J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

[2] J. Besag. On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society, Series B*, 48:259–302,



(a) Original image. (b) Manual segmentation result.

(c) By traditional model and simulated annealing. (d) By edge weighted model and ICRM.

Figure 5. Segmentation of a SAR image of Baffin Bay

1986.

[3] Y. Boykov, O. Veksler, and R. Zabih. Markov Random Fields with Efficient Approximations. In *Computer Vision and Pattern Recognition, 1998*, pages 648–655, 23-25 June 1998.

[4] R. Chellappa. Two-dimensional Discrete Gaussian Markov Random Field Models for Image Processing. *Progress in Pattern Recognition*, pages 79–112, 1985.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[6] H. Derin and H. Elliott. Modeling and Segmentation of Nosiy and Textured Images Using Gibbs Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):39–55, 1987.

[7] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2001.

[8] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

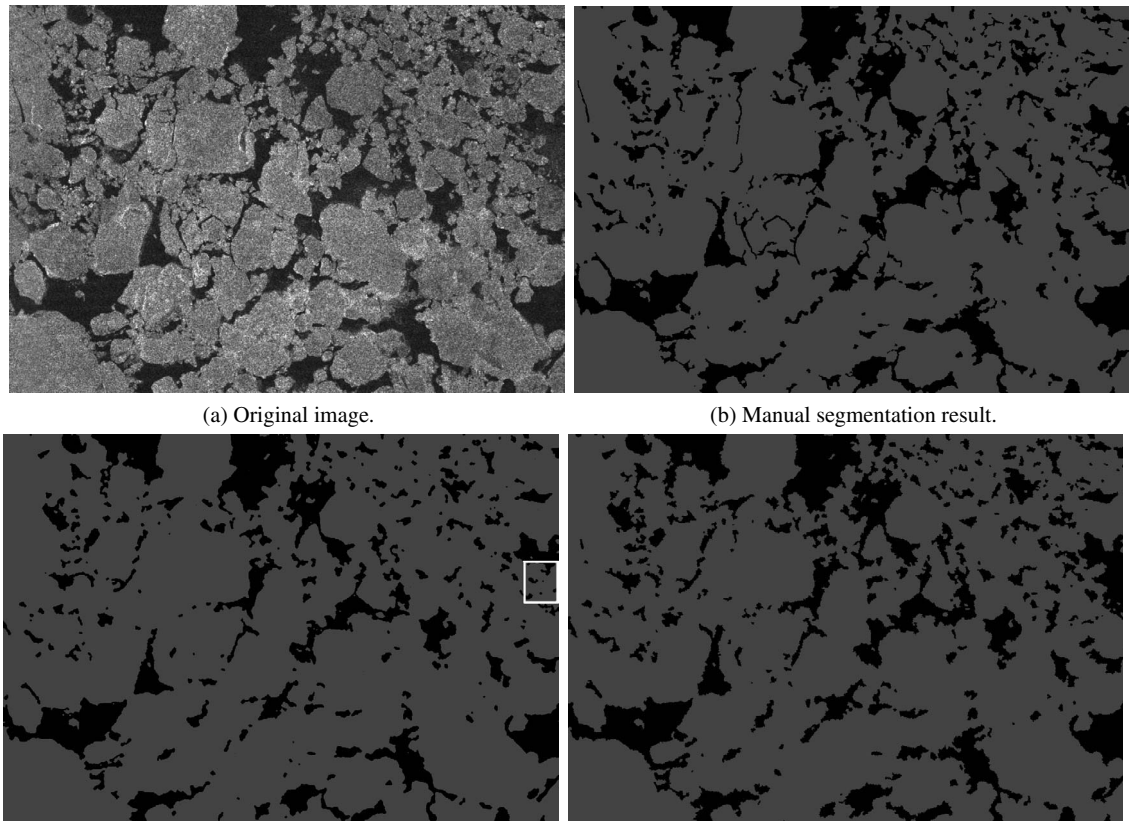
[9] K. Haris, S.N. Efstratiadis, N. Maglaveras, and A.K. Katsaggelos. Hybrid Image Segmentation Using Watersheds and Fast Region Merging. *IEEE Transactions on Image Processing*, 7(12):1684–1699, 1998.

[10] R. Huang, V. Pavlovic, and D.N. Metaxas. A Graphical Model Framework for Coupling MRFs and Deformable Models. In *Computer Vision and Pattern Recognition, 2004*, volume 2, pages 739–746, 27 June - 2 July 2004.

[11] S.Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer, 2001.

[12] B. McCane and T. Caelli. Multi-scale Adaptive Segmentation Using Edge and Region Based Attributes. In *Knowledge-Based Intelligent Electronic Systems*, volume 1, pages 72–80, May 1997.

[13] A. Sarkar, M.K. Biswas, and K.M.S. Sharma. A Simple Unsupervised MRF Model Based Image Segmentation Ap-



(a) Original image. (b) Manual segmentation result.

(c) By traditional model and simulated annealing. (d) By edge weighted model and ICRM.

Figure 6. Segmentation of a SAR image of gulf of St. Lawrence

proach. *IEEE Transaction on Image Processing*, 9(5):801–812, 2000.

[14] J.C. Tilton. Hybrid Image Segmentation for Earth Remote Sensing Data Analysis. In *International Geoscience and Remote Sensing Symposium*, volume 1, pages 703–705, May 1996.

[15] Z. Tu and S.C. Zhu. Image Segmentation by Data-Driven Markov Chain Monte Carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):657–673, 2002.

[16] L. Vincent and P. Soille. Watershed in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):583–598, 1991.

[17] 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.