Remote Sens. **2014**, *6*, 6727-6764; doi:10.3390/rs6086727

OPEN ACCESS

remote sensing

ISSN 2072-4292

www.mdpi.com/journal/remotesensing

Article

Efficient Learning of Spatial Patterns with Multi-Scale Conditional Random Fields for Region-Based Classification

Mitchel Alioscha-Perez [1] and Hichem Sahli [1,2,*]

[1] Electronics and Informatics Department (ETRO), Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Brussels, Belgium; E-Mail: maperezg@etro.vub.ac.be

[2] Interuniveristy Microelectronics Center (IMEC), Kapeldreef 75, BE-3001 Leuven, Belgium

* Author to whom correspondence should be addressed; E-Mail: hsahli@vub.ac.be;
Tel.: +32-2-629-2916; Fax: +32-2-629-2883.

Received: 19 March 2014; in revised form: 25 June 2014 / Accepted: 30 June 2014 / Published: 24 July 2014

Abstract: Automatic image classification is of major importance for a wide range of applications and is supported by a complex process that usually requires the identification of individual regions and spatial patterns (contextual information) among neighboring regions within images. Hierarchical conditional random fields (CRF) consider both multi-scale and contextual information in a unified discriminative probabilistic framework, yet they suffer from two main drawbacks. On the one hand, their current classification performance still leaves space for improvement, mostly due to the use of very simple or inappropriate pairwise energy expressions to model complex spatial patterns; on the other hand, their training remains complex, particularly for multi-class problems. In this work, we investigated alternative pairwise energy expressions to better account for class transitions and developed an efficient parameters learning strategy for the resultant expression. We propose: (i) a multi-scale CRF model with novel energies that involves information related to the multi-scale image structure; and (ii) an efficient maximum margin parameters learning procedure where the complex learning problem is decomposed into simpler individual multi-class sub-problems. During experiments conducted on several well-known satellite image data sets, the suggested multi-scale CRF exhibited between a 1% and 15% accuracy improvement compared to other works. We also found that, on different multi-scale decompositions, the total number of regions and their average size have a direct impact on the classification results.

Keywords: conditional random fields (CRF); multi-class maximum margin; standard piecewise training; image segmentation; image classification

1. Introduction

In this work, we consider the following problem: after segmenting an image into several regions, classify each region into one of the predefined classes. We refer to this process as region-based image labeling. The result is both a segmentation of the image and labeling of each region (segment) as a given object class. Automatic (region-based) image labeling is of importance for several applications, such as scene understanding, image data base querying, *etc.*

In natural scenes, different classes are dependent across similar pixels or regions and also on their co-occurrence. It is not surprising to find regions of certain classes (*i.e.*, river and tree) located next to each other, but it is very unlikely to find some aberrant co-occurrences, such as a building in the middle of a lake. Spatial patterns provide valuable contextual information for achieving consistent labeling. Moreover, since segmentation could provide very small (or too big) ambiguous regions, multi-scale analysis allows one to determine the objects' class more accurately [1,2] in the most appropriate scale within a range (multi-scale) [3].

One of the popular choices to address image classification, including contextual and multi-scale information, has been Markov random fields (MRF) [4,5]. Recently, a discriminative probabilistic framework, named conditional random fields (CRF) [6,7], has been proposed for the same purpose [1,8,9]. However, despite the undeniable improvements of CRF with respect to MRF, two main issues are still common in most of the CRF state-of-the-art models.

The first issue is that classification results, in several cases, are still very low. We associate this problem with an inappropriate pairwise energy expression that fails to model complex spatial patterns (*i.e.*, class transition detections or co-occurrence patterns). This problem can be alleviated by defining proper features for the energies and by defining the appropriate framework to allow more complex energy expressions.

The work of [10] suggests the use of non-linear (exponential) potentials. Similarly, in this work, we model each potential as an unknown function $\psi_c : \mathcal{Y} \to \mathbb{R}$ (of unknown distribution) in a set of possible functions $\psi_{\mathcal{C}} \in \mathcal{H}$, where $\mathcal{H}$ is a Hilbertian space endorsed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. We also introduce a set of additional weighting parameters (variables) in order to control the potential combination.

The work of Gould *et al.* [11] discusses the design of pairwise energies and augments these with relative location information, but their segmentation strategy does not allow the use of multi-scale analysis. The authors in [9] use multi-scale information to express pairwise potentials, despite their local dependencies (intra-scale) involving image structure information (*i.e.*, number of neighbors); their long-range dependencies (multi-scale or inter-scale) only use color distances. We also make use of spectral feature differences in the form of distances, and we go further by including image hierarchical structure (multi-scale) information, such as the scale of analysis, lowest and highest spectral (color) distances to neighboring regions, the number of siblings (the regions sharing the same parent region on the next coarser scale) and the region's lifetime. We follow a strategy more close to [12], where the region locations are determined by the multi-scale information (inside, tangential, non-tangential, *etc.*), but based on hierarchical definitions (parents, siblings, neighbors), since our segmentation strategy guarantees a non-overlapping region's hierarchical decomposition.

The first contribution of this work is a general energy expression that considers:

- Possible non-linear potential functions $\psi_c(\cdot;\theta) = \langle\theta, f_\Phi\rangle_{\mathcal{H}}$ (inner product) that differ from the usual $\psi_c(\cdot;\theta) = \theta^T f$ (dot product), where f is a feature vector;
- A weighted mixture of potentials $E^W(\cdot;\theta,d) = \sum_c d_c\psi_c(.;\theta), d_c \geq 0$;
- Simpler (in number of features), yet rich and expressive pairwise energies involving contextual and multi-scale information.

The second issue, present in most CRF models, is related to the complexity of the parameter learning stage. The conditional maximum likelihood estimation (CMLE) approach requires approximations to estimate the partition function [8,13,14] and achieves relatively simple and accurate parameter learning using gradient descendant optimization methods [15]. Maximum margin confidence approaches do not involve the partition function on their formulation, but poses a challenging quadratic program (QP) for the parameter learning, usually solved with (online) sub-gradient [10,16] or cutting-plane methods [17–19].

When using any of them in the context of CRF parameter learning, another challenge is related to the necessity of solving an inference problem at each iteration during the optimization loop, which increases the computational burden.

Other training strategies, such as the piecewise training framework [20] and its variants [21], proposed simpler formulations to the parameter learning problem and have been shown to be very competitive [19,21]. They have been extensively used in the context of CMLE [11,19,21–23].

An example is the work in [24], where the authors start from a decomposition on the dual formulation of a max-margin QP [25] and introduce a piecewise pseudo-likelihood (PWPL) estimator [21] in order to reduce the (per iteration) inference complexity during training.

In this paper, we consider a particular set of values for the energy weighting coefficients (d_c), where the final energy form involves only single sites and their corresponding pairwise factors, while in [24], the single-site factors of neighboring sites are involved, too. Then, in order to learn the parameters, we decoupled the associated QP into simpler individual multi-class sub-problems that can be efficiently solved dually using the Crammer and Singer fixed-point method [26].

The second contribution of this work is a more general energy expression considering contextual and multi-scale information, in a max-margin parameter learning approach within the piecewise framework, which provides a novel strategy to tackle the resultant parameter learning problem. The proposed energy expression connects to and generalizes other works, reaching the same expression as in [22,24] for different weighting schemes (values for the weights).

The remainder is organized as follows. The multi-scale CRF formulation is provided in Section 2. The proposed training strategy is detailed in Section 3, while the inference process is described in Section 4. Experimental results, using remote sensing images, are reported and discussed in Section 5, and conclusions are given in Section 6.

2. Multi-Scale Conditional Random Fields for Image Classification

The CRF formulation model directly the labels conditioned on observations, decomposing the energy expression according to the different clique order. The clique decomposition theorem defines the overall form of a general CRF model as follows:

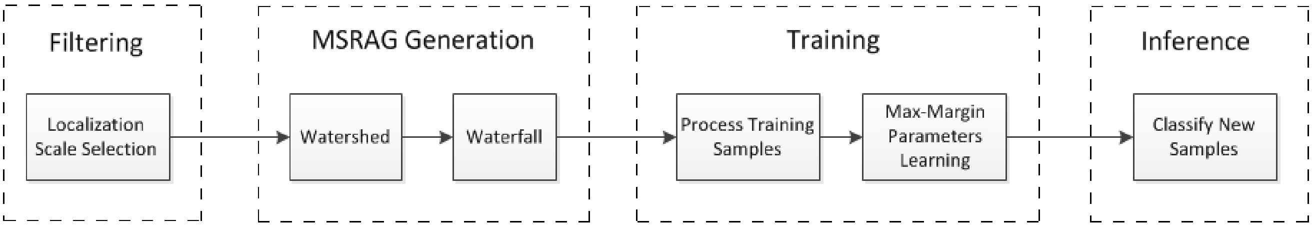
$$P(x \mid y, \theta) \quad = \quad \frac{1}{Z(y,\theta)} \exp\left(\sum_{c \in \mathcal{C}} \psi_c(x_c, y, \theta) \right) \tag{1}$$

where $Z(y, \theta)$ is a normalization function named the partition function.

The above model is known to follow an exponential distribution [4], while the potentials ψ_c are functions defined over cliques $c \in \mathcal{C}$, usually assumed to be linear and expressed as an inner product between features y and model parameters θ [1,7,8,22]; x_c are the labels limited to factors of order $c \in \mathcal{C}$.

This general CRF model can be specified for (defined on) the multi-scale region-based image labeling context. This includes an appropriate potential definition, features for the energy expression, among others. For the region-based image labeling, there are some other stages involved in the overall process, such as the image pre-processing (filtering), segmentation, parameter learning and classification. The general framework is illustrated in Figure 1.

Figure 1. Stages involved in the proposed multi-scale region-based image labeling process.



2.1. PDE-Based Smoothing

In order to avoid blurring and the delocalization of the image features, an image adaptive scale-space filter is used. In this work, we opted for a method that guides the filtering process in such a way that intra-region smoothing is preferred over inter-region smoothing and edges are gradually enhanced. The employed filter [27], belongs to the class of nonlinear anisotropic diffusion filters. It is a combination of the Catté *et al.* [28] regularized Perona and Malik filter [29].

Let $I = \{I^{(1)}, I^{(2)}, \ldots, I^{(\mathcal{R})}\}$ be a vector-valued image defined on a finite domain Ω. The multiscale tower (u) of I is governed by the following system of coupled parabolic partial differential equations (PDEs):

$$\begin{cases} \partial_t u^{(r)} & = \operatorname{div}\left[g\left(|\nabla u_\sigma|\right) \frac{\nabla u^{(r)}}{|\nabla u^{(r)}|} \right] \quad \forall r = 1, 2, \ldots, \mathcal{R} \\ u\left(t=0\right) & = I \\ \partial_{\mathbf{n}} u & = 0 \qquad \text{on } \delta\Omega \end{cases} \tag{2}$$

where $u^{(r)}$ represents the $r-th$ image band, t is the continuous scale parameter, $\delta\Omega$ is the image boundary (with $\mathbf{n}$ denoting the normal direction to it) and σ is a regularization parameter, which ensures the well-posedness of the above system; g is the Lorentzian edge stopping function [30], which is formulated as:

$$g(|\nabla u_\sigma|) = \frac{1}{1 + \frac{|\nabla u_\sigma|^2}{k^2}} \tag{3}$$

The so-called contrast parameter k in Equation (3) separates backward from forward diffusion and is estimated using the cumulative histogram of the regularized gradient magnitude ($|\nabla u_\sigma|$) [27]. In this work, a discrete version of the multiscale tower u, denoted as $U = \{u_0, \ldots, u_n, \ldots, u_N\}$, is obtained by applying the natural scale-space sampling method [31]. From the scale space tower, we select a scale, referred to as the localization scale, at which the diffusion filter has removed the noise without affecting or dislocating the edges of the salient features in the image, and we apply to it the multi-scale region adjacency graph (MSRAG) generation process described in Section 2.2. The localization scale is determined using a maximum correlation criterion proposed in [32].

The maximum correlation criterion selects the scale that maximize the correlation between the original image and the diffused image:

$$s_0 = \arg\max_{t_j} [\mathcal{C}_{\mathrm{m}\rho}(t_j)] = \arg\max_{t_j} \left[\sigma[\mathbf{u}(t)] - \frac{\sigma[\mathbf{n}_{\mathrm{out}}(0)]}{\sigma[\mathbf{n}(0)]} \cdot \sigma[\mathbf{n}_{\mathrm{out}}(t)]\right] \tag{4}$$

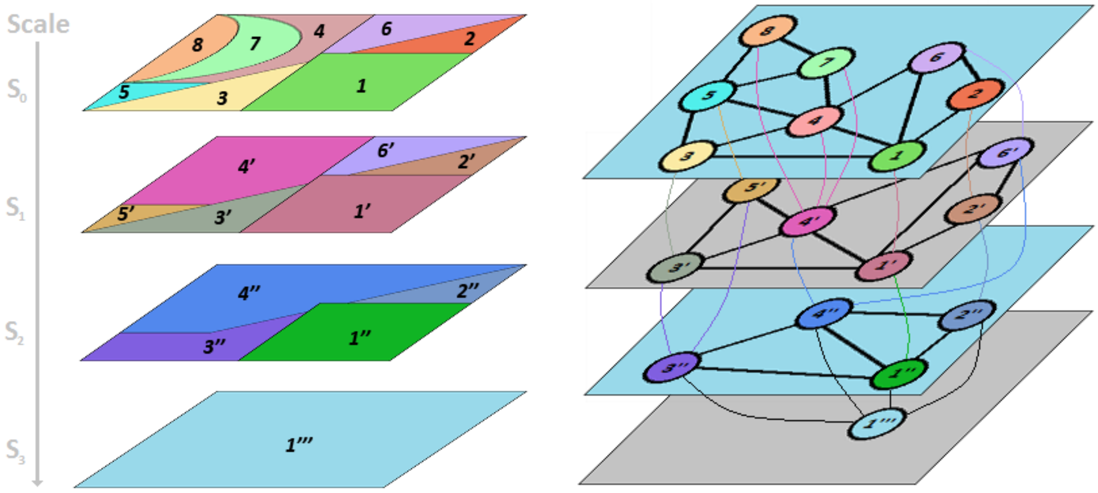
where $\sigma[\mathbf{u}(t)]$ and $\sigma[\mathbf{u}(0)]$ are the standard deviation of both the diffused image at the $t - \mathrm{th}$ scale and the noisy image, respectively, and $\sigma[\mathbf{n}_{\mathrm{out}}(t)]$ and $\sigma[\mathbf{n}_{\mathrm{out}}(0)]$ are the standard deviation of the outlier noise of both the diffused image at the $t - \mathrm{th}$ scale and the noisy image, respectively. The values of $\sigma[\mathbf{n}_{\mathrm{out}}(t)]$ and $\sigma[\mathbf{n}_{\mathrm{out}}(0)]$ are estimated by using the median absolute deviation of the gradient of the diffused and original images (see [33] for details).

2.2. Multi-Scale Region Adjacency Graph (MSRAG)

In this work, the image classification is treated as a hierarchical labeling problem applied to the nodes S of a multi-scale region adjacency graph (MSRAG). The MSRAG represents a nested hierarchy of partitions, $\mathcal{S}_n = \{r_1^n, r_2^n, \ldots, r_{mn}^n\}$; $n = 1, \ldots, N$ (r_j^n being disjoint regions), which preserves the inclusion relationship $\mathcal{S}_n \supseteq \mathcal{S}_{n-1}$, implying that each atom of the set $\mathcal{S}_n$ is a disjoint union of atoms from the set $\mathcal{S}_{n-1}$. In other words, a partition at a coarse level is obtained by merging regions of the finer partition. The MSRAG is defined as a graph $\mathcal{G} = (S, E)$, with the set of nodes S being the partitions at each scale, $S = S_1 \bigcup S_2 \bigcup \cdots \bigcup S_N$, and the edges E are the set of common boundaries between the regions at each scale (intra-scale) and the inter-scale (parent-child) links; see Figure 2.

Hierarchical feature representation through multi-scale segmentation offers new possibilities in object-oriented and multi-scale image analysis [34]. State-of-the art, object-based and object-oriented segmentation algorithms are often based on region growing and merging, or linear and non-linear scale-space.

Figure 2. The multi-scale region adjacency graph (MSRAG) defines the hierarchical structure of an image. (**a**) Nested hierarchy of partitions; (**b**) multi-scale region adjacency graph (MSRAG).



(a) (b)

In the proposed framework, starting from the localization scale s_0 (obtained using the PDE of Section 2.1), we follow the approach of [35], where the input multi-spectral image, u_0, is first segmented using the watershed transform. Then, the waterfall algorithm [36] is used for producing a nested hierarchy of partitions: the multi-scale region adjacency graph (MSRAG); as illustrated in Figure 2. Such a hierarchy preserves the topology of the initial watershed lines and extracts homogeneous objects of a larger scale. The waterfall algorithm removes from the current partition (hierarchical level) all of the boundaries completely surrounded by higher boundaries. Thus, the saliency of a boundary is measured with respect to its neighborhood. The iteration of the waterfall algorithm ends with a partition of only one region. To apply the watershed, the gradient of the multi-spectral image is obtained by combining, using the approach of [37], the gradients of the texture (orientations) and the DiZenzo [38] multi-spectral gradient. For producing the nested hierarchy, we use the approach proposed in [35], where the saliency measure of a boundary is based on a collection of energy functions used to characterize desired single-segment properties and pair-wise segment properties. The single segment properties includes area, convexity, compactness and color variances within the segment. The pair-wise properties includes color mean differences between two segments and edge strength. In [35], the saliency measure, $E(\tilde{r} = r_i \cup r_j | r_i, r_j)$, of a boundary between two neighboring segments r_i and r_j (being the cost of merging the regions r_i and r_j), has been defined as:

$$E(\tilde{r} = r_i \cup r_j \mid r_i, r_j) = E(\tilde{r}) + E(r_i, r_j) \tag{5}$$

The single segment properties, $E(\tilde{r})$, is expressed as:

$$\begin{aligned} E(\tilde{r}) \;\; &= E_{area}(r) \cdot \frac{1}{E_{\text{hom}}(\tilde{r})} \cdot \sum_c E_{\text{var}_c}(\tilde{r}) \cdot \\ &(1 + |E_{\text{conv}}(\tilde{r})|)^{\text{sign}(E_{\text{conv}}(\tilde{r}))} \cdot (1 + |E_{\text{comp}}(\tilde{r})|)^{\text{sign}(E_{\text{comp}}(\tilde{r}))} \end{aligned} \tag{6}$$

and the pair-wise property as:

$$E(r_i, r_j) = E_{\text{edge}}(r_i, r_j) \cdot E_{\text{CMDif}}(r_i, r_j) \tag{7}$$

The different energies associated with segment properties are given in Appendix (C), and additional details can be found in [35].

2.3. Hierarchical Labeling for Region Classification

The labeling is performed using a finite set $\mathcal{L} = \{1, \ldots, L\}$ of interpretation labels. We consider a couple of random fields (X, Y) on $\mathcal{G}$, with $X = \{X^n, \forall n \in [1, \ldots, N]\}$ and $X^n = \{X^n_s, \forall s \in S_n\}$ being the label field and $Y = \{Y^n, \forall n \in [1, \ldots, N]\}$ and $Y^n = \{Y^n_s, \forall s \in S_n\}$ the observations field. In our current implementation Y^n_s represents the region properties at certain site s and scale n. A similar notation holds for the realizations: $x = \{x^n_s, \forall n \in [1, \ldots, N] \forall s \in S_n\}$, $x^n_s \in \mathcal{L}$ and $y = \{y^n_s, \forall n \in [1, \ldots, N] \forall s \in S_n\}$, $y^n_s \in \mathbb{R}^{p \times \mathcal{B}}$, with p the number of pixels in the region r^n_s and $\mathcal{B}$ the number of spectral bands of the image.

We define the general form of the proposed multi-scale CRF for determining the optimal labeling as:

$$P(x \mid y, \theta) = \frac{1}{Z(y, \theta)} \exp\left(\sum_{n=1}^{N} \sum_{s \in S_n} d_1 \psi_A(x^n_s, y, \omega) + \right. \\ \left. + \sum_{n=1}^{N-1} \sum_{s \in S_n} d_2 \psi_L(x^n_s, x^{n+1}_{\overline{s}}, y, \beta^L) + \sum_{n=1}^{N} \sum_{i \in S_n} \sum_{j \in \eta(i)} d_3 \psi_I(x^n_i, x^n_j, y, \beta^I) \right) \tag{8}$$

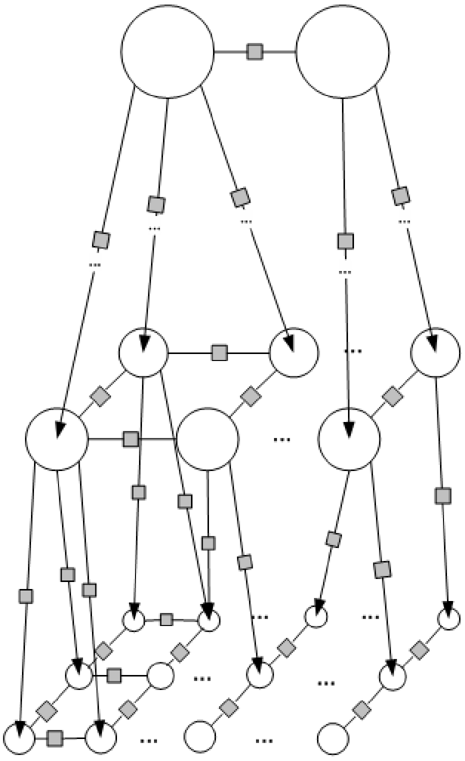
with $\overline{s}$ the parent of a node s, $\eta(i)$ the set of neighbors of site i at the considered level and the parameters $\theta = (\omega, \beta^I, \beta^L)^T$ defined in order to differentiate intra-scale and inter-scale spatial patterns; the weighting parameters $d_m \in \mathbb{R}^+$ regulate each potential individual contribution to the overall energy expression. The pairwise potentials $\psi_I(\cdot; \beta^I)$ and $\psi_L(\cdot; \beta^L)$ encode contextual information between neighbors within the same scale and between parent/child regions at any two consecutive scales, respectively; while $\psi_A(\cdot; \omega)$ is the association potential, which provides local evidence.

The considered model interactions are approximately represented in Figure 3. In order to make the representation simpler to follow, we illustrate only parent-child and four neighbors, while in real segmentation, they usually have any amount of neighbors. This structure falls within a research line of higher order potential design [39,40].

The partition function $Z(y, \theta)$ is defined as the sum of per-clique model energy for every possible realization x, $Z(y, \theta) = \sum_x \prod_{c \in \mathcal{C}} \psi_c(x_c, y, \theta)$, where $\mathcal{C}$ account for up to second order cliques. The number of possible values combinations for x is equal to the number of different labels exponential to the number of sites ($|\mathcal{L}|^{|S|}$); due to this, the exact computation of $Z(y, \theta)$ becomes very easily intractable, even for small graphs with a low number of labels.

For the labeling, we consider inter-scale interactions directionally from parents to child, aiming to collect classification results at the finest scale (see Figure 3), which provides the final classification map.

Figure 3. In the considered structure, during classification, the information flows from coarser (higher) to finer (lower) scales to regulate long-range labels consistency and flows within the same scale to capture local interactions between neighboring regions.



2.4. Potential Definition

The different potentials are defined as follows:

$$\psi_A(x_s, y, \omega) = \sum_{l \in \mathcal{L}} 1_{(x_s=l)} \omega(l; f_y) \tag{9}$$

$$\psi_L(x_s, x_{\bar{s}}, y, \beta^L) = \sum_{l,k \in \mathcal{L}} 1_{(x_s=l)} 1_{(x_{\bar{s}}=k)} 1_{(l=k)} \beta^L(l, k; h_y^L) \tag{10}$$

$$\psi_I(x_s, x_j, y, \beta^I) = \sum_{l,k \in \mathcal{L}} 1_{(x_i=l)} 1_{(x_j=k)} 1_{(l=k)} \beta^I(l, k; h_y^I) \tag{11}$$

where f_y is a unary feature vector, while h_y^L and h_y^I are pairwise feature vectors. The function $1_{(a=b)}$ returns one if the condition $(a = b)$ is met and zero otherwise. The separation among the different classes is determined as the following inner product:

$$\omega(l; f_y) = \langle \omega_l, f_{y,\Phi} \rangle_{\mathcal{H}} \tag{12}$$

$$\beta^L(l, k; h_y^L) = \left\langle \beta^L_{1(l=k)}, h^L_{y,\Phi} \right\rangle_{\mathcal{H}} \tag{13}$$

$$\beta^I(l, k; h_y^I) = \left\langle \beta^I_{1(l=k)}, h^I_{y,\Phi} \right\rangle_{\mathcal{H}} \tag{14}$$

being $w_l \in \mathcal{H}$ a total of l unknown parameters ($l \in \mathcal{L}$), while $\beta^L_{1(l=k)}$ and $\beta^I_{1(l=k)}$ are two and two unknown parameters, respectively (four in total with $1_{(l=k)} \in \{0, 1\}$).

The (mapped) features $f_{y,\Phi}$, $h^L_{y,\Phi}$ and $h^I_{y,\Phi}$ are the feature vectors (f_y, h^L_y and h^I_y) mapped by a function $\Phi : \mathcal{Y} \rightarrow \mathcal{H}$. In practice, neither the mapping function $\Phi(\cdot)$ nor the space $\mathcal{H}$ needs to be explicitly defined; it is sufficient to define a symmetric and positive definite kernel function $K : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$, to inherit both the hypothesis space $\mathcal{H}$ as its unique reproducing kernel Hilbert space (the Moore–Aronszajn theorem [41]) and the mapping function $\Phi(\cdot) = K(y, \cdot)$.

Note that the distribution of the potentials $\psi_c(\cdot; \theta)$ is assumed to be unknown. For the particular case of using linear kernel functions, our work will recover the usual linear potential functions, and for any other kernel functions, we will recover more complex potential functions. A bias value can be easily introduced in Equations (12)–(14) (if wanted or necessary) by making use of algebraic methods.

In this work, we consider class transition detections for pairwise potential modeling $\{1_{(x_a=l)} 1_{(x_b=k)} 1_{(l=k)}, \forall\, l, k \in \mathcal{L} \times \mathcal{L}\}$, which can be seen as two binary models: a binary model (transition occur: yes or no) for the intra-scale potential and a binary model (correct region merging: yes or no) for the inter-scale potential.

However, the model has been designed for a broader multi-class pairwise potential definition $\{1_{(x_a=l)} 1_{(x_b=k)}, \forall\, l, k \in \mathcal{L} \times \mathcal{L}\}$ as two multi-class co-occurrence models: a multi-class model of $L \times L$ classes (the transition of class a to b occurs: yes or no?) for the intra-scale potential and a multi-class model of $L \times L$ classes (correct region merging of class a to b: yes or no) for the inter-scale potential.

2.4.1. Unary Local Features

The association potential, $\psi_A(\cdot; \omega)$, is defined as the probability of a site to acquire a certain label considering the site observations independently. The single-site observation feature vector f_y is obtained as a concatenation of different features making use of an observation y. The following features have been considered in the current implementation:

1. Spectral: the spectral signature of a region y is characterized for each channel by its mean, standard deviation, kurtosis and skewness.
2. Textural: the texture of a region is characterized using the local binary pattern [42], from which we compute the region's mean value, repeating the process using eight different radius (neighborhood) values.
3. Morphological: for the shape information, we compute the following morphological features: elongation, area to length ratio and extent.
4. Scale: the actual scale of the region in question.

Considering a color RGB image, the resultant feature vectors will contain 24 components.

2.4.2. Inter-Scale Pairwise Features

The inter-scale potential, $\psi_L(\cdot; \beta^L)$, aims at modeling the probability distribution of a good region merging having taken place (detecting errors on segmentation) during the segmentation process. This potential function is based on the observations of the region in question (named the actual region) and other additional information, such as:

1. its parent region: a good merging is supposed to keep spectral homogeneity between the actual region and its parent;

2. all of its siblings: a region that is merged with several siblings is supposed to keep spectral homogeneity with each one of them;
3. the number of siblings: the greater the number of regions that are merged (the number of siblings), the more likely it is for a bad merging to occur;
4. the scale of the actual region: in hierarchical segmentation, higher levels are more likely to exhibit regions that combine multiple objects (bad merging);
5. the lifetime of the actual region: whenever a region exhibits a high lifetime (strong edges), it is because it can be well differentiated (in terms of spectral homogeneity) from its neighboring regions; therefore, the merging of regions with a high lifetime implies a bad merging detection.

For the inter-scale pairwise energy expression, we combine all of the previous information in a five-dimensional vector as follows:

$$h_y^L = \left[D(y_s, y_{\bar{s}}), \sum_{j \in \varsigma(s)} D(y_s, y_j), |\varsigma(s)|, n, \text{lifetime}(s) \right]^T \tag{15}$$

$\varsigma(s)$ being all of the siblings (which have a parent region in common) of the region in question; lifetime(s) is the number of scales at which a region remains the same (without being merged) according to the MSRAG structure; n denotes the level of the actual region in the MSRAG; $|\varsigma(s)|$ denotes the number of siblings of the region in question; and $D(\cdot,\cdot)$ is the Bhattacharyya distance between any two regions' histogram.

2.4.3. Intra-Scale Pairwise Features

The intra-scale potential, $\psi_I(\cdot;\beta^I)$, aims at modeling the probability distribution of two regions sharing the same label. This potential function is based on the observations of the two regions in question, but also benefits from additional information, such as:

1. the corresponding neighbor: a region that shares the same label with a neighbor is supposed to keep spectral homogeneity;
2. lowest neighbor distance: the lowest spectral distance from all neighbors provides a measurement of how homogeneous the surrounding area that contains the region in question is;
3. highest neighbor distance: the highest spectral distance from all neighbors provides a measurement of how homogeneous the surrounding area that contains the region in question is;
4. the scale of the actual region: in hierarchical segmentation, higher levels are more likely to exhibit regions that combine multiple objects (bad merging).

For the intra-scale pairwise energy expression, we concatenate all of the previous information in a four-dimensional vector as follows:

$$h_y^I = \left[D(y_s, y_{\bar{s}}), \min_{j \in \eta(s)} D(y_s, y_j), \max_{j \in \eta(s)} D(y_s, y_j), n \right]^T \tag{16}$$

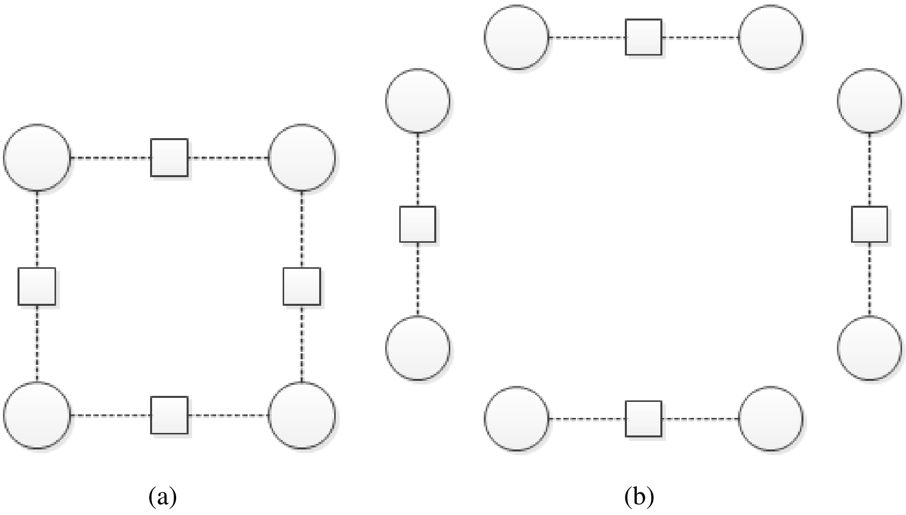
n being the actual processing scale of site s in the MSRAG and $D(\cdot,\cdot)$ the Bhattacharyya distance between the two regions' histogram.

3. Parameter Learning

3.1. Piecewise Training Framework

The piecewise training framework has been proposed by [20] in order to reduce the complexity of the partition function estimation in CMLE approaches, consisting of dividing the full model into pieces $\mathcal{P} \in \mathcal{T}$, according to a set of piece types $\mathcal{T}$ [21]. An illustration of the standard piecewise framework ($\mathcal{P} = \mathcal{C}_2$) can be seen in Figure 4b.

Figure 4. Piecewise pairwise factors of a Markovian model are considered individual pieces of the same original model. (**a**) Original model. (**b**) Standard piecewise training of the same model (pairwise factors).



From the partitions $\mathcal{P}$, a new family of upper bounds $\log Z_{\mathcal{P}}(y, \theta)$ can be used to provide a simpler (tractable) estimator [20] for the minimization of the log-partition function $\log Z(y, \theta) \propto \sum_{\mathcal{P} \in \mathcal{T}} \log Z_{\mathcal{P}}(y, \theta)$. This allows one to re-write the log-likelihood of Equation (1) as follows:

$$\begin{aligned} \log P(x \mid y, \theta) &= \log \frac{1}{Z(y, \theta)} \exp \left(\sum_{c \in \mathcal{C}} \psi_c(x_c, y, \theta) \right) \\ &= \sum_{\mathcal{P} \in \mathcal{T}} \sum_{c \in \mathcal{P}} \psi_c(x_c, y, \theta) - \sum_{\mathcal{P} \in \mathcal{T}} \log Z_{\mathcal{P}}(y, \theta) \\ &= \sum_{\mathcal{P} \in \mathcal{T}} \left(\sum_{c \in \mathcal{P}} \psi_c(x_c, y, \theta) - \log Z_{\mathcal{P}}(y, \theta) \right) \end{aligned} \qquad (17)$$

where $\psi_c(\cdot; \theta)$ consists of the per clique and per piece-type energy given θ and $Z_{\mathcal{P}}(y, \theta)$ the local partition function estimator per model piece-type $\mathcal{P}$ [21]. Based on Equation (18), gradient descendant techniques provide a simple way to learn θ.

3.2. Max-Margin Parameter Learning

Within the max-margin parameter learning framework, the problem is posed differently. The parameter learning problem aims to find a θ, such that, for the observed labels $\hat{x}$ and observations y in training samples, $P(\hat{x} \mid y, \theta)$ has to be the highest possible, among all possible values of x:

$$P(\hat{x} \mid y, \theta) \geq P(x \mid y, \theta) \quad \forall x \neq \hat{x} \tag{18}$$

This will guarantee that for any possible value x, $\arg\max_x P(x \mid y, \theta)$ will attain the maximum when $x = \hat{x}$. Considering the clique decomposition theorem (Equation (1)), the previous inequality takes the following form:

$$\begin{aligned}\frac{1}{Z(y,\theta)} \exp\left(\sum_{\mathcal{P}\in\mathcal{T}} \sum_{c\in\mathcal{P}} \psi_c(\hat{x_c}, y, \theta)\right) \geq \\ \frac{1}{Z(y,\theta)} \exp\left(\sum_{\mathcal{P}\in\mathcal{T}} \sum_{c\in\mathcal{P}} \psi_c(x_c, y, \theta)\right)\end{aligned} \tag{19}$$

therefore:

$$\sum_{\mathcal{P}\in\mathcal{T}} \sum_{c\in\mathcal{P}} \psi_c(\hat{x_c}, y, \theta) \geq \sum_{\mathcal{P}\in\mathcal{T}} \sum_{c\in\mathcal{P}} \psi_c(x_c, y, \theta) \quad \forall x \neq \hat{x} \tag{20}$$

where the partition $Z(y, \theta)$ function is no longer necessary to be computed.

In the max-margin learning approach, we want that the previous in-equivalence Equation (20) not only holds, but also has the maximum margin. Assuming the following notations:

$$E_{\mathcal{P}} = -\sum_{c\in\mathcal{P}} \psi_c(x_c, y, \theta) \tag{21}$$

$$\hat{E_{\mathcal{P}}} = -\sum_{c\in\mathcal{P}} \psi_c(\hat{x_c}, y, \theta) \tag{22}$$

and introducing a generic loss function $\ell(x, \hat{x})$, the parameter learning can be formulated as the following optimization problem:

$$\begin{aligned} \max_{\lambda > 0,\, \theta} \quad & \lambda \\ \text{s.t.} \quad & \sum_{\mathcal{P}\in\mathcal{T}} E_{\mathcal{P}} - \sum_{\mathcal{P}\in\mathcal{T}} \hat{E_{\mathcal{P}}} \geq \lambda \ell(x, \hat{x}) \\ & \|\theta\|_2 \leq 1 \qquad\qquad \forall x \neq \hat{x} \end{aligned} \tag{23}$$

which is similar to the max-margin parameter learning problems posed in [18] without considering noisy data (hard margin). The application of cutting-planes or sub-gradient methods are a common choice to solve this problem, requiring a full inference on every iteration step [16,18]. This is the first drawback associated with the max-margin parameter learning. The second drawback is related to the number of constraints included by $x \neq \hat{x}$ in Equation (23), which is an exponential value in the number of sites equal to $|\mathcal{L}|^{|S|} - 1$.

3.3. Proposed Efficient Max-Margin Parameter Learning

In our proposed learning strategy, we avoided using sub-gradient and cutting-plane methods by decomposing the original problem into several individual multi-class sub-problems that can be efficiently solved dually, and we significantly reduced the number of constraints included by $x \neq \hat{x}$. Starting from Equation (23), a usual assumption [43] is that, likewise the per-clique energy functions $\psi_c(\cdot;\theta)$, the loss function $\ell(x,\hat{x})$ is decomposable according to the different cliques' order. Using the notation $\ell_{\mathcal{P}} = \ell_{\mathcal{P}}(x_{\mathcal{P}}, \hat{x_{\mathcal{P}}})$, we can state that:

$$\ell(x,\hat{x}) = \sum_{\mathcal{P}\in\mathcal{T}} \ell_{\mathcal{P}}(x_{\mathcal{P}}, \hat{x_{\mathcal{P}}}) \tag{24}$$

$$= \sum_{\mathcal{P}\in\mathcal{T}} \ell_{\mathcal{P}} \tag{25}$$

Then, for any model partition (as simple or as complex as wanted) under $\mathcal{T}$, we propose to decouple our formulation of Equation (23), decomposing it into the following set of individual multi-class sub-problems (see Appendix A):

$$\forall \mathcal{P} \in \mathcal{T} \quad \begin{cases} \min\limits_{\theta_{\mathcal{P}}} & \frac{1}{2}\|\theta_{\mathcal{P}}\|_2^2 + C\sum_i \xi_i \\ \textit{s.t.} & E_{\mathcal{P}} - \hat{E_{\mathcal{P}}} \geq \ell_{\mathcal{P}} - \xi_i, \quad \forall_{i=1}^{N_{\mathcal{P}}}, \forall x_i \in \Omega_{\mathcal{P}} \end{cases} \tag{26}$$

where $\Omega_{\mathcal{P}}$ is the set of label combinations for a $\mathcal{P}$-order clique (see Appendix A) and $N_{\mathcal{P}}$ the total number of related factors [21]; $x_i \in \Omega_{\mathcal{P}}$ represents the i-th factor ($i = \{1,\ldots,N_{\mathcal{P}}\}$) possible labels, ξ_i are slack variables and $\hat{x}_i$ along with y_i the corresponding (i-th) ground-truth and observations, respectively, present in the training samples.

Note that, since $\mathcal{T}$ defines the relations of each site with its neighborhood in a Markov blanket (the set of its parent, children and brothers), it will determine the number of constraints per each i-th site, as well as the energy form related to each individual site i. Since they have a direct impact on the model complexity, simple model partitions, such as piecewise (PW) or piecewise-pseudolikelihood (PWPL), are often preferred, but any structure can be defined for $\mathcal{T}$.

The per-site energy expression, usually defined as $E(\cdot;\theta) = \sum \psi_c(\cdot;\theta)$, can take different forms and vary according to the site's relations with its neighbors in the Markov blanket. It is known that such simple energy expressions are sometimes inappropriate, and the use of extra weighting terms is necessary to adjust the potential combination [9,18,22,39].

We propose a more general formulation $E^W(\cdot;\theta,d)$ for the per-site energy expression, which involves a weighted mixture of potential functions, resulting in a weighted combination of local, contextual and multi-scale information:

$$\begin{aligned} E^W(x_i,y;\theta,d) = d_1\psi_A(x_i,y,\theta) + \\ \sum_{j\in C_{\text{neighbor}}|j\in\eta(i)} \big(d_4\psi_A^I(x_j,y,\theta) + d_2\psi_I(x_j,x_i,y,\theta)\big) + \\ \sum_{j\in C_{\text{parent/child}}|j\in\eta(i)} \big(d_5\psi_A^L(x_j,y,\theta) + d_3\psi_L(x_j,x_i,y,\theta)\big) \end{aligned} \tag{27}$$

In order to establish connections with other energy expressions, we defined the model structure $\mathcal{G}$ according to some MSRAG and considered, without loss of generality, some model partition, such as the PWPL proposed in [21] and assumed in [24].

As can be seen, $\psi^I_A(\cdot)$ and $\psi^L_A(\cdot)$ are unknown potential functions involving single factors of neighbors within the same scale and parent/child scales, respectively. The weighting terms $d_m, m = 1, \ldots, 5$ control the mixture of the different potential information.

The proposed energy expression, E^W, generalizes the one used in the works of Alahari *et al.* [24], referred to as E^{PWPL}. Note that assuming a fixed set of weights $d = (1, 1, 0, 1, 0)$, both energy expressions E^W and E^{PWPL} become equivalents: $E^W \equiv E^{PWPL}$. It also generalizes the one used in [22] within the context of CMLE and referred to as E^{PW}, since, for the particular case of $d = (1, 1, 0, 0, 0)$, both energy expressions E^W and E^{PW} become equivalents: $E^W \equiv E^{PW}$.

Then, we introduce the weighted potential mixture into our learning strategy to obtain the following set of QP problems:

$$\forall \mathcal{P} \in \mathcal{T} \quad \begin{cases} \min\limits_{\theta_\mathcal{P}} & \frac{1}{2}\|\theta_\mathcal{P}\|_2^2 + C\sum_i \xi_i \\ s.t. & E^W_\mathcal{P} - \hat{E}^W_\mathcal{P} \geq \ell_\mathcal{P} - \xi_i, \quad \forall_{i=1}^{N_\mathcal{P}}, \forall x_i \in \Omega_\mathcal{P} \end{cases} \tag{28}$$

where:

$$E^W_\mathcal{P}(x_i, y, \theta) \;=\; -\sum_{c \in \mathcal{P}} d_c \psi_c(x_i, y, \theta) \tag{29}$$

We considered values for d, such that $\forall_{m=1}^{5} d_m \in \mathbb{R}^+$ with $d_4 = d_5 = 0$ and solved the parameter learning in a max-margin approach, where each resultant problem of Equation (28) takes the form of a standard multiclass support vector machine (SVM) model, very similar to the work of Crammer and Singer [26], that can be efficiently solved dually using the fixed-point method.

Similar to [39] for max-margin and to [22] for CMLE, we assume a fixed value for d during the training stage, and we teach them afterword on a validation set.

Since each individual problem is solved dually, the obtained set of optimal dual variables (α^*-coefficients) are related, under a strong duality assumption (see [25,26]), to the primal variables (ω, β^L and β^I) by the Karush–Khun–Tucker (KKT) conditions. Therefore, the optimal ω^l is the following:

$$\forall l \in \mathcal{L} \qquad (\omega^l)^* = \sum_{i|\hat{x}_i = l} \alpha^*_{il}\, y_i \tag{30}$$

where $\alpha^*_{i,l}$ are the (already optimal) i^{th} dual variables (also known as support vectors) for the class l and y_i the observation of the $i-th$ sample; the same for β_I, β_L. For more details, we refer to [26].

The following algorithm details the proposed training strategy:

```
function LearnParams( $y, \hat{x}, \mathcal{G}, \mathcal{T}$ )
    for ($\mathcal{P} \in \mathcal{T}$) do
        $y_{\mathcal{P}} \leftarrow$ StructuredObservations($y, \mathcal{G}, \mathcal{P}$)      ▷ Defined $y_n^s$ in Section 2.3
        $\hat{x}_{\mathcal{P}} \leftarrow$ StructuredSamples($\hat{x}, \mathcal{G}, \mathcal{P}$)      ▷ Detailed in Section 3.4

        $\alpha_{\mathcal{P}} \leftarrow$ MulticlassSVM($y_{\mathcal{P}}, \hat{x}_{\mathcal{P}}$)      ▷ Dual of Equation (28)
        $\omega_{\mathcal{P}} \leftarrow$ ToPrimal($\alpha_{\mathcal{P}}, y_{\mathcal{P}}, \hat{x}_{\mathcal{P}}$)      ▷ Using Equation (30)
    end for

    $\theta^* = (\omega_{\mathcal{P}_1}, \ldots, w_{\mathcal{P}_N})^T, \quad N = |\mathcal{T}|$
    return $\theta^*$
end function
```

3.4. Structuring the Training Samples According to the MSRAG

Regarding the observations, since the training data is a limited set of labeled pixels (regions) from the original image, it is structureless and has no related hierarchical information. Therefore, in our model, they are represented by the observations on the first (finest) scale y_s^1 only.

In order to obtain the training samples at the other (higher) scales, we propagate this information to the next coarser scales (see Figure 5), considering that a parent region will have the same label as its child, only if all of the children share the same label. When children have different labels, we select the most represented class, by counting the number of pixels related to each class and taking the highest one:

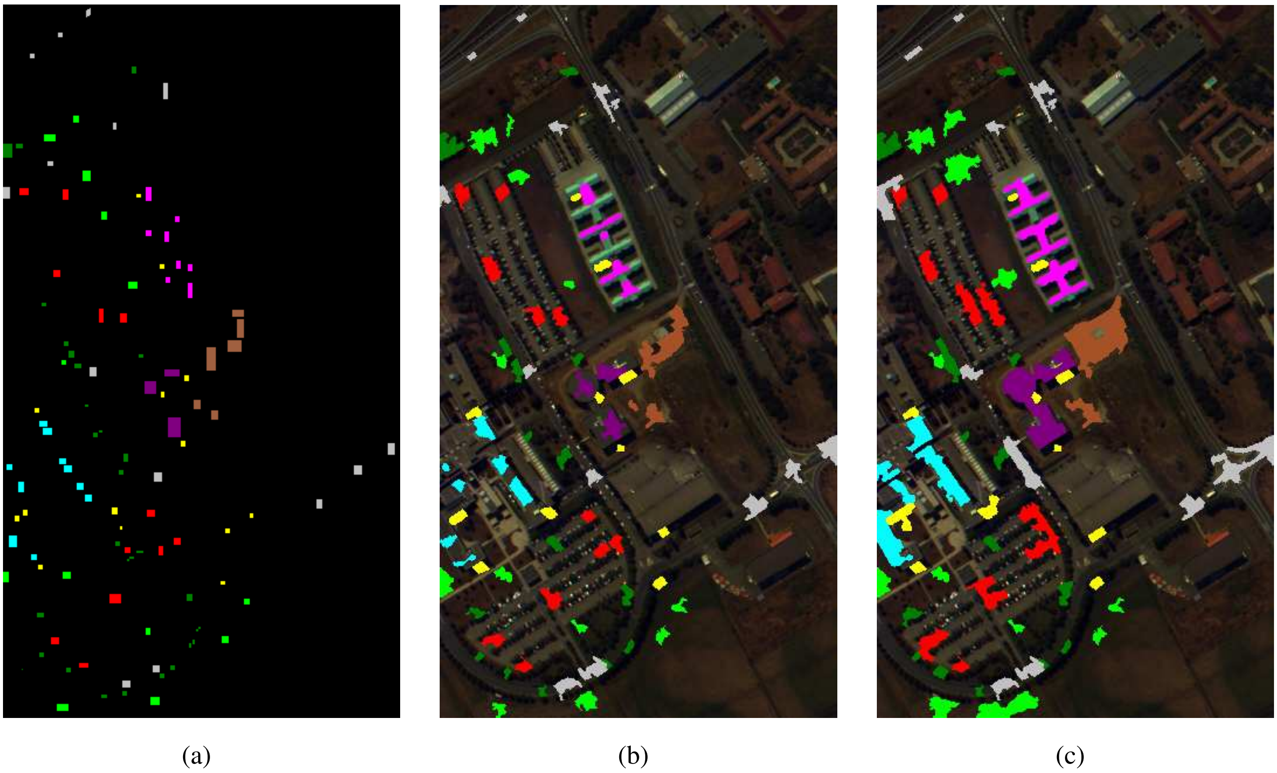
$$\arg\max_{l} \sum_{i|\bar{i}=s} |y_i^n| 1_{\hat{x}_i^n=l} \tag{31}$$

$1_{\hat{x}_i^n=l} = 1$ if the region $r_{s_i}^n$ has been labeled with a label l on the ground-truth or zero otherwise, and $|y_i^n|$ is the number of pixels inside the region in question. Then, a region at a higher scale will inherit this class according to the following condition:

$$\hat{x}_s^{n+1} = \begin{cases} l & \text{if} \quad 1 - \frac{\sum_{i|\bar{i}=s} |y_i^n| 1_{\hat{x}_i^n=l}}{|y_s^{n+1}|} \geq th \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

The above condition guarantees (for the training stage only) that a parent region will receive the same label of the class most represented on its children regions, only if the total number of known pixels on its children covers more than a certain percent of the parent region. Such a percent is given by the threshold th; otherwise, it will remain as an unknown (non-labeled, represented by $x_s^{n+1} = 0$) region, and therefore, it will not be part of the training sample set. A similar threshold has been used in [39] for the same purpose. We usually keep this value as $th = \frac{1}{L}$, where L is the total number of available classes.

Figure 5. Information flows from the finest to coarsest scales in order to conform the training sample set structured according to the MSRAG. (**a**) The provided training samples. (**b**) Training at Scale 1. (**c**) Training at Scale 5.



4. Inference

Once the optimal parameters θ^* haven been learned, we can use them to label new samples, by finding the most likely label configuration given the observations and parameters: $\arg\max_{x \in X} P(x \mid y, \theta^*)$, where:

$$P(x \mid y, \theta^*) = \exp\left(\sum_{n=1}^{N} \sum_{s \in S_n} d_1 \psi_A(x_s^n, y, \omega^*) + \right. \\ \left. + \sum_{n=1}^{N-1} \sum_{s \in S_n} d_2 \psi_L(x_s^n, x_{\bar{s}}^{n+1}, y, \beta^{L^*}) + \sum_{n=1}^{N} \sum_{i \in S_n} \sum_{j \in \eta(i)} d_3 \psi_I(x_i^n, x_j^n, y, \beta^{I^*}) \right) \quad (33)$$

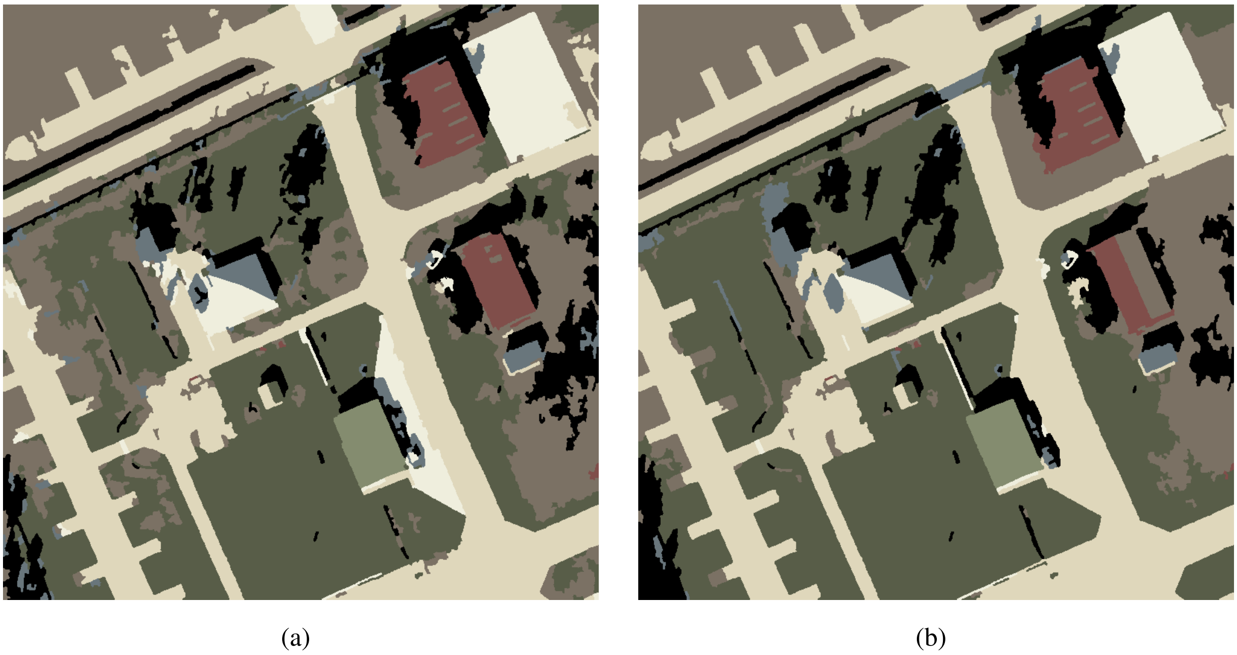
The inference process can be performed using several methods in the case of multiclass problems structured as a general graph with loops. Among them, we can find: message passing algorithms, such as loopy believe propagation (LBP) [44] or tree re-weighted LBP (TR-LBP) [45], and move making algorithms (graph-cut), such as α/β swap [46] or α-expansion [46].

We selected the LBP method for the inference process, which approximate the marginalized posterior by a value named believes, computed using a message passing algorithm [44].

Initially, each site belief is equal to the association potential. Then, the neighbors and parent of each region start to iteratively influence the site belief value, until all of the belief values settle, providing the final belief value as an approximation of the marginalized posterior.

Note that different classification maps are obtained from the inference process on different scales, and coarser scale results have an influence on finer classification maps during inference (see Figure 6). This influence consists in the enforcement of long-range dependencies from coarser scales to finer scales.

Figure 6. Inference at coarser (higher) scales will influence inference on finer (lower) scales, enforcing long-range dependencies. (**a**) Inference result at Scale 1. (**b**) Inference result at Scale 9.



The convergence properties of LBP have been associated with the properties of Bethe free energy and converge to stationary points of this free energy approximation; see [44] for further details. Recalling Equation (33), it might be possible that the terms of each independent potential are correlated. This can lead to over-counting problems during the inference. Once the parameter θ^* has been found for a fixed weighting scheme, likewise in [9,18,22,39], we learn the weights d_1, d_2, d_3 via cross-validation.

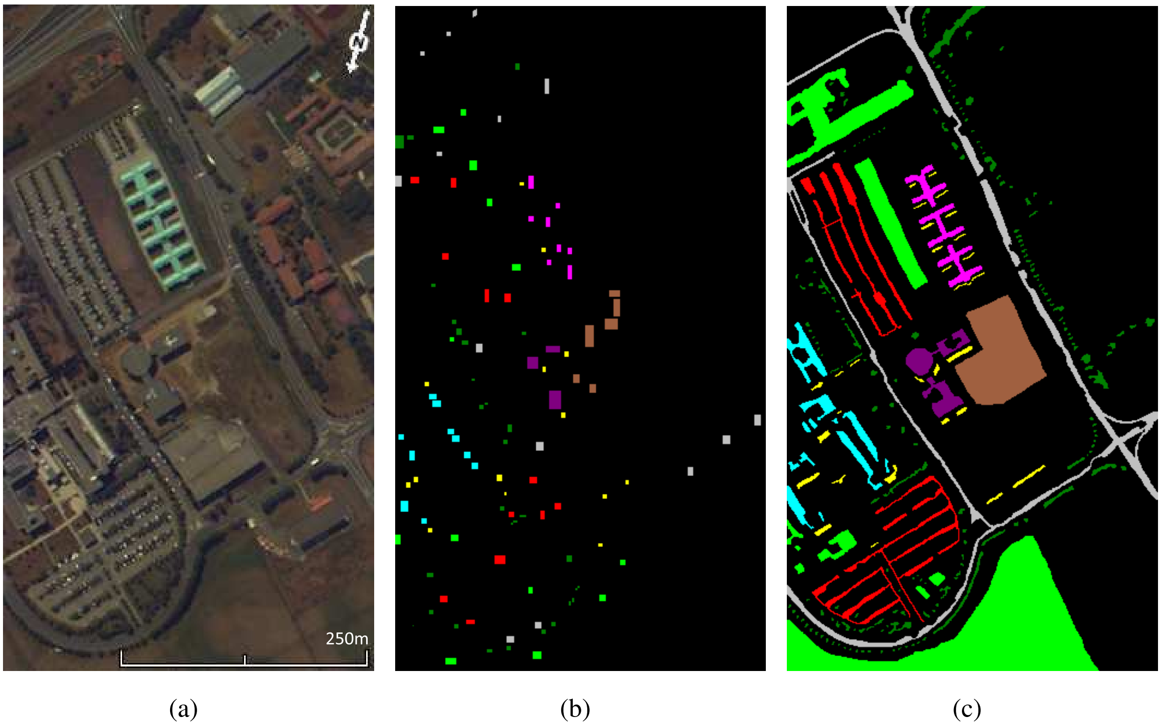
5. Experimental Results and Discussion

For the evaluation of the proposed model, we selected two well-known remote sensing datasets, namely the Pavia City Center and Pavia University (both images available online at http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes), and a high resolution aerial RGB image. The Pavia images are used to compare the proposed method to state-of-art ones, and the high resolution image is used to analyze the effectiveness of our multi-scale CRF approach.

5.1. Pavia University Dataset

For the case of the Pavia University dataset (Figure 7a), both the training samples (Figure 7b) and the ground-truth (Figure 7c) were available and used in the work of Aksoy [47], where a pixel-based and a region-based classification method have been compared. The details of the training sample pixel distribution can be found in [47].

Figure 7. The Pavia University dataset (45°11′39.84″N 9°08′06″E). (**a**) Image. (**b**) Training samples. (**c**) Testing samples.



(a) (b) (c)

The segmentation scheme described in Section 2.2 has been applied to Pavia University using three spectral bands (Bands 68, 30 and 2), out of the 103 available bands. We further used the five first obtained scales in order to define the different potentials, learn the parameters and perform the classification. It has to be noted that state-of-the-art hierarchical methods use a different number of scales: two in [39], three in [40] and four in [1]; however, we found that scales higher than five usually do not provide complementary information, due to the limited ground-truth.

Results using the the proposed multi-scale CRF model, the pixel-level Bayesian (P-L-B) model of [47] and the region-based Bayesian (R-B-B) model of [47] are given in Tables 1 and 2.It can be observed that the proposed approach outperforms both the previous results, as well as other region-based approaches, despite the fact that our proposed model was trained using only three spectral bands, out of the 103 available bands used in [47].

Table 1. Pavia University dataset: performance evaluation of the proposed approach in comparison with previous results.

Model	**# of Used Spectral Bands**	**Accuracy (%)**
Pixel-Level Bayesian [47]	10 (PCA from 103)	74.92
Quadratic Gaussian [47]	10 (PCA from 103)	81.27
Region-Level Bayesian [47]	10 (PCA from 103)	84.44
Proposed CRF model	3 (68, 30 and 2)	**87.16**

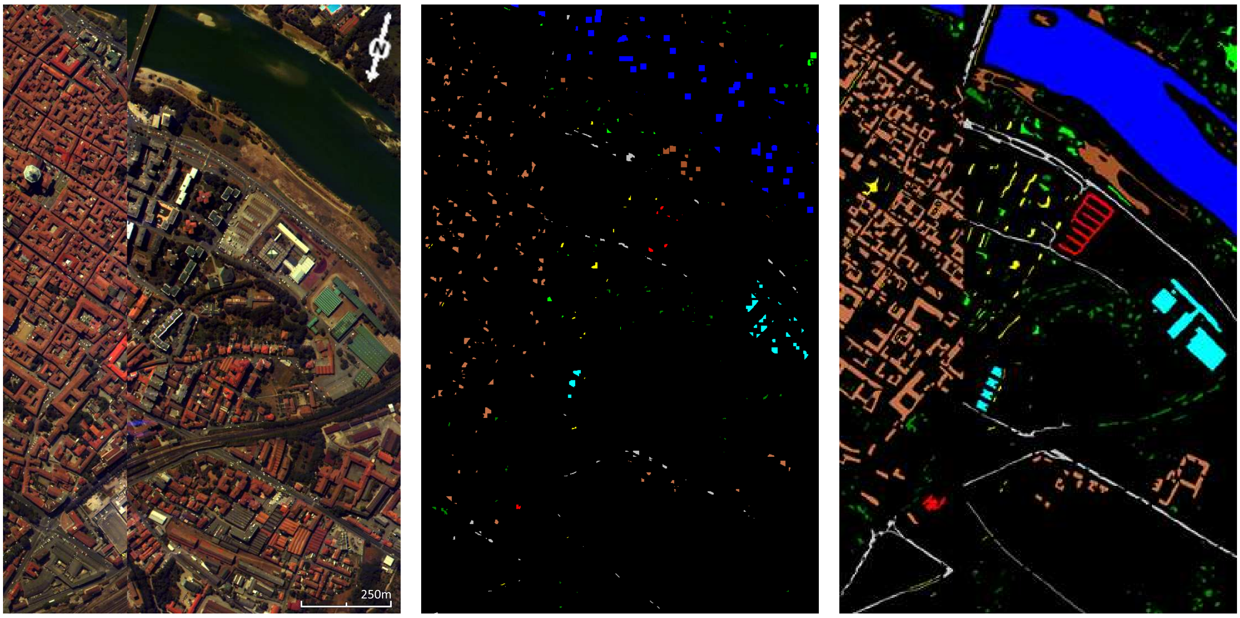
Table 2. Pavia University dataset: performance evaluation of the proposed approach in comparison with previous results; per-class accuracy (%) and overall accuracy (%). P-L-B, pixel-level Bayesian; R-B-B, region-based Bayesian.

Class	**Proposed**	**Models in [47]**	
	CRF	**P-L-B**	**R-B-B**
C1-Asphalt	86.38	61.00	69.67
C2-Meadows	95.91	78.87	92.48
C3-Gravel	57.22	69.84	64.79
C4-Trees	83.29	95.53	95.99
C5-Metal Sheets	90.04	99.70	99.93
C6-Bare Soil	76.09	74.47	79.82
C7-Bitumen	67.51	62.33	77.67
C8-Bricks	83.65	56.06	71.56
C9-Shadow	95.35	97.47	98.10
Overall Accuracy	**87.16**	74.92	84.44

5.2. Pavia Center Dataset

In the case of the Pavia Center dataset, the ground-truth was available, but not the training samples. As consequence, we decided to take 10% of the ground-truth as the training samples. We used a random uniform distribution for determining: (1) squares of a random size; and (2) to locate these squares in random positions within the image. During the random training selection, we also granted that all of the classes were roughly represented in a balanced way, resulting in 10% of each class selection conforming to 10% of the ground-truth. The randomly obtained training samples are detailed in Table 3 and are illustrated in Figure 8b. As such, compared to [47], we use less training pixels for the following five out of the considered nine classes: trees, meadows, bricks, bare soil and shadow.

Figure 8. Pavia Center dataset (45°11′0.96″N 9°08′25.44″E). (**a**) Image. (**b**) Training samples. (**c**) Testing samples.



(a) (b) (c)

Results using the the proposed multi-scale CRF model, the pixel-level Bayesian (P-L-B) model of [47] and the region-based Bayesian (R-B-B) model of [47] are given in Table 3.

Table 3. Pavia Center dataset: description of the randomly selected samples, and performance evaluation (as accuracy %) of the proposed approach.

Class	**Training Pixels**	**Total Pixels**	**(%)**	**Proposed**	**Models in [47]**	
				CRF	**P-L-B**	**R-B-B**
C1-Water	4410	65,971	06.68	99.98	99.85	99.74
C2-Trees	713	7598	09.38	89.07	84.5	81.71
C3-Meadows	346	3090	11.20	69.19	87.96	95.21
C4-Bricks	174	2685	06.48	76.16	83.35	83.69
C5-Bare Soil	800	6584	12.15	95.37	78.77	93.24
C6-Asphalt	1058	9248	11.44	90.27	85.39	93.74
C7-Bitumen	1202	7287	16.50	95.1	83.18	89.28
C8-Tiles	4386	42,826	10.24	99.72	97.67	98.96
C9-Shadow	268	2863	09.36	89.31	82.82	86.20
Overall	13,357	148,152	09.02	97.20	94.89	96.76

From Tables 2 and 3, it is apparent and confirmed according to the measured accuracy considering the available ground-truth that the multi-scale CRF-based strategy is generally more effective for classification compared to state-of-art Bayesian (region and pixel) approaches.

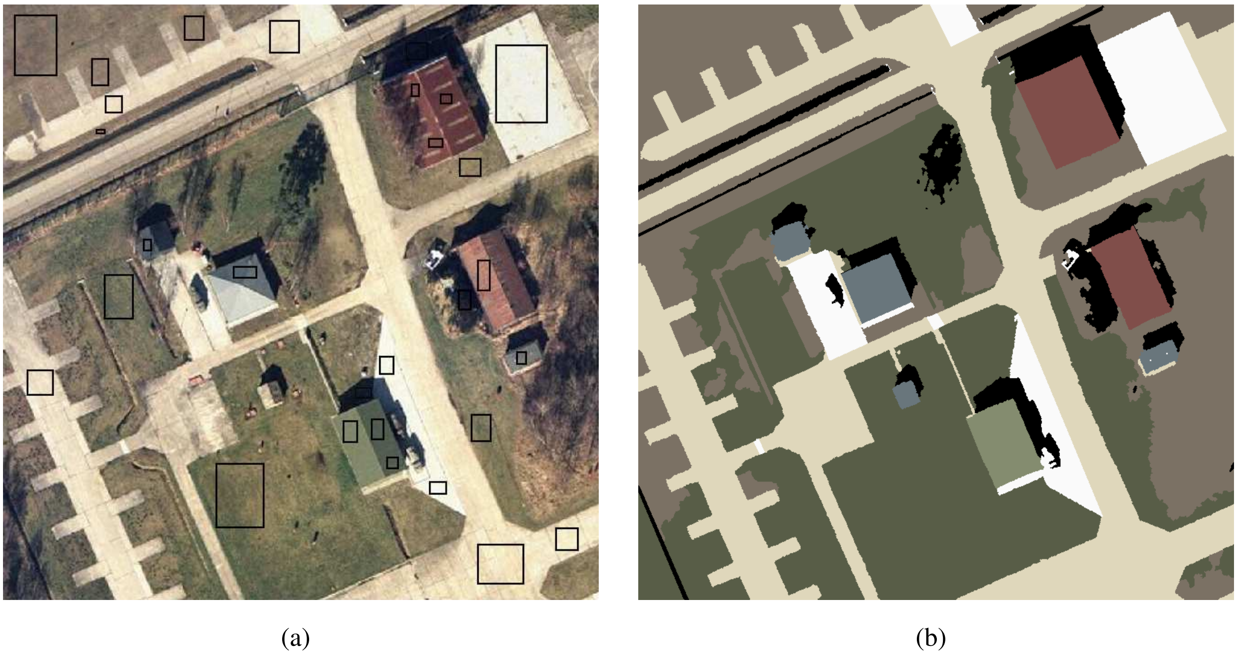
5.3. Airborne RGB Color Image

In order to asses the effectiveness of the proposed multi-scale CRF-based classification, we considered an areal color (RGB) image, as illustrated in Figure 9a, with nine classes, namely: shadow, dark roads and paths, light roads and paths, grass land, blue rooftop, red rooftop, green rooftop and arid land. We conducted an assessment process consisting of: (i) evaluating the importance of the MSRAG structure, *i.e.*, hierarchical segmentation; and (ii) analyzing the discriminative power of the different models along with the impact of the potentials on the result. For all of the experiments, we used the same training data of Figure 9a.

5.3.1. Segmentation and MSRAG Generation Analysis

The proposed multi-scale CRF-based classification can be applied to any hierarchical decomposition of the image. In this section, we evaluate the classification results using different hierarchical decomposition approaches. Namely, the proposed MSRAG approach of Section 2.2 and a scale-space-based hierarchy, denoted MRAT (multiscale region adjacency tree), as described in [3].

Figure 9. The considered areal color (RGB) image. (**a**) Training samples. (**b**) Testing samples.



The creation of the MRAT is based on the creation of the multi-scale tower using the PDE approach of Section 2.1. Then, at a manually selected localization scale, u_0, a gradient watershed transformation is performed to detect a set of regions with well-localized contours. To create the hierarchy, the regions at the localization scale are tracked across the multi-scale tower, and a parent-child linking [27] process is made. This is achieved by spatially projecting the regional minima at each scale into the coarser one. Using the duality between the regional minima of the gradient and the catchment basins of the watershed, each regional minimum residing in a given scale is linked with a regional minimum in the coarser scale, if and only if its projection falls in the catchment basin of that regional minimum (for more details, refer to [27]). It has to be noted that, in [3], the MRAT has the form of a truncated tree, with each node (apart from the ones at the highest scale) having a unique predecessor (its parent) and not considering intra-scale links (edges denoting the region adjacency at a given scale).

An illustration of the different hierarchical segmentations is provided in Figures 10 and 11 along with the respective classification results. The multi-scale CRF classification applied to the considered multi-scale region adjacency graphs can be compared under visual inspection in Figure 12, and their per-class accuracy is detailed in Table 4.

In our experiments, we considered the following multi-scale region adjacency graphs:

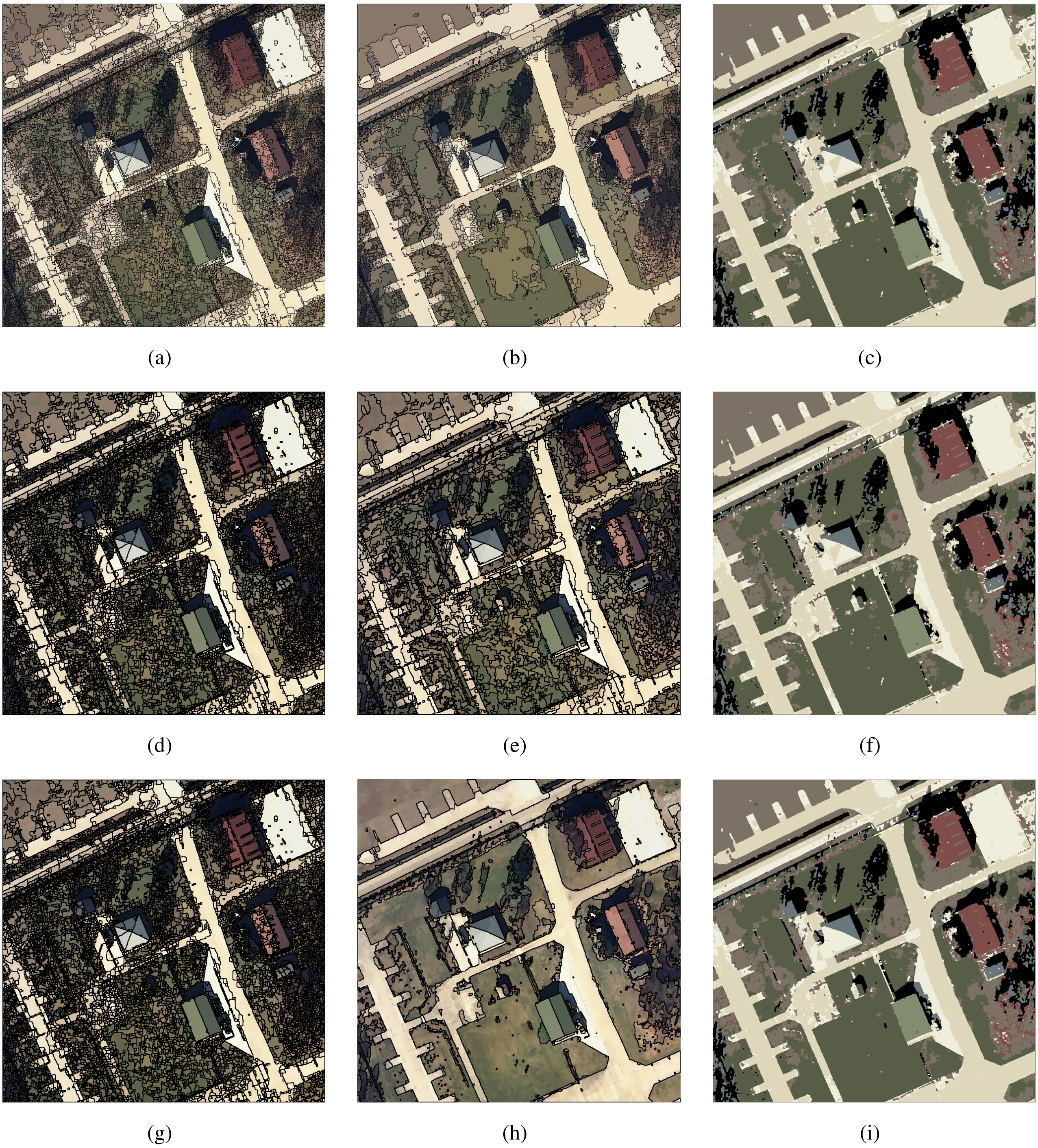
- A-MRAT (augmented MRAT): this being the MRAT produced using the approach of [3], where the localization scale was selected manually, and we added the intra-scale edges/links to conform to our definition of MSRAG (see Figure 2);
- MSRAG A–D: four multi-scale region adjacency graphs obtained starting from the same localization scale used in the A-MRAT and applying the hierarchical segmentation scheme of Section 2.2 with an extra step of small regions merging before the waterfall process, taking into account an expected granularity during the waterfall, with different parameters;
- MSRAG E: this multi-scale region adjacency graph has been obtained by applying (on the original image) the PDE filtering and localization scale selection of Section 2.1 followed by the hierarchical segmentation scheme of Section 2.2.

The resultant classification accuracy is given in Table 5, along with the regions and average region size, per scale and per multi-scale region adjacency graph.

Table 4. Per-class accuracy (%) and overall accuracy (%) of the multi-scale CRF applied to the considered multi-scale region adjacency graphs.

Class	**A-MRAT**	**MSRAG A**	**MSRAG B**	**MSRAG C**	**MSRAG D**	**MSRAG E**
C1-Shadow	83.98	85.47	89.49	92.15	90.52	97.30
C2-Roads/Path (Dark)	79.65	77.52	75.55	87.87	85.36	87.49
C3-Grass land	82.88	79.90	80.19	68.05	80.18	83.22
C4-Rooftop 1 (Blue)	47.00	46.16	48.05	49.30	52.60	49.58
C5-Rooftop 2 (Red)	77.76	79.83	79.88	79.67	77.43	79.23
C6-Rooftop 3 (Green)	96.06	96.06	96.06	97.75	98.25	97.25
C7-Arid land	61.73	64.63	65.62	81.60	77.84	73.86
C8-Roads/Path (White)	71.19	71.86	77.80	74.99	69.49	74.55
Overall Accuracy	75.49	74.92	75.39	78.76	80.52	**81.45**

Figure 10. Scales and classification maps of the augmented multiscale region adjacency tree (A-MRAT), MSRAG A and MSRAG B. The first two columns provide regions at Scales 1 and 9, respectively, while the last column provides the classification map. (**a**) A-MRAT Scale 1. (**b**) A-MRAT Scale 9. (**c**) A-MRAT class-map. (**d**) MSRAG A Scale 1. (**e**) MSRAG A Scale 9. (**f**) MSRAG A class-map. (**g**) MSRAG B Scale 1. (**h**) MSRAG B Scale 9. (**i**) MSRAG B class-map.

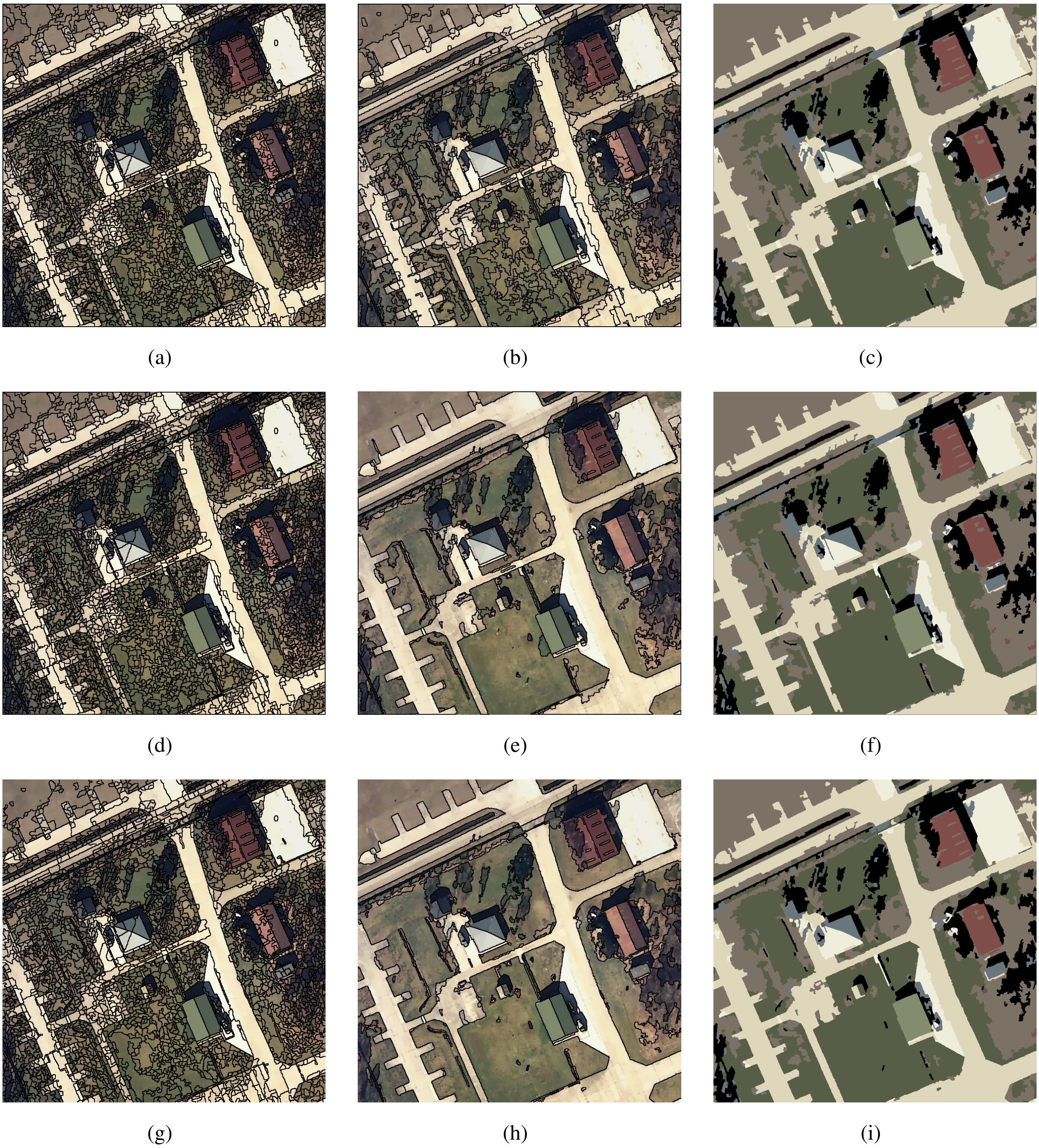


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 11. Scales and classification map of the MSRAG C, MSRAG D and MSRAG E. The first two columns provide regions at Scales 1 and 9, respectively, while the last column provides the classification map. (**a**) MSRAG C Scale 1. (**b**) MSRAG C Scale 9. (**c**) MSRAG C class-map. (**d**) MSRAG D Scale 1. (**e**) MSRAG D Scale 9. (**f**) MSRAG D class-map. (**g**) MSRAG E Scale 1. (**h**) MSRAG E Scale 9. (**i**) MSRAG E class-map.

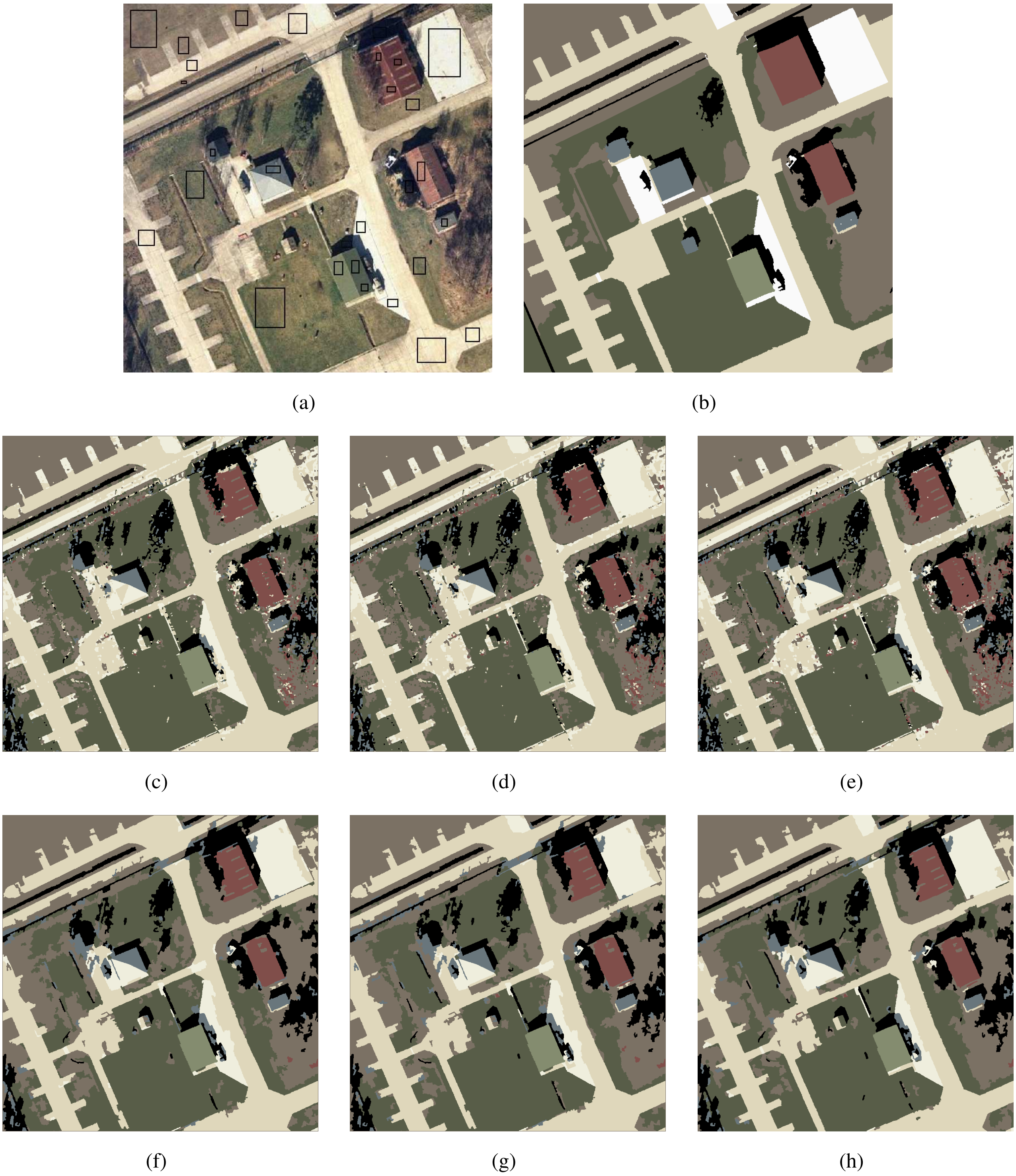


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 12. Different MSRAG generations will provide different classification results. (**a**) Training samples. (**b**) Testing samples. (**c**) A-MRAT. (**d**) MSRAG A. (**e**) MSRAG B. (**f**) MSRAG C. (**g**) MSRAG D. (**h**) MSRAG E.



(a) (b)

(c) (d) (e)

(f) (g) (h)

Table 5. The different hierarchical segmentation strategies provide different numbers of regions, as well as average region sizes (between brackets) per scale, which influence the classification results of the CRF model.

Scale	**A-MRAT**	**MSRAG A**	**MSRAG B**	**MSRAG C**	**MSRAG D**	**MSRAG E**
1	14,870 (43)	14,870 (43)	14,870 (43)	3224 (199)	3224 (199)	3081 (208)
2	12,997 (49)	11,781 (54)	6776 (94)	2499 (256)	1475 (434)	1417 (452)
3	11,481 (56)	9941 (64)	4253 (150)	2073 (309)	911 (703)	886 (722)
4	10,155 (63)	8680 (74)	3150 (203)	1767 (362)	659 (971)	647 (989)
5	8953 (71)	7773 (82)	2523 (254)	1381 (414)	526 (1217)	522 (1226)
6	7926 (81)	7106 (90)	2170 (295)	1241 (463)	440 (1455)	442 (1448)
7	6990 (92)	6557 (98)	1933 (331)	1130 (516)	387 (1654)	383 (1671)
8	6145 (104)	6096 (105)	1761 (363)	1035 (566)	348 (1839)	340 (1882)
9	5397 (119)	5710 (112)	1628 (393)	963 (618)	316 (2025)	303 (2112)
Total	84,914	78,514	39,074	15,313	8286	8021
Accuracy	75.49	74.92	75.39	78.49	80.52	**81.45**

As can be seen, multi-scale region adjacency graphs with a lesser number of regions seem to be better for the proposed multi-scale CRF model, as far as the reduction of the number of regions not incurring the addition of excessive region merging errors on the initial scales. Additionally, there seems to be an influence of the average region's size over the accuracy of the proposed multi- scale CRF model classification results.

It is worth noting that the optimal region's size depends on the size of objects present in the image and naturally varies from one image to another. Moreover, to better characterize the spectral, textural and morphological features of a region, the larger the region is, the better the characterization is. In summary, the experimental results suggest a direct influence of the total number of regions and their average size on the classification results. The obtained results also suggest that the proposed segmentation scheme using the PDE smoothing and localization scale selection of Section 2.1 and the multi-scale segmentation scheme of Section 2.2 produces satisfactory results.

5.3.2. Different Classification Models

For the valuation of the discriminative power of the proposed multi-scale CRF potentials, which involves both intra-scale and inter-scale potentials trained using the above-described max-margin training strategy, we compared it with: (i) a region-based support vector machine (SVM) model; (ii) the MRF classification model of [3]; and (iii) a CRF model using standard potentials.

Classification accuracy results are detailed in Table 6, using the ground-truth of Figure 13b. The classification maps of the MRF and the proposed CRF, using the A-MRAT hierarchy, are illustrated in Figure 13c,d, respectively.

Table 6. Per-class accuracy (%) and overall accuracy (%) of the proposed CRF model considering different MSRAG structures and the region-based multi-scale Markov random fields (MRF) of [3]; the highest per-class accuracy has been underlined. Std. CRF refers to a CRF model with standard energies/potentials.

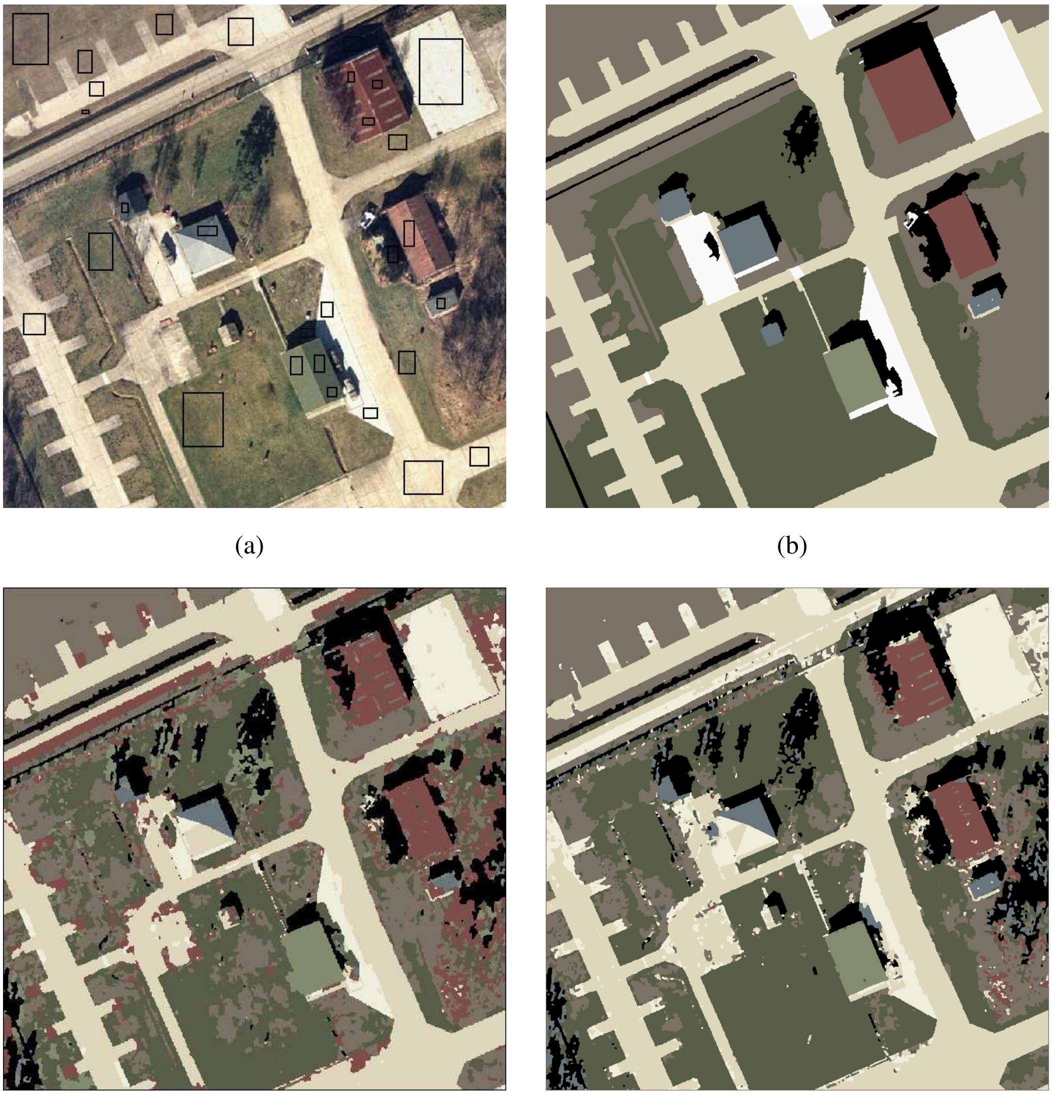
Class	**SVM**	**MRF [3]**	**Std. CRF A-MRAT**	**Proposed CRF A-MRAT**	**Proposed CRF MSRAG E**
C1-Shadow	84.49	85.42	83.45	83.98	97.30
C2-Roads/Path (Dark)	82.98	80.80	64.31	79.65	87.49
C3-Grass land	76.94	60.23	42.54	82.88	83.22
C4-Rooftop 1 (Blue)	48.47	40.32	48.34	47.00	49.58
C5-Rooftop 2 (Red)	80.77	79.93	76.08	77.76	79.23
C6-Rooftop 3 (Green)	00.00	98.74	00.00	96.06	97.25
C7-Arid land	60.99	56.31	54.35	61.73	73.86
C8-Roads/Path (White)	62.21	63.26	68.51	71.19	74.55
Overall Accuracy	72.78	66.88	55.71	75.49	**81.45**

The SVM is a discriminative unstructured model, which has been widely used in remote sensing applications [48]. The related SVM classification is parameterized by a penalty factor C and the kernel parameters. For our tests, we used the LibSVM (available at http://www.csie.ntu.edu.tw/ cjlin/libsvm/) package, a linear kernel $(K(x, x) = \langle x, x \rangle)$ and a hyper-parameter $C = 0.1$. Both training and classification were performed over the finest scale of the hierarchy; each region was described by its color, texture and shape features.

For the MRF, we used the implementation of the multiscale region-based classification of Katartzis *et al.* [3], where the authors proposed a causal Markovian model, defined on the hierarchy of a multiscale region adjacency tree (MRAT). As mentioned above, their proposed MRAT graph has the form of a truncated tree, with each node (apart from the ones at the highest scale) having a unique predecessor (its parent). It has to be noted that this model does not consider intra-scale information. In [3], the observation model is a unary region property that represents the spectral signature of the region at the original image. The parameters were learned using a maximum likelihood approach (using expectation-maximization) and using an efficient exact inference procedure thanks to the tree-like structure of the model.

As standard CRF, we used the Undirected Graphical Models (UGM) matlab code [49]. The model is based on Equation (1) using linear unary and pairwise energies $E(\cdot;\theta) = \theta^T f$. We adapted the implementation to the proposed multi-scale structure and used the max-margin parameters learning strategy of Section 3.3, however, keeping the original standard energies.

Figure 13. The proposed multi-scale CRF exhibits superior classification results compared with MRF for the airborne image. (**a**) Training samples. (**b**) Testing samples. (**c**) Multi-scale MRF (MRAT) [3]. (**d**) Proposed multi-scale CRF (A-MRAT).



(a) (b)

(c) (d)

As can be seen from Figure 13, the MRF model of [3] does not exhibit the salt and pepper effect. This is due to the use of the region-based approach and the long-range dependencies enforced using the multi-scale information. However, it is unable to properly enforce spatial consistency, as it does not model intra-scale dependencies.

From Table 6, the higher performance of the proposed multi-scale CRF model compared to the MRF of [3] has to be related to: (i) the CRF's discriminative nature [7] compared to the generative nature [6,7] of MRF, which is also confirmed when comparing to the SVM results; (ii) the design of the unary and pairwise potentials; and (iii) the type of MSARG.

The highest per-class performance has been underlined in Table 6. It is worth noting that, using the MRAT as defined in [3], MRF is capable of achieving slightly better performance in one of the eight classes compared to the CRF models and in five of eight classes compared to the CRF applied on the

A-MRAT. However, the accuracy improvements of the MRF compared to the CRF models is in the order of 1.5% in the highest case, while the proposed CRF, using MSRAG E, is capable of improving the classification results in the order 22% of the accuracy value for certain classes (*i.e.*, C3). The low overall accuracy of the standard CRF can be due to the parameter settings in our implementation.

5.3.3. The Impact of the Intra- and Inter-Scale Potentials

The above comparison between the MRF and CRF models does not use the same features to characterize the classes. Indeed, for MRF, only the spectral signature of the regions has been used, while, for the CRF, each region was characterized by its spectral, texture and shape information. Therefore, in order to measure the real impact of each individual potential (intra- and inter-scale) in the classification results, we considered different sets of values for the potential combination of the energy expression Equation (33). The obtained results are detailed in Table 7 using the considered areal color image.

The results indicate that for the proposed model, for the considered image and model structure (MSRAG), the multi-scale information is valuable. This can be seen comparing the results that involve multi-scale information (d_2 and d_3 used, with the second column as Yes) with the others that do not involve it (only d_1 used, with the second column as No).

From the potentials' influence point of view, it is clear that their different combinations on the energy expression of Equation (33) lead to different classification results. However, each potential contribution will vary depending on the image nature and on the training samples.

For instance, if certain images exhibit clear class transition patterns and these patterns are present in the training samples, then we expect the intra- scale potential to be influential on the energy expression in order to improve the classification results. If, on the contrary, no clear class transition patterns can be distinguished in the image or if they are not present in the training samples, then we should not expect the intra-scale potential to provide valuable information on the energy expression. A similar analysis can be done for the inter-scale potential.

Table 7. The potentials mixture weights ($d_m, m = 1, \ldots, 3$ in Equation (33)) establish the final performance of the multi-scale CRF classifier.

Solution #	**Multi-Scale Information**	**Unary (d_1)**	**Pairwise**		**CRF Acc**
			Intra-Scale (d_2)	**Inter-Scale (d_3)**	
1	No	1.0	0.0	-	78.30
2	Yes	1.0	0.0	0.0	80.30
3	Yes	1.0	0.0	1.0	80.30
4	Yes	1.0	0.1	1.0	80.52
5	No	1.0	0.3	-	78.87
6	No	6.0	0.3	-	79.17
7	Yes	6.0	0.3	0.0	81.45
8	Yes	6.0	0.3	11.0	81.45

5.3.4. Computational Complexity

All of the reported experiments of Section 5.3.2. have been made on an Intel Core i7 at 3.40 Ghz PC with 8 GB of RAM. Table 8 summarizes the computational costs of the evaluated approaches, considering the same hierarchical segmentation as the input.

Table 8. Execution time for the parameter learning and inference stage of different models: SVM as Support Vector Machines; Prop. CRF as the proposed CRF model; Std. CRF (Max-Margin) as CRF model with standard energies/potentials trained using max-margin approach; Std. CRF (Max-Likelihood) as CRF model with standard energies/potentials trained using max-likelihood approach.

Stage	**SVM**	**MRF**	**Prop. CRF**	**Std. CRF (Max-Margin)**	**Std. CRF (Max-Likelihood)**
Params Learning	-	-	3 s	36 s	>4000 s *
Inference	-	-	340 s	716 s	757 s
Total	<1 s	789 s	343 s	752 s	> 4500 s

* Ended before convergence after 150 iterations using l-BFGS descent (Broyden–Fletcher–Goldfarb–Shanno).

For all evaluated models, we assume that they share the same computational complexity for the hierarchical segmentation stage, which can be more or less efficient (it can vary depending on the used segmentation scheme). As can be seen, SVM applied on the localization scale is the most computational efficient, since it is simple, but cannot benefit from any available contextual information (hierarchy). For the MRF and CRF models, the major difference in computational complexity is due to the parameter learning and inference process for multiclass problems. The computational complexity is dependent on: (i) the parameter learning approach (maximum likelihood or max-margin); and (ii) the inference process according to the type of input structure of the hierarchy (with or without loops). From the theoretical point of view, likewise in [24], the proposed parameter learning is much more efficient, mainly due to the reduction of the number of constraints ($\forall x \neq \hat{x}$ in Equation (23)) to a much less number of constraints ($\forall x_{\mathcal{P}} \neq \hat{x_{\mathcal{P}}}, \forall \mathcal{P} \in \mathcal{T}$ in Equation (38) in Appendix A); but it is also due to the proposed problem decomposition into multiple individual problems that can be solved using the fixed-point method (see [26] for the details on the method's efficiency). The execution time values detailed in Table 8 confirm the suitability of the proposed max-margin parameter learning strategy. The proposed CRF and the standard CRF models, both involving loopy structures and trained using max-margin approach, perform similar to the MRF with a loopy-free, tree-like structure. The proposed CRF also largely outperforms the standard CRF model when trained in a traditional maximum likelihood fashion. For the CRF models, the reported execution time for the inference process accounts for a fixed set of values for $d_1, \ldots, d_5$.

6. Conclusions

The contributions of this paper are a multi-scale CRF model with novel energies along with an efficient parameter learning strategy, where the complex learning problem is decomposed into simpler, individual multi-class sub-problems.

The paper described the main principles of our method and illustrated classification results on a set of remote sensing images, together with qualitative and quantitative comparisons with pixel-based techniques, region-based approaches that follow a Bayesian–Markovian framework, either on hierarchical structures (MRF) or the original image lattice, region-based SVM and standard CRF models. The extensive experimental results validated the suitability of the model for region-based image labeling.

We found that there is a relationship between the obtained segmentation and the classification process. The performed experiments suggested that this influence is related (although not exclusively) to the created hierarchy (parent-child link), the total number of regions and their average size, as well as the features used.

In the future, we will be investigating how to pose the per-clique weight (d) and potential (θ) parameter learning as one joint optimization problem and how to solve it efficiently. We are also planning to compare the model accuracy and performance with respect to other (more recent) state-of-the-art models in other (not necessarily aerial) type of images (e.g., [2]).

Author Contributions

Mitchel Alioscha-Perez designed and implemented the methodology, conducted the experiments and analysis with guidance from Hichem Sahli, who defined the energy terms. The manuscript was written and revised by Mitchel Alioscha-Perez. Hichem Sahli was responsible for the research design and contributed to editing and review of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix

A. Decomposition of QP

Starting from the problem (23), recalling from Equations (21) and (22) that $E_{\mathcal{P}}$ and $\hat{E_{\mathcal{P}}}$ are dependant on θ, and considering Equation (25), we obtain the following optimization problem:

$$\max_{\lambda>0,\theta} \quad \lambda \tag{34}$$

$$\textit{s.t.} \quad \sum_{\mathcal{P}\in\mathcal{T}} \left(E_{\mathcal{P}} - \hat{E_{\mathcal{P}}} \right) \geq \sum_{\mathcal{P}\in\mathcal{T}} \lambda \ell_{\mathcal{P}} \qquad \forall x \neq \hat{x} \tag{35}$$

$$\|\theta\|_2 \leq 1$$

From Corollary 1 (see Appendix B), under the maximum margin criteria, we can redefine constraints (35) as a set of $\mathcal{P}$ constraints $E_{\mathcal{P}} - \hat{E_{\mathcal{P}}} \geq \lambda \ell_{\mathcal{P}}, \forall \mathcal{P} \in \mathcal{T}$, as follows:

$$\max_{\lambda > 0, \theta} \quad \lambda \tag{36}$$

$$\textit{s.t.} \quad E_{\mathcal{P}} - \hat{E_{\mathcal{P}}} \geq \lambda \ell_{\mathcal{P}} \qquad \forall x \neq \hat{x}, \forall \mathcal{P} \in \mathcal{T} \tag{37}$$

$$\|\theta\|_2 \leq 1$$

Considering a model partition (*i.e.*, Piecewise)) [20,21], labels can be structured as $\forall x \neq \hat{x} \propto \forall x_{\mathcal{P}} \neq \hat{x_{\mathcal{P}}}, \forall \mathcal{P} \in \mathcal{T}$; then:

$$\begin{aligned} \max_{\lambda > 0, \theta} \quad & \lambda \\ \textit{s.t.} \quad & E_{\mathcal{P}} - \hat{E_{\mathcal{P}}} \geq \lambda \ell_{\mathcal{P}} \qquad \forall x_{\mathcal{P}} \neq \hat{x_{\mathcal{P}}}, \forall \mathcal{P} \in \mathcal{T} \\ & \|\theta\|_2 \leq 1 \end{aligned} \tag{38}$$

Transforming to eliminate λ [50], we obtain the following QP:

$$\begin{aligned} \min_{\theta} \quad & \frac{1}{2}\|\theta\|_2^2 \\ \textit{s.t.} \quad & E_{\mathcal{P}} - \hat{E_{\mathcal{P}}} \geq \ell_{\mathcal{P}} \qquad \forall x_{\mathcal{P}} \neq \hat{x_{\mathcal{P}}}, \forall \mathcal{P} \in \mathcal{T} \end{aligned} \tag{39}$$

Taking into account the per-potential type ($\mathcal{P}$) block structure of θ, the equivalence $\|\theta\|_2^2 = \sum_{\mathcal{P}} \|\theta_{\mathcal{P}}\|_2^2$ [51] leads us to:

$$\min_{\theta} \quad \frac{1}{2}\sum_{\mathcal{P}} \|\theta_{\mathcal{P}}\|_2^2 \tag{40}$$

$$\textit{s.t.} \quad E_{\mathcal{P}} - \hat{E_{\mathcal{P}}} \geq \ell_{\mathcal{P}} \qquad \forall x_{\mathcal{P}} \neq \hat{x_{\mathcal{P}}}, \forall \mathcal{P} \in \mathcal{T} \tag{41}$$

where the problem decouples as:

$$\forall \mathcal{P} \in \mathcal{T} \begin{cases} \min\limits_{\theta_{\mathcal{P}}} & \frac{1}{2}\|\theta_{\mathcal{P}}\|_2^2 \\ \textit{s.t.} & E_{\mathcal{P}} - \hat{E_{\mathcal{P}}} \geq \ell_{\mathcal{P}}, \quad \forall x_{\mathcal{P}} \end{cases} \tag{42}$$

Note that for the case of $x_{\mathcal{P}} = \hat{x_{\mathcal{P}}}$, Constraint (41) is met, since both sides of the inequality ($E_{\mathcal{P}} - \hat{E_{\mathcal{P}}}$ and $\ell_{\mathcal{P}}$) become zero; therefore, we considered $\forall x_{\mathcal{P}}$ instead of $\forall x_{\mathcal{P}} \neq \hat{x_{\mathcal{P}}}$.

Let us define $\Omega_{\mathcal{P}}$ as the set of label combinations for any clique-order $\mathcal{P}$ and $N_{\mathcal{P}} = |\mathcal{P}|$ the total number of (related) factors [21]; in our problem, $\Omega_{\mathcal{P}=\text{site}} \in \{1, \ldots, L\}$, $\Omega_{\mathcal{P}=\text{neighbor}} \in \{0, 1\}$ and $\Omega_{\mathcal{P}=\text{parent/child}} \in \{0, 1\}$. For each i-th factor ($i = \{1, \ldots ., N_{\mathcal{P}}\}$), we associate a possible label $x_i \in \Omega_{\mathcal{P}}$ with the corresponding observation y_i and the corresponding true label $\hat{x}_i$ from the training data:

$$\forall \mathcal{P} \in \mathcal{T} \begin{cases} \min\limits_{\theta_{\mathcal{P}}} & \frac{1}{2}\|\theta_{\mathcal{P}}\|_2^2 \\ \textit{s.t.} & E_{\mathcal{P}} - \hat{E_{\mathcal{P}}} \geq \ell_{\mathcal{P}}, \ \forall_{i=1}^{N_{\mathcal{P}}}, \forall x_i \in \Omega_{\mathcal{P}} \end{cases} \tag{43}$$

The introduction of slack variables ξ will allow one to assume errors on the obtained separating hyperplane (soft margin):

$$\forall \mathcal{P} \in \mathcal{T} \begin{cases} \min\limits_{\theta_{\mathcal{P}}, \xi_i} & \frac{1}{2}\|\theta_{\mathcal{P}}\|_2^2 + C\sum_i \xi_i \\ \textit{s.t.} & E_{\mathcal{P}} - \hat{E_{\mathcal{P}}} \geq \ell_{\mathcal{P}} - \xi_i, \quad \forall_{i=1}^{N_{\mathcal{P}}}, \forall x_i \in \Omega_{\mathcal{P}} \end{cases} \tag{44}$$

B. Necessary and Sufficient Conditions for Maximum Margin

Proposition 1: Constraints $x_n \geq l_n \forall n$ are sufficient (although not necessary) conditions for the constraint $\sum x_n \geq \sum l_n$ to be satisfied.

Proof: This is a property of inequalities in algebra. Let $\{x_n, l_n\}_{n=1}^{N}$ be a set of arbitrary real numbers satisfying that $x_n \geq l_n \forall n$:

$$\begin{aligned}
x_1 \geq l_1 &\implies x_1 + l_2 \geq l_1 + l_2 \\
x_2 \geq l_2 &\implies x_1 + x_2 \geq l_1 + l_2 \\
\cdots & \\
x_N \geq l_N &\implies x_1 + \cdots + x_N \geq l_1 + \cdots + l_N \\
&\implies \sum x_n \geq \sum l_n \qquad (45)
\end{aligned}$$

Proposition 2: The constraint $\sum x_n \geq \sum l_n$ admits a representation of the form $\sum_{u \in U} x_u \geq \tilde{l}$, where $U = \{n \mid x_n < l_n\}$ are indexes related to unsatisfied constraints, $U \subseteq A, A = \{1, \ldots, N\}$ and $\tilde{l}$ is a slack variable.

Proof: Let $S = A \setminus U$ be the relative complement of U in A. Let us consider the residual $(x_p - l_p)$, where $x_p = \sum_{s \in S} x_s$ and $l_p = \sum_{s \in S} l_s$, part of the following slack definition:

$$\tilde{l} = \sum_{u \in U} l_u - (x_p - l_p) \qquad (46)$$

then:

$$\begin{aligned}
\sum_{u \in U} x_u \geq \tilde{l} &\iff \sum_{u \in U} x_u \geq \sum_{u \in U} l_u - \sum_{s \in s} x_s + \sum_{s \in S} l_s && \text{due to Equation (46)} \\
&\iff \sum_{n \in A} x_n \geq \sum_{n \in A} l_n && \text{since } A = S \bigcup U
\end{aligned}$$

Theorem 1: Let $x \in \mathbb{R}^N = \{x_n \mid \forall n \in A\}$ be a solution of a problem P satisfying the constraint $\sum x_n \geq \sum l_n, l_n \in \mathbb{R}$, where $A = \{1, \ldots, N\}$. Let $M : \mathbb{R}^N \rightarrow \mathbb{R}, M(x) = \sum x_n - \sum l_n$ be a solution margin of P associated with x, $U = \{n \mid x_n < l_n\}$ and $S = A \setminus U$ be the relative complement of U in A. Any solution margin $M_{U \neq \varnothing}$ of P, such that $U \neq \varnothing$, will be bounded above by another solution margin $M_{U = \varnothing}$ of P, where $U = \varnothing$.

Proof:

$$M = \sum_{s \in S}(x_s - l_s) + \sum_{u \in U}(x_u - l_u) \qquad \text{where } (x_s - l_s) \geq 0 \text{ by definition of S}$$

$$\text{if} \quad U = \varnothing$$
$$\text{then} \sum_{u \in U}(x_u - l_u) \text{ vanishes (zero)}$$

$$\text{if} \quad U \neq \varnothing \iff x_u - l_u < 0 \quad \forall u \in U$$
$$\text{then} \sum_{u \in U}(x_u - l_u) < 0$$

therefore, $M_{U \neq \varnothing} < M_{U = \varnothing}$.

Corollary 1: For any candidate solution x of a problem P, the constraints $x_n \geq l_n \forall n$ are: (i) sufficient conditions to satisfy the constraint $\sum x_n \geq \sum l_n$; and (ii) necessary and sufficient conditions to maximize the margin $\sum(x_n - l_n)$.

Proof:

Proposition 1 is the proof of (i), while Proposition 2 confirms the possible existence of solutions x that do not necessarily satisfy $x_n \geq l_n \forall n$. However, a consequence of Theorem 1 is that only solutions that satisfy $x_n \geq l_n \forall n$ ($U = \varnothing$) will guarantee the maximum margin $M = \sum(x_n - l_n)$, which proves (ii).

C. Single and Pairwise Energy of Segments

- Area

$$E_{area}(r_i) = 0.002 \cdot \frac{N \cdot M}{Area(r_i)} \tag{47}$$

where $N \times M$ is the image size and $Area(r_i)$ is the area of the region. The 0.002 term means that regions with a 0.2% area of the whole image have energy 1.0.

- Convexity

$$E_{\text{conv}}(r_i) = \frac{Area(ConvexHull(r_i))}{Area(r_i)} - 1.25 \tag{48}$$

represents the region convexity energy. We assume that regions with convexity larger than 1.25 are not preferred, and those with convexity energy smaller than 1.25 are desired. Therefore, the offset for the convexity energy is set to -1.25.

Compactness

$$E_{\text{comp}}(r_i) = \frac{Perimeter(r_i)^2}{4\pi Area(r_i)} - 1.25 \tag{49}$$

represents the region compactness energy.

The compactness energy $\frac{Perimeter(r_i)^2}{4\pi Area(r_i)}$ is always greater than or equal to one (one for a circle, $4/\pi$ for a square). We assume that regions with compactness larger than 1.25 are not preferred. Again, the offset for compactness energy is set to -1.25.

Homogeneity

$$E_{\text{hom}}(r_i) = 1 - \sigma(r_i)V(r_i) \tag{50}$$

represents the region's intensity (I) homogeneity. Homogeneity is largely related to the local information extracted from an image and reflects how uniform a region is.

Color variance

$$E_{\text{var}_c}(r_i) = \frac{1}{15}\sigma_c(r_i) \tag{51}$$

represents the color homogeneity of a region, with $\sigma_c(r_i)$ the standard deviation of the color $c \in \{L, a, b\}$ within region r_i. The normalization factor for color variances is derived from statistical analysis of the color variance results on the image data base [52].

Color contrast

$$E_{\text{CMDif}}(r_i, r_j) = \sqrt{(\Delta\mu_L(r_i, r_j))^2 + (\Delta\mu_a(r_i, r_j))^2 + (\Delta\mu_b(r_i, r_j))^2} - T_d \tag{52}$$

represents the normalized mean color difference. The normalization factor T_d (for color merging) is estimated as $T_d = \mu_d - \sigma_v$, with μ_d the mean of the color differences D_is, and $\sigma_v = \sqrt{1/n \sum_{i=1}^{n}(D_i - \mu_d)^2}$ the standard deviation of the $n = \frac{k(k-1)}{2}$ color differences between the k regions.

Edgeness $E_{\text{edge}}(r_i, r_j)$ represents the dynamics of the contour of the edge between r_i and r_j.

References

1. Reynolds, J.; Murphy, K.P. Figure-ground segmentation using a hierarchical conditional random field. In Proceedings of the 2007 Fourth Canadian Conference on Computer and Robot Vision, Montreal, QC, Canada, 28–30 May 2007; pp. 175–182.
2. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning hierarchical features for scene labeling. *IEEE Trans. Patt. Anal. Mach. Intell.* **2013**, *35*, 1915–1929.
3. Katartzis, A.; Vanhamel, I.; Sahli, H. A hierarchical Markovian model for multiscale region-based classification of vector-valued images. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 548–558.
4. Besag, J. On the statistical analysis of dirty pictures. *J. R. Stat. Soc. Ser. B* **1986**, *48*, 259–302.
5. Geman, S.; Geman, D. Stochastic relaxation, Gibbs distributions and the bayesian restoration of images. *J. Appl. Stat.* **1993**, *20*, 25–62.

6. Lafferty, J.; McCallum, A.; Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2001; pp. 282–289.
7. Kumar, S.; Hebert, M. Discriminative random fields: A discriminative framework for contextual interaction in classification. In Proceedings of the 2003 Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; pp. 1150–1157.
8. He, X.; Zemel, R.S.; Carreira Perpinan, M.A. Multiscale conditional random fields for image labeling. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July 2004; pp. 695–702.
9. Yang, M.Y.; Förstner, W.; Drauschke, M. A Hierarchical conditional random field model for labeling and classifying images of man-made scenes. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops, Barcelona, Spain, 6–13 November 2011; pp. 196–203.
10. Munoz, D.; Bagnell, J.A.; Vandapel, N.; Hebert, M. Contextual classification with functional max-margin markov networks. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 975–982.
11. Gould, S.; Rodgers, J.; Cohen, D.; Elidan, G.; Koller, D. Multi-class segmentation with relative location prior. *Int. J. Comput. Vis.* **2008**, *80*, 300–316.
12. Su, X.; He, C.; Feng, Q.; Deng, X.; Sun, H. A supervised classification method based on conditional random fields with multiscale region connection calculus model for SAR image. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 497–501.
13. Martinis, S.; Twele, A. A hierarchical spatio-temporal Markov model for improved flood mapping using multi-temporal X-band SAR data. *Remote Sens.* **2010**, *2*, 2240–2258.
14. Kasetkasem, T.; Rakwatin, P.; Sirisommai, R.; Eiumnoh, A. A joint land cover mapping and image registration algorithm based on a Markov random field model. *Remote Sens.* **2013**, *5*, 5089–5121.
15. Vishwanathan, S.V.N.; Schraudolph, N.N.; Schmidt, M.W.; Murphy, K.P. Accelerated training of conditional random fields with stochastic gradient methods. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 969–976.
16. Ratliff, N.D.; Bagnell, J.A.; Zinkevich, M.A. Subgradient methods for structured prediction. In Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AIStats), San Juan, Puerto Rico, 21–24 March 2007; pp. 380–387.
17. Tsochantaridis, I.; Hofmann, T. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.* **2005**, *6*, 1453–1484.
18. Szummer, M.; Kohli, P.; Hoiem, D. Learning CRFs using graph cuts. In Proceedings of the 10th European Conference on Computer Vision: Part II, Marseille, France, 12–18 October 2008; pp. 582–595.
19. Nowozin, S.; Gehler, P.; Lampert, C. On parameter learning in CRF-based approaches to object class image segmentation. In Proceedings of the 11th European Conference on Computer Vision: Part VI, Heraklion, Crete, Greece, 5–11 September 2010; pp. 98–111.

20. Sutton, C.; Mccallum, A. Piecewise training for undirected models. In Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05), Edinburgh, Scotland, 26–29 July 2005; pp. 568–575.
21. Sutton, C.; Mccallum, A. Piecewise pseudolikelihood for efficient training of conditional random fields. In Proceedings of the 24th International Conference on Machine learning, Corvallis, OR, USA, 20–24 June 2007; pp. 863–870.
22. Zhong, P.; Wang, R. Learning conditional random fields for classification of hyperspectral images. *IEEE Trans. Image Process.* **2010**, *19*, 1890–1907.
23. Koltun, V. Efficient inference in fully connected CRFs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst.* **2011**, arXiv:1210.5644.
24. Alahari, K.; Russell, C.; Torr, P.H.S. Efficient piecewise learning for conditional random fields. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 895–901.
25. Kumar, M.P.; Zisserman, A.; Torr, P.H.S. Efficient discriminative learning of parts-based models. In Proceedings of the 2009 IEEE 12th Conference on Computer Vision and Pattern Recognition, Kyoto, Japan, 29 September–2 October 2009; pp. 552–559.
26. Crammer, K.; Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.* **2001**, *2*, 265–292.
27. Vanhamel, I.; Pratikakis, I.; Sahli, H. Multiscale gradient watersheds of color images. *IEEE Trans. Image Process.* **2003**, *12*, 617–626.
28. Catté, F.; Lions, P.L.; Morel, J.M.; Coll, T. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM J. Numer. Anal.* **1992**, *29*, 182–193.
29. Perona, P.; Malik, J. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **1990**, *12*, 629–639.
30. Black, M.; Sapiro, G.; Marimont, D.; Heeger, D. Robust anisotropic diffusion. *IEEE Trans. Image Process.* **1998**, *7*, 421–432.
31. Koenderink, J. The structure of images. *Biol. Cybernet.* **1984**, *50*, 363–370.
32. Vanhamel, I.; Alrefaya, M.; Sahli, H. Multi scale representation for remotely sensed images using fast anisotropic diffusion filtering. In Proceedings of the 2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Honolulu, HI, USA, 25–30 July 2010; pp. 2222–2225.
33. Vanhamel, I.; Mihai, C.; Sahli, H.; Katartzis, A.; Pratikakis, I. Scale selection for compact scale-space representation of vector-valued images. *Int. J. Comput. Vis.* **2008**, *84*, 194–204.
34. Blaschke, T. Object based image analysis for remote sensing. *J. Photogramm. Remote Sens.* **2010**, *65*, 2–16.
35. Geerinck, T.; Sahli, H.; Henderickx, D.; Vanhamel, I.; Enescu, V. Modeling attention and perceptual grouping to salient objects. In *Attention in Cognitive Systems*; Paletta, L., Tsotsos, J.K., Eds.; Springer-Verlag: Berlin, Germany, 2009; pp. 166–182.
36. Marcotegui, B.B.S. Fast implementation of waterfall based on graphs. In *40 Years On Mathematical Morphology*; Springer-Verlag: Berlin, Germany, 2005; pp. 177–186.
37. O'Callaghan, R.J.; Bull, D.R. Combined morphological-spectral unsupervised image segmentation. *Image Process.* **2005**, *14*, 49–62.

38. DiZenzo, S. A note on the gradient of a multi-image. *Comput. Vis. Graph. Image Process.* **1986**, *33*, 116–125.
39. Ladicky, L.; Russell, C.; Kohli, P.; Torr, P.H.S. Associative hierarchical CRFs for object class image segmentation. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September 2009–2 October 2009; pp. 739–746.
40. Kohli, P.; Torr, P. Robust higher order potentials for enforcing label consistency. *Int. J. Comput. Vis.* **2009**, *82*, 302–324.
41. Aronszajn, N. Theory of reproducing kernels. *Trans. Am. Math. Soc.* **1950**, *68*, 337–404.
42. Zhou, H.; Wang, R.; Wang, C. A novel extended local-binary-pattern operator for texture analysis. *Inf. Sci.* **2008**, *178*, 4314–4325.
43. Komodakis, N. Efficient training for pairwise or higher order CRFs via dual decomposition. In proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 20–25 June 2011; pp. 1841–1848.
44. Yedidia, J.; Freeman, W. Understanding belief propagation and its generalizations. In *Exploring Artificial Intelligence in the New Millennium*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2003; pp. 236–269.
45. Wainwright, M. Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching. *Workshop Artif. Intell. Stat.* **2003**, *21*, 97–104.
46. Boykov, Y.; Kolmogorov, V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *Science* **2004**, *26*, 1124–1137.
47. Aksoy, S. Spatial techniques for image classification. In *Signal and Image Processing for Remote Sensing*; CRC Press: Boca Raton, FL, USA, 2006; pp. 491–513.
48. Petropoulos, G.P.; Kalaitzidis, C.; Vadrevu, K.P. Support vector machines and object-based classification for obtaining land-use/cover cartography from Hyperion hyperspectral imagery. *Comput. Geosci.* **2012**, *41*, 99–107.
49. Schmidt, M. UGM: Matlab Code for Undirected Graphical Models. Available online: http://www.cs.ubc.ca/~schmidtm/Software/UGM.html (accessed on 20 April 2011).
50. Taskar, B.; Guestrin, C.; Koller, D. Max-Margin Markov networks. *Adv. Neural. Inf. Process. Syst.* **2004**, *16*, 25–32.
51. Bach, F.R.; Lanckriet, G.R.G.; Jordan, M.I. Multiple kernel learning, conic duality, and the SMO algorithm. In Proceedings of the Twenty-First International Conference on Machine Learning-ICML '04, Banff, AB, Canada, 4–8 July 2004; pp. 6–13.
52. Luo, J.; Guo, C. Perceptual grouping of segmented regions in color images. *Patt. Recogn.* **2003**, *36*, 2781–2792.