An Overview and Comparison of Smooth Labeling Methods for Land-Cover Classification

Konrad Schindler* *Member, IEEE*

***Abstract*—An elementary piece of our prior knowledge about images of the physical world is that they are spatially smooth, in the sense that neighboring pixels are more likely to belong to the same object (class) than to different ones. The smoothness assumption becomes more important as sensor resolutions keep increasing, both because the radiometric variability within classes increases and because remote sensing is employed in more heterogeneous areas (e.g. cities), where shadow and shading effects, a multitude of materials etc. degrade the measurement data and prior knowledge plays a greater role. This paper gives a systematic overview of image classification methods, which impose a smoothness prior on the labels. Both local filtering-type approaches and global random field models developed in other fields of image processing are reviewed, and two new methods are proposed. Then follows a detailed experimental comparison and analysis of the presented methods, using two different aerial datasets from urban areas with known ground truth. A main message of the paper is that when classifying data of high spatial resolution, smoothness greatly improves the accuracy of the result – in our experiments up to 33%. A further finding is that global random field models outperform local filtering methods, and should be more widely adopted for remote sensing. Finally, the evaluation confirms that all methods already over-smooth when most effective, pointing out that there is a need to include more and more complex prior information into the classification process.**

***Index Terms*—remote sensing, computer vision, image analysis, machine learning**

I. Introduction

Classification of image content into semantically defined classes (pixel labeling) is a basic problem of remote sensing with imaging sensors. Since the classes of interest are problem-specific and thus in most cases defined by the user, the prevalent approach is supervised classification. The general process is the following: the user marks example regions for each class, usually referred to as *training data*; each pixel (or small segment) in the training data is represented by a *feature vector* which is computed from the raw intensities following a fixed recipe; the feature vectors and labels then serve as input to some *statistical learning* framework, resulting in a function which maps feature vectors to probabilities (or more generally scores) of belonging to the different classes; finally, if a hard decision is needed the class with the highest probability is selected.

A main conceptual limitation of the described approach is that each pixel (or segment) is treated as independent of all others, although there clearly are correlations between the labels of nearby pixels. The problem becomes worse in images where regions of the same class exhibit strong spectral variations, like for example high-resolution images of urban areas. In such environments independent classification still only reaches limited accuracy, and it becomes particularly important to take into account the dependencies between nearby pixels.

The most basic – and in many cases also the most important – form of dependence is the observation that nearby pixels tend to have the same label (the "smoothness assumption"), and that this tendency is stronger if their observed radiometric intensities are similar ("contrast-sensitive smoothness"). The necessity to enforce label smoothness increases with growing sensor resolution: as the ground sampling distance (GSD) decreases, smaller details become visible, increasing the spectral variability within a class.

Formally speaking, under the smoothness assumption the individual pixels are not independent random variables, but form a *random field*, hence their joint likelihood cannot be factorized into per-pixel decisions [1]. Unfortunately, exact inference in such random fields – i.e. finding the globally most likely configuration of labels under smoothness (or similar) constraints – is computationally intractable (NP-hard for >2 classes). While the problem has long been recognized in remote sensing, little systematic work exists about how to mitigate it.

In related fields like medical imaging and image restoration, there are two main strategies. The first one is to model only short-range interactions between pairs of immediate neighbors (where neighborhood is typically defined as a 4- or 8-neighborhood in the pixel grid, or as Voronoi neighborhood in an irregular set of segments). For such models, efficient inference algorithms exist to approximately maximize the joint posterior likelihood. The most prominent examples are *graph cuts* [2], widely used in computer vision and image processing, e.g. [3]; and *message passing* algorithms [4], [5], [6] used in various branches of signal processing and pattern recognition, e.g. [7], [8]. Another popular algorithm for approximate inference in random fields is *semi-global labeling*. That method has so far been used exclusively for image matching in computer vision and photogrammetry under the name of *semi-global matching* [9]. In the present paper it is adapted for the first time to the classification problem.

The second strategy is to model dependencies over bigger neighborhoods.[1] In such higher-order random fields the above global inference algorithms quickly reach their limits, and for realistic image sizes one has to resort to locally smoothing

*Photogrammetry and Remote Sensing Group, ETH Zürich, 8093 Zürich, Switzerland

[1]Note that such models can in principle always be reduced to models with only pairwise dependencies using auxiliary variables, e.g. [10], [11], but in most cases that reduction is of merely theoretical interest, since the problem size grows exponentially.

the per-pixel likelihoods based only on nearby values. Such an inference scheme based on a combination of values in a pixel's neighborhood is equivalent to a (usually non-linear) spatial filtering.

Sensible candidates include the wide-spread *majority filter* (e.g. [12]); *Gaussian smoothing* of the per-class probabilities to isotropically propagate information to the neighborhood; and *bilateral filtering* [13], so that the spatial distribution of the class probabilities is taken into account. Additionally, an *edge-aware filter* is proposed in the present paper, which is a variant of the bilateral filter that takes into account the radiometric similarity of neighbors.

The goal of this paper is a systematic evaluation and comparison of the mentioned algorithms for pixel-wise image classification under the smoothness assumption. It is demonstrated that any type of smoothness prior is beneficial in terms of classification accuracy w.r.t. ground truth, with the best results obtained by graph cuts, followed by semi-global labeling. The gains are significant – in our experiments up to 33% in classification accuracy (κ).

II. Methods

This section briefly reviews the methods and algorithms employed in the subsequent evaluation. For details please refer to the original publications or textbooks, as cited in the corresponding subsections.

A. Per-pixel classification

No matter how the final inference is performed one first has to estimate for every pixel the probability of belonging to each of the possible classes (in random field terminology the "unary potentials" or for short "unaries", see below). A wide variety of classifiers for that purpose exist, and any one of them can be used in conjunction with the smoothing methods investigated in the present work. The only condition is that the classifier delivers for each class a probability that the test sample belongs to that class. For generative classifiers this requirement is fulfilled by construction; for discriminative methods, it can in most cases be met by mapping the test samples' distances from the decision boundaries to pseudo-probabilities.

Obviously the results of all tested methods will depend on the classifier used to obtain the per-pixel probabilities. In this study two different classifiers are employed: on one hand *random forests (RFs)* as a representative of modern discriminative methods, and on the other hand the classical (Gaussian) *maximum-likelihood (ML)* method.

Random forests [14] serve as main classifier in the evaluation, as an example of a state-of-the-art discriminative method. Random forests, i.e. ensembles of randomized decision trees, are efficient to evaluate, and their good classification performance has been confirmed in many studies in remote sensing as well as image processing in general, e.g. [15], [16], [17]. A random forest is an ensemble of many decision trees, which have been trained on randomly selected subsets of the training data and/or with some randomization in the choice of decision functions for the individual nodes, in order to de-correlate the individual trees. Each tree yields a conditional probability distribution over the possible class labels given the observed data, and these distributions are then averaged over all trees to regularize the classifier and prevent over-fitting. Random forests have the attractive properties that they are efficient to evaluate for a complex non-linear classifier, and that they are inherently suitable for multi-class problems. In the author's experience the method is representative of modern discriminative classification and other popular alternatives such as support vector machines [18] or AdaBoost [19] would yield quite similar results on average (although there may of course be differences on individual datasets). This is also confirmed by other researchers, e.g. [20], [21].

As a representative of the classical methods widely used in operational remote sensing pipelines, the classical maximum likelihood classifier (e.g. [12]) is also tested, i.e. a simple generative model which fits a multivariate Gaussian distribution to the training samples of each class, and evaluates the class probabilities of a test sample under each class model. It can be shown that the decision boundary between any two classes in such a model is a quadratic function, which is why the method is commonly referred to as *quadratic discriminant analysis* in statistics and machine learning texts, e.g. [22].

B. Preliminaries and terminology

Before moving on to smooth labeling methods this section introduces some notation and terminology. As usual, the pixel values of an image with k channels are viewed as samples of a non-parametric function $I : \mathbb{R}^2 \to \mathbb{R}^k$. The number of pixels is denoted by n, and individual pixel locations are referred to by two-dimensional vectors, denoted with lowercase bold letters $\mathbf{x}$. The aim of classification is to assign each image pixel one of l possible class labels c_i, to obtain a new single-channel image, the thematic map $C : \mathbb{R}^2 \to \{c_1 \dots c_l\}$. Finding the thematic map with the highest probability amounts to searching the labeling which maximizes the probability $P(C|I) \sim P(I|C)P(C)$, respectively minimizes its negative log-likelihood or "energy",

$$\begin{aligned} -\log P(C|I) &= -\log P(I|C) - \log P(C) + \text{const.}\,, \\ E(I,C) &= E_{\text{data}}(I,C) + E_{\text{smooth}}(I,C)\,. \end{aligned} \tag{1}$$

The energy consists of two parts: a "data term" which describes how likely a certain label is at each pixel given the observed data, and decreases as the labeling fits the observed data better; and a "smoothness term" which describes the likelihood of a certain label configuration, and decreases as the labeling gets smoother.

Without a smoothness prior the second term vanishes and classification decomposes into per-pixel decisions which can be taken individually, $P(C|I) \sim P(I|C) = \prod_n P(I(\mathbf{x})|C(\mathbf{x}))$, respectively $E(I,C) = \sum_n E(I(\mathbf{x}), C(\mathbf{x}))$. For convenience of notation the unary potential at a specific pixel and class label, $E(I(\mathbf{x})|C(\mathbf{x}) = c_i)$, will be abbreviated as $E(\mathbf{x}, c_i)$.

If smoothness is included, the labels at different locations $\mathbf{x}$ are no longer independent, but form a *random field*. The energy of a given pixel depends not only on its data $I(\mathbf{x})$, but also

on the labels of other pixels in its neighborhood or "clique". Since different cliques interact through common pixels, they can no longer be treated independently.

In general, finding the labeling that globally minimizes $E(I, C)$ is intractable – since there is no factorization into smaller problems one would, at least conceptually, have to check all l^n possible labelings.

For random fields with only pairwise cliques (called first-order random fields) efficient approximation methods exist to find good minima, see Sec. III. Such random fields are often represented as graphs: every pixel corresponds to a node with an associated unary potential, and every clique corresponds to an edge linking the corresponding node pair, with an associated pairwise potential.

Random fields with larger cliques have greater modeling power, but the practically relevant optimization schemes are no longer viable. Except for the special case of very sparse higher-order potentials – meaning that almost all label combinations in the clique have the same likelihood, e.g. [11] – one has to resort to optimization within local neighborhoods.

In this paper both strategies mentioned above are tested: on the one hand *filtering methods*, which allow for large cliques (practical sizes range from 25-500), but are not amenable to approximate global optimization. Instead, the optimization is decoupled, either by first labeling without smoothness constraint and then locally smoothing the labels (in the case of the majority filter), or by directly smoothing the unaries before selecting the labels (for the remaining methods).

On the other hand *global methods* are evaluated, which use only a first-order random field, but for which recent advances allow one to find strong (although still not global) minima of the energy over the whole image. In practice, the random fields should be restricted not only to pairwise cliques, but also to low connectivity (i.e. only few neighbors per pixel). Long-range cliques beyond the immediate neighbors in the pixel grid typically bring only marginal improvements at much higher computational cost. Following the mainstream literature the conventional 4- and 8-neighborhoods are used.

It should be noted that labeling based on segments ("object-based classification") [23] can also be interpreted as a smoothness prior, which forces all pixels within a segment to have the same label. This strategy is however not further considered in the present article. The main weakness of classifying segments (or "super-pixels") is that the classification cannot split segments, while it is in no way ensured that the segmentation is actually aligned with the class boundaries, since it is performed before any class information is available. The situation can be alleviated by using multiple putative segmentations, e.g. [24], or hierarchies of segmentations, e.g. [25], nevertheless segmenting into class-aligned regions before one has determined which features actually carry the class information remains a problem. Random fields with sparse higher-order potentials are interesting in this context, since they allow one to take into account segments as soft constraints, through potentials which penalize the assignment of different labels within a segment [26]. This could potentially help to bridge the gap between pixel-based and segment-based labeling.

C. Filtering methods

An obvious way to enforce smoothness in a field of class probabilities is to filter it: a filter kernel is run over the data in a sliding window fashion and the image values inside the window $\mathcal{W}$ are combined in some way to generate the output value for the center pixel.[2] In random field terminology, the pixels which fall inside the filter window at any given image location form a clique, thus the filtering can be seen as an approximation which decouples smoothness from label inference, by locally enforcing the smoothness constraint on the unaries, respectively the labels.

In terms of computational complexity, all filtering approaches share the property that each class likelihood c_i at each pixel $\mathbf{x}$ has to be visited once, thus all approaches have complexity $\mathcal{O}(l \cdot n)$ – for a given label set the computational cost grows linearly with the number of pixels, and even huge remote sensing images can be processed in acceptable time. The following methods have been tested in our evaluation:

Majority filter. This long-standing and popular method has been included as a baseline. It first converts the class probabilities to a label image by assigning each pixel to the most likely class,

$$B(\mathbf{x}) = \arg\min_i \left[E(\mathbf{x}, c_i)\right] \ . \tag{2}$$

Then this "raw label image" B is converted to a final result by taking at each location a majority vote in a local neighborhood,

$$C(\mathbf{x}) = \arg\max_i \Big[\sum_{\mathbf{u}\in\mathcal{W}} \delta\big(B(\mathbf{u}) = c_i\big) \Big] \ , \tag{3}$$

where $\delta(\cdot)$ is the Dirac delta function, such that the sum in the above equation counts the number of pixels inside $\mathcal{W}$ which have class c_i.

Note that the majority filter does not use the original class likelihoods $P(\mathbf{x}, c_i)$. For example, if in a 5×5 neighborhood 13 pixels have a probability of 51% for class *grass* and 49% for class *tree*, and the other 12 pixels have a 99% probability for *tree*, the voting will nevertheless prefer *grass*. There are variants which give pixels closer to the center more voting power, but typically yield similar results.

Gaussian filter. Another obvious route is to apply Gaussian smoothing in image space to the unary potentials for each class, so as to reduce local fluctuations of the per-pixel likelihoods. Denoting the zero-mean Gaussian density function with variance σ^2 by $\mathcal{G}_\sigma(\cdot)$, the output value M is computed by

$$M(\mathbf{x}, c_i) = \frac{1}{Z(\sigma)} \sum_{\mathbf{u}\in\mathcal{W}} E(\mathbf{u}, c_i) \cdot \mathcal{G}_\sigma\big(\|\mathbf{x} - \mathbf{u}\|\big) \ , \tag{4}$$

with a normalization factor Z which ensures the weights sum to 1. Gaussian smoothing corresponds to the assumption that the probabilities for a certain object class change slowly within a neighborhood, and that the correlation depends only on

[2]Note, the discussion is not restricted to linear filtering (convolution). In fact the filters mentioned in the following are not linear, with the exception of Gaussian smoothing.

how far two pixels are from each other. The class label is then decided by picking the class with the highest probability according to the filtered probability maps,

$$C(\mathbf{x}) = \arg\min_i \big[M(\mathbf{x}, c_i)\big] \ . \tag{5}$$

Bilateral filter. A weakness of the Gaussian filter is that it is isotropic, meaning that two neighbors with the same offset from the center pixel have the same influence, independent of their class likelihoods. This results in a strong tendency to blur object boundaries (in the same way that Gaussian smoothing of intensity images blurs radiometric discontinuities). A well-proven way to overcome that behavior is bilateral filtering [13]. In this non-linear filter, a pixel's influence is not only determined by its (Gaussian-weighted) spatial distance to the center pixel, but also by the (also Gaussian-weighted) difference of the function values – in our case the unaries – meaning that neighbors with similar value have more influence than those with very different values,

$$M(\mathbf{x}, c_i) = \frac{1}{Z(\sigma,\tau)} \sum_{\mathbf{u}\in\mathcal{W}} E(\mathbf{u}, c_i) \cdot \mathcal{G}_\sigma\big(\|\mathbf{x}-\mathbf{u}\|\big) \cdot \mathcal{G}_\tau\big(E(\mathbf{x}, c_i) - E(\mathbf{u}, c_i)\big) \tag{6}$$

In this way smoothing is stronger within relatively homogeneous neighborhoods and weak across discontinuities, and boundaries are preserved better. Bilinear filtering is very popular in image editing and image processing [27], [28] and has also been successfully applied to remote sensing images [29], [30]. Very efficient implementations exist [31], notably also in image editing packages like *Adobe Photoshop*. Note that bilateral filtering (as well as edge-aware filtering, see next paragraph) contain Gaussian filtering as a special case when $\tau \to \infty$.

Edge-aware filter. This filter is a variation of bilateral filtering, which has been developed for the present investigation. The underlying idea is the following: having the influence of a pixel depend on *similarity of class probabilities* may not be desirable, since probability maps in many cases do not exhibit sharp boundaries even when the object class changes. Instead, the edge-aware filter goes back to the original image intensities and measures their difference to determine the influence,

$$M(\mathbf{x}, c_i) = \frac{1}{Z(\sigma,\tau)} \sum_{\mathbf{u}\in\mathcal{W}} E(\mathbf{u}, c_i) \cdot \mathcal{G}_\sigma\big(\|\mathbf{x}-\mathbf{u}\|\big) \cdot \mathcal{G}_\tau\big(I(\mathbf{x}) - I(\mathbf{u})\big) \tag{7}$$

For multi-channel images the maximum difference over all channels is used, which turned out to work best. The edge-aware filter favors regions of similar radiometric values to have the same class label, even if the per-pixel classification is uncertain or ambiguous.

III. Global methods

This class of methods uses only a small number of pairwise cliques between neighboring pixels to encode the smoothness, corresponding to a first-order random field with low connectivity. The goal is to maximize the posterior over the entire random field, respectively find the minimum of its negative log-likelihood, $\arg\min_C E(I, C)$, with

$$\begin{aligned} E(I, C) = &\sum_{\mathbf{u}\in\mathcal{I}} E_{\text{data}}\big(\mathbf{u}, C(\mathbf{u})\big) + \\ &+ \lambda \cdot \sum_{\mathbf{u},\mathbf{v}\in\mathcal{N}} E_{\text{smooth}}\big(\mathbf{u}, \mathbf{v}, C(\mathbf{u}), C(\mathbf{v})\big) \ . \end{aligned} \tag{8}$$

Note, in spite of having only connections between neighbors the optimization propagates information over larger distances. The problem is NP-hard, but strong approximate optimization algorithms exist [2], [6], [7].

Here, the two most basic and most popular smoothness priors E_{smooth} are chosen, the *Potts model* and the *contrast-sensitive Potts model*. In the simpler Potts model, two neighboring pixels pay no penalty if they are assigned the same label, and the same penalty for any combination of different labels,

$$\begin{aligned} E_{\text{smooth}}\big(C(\mathbf{u}) = C(\mathbf{v})\big) &= 0 \\ E_{\text{smooth}}\big(C(\mathbf{u}) \neq C(\mathbf{v})\big) &= 1 \ , \end{aligned} \tag{9}$$

where the choice of the constant (here 1) is arbitrary since its contribution to the energy (8) is rescaled by another constant λ.

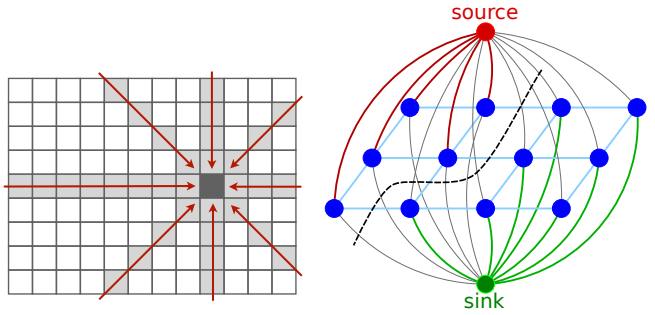
In the contrast-sensitive version, the penalty for a change of label additionally depends on the intensity gradient (contrast) between the two neighbors, but is still 0 for identical labels and equally high for any pairing of distinct labels. The penalty is higher (stronger smoothing) if the pixels have similar intensities, respectively a small intensity gradient. The gradients images $\dot{I}$ are estimated independently for each direction, with Gaussian derivative filters with standard deviation 0.5 pixels,

$$\begin{aligned} &\mathcal{G}_{0.5} \odot \begin{bmatrix} -1 & 1 \end{bmatrix} \qquad \mathcal{G}_{0.5} \odot \begin{bmatrix} -1 \\ 1 \end{bmatrix} \\ &\mathcal{G}_{0.5} \odot \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mathcal{G}_{0.5} \odot \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \ . \end{aligned} \tag{10}$$

For example, the gradient image in horizontal (row) direction is $\dot{I} = \|\mathcal{G}_{0.5} \odot \begin{bmatrix} -1 & 1 \end{bmatrix} \odot I\|$. As the images used in this study have multiple channels, the largest (absolute) gradient over all channels is used (including the height, see Sec. IV). Taking the maximum across channels is a common strategy in gradient-based image processing, e.g. [32], and proved to work well in the present study, too. To map raw gradients to pairwise potentials, a truncated linear potential function has been chosen: the penalty is highest when the gradient is 0, and decreases linearly until at a certain gradient strength it becomes zero. Here that point is chosen at ϕ=70% of the largest gradient $\dot{I}_{\text{MAX}}$ in the whole image. Label changes at strong discontinuities ($>$70% of $\dot{I}_{\text{MAX}}$) do not incurr any penalty:

$$\begin{aligned} E_{\text{smooth}}\big(C(\mathbf{u}) = C(\mathbf{v})\big) &= 0 \\ E_{\text{smooth}}\big(C(\mathbf{u}) \neq C(\mathbf{v})\big) &= \max\big(0, 1 - \frac{2-\phi}{\dot{I}_{\text{MAX}}} \dot{I}_{\mathbf{u}\to\mathbf{v}}\big) \ , \end{aligned} \tag{11}$$

again with an arbitrary maximum of 1, to be rescaled by λ in the energy (8).



(a) semi-global labeling (b) graph cut

Fig. 1: Global optimization methods for random fields. (a) Semi-global labeling sums the minimal energies of several 1-dimensional Markov chains converging at a pixel. (b) Graph cuts convert the optimization to a sequence of minimum cut problems on a graph.

Two optimization algorithms are tested, the well-established α-expansion graph cut method and semi-global labeling, a variant of the semi-global matching method for dense stereo correspondence. Importantly, the computational complexity of these methods is similar to filtering: semi-global labeling has complexity $\mathcal{O}(l \cdot n)$; for α-expansion the empirical complexity for graphs with regular and low connectivity, such as those encountered in image processing, is $\mathcal{O}(l^2 \cdot n)$ [33], so still linear in the image size.

Semi-global labeling. This method is a variant of the popular semi-global matching (SGM) method for dense stereo [9], [34]. SGM has been a big success for image matching. However, in spite of the fact that it is a general-purpose approximation method for random field inference, it has to the author's knowledge not been applied to labeling before. Here, the optimization algorithm of SGM is for the first time applied to labeling under the Potts model, and that approach will be termed "semi-global labeling".

The principle is straight-forward: it is well known that for a 1-dimensional random field I^1 (a Markov chain) the labeling problem can be solved to global optimality by Viterbi decoding [35], i.e. using dynamic programming to compute the label sequence from the beginning to the end of the chain which has the lowest cumulative labeling cost $E(I^1, C)$. Viterbi decoding has long been exploited in image processing by applying it row by row, e.g. [36], however this leads to streaking artifacts because the correlations across rows are disregarded.

Therefore, semi-global labeling uses multiple scanning directions. The costs of all paths converging at a given pixel are summed, and the label with the lowest cumulative cost is selected.

$$\begin{aligned} E(\mathbf{x}, c_i) &= \sum_{\psi \in \text{paths}} E\big(I^1(\psi, \mathbf{x}), c_i\big) \ , \\ C(\mathbf{x}) &= \arg\min_i \big[E(\mathbf{x}, c_i)\big] \end{aligned} \tag{12}$$

The linear scan path to pixel $\mathbf{x}$ from direction ψ is denoted as $I^1(\psi, \mathbf{x})$. Typically eight evenly spaced directions are used, corresponding to a random field with pairwise cliques in an 8-neighborhood. See Fig. 1(a) for an illustration.

Graph cuts. This labeling method is widely used in computer vision and medical image processing, but surprisingly it does not seem to have been adopted for remote sensing yet, with few exceptions [37], [38], [39], [40]. The method is based on the observation that the *binary* labeling problem with only two labels $c \in [0, 1]$ can be solved to global optimality (under certain conditions which the Potts model fulfills). To that end the graph of the random field is augmented with a source and a sink node, which represent the two labels and are connected to all pixels by edges representing the unary potentials. A large additive constant on those terms guarantees that the minimum cut of the augmented graph into two unconnected parts leaves each node connected to only the source or the sink. Computing the minimum cut is equivalent to finding the maximum flow from source to sink, for which fast algorithms exist, e.g. [33]. For an illustration see Fig. 1(b).

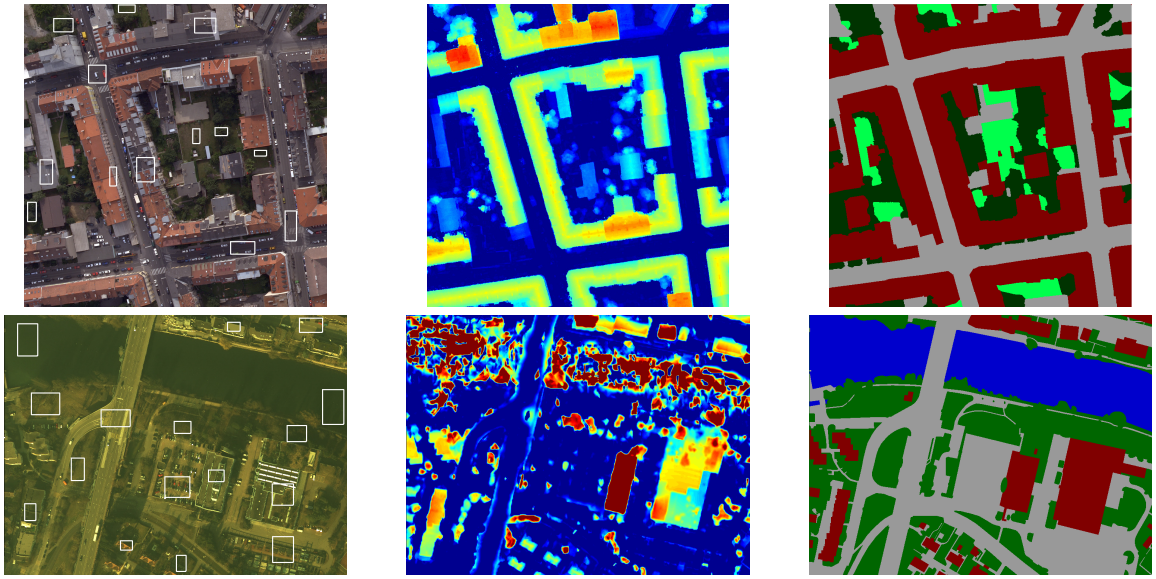
With the binary graph cut as building block, multi-label problems can be solved approximately by the α-expansion method [2]: each label α is visited in turn and a binary labeling is solved between that label and all others, thus flipping the labels of some pixels to α; these expansion steps are iterated until convergence. The algorithm directly returns a labeling C of the entire image which corresponds to a minimum of the energy $E(I, C)$ in Eq.(8).

For graph cuts using cliques in the 4-neighborhood is in most cases sufficient. The 8-neighborhood tends not to perform a lot better, although in theory the higher number of available directions should reduce metrication artifacts. Both versions are tested in the following.

IV. Datasets

Two different datasets are used for the experiments. Both were recorded in the visible spectrum with aerial mapping cameras and show challenging urban areas. In addition to the (pan-sharpened) RGB channels a dense height map was derived by photogrammetric triangulation and filtered to a bare-earth height map. The difference between them (the relative height over ground) was then used as a fourth channel for classification. It is well-documented that using height greatly improves classification in urban areas, e.g. [41], [42]. All processing was fully automatic, without correcting errors in the height models: manually editing the height model is so labor-intensive that arguably one might as well manually classify the data; it thus appears more adequate not to include any manual intervention, and let the classification deal with the height errors.

The first dataset Graz, from the city of Graz/Austria, has 800×800 pixels and shows an urban residential area (see Fig 2). The GSD is $\approx$25 cm. The entire image was manually classified into the four classes {*road, building, grass, tree*}, where roads also include all other sealed areas at ground level. Note, the image was not ortho-rectified, instead the height map was reprojected to the original camera viewpoint to achieve co-registration in the camera coordinate frame. A number of



(a) (b) (c)

Fig. 2: Test data sets GRAZ (top) and ZÜRICH (bottom). (a) Color image, overlaid with training areas used for classifier learning. (b) Height over ground. Gross height errors in water and shadow areas were not corrected. (c) manually annotated ground truth.

rectangular regions were selected as training areas, covering $\approx 3.2\%$ of the image (see Fig. 2).

The second dataset ZÜRICH, from the city of Zürich/Switzerland, has 695×920 pixels and shows a mixed industrial/residential area (see Fig 2). The image has been ortho-rectified with a GSD of 40 cm. The entire image was manually classified into the four classes {*road, building, vegetation, water*}, where roads also include all other sealed areas at ground level and vegetation includes other unsealed areas at ground level. A number of rectangular regions, covering $\approx 6.6\%$ of the image, serve as training data (see Fig. 2).

V. EXPERIMENTAL EVALUATION

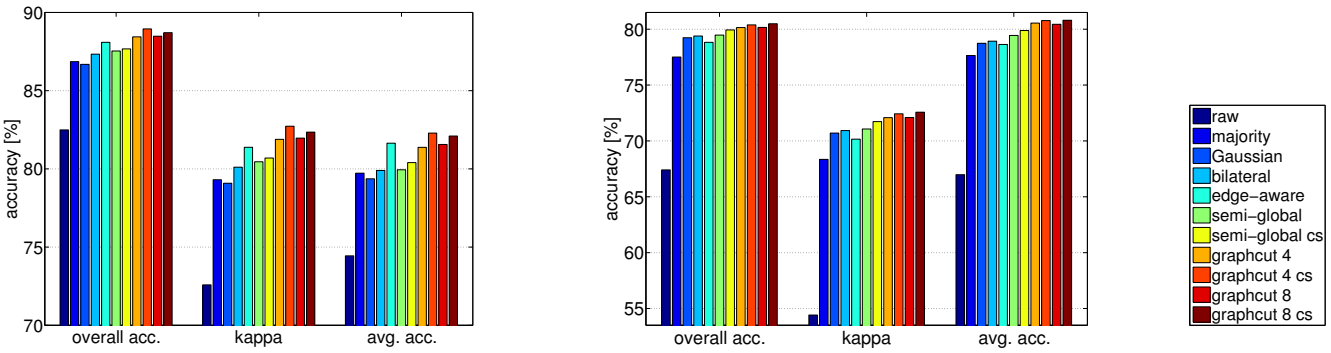
This section presents a detailed quantitative evaluation and comparison of the different smooth labeling schemes listed above. The experimental protocol was designed to guarantee a meaningful comparison between the different smooth labeling methods. Thus, care was taken to use identical inputs (unary potentials) and only vary the smoothing method. The output of independent per-pixel classification serves as baseline. All implementation and processing was done in *Matlab*, and third-party routines were only used when available in source code, to rule out the possibility that undocumented processing steps in black-box implementations influence the results. For each method separately, the optimal parameters were determined (see Sec. V-C), so that their best-case performance is compared.

Any quantitative research about classification faces the problem what metrics to use to measure the generic "goodness" of a thematic map. While full confusion matrices obviously are the most complete approach, they are difficult to compare and rank. It thus seemed more useful to base the tables and figures on scalar measures. The two most frequently used measures are the overall accuracy (total number of correct pixels / image size) and the κ-value (difference to chance). These two are highly correlated for both datasets used here, as is typically the case. On the contrary, the average accuracy (mean fraction of correct pixels per class) gives the same importance to each class irrespective of its pixel count. It thus helps to detect situations in which an overall improvement is achieved at the cost of greatly deteriorating the correctness of less frequent classes. For completeness, user's and producer's accuracies are nevertheless given for the main comparison.

A. Quantitative results

Overall accuracies, κ-values and average accuracies for both datasets are displayed in Fig. 3. The raw random forest classification achieves a κ of 72.6% for the GRAZ data set, and only 54.4% for the more difficult ZÜRICH data set. Note, these values, and all others derived on the basis of random forests, vary slightly over different runs with a standard deviation of $\approx 0.2\%$, as a consequence of the randomness in the classifier. The reported values are means over 20 runs. Their pairwise differences are all statistically significant (pairwise t-test that means differ, 20 samples, significance level 95%), except for the difference between graph cuts with 4- and 8-neighborhood *without* contrast-sensitivity (on both datasets).

It is immediately visible from the figures that all smooth labeling methods employed here significantly improve performance, and that the performance gap between the best and worst smoothed result is smaller than the one from the worst one to the raw results. In other words, the prior assumption of label smoothness is clearly justified for high-resolution images, and at least for the data and class nomenclature in this study, all classes gain from the smoothness assumption –



(a) GRAZ (b) ZÜRICH

Fig. 3: Classification accuracy using different smooth labeling methods, with random forests as data term. Please note the different scales of the y-axis. For the global methods "cs" denotes the contrast-sensitive version.

see user's and producer's accuracies in Tab. I. While there are noticeable differences between different ways of implementing the smoothness prior, any method is preferable to neglecting it.

Going one step further, global methods dominate local filtering, and in particular contrast-sensitive graph cuts consistently give the best results for both datasets on all three performance measures, reaching κ=82.7% for GRAZ, respectively κ=72.6% for ZÜRICH (i.e. relative improvements over the raw per-pixel results of 14%, respectively 33%). On the contrary, the difference between using the 4- or 8-neighborhood is very small. Moreover, the contrast-sensitive version consistently outperforms the simple Potts model in all runs, on average by 0.6% with the 4-neighborhood and by 0.4% with the 8-neighborhood.

It is important to note that graph cuts were selected here as a representative for an entire class of algorithms for approximate global random field inference, and it can be expected with high certainly that message passing methods like loopy belief propagation [7] or tree-reweighted message passing [6] would give practically identical results for the Potts model [43], [44].

Semi-global labeling also performs rather well (κ=80.7% for GRAZ, κ=71.7% for ZÜRICH), although it is dominated by graph cuts. Again the contrast-sensitive version is consistently better, on average by 0.5%. Semi-global labeling may nevertheless be an interesting alternative for certain applications, since it offers some advantages in terms of implementation. Firstly, for problems with many classes graph cuts (as well as message passing) may become too slow, since their complexity grows quadratically with the number of labels. Secondly, and more importantly, they cannot process huge images because they have a rather large memory footprint and cannot be parallelized well. Processing images in (overlapping) tiles offers a viable solution, but requires proper treatment of the overlap areas, which complicates the implementation. To the author's knowledge this paper is the first one to consider semi-global labeling, thus further research is needed to clarify the potential of the method.

The conclusions regarding filtering methods are less clear-cut, it appears that they are more sensitive to the characteristics of individual datasets. The most consistent filtering approach is the bilateral filter (κ=80.1% for GRAZ, κ=70.9% for ZÜRICH). In particular, it outperforms Gaussian filtering in all measures on both datasets, confirming the intuition that anisotropic smoothing is more appropriate. The differences between the two methods are small for the ZÜRICH dataset, which makes sense since they only differ if there are discontinuities in the data, while the class probability maps for the ZÜRICH image are rather diffuse due to the low contrast.

The results for the majority filter and Gaussian smoothing – probably the most widely used smoothing methods in operational remote sensing – are mixed: on the GRAZ image, which is crisp and has good contrast, Gaussian filtering does not work as well as other smoothing methods (κ=79.1%), whereas on the more blurry and low-contrast ZÜRICH image it is competitive with other filtering approaches (κ=70.7%). On the contrary, the majority filter works moderately well on the GRAZ data (κ=79.3%), whereas it is by some margin the weakest method on the ZÜRICH data (κ=68.3%), which has much noisier unaries.

A similar behavior can be observed for the proposed edge-aware filter. On the contrast-rich GRAZ data it is clearly the best filtering method (κ=81.4%), and even competitive with global inference. In particular, it stands out for high average accuracy, i.e. it performs well also for *trees* and *grass*, which are the smaller classes in terms of pixel count (14.5%, respectively 7.0% of the image area). On the contrary, edge-aware filtering does not cope well with the low radiometric quality of the ZÜRICH image, and delivers less correct results than the Gaussian and bilateral filters (κ=70.2%), although it is not as badly affected as the majority filter. Again, the method has not appeared before in the literature in this form, so further research is required to reach well-founded conclusions.

Overall, the author at this stage recommends to use graph cuts (or alternatively, message passing) as smooth labeling algorithm. They should become a standard option in classification toolboxes, and have not yet reached the popularity they deserve in remote sensing, despite their well-documented success in other domains of image processing. For very large problems, semi-global labeling may become a viable

alternative, subject to further evaluation.

If filtering is more appropriate because of performance or implementation issues, the bilateral filter is the best bet. It consistently works as good or better than majority filtering and Gaussian filtering, and has no disadvantages compared to them. Edge-aware filtering appears to be the best alternative in high-quality images with sufficient contrast, but requires further investigation.

B. Qualitative results

Figure 4 shows the classification results for all tested methods on the GRAZ dataset. Beyond the statistics given in the previous section, this section discusses some qualitative observations.

The quantitatively best results with filtering methods are achieved with excessively large filter kernels, a further indication that the raw unaries for high-resolution images are overly noisy and smoothness should be enforced. This is in fact not surprising: at high resolutions the small GSD means that more pixels are needed to cover the same surface area, so that larger neighborhoods of pixels will have the same label, which is precisely the assumption made by smoothness priors. Additionally, at high resolutions there is no longer a simple relation between semantic class and radiometric intensity. The fine structure of a class (shading effects due to the geometry, textures, within-class material differences etc.) becomes apparent, such that pixels within each class can have very different image intensities, leading to more diffuse and ambiguous unary potentials. The strong smoothness prior is thus not only admissible, but also required. Unfortunately, while the large filters yield more correct labels overall, they nevertheless incorrectly distort the class boundaries and cause over-smoothed, "blobby" thematic maps. For the high-contrast GRAZ image the over-smoothing with the quantitatively best-performing setting is largely avoided by the edge-aware filter, at the price of visually somewhat noisy *street* and *building* layers.

The classification results of the global methods SGL and graph cuts are visible crisper, showing both geometrically more correct class boundaries and more homogeneous regions – see Fig. 5. That said, a careful examination reveals that all methods, including the global ones, over-smooth small structures at their best settings.

It appears that this is a limitation of the rather simple Potts-type models which underly all methods compared in this work, not the optimization method used for inference.[3] Although variations of the Potts model (e.g. using different pairwise potentials) may certainly still yield performance gains, the smoothness assumption alone quite clearly ignores a large amount of useful prior knowledge. More complex priors could for example include assumptions about typical shapes and topologies of certain object classes, such as the piecewise straightness and network connectivity of the road system. It is however largely unclear how to perform efficient inference in such models – even relatively simple prior assumptions about relative position, orientation and layout of geometric elements lead to mathematically complex and computationally expensive optimization problems, e.g. [45], [46].

Going further, one may even question what the right "mapping scale" for a given set of classes is, and whether a unique minimum dimension exists, above which objects should be preserved. As an example, it is often desired that cars (width $\approx 1.7\,\mathrm{m}$) are merged into the street layer, whereas in the same application one may not want to merge a $1.2\,\mathrm{m}$ wide bicycle lane into the surrounding vegetation. In a bottom-up processing scheme without a powerful geometric scene model such high-level constraints seem difficult to impose.

C. Sensitivity of parameters

Any smooth labeling method has at least one parameter more than independent per-pixel labeling with the same unaries, namely the strength of the smoothness prior compared to the observation (or in random field terminology the weight of the pairwise term relative to the unary term). Anisotropic methods must have at least one further parameter to characterize the anisotropy. In our case, that is the width τ for the bilateral and edge-aware filters. This section empirically evaluates how these parameters influence the result, respectively how sensitive (and thus how difficult to set) they are. For completeness it should be noted that contrast-sensitive global methods are also anisotropic. The corresponding parameter(s) appear in the mapping from gradients to pairwise potentials e.g. (Eq. 11), but are not further investigated here.

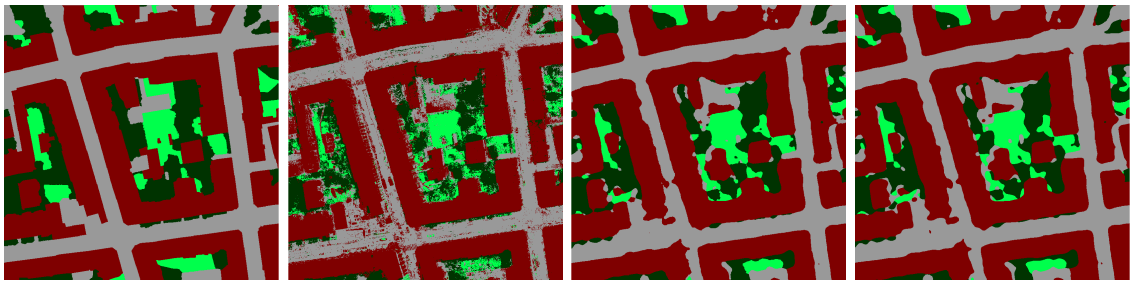
Figure 6 shows the κ-values and average accuracies of different methods over a range of smoothing parameters around the optimal value (the overall accuracy is not shown for clarity, since its values correlate closely with κ). The most important finding from a practical point of view is that all methods are rather stable, meaning that the classification accuracy changes smoothly as a function of the parameter and degrades slowly as one moves away from the optimal setting. That means that there is a relatively wide range over which the correctness of the result is similar, such that choosing a satisfactory value for a given problem (or learning it from data, if enough reference data is available) is easy.

To keep the graphs readable, only three settings are depicted (as separate curves) for the range parameter τ of the bilateral and edge aware filters. Overall, the satisfactory range for that parameter is usually slightly smaller. To understand the influence of that parameter, consider the following extreme cases: a much too low τ will put all the weight on those pixels in the spatial neighborhood, whose value is very similar to the central pixel anyway, and thereby reduce the effect of the filter until ultimately at $\tau = 0$ the filter has no more effect, independent of the kernel size; very high τ will make the weights $\mathcal{G}_\tau$ increasingly diffuse, until ultimately they converge to a uniform distribution, and the filter degenerates to Gaussian smoothing (see Figure 6). Fortunately, the fact that the choice of τ depends directly on the distribution of the input data (i.e. the unaries, respectively the raw intensities), allows one to examine that distribution in order to find the range of sensible values.

[3] This is also confirmed by evidence from other application problems such as stereo or image restoration, where graph cuts and similar optimization schemes regularly find solutions with lower energies than the ground truth [44].

	GRAZ								ZÜRICH							
	building		**road**		**grass**		**tree**		**building**		**road**		**vegetation**		**water**	
	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA
raw	87.5	92.6	80.9	77.6	58.9	57.6	78.2	70.1	68.7	79.0	59.5	70.0	59.7	52.2	82.0	76.4
majority	88.6	95.3	87.2	83.8	72.5	64.6	86.3	75.7	69.7	91.2	85.6	72.6	78.6	58.1	91.3	89.7
Gaussian	89.5	94.5	85.1	84.4	73.6	62.1	84.6	76.6	71.9	90.6	88.9	72.3	76.9	64.8	94.6	87.1
bilateral	90.1	95.1	86.0	85.3	74.0	62.0	85.2	77.4	72.1	90.6	88.8	72.7	77.3	64.9	94.4	87.4
edge-aware	90.6	94.8	88.1	85.1	76.9	63.3	83.4	83.3	69.0	92.1	79.9	77.1	77.5	54.2	93.3	86.1
semi-global	90.5	94.8	86.6	86.3	74.6	59.3	84.1	79.5	70.2	93.1	88.1	74.7	82.5	59.9	93.7	91.2
semi-global cs	90.9	94.7	86.5	86.2	74.6	60.5	84.0	80.3	71.0	92.5	88.7	74.5	81.9	61.4	93.5	91.1
graphcut 4	91.2	95.2	88.2	87.4	75.3	63.6	84.8	83.4	70.7	92.3	87.3	77.0	83.1	61.2	94.2	91.3
graphcut 4 cs	91.8	95.3	88.6	87.4	76.6	63.6	85.1	83.4	70.8	92.1	86.5	76.2	83.8	60.6	94.2	91.5
graphcut 8	91.1	95.1	88.0	87.1	76.6	63.3	85.4	81.6	70.5	92.8	86.5	76.2	83.8	60.6	94.2	91.5
graphcut 8 cs	91.7	95.3	88.7	87.3	75.9	64.4	84.9	82.6	71.5	92.5	87.1	76.3	83.5	62.4	93.8	91.5

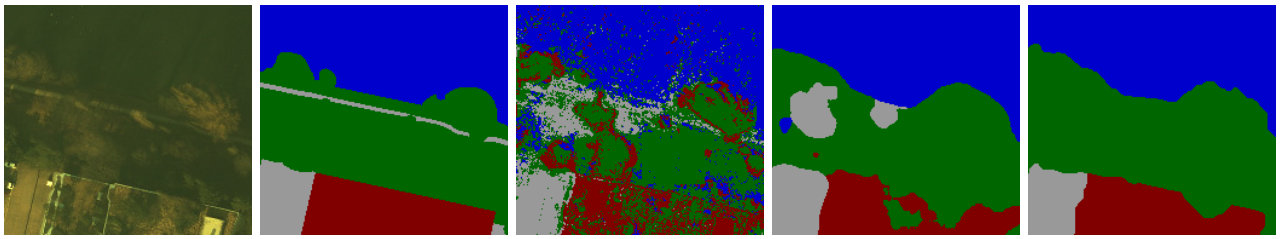
TABLE I: User's and producer's accuracies of different methods (in %). For the global methods the contrast-sensitive version is denoted by "cs".



(a) ground truth (b) raw random forest (c) majority (d) Gaussian

(e) bilateral (f) edge-aware (g) semi-global cs (h) graphcut 4 cs

Fig. 4: Visual comparison for GRAZ. Results for all methods are shown with optimal parameter setting according to the quantitative evaluation. Noteworthy details include the amount of noise in the raw image (b); the tendency of isotropic methods to over-smooth when effective (c,d) and the visibly superior performance of global methods, especially graph cuts (h).



(a) RGB image (b) ground truth (c) raw random forest (d) majority (e) graphcut 4 cs

Fig. 5: Visual comparison for a detail from the ZÜRICH data, showing the impact of smooth labeling. All methods tend to over-smooth in difficult regions where the data is ambiguous, but there are great differences in quality between the worst and best smooth labeling.

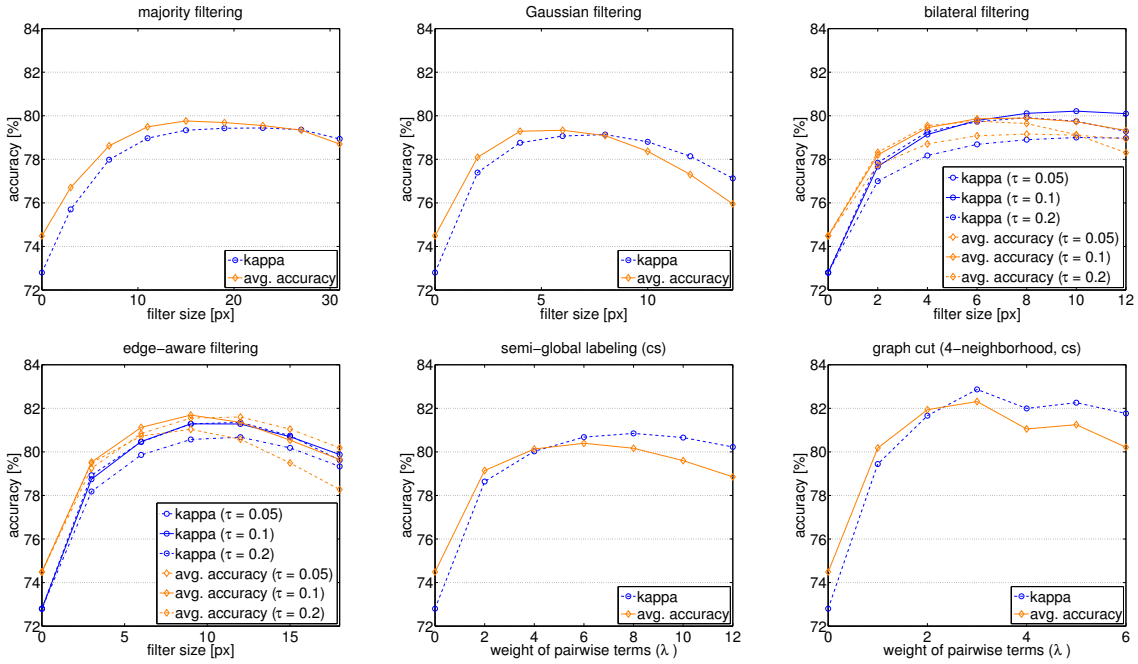
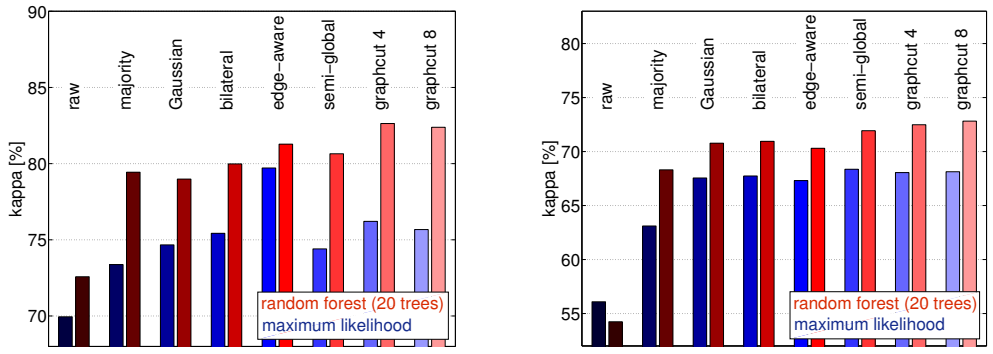


Fig. 6: Dependence of classification accuracy (κ and average) on parameters. Overall, any type of smoothing helps significantly, and all methods are reasonably stable, yielding comparable results over a range of parameter settings (factor 2-3). The (dimensionless) weight for the global methods is given as defined in equation (8), i.e. for a unary weight of 1.



(a) GRAZ (b) ZÜRICH

Fig. 7: Classification accuracy (κ-values) using different classifiers to estimate the per-pixel class probabilities. For semi-global labeling and graph cuts only the contrast-sensitive versions are shown. Please note the different scales of the y-axis.

D. Influence of data term

One may of course ask whether the influence of smooth labeling is the same for different ways of estimating the unary potentials. Since the success of smoothing depends on the unaries, and in particular not only on their ranking, but also one their relative values, some ways of estimating them may be more amenable than others to smoothing.

To that end the quantitative experiments have been repeated, but this time using maximum-likelihood (ML) classification to estimate the unaries, instead of the random forest (RF). It turns out that the differences are significant. Optimal parameter settings for the different methods were determined separately for the two classifiers, to ensure that the smooth labeling is tuned correctly for the specific data term and the comparison remains fair.

On the GRAZ data, RF as such already works noticeably better than ML classification, and the difference between the two methods is amplified by smoothing: usually the gap between them is bigger when using smooth labeling, than it is on the raw per-pixel classification, see Fig. 7(a). There is one exception, namely the edge-aware filter, which combines exceptionally well with ML classification, and achieves the best result with that classifier (although still inferior to RF). A detailed analysis revealed that the edge-aware filter was more successful in cases where the ML unaries are rather ambiguous over larger regions – seemingly the higher-order cliques come

into play here. The effect does not appear with RF, because it delivers less ambiguous unaries. In spite of the exception, RF supports smooth labeling better overall.

Even stronger evidence comes from the ZÜRICH data, see Fig. 7(b): by itself, the ML classifier works a bit better here than RF, but even so RF better supports smooth labeling. It appears that the per-pixel probabilities estimated by ML are much noisier, so that the smoothness prior cannot exploit them as well. Although starting from unary potentials which at the pixel level yield better classification accuracy, the smooth labeling results lag behind those based on RF, with all evaluated methods.

E. Implicit smoothness through larger support regions

A further possible strategy to enforce smoothness is to augment the feature set of every pixel to also include features from neighboring pixels, so that the information about their classification is implicitly available to the per-pixel classifier. In that way, the unaries are expected to already be smoother and yield more correct labels, which would avoid the computational burden of explicit smoothing.[4] There is evidence that in schemes based on super-pixels (small segments) that strategy performs as well as an explicit smoothness prior [47].

To test this strategy, the classification of the raw unaries (computed with the random forest classifier) is repeated, but this time supplying features from increasing neighborhoods to the classifier. Although the additional information does bring small gains, it is not competitive with explicit smoothing for the data used here – see Fig. 8. Moreover, larger support regions even deteriorate the suitability of the unaries for subsequent smooth labeling, as can be seen from the performance of graph cuts.

A closer inspection reveals that this is due to the fact that large support regions have an unwanted side effect on small structures and near object boundaries: in these locations, features from across the object boundaries are included, thus diluting the unaries and aggravating ambiguities. Since high-resolution aerial or satellite images have many small objects and thus also many boundaries, the damage outweighs the – clearly observable – smoothing effect.

VI. CONCLUDING REMARKS

The motivation for the present work has been twofold: first, attempt a systematic overview of classification methods which model spatial smoothness of the labels in a thematic map, and which are potentially relevant for remote sensing imagery. Second, perform an experimental comparison of these methods for the problem of classifying images of high spatial (and consequently low-to-moderate spectral) resolution, and – as far as possible – extract guidelines when and how to use them. Along the way two additional methods have been proposed: the edge-aware filter, which is an extension of the bilateral filter, and semi-global labeling, an adaptation of semi-global matching to the labeling problem.

[4]Note, adding observations from neighboring pixels also gives the classifier access to local texture information, so there is a potential for improved classification even without the smoothing effect.

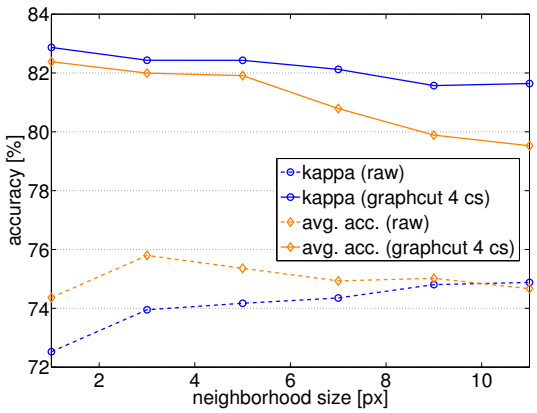


Fig. 8: Dependence of classification accuracy on the window size used for the unaries. Depicted are the accuracies achieved for GRAZ by both raw random forest classification and graph cuts, when feeding the classifier features from progressively larger neighborhoods. Larger regions improve the results, but are no replacement for explicit smoothness, and even impair subsequent smoothing.

To conclude, the main outcomes of the study are summarized. First and foremost, a smoothness prior is imperative to reach high classification accuracy in high-resolution images, since most land-cover classes have strongly varying and partially ambiguous spectral signatures when sampled at small GSD. *Any type of smoothing significantly improves the result when classifying high-resolution images.* Empirically, gains of up to 33% have been observed, but even comparatively weak smoothing methods bring $\approx 10\%$ with suitable unaries. The strongest recommendation is that smoothness should be taken into account – that is more important than how exactly it is done.

Still, there are significant differences between different smoothing schemes. Among the filtering-type methods, the bilateral filter should be preferred over classical filters like majority voting or Gaussian smoothing. The method is well established and understood, and fast algorithms exist. The proposed edge-aware filter has shown potential, but further research is needed to better clarify its strengths and limitations. Other anisotropic filters, which have not been included in the study, may also be applicable and should be investigated.

The global random field methods, which image processing researchers have developed over the past decade, consistently outperform local filtering and should be the method of choice. In particular, graph cuts or an equivalent technique like tree-reweighted message passing should become a standard option of classification toolboxes. The proposed semi-global labeling method does not quite reach the performance of graph cuts, but nevertheless works better than local filters, and scales to almost arbitrary image sizes [9], whereas graph-based methods will inevitably require a subdivision or tiling scheme for huge images in the Giga-pixel range, with the associated complications to guarantee consistency at tile boundaries.

Finally, all methods over-smooth when most effective. This could be expected, but nevertheless provides further evidence

that smoothness alone, while already yielding very significant improvements, is a too simple prior, and more sophisticated image models are required, which can represent more complex a-priori assumptions about the observed world.

REFERENCES

[1] S. Z. Li, *Markov Random Field Modeling in Image Analysis*, 3rd ed. Springer Verlag, 2009.

[2] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[3] L. Grady and G. Funka-Lea, "Multi-label image segmentation for medical applications based on graph-theoretic electrical potentials," in *Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis*, 2004.

[4] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Inc., 1988.

[5] M. J. Wainwright, T. Jaakkola, and A. S. Willsky, "MAP estimation via agreement on trees: message-passing and linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 11, pp. 3697–3717, 2005.

[6] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, 2006.

[7] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," *International Journal of Computer Vision*, vol. 70, no. 1, 2006.

[8] R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng, "Turbo decoding as an instance of Pearl's "belief propagation" algorithm," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 2, pp. 140–152, 1998.

[9] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[10] E. Boros and P. L. Hammer, "Pseudo-boolean optimization," *Discrete Applied Mathematics*, vol. 123, no. 1-3, pp. 155–225, 2002.

[11] C. Rother, P. Kohli, W. Feng, and J. Jia, "Minimizing sparse higher order energy functions of discrete variables," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL*, 2009.

[12] T. M. Lillesand, R. W. Kiefer, and J. W. Chipman, *Remote Sensing and Image Interpretation*. John Wiley and Sons, 2003.

[13] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th International Conference on Computer Vision, Bombay, India*, 1998.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] R. Marée, P. Geurts, J. Piater, and L. Wehenkel, "Random subwindows for robust image classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA*, 2005.

[16] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forests for land cover classification," *Pattern Recognition Letters*, vol. 27, no. 4, pp. 294–300, 2006.

[17] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK*, 2008.

[18] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[19] Y. Freund and R. E.Shapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[20] D. Meyer, F. Leisch, and K. Hornik, "The suppor vector machine under the test," *Neurocomputing*, vol. 55, no. 1-2, pp. 169–186, 2003.

[21] J. C.-W. Chan and D. Paelinckx, "Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery," *Remote Sensing of Environment*, vol. 112, no. 6, pp. 2999–3011, 2008.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer Verlag, 2009.

[23] U. C. Benz, P. Hofmann, G. Willhauck, I. Lingenfelder, and M. Heynen, "Multi-resolution, object-oriented fuzzy analysis of remote sensing data for gis-ready information," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 58, no. 3-4, pp. 239–258, 2004.

[24] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *International Journal of Computer Vision*, vol. 75, no. 1, 2007.

[25] M. Y. Yang and W. Förstner, "A hierarchical conditional random field model for labeling and classifying images of man-made scenes," in *IEEE/ISPRS Workshop in Computer Vision for Remote Sensing of the Environment*, 2011.

[26] L. Ladický, C. Russell, P. Kohli, and P. H. S. Torr, "Associative hierarchical crfs for object class image segmentation," in *Proc. 12th International Conference on Computer Vision, Kyoto, Japan*, 2009.

[27] A. Ansar, A. Castano, and L. Matthies, "Enhanced real-time stereo using bilateral filtering," in *3D Data Processing, Visualization and Transmission*, 2004.

[28] J. Xiao, H. Cheng, H. Sawhney, and C. R. amd M. Isnardi, "Bilateral filtering-based optical flow estimation with occlusion detection," in *Proc. 9th European Conference on Computer Vision, Graz, Austria*, 2006.

[29] X. Niu, "A semi-automatic framework for highway extraction and vehicle detection based on a geometric deformable model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 61, no. 3-4, pp. 170–186, 2006.

[30] B. Sirmaçek and C. Unsalan, "Urban-area and building detection using sift keypoints and graph theory," *IEEE Transactions on Geosciences and Remote Sensing*, vol. 47, no. 4, pp. 1156–1167, 2009.

[31] S. Paris, P. Kornprobst, J. Tumblin, and F. Durand, "Bilateral filtering: Theory and applications," *Foundations and Trends in Computer Graphics and Vision*, vol. 4, no. 1, pp. 1–73, 2009.

[32] N. Dalal and W. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA*, 2005.

[33] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.

[34] S. K. Gehrig, F. Eberli, and T. Meyer, "A real-time low-power stereo vision engine using semi-global matching," in *International Conference on Computer Vision Systems*, 2009.

[35] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[36] Y. Ohta and T. Kanade, "Stereo by intra- and inter-scanline search using dynamic programming," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 7, no. 2, pp. 139–154, 1985.

[37] C. Poullis and S. You, "Delineation and geometric modeling of road networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 2, pp. 165–181, 2010.

[38] J. Borges, J. Bioucas-Dias, and A. Marçal, "Hyperspectral image segmentation with discriminative class learning," *IEEE Transactions on Geosciences and Remote Sensing*, vol. 49, no. 6, pp. 2151–2164, 2011.

[39] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image classification using multinomial logistic regression with active learning," *IEEE Transactions on Geosciences and Remote Sensing*, vol. 49, no. 10, pp. 3947–3960, 2011.

[40] P. Zhong and R. Wang, "Modeling and classifying hyperspectral imagery by CRFs with sparse higher order potentials," *IEEE Transactions on Geosciences and Remote Sensing*, vol. 49, no. 2, pp. 688–705, 2011.

[41] F. Lafarge, X. Descombes, J. Zerubia, and M. Pierrot-Deseilligny, "Automatic building extraction from DEMs using an object approach and application to the 3d-city modeling," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 63, no. 3, pp. 365–381, 2008.

[42] S. Kluckner, T. Mauthner, P. M. Roth, and H. Bischof, "Semantic classification in aerial imagery by integrating appearance and height information," in *Asian Conference on Computer Vision*, 2009.

[43] M. F. Tappen and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters," in *Proc. 9th International Conference on Computer Vision, Nice, France*, 2003.

[44] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 1068–1080, 2008.

[45] R. Stoica, X. Descombes, and J. Zerubia, "A Gibbs point process for road extraction from remotely sensed images," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 121–136, 2004.

[46] G. Perrin, X. Descombes, and J. Zerubia, "2d and 3d vegetation resource parameters assessment using marked point processes," in *Proc. International Conference on Pattern Recognition*, 2006.

[47] A. Lucchi, Y. Li, X. Boix, K. Smith, and P. Fua, "Are spatial and global constraints really necessary for segmentation?" in *Proc. 13th International Conference on Computer Vision, Barcelona, Spain*, 2011.