# Super-Pixel Class Segmentation in Large-Scale Aerial Imagery*

Stefan Kluckner, Michael Donoser, and Horst Bischof

Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{*kluckner,donoser,bischof*}*@icg.tugraz.at*

***Abstract***
*This paper proposes a super-pixel driven class segmentation for large-scale image understanding in aerial imagery. We demonstrate how super-pixels can be exploited to improve classification performance without introducing too much additional computational costs. Therefore, we explore a reasonable arrangement of processing steps ranging from a pure super-pixel classification to a grouping of confidences within an image segment. Moreover, we outline powerful yet low-dimensional region descriptors, capable of integrating multiple feature modalities such as color, texture or elevation measurements, together with random forests as classifiers, providing probabilistic class assignments. We first evaluate our approach on the Microsoft Research Cambridge (MSRC) image collection, which demonstrates state-of-the-art results for an averaged class recognition even without considering any additional context analysis. Then, we demonstrate our super-pixel driven class segmentation in aerial imagery.*

## 1. Introduction

Class specific image segmentation, also referred to as semantic classification, is of major scientific interest in the computer vision community. Most semantic interpretation pipelines [12, 16, 9, 10, 4] assign a specific object class label to each pixel in an image. Typically, huge visual variabilities of objects, but also occlusions, viewpoint and scale changes, and illumination variations make accurate classification challenging. In addition, the majority of the proposed approaches are only evaluated on collections of low-resolution images such as MSRC [12], where each image has a size of approximately 1-2 MPixels. In this work, we rather aim to perform large-scale classification in aerial imagery, where an imagery consists of hundreds of images and each image has a resolution of more than 85 MPixels. Figure 1 shows a comparison of a single aerial image of *Graz* with a resolution of $11500 \times 7500$ pixels and a montage of all $532$ MSRC images. Note that one single aerial image contains more pixels than the complete image collection. Hence classification algorithms must offer low computational costs in order to cope with the immense amount of data. Additionally, the huge object variability requires a high generalization capability to ensure an accurate classification.

Due to remarkable efficiency, many classification approaches are built on extracting features within rectangular image patches [17, 10, 11, 6], which are then directly trained and evaluated with strong classifiers. The proposed features are computed efficiently e. g. by applying integral structures [17, 14]

*This work was financed by the FFG Project APAFA (813397), the Austrian Science Fund Project W1209 under the doctoral program Confluence of Vision and Graphics and the Virtual Earth Academic Research Collaboration funded by Microsoft.

**Figure 1. A comparison of an aerial image, taken from *Graz*, with a dimension of $11500 \times 7500$ pixels and a montage of all images from MSRC [12]. One single aerial image contains more pixels than the complete MSRC image collection.**

or even by sampling raw pixel values [11]. Furthermore, random forests (RF) as classifiers [1] are frequently used to provide fast and accurate results in multi-class segmentation tasks [11, 10, 6].

Working directly with image patches allows to integrate local context into the class segmentation, but classification has to be done for each pixel. In addition, this approach does not preserve object boundaries. In order to achieve an accurate delineation of objects in images, a number of recent approaches utilize conservative over-segmentation techniques. Malisiewics and Efros [8] showed that a correct spatial support, provided by unsupervised segmentation methods, significantly improves the recognition performance over approaches using image patches. A variety of methods, obtaining state-of-the-art performance on evaluation datasets [12], directly integrate image partition methods into semantic classification or object localization/detection [9, 3, 4, 2, 5]. In [9, 3, 4] the authors utilized multiple image segmentations. It is obvious, that the generation of multiple segmentations induces enormous computational complexity and is impractical for aerial image classification. Just recently, Fulkerson et al. [2] proposed to use super-pixels, rapidly generated by quickshift [15], for object segmentation and localization. More important, these super-pixels accurately preserve object boundaries (as shown in Figure 2). In addition, subsequent stages benefit from the reduced image grid in terms of computational complexity. In contrast to [2], where a bag-of-word model [13] is used to describe image segments, we explore two concepts in order to get confident probabilistic class assignments for each super-pixel in the image at runtime.

The first concept (we refer to it as *SuperPixel*) relies on a direct computation of statistical features for each super-pixel. Figure 2 shows computed super-pixels for some MSRC images. We first show how to use statistical features for compact description of multiple modalities within super-pixels. Then, super-pixel feature instances, together with their class labels (taken from the available ground truth maps), are learned with a strong classifier. Please note, this concept also involves a conservative super-pixel segmentation for the training process, but feature extraction and classification can be performed efficiently on a reduced image grid.

The second method (we refer to it as *SpatialSupport*) is based on extracting features within rectangular patches located at each pixel. We train a classifier where each obtained feature vector is assigned the most frequent class label within the rectangular patch analyzing the provided ground truth. Contrary to *SuperPixel*, where the classifier directly yields a class distribution for a segment, we consider a super-pixel more like a region providing important spatial support. At runtime, the classifier computes confidences for each pixel considering a small rectangular neighborhood by using

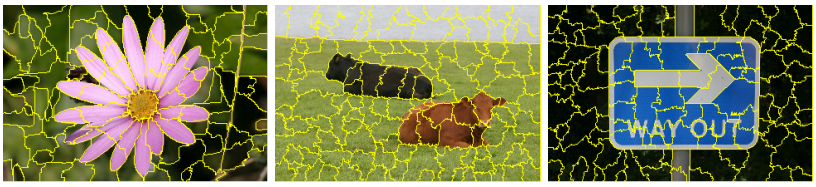


**Figure 2. Computed super-pixels using quickshift [15] for images extracted from the MSRC dataset [12]. These super-pixels preserve most of the object boundaries.**

a sliding window approach [17]. Then, an aggregation of confidences within a super-pixel results in an averaged class distribution for each segment. This method does not require any segmentation during training (which significantly simplifies the training process) and benefits from rapid feature computation within patches by using integral structures [14].

For both approaches we use Sigma Points [6] as feature representation and a random forest [1] as classifier (Sec. 2.). Considering the class distributions for each super-pixel in the image, a conditional random field (CRF) stage [7] refines the result on an adjacency graph for both concepts (Sec. 3.). In the experimental section (Sec. 4.), we first investigate the performance of both concepts on the well-known MSRC dataset. Then, we discuss the application of super-pixels in the context of large-scale class segmentation in aerial imagery. Finally, Sec. 5. concludes our approach and gives an outlook on future work.

## 2. Super-Pixel Class Segmentation

In this section we outline our powerful feature representation. We use Sigma Points, which describe various modalities such as color, filter responses, elevation measurements, etc. within defined image regions. Moreover, we discuss random forests as classifiers providing probabilistic class assignments.

### 2.1. Region Description using Sigma Points

Tuzel et al. [14] presented a compact region descriptor, based on covariance matrices of low-level image features, for rapid object detection. They also proposed a fast construction for arbitrary rectangles by exploiting an extension of integral structures [17]. Besides rectangular images patches, in our case a region can also be a super-pixel generated by quickshift [15]. While the feature extraction on local patches, defined by a center pixel, width and height, is performed using integral images, features for super-pixels, which vary in size and shape, are generated in one pass by using coordinate lists. A covariance matrix provides a low-dimensional description of statistics over $d$ modalities. The diagonal elements of the covariance matrix are the variances of the feature attributes in one channel, whereas the off diagonal elements capture the correlation values between the involved modalities. Thus the statistics up to second order of collected feature vectors can be represented by a mean vector $\mu \in \mathbf{R}^d$ and a covariance matrix $\Sigma \in \mathbf{R}^{d \times d}$.

The space of covariance matrices is not a vector space, therefore, simple arithmetic differences between the elements do not measure the real distance between two matrices. Thus covariance descriptors can not be used straightforwardly in e. g. random forests, where simple attribute comparisons are used to construct the classifier. Instead of exploiting manifolds [14] for valid covariance similarity measurements, we use Sigma Points [6], which represent individual covariance matrices in the Euclidean vector space. The idea relies on extracting specific samples of a given distribution, charac-

terized by $\mu$ and $\Sigma$, and offers a simple concept for combining first and second order statistics, since the mean vector describes an offset in the Euclidean vector space. We construct the set of Sigma Points[1] as follows:

$$\mathbf{p}_0 = \mu \qquad \mathbf{p}_i = \mu + \alpha(\sqrt{\Sigma})_i \qquad \mathbf{p}_{i+d} = \mu - \alpha(\sqrt{\Sigma})_i, \tag{1}$$

where $(\sqrt{\Sigma})_i$ with $i = 1 \dots d$ defines the $i$-th column of the required matrix square root. Due to symmetry of the covariance matrix, we apply the Cholesky factorization to compute the matrix square root of $\Sigma$ efficiently. The term $\alpha$ defines a weighting for the elements in the covariance matrix and is set to $\alpha = \sqrt{d}$. Then, a resulting region descriptor $\mathbf{P} = \{\mathbf{p}_0, \dots, \mathbf{p}_{2d}\}$ consists of $2d+1$ concatenated Sigma Points $\mathbf{p}_i \in \mathbf{R}^d$ and has a dimension of $\mathbf{P} \in \mathbf{R}^{d(2d+1)}$. For details we refer to [6].

### 2.2. Random Forests

Random forests [1, 11] offer a powerful method to classify high-dimensional data by using simple attribute comparisons. Forests are inherently multi-class and can handle noise and errors in the labeled training data. We train each decision tree in a supervised manner. A feature instance $(\mathbf{P}_i, y_i)$ of the training set consists of a region description $\mathbf{P}_i$, represented by the Sigma Points, and the class label $y_i$. The split nodes are learned from a random subset of the training data (which speeds up the training process) by using a greedy strategy. Each split criterion then minimizes the weighted information gain [11], considering the class distributions estimated from target labels falling into left and right children nodes. After construction by using the subset of training samples, each tree is refined with the complete set of instances in order to generate the final leaf node's class distributions. This technique enables a sophisticated handling of a large amount of data and improves the generalization capability [11]. At runtime, the classifier is evaluated by parsing down a test vector $\mathbf{P}_j$ in the forest and accumulating the distributions in the reached leaf nodes $L$ yielding an averaged class histogram $H(\mathbf{c}|L)$. Implementation details can be found in [6, 11, 1].

## 3. Refined Labeling

Although the Sigma Point representation captures some local context information, each pixel or super-pixel in the image space is interpreted independently. In order to incorporate spatial dependencies between the nodes defining the image grid, e. g. CRF formulations [10, 12, 6] are used to enforce an evident final class labeling. In this work, we apply a CRF stage defined on super-pixel neighborhoods. Let $G(S, E)$ be the adjacency graph with a super-pixel node $s_i \in S$ and a pair $(s_i, s_j) \in E$ be an edge between the segments $s_i$ and $s_j$. Then, a label energy can be defined with respect to the class labels $\mathbf{c}$ as

$$E(\mathbf{c}|G) = \sum_{s_i \in S} D(s_i|c_i) + \omega \sum_{(s_i,s_j)\in E} V(s_i, s_j|c_i, c_j). \tag{2}$$

The term $D(s_i|c_i) = -\log\left(H(\mathbf{c}|s_i)\right)$ denotes the unary potential, directly provided by the output of the RF (for *SuperPixel*) or obtained by aggregating confidences within a super-pixel with $H(\mathbf{c}|s_i) = 1/|s_i| \sum_{(x,y)\in s_i} H(\mathbf{c}|L, (x, y))$ (for *SpatialSupport*). The factor $\omega$ controls the influence of the regularization and is estimated during the training process. In order to integrate the region sizes into the minimization process (preferring larger regions), we compute the pairwise edge term $V(s_i, s_j|c_i, c_j)$ between the super-pixels $s_i$ and $s_j$ with

$$V(s_i, s_j|c_i, c_j) = \frac{b(s_i, s_j)}{1 + g(s_i, s_j)} \delta(c_i \neq c_j), \tag{3}$$

[1] An implementation can be downloaded at http://www.icg.tugraz.at/Members/kluckner.

where $b(s_i, s_j)$ counts the number of common boundary pixels of two given segments and $g(s_i, s_j)$ is the $L^2$ norm of the mean color distance vector. In this work we minimize the energy defined in Equation 2 by using the efficient primal-dual strategy of Komodakis and Tziritas [7].

# 4. Experiments

We first apply our two concepts *SuperPixel* and *SpatialSupport* to the MSRC images [12]. The dataset includes $532$ images with $21$ object classes. As suggested in [12], $276$ images are used for training and the remaining $256$ for testing. In all experiments we extract the target labels considering the available ground truth data, which provides class information at the pixel level.

We compute the Sigma Points, describing a super-pixel or an image patch ($21 \times 21$ pixels), over CIELab colors and the absolute values of the first order derivatives yielding a feature vector $\mathbf{P}_i$ with a dimension of $55$ for $d = 5$. We train an RF with $10$ trees, each with a maximum depth of $15$. In order to obtain precise split decisions near the leaf nodes, we linearly increase the number of greedy iterations according to the corresponding node's depth in the tree. Each tree in the forest is trained on a set of $60\,000$ randomly selected feature instances. Due to unbalanced quantity of samples for each class, we apply an inverse class weighting similar to [11]. In order to generate small segments, not limited to a size or number, we apply quickshift [15] to a five-dimensional vector consisting of image coordinates and CIELab color. The parameters for quickshift are set to $\sigma = 2$ and $\tau = 8$.

In Table 1, we compare the results, obtained for *SuperPixel* and *SpatialSupport*, in terms of correctly classified pixels with state-of-the-art rates [9, 12]. We report both the overall per-pixel classification rate (i. e. the accuracy of all pixels correctly classified) and the average of class specific per-pixel percentages, which gives a more significant measurement due to varying quantity of labeled pixels for each class. The overall results, adding a CRF stage, are given for $\omega = 4.5$. Interestingly, the initial classification, performed at the pixel level only using image patches, obtains a rate of $55.4\%$ correctly classified pixels. This rate improves to $61.2\%$ by exploiting super-pixels as spatial support. We assume that an improved integration of local context information due to exceeding the object boundaries and high redundancy within a super-pixel obtained by classification at the pixel-level, cause the significant increase of correctly classified pixels. The CRF stage further improves the classification rates yielding a consistent final labeling of the super-pixels. Please note, the results presented in this work do not consider e. g. global context information [3] or location priors [4], which would avoid impossible object constellations within an image like that a cow stands on a table.

An evaluation of computational costs at runtime shows that the generation of the super-pixels for a MSRC image consumes most of the time (approximately $680$ ms) in both concepts, *SuperPixel* and *SpatialSupport*. Feature extraction, classification and refined labeling on a reduced image grid ($100$ ms) runs slightly faster than evaluating all pixels in the test image by using efficient integral structures ($160$ ms). For comparison, our suggested concepts take about $1$ second per image, which is $6$ times faster than the approach of [12]. Taking into account the computational costs and classification accuracy we conclude that the method *SpatialSupport* performs better than *SuperPixel* on small images. In the following, we perform large-scale class segmentation in aerial imagery.

## 4.1. Classification in Aerial Imagery

We use high resolution color images ($11500 \times 7500$ pixels) taken with the *Microsoft UltraCam*. Due to the quantity (e. g. the imagery of *Graz* consists of $155$ highly overlapping images) and the dimension

| | RF+Segments | | with CRF/overall | |
|---|---|---|---|---|
| | pixel level | class avg | pixel level | class avg |
| *SuperPixels* | 54.3 | 43.2 | 67.8 | 57.5 |
| *SpatialSupport* | 61.2 | 51.5 | **68.6** | **59.6** |
| Pantofaru et al. [9] | | | 74.3 | 60.3 |
| Shotton et al. [12] | | | 72.2 | 57.7 |

**Table 1. Classification accuracies on the MSRC dataset with $21$ classes: The results are given in terms of correctly classified pixels at the pixel level and the averaged accuracy for each object class. A comparison to [9] and [12] shows state-of-the-performance for the class average. Please note, contrary to [9] or [4], our approach does not include global context or spatial information.**

of the images, a partition of each image into super-pixels for training and testing is impractical to solve. A full segmentation into super-pixels, without a classification procedure, of a single image (divided into $100$ tiles) takes about $25$ minutes, which scales to a couple of hours for a complete dataset. Hence we rather concentrate on the concept of *SpatialSupport*, where the super-pixels give spatial support for a classification at the pixel level. In addition, we aim to compute the super-pixels on ortho-projected color image tiles, resulting from a fusion step similar to the concept proposed by Zebedin et al. [18]. Taking into account camera data and 3D depth information, this step enables a pixel-wise fusion of multiple highly redundant images into a common 3D coordinate system forming a reduced set of data. For each pixel on ground, the fusion provides up to ten observations for color, height and assigned class probabilities.

We consider a segmentation into five object classes (*building*, *tree*, *water*, *grass* and *street*). Figure 3 shows the color coding of the classes (bottom row) and examples of two corresponding color images and pixel-wise classifications, providing redundant information from different viewpoints (top row). For training we manually label parts of six perspective images providing the ground truth data. Additionally, we integrate absolute elevation measurements into the Sigma Points feature representation yielding a significant improvement in classification accuracy as shown in [6]. The feature vector then consists of $78$ attributes, since a number of $d = 6$ modalities is considered. The size of the rectangular neighborhood is set to $7 \times 7$ pixels. We train an RF with $8$ trees, each with a depth of $12$. At runtime we apply the classifier to single aerial images, yielding a highly redundant semantic interpretation from different viewpoints.

A classification of a single aerial image at full resolution takes approximately $3$ minutes on a dual core machine. The fusion step for color and classification, also including the super-pixel generation, of $8$ different viewpoints covering an area of $100 \times 100$ meters lasts less than $4$ minutes. For evaluation we use additionally labeled test data, covering an area of approximately $400 \times 400$ meters. The classification rates obtained for *Graz* are summarized in Table 2. Figure 3 shows results for two fused image tiles of *Graz*. While the raw pixel-wise fusion of the class probabilities shows high granularity in the most likely assigned classes and blurred object boundaries due to inaccurate 3D information, an integration of super-pixels enhances the result significantly. Applying the CRF stage further improves the final semantic classification.

## 5. Conclusion

This work has explored how super-pixels can be exploited in an efficient manner to improve class segmentation accuracy. We have presented two concepts *SpatialSupport* and *SuperPixel* in order to

| | *building* | *water* | *grass* | *tree* | *street* |
|---|---|---|---|---|---|
| pixel-level | 88.5 | 74.2 | 89.4 | 87.3 | 91.7 |
| *SpatialSupport* | 90.6 | 82.0 | 91.8 | 91.5 | 94.4 |
| with CRF | **92.7** | **85.8** | **94.4** | **92.6** | **95.3** |

**Table 2. Class segmentation accuracy in terms of correctly classified pixels on hand-labeled test data. It can be seen clearly that a partition into small image segments, providing spatial support, improves accuracy.**

compute reliable class probabilities for each image segment. Since a super-pixel computation consumes a multiple of the computation time, compared to feature extraction, classification and refined labeling, we have discussed how to apply the concepts to aerial imagery maintaining low computational costs. In the experimental evaluation, we have shown state-of-the-art performance on MSRC images without using location priors or global context cues. In addition, we have demonstrated that the concept of *SpatialSupport* improves the accuracy significantly taking into account the classification at the pixel level. Due to superior initial classification accuracy on the MSRC dataset, future work will focus on integrating context cues to handle valid class constellations within a test image. Moreover, we will investigate the performance of Sigma Points applied to a bag-of-words model.

## References


[1] Leo Breiman. Random Forests. In *ML*, pages 5–32, 2001.

[2] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Class Segmentation and Object Localization with Superpixel Neighborhoods. In *ICCV*, 2009.

[3] Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. Object Categorization using Co-Occurrence, Location and Appearance . In *CVPR*, 2008.

[4] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. Multi-class Segmentation with Relative Location Prior. *IJCV*, 80(3):300–316, 2008.

[5] Chunhui Gu, Joseph J. Lim, Pablo Arbelaez, and Jitendra Malik. Recognition using Regions. In *CVPR*, 2009.

[6] Stefan Kluckner, Thomas Mauthner, Peter M. Roth, and Horst Bischof. Semantic Classification in Aerial Imagery by Integrating Appearance and Height Information. In *ACCV*, 2009.

[7] Nikos Komodakis and Georgios Tziritas. Approximate Labeling via Graph Cuts Based on Linear Programming. *PAMI*, 29(8):1436–1453, 2007.

[8] Tomasz Malisiewicz and Alexei A. Efros. Improving Spatial Support for Objects via Multiple Segmentations. In *BMVC*, 2007.

[9] Caroline Pantofaru, Cordelia Schmid, and Martial Hebert. Object Recognition by Integrating Multiple Image Segmentations. In *ECCV*, 2008.

[10] Florian Schroff, Antonio Criminisi, and Andrew Zisserman. Object Class Segmentation using Random Forests. In *BMVC*, 2008.

[11] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic Texton Forests for Image Categorization and Segmentation. In *CVPR*, 2008.

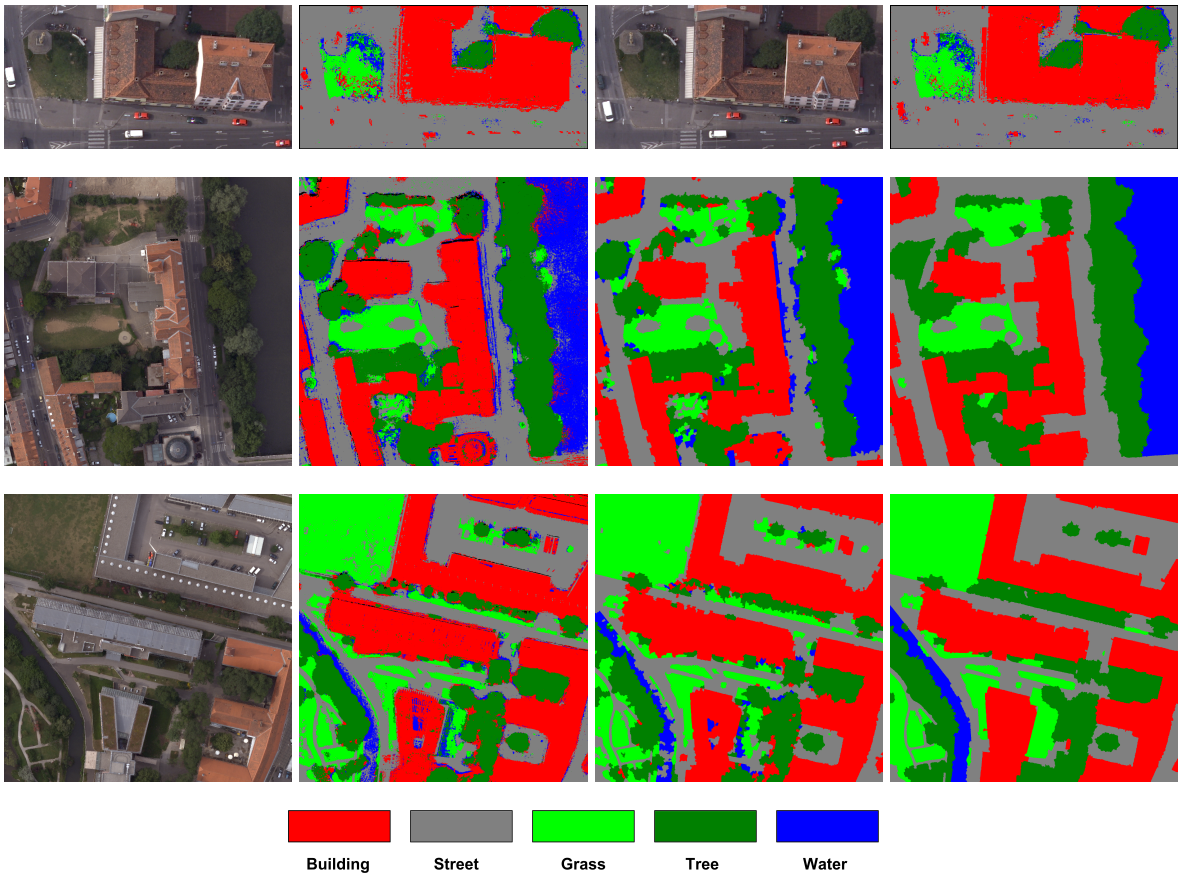


**Figure 3. Results for *Graz*: The top row shows a class segmentation at the pixel level obtained for two corresponding color images. We fuse the redundant information by using a projection to a common 3D coordinate system (second and third row): The first column (from left to right) shows the fused color images. The next column gives the most likely class at the pixel level, fused from multiple observation. The third column depicts the most likely class at super-pixel level using the concept of *SpatialSupport*. The results obtained with the CRF stage are presented in the last column. The last row gives the color coding of the five classes. Best viewed in color.**

[12] Jamie Shotton, John Winn, Carsten Rother, and Antonio Crimininsi. TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. In *ECCV*, 2006.

[13] Josef Sivic and Andrew Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, 2003.

[14] Oncel Tuzel, Fatih Porikli, and Peter Meer. Region Covariance: A fast Descriptor for Detection and Classification. In *ECCV*, 2006.

[15] Andrea Vedaldi and Stefano Soatto. Quick Shift and Kernel Methods for Mode Seeking. In *ECCV*, 2008.

[16] Jakob Verbeek and Bill Triggs. Region Classification with Markov Field Aspect Models. In *CVPR*, 2007.

[17] Paul Viola and Michael Jones. Robust Real-time Object Detection. *IJCV*, 2(57):137–154, 2004.

[18] Lukas Zebedin, Andreas Klaus, Barbara Gruber-Geymayer, and Konrad Karner. Towards 3D Map Generation from Digital Aerial Images. *IJPRS*, 60:413–427, 2006.