Semantic Classification in Aerial Imagery by Integrating Appearance and Height Information*

Stefan Kluckner, Thomas Mauthner, Peter M. Roth, and Horst Bischof

Institute for Computer Graphics and Vision,
Graz University of Technology, Austria
{kluckner,mauthner,pmroth,bischof}@icg.tugraz.at
http://www.icg.tugraz.at

Abstract. In this paper we present an efficient technique to obtain accurate semantic classification on the pixel level capable of integrating various modalities, such as color, edge responses, and height information. We propose a novel feature representation based on Sigma Points computations that enables a simple application of powerful covariance descriptors to a multi-class randomized forest framework. Additionally, we include semantic contextual knowledge using a conditional random field formulation. In order to achieve a fair comparison to state-of-the-art methods our approach is first evaluated on the MSRC image collection and is then demonstrated on three challenging aerial image datasets Dallas, Graz, and San Francisco. We obtain a full semantic classification on single aerial images within two minutes. Moreover, the computation time on large scale imagery including hundreds of images is investigated.

1 Introduction

Internet driven initiatives, like *Google Earth* and *Virtual Earth*, collect an enormous amount of aerial and satellite images in order to automatically construct 3D worlds of urban environments because of the demand for fast realistic 3D modeling, cartography, navigation support, etc. These location-aware applications on the internet push the development of efficient, accurate, and automatic technologies. The first step is to acquire high resolution images. In particular, the *Microsoft Ultracam* takes multi-spectral images in overlapping strips, resulting in high redundancy, which adheres every visible spot of urban environments from many different camera viewpoints. The high redundancy within the data enables methods for automatic height data generation [1] or full photo-realistic 3D modeling [2]. In contrast to photo-realistic 3D modeling, where the model consists of millions of triangles with fitted texture extracted from aerial images, we aim

* This work was supported by the Austrian Science Fund Projects W1209 and P18600 under the doctoral program Confluence of Vision and Graphics, by the FFG projects APAFA (813397) and AUTOVISTA (813395), financed by the Austrian Research Promotion Agency, and by the Austrian Joint Research Project Cognitive Vision under the projects S9103-N04 and S9104-N04.

for synthetic modeling, i.e., based on the information directly derived from the images to build a virtual model of a city. In addition, a synthetic model reduces the problem of privacy violations due to modeling the semantic interpretation instead of the realistic appearance.

Due to high variability in aerial imagery, automatic classification and semantic description still pose an unsolved task in computer vision. We aim to use appearance cues, such as color, edge responses, and height information for accurate semantic classification into five classes. For instance, using a combination of color and height data successfully separates the street regions from gray-valued roof tops or distinguishes between green areas and trees. Figure 1(a) shows corresponding color and height images, extracted from the dataset *San Francisco*. The classification of aerial images into several classes provides a semantic knowledge of the objects on ground and approves a specified post-processing to build up a semantic 3D world, where each object is modeled according to its obtained interpretation. A semantic description of a small sub-image is illustrated in Fig. 1(b).

Fig. 1. A pair of images extracted from the dataset *San Francisco* consisting of color and height information, and the corresponding semantic description of the sub-image (highlighted rectangles).

In [3], the authors proposed an appearance driven approach to exploit color and infrared data for initial classification. Several methods concentrate on extracting single object classes, e.g., buildings by integrating only LIDAR data [4] or height models [5]. The tight integration of 3D data into image classification, as additional information source, is still a new and upcoming field of research. Hoiem [6] extracted 3D information, such as surface orientation or vanishing lines, from single images to improve 2D object recognition. Recent approaches [7, 8] include SfM to improve the interpretation in street side images. In this work, we exploit dense matching results [1] together with appearance features to obtain an accurate semantic interpretation.

Shotton et al. [9] proposed simple color value differences in a small neighborhood for initial semantic classification on the pixel level using a randomized forest (RF) classifier [10]. Schroff et al. [11] extended this approach by including multiple feature types for an improved RF classification. Strong low level fea-

ture representations, such as SIFT [12], histograms of oriented gradients [13], or various types of filter responses [14–16] are widely used in appearance driven supervised classification. However, a compact combination of different feature cues is computationally very expensive. In addition, an integration into a common classification framework requires sophisticated techniques.

Thus, our work has three main contributions: To allow an efficient semantic classification, we first introduce a novel technique to obtain a powerful feature representation, derived from compact covariance descriptors [17] which is directly applicable to RF classifiers. Covariance matrices [17] can be efficiently computed and provide an intuitive integration of various feature channels. Since the space of covariance matrices does not form a Euclidean vector space [17], this representation can not be directly used for most machine learning techniques. To overcome this drawback, manifolds [18, 17, 19] are typically utilized, which, however, is computationally expensive. In contrast to calculating similarity between covariance matrices on Riemannian manifolds [18], we present a simple concept for mapping individual covariance descriptors to Euclidean vector space. The derived representation enables a compact integration of appearance, filter responses, height information etc. while the RF efficiently performs a multi-class classification task on the pixel level. Second, we introduce semantic knowledge by applying an efficient conditional random field (CRF) stage incorporating again several feature cues and co-occurrence information. To demonstrate the state-of-the-art performance we present quantitative results on the *Microsoft Research Cambridge* dataset MSRC-9 [15] by integrating visual appearance cues, such as color and edge information. Third, we apply our proposed method to real world aerial imagery, performing large scale semantic classification. We extend the novel feature representation with available height data as an additional cue and investigate the classification accuracy in terms of correctly classified pixels. Labeled training data, representing five annotated classes (building, tree, waterbody, green area and streetlayer), provides the input for the training process.

The remainder of this paper is structured as follows. Section 2 describes the derived covariance region descriptor in detail, illustrates the application to the RF framework, and also addresses the integration of the contextual constraints. Section 3 highlights the included feature cues and presents results on the MSRC-9 dataset and various real world aerial images. Finally, Sec. 4 concludes our work and gives an outlook on future work.

2 Semantic Classification

In this section we highlight the semantic classification pipeline including the feature representation based on covariance descriptors and *Sigma Points*, respectively, the straight forward application to a multi-class RF framework, and the CRF stage to handle the contextual constraints.

2.1 Approximated Covariance Representation

Tucel et al. [17, 19] presented a compact feature representation based on covariance matrices for rapid object detection and classification. In fact, covariance descriptors [17] provide a low-dimensional feature representation that simply integrates multiple feature channels, such as color, filter responses, height information, etc. and also exploits the correlation between them. The diagonal elements provide the variances of the feature attributes in one channel, whereas the off diagonal elements capture the correlation values between the different feature modalities. The statistics up to second order of N independent and identically distributed feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ can be represented by the sample mean $\mu = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$ and the sample covariance $\Sigma \in \mathbb{R}^{d\times d}$:

$$\Sigma = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T. \tag{1}$$

As shown by Tuzel et al. [17] the concept of integral images [16] can be applied to compute covariance descriptors on a rectangular image grid in constant time: Given a multi-channel feature image I of the dimension $w \times h \times d$, any $n \times m$ rectangular region $R \subseteq I$ can be represented by a $d \times d$ covariance matrix Σ. An extension of common integral images to higher dimensions incorporating additional tensor integral images, enables the computation of symmetric covariance matrices using the law of total variance. Implementation details can be found in [17].

Because of the missing symmetry requirement the space of covariance matrices is non-Euclidean [17]. Hence, standard machine learning methods, which require similarity computations can not be used directly. Instead of exploiting computationally costly manifolds [17, 19] to obtain a valid covariance similarity measurement, we propose a technique to represent individual covariance matrices directly on Euclidean vector space. Julier et al. [20] proposed the unscented transform (UT), which approximates a single distribution by sampling instead of approximating an arbitrary non-linear function by mapping to manifolds [18]. The UT provides an efficient estimator for the probability distribution and was successfully applied to unscented Kalman filtering [21], where it overcomes the drawbacks of truncated (second order) Taylor expansions. In the d-dimensional case the UT relies on constructing a small set of $2d+1$ specific vectors $\mathbf{s}_i \in \mathbb{R}^d$, also referred to as *Sigma Points* [20]. We construct the set of *Sigma Points* as follows:

$$\mathbf{s}_0 = \mu \qquad \mathbf{s}_i = \mu + \alpha(\sqrt{\Sigma})_i \qquad \mathbf{s}_{i+d} = \mu - \alpha(\sqrt{\Sigma})_i, \tag{2}$$

where $i = 1 \ldots d$ and $(\sqrt{\Sigma})_i$ defines the i-th column of the required matrix square root $\sqrt{\Sigma}$. The scalar α defines a constant weighting for the elements in the covariance matrix and is set to $\alpha = \sqrt{2}$ for Gaussian input signals [20].

In contrast to Monte Carlo methods, where test vectors are selected at random, the construction of the *Sigma Points* can be seen as an efficient mapping of

a specified set of test vectors $t_i \in \mathbb{R}^d$ that deterministically sample the intersections of an unit hypersphere with a d-dimensional Cartesian coordinate system. Here, the mean vector $t_0 = \mu$ represents the origin. The computed statistics of these points $\mathbf{s}_i$ accurately capture the original information about Σ up to third order for Gaussian and up to second order for non-Gaussian inputs [21]. Figure 2 illustrates the specified sampling of the test vectors and the mapping for a simplified 2D case.

Since covariance matrices Σ are positive semi-definite by definition, we first perform a simple regularization $\Sigma = \Sigma + \epsilon\mathbf{I}$, where $\mathbf{I}$ is the identity matrix and $\epsilon =$ 1e-6, to obtain symmetric positive definite matrices. Due to symmetry and positive definiteness of the regularized covariance matrices, the efficient Cholesky factorization can be applied to compute the matrix square root by decomposing $\Sigma = LL^T$. Then, L corresponds to $\sqrt{\Sigma}$ and is a lower triangular matrix. In principle any method for square root factorization can be used, however, the Cholesky decomposition requires the lowest mathematical operations yielding a complexity of $O(n^3/3)$.

The resulting feature representation $\mathcal{S}^k = \left(\mathbf{s}_0^k, \mathbf{s}_1^k \dots \mathbf{s}_{2d}^k\right)$ is obtained by concatenation of the *Sigma Points* and captures both, first and second order statistics, which are given by the mean and covariance information. Each of these generated vectors $\mathbf{s}_i^k \in \mathbb{R}^d$ describe Euclidean space, therefore, element-wise distance computations between corresponding samples of a given distribution are feasible. The construction pipeline for the set of *Sigma Points* is summarized in Algorithm 1.

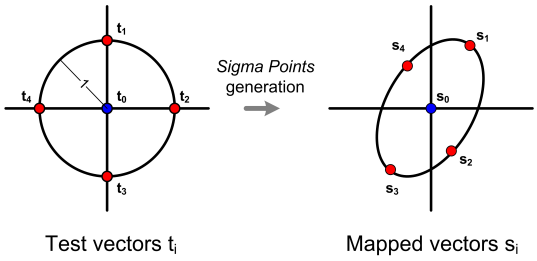


Fig. 2. The mapping of a fixed set of test vectors $\mathbf{t}_i$ to the *Sigma Points* $\mathbf{s}_i$ given in a second coordinate system, representing the original characteristics of the covariance matrix $\Sigma = LL^T$.

The structure of this feature representation $\mathcal{S}^k$ perfectly fits the concept of randomized forest classifiers, where the learning and evaluation strategy is based on comparing randomly selected attributes of an available representation. Note that, since a reference representation is missing, similarity measurements, such as the Foerstner metric [18] are intractable to directly use in decision trees. In the following section we show how our representation can be applied straight forward to a RF framework.

Algorithm 1 Construction of our proposed feature representation based on *Sigma Points*.

Require: Mean vector μ^k and covariance matrix Σ^k
1: Perform a simple regularization $\Sigma^k = \Sigma^k + \epsilon\mathbf{I}$
2: Compute matrix square root $\Sigma^k = LL^T$
3: Compute s_i^k according to (2)
4: Construct the set of *Sigma Points* $\mathcal{S}^k = \left(\mathbf{s}_0^k, \mathbf{s}_1^k \dots \mathbf{s}_{2d}^k\right)$

2.2 Randomized Forest Framework

Randomized forests [10] have proven to give robust and accurate classification results for multi-class tasks [9, 11, 22]. An RF consists of an ensemble of binary decision trees, where the nodes of each tree include split criteria that give the direction of branching left and right down the tree until a leaf node is reached. Each leaf node l_i in a given maximal depth D contains a learned class distribution $P(\mathbf{c}|l_i)$. By averaging the decisions over all T trees in a forest the resulting accumulated probabilities yield an accurate class distribution $P(\mathbf{c}|L) = \frac{1}{T}\sum_{i=1}^{T} P(\mathbf{c}|l_i)$. To rapidly grow each tree of the forest, the split node criteria are learned using only a subset $\mathcal{S}'$ of the whole training data $\mathcal{S}$. For training a class label c_k is assigned to each feature representation $\mathcal{S}^k \in \mathcal{S}$. The learning proceeds from the root node top-down by tiling the available subset at each split node into left and right sets. Proposed splitting decisions in [9, 22] are achieved by comparing two or multiple randomly chosen elements $\mathbf{s}_i^k$ and $\mathbf{s}_j^k$ of the given feature sample $\mathcal{S}^k$. In our implementation we follow a strategy similar to [22], randomly taking into account the correct corresponding dimension $a \in \{1 \dots d\}$ selecting two weighted elements i and j according to

$$\alpha\mathbf{s}_i^k(a) + \beta\mathbf{s}_j^k(a) = \begin{cases} > \gamma, & split\ left \\ \leq \gamma, & split\ right \end{cases} . \tag{3}$$

Here, α, β, and γ denote the greedy-optimized parameters that minimize the information gain with respect to the training labels [9, 22]. We take the numbers of split node tests as suggested in [22]. Once the forest is trained, we evaluate the classifier at each pixel location by parsing down the extracted feature representation in the forest and accumulating the class distribution to obtain an overall probability map $P(\mathbf{c}|L)$.

2.3 Incorporating Contextual Information

Although our feature representation includes a spatial neighborhood of $n \times m$ implicitly, each pixel is classified independently. In this work we apply an efficient conditional random field (CRF) stage based on linear programming [23] to incorporate semantic contextual constraints yielding a smooth labeling of the final image classification. In addition, we include edge information into the four-connected graph to exactly capture the real object boundaries.

In order to obtain the contextual semantic information, we construct a dataset dependent co-occurrence matrix by counting the frequency of class labels in the training images within randomly chosen rectangular sub-windows [9]. The frequency counts can be performed quickly on single images using integral structures [16]. Furthermore, we follow the concept of [24] to compute a normalized co-occurrence matrix $\theta(c_i, c_j)$ representing the pairwise semantic contextual information of the grid nodes i and j. The application of the CRF allows us to include the posterior class distribution $P(\mathbf{c}|L)$, the likelihood co-occurrence matrices $\theta(\cdot)$ and an edge penalty function to preserve the object boundaries. Given a four-neighborhood connected image I we define an energy with respect to the class labels according to

$$E(\mathbf{c}) = \sum_i D(c_i) + \sum_{i,j} w_{ij} V(c_i, c_j), \tag{4}$$

where $D(c_i)$ denotes the data term, including the unary potentials according to $D(c_i) = -\log(P(c_i|L))$ at grid node i. The pairwise class potentials are computed according to $V(c_i, c_j) = -\log(\theta(c_i, c_j))\delta(c_i \neq c_j)$ and include the semantic knowledge. The weight w_{ij} describes an edge penalty term between the nodes i and j. Following the concept suggested in [11], where the authors used color distance computations to capture the object boundaries, we exploit the height information in case of the aerial images. Thus, the weight is constructed with $w_{ij} = \exp\left(-\lambda \|h_i - h_j\|^2\right)$, where h_i, h_j are the height values at the neighboring graph nodes. λ defines a factor and is learned while training. In this work we apply the strategy of Komodakis et al. [23] to minimize the energy defined in (4). In the experimental evaluation we present overall results incorporating semantic contextual information into the classification pipeline.

3 Experimental Evaluation

Due to efficient computation of our region based covariance representation, we exploit several feature cues incorporating a small spatial neighborhood. First, we construct the required integral images to compute the covariance descriptors including the feature cues, such as color channels, first derivatives in x and y direction, and the height values. Then, the feature instances are constructed according to our proposed concept (see Sec. 2.1). The collected samples provide the representation for training and testing. In the following, we first evaluate our classification pipeline on the standard MSRC-9 [15] evaluation dataset. By integrating appearance and height information we illustrate the application to real world aerial imagery and investigate large scale capability.

3.1 Experiments on MSRC-9 Dataset

In our first experiment we use the MSRC-9 dataset with nine on the pixel level labeled classes to provide results for a comparison to state-of-the-art approaches [11, 25]. For the training and the testing procedure we randomly split

the dataset including a total number of 240 images, 120 training and 120 test images. The training samples, consisting of the set of *Sigma Points* $\mathcal{S}$ and a target label vector c, are regularly collected on a 5×5 grid with a small spatial neighborhood of $n = m = 21$ pixels. The corresponding label is extracted by considering the available ground truth images. Confirming the observation in [9], the CIELab color space generalizes better than raw RGB values. The first derivatives are computed on the L-channel. We apply small synthetic affine distortions to the training images capturing an invariance to shape deformations [9]. In addition, we extend the test images, according to the spatial neighborhood, to obtain class probabilities at the image borders. Due to randomness of our approach, we repeat the experiment 20 times independently to obtain meaningful averaged classification rates. In this work, we choose a relatively small size of the forest ($T = 15$ trees and a maximum depth of $D = 10$) to provide both, efficiency in testing and classification accuracy.

Our pixel-wise RF classification returns rates of 64.2% using only color and 71.1% integrating both, color and derivative information. The feature representation $\mathcal{S}_i$ at a pixel i integrating only color yields a concatenated vector with a dimension of 21 attributes, while an extension to include derivatives increases the size to 55. In [11] rates of 72.2% are given for only incorporating color information, however using a forest with 20 trees each with a maximum depth of 20. Running the full classification cue including the CRF stage achieves an average classification performance of 84.2%, while in [11] and [25] rates of 87.2% and 84.9% are reported, respectively. Running the full classification cue, consisting of the feature extraction, the evaluation of the classifier at each pixel and the integration of semantic knowledge using the CRF stage, on a single image requires less than 2 seconds on a standard single core PC. Figure 3 depicts a selection of semantic classification results on the MSRC-9 dataset.

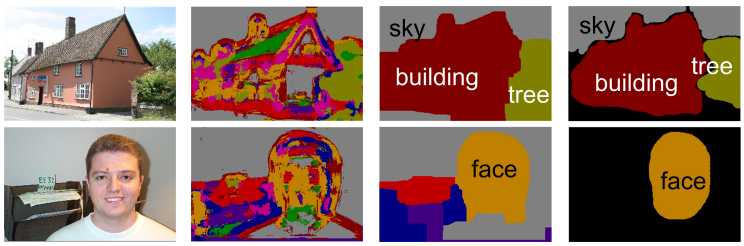


Fig. 3. A selection of results on the MSRC-9 dataset. From left to right: color images, pure pixel-wise classifications, final result using RF and CRF and ground truth labeling.

Considering the results of our first experiment on the MSRC dataset, we conclude that an integration of color and derivatives enhances the classification rates significantly. Including semantic contextual information, using an efficient CRF stage further improves the results. The comparison shows that our throughout simple approach is competitive with existing methods [11, 25].

3.2 Experiments on Aerial Images

The second experiment evaluates the classification pipeline on huge real world aerial images. We apply separately trained RF classifies to single aerial images performing a semantic classification into five classes (building, tree, waterbody, green area and streetlayer) on the level of pixels.

In this work, we perform experiments on three different datasets, generated by the *Microsoft Ultracam*: The dataset *Dallas* includes large building structures and gray valued areas, *Graz* shows a colorful characteristic with challenging building blocks, and the images of *San Francisco* have mainly suburban appearance. The color images have a dimension of $11.5K \times 7.5K$ pixels and provide a ground sampling distance (GSD) of 8 cm (*Graz*) and 15 cm (*Dallas*, *San Francisco*). Due to high redundancy a dense matching process [1], taking into account three adjacent images, yields range images representing the surface model. Subtracting the surface model from the extracted ground plane using, e.g., [26] produces the relative height information that is directly applicable to our classification procedure as an additional feature channel. The dimension of the resulting feature vector increases to 78, if CIELab color, derivative, and height information are integrated. Figure 1(a) shows a pixel synchronous pair of a color and the corresponding height image. For each dataset we independently label three images providing the training labels on the pixel level. Additionally, we generate two non-overlapping images as ground truth data for testing. Similar to the MSRC training process the target labels are then collected taking into account these training maps.

In case of the aerial images we compute our feature representation integrating the color, texture, and height information and train an RF classifier with 15 trees and maximum depths of 10 separately for each dataset. The dimension of the spatial neighborhood is set according to the datasets GSD with $n = m = 2(50/GSD + 1)$. The trained RFs are evaluated at each pixel location using a fourth of the full image resolution. The obtained classification rates for the three datasets are summarized as confusion matrices in Fig. 5. A combination of color, derivatives, and height information results in averaged rates of 92% (*Dallas*), 93% (*Graz*), and 88% *San Francisco*. For instance, using only color and derivative cues yields low classification accuracies of 79% (*Dallas*), 78% (*Graz*), and 73% *San Francisco*.

Figure 4 depicts a full semantic classification including the CRF stage of a single image taken from the *Graz* dataset. The feature extraction and pixel-wise classification of a single aerial image of *Graz* covering an area of approximately 0.5 km^2 requires about 35 seconds, the CRF stage increases the computation time to approximately 80 seconds. This scales to an overall computation of about 1.5 hours on a standard PC given a complete dataset, e.g., *Graz* with 155 images. Note that for a full dataset processing the CRF stage can be applied to a fused classification result instead of using the per-image classification, which speeds up the computation drastically. Figure 6 illustrates a selection of classified subimages extracted from full processing steps.

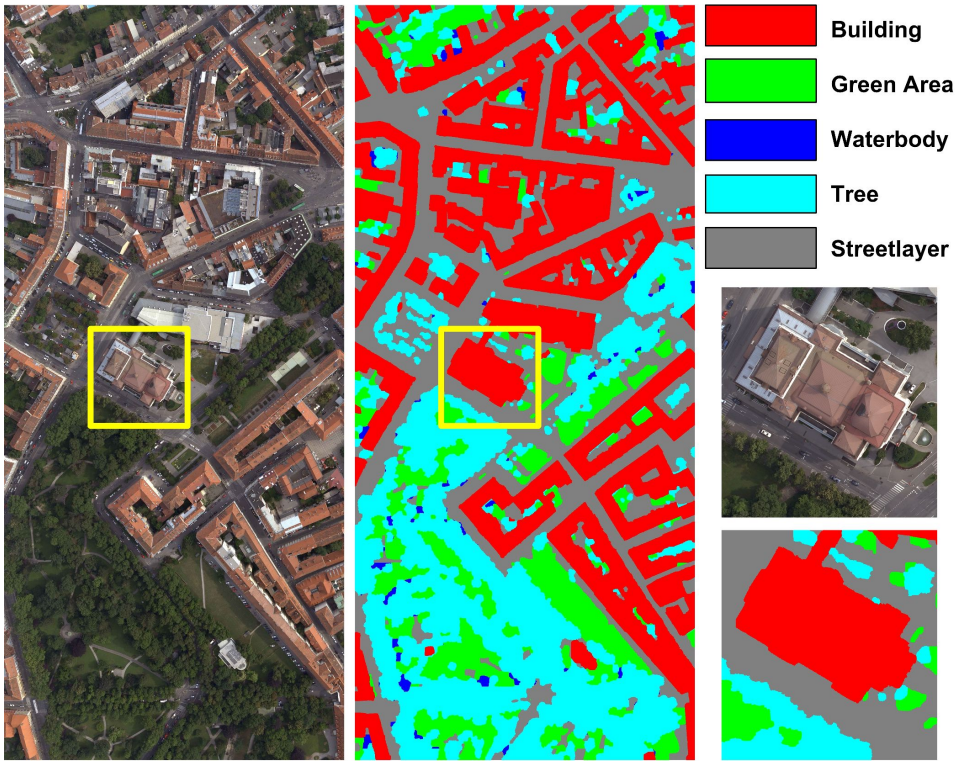
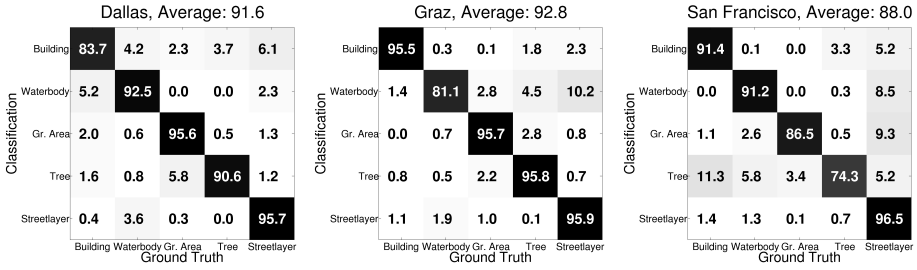


Fig. 4. Full semantic classification of a single image taken of the dataset *Graz*. The image provides a ground sampling of 8 cm and covers an area of approximately 0.5 km^2.



(a) *Dallas* (b) *Graz* (c) *San Francisco*

Fig. 5. Computed confusion matrices on the three aerial image datasets. We obtain classification rates of approximately 90% on the three challenging datasets. The low gray-valued buildings in *Dallas* are sometimes mixed with the streetlayer class which can be caused by inaccurate terrain models. Due to similar spectral ranges small shadow regions in the streets are classified as waterbody in *Graz*. Many small trees inside of courtyards and the hilly terrain in *San Fransisco* explain the relatively low classification rate for trees.

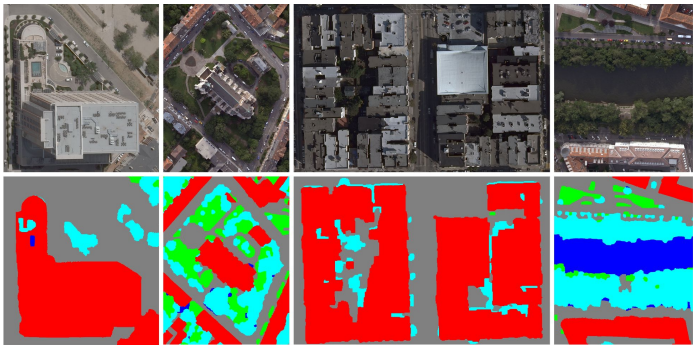


Fig. 6. Representative sub-images extracted from full semantic classification results. From left to right: a hotel complex with a pool/trees on the top in *Dallas*, a church surrounded with vegetation in *Graz*, a typical building block of *San Fransisco*, and a detail showing a river in *Graz*.

4 Conclusion

This work has proposed an efficient approach for semantic classification of images by integrating multiple types of feature modalities, such as appearance, edge responses, and height information. We presented a novel feature representation based on covariance matrices and *Sigma Points*, respectively, that can be directly applied to multi-class RF classifiers. By including contextual information using a CRF stage, we achieved an accurate semantic description of test images on the pixel level. We performed experiments on the MSRC dataset and on huge real world aerial images and demonstrated accurate classification results with low computational costs. Further work will investigate the influence of additional data cues, like infrared and pan-chromatic images, on the classification quality. In addition, we work on exploiting the redundancy by fusing multiple image classification results of different viewpoints.

References

1. Klaus, A., Sormann, M., Karner, K.: Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In: Proceedings ICPR. (2006)
2. Zebedin, L., Bauer, J., Karner, K., Bischof, H.: Fusion of feature- and area-based information for urban buildings modeling from aerial imagery. In: Proceedings ECCV. (2008)
3. Zebedin, L., Klaus, A., Gruber-Geymayer, B., Karner, K.: Towards 3D map generation from digital aerial images. International Journal of Photogrammetry and Remote Sensing **60** (2006) 413–427
4. Matei, B.C., Sawhney, H.S., Samarasekera, S., Kim, J., Kumar, R.: Building segmentation for densely built urban regions using aerial LIDAR data. In: Proceedings CVPR. (2008)

5. Lafarge, F., Descombes, X., Zerubia, J., Pierrot-Deseilligny, M.: Automatic building extraction from dems using an object approach and application to the 3D-city modeling. International Journal of Photogrammetry and Remote Sensing **63**(3) (2008) 365–381
6. Hoiem, D., Stein, A., Efros, A.A., Hebert, M.: Recovering occlusion boundaries from a single image. In: Proceedings ICCV. (2007)
7. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: Proceedings ECCV. (2008)
8. Cornelis, N., Leibe, B., Cornelis, K., Van Gool, L.: 3D city modeling using cognitive loops. In: International Symposium on 3DPVT. (2006)
9. Shotton, J., Johnson, M., Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: Proceedings CVPR. (2008)
10. Breiman, L.: Random forests. Machine Learning (2001) 5–32
11. Schroff, F., Criminisi, A., Zisserman, A.: Object class segmentation using random forests. In: Proceedings BMVC. (2008)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
13. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings CVPR. (2005)
14. Shotton, J., Winn, J., Rother, C., Crimininsi, A.: Textonboost: Joint appearance, shape and conext modeling for muli-class object recognition and segmentation. In: Proceedings ECCV. (2006)
15. Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary. In: Proceedings ICCV. (2005)
16. Viola, P., Jones, M.: Robust real-time object detection. International Journal of Computer Vision (2002)
17. Tuzel, O., Porikli, F., Meer, P.: Learning on lie groups for invariant detection and tracking. In: Proceedings CVPR. (2008)
18. Foerstner, W., Moonen, B.: A metric for covariance matrices. Technical report, Department of Geodesy and Geoinformatics, Stuttgart University (1999)
19. Tuzel, O., Porikli, F., Meer, P.: Human detection via classification on riemannian manifolds. In: Proceedings CVPR. (2007)
20. Julier, S., Uhlmann, J.K.: A general method for approximating nonlinear transformations of probability distributions. Technical report, Robotics Research Group, Department of Engineering Science, University of Oxford (1996)
21. Julier, S., Uhlmann, J.K.: A new extension of the kalman filter to nonlinear systems. In: International Symposium of Aerospace/Defense Sensing, Simulations and Controls. (1997)
22. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: Proceedings ICCV. (2007)
23. Komodakis, N., Tziritas, G.: Approximate labeling via graph cuts based on linear programming. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(8) (2007) 1436–1453
24. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: Proceedings ICCV. (2006)
25. Verbeek, J., Triggs, B.: Scene segmentation with CRFs learned from partially labeled images. In: Advances in NIPS. (2007)
26. Champion, N., Boldo, D.: A robust algorithm for estimating digital terrain models from digital surface models in dense urban areas. In: Proceedings ISPRS Commission 3 Symposium, Photogrammetric Computer Vision. (2006)