https://doi.org/10.2352/ISSN.2470-1173.2018.17.AVM-346

# Context Aware Hyperspectral Scene Analysis

***Christian Winkens, Dietrich Paulus, Active Vision Group, Institute for Computational Visualistics, University of Koblenz-Landau, Germany***

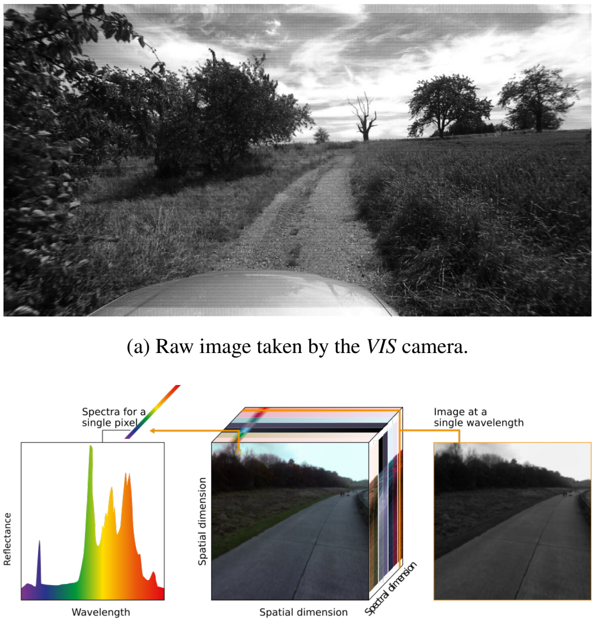
## Abstract

*Hyperspectral imaging increases the amount of information incorporated per pixel in comparison to normal color cameras. Conventional hyperspectral sensors as used in satellite imaging utilize spatial or spectral scanning during acquisition which is only suitable for static scenes. In dynamic scenarios, such as in autonomous driving applications, the acquisition of the entire hyperspectral cube at the same time is mandatory. In this work, we investigate the eligibility of novel snapshot hyperspectral cameras in dynamic scenarios such as in autonomous driving applications. These new sensors capture a hyperspectral cube containing 16 or 25 spectra without requiring moving parts or line-scanning. These sensors were mounted on land vehicles and used in several driving scenarios in rough terrain and dynamic scenes. We captured several hundred gigabytes of hyperspectral data which were used for terrain classification. We propose a random-forest classifier based on hyperspectral and spatial features combined with fully connected conditional random fields ensuring local consistency and context aware semantic scene segmentation. The classification is evaluated against a novel hyperspectral ground truth dataset specifically created for this purpose.*

## Introduction

Environment perception and analysis is crucial for autonomous driving, especially in off-road scenarios. Given sensor data, the correct semantic interpretation of a scene is a key factor for successful autonomous navigation. The use of hyperspectral sensors brings an advantage, as it allows a more detailed view of the composition and surface of materials, plants and floor coverings than conventional cameras, like shown in figures below. Researchers use so called hyperspectral line-scanning sensors mounted on satellites or planes (AVIRIS) for acquiring spectral data which provides static information of the Earths surface and allows only offline analysis. The drawback of established sensors are the scanning requirements for constructing a hyperspectral-cube (hypercube) of a scene like displayed in figure 1b. This leads to slow acquisition and motion artifacts when observing dynamic scenes like driving scenarios. This drawback can be overcome with novel, snapshot-mosaic (SSM) imaging sensors, which capture a whole spectrum in one shot. With these sensors it's possible to mount hyperspectral cameras on unmanned land vehicles and utilize them for hyperspectral scene analysis and autonomous navigation. This is an exciting and promising new application scenario, which has not been explored before.

In this work we investigate the use of novel sensors on unmanned land vehicles for drivability and scene analysis. Therefore, we utilize machine learning techniques to classify the captured spectral reflectances and make use of and combine established supervised classifiers to recognize different classes, which can be seen as environmental perception. Therefore we propose a classification pipeline which uses both spectral and neighborhood characteristics to achieve a consistent segmentation result. We utilize a Random Forest classifier to get an initial per-pixel classification which serves as input for an adapted fully connected conditional random field that establishes pairwise potentials on all pairs of pixels and enhances segmentation results.



(a) Raw image taken by the *VIS* camera.

(b) A schematic representation of a hypercube and an interpolated plot of a single data point (hyperpixel).

Figure 1: Raw image *VIS* camera with visible mosaic pattern. And a schematic representation of a hypercube used in this work.

By using this pipeline we examine the use of hyperspectral data for dynamic scene understanding especially in autonomous driving scenarios, as hyperspectral data allows a more detailed view of the composition of materials, plants and floor coverings.

The remainder of this paper is organized as follows. In the following section an overview of common algorithms for feature extraction and spectral classification is given. Then our general setup is presented in the Sensors section. Our feature extraction and classification approach is described in detail in the Scene Analysis section. And in the Experiments section we present our results on our new hand-labeled dataset. Finally a conclusion of our work is given in the last section.

## Related Work

The standard procedure for image-based scene-segmentation is defined by capturing regular RGB images and trying to identify different classes, like Chetan et al. [Chetan et al., 2010] and others did. They used color information and some features like local binary patterns (LBP) and trained different supervised classifiers. Shotton et al. [Shotton et al., 2009] introduced in 2009 a novel method for efficient recognition and semantic scene understanding from image data. They proposed new features which capture layout, texture and context information. The pipeline is also coupled with a conditional random field, which enhances classification accuracy. A patch-wise scene classification for urban street scenes is proposed by Ess et al. [Ess et al., 2009] utilizing a superpixel representation. Fulkerson et al. [Fulkerson et al., 2009] proposed a method to identify and localize objects in images. They trained a classifier on histograms of local features which where computed from superpixels. The classifier is then regularized by aggregating histograms from the neighbors of each superpixel. Finally the segmentation result is further enhanced by operating a conditional random field on the superpixel graph. An approach using local descriptors like SIFT [Lowe, 2004], Local Binary Patterns [He and Wang, 1990, Ojala et al., 1996] and enriching them with additional image information was presented by Carreira et al. [Carreira et al., 2012]. The enriched features were coupled with second-order pooling over free-form regions for semantic scene segmentation which produced good results on Pascal VOC 2011 dataset. In addition Wojek et al. [Wojek et al., 2013] performs 3D scene understanding from urban traffic scenes by utilizing a probabilistic scene model and a monocular camera. Scharwaechter et al. [Scharwächter et al., 2013] combine Stixels and a multi-cue bag-of-features classification scheme for semantic segmentation on grayscale and depth data. Recently Chen et al. [Chen et al., 2016] proposed *Deeplab*, a system which uses trained networks on image classification for semantic scene segmentation. They combined convolutional neural networks and fully-connected conditional random fields for detailed segmentations. Nearly all algorithms for semantic scene analysis use RGB data for classification. But, in recent years, hyperspectral imaging and classification has gained additional interest. Hyperspectral data allows for a more detailed insight into the composition and nature of objects and materials like plants and soil than standard RGB data. Given hyperspectral data, the goal of classification is to assign a unique label to each reflectance-vector so that it is well-defined by a given class. Unfortunately most supervised classifiers suffer from the Hughes effect [Hughes, 1968], especially when dealing with high-dimensional hyperspectral data. To deal with this issue, Melgani et al. [Melgani and Bruzzone, 2004] and Camps-Valls et al. [Camps-Valls and Bruzzone, 2005] introduced support vector machines with adequate kernels for hyperspectral classifications. But there are other algorithms which are suitable for hyperspectral data processing and analysis. Cavigelli et al. [Cavigelli et al., 2016] analyzed the potential of multispectral sensors in combination with deep neural-nets for semantic classification. The combination of RGB and multispectral data, using the same hyperspectral snapshot cameras, was evaluated by Cavigelli et al. [Cavigelli et al., 2016] on data with static background and a very small dataset utilizing deep neural networks. Furthermore there were some research on hyperspectral terrain-classification using random-forests [Winkens et al., 2017c].

## Sensors

The sensors used in this work utilize a specific filter mosaic structure, which has a per pixel design developed by IMEC [Geelen et al., 2014]. The filters are arranged in a rectangular mosaic pattern of $n$ rows and $m$ columns, which is repeated $w$ times over the width and $h$ times over the height of the sensor We used two different camera models, the MQ022HG-IM-SM4X4-VIS (*VIS*) which captures the visible spectrum 470–630 nm and the MQ022HG-IM-SM5X5-NIR (nir) which is designed for the near-infrared range 600-975 nm. The *VIS* camera has a $4 \times 4$ mosaic pattern and the *NIR* $5 \times 5$ which results in a spatial resolution of approx. $512 \times 272$ pixels $(4 \times 4)$ and $409 \times 217$ pixels $(5 \times 5)$. The cameras provide images in a lossless format with 8 bits per sample. There- fore the raw data captured by the camera needs a special preprocessing to construct a hypercube with spectral reflectances from the raw data like seen in figure 1. Preprocessing consists of cropping the raw-image to the valid sensor area, removing the vignette and converting to a three dimensional image, which we call a hypercube, like describe in [Winkens et al., 2017c].

## Scene Analysis

We have chosen to utilize a Random Forest (RF) as a supervised classifier, because it's fast to train and delivers remarkable results even in spectral classification [Winkens et al., 2017c]. Supervised learning techniques like Random Forest make use of training sets, which consist of a set of sample feature vectors coupled with a corresponding labeling. The labels $c \in \mathcal{C}$ are user-defined classes which are normally represented by integer numbers.

Given a set of $N$ corresponding training pairs the aim is to find a function $\gamma$ which generalizes well enough to new data, so accurate predictions for previously unseen data can be computed.

$$\gamma(\mathbf{x}) = c \tag{1}$$

In this process a classifier might generate a model which is a representation of the given problem from which a classification can be deduced. An accurate model yields better results for unseen data but highly depends on the training data.

Random Forests belong to the group of ensemble classifiers and utilize a set of Decision Tree classifiers to learn a robust model. Each classifier is trained on its own subset of training data which is generated by bagging. Bagging is a common approach where samples are randomly drawn with replacement from the original dataset to generate a new distribution of the data. This prevents overfitting and yields different patterns in the input data. The decision trees are unbalanced binary trees. A single decision tree is composed of several nodes, an unique root node, a set of internal nodes and a set of leaves. They form an decision space with the leafs representing a class assignment.s Furthermore these decision trees only use a random subset of the features for every decision node to further increase their diversity. These decision trees, form a Random Forest, which are used to classify the generated subsets. The result of the classification is obtained by majority voting. In order to semantic scene analysis, a suitable model must be trained using a Random Forest classifier. Since two cameras with different wavelength sensitivities were used here, two separate models need to be trained. As already mentioned in section , a pre-processed image forms a hypercube with a spectrum of 16 or 25 spectral reflectances for each pixel defined as $\chi$. For training, the annotated hypercubes are first dissected and filtered. We use the normalized spectrum as a feature, which reduces the influence of scene illumination and other irregularities

[Winkens et al., 2017a]. The normalization is computed by the sum of the spectrum's $n$ values:

$$\chi^S(x,y) = \sum_{i=1}^{n} \mathcal{B}_i(x,y) \tag{2}$$

Hence the normalized spectrum at image position $(x,y)$ is computed as

$$\mathcal{B}'_k(x,y) = \frac{\mathcal{B}_k(x,y)}{\chi^S(x,y)} \tag{3}$$

for each spectral band $k$. As input data, a Random Forest now receives an annotated spectrum $\chi$ with $|\mathcal{L}_\lambda|$ normalized spectral bands as a feature vector which corresponds to a per pixel classification of an image. Since the data is only classified pixel by pixel by the Random Forest, results are subject to a certain amount of noise because no neighborhood information is used. Therefore, the classification results provided by the Random Forest are used as input for a fully-connected CRF in a second classification step.
A Conditional Random Field (CRF) [Lafferty et al., 2001] defines smoothness terms that require the equality of labels of neighboring pixels, which can also model contextual relations between objects. CRF models consist of unary potentials defined on single pixels or image patches and pairwise potentials defined on adjacent pixels or patches. This results in a neighborhood structure encoded in a CRF. However, this structure is very limited in its ability to model far-reaching connections and relationships. As a result, object edges are usually subject to excessive smoothing. In order to improve segmentation, Koltun et al. [Koltun, ] extended the basic CRF-framework to include hierarchical connectivity, which is visualized in figure 2.
In a refinement step we use a fully connected conditional random field implementation as proposed by Krähenbühl et al. [Krähenbühl and Koltun, 2011, Krähenbühl and Koltun, 2013] which delivers a highly efficient approximate inference algorithm for fully connected CRF models.

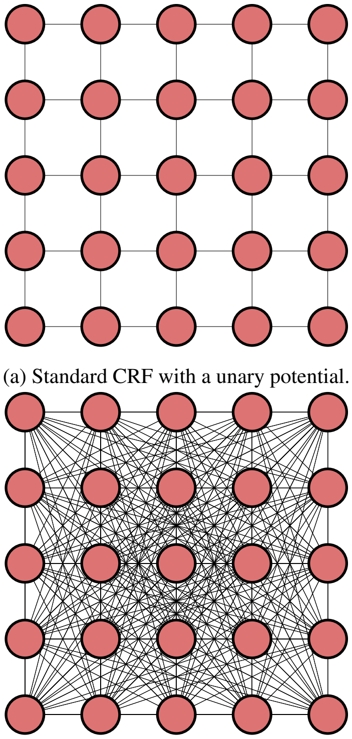
**Fully Connected CRF** A fully connected CRF is defined over a set $X = \{X_1, \ldots, X_n\}$ of variables whose domain is defined by a set of labels $\mathbb{L} = \{l_1, \ldots, l_k\}$ conditioned on the image $I$. The Gibbs energy of a labeling $\rho \in \mathbb{L}^{\mathcal{N}}$ is defined as

$$E(\rho|I) = \sum_i \psi_u(\rho_i) + \sum_{i<j} \psi_p(\rho_i, \rho_j)$$

with $i$ and $j$ ranging from 1 to $\mathcal{N}$. Here $\psi_u(\rho_i) = -\log P(\rho_i)$ defines the unary potential where $P(\rho_i)$ is the label assignment probability at pixel $i$ which is normally computed and provided by a classifier. In our work we used the Random Forest classifier to get per pixel label assignment probabilities as described above. The pairwise potentials $\sum_{i<j} \psi_p(\rho_i, \rho_j)$ are modeled as mixtures of kernels in feature space

$$\psi_p(\rho_i, \rho_j) = \mu(\rho_i, \rho_j) \sum_{m=1}^{K} w^{(m)} k^{(m)}(f_i, f_j)$$

where $k^{(m)}$ defines a gauss kernel and the vectors $f_i$ and $f_j$ are feature vectors in a $w^{(m)}$ feature space. For multi-class problems, two kernel potentials are defined which contain feature vectors $I_i$ and $I_j$ like pixel color and pixel positions $p_i$ and $p_j$.



(a) Standard CRF with a unary potential.

(b) Dense-CRF with unary and pairwise potential.

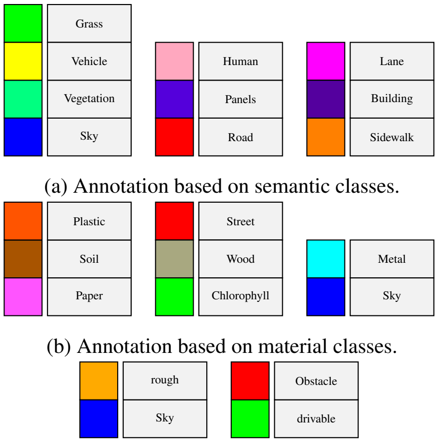
Figure 2: Schematic description of different CRF.

$$k(f_i, f_j) = w^{(1)} \exp\left(-\frac{|p_i - p_j|^2}{2\sigma_\alpha^2} - \frac{|I_i - I_j|^2}{2\sigma_\beta^2}\right) + w^{(2)} \exp\left(\frac{|p_i - p_j|^2}{2\sigma_\phi^2}\right)$$

A kernel defines the appearance probability that pixels with the same color belong to the same class $\sigma_\alpha^2$ and $\sigma_\beta^2$ control proximity and similarity. The smoothness kernel, on the other hand, removes small isolated regions. The pairwise potential term has a form that allows for efficient inference while using a fully connected graph. In our work we used a six dimensional feature space $w^6$ consisting of pixel positions, a local binary pattern feature and 3 selected spectral bands representing RGB colors.

## Experiments

As far as we know, there is no publicly available data set with hyperspectral data recorded by MQ022HG-IM-SM4X4-VIS camera and MQ022HG-IM-SM5X5-NIR camera, which use snapshot mosaic technique to acquire hyperspectral data. So we had to build a new dataset on our own, which has recently been published [Winkens et al., 2017b] and is publicly available. We equipped a standard car with the cameras manufactured by Ximea and collected several hours of data driving through suburban and rural areas, from which we selected a subset for labeling hyperspectral data. So we published a freely available synchronized and calibrated autonomous driving dataset capturing different scenarios. To the best of our knowledge its the first dataset including snapshot mosaic hyperspectral hyperspectral data from the visible to



(a) Annotation based on semantic classes.

(b) Annotation based on material classes.

(c) Annotation based on drivability classes.

Figure 3: Introduced annotation classes

the near-infrared range. We provide semantic, material and drivability labels to examine the use of hyperspectral data for semantic scene understanding especially in autonomous driving scenarios as shown in figure 3 and figure 4. The results of our experiments on semantic labeled *VIS* data are displayed in figure 3 and figure 5 with some examples shown in figure 6. Furthermore our results on semantic labeled *NIR* data are displayed in figure 7 and figure 8.

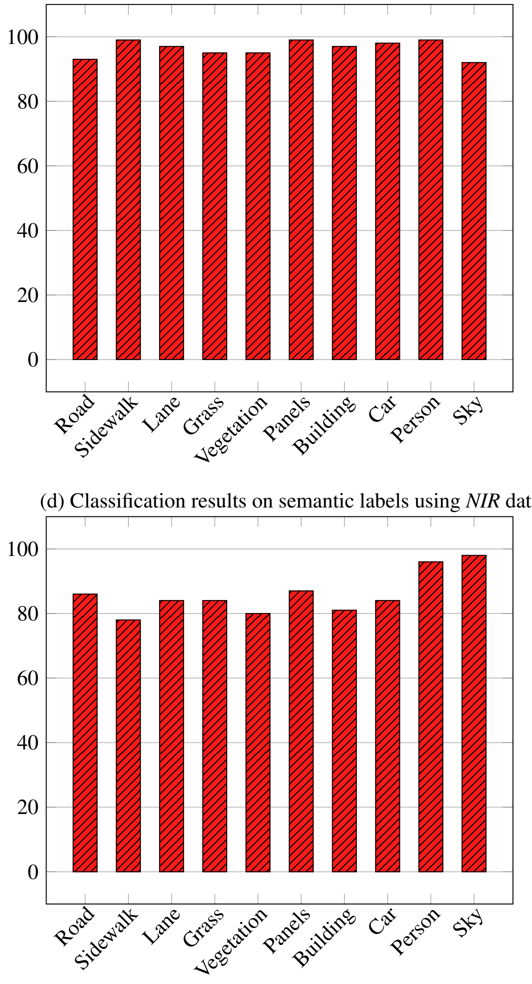
Taking the results of the CRF refinement into account, it can be seen that the segmentation is improved by adding neighborhood information. Many outliers have been removed from the road and other surfaces. Looking at the results in figure 5, it can be seen that by using a CRF the differentiation between vegetation and grass increases. This is particularly important for autonomous off-road driving. In addition, the classification of trafficable roads can be carried out very reliably. Our results demonstrate that use of fully connected CRF can increase accurate pixel-level classification performance in hyperspectral scene segmentation.

## Conclusion

We proposed a hyperspectral scene analysis pipeline which combines a per-pixel classification with context aware fully connected conditional random fields. The use of CRF allows the integration of context information in the classification process which enables local consistency. The combination of hyperspectral data and dense pixel-level connectivity leads to a more accurate pixel-level classification performance as our experiments indicate. Experiments were carried out on a novel hyperspectral ground truth dataset which is freely available now. In order to improve the classification results further, we plan to add 3D laser data to improve classification performance.

## Acknowledgments

This work was partially funded by Wehrtechnische Dienststelle 41 (WTD), Koblenz, Germany.



(d) Classification results on semantic labels using *NIR* data.

(e) Classification results on semantic labels using *VIS* data.

Figure 3: Classification results of our proposed classification pipeline.

## References

[Camps-Valls and Bruzzone, 2005] Camps-Valls, G. and Bruzzone, L. (2005). Kernel-based methods for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 43(6):1351–1362.

[Carreira et al., 2012] Carreira, J., Caseiro, R., Batista, J., and Sminchisescu, C. (2012). Semantic segmentation with second-order pooling. *Computer Vision–ECCV 2012*, pages 430–443.

[Cavigelli et al., 2016] Cavigelli, L., Bernath, D., Magno, M., and Benini, L. (2016). Computationally efficient target classification in multispectral image data with deep neural networks. In *SPIE Security+ Defence*, pages 99970L–99970L. International Society for Optics and Photonics.

[Chen et al., 2016] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2016). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*.

[Chetan et al., 2010] Chetan, J., Krishna, M., and Jawahar, C. (2010). Fast and spatially-smooth terrain classification using monocular camera. In *Pattern Recognition (ICPR), 2010 20th International Confer-*

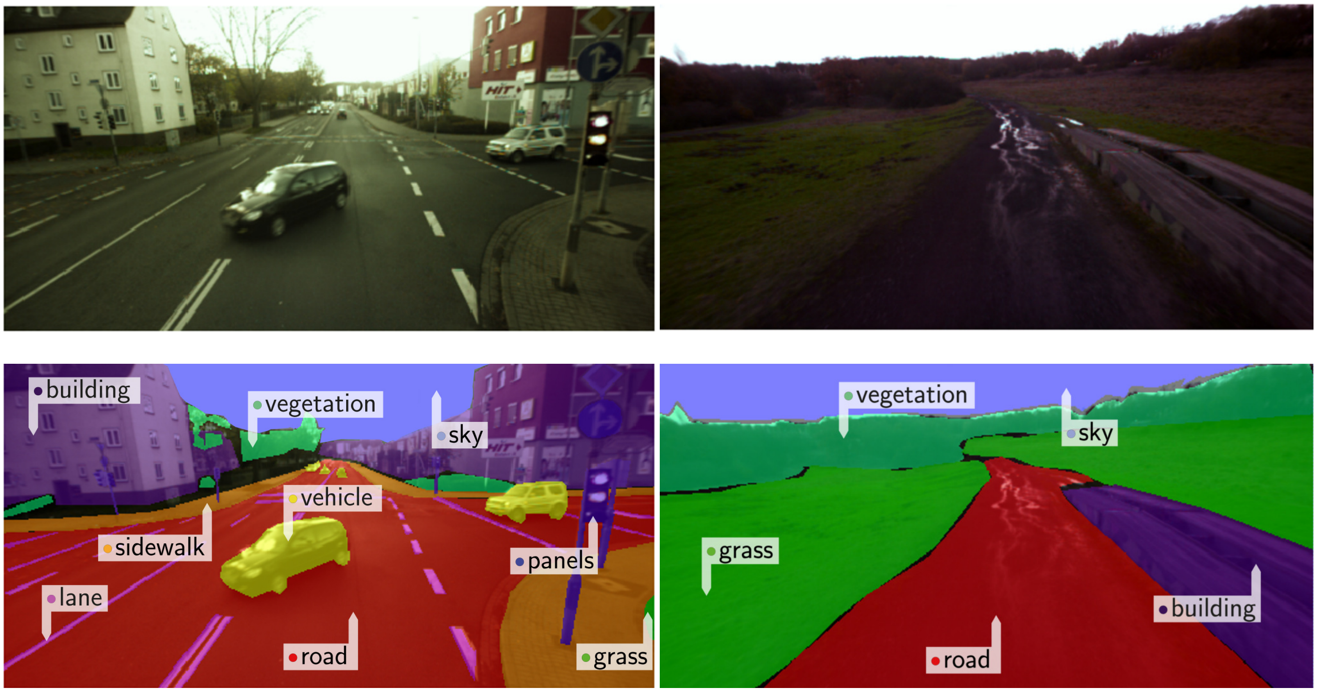


Figure 4: Some Example of our dataset with semantic annotations.

*ence on*, pages 4060–4063. IEEE.

[Ess et al., 2009] Ess, A., Müller, T., Grabner, H., and Van Gool, L. J. (2009). Segmentation-based urban traffic scene understanding. In *BMVC*, volume 1, page 2.

[Fulkerson et al., 2009] Fulkerson, B., Vedaldi, A., and Soatto, S. (2009). Class segmentation and object localization with superpixel neighborhoods. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 670–677. IEEE.

[Geelen et al., 2014] Geelen, B., Tack, N., and Lambrechts, A. (2014). A compact snapshot multispectral imager with a monolithically integrated per-pixel filter mosaic. In *Spie Moems-Mems*, pages 89740L–89740L. International Society for Optics and Photonics.

[He and Wang, 1990] He, D.-C. and Wang, L. (1990). Texture unit, texture spectrum, and texture analysis. *IEEE transactions on Geoscience and Remote Sensing*, 28(4):509–512.

[Hughes, 1968] Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory*, 14(1):55–63.

[Koltun, ] Koltun, P. K. V. Efficient inference in fully connected crfs with gaussian edge potentials.

[Krähenbühl and Koltun, 2011] Krähenbühl, P. and Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117.

[Krähenbühl and Koltun, 2013] Krähenbühl, P. and Koltun, V. (2013). Parameter learning and convergent inference for dense random fields. In *International Conference on Machine Learning*, pages 513–521.

[Lafferty et al., 2001] Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

[Melgani and Bruzzone, 2004] Melgani, F. and Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on geoscience and remote sensing*, 42(8):1778–1790.

[Ojala et al., 1996] Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59.

[Scharwächter et al., 2013] Scharwächter, T., Enzweiler, M., Franke, U., and Roth, S. (2013). Efficient multi-cue scene segmentation. In *German Conference on Pattern Recognition*, pages 435–445. Springer.

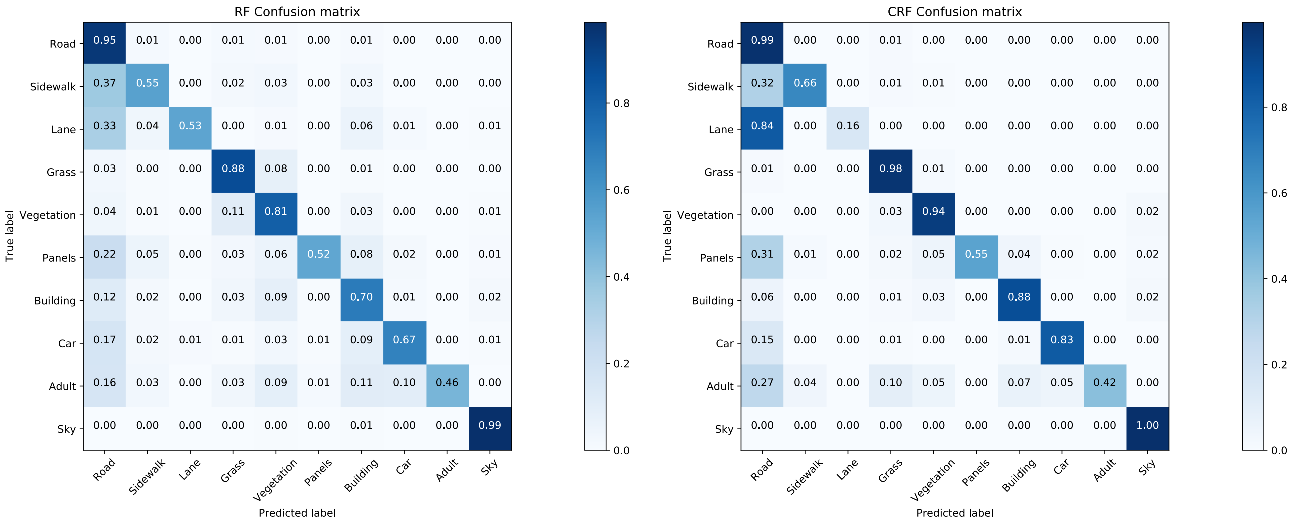
[Shotton et al., 2009] Shotton, J., Winn, J., Rother, C., and Criminisi, A. (2009). Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23.

[Winkens et al., 2017a] Winkens, C., Kobelt, V., and Paulus, D. (2017a). Robust features for snapshot hyperspectral terrain-classification. In *International Conference on Computer Analysis of Images and Patterns*, pages 16–27. Springer.

[Winkens et al., 2017b] Winkens, C., Sattler, F., Adams, V., and Paulus, D. (2017b). Hyko: A spectral dataset for scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 254–261.

[Winkens et al., 2017c] Winkens, C., Sattler, F., and Paulus, D. (2017c). Hyperspectral terrain classification for ground vehicles. In *12th International Conference on Computer Vision Theory and Applications (VISAPP)*.

[Wojek et al., 2013] Wojek, C., Walk, S., Roth, S., Schindler, K., and Schiele, B. (2013). Monocular visual scene understanding: Understanding multi-object traffic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 35(4):882–897.



(a) Confusion matrix of random-forest classification on semantic labels using *VIS* data.

(b) Confusion matrix of combined random-forest and crf classification on semantic labels using *VIS* data.

Figure 5: Confusion matrices of random forest and combined classification on *VIS* data.

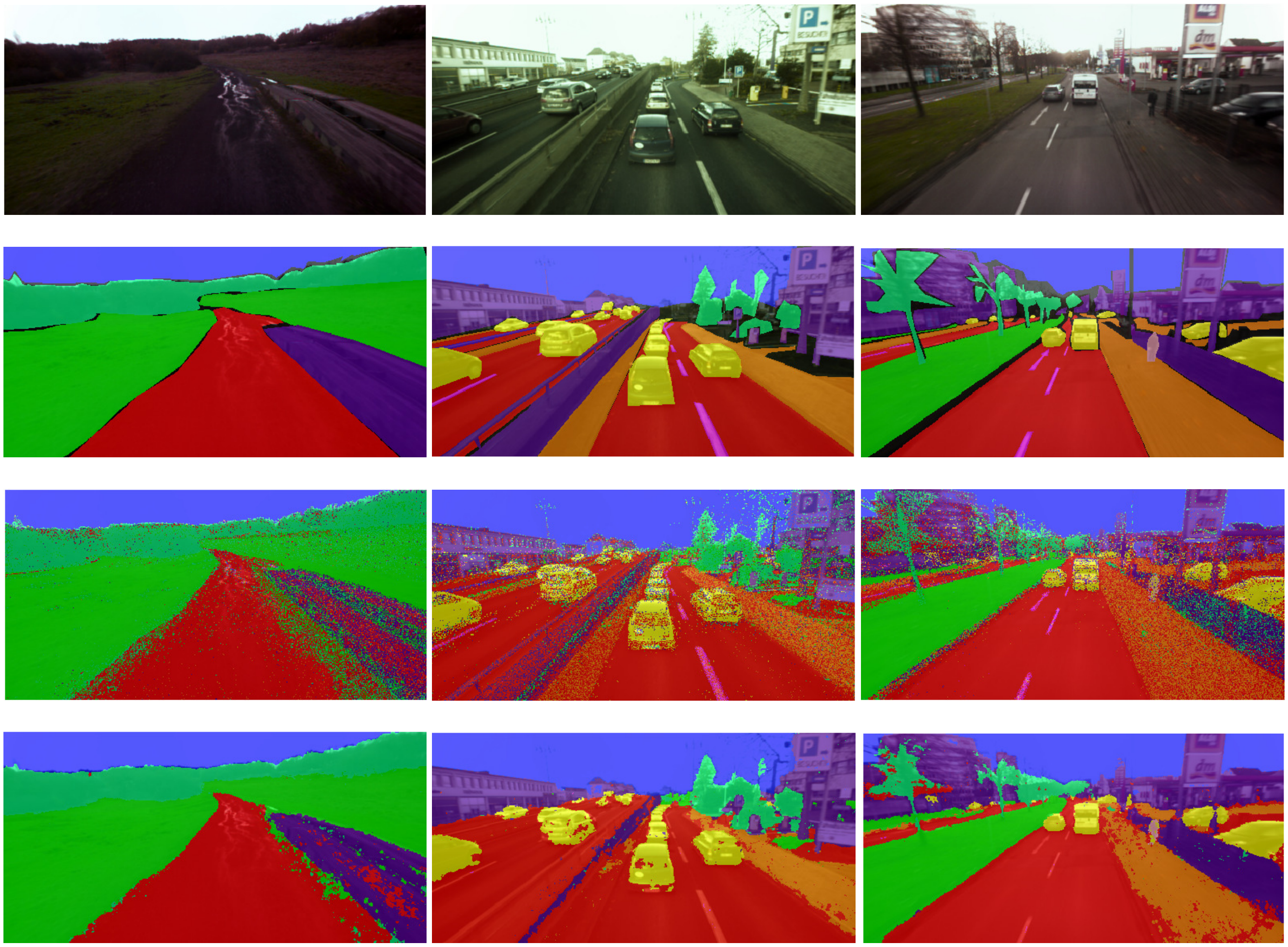
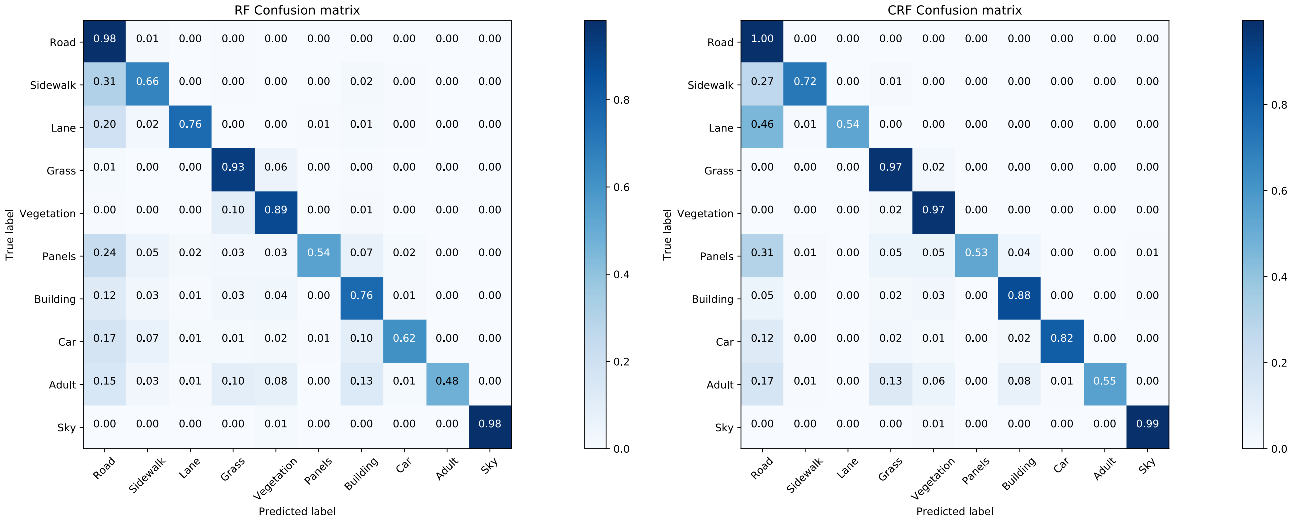


Figure 6: Comparison of classification results on *VIS* data with semantic labels. The rows show a RGB representation of the hyperspectral input image, the ground truth, the rand-forest classification and the combined classification.



(a) Confusion matrix of random-forest classification on semantic labels using *NIR* data.

(b) Confusion matrix of combined random-forest and crf classification on semantic labels using *VIS* data.

Figure 7: Confusion matrices of random forest and combined classification on *NIR* data.

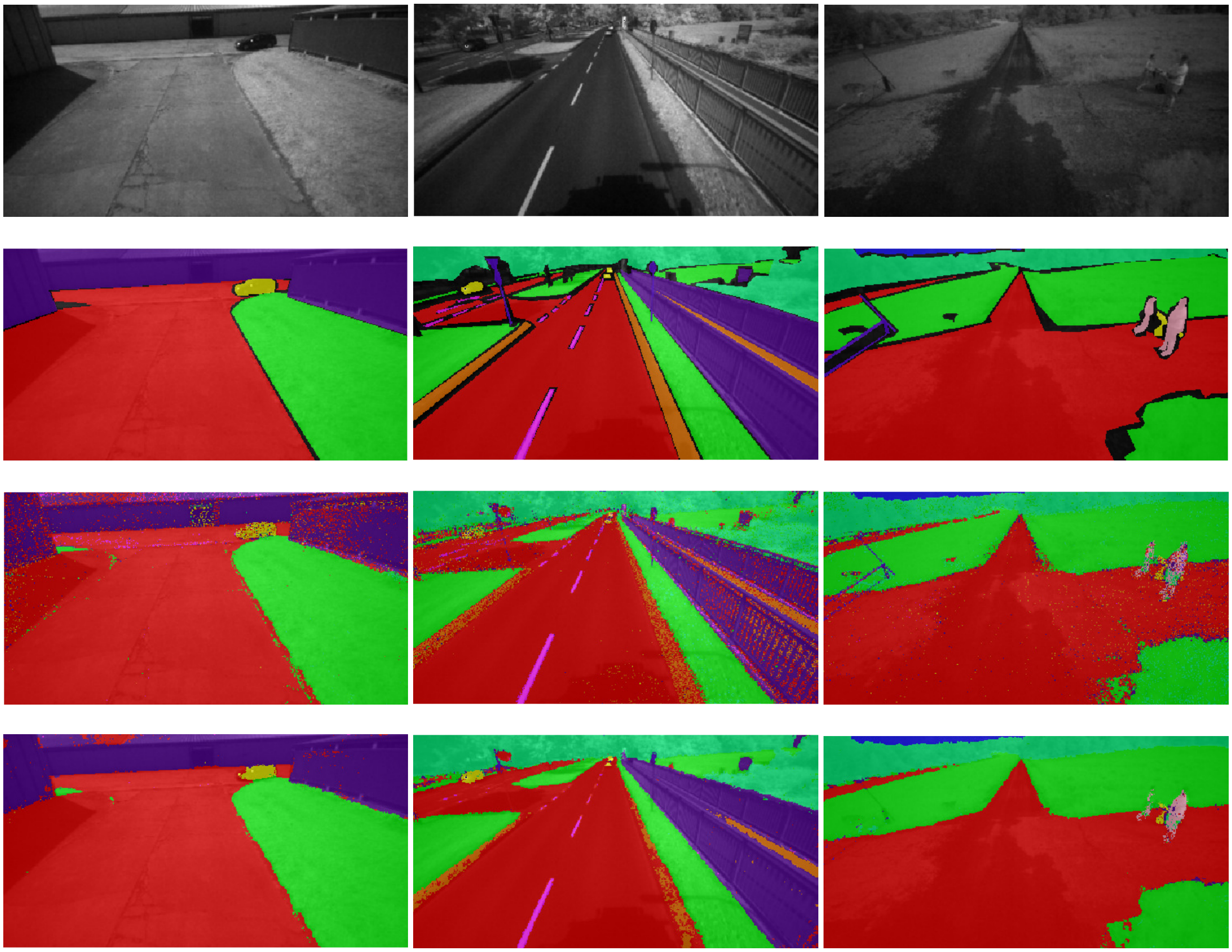


Figure 8: Comparison of classification results on *NIR* data with semantic labels. The rows show a grey value representation of the hyperspectral input image, the ground truth, the rand-forest classification and the combined classification.