FUSION OF HYPERSPECTRAL, MULTISPECTRAL, COLOR AND 3D POINT CLOUD INFORMATION FOR THE SEMANTIC INTERPRETATION OF URBAN ENVIRONMENTS

Martin Weinmann[1], Michael Weinmann[2]

[1] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Germany - martin.weinmann@kit.edu
[2] Institute of Computer Science II – Computer Graphics, University of Bonn, Germany - mw@cs.uni-bonn.de

Commission II, ICWG II/III

KEY WORDS: Scene Interpretation, Classification, Data Fusion, Multispectral, Hyperspectral, 3D, Aerial Sensor Platform

ABSTRACT:

In this paper, we address the semantic interpretation of urban environments on the basis of multi-modal data in the form of RGB color imagery, hyperspectral data and LiDAR data acquired from aerial sensor platforms. We extract radiometric features based on the given RGB color imagery and the given hyperspectral data, and we also consider different transformations to potentially better data representations. For the RGB color imagery, these are achieved via color invariants, normalization procedures or specific assumptions about the scene. For the hyperspectral data, we involve techniques for dimensionality reduction and feature selection as well as a transformation to multispectral Sentinel-2-like data of the same spatial resolution. Furthermore, we extract geometric features describing the local 3D structure from the given LiDAR data. The defined feature sets are provided separately and in different combinations as input to a Random Forest classifier. To assess the potential of the different feature sets and their combination, we present results achieved for the MUUFL Gulfport Hyperspectral and LiDAR Airborne Data Set.

1. INTRODUCTION

The acquisition, exploration and analysis of natural and built environments represents a topic of great interest in photogrammetry and remote sensing. For data acquisition, different sensors can be used which allow capturing different characteristics of a scene (e.g. geometric or radiometric data) in different data representations (e.g. point clouds, meshes or imagery) with different resolutions. The use of such individual types of geospatial data or their combination, in turn, has received much attention in recent years, particularly regarding the acquisition and analysis of urban scenes which provide a rich diversity of both natural and man-made objects. In this context, however, most investigations focus on the use of aerial imagery and the corresponding Digital Surface Model (DSM) (Rottensteiner et al., 2012; Gerke and Xiao, 2014; Audebert et al., 2018; Liu et al., 2017; Chen et al., 2018), thereby neglecting other potentially useful types of geospatial data.

In this paper, we address the semantic interpretation of urban environments on the basis of multi-modal data acquired from aerial sensor platforms. The considered types of data comprise RGB color imagery, hyperspectral data and LiDAR data. From the radiometric data, we extract features based on the reflectance values corresponding to the respective spectral bands, and we also consider the effect of different transformations to potentially better data representations (Weinmann and Weinmann, 2018). For the transfer of RGB color imagery to a potentially better data representation, we apply transformations derived via color invariants, normalization procedures or specific assumptions about the color representation of a scene. For the transfer of hyperspectral data to a potentially better data representation, we take into account that the spectral bands of the given hyperspectral data are located directly next to each other with respect to the electromagnetic spectrum. Consequently, the acquired reflectance values of neighboring spectral bands tend to be strongly correlated. This kind of redundancy typically decreases the quality of the achieved classification results, so that approaches for dimensionality reduction or band selection are commonly involved. While dimensionality reduction techniques focus on transforming the given data into a new space of lower dimensionality (van der Maaten et al., 2009), band selection techniques allow conclusions about relationships with respect to physical properties as they retain a subset of the original spectral bands (Guyon and Elisseeff, 2003; Saeys et al., 2007) which, in turn, can further be used for conclusions regarding a diversity of environmental applications. For dimensionality reduction, we apply a standard encoding of hyperspectral data using Principal Component Analysis (PCA). For band selection, we apply Correlation-based Feature Selection (Hall, 1999). Furthermore, we also apply a transformation of given hyperspectral data to multispectral Sentinel-2-like data of the same resolution (Weinmann et al., 2018). Such a transformation has already been proposed for the simulation of Sentinel-2 and other multispectral imagery (Thonfeld et al., 2012) and been applied for assessing land use (Elbertzhagen et al., 2012) or for geological and soil analyses (van der Meer et al., 2014). From the geometric data, we extract a set of commonly used low-level geometric features to describe the local neighborhood around each 3D point of the LiDAR data (Weinmann, 2016). For each 3D point, these geometric features are derived from the spatial arrangement and thus the coordinates of that 3D point and other 3D points within its local neighborhood, where a locally adaptive neighborhood definition is involved. Finally, the different sets of radiometric and geometric features are provided separately and in different combinations as input to a Random Forest classifier (Breiman, 2001).

This paper represents an extension of our previous work (Weinmann and Weinmann, 2018) with a particular focus on

1) further options regarding the involved sets of features and 2) a comprehensive analysis of the potential of different feature sets and their combination for urban scene interpretation. The choice of using hand-crafted features and a standard classifier is motivated by a scenario where only few training data are available and modern deep learning approaches therefore cannot be trained appropriately.

After briefly summarizing related work in Section 2, we present our framework for the semantic interpretation of urban environments on the basis of multi-modal data in Section 3. Subsequently, in Section 4, we present the results achieved with our framework on a benchmark dataset. These results are discussed in detail in Section 5. Finally, in Section 6, we provide concluding remarks and suggestions for future work.

2. RELATED WORK

Many investigations addressing the semantic interpretation of urban environments rely on data in the form of true orthophotos and the corresponding DSMs (Rottensteiner et al., 2012; Gerke, 2014). In this regard, the traditional approaches focus on extracting hand-crafted features and using standard classifiers such as Random Forests (Weinmann and Weinmann, 2018; Gerke and Xiao, 2014) or Conditional Random Fields (CRFs) (Gerke, 2014). In recent years, however, the use of modern deep learning techniques has become more and more popular, as such techniques allow jointly performing feature learning and classification with respect to the given classification task. Among a diversity of proposed network architectures, exemplary approaches rely on the use of a fully convolutional network (Sherrah, 2016), an encoder-decoder architecture (Volpi and Tuia, 2017) or an adaptation of the ResNet architecture (Chen et al., 2018) for the semantic interpretation of urban environments based on true orthophotos and the corresponding DSMs. Furthermore, different strategies have been proposed to fuse such multi-modal geospatial data within a deep learning framework (Marmanis et al., 2016; Audebert et al., 2016; Audebert et al., 2018; Liu et al., 2017).

While most of the related approaches focus on the classification part, only little attention has been paid to the given input data. Few investigations involve basic hand-crafted features represented by the Normalized Difference Vegetation Index (NDVI) and the normalized Digital Surface Model (nDSM) (Gerke, 2014; Audebert et al., 2016; Audebert et al., 2018; Liu et al., 2017). However, other types of radiometric or geometric features which can be extracted from a local neighborhood (Gerke and Xiao, 2014; Weinmann and Weinmann, 2018) have only rarely been involved, although, in the context of classifying aerial imagery based on given true orthophotos and the corresponding DSMs, it has recently been demonstrated that the additional consideration of such hand-crafted radiometric and geometric features on a per-pixel basis may lead to improved classification results (Chen et al., 2018).

In addition to the aforementioned progress, the use of hyperspectral data has also come into the focus of research on environmental mapping over years (Plaza et al., 2009; Camps-Valls et al., 2014), as such information e.g. allows distinguishing different types of vegetation (Bradley et al., 2018; Weinmann et al., 2018) and different materials (Ilehag et al., 2017). This, in turn, can be helpful if the corresponding geometric structure of the observed scene is similar. For instance, the use of co-registered hyperspectral imagery and LiDAR data for scene analysis has been proposed for tree species classification (Puttonen et al., 2010) as well as for civil engineering and urban planning applications (Brook et al., 2010) and in terms of semantic scene interpretation (Weinmann and Weinmann, 2018).

When using high-dimensional hyperspectral data, however, the high degree of redundancy contained in these data typically decreases the predictive accuracy of a classifier (Melgani and Bruzzone, 2004; Bradley et al., 2018), so that approaches for dimensionality reduction or band selection are commonly involved. In this context, dimensionality reduction techniques focus on deriving a new data representation based on fewer, but potentially better features extracted from the given data representation. For this purpose, standard approaches are represented by variants of Principal Component Analysis (PCA) (Licciardi et al., 2012) or Independent Component Analysis (ICA) (Wang and Chang, 2006; Villa et al., 2011). The use of such techniques typically allows a mapping of the given space spanned by the complete set of hyperspectral bands to a new space of lower dimensionality without a significant lack of information. However, the new space is spanned by meta-features, so that derived results hardly allow concluding about relationships with respect to physical properties as e.g. possible when considering the center wavelengths of involved spectral bands. In contrast, band selection techniques focus on retaining the most relevant and most informative spectral bands, while discarding less relevant and/or redundant spectral bands which, in turn, typically allows gaining predictive accuracy, improving computational efficiency with respect to both time and memory consumption, and retaining meaningful features with respect to the given classification task (Saeys et al., 2007; Guyon and Elisseeff, 2003).

3. METHODOLOGY

Our framework receives multi-modal data comprising co-registered RGB color imagery, hyperspectral data and LiDAR data as input and involves the major steps of feature extraction and classification, leading to an output in the form of respectively classified data. In the scope of this paper, we focus on different types of hand-crafted features. We define radiometric features on the basis of the given RGB color imagery and transformations to potentially better data representations (Section 3.1). Furthermore, we involve radiometric features defined on the basis of the given hyperspectral data and different encodings/transformations to potentially better data representations (Section 3.2). Besides the radiometric features, we also involve geometric features extracted from the given LiDAR data (Section 3.3). Finally, the different sets of radiometric and geometric features are provided separately and in different combinations as input to a Random Forest classifier (Section 3.4).

3.1 Definition of Radiometric Features on the Basis of RGB Color Imagery

On the one hand, we straightforwardly use pixel-wise radiometric features directly extracted from RGB color imagery which is typically used for a diversity of applications. On the other hand, we take into account that such RGB color representations are less robust with respect to changes in illumination, and we therefore also consider several transformations of RGB color imagery to a potentially better data representation. Each of these transformations results in

an image, where each pixel is characterized in a feature space spanned by three radiometric features (Figure 1).

3.1.1 RGB Color Imagery: We define a feature set $\mathcal{S}_{\text{RGB}}$ which comprises the reflectance values corresponding to the red (R), green (G), and blue (B) channels in the visible spectrum:

$$\mathcal{S}_{\text{RGB}} = \{R, G, B\} \tag{1}$$

3.1.2 Transformation of RGB Color Imagery to Chromaticity Values: We define a feature set $\mathcal{S}_{\text{RGB,norm}}$ which relies on color invariants in the form of normalized colors also known as chromaticity values (Gevers and Smeulders, 1999):

$$\mathcal{S}_{\text{RGB,norm}} = \left\{ \frac{R}{R+G+B}, \frac{G}{R+G+B}, \frac{B}{R+G+B} \right\} \tag{2}$$

Such color invariants are insensitive to surface orientation, illumination direction and illumination intensity.

3.1.3 Transformation of RGB Color Imagery to $c_1c_2c_3$ Space: We define a feature set $\mathcal{S}_{c_1c_2c_3}$ which relies on photometric color invariants for matte, dull surfaces (Gevers and Smeulders, 1999):

$$\mathcal{S}_{c_1c_2c_3} = \left\{ \arctan\left(\frac{R}{\max(G,B)}\right), \arctan\left(\frac{G}{\max(R,B)}\right), \arctan\left(\frac{B}{\max(R,G)}\right) \right\} \tag{3}$$

These color invariants are invariant to viewing direction, surface orientation, illumination direction and illumination intensity under the assumption of a dichromatic reflection model with white illumination.

3.1.4 Transformation of RGB Color Imagery to $l_1l_2l_3$ Space: We define a feature set $\mathcal{S}_{l_1l_2l_3}$ which relies on photometric color invariants for both matte and shiny surfaces (Gevers and Smeulders, 1999):

$$\mathcal{S}_{l_1l_2l_3} = \left\{ \frac{(R-G)^2}{(R-G)^2+(R-B)^2+(G-B)^2}, \frac{(R-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2}, \frac{(G-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2} \right\} \tag{4}$$

Like the $c_1c_2c_3$ color invariants, the $l_1l_2l_3$ color invariants are invariant to viewing direction, surface orientation, illumination direction and illumination intensity under the assumption of a dichromatic reflection model with white illumination. In addition, the $l_1l_2l_3$ color invariants are invariant to effects of surface reflection in the form of specular highlights.

3.1.5 Transformation of RGB Color Imagery via Comprehensive Color Image Normalization: We define a feature set $\mathcal{S}_{\text{CCIN}}$ which relies on Comprehensive Color Image Normalization (CCIN) (Finlayson et al., 1998). This approach is based on normalization procedures with respect to either lighting geometry or illuminant color, which are applied together and iteratively until convergence to a unique comprehensively normalized image.

3.1.6 Transformation of RGB Color Imagery via Gray-World Assumption: We define a feature set $\mathcal{S}_{\text{GW}}$ which relies on the gray-world (GW) hypothesis assuming that the average reflectance of surfaces in the scene is achromatic (Buchsbaum, 1980).

3.1.7 Transformation of RGB Color Imagery via Edge-Based Color Constancy: We define a feature set $\mathcal{S}_{\text{EBCC}}$ which relies on Edge-Based Color Constancy (EBCC) (van de Weijer et al., 2007). In this context, color constancy is defined as the ability to measure colors of objects independent of the color of the light source, and EBCC makes use of the gray-edge hypothesis assuming that the average edge difference in a scene is achromatic.

3.2 Definition of Radiometric Features on the Basis of Hyperspectral Data

For the sake of comparison, we straightforwardly involve radiometric features given with the original hyperspectral data. In addition, we take into account that the spectral bands of hyperspectral data are directly next to each other so that the acquired reflectance values of neighboring spectral bands tend to be strongly correlated and thus contain a high degree of redundancy which, in turn, typically has a detrimental effect on the classification results (Melgani and Bruzzone, 2004; Keller et al., 2016; Bradley et al., 2018). Consequently, we also involve several encodings/transformations of the given hyperspectral data.

3.2.1 Hyperspectral Data: We define a feature set $\mathcal{S}_{\text{HSI}}$ which comprises the reflectance values corresponding to a multitude of spectral bands, i.e. the reflectance values I across N_B spectral bands:

$$\mathcal{S}_{\text{HSI}} = \{I_1, \ldots, I_{N_B}\} \tag{5}$$

3.2.2 PCA-based Encoding of Hyperspectral Data: We define a feature set $\mathcal{S}_{\text{HSI,PCA}}$ by focusing on dimensionality reduction via the standard Principal Component Analysis (PCA). To reduce redundancy, the PCA uses an orthogonal transformation transferring the given hyperspectral data to a new space spanned by linearly uncorrelated variables which are referred to as principal components (PCs). The PCs are sorted with respect to the covered variability, so that the most relevant information is preserved in the first few PCs. In our work, we select the first few PCs covering 99.9% of the variability of the given training data, and we assume that all information preserved in the remaining PCs does not significantly contribute to the variability of the given data and can hence be discarded.

3.2.3 CFS-based Band Selection from Hyperspectral Data: We define a feature set $\mathcal{S}_{\text{HSI,CFS}}$ by focusing on band subset selection for which we apply Correlation-based Feature Selection (CFS) (Hall, 1999), i.e. we aim at reducing the redundancy preserved in the given hyperspectral data by selecting a set of spectral bands ("features") that seem to be particularly relevant with respect to the considered classification task. The CFS takes into account 1) the correlation between spectral bands and classes to identify relevant spectral bands and 2) the correlation among spectral bands to identify and discard redundant spectral bands.

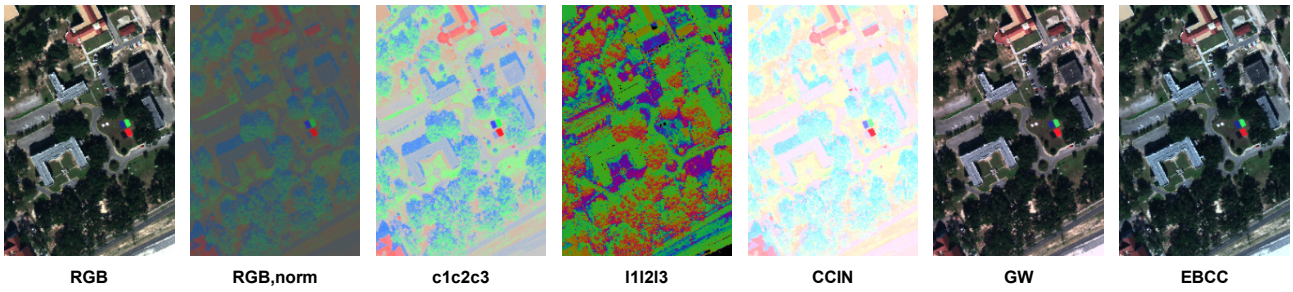

Figure 1. RGB color image and transformations derived via normalized colors, $c_1c_2c_3$ space, $l_1l_2l_3$ space, Comprehensive Color Image Normalization (CCIN), Gray-World (GW) assumption and Edge-Based Color Constancy (EBCC).

3.2.4 Transformation of Hyperspectral Data to Multispectral Sentinel-2-like Data: We define a feature set $\mathcal{S}_{\text{HSI}\to\text{S2}}$ with which we aim at reducing the redundancy preserved in the given hyperspectral data by applying a transformation to multispectral data, where the neighboring spectral bands are well-separated by a sufficiently large margin across wavelengths. More specifically, we apply a transformation of the given hyperspectral data to multispectral Sentinel-2-like data of the same spatial resolution (Weinmann et al., 2018). In this context, the spectral bands of the hyperspectral data that are within the individual spectral bands defined for Sentinel-2 data are used to compute a weighted mean of the respective reflectance values, where the weights are determined via a linear interpolation based on the Sentinel-2 Spectral Response Functions (S2-SRFs) normalized to 1 as shown in Figure 2. After performing this transformation, we take into account that the atmospheric transmission is generally low for the spectral bands B_1, B_9 and B_{10} of Sentinel-2 data. This can be attributed due to ozone, oxygen or water vapor which strongly affect the atmospheric transmissivity at certain wavelengths (Weinmann et al., 2018). Furthermore, we account for the overlap of the spectral bands B_8 and B_{8a}, where the former is much wider and can hence be considered as less characteristic. Consequently, we discard the reflectance values corresponding to the spectral bands B_1, B_8, B_9 and B_{10}, and we accordingly focus on the remaining spectral bands.

3.3 Extraction of Geometric Features

From the LiDAR data, we extract geometric features which describe the local surface structure around a 3D point $\mathbf{X}$, i.e. the spatial arrangement of 3D points within a local neighborhood. We specify this local neighborhood for each 3D point by applying spherical neighborhoods defined by one scale parameter represented by the number k of nearest neighbors to be considered. This neighborhood definition allows for a variable radius and thus a variable absolute size which might be preferable in case of strongly varying point density. Instead of the straightforward solution of selecting an identical value of the scale parameter k for each point of the LiDAR data, we take into account that a suitable size of the local neighborhood depends on the local 3D structure and the classes of interest (Weinmann, 2016). Consequently, we focus on a data-driven approach for optimal neighborhood size selection for which we apply eigenentropy-based scale selection (Weinmann, 2016). To determine the optimal size of the local neighborhood for each individual 3D point, this approach relies on taking different values k (here: $k = 10, \ldots, 100$) and, for each k, calculating the eigenentropy (i.e. the disorder of 3D points) based on the eigenvalues of the respective 3D structure tensor. Finally, the locally optimal neighborhood size is derived by selecting the value for k that yields the minimum eigenentropy. Based on the defined local neighborhoods, we derive the 3D structure tensor and its eigenvalues for each 3D point of the LiDAR data. The eigenvalues, in turn, are used to extract the geometric features of linearity, planarity, sphericity, omnivariance, anisotropy, eigenentropy, sum of eigenvalues and local surface variation (West et al., 2004; Pauly et al., 2003). Further geometric features are defined by the height of $\mathbf{X}$, the radius of the local neighborhood, the local point density, the verticality and the maximum difference and standard deviation of the height values of those points within the local neighborhood of $\mathbf{X}$.

3.4 Supervised Classification

The defined sets of radiometric and geometric features are considered separately and in different combinations. For each case, the respective features are concatenated to a feature vector and provided as input to a classifier.

For classification, we assume that only few training data are available, which is realistic for practical applications involving hyperspectral data. Consequently, we focus on a standard supervised classification of given feature vectors. As classifier, we use a Random Forest classifier (Breiman, 2001) as representative of modern discriminative methods. Such a classifier relies on strategically generating a set of weak learners in the form of decision trees via bootstrap aggregating ("bagging") (Breiman, 1996), where a predefined number of these weak learners are trained independently from each other on subsets of the training data which are randomly drawn with replacement. The random sampling results in randomly different decision trees and thus in diversity in terms of de-correlated hypotheses across the individual trees. Given the trained classifier, the classification of an unseen feature vector takes into account that each decision tree casts a vote for one of the class labels and that the majority vote across the individual votes represents a robust classification output.

The most important parameters of a Random Forest classifier are the number N_T of involved decision trees, the minimum number N_S of samples to allow a tree node to be split, the number N_a of active variables to be used for the test in each tree node, and the maximum tree depth D_{max}. To appropriately specify these internal parameters of the classifier, we conduct a grid search for N_T on a suitable subspace, while the remaining settings are defined following the recommendations of the openCV implementation. Accordingly, a node is only split if it is reached by at least $N_S = 20$ training samples, while the number N_a of active variables for each test is set to $N_a = \sqrt{N_F}$ with N_F being the number of features considered for the respective case. The maximum tree depth is set to $D_{\text{max}} = 15$.

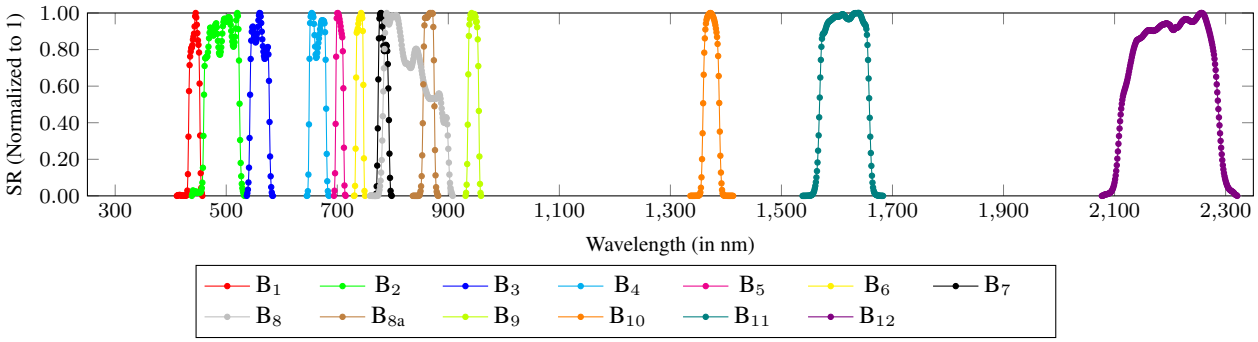

Figure 2. Visualization of the Sentinel-2 Spectral Response Functions (S2-SRFs), i.e. the measured spectral response (SR) for each band of the Sentinel-2 MultiSpectral Instrument (S2-MSI).

4. EXPERIMENTAL RESULTS

In the following, we first describe the used dataset (Section 4.1). Subsequently, we explain implementation details (Section 4.2) and summarize the conducted experiments (Section 4.3). Finally, we present the achieved results (Section 4.4).

4.1 Dataset

For performance evaluation, we use the MUUFL Gulfport Hyperspectral and LiDAR Airborne Data Set (Gader et al., 2013; Zare et al., 2016) comprising co-registered RGB, hyperspectral and LiDAR data acquired in 2010 with an aerial sensor platform from altitudes of about 3500-6700 ft over the University of Southern Mississippi Gulf Park Campus in Long Beach, Mississippi, USA. More specifically, the hyperspectral data were acquired with an ITRES Compact Airborne Spectrographic Imager (CASI-1500) delivering measurements on 72 spectral bands. These spectral bands cover the wavelength interval between 367.7 nm and 1043.4 nm (i.e. the visible and near-infrared (VNIR) domain) with a spectral sampling varying from 9.5 nm to 9.6 nm. For the analyses, it has been taken into account that the first four and the last four of these spectral bands correspond to parts of the electromagnetic spectrum where the atmospheric transmission is generally low due to effects arising from ozone, oxygen or water vapor in the atmosphere. Consequently, the measurements corresponding to those spectral bands have been discarded, and only the information preserved in the remaining 64 spectral bands has been released. As a complementary type of data, the LiDAR data were acquired with an Optech Gemini Airborne Laser Terrain Mapper (ALTM) relying on a laser with a wavelength of 1064 nm. Both types of data have been co-registered and transferred to a discrete image grid of 325×220 pixels, where each pixel corresponds to an area of $1\,\text{m} \times 1\,\text{m}$ (i.e. all data are given with a ground sampling distance of 1 m). Accordingly, the study area has a size of 7.15 ha.

Together with the co-registered RGB, hyperspectral and LiDAR data, a reference labeling with respect to 11 semantic classes as well as a further class for unlabeled data is provided (Du and Zare, 2017) as shown in Figure 3. This reference labeling addresses a variety of ground cover types (trees, grass, dirt, sand, water, etc.) and structural scene elements (road, buildings, sidewalk, curb, etc.).

4.2 Implementation

We implemented our framework in Matlab and used external software packages for the CFS implementation (Zhao et al., 2010) and for the Random Forest implementation (Liaw and Wiener, 2002). All experiments are run on a standard laptop computer (Intel Core i7-6820HK, 2.7 GHz, 16 GB RAM).

4.3 Experiments

For our experiments, we split the given dataset into disjoint sets of 1) training examples used for training the classifier and 2) test examples used for performance evaluation. In this context, we discard all pixels labeled as C12 ("unlabeled") and then randomly select an identical number of 100 examples for each of the 11 remaining classes as training data, while the 52,587 remaining examples are used as test data. The relatively small number of training examples per class can be considered as realistic for practical applications.

In our experiments, we provide the different sets of radiometric and geometric features separately and in different combinations as input to a Random Forest classifier. In particular, we focus on the use of different sets of radiometric features with and without additional 3D shape information. For each case, the classifier is trained on the training data, while performance evaluation is done on the test data. As evaluation metrics, we consider the overall accuracy (OA), the κ-index (κ), the mean F_1-score across all classes (mF_1) and the mean Intersection-over-Union (mIoU).

4.4 Results

Using the defined sets of radiometric and geometric features separately and in different combinations as input to the Random Forest classifier, we achieve the classification results summarized in Table 1 and visualized in Figure 4 for the complete scene. The derived results vary from about 50 % to about 79 % in OA, and they clearly indicate the potential of the different feature sets as the basis for classification.

The derived results reveal that the use of RGB color information or 3D shape information alone results in an OA of about 50-51 % and is therefore not sufficient to achieve reasonable classification results. Using transformations of RGB color imagery to a potentially better data representation, a significant gain of about 10-11 % in OA may be achieved when using chromaticity values ($\mathcal{S}_{\text{RGB,norm}}$), the $c_1c_2c_3$ space ($\mathcal{S}_{c_1c_2c_3}$) or Comprehensive Color Image Normalization ($\mathcal{S}_{\text{CCIN}}$) instead of the original RGB color imagery ($\mathcal{S}_{\text{RGB}}$). Using the gray-world hypothesis ($\mathcal{S}_{\text{GW}}$) or the gray-edge hypothesis ($\mathcal{S}_{\text{EBCC}}$) also leads to improved classification results in comparison to the use of the original RGB color imagery ($\mathcal{S}_{\text{RGB}}$), where the improvement is about 3-4 % in OA. Only the use of the $l_1l_2l_3$

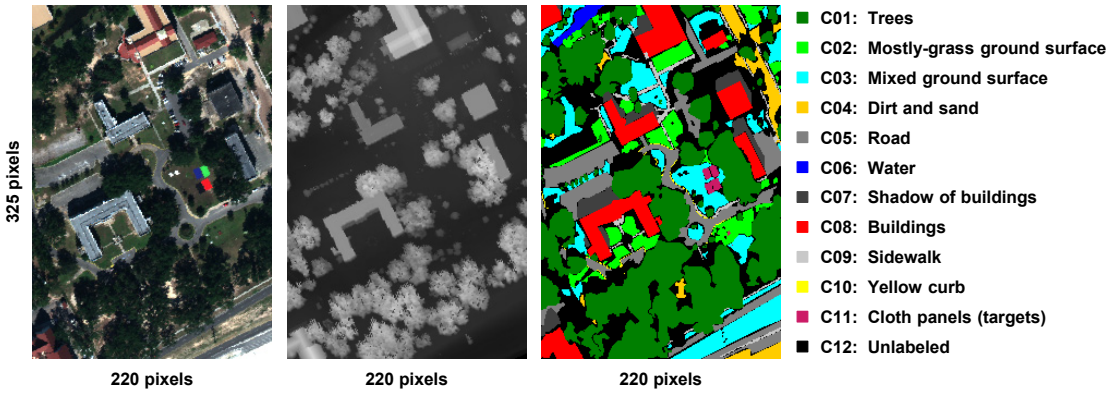

Figure 3. RGB color image, the corresponding LiDAR data represented digital surface model (DSM) and the given reference labeling with respect to 11 semantic classes as well as a further class for unlabeled data.

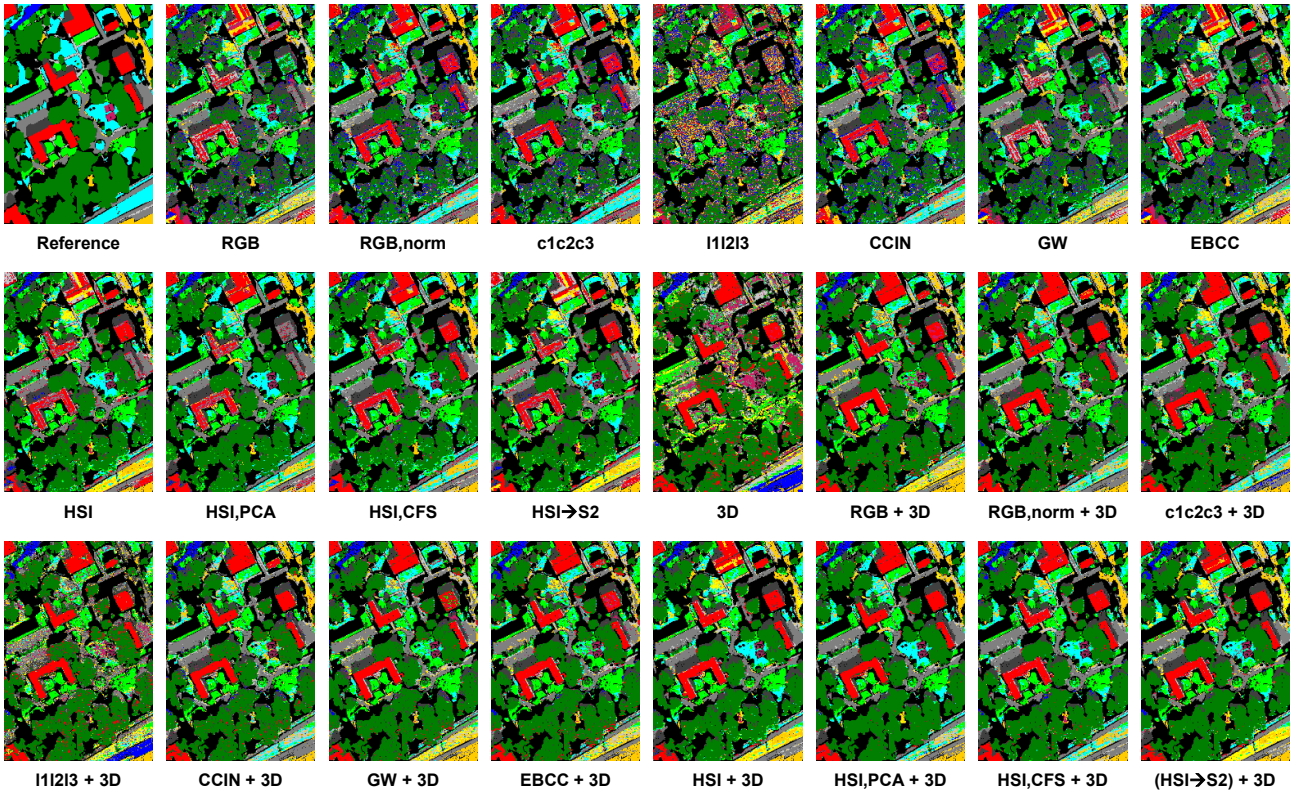

Figure 4. Reference labeling and the classification results achieved for different feature sets and their combination.

space ($\mathcal{S}_{l_1 l_2 l_3}$) does not seem to be appropriate for the given classification task.

The use of hyperspectral data seems to be even more beneficial. When using the original hyperspectral data ($\mathcal{S}_{\text{HSI}}$), the improvement is about 18 % in OA in comparison to the case in which the original RGB color imagery is used. However, the PCA-based encoding of the original hyperspectral data ($\mathcal{S}_{\text{HSI,PCA}}$) yields a further improvement of about 7 % in OA, whereas the CFS-based band selection from hyperspectral data ($\mathcal{S}_{\text{HSI,CFS}}$) yields an improvement of about 5 % in OA in comparison to the use of the original hyperspectral data ($\mathcal{S}_{\text{HSI}}$). Interestingly, the use of multispectral Sentinel-2-like data on the same spatial resolution ($\mathcal{S}_{\text{HSI}\rightarrow\text{S2}}$) leads to similar results to those achieved via the original hyperspectral data.

Finally, the consideration of geometric features in addition to radiometric features yields a significant improvement. This improvement is about 3-8 % in OA in case of involved radiometric data in the form of hyperspectral data and different encodings/transformations, whereas the improvement is about 11-24 % in OA in case of involved radiometric data in the form of RGB color imagery and different transformations.

5. DISCUSSION

The finding that the use of RGB color information or 3D shape information alone is not sufficient to achieve reasonable classification results might seem obvious as some of the classes exhibit similar RGB colors and/or a similar geometric behavior when focusing on the local structure on the given raster with 1 point/m^2 (Figure 3). The RGB color information allows for the separation of classes with a different appearance, even if they exhibit a similar geometric behavior, and thus allows separating natural ground cover classes such as “mixed ground surface” and “dirt and sand” from man-made ground cover classes such as “road” and “sidewalk”. Using the different transformations

Feature Set	N_F	OA	κ	mF$_1$	mIoU
$\mathcal{S}_{\text{RGB}}$	3	51.31	43.02	39.60	26.82
$\mathcal{S}_{\text{RGB,norm}}$	3	62.73	55.14	49.62	35.52
$\mathcal{S}_{c_1c_2c_3}$	3	61.94	54.27	51.06	36.70
$\mathcal{S}_{l_1l_2l_3}$	3	37.17	25.82	21.53	13.98
$\mathcal{S}_{\text{CCIN}}$	3	61.16	53.37	47.69	33.76
$\mathcal{S}_{\text{GW}}$	3	54.66	45.83	40.47	27.53
$\mathcal{S}_{\text{EBCC}}$	3	55.70	47.07	41.75	28.67
$\mathcal{S}_{\text{HSI}}$	64	68.96	61.04	54.82	40.36
$\mathcal{S}_{\text{HSI,PCA}}$	15	76.05	69.40	64.71	50.14
$\mathcal{S}_{\text{HSI,CFS}}$	11	74.31	67.77	63.83	48.32
$\mathcal{S}_{\text{HSI}\rightarrow\text{S2}}$	7	67.80	59.78	54.29	39.99
$\mathcal{S}_{\text{3D}}$	14	49.76	36.94	26.69	18.44
$\mathcal{S}_{\text{RGB + 3D}}$	17	69.54	61.32	51.34	38.14
$\mathcal{S}_{\text{RGB,norm + 3D}}$	17	74.59	67.70	59.02	44.92
$\mathcal{S}_{c_1c_2c_3\text{ + 3D}}$	17	73.88	66.40	56.69	42.85
$\mathcal{S}_{l_1l_2l_3\text{ + 3D}}$	17	60.94	50.47	38.92	27.60
$\mathcal{S}_{\text{CCIN + 3D}}$	17	72.46	64.99	58.05	43.61
$\mathcal{S}_{\text{GW + 3D}}$	17	69.41	61.12	52.51	38.97
$\mathcal{S}_{\text{EBCC + 3D}}$	17	69.70	61.59	51.56	38.03
$\mathcal{S}_{\text{HSI + 3D}}$	78	74.88	68.36	61.91	48.57
$\mathcal{S}_{\text{HSI,PCA + 3D}}$	29	79.24	73.17	69.93	57.08
$\mathcal{S}_{\text{HSI,CFS + 3D}}$	25	79.04	73.24	65.87	51.65
$\mathcal{S}_{(\text{HSI}\rightarrow\text{S2})\text{ + 3D}}$	21	75.34	68.52	60.23	46.82

Table 1. Number N_F of involved features and achieved classification results (OA, κ, mF$_1$ and mIoU in %) when using different feature sets as the basis for classification.

of RGB color imagery leads to a gain in OA in most cases which, in turn, indicates that most of the applied techniques relying on color invariants, normalization procedures or specific assumptions about the color representation of a scene provide a better basis for the considered classification task. This can be attributed to the fact that RGB color representations are less robust with respect to changes in illumination, whereas some of the applied transformations even provide invariance properties with respect to changes in viewing direction, object geometry and illumination.

Using hyperspectral data seems to generally allow for a much better differentiation of the defined classes due to the fact that they contain reflectance information across numerous spectral bands reaching from the visible domain to the near-infrared domain. However, it can be observed that involving techniques for dimensionality reduction (e.g. PCA) or feature selection (e.g. CFS) yields a significant improvement of about 5-7 % in OA which, in turn, indicates that reducing the redundancy contained in the reflectance values of neighboring spectral bands has a beneficial impact on classification.

Interestingly, the use of multispectral Sentinel-2-like data on the same resolution leads to classification results of almost the same quality as given for the use of the original hyperspectral data. This indicates that multispectral Sentinel-2-like data already provide a good source of information for the considered classification task involving a variety of ground cover types and structural scene elements. Such an effect has recently also been reported for land cover and land use classification (Weinmann et al., 2018). This is of particular interest as multispectral Sentinel-2 data can currently be acquired with lower spatial resolution, but with short revisit times which, in turn, allows for multitemporal analyses taking into account seasonal changes, growth cycles or other dynamic processes.

In contrast, the 3D shape information does not allow separating classes with a similar geometric behavior as given for the classes "mostly-grass ground surface", "mixed ground surface", "dirt and sand", "road", "sidewalk" and "water", yet the 3D shape information allows the differentiation of classes with different geometric characteristics which, in particular, leads to a much better recognition of the class "buildings" (Figure 4).

The significant improvement obtained when using a combination of radiometric and geometric features indicates the synergetic effect of using complementary types of features. While the radiometric features allow reasoning about materials, the geometric features allow reasoning about the shape of objects. The significant gain in OA, κ, mF$_1$ and mIoU (Table 1) indicates both a better overall performance and a significantly improved recognition of instances across all classes.

6. CONCLUSIONS

In this paper, we have addressed the semantic interpretation of urban environments on the basis of multi-modal data acquired from aerial sensor platforms. In this context, we have assessed the potential of RGB color imagery, hyperspectral data and LiDAR data as well as a variety of potentially better data representations derived from these. In this regard, we have transformed the RGB color imagery via techniques relying on color invariants, normalization procedures or specific assumptions about the color representation of a scene. For the hyperspectral data, we have taken into account specific encodings/transformations derived via dimensionality reduction or band selection techniques which have proven beneficial in numerous investigations. Furthermore, we have transformed the hyperspectral data to high-resolution multispectral Sentinel-2-like data. Using different sets of radiometric and/or geometric features separately and in different combinations as input to a Random Forest classifier, we have demonstrated that particularly the use of shape information in combination with hyperspectral information or respective data encodings/transformations leads to classification results of rather good quality, even for a challenging scene acquired with a low spatial resolution of 1 point/m^2. However, the results have also revealed further interesting insights. On the one hand, commonly involved combinations of radiometric and geometric features such as the combination of RGB color imagery and 3D data should be adapted by a transformation of the considered color space. On the other hand, the use of the original high-dimensional hyperspectral data should be avoided, because different encodings/transformations of these lead to classification results of similar or even better quality.

In future work, we plan to address the fact that neighboring data points tend to be strongly correlated, i.e. class labels typically change only at the boundaries of objects, while most of the data points within a local neighborhood belong to the same class. Accordingly, a class change should be unlikely for neighboring data points and rather be related to a change of the respective features. To achieve a smoother labeling, we plan to use either smoothing techniques or spatial regularization techniques (Schindler, 2012; Landrieu et al., 2017).

REFERENCES

Audebert, N., Le Saux, B., Lefèvre, S., 2016. Semantic segmentation of Earth observation data using multimodal and multi-scale deep networks. *Proc. 13th Asian Conference on Computer Vision*, I, 180–196.

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140, 20–32.

Bradley, P. E., Keller, S., Weinmann, M., 2018. Unsupervised feature selection based on ultrametricity and sparse training data: a case study for the classification of high-dimensional hyperspectral data. *Remote Sensing*, 10(10), 1564:1–1564:35.

Breiman, L., 1996. Bagging predictors. *Machine Learning*, 24(2), 123–140.

Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), 5–32.

Brook, A., Ben-Dor, E., Richter, R., 2010. Fusion of hyperspectral images and lidar data for civil engineering structure monitoring. *Proc. 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 1–5.

Buchsbaum, G., 1980. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310(1), 1–26.

Camps-Valls, G., Tuia, D., Bruzzone, L., Benediktsson, J. A., 2014. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. *IEEE Signal Processing Magazine*, 31(1), 45–54.

Chen, K., Weinmann, M., Gao, X., Yan, M., Hinz, S., Jutzi, B., Weinmann, M., 2018. Residual shuffling convolutional neural networks for deep semantic image segmentation using multi-modal data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2, 65–72.

Du, X., Zare, A., 2017. Technical report: scene label ground truth map for MUUFL Gulfport Data Set. Technical report, University of Florida, Gainesville, FL, USA.

Elbertzhagen, I., Thonfeld, F., Menz, G., 2012. SVM-based agricultural land use assessment using Sentinel-2 simulation data. *Proc. Sentinel-2 Preparatory Symposium*.

Finlayson, G. D., Schiele, B., Crowley, J. L., 1998. Comprehensive colour image normalization. *Proc. European Conference on Computer Vision*, 475–490.

Gader, P., Zare, A., Close, R., Aitken, J., Tuell, G., 2013. MUUFL Gulfport Hyperspectral and LiDAR Airborne Data Set. Technical report, REP-2013-570, University of Florida, Gainesville, FL, USA.

Gerke, M., 2014. Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen). Technical report, ITC, University of Twente, Twente, The Netherlands.

Gerke, M., Xiao, J., 2014. Fusion of airborne laserscanning point clouds and images for supervised and unsupervised scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 87, 78–92.

Gevers, T., Smeulders, A. W. M., 1999. Color based object recognition. *Pattern Recognition*, 32(3), 453–464.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.

Hall, M. A., 1999. Correlation-based feature subset selection for machine learning. PhD thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

Ilehag, R., Weinmann, M., Schenk, A., Keller, S., Jutzi, B., Hinz, S., 2017. Revisiting existing classification approaches for building materials based on hyperspectral data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-3/W3, 65–71.

Keller, S., Braun, A. C., Hinz, S., Weinmann, M., 2016. Investigation of the impact of dimensionality reduction and feature selection on the classification of hyperspectral EnMAP data. *Proc. 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 1–5.

Landrieu, L., Raguet, H., Vallet, B., Mallet, C., Weinmann, M., 2017. A structured regularization framework for spatially smoothing semantic labelings of 3D point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 132, 102–118.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. *R News*, 2/3, 18–22.

Licciardi, G., Marpu, P. R., Chanussot, J., Benediktsson, J. A., 2012. Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles. *IEEE Geoscience and Remote Sensing Letters*, 9(3), 447–451.

Liu, Y., Piramanayagam, S., Monteiro, S. T., Saber, E., 2017. Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order CRFs. *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1561–1570.

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., Stilla, U., 2016. Semantic segmentation of aerial images with an ensemble of CNNs. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, III-3, 473–480.

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

Pauly, M., Keiser, R., Gross, M., 2003. Multi-scale feature extraction on point-sampled surfaces. *Computer Graphics Forum*, 22(3), 81–89.

Plaza, A., Benediktsson, J. A., Boardman, J. W., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Marconcini, M., Tilton, J. C., Trianni, G., 2009. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113, S110–S122.

Puttonen, E., Suomalainen, J., Hakala, T., Räikkönen, E., Kaartinen, H., Kaasalainen, S., Litkey, P., 2010. Tree species classification from fused active hyperspectral reflectance and lidar measurements. *Forest Ecology and Management*, 260(10), 1843–1852.

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U., 2012. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, I-3, 293–298.

Saeys, Y., Inza, I., Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507–2517.

Schindler, K., 2012. An overview and comparison of smooth labeling methods for land-cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11), 4534–4545.

Sherrah, J., 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. arXiv:1606.02585.

Thonfeld, F., Feilhauer, H., Menz, G., 2012. Simulation of Sentinel-2 images from hyperspectral data. *Proc. Sentinel-2 Preparatory Symposium*.

van de Weijer, J., Gevers, T., Gijsenij, A., 2007. Edge-based color constancy. *IEEE Transactions on Image Processing*, 16(9), 2207–2214.

van der Maaten, L. J. P., Postma, E. O., van den Herik, H. J., 2009. Dimensionality reduction: a comparative review. Technical report, Tilburg University, Tilburg, The Netherlands.

van der Meer, F. D., van der Werff, H. M. A., van Ruitenbeek, F. J. A., 2014. Potential of ESA's Sentinel-2 for geological applications. *Remote Sensing of Environment*, 148, 124–133.

Villa, A., Benediktsson, J. A., Chanussot, J., Jutten, C., 2011. Hyperspectral image classification with independent component discriminant analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 49(12), 4865–4876.

Volpi, M., Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2), 881–893.

Wang, J., Chang, C.-I., 2006. Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 44(6), 1586–1600.

Weinmann, M., 2016. *Reconstruction and analysis of 3D scenes – From irregularly distributed 3D points to object classes*. Springer, Cham, Switzerland.

Weinmann, M., Maier, P. M., Florath, J., Weidner, U., 2018. Investigations on the potential of hyperspectral and Sentinel-2 data for land-cover / land-use classification. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-1, 155–162.

Weinmann, M., Weinmann, M., 2018. Geospatial computer vision based on multi-modal data – How valuable is shape information for the extraction of semantic information? *Remote Sensing*, 10(1), 2:1–2:20.

West, K. F., Webb, B. N., Lersch, J. R., Pothier, S., Triscari, J. M., Iverson, A. E., 2004. Context-driven automated target detection in 3-D data. *Proc. SPIE*, 5426, 133–143.

Zare, A., Jiao, C., Glenn, T., 2016. Multiple instance hyperspectral target characterization. arXiv:1606.06354v2.

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., Liu, H., 2010. Advancing feature selection research – ASU feature selection repository. Technical report, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA.