PFG 2012 / 4, 0371–0384
Stuttgart, August 2012

Article

Kernel Composition with the one-against-one Cascade for Integrating External Knowledge into SVM Classification

Andreas Ch. Braun, Uwe Weidner, Boris Jutzi & Stefan Hinz, Karlsruhe

Keywords: data fusion, knowledge integration, SVM, hyperspectral, airborne laserscanning

Summary: This work focuses on two main questions. How can data fusion be performed before SVM (support vector machine) classification? And secondly: how can the one-against-one cascade be exploited to use information selectively thus integrating human knowledge? Kernel composition represents a specialized method for fusing data on the feature level. Its main advantage is given by the fact that it reduces the Hughes phenomenon (performance decrease due to high dimensionality) because it abstains from raising dimensionality in the feature space. Since the paper focuses on hyperspectral data, a specialized kernel based on the spectral angle is employed and evaluated. Two application schemes are presented. At first, hyperspectral data are fused with laserscanning data by taking into account explicit knowledge on roof geometries. Secondly, a spectral-spatial framework for hyperspectral data is presented which integrates implicit knowledge on the relevance of spatial context into classification. Both approaches are promising as they obtain higher classification accuracies when integrating external knowledge. The innovation of the contribution is that data fusion with a second source of data via kernel composition is combined with a modification of the one-against-one cascade which allows integration of human knowledge.

Zusammenfassung: *Verknüpfung von Kernfunktionen mit der eins-gegen-eins Kaskade für die Einbindung von Wissen in die SVM Klassifizierung.* Dieser Beitrag vertieft zwei Hauptfragen. Wie kann die Datenfusion für die SVM Klassifizierung vorgenommen werden? Und zweitens: wie kann die eins-gegen-eins Kaskade genutzt werden, um Information selektiv zu nutzen und menschliches Wissen einzubringen? Die Verknüpfung von Kernfunktionen stellt eine spezielle Methode der Datenfusion für kernbasierte Klassifikatoren wie Stützvektormaschinen (support vector machines, SVM) dar. Ihr Hauptvorteil besteht darin, dass dadurch das Hughes Phänomen (Performanzverlust durch hohe Dimensionalität) reduziert wird, indem sie es vermeidet, die Dimensionalität des Merkmalsraums zu erhöhen. Da sich der Beitrag mit hyperspektralen Daten beschäftigt, wird eine spezielle Kernfunktion, die auf dem spektralen Winkel basiert, eingesetzt und bewertet. Zwei Anwendungsschemata werden vorgestellt. Zuerst werden Hyperspektraldaten mit Laserscanningdaten fusioniert, wobei explizites Wissen über Dachgeometrien genutzt wird. Danach wird ein spektral-räumlicher Klassifizierungsansatz vorgestellt, welcher implizites Wissen über die Relevanz des räumlichen Kontextes in die Klassifizierung einbringt. Beide Ansätze sind vielversprechend, da sie höhere Klassifizierungsgenauigkeiten erzielen, wenn Wissen genutzt wird. Die Innovation des Beitrages ist, dass eine zweite Datenquelle über die Verknüpfung von Kernfunktionen kombiniert wird mit einer Modifikation der eins-gegen-eins Kaskade, die es erlaubt, Wissen zu integrieren.

1 Introduction

For many applications in remote sensing, it is beneficial to use multiple data sources (Amarsaikhan & Douglas 2004). By rendering classification results more reliably, a better interpretation of the resulting maps is ensured, which finally can help to make better decisions. Various approaches for multisource classification have been pub-

www.schweizerbart.de
DOI: 10.1127/1432-8364/2012/0124
1432-8364/12/0124 $ 3.50

lished based on state-of-the-art classifiers like Markov random fields (Solberg et al. 1996), neural networks (Paola & Schowengerdt 1995), fuzzy classifiers (Binaghi et al. 1997), or combined classifiers (Huang & Lees 2004). Many approaches based on support vector machines (SVM) have been published as well. SVMs are a group of supervised learning methods that can be applied to classification or regression (Burges 1998). Watanachaturaporn et al. (2008) use an approach based on feature concatenation. Halldorsson et al. (2003) propose a multisource framework for SVM which modifies the RBF (radial basis function) kernel by using the distances of data points in various sources as features. Waske et al. (2007) classify two data sources separately. Then, the outputs of the SVM decision functions are used as new features which are again classified. Approaches for the fusion of hyperspectral with laserscanning data are presented by e.g. Jones et al. (2010) and Voss & Sugumaran (2008). A crucial prerequisite of all of these classification approaches is data fusion. For kernel based classifiers (like SVMs) a specialized method is offered by kernel composition. It exploits the fact, that kernel functions can be combined (e.g. by addition, multiplication, or weighting) to form new kernels. Thus, different kernels computed on different data sources can be combined. The first main focus of this contribution is to exemplify multisource classification by data fusion via kernel composition using SVMs. A second focus is also discussed. Adapting the SVM from binary to multiclass classification is frequently done via the one-against-one (OaO) cascade. Instead of considering all of the n classes in one step, the OaO cascade considers only two classes at a time. For each of the $\boldsymbol{n} \cdot (\boldsymbol{n} - 1)/2$ combinations of classes, an individual SVM is trained. Each SVM assigns a class label to each point. Afterwards, the final class label for each point is the label that has been most frequently assigned by the individual SVMs (Braun et al. 2010). Since the cascade considers only two of the n classes, it can be exploited to use additional sources of information selectively, i.e. only when the human operator considers it discriminative for the distinction of two particular classes. Thus, external knowledge on the classes available to the operator can be considered.

Both aspects are demonstrated and discussed on two application schemes. The first is multisource classification of hyperspectral and laserscanning data based on our previous work in Braun et al. (2011). The second is spectral-spatial classification on a hyperspectral benchmark dataset. Many researchers have made much effort on integrating the spatial context of pixel neighbourhoods into classification (Plaza et al. 2009). These spectral-spatial approaches slightly differ from the multisource case (fusion of hyperspectral and laserscanning data) since the spatial data source is derived from the spectral data source. However, an important issue which both have in common is the necessity to fuse two different sources of information. As a consequence, both multisource and spectral-spatial approaches need methodologies for data fusion. The main objectives of this paper are therefore data fusion and the way in which the OaO cascade can be used to integrate knowledge into classification. In the first application scheme, the knowledge is explicit knowledge of the operator about the shape of roofs. In the second scheme, it is implicit knowledge on the separability of classes. The OaO cascade allows using this knowledge in SVM classification. The remainder of the paper is organized as follows. Section 2 presents a mathematical introduction to the methods used. Section 3 outlines the data preprocessing, data fusion and classification of an urban dataset composed of hyperspectral and laserscanning data. In section 4, the transferability of the proposed framework will be shown for a spectral-spatial classification approach. In section 5, both classification approaches are discussed and compared in a synoptic manner, while section 6 concludes the paper.

2 Mathematical Foundations

This section provides briefly the mathematical foundations of the methods used. Kernel composition utilizes the Mercer property that kernels can be combined by addition, multiplication or ratio formation. SVMs ensure generalization by accepting a small amount of errors on the training data. This paper uses the ν-SVM by Schölkopf et al. (2000). The choice of data preprocessing and error handling strategy used herein is based on Braun et al. (2011), where a more thorough discussion on the different types can be found.

2.1 *Kernels and the Support Vector Machine*

Given a dataset X with n data points, kernel matrices $K(x_i,x_j)$ are the result of kernel functions applied over all n^2 pairs of data tuples. The outcome of a kernel function $K(x_i,x_j) = f\delta(x_i,x_j)$ is a similarity measure for the two training data x_i and x_j depending on some distance metric δ. φ is usuallythe Euclidean distance (Mercier & Lennon 2003). However, kernel functions can be modified e.g. by introducing different similarity measures (Amari & Wu 1999). To model complex distributions of the training data in the feature space, $f\varphi$ is usually some non-linear function. The most frequently applied family of non-linear functions are Gaussian radial basis functions (RBF) (Mercier & Lennon 2003, Amari & Wu 1999). The closer two points in the feature space are, the higher the resulting kernel value is. Given these facts, the kernel matrix simply represents the similarity between the points of the training dataset. To understand how the kernel matrix is used in SVM classification, it is helpful not to look at the primal, but the dual formulation of the SVM problem. The dual problem is given by (1).

$$\text{maximize: } \sum_i \lambda_i - \frac{1}{2}\sum_i\sum_j \lambda_i\lambda_j y_i y_j K(x_i,x_j) \tag{1}$$

with

$\lambda_i > 0 \ \forall$ support vectors,
$\lambda_i = 0 \ \forall$ other points.

The Lagrange multipliers λ_i are only greater than zero for the support vectors (SVs) (see constraint of (1)). Hence, only training data which are SVs contribute to the solution of (1). For all cases where at least one of the points is not a support vector, $\lambda_i \cdot \lambda_j = 0$. Thus, the double sum term of (1) is also zero. Hence, if at least one of the points is not a support vector, the couple of points does not contribute to maximizing (1). The class labels y_i are in $\{-1,1\}$. Since the second part of (1) is subtracted, only points with different class labels can maximize the term. Since their product $y_i \cdot y_j = -1$, the double sum term is made positive and the double sum of this combination of points can contribute to maximizing (1). The problem is usually optimized by an optimization procedure. This procedure sequentially assigns λ_i to the training data. It converges if points are chosen as SVs which have different class labels but are found close to each other in the feature space. Such points will yield a high value in the kernel matrix $K(x_i,x_j)$. Their double sum of (1) will thus yield a high value and contribute strongly to maximizing (1). To conclude, it can be said that the similarity values of the kernel matrix are used for finding the best suited training points as SVs. By not letting non-SV points (which have $\lambda_i = 0$) influence (1), a sparse solution is found which only depends on the SVs (which have $\lambda_i > 0$).

2.2 *Spectral Angle Kernels*

In Braun et al. (2011) the linear kernel is used predominantly. It represents a dot product of the features and, thus, does not induce a high dimensional reproducing kernel Hilbert space (RKHS). However, the input feature space is already high-dimensional for hyperspectral datasets. Chances are that a good separation is achieved without transformation. Statistical learning theory teaches that unnecessarily complex functions raise the upper bound on the generalization error, a principle similar to Occam's razor. Therefore, it is beneficial to evaluate more simple functions – like the linear kernel – before trying more complex ones. Another advantage of using the linear kernel for kernel composition is its simple computation. Since the resulting kernel matrices are huge, computational load should be kept as low as possible. Good results were achieved by using this kernel function. On the other hand, the spectral angle (3) is considered a highly valuable measure of similarity for hyperspectral data due to their high dimensionality. One main advantage is that the spectral angle is insensitive with respect to differences in illumination. Thus, brighter and darker parts of the same material are more likely to be assigned to the same class if such a distance measure is used. Mercier & Lennon (2003) and Honeine & Richard (2010) propose a methodology for integrating the spectral angle into SVM classification. By simply replacing the $\|x_i – x_j\|$ in the RBF (2) with the spectral angle (3), they obtain a spectral kernel which combines the illumination insensitivity of the spectral angle distance with the discriminative power of kernel-based classification:

$$K_{RBF} = \exp\left(\frac{-\| x_i - x_j \|}{2\gamma}\right) \quad (2)$$

$$\alpha(x_i, x_j) = \arccos\left(\frac{x_i \cdot x_j}{\| x_i \| \times \| x_j \|}\right) \quad (3)$$

2.3 *Kernel Composition for Data Fusion*

For both application schemes presented in this paper, two information domains are available that have to be fused. Kernel matrices based on kernel functions are the representation of the similarity among input data and are used to identify the support vectors. According to Mercer's theorem, kernel matrices can be combined. Thus, after computing one kernel on the domain A and one kernel on the domain B, data fusion can be performed for instance by adding the kernels from A and B. Various methods for kernel composition are available (Camps-Valls et al. 2005). Herein, four composed kernels will be exemplified: the direct summation kernel (4), the weighted summation kernel (5), the product kernel (6) and the cross-information kernel (7). The direct summation kernel is a simple addition of two kernels. The weighted summation kernel introduces weighting factors τ_1 and τ_2. By setting these parameters, the user can define to which extent the information in each domain is considered as significant for the classification problem. For instance, if the user considers the first domain to be more relevant for the classification problem, he or she could set $\lambda_1 = 0.8$ and $\lambda_2 = 1–\lambda_1$. Thus, the similarity of two features in the first domain influences their kernel values much more than their similarity in the second domain. The decision whether a point belongs to a certain class would therefore strongly depend on the first domain, while the second domain would influence it much less. Under these circumstances, the second domain may only be decisive if a point yields very comparable decision values for two classes based on the first domain. As mentioned above, this option is not available for concatenating features. (7) is called the cross-information kernel. It consist of four single kernels where the last two K_{AB} and K_{BA} allow the incorporation of the mutual information between the data sources A and B (e.g. differences between the values of both data sources for a particular data point). The most important advantage of kernel composition over concatenation is that the Hughes phenomenon is circumvented. The Hughes phenomenon is an accuracy decrease which occurs when the dimensionality of the feature space rises (Hughes 1968). Given a certain number of training samples, the predictive power of classification algorithms tends to decline as the dimensionality increases, e.g. because the feature space becomes more and more empty and statistics in the feature space cannot be estimated properly any more. For a thorough discussion see Landgrebe (1997). Concatenation fuses data in the feature space, thus raising dimensionality. Kernel composition fuses data in the RKHS. Since this space is high-dimensional (possibly infinite-dimensional), introducing extra features affects the solvability of the separation problem to a much smaller extent. On the other hand, it could be argued that concatenation has one advantage over kernel composition. If the data are not separable in all single kernel RKHS, there is a possibility that the separability in the composite RKHS is reduced, thus making concatenation the more favourable method in this case. For this reason, both approaches presented herein use kernel composition only if separability is assumed in the individual kernel spaces. If a certain information domain and its respective kernel are not considered representative for the separation, they are not used.

$$K_C(x_i^C, x_j^C) = K_A(x_i^A, x_j^A) + K_B(x_i^B, x_j^B) \quad (4)$$

$$K_C(x_i^C, x_j^C) = \mu K_A(x_i^A, x_j^A) + (1-\mu) K_B(x_i^B, x_j^B) \quad (5)$$

$$K_C(x_i^C, x_j^C) = K_A(x_i^A, x_j^A) \times K_B(x_i^B, x_j^B) \quad (6)$$

$$K_C(x_i^C, x_j^C) = K_A + K_B + K_{AB}(x_i^A, x_j^B) + K_{BA}(x_i^B, x_j^A) \quad (7)$$

2.4 *Exploitation of the one-against-one Cascade*

Since SVMs are binary classifiers, their original formulation is only able to solve classification problems distinguishing two classes. For our first application scheme, 14 subclasses need to be distinguished, though. Thus, the 14-class problem needs to be split up into various two class problems. The OaO strategy considers each of the 91 permutations of the classes separately – leading to 91 training and classification steps. For each step, a model is trained to separate a subset of two classes, considering e.g. the training pixels of class 6 and class 9. During classification, this SVM assigns either 6 or 9 as a label to each pixel. Each pixel is labelled by all 91 models and, thus, receives 91 labels, so that a 1×91 label vector v_i for each pixel is produced. The final class membership for the ith pixel is decided by $mode(v_i)$, i.e. the label most frequently assigned to the pixel. As mentioned above, two different information domains are available for each application scheme in this paper. Consider the urban image scenario in which different roof material classes are to be separated. While information of the airborne laserscanning (ALS) data can be helpful to distinguish between sloped brick roofs and flat gravel roofs, it will not facilitate the separation of two sloped roof classes significantly. At this point, the OaO cascade can be exploited. As it does not consider all 14 classes in one step, the cascade can be used to recognize whether the user considers geometry as significant for a given classification step or not. When separating two roof material classes with different geometries (sloped vs. flat) the spectral domain is fused with the geometric domain. In contrast, when separating similar roof geometries (sloped vs. sloped or flat vs. flat), the spectral domain is used exclusively. In this way, data known not to contain any new information can be omitted and the classifier training is focused on relevant information. Since this framework is very general, it can be applied to different kinds of data and different knowledge.

3 Fusion and Classification of Hyperspectral and ALS data

Within this section, improvements of our previous work on data fusion are presented. The application scheme is based on the fusion of hyperspectral and laserscanning data. The approach in Braun et al. (2011) is enhanced by using the spectral angle kernel for the hyperspectral data and a radial basis function kernel for the ALS data. Moreover, further kernel composition approaches are employed.

3.1 *Data Preprocessing*

An image from the city of Karlsruhe taken by the HyMap sensor in 2003 with a spatial resolution of $4\times4\ m^2$ and 126 spectral channels provides the hyperspectral information. A laserscan with $1\times1\ m^2$ resolution from 2002 delivers geometrical information. Apart from the first and last pulse information, the gradient and curvature of the first pulse are computed. This information is used to distinguish sloped roofs from flat roofs. Hence, six information channels are derived from the ALS data. The data of HyMap is resampled to the spatial resolution of the laser scan, using nearest neighbourhood interpolation. To reduce computational cost, and to allow for a later comparison with Lemp & Weidner (2004) and Braun et al. (2011), a *605×987* pixel subset is chosen which shows the campus of the Karlsruhe Institute of Technology (KIT). A z-transformation (i.e. normalization by mean and standard deviation) on each layer is performed to ensure comparability. The material classes to be distinguished are: brick (red in the classification results), copper (green), gravel (brown), slate (dark blue), zinc (light blue) and stone plates (grey). For each class, various training areas are defined. With respect to roof geometry, a sloped roof class and flat roof class are distinguished by integration of spatial information during the OaO cascade. However, the distinction is implicit in the sense that no classification map on sloped and flat roofs is produced. The relationship between the classes can be seen in Fig. 1. 200 points are randomly chosen from these areas for training and 100 for validation. Mean-shift segmentation is applied to the first pulse information. In order to be able to decide whether segments correspond to roofs or not, a 'roof mask' is calculated from laserscanning and hyperspectral data. First, a normalized digital surface model (nDSM) is computed by subtracting the terrain model from the surface model. All pixels having nDSM heights below two metres are considered to be non-roof pixels; this process will remove streets and other low man-made objects. Furthermore, the normalized difference vegetation index is computed on the hyperspectral data to mask out the remaining vegetated areas. Only segments being consistent with the roof mask thus created are considered in the subsequent processes. We start with a pixel-based classification. Afterwards, the results of the pixel-based classification are transferred to the roof segments. This is done by assigning each roof segment the class label most frequently assigned to the pixels it consists of.

3.2 *Selection of Single Kernels*

In a first step, we evaluated the capacities of single kernel functions on each domain. We trained a SVM using only one data domain to identify the best suited kernel for hyperspectral and ALS data separately. We evaluated the capacities of the linear, the polynomial, the radial basis function, the sigmoid and the spectral angle kernel by training, using five-fold cross validation to optimize ν and the respective kernel parameters (i.e. γ for RBF, sigmoid and spectral angle kernel, the degree for the polynomial and none for the linear kernel) simultaneously. Validation was performed on an independent set of validation data, entirely unknown to the classifier. The range of the

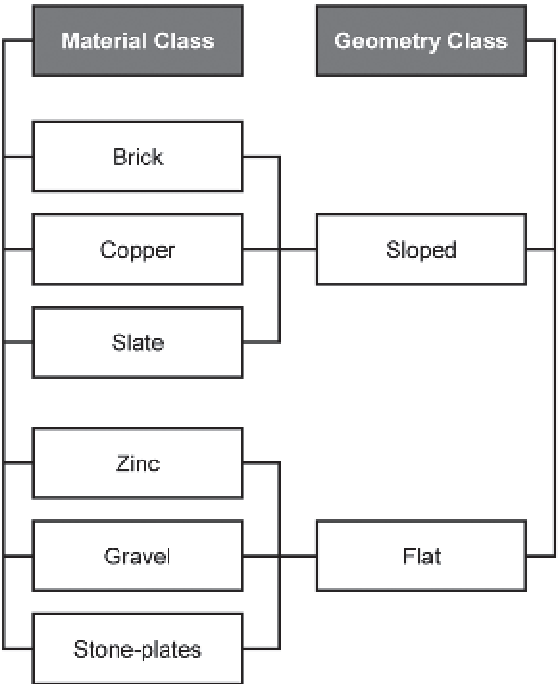

Fig. 1: Classes of the urban dataset and their relationship; left: material classes, right: geometry classes.

Tab. 1: Overall accuracy values using different kernel functions on a single data domain: urban dataset application scheme (RBF = radial basic function).

Kernel	**Hyper-spectral**	**Laser-scanning**
Linear	80.10 %	27.00 %
Polynomial	80.20 %	**55.90 %**
RBF	80.00 %	51.00 %
Sigmoid	53.90 %	25.30 %
Spectral angle	**84.80 %**	26.80 %

grid search for γ was $[2^{-15}, 2^{5}]$, ν was tuned in [0, 1], the polynomial degree was tuned in [1, 10]. For each tuning interval, 10 steps were evaluated (e.g. polynomial degree: 1,2,3,...,10).

As Tab. 1 reveals, the best suited kernel for the hyperspectral data is the spectral angle kernel. This finding empirically confirms the advantages which were theoretically described in section 2.2. The optimal kernel for laserscanning data is the polynomial kernel. The overall accuracy values yielded for the classification based on laserscanning data only were low, which is not surprising keeping in mind the features derived from laserscanning described above are merely enough to separate the spectrally different roof material classes. The sigmoid kernel performed poorly for both domains.

3.3 *Classification of Fused Dataset*

After identifying the most suitable single kernels, we performed kernel composition. Note that the cross-information kernel requires both input data domains to have the same dimensionality. Since the hyperspectral data consists of 126 features while the ALS data consists only of 6, this constraint is not fulfilled. Therefore, a principal component analysis (PCA) was computed on the hyperspectral data and only the first six principal components are used. The PCA is employed only for the cross-information kernel, because for the other kernel composition approaches the two data domains do not need to have the same dimensionality. In the case of the spectral angle kernel, feature reduction via PCA is not helpful. Since the spectral angle is especially designed to exploit the information of the numerous wavelengths in a hyperspectral dataset, it does not seem to be appropriate to reduce by discarding PCAs with lower eigenvalues. According to Braun et al. (2011), we performed data fusion by kernel composition only for the OaO steps where a sloped roof class is separated from a flat roof class. The selection scheme follows these short heuristics:

if geometry of roof_class 1 ≠ geometry of roof_class 2:

use hyperspectral and ALS data jointly via kernel composition

else: use hyperspectral data only via single kernel

After applying these heuristics, we checked for how many steps of the OaO cascade the second domain had been used. It had been used for 40 out of the 91 OaO steps (i.e. in ~44 % of the steps). As Tab. 2 reveals, all data fusion approaches yielded higher overall accuracy values than did the best single kernel. The best approach is the direct summation kernel which yielded 86.9 % overall accuracy, thus improving the best result published in former work by 0.3 %. An attempt to use both data sources at each step yielded 79.6 % overall accuracy, thus validating the assumption that selective usage is beneficial. These findings confirm the conclusions on selective usage thoroughly discussed in Braun et al. (2011). Running the algorithm concatenating the features yielded 84.1 % overall accuracy. The classification results of using only the hyperspectral domain are given in Fig. 2

Tab. 2: Overall accuracy values of different fusion approaches for urban dataset application scheme.

Classification approach	**Overall accuracy**
Only hyperspectral domain	84.80 %
Concatenation	84.10 %
Direct summation (selective) $\tau_1=1, \tau_2=1.$	**86.90 %**
Weighted summation (selective) $\tau_1=0.8, \tau_2=0.2.$	85.50 %
Cross-information kernel (selective)	85.60 %
Product kernel (non selective)	85.90 %

(3) and the result of the direct summation kernel is shown in Fig. 2 (4). Due to the quite high classification accuracies, both results are visually similar in their major part. Fig. 2 (2) visualizes the differences between the two results. The result based on hyperspectral information contains many roofs classified as stone-plate roofs (grey) or zinc roofs (bright blue) which in reality belong to other classes – mostly gravel (brown). The method based on the direct summation kernel of hyperspectral and ALS data classified the circular building in the north-east as brick (red), although it is a zinc roof. A McNemar test (Foody 2004) was used to check the significance of the improvement. The value of the McNemar test $|z|$ indicates the significance of differences between classification results. A difference is considered as significant if $|z| \geq 1.96$. The results based solely on hyperspectral data were tested against the proposed selective kernel composition approach using the McNemar test. The test yielded a value of $|z| = 18.9$. Thus, the advantage of the proposed selective kernel composition approach is highly significant.

In order to assess the performance of the two approaches to individual classes, producer's and user's accuracies were compared. As Tab. 3 reveals, the laserscanning data is helpful for the producer accuracies of gravel and slate. For copper and stone-plates, no differences were found, for brick and zinc, the laserscanning data impairs the producer accuracy slightly. The average improvement over all six classes is 1.81 percent points. The producer's accuracies of some classes of both sloped and flat roofs were improved, others slightly impaired. Hence, no clear trend with respect to roof geometry can be observed. Results for the user's accuracy are similar. Brick, zinc and stone-plates were improved by laserscanning, the copper class was not changed and gravel

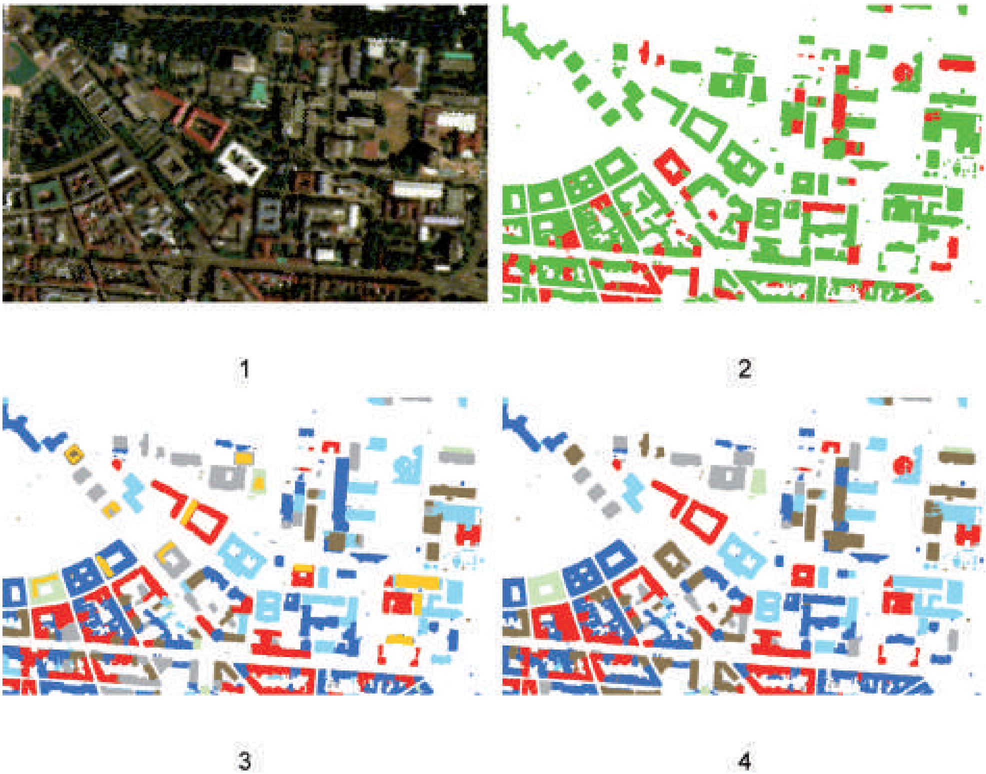

Fig. 2: Urban dataset, 1: True colour view of the hyperspectral data, 2: Agreement (green: equal labels, red: different labels), 3: Classification based on hyperspectral information only (orange fills: training areas), 4: Proposed direct summation approach (fusion of hyperspectral and ALS data).

Tab. 3: Producer's (PA) and user's (UA) values and their differences for the two approaches (HYP = hyperspectral data, LS = laserscanning data).

Class	PA (HYP/LS)	PA (HYP)	δ(PA)	UA (HYP/LS)	UA (HYP)	δ(UA)
Brick	85.17	87.70	-2.53	100.00	94.86	**5.14**
Copper	87.28	87.28	0.00	100.00	100.00	0.00
Gravel	100.00	89.73	**10.27**	62.30	66.69	-4.39
Slate	85.00	77.72	**7.28**	81.28	83.74	-2.46
Zinc	87.56	91.67	-4.11	97.38	88.77	**8.62**
Stone-plates	74.61	74.61	0.00	76.81	64.79	**12.02**
			Ø=1.81			Ø=3.15

and slate were slightly impaired. Again, no clear trend with respect to roof shape can be observed. The average improvement is 3.14 percent points.

4 Fusion and Classification of Spectral Spatial data

To show the transferability of the methodology developed above to a different application scheme, a similar framework of a different dataset is developed. This approach is based on a spectral-spatial classification of the well-known AVIRIS Indian Pines dataset provided in the MultiSpec toolbox (Biehl & Landgrebe 2002). The Indian Pines dataset consists of a *145x145* pixel image of an agricultural area taken by the AVIRIS sensor. This dataset is known as a difficult benchmark for hyperspectral classification due to the high spectral similarity of the classes included (different crops at an early phenological stage). It has 220 spectral bands and includes 16 land use classes. Many bands are quite noisy so some authors remove some bands from the dataset before classification. The dataset comes with a ground truth image in which around 49 % of the pixels are labelled. Nonetheless, some classes consist of very few labelled points (e.g. only 20 points). For this contribution, all 16 classes were used and no bands were removed in order to maintain a high dimensionality and the noise level typical for many hyperspectral sensors. The ground truth data were randomly split into training data (20 %) and evaluation data (80 %).

The spatial context is produced by a simple *3x3* median filter on each channel. The resulting feature vector is then fused with the spectral data. For each class, 20 % of the data labelled by ground truth are used for training and the rest for classification. Selective data fusion via kernel composition of spectral and spatial data by exploiting the OaO cascade is used. However, unlike in section 3, no explicit knowledge is available on whether to use spectral information for a certain classification step or not. Thus, this knowledge is created implicitly after two basic classifications. Again, we evaluated the capacities of the single kernels on the two data domains. The tuning of hyperparameters was performed identically to section 3.2. The polynomial kernel is identified to be ideal for the spectral domain and the RBF kernel to be ideal for the spatial domain (Tab. 4). The overall accuracies yielded were higher on the spatial domain for each kernel except for the sigmoid kernel whose performance is again poor. We believe that

Tab. 4: Overall accuracy values using different kernel functions on a single data domain: spectral-spatial application scheme.

Kernel	Spectral domain	Spatial domain
Linear	79.50 %	87.80 %
Polynomial	**81.10 %**	89.40 %
RBF	79.80 %	**90.10 %**
Sigmoid	38.70 %	38.40 %
Spectral angle	73.80 %	80.20 %

this is due to the lack of preprocessing steps of the spectral domain, which therefore is classified with all the noise it contains. For this reason, we decided to use the features derived by median filtering as our main domain. Unlike for the first classification approach, the spectral angle kernel did not produce the best results on the spectral domain. Having identified the best kernels, we classified the AVIRIS dataset using only the spectral information of the original channels first. Afterwards, a confusion matrix for the result based on spectral data only is computed (it is called **CMsc**). In a second step, we classified, but using the spectral and spatial information jointly at each step of the OaO cascade. Again, a confusion matrix based on the joint usage of spectral and spatial data is computed (it is called **CMscsp**). From these two confusion matrices, the knowledge of whether to use the spatial information is created by simple and straightforward heuristics:

```
if CMscsp(i,j) > CMsp(i,j) or CMscsp(j,i) > CMsp(j,i) with i≠j:
    use spectral and spatial information jointly via kernel composition
else: use spatial information only via single kernel
```

Tab. 5: Overall accuracy values of different fusion approaches for spectral-spatial application scheme.

Classification approach	**Overall accuracy**
Only spatial domain	90.10 %
Direct summation (non selective) *$\tau_1=1, \tau_2=1.$*	83.70 %
Weighted summation (non selective) *$\tau_1=0.6, \tau_2=0.4.$*	84.00 %
Cross-information (non selective)	74.60 %
Product kernel (non selective)	90.70 %
Product kernel (selective)	**91.60 %**

Given 16 classes, the OaO cascade consists of 120 classification steps. Applying this heuristic, the spectral information proved relevant for 49 steps. A third classification was performed, using the spectral and spatial information jointly for these 49 steps only. Various kernel composition approaches were evaluat-

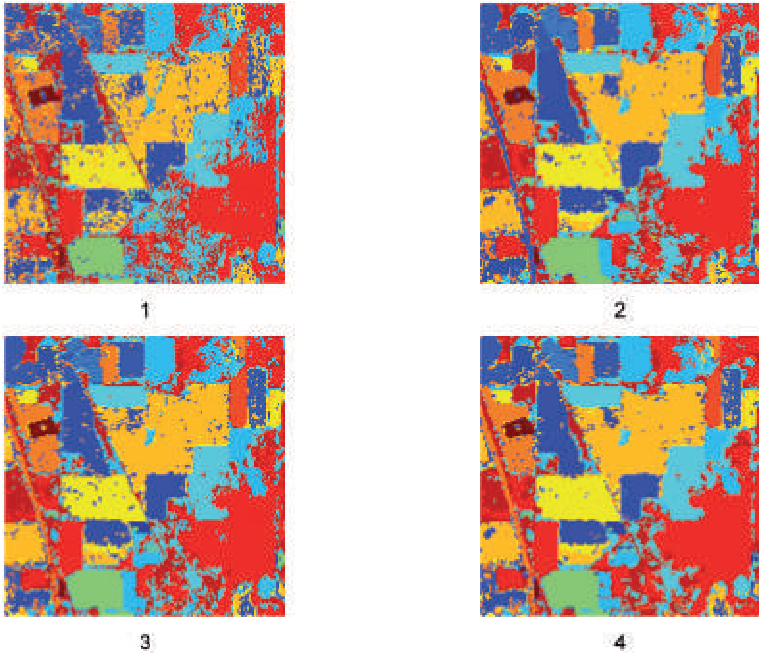

Fig. 3: Result for AVIRIS, 1: Spectral data, 2: Spatial data, 3: Spectral, spatial data fused at each step, 4: Spectral, spatial data with selective kernel composition.

ed. As can be seen in Tab. 5 all except one approach which use the two domains non-selectively produce worse results than using only the median filtering (spatial). The cross-information kernel performs worst of all approaches, confirming the findings of Camps-Valls et al. (2005, 2008) who state that more sophisticated kernels tend to produce worse results. Again, this finding underlines the care that should be taken when fusing data. However, the non-selective product kernel produces a slightly higher overall accuracy than using the approach based on one domain only. As can be seen, spectral information is helpful for most, but not for all the classes. Finally, when using a product kernel only for the 49 steps described above, an overall accuracy value of 91.6 % is obtained, which is the best value for all approaches presented for this dataset. The McNemar test is used to check the significance of the improvements yielded. The value for the test statistic when testing the selective product kernel against the single kernel (only spatial domain) is 2.9, thus, the improvement is significant. When testing the selective product kernel against the non-selective product kernel, a value for the test statistic of 2.7 is obtained, indicating that the improvement caused by selective usage was significant as well.

Fig. 3 shows a comparison of classification results. Fig. 3 (1) shows the result for a classification based on the spectral data only and Fig. 3 (2) shows the result obtained by using the spatial data. Fig. 3 (3) shows the result obtained by fusing spectral and spatial information at each OaO step. Fig. 3 (4) shows the result produced by selectively integrating the spatial information applying the heuristics given above. Note that the images based on using spatial context appear to be much smoother while the borders of different crop fields are preserved. The results are much less affected by salt-and-pepper classification noise. While using spatial information at each step already seems to raise classification performance, the selective usage of the spatial context seems to be even more suitable. The right image is even less affected by salt-and-pepper noise and the field borders are more clearly outlined (southeast quarter of the image).

5 Discussion

The results presented in Braun et al. (2011) could be enhanced by various means. First of all, the spectral angle kernel computed on the hyperspectral data improved the performance in terms of the best accuracy achieved by 4.7 % compared to the previous approach. As the spectral angle is a very well suited measure of similarity for hyperspectral data its integration into SVM combines the advantages of traditional classification approaches for hyperspectral data (the spectral angle mapper) with the advantages of kernel learning. Our results indicate that the spectral angle kernel should be taken into account for hyperspectral classification approaches. Rong et al. (2006) provide an insightful overview on selection of kernel functions. However, their paper does not include the spectral angle kernel. Fauvel et al. (2006) present a comparison taking into account the spectral angle kernel. Their results indicate a slight advantage of RBF over the spectral angle kernel. However, their dataset requires the distinction of classes which mainly differ in intensity (trees, meadows). Furthermore, the best classification accuracy achieved by an approach based on selective data fusion (Braun et al. 2011) could be raised to 86.9 % in this paper. Individual values of producer's and user's accuracies were improved for some classes and impaired for others. However, both producer's and user's accuracy were improved on average. No clear trend between improvements and roof geometries could be observed. This was to be expected, because each class had to be distinguished from classes having both the same and different roof shapes. The highest classification accuracy is achieved by selective data fusion based on the direct summation kernel. Just as Camps-Valls et al. (2005, 2008) we found that simple kernels tend to produce better results than more sophisticated ones. Results indicate that the single kernel type should be selected individually for each input domain. Although it is also possible to use the same kernel type for both domains, higher accuracy values and a more flexible approach are achieved by an individual selection. In our case, the spectral angle kernel and the polynomial kernel proved to be best suited for the fusion of hyperspectral and ALS data while the polynomial and the RBF kernel were optimal for the spectral-spatial approach. The result fails to outperform the overall accuracy (over 90 %) yielded by Bähr et al. (2005) on the same dataset. However, he and his group used a rule-based object-oriented approach. In such a design, knowledge on the classification problem can be integrated in a much more flexible manner than in the machine learning approach implemented herein. Furthermore, rather complex knowledge can be easily integrated using more complex rules (e.g. combinations of thresholds). These factors seem to explain the advantage in overall accuracy by Bähr et al. (2005) over the machine learning approach yielded herein. On the other hand, the object-oriented approach may require a thorough redesign of rules when applied to a different image scene. In contrast, the machine learning approach requires defining new training areas at most. Thus, despite its lower overall accuracy, it may be a valuable alternative to object-oriented designs since it can be applied to new images more easily.

To prove the transferability of our methodology to a different problem and to work towards a more general framework, we have presented a second application scheme. A spectral-spatial classification of AVIRIS Indian Pines reveals similar tendencies as the fusion of hyperspectral and laserscanning data. Of course, in this case the two domains are much more correlated than for the hyperspectral – ALS dataset. The median filter on the hyperspectral channels produces more accurate results than the original channels. Given the considerable noise level and the low separability of the classes in this dataset, smoothing the image by median filtering produces more robust results. The product kernel performs slightly better when applied in a non-selective framework and considerably better when applied in a selective manner. According to the McNemar test, these advantages are significant. The images appear visually much smoother and less affected by salt-and-pepper noise. In contrast to the hyperspectral-ALS approach, implicit knowledge is generated by comparing confusion matrices. These findings show the transferability of our framework to differently posed problems and constitutes an important step towards a general framework for knowledge integration. The results confirm that the approach is applicable to different data types and different types of knowledge. Using the second data domain in a naïve way does not necessarily increase classification accuracy; as some fusion approaches perform worse than the approach based on a single domain only, thorough consideration needs to be made not only on how to fuse the data but also on when to use the second information domain.

6 Conclusion

Kernel composition proves to be a potent method for data fusion when using SVM. Best results are achieved with the direct summation and the product kernel. The OaO cascade is a useful way to integrate the second information domain selectively. Knowledge is exploited to use information selectively and proves suitable to enhance classification results for both application schemes described in this paper. The spectral kernel is used for hyperspectral data since it integrates a well-established measure of similarity into SVM classification. Furthermore, an approach for spectral-spatial classification is presented for the well-known AVIRIS Indian Pines datasets. Spatial context is used for classification. Implicit knowledge on the relevance of the spatial context is integrated. The transfer from fusion of hyperspectral and ALS data based on given explicit knowledge to spectral-spatial classification based on implicit knowledge highlights the transferability of the proposed framework.

References

Amari, S. & Wu, S., 1999: Improving support vector machine classifiers by modifying kernel functions. – Neural Computation **12:** 783–789.

Amarsaikhan, D. & Douglas, T., 2004: Data fusion and multisource image classification. – International Journal of Remote Sensing **25** (17): 3529–3539.

Bähr, H.-P., Lemp, D. & Weidner, U., 2005: Hyperspectral Meets Laserscanning: Image Analysis of Roof Surfaces. – ISPRS Workshop High Resolution Earth Imaging for Geospatial Information **XXXVI** (I/W3): CD-ROM.

Biehl, L. & Landgrebe, D., 2002: MultiSpec – a tool for multispectral-hyperspectral image data analysis. – Computers and Geoscience **28** (10): 1153–1159.

Binaghi, E., Madella, P., Grazia Montesano, M. & Rampini, A., 1997: Fuzzy contextual classification of multisource remote sensing images. – IEEE Transactions on Geoscience and Remote Sensing **35** (2): 326–340.

Braun, A.C., Weidner, U. & Hinz, S., 2010: Support Vector Machines for Vegetation Classification – A Revision. – PFG **2010** (4): 273–281.

Braun, A.C., Weidner, U., Jutzi, B. & Hinz, S., 2011: Integrating model knowledge into SVM classification – Fusing hyperspectral and laserscanning data by kernel composition. – Heipke, C., Jacobsen, K., Rottensteiner, F., Müller, S. & Sörgel, U. (eds.): High-resolution earth imaging for geospatial information. – International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences **38** (4/W19): CD-ROM.

Burges, C.J.C., 1998: A Tutorial on Support Vector Machines for Pattern Recognition. – Data Mining and Knowledge Discovery **2** (2): 121–167.

Camps-Valls, G., Gomez-Chova, L., Muñoz-Marí, J., Vila-Francés, J. & Calpe-Maravilla, J., 2005: Composite Kernels for Hyperspectral Image Classification. – IEEE Geoscience and Remote Sensing Letters **3** (1): 93–97.

Camps-Valls, G., Gomez-Chova, L., Muñoz-Marí, J., Rojo-Alvarez, J.L. & Martinez-Ramon, M., 2008: Kernel-Based Framework for Multitemporal and Multisource Remote Sensing Data Classification and Change Detection. – IEEE Transactions on Geoscience and Remote Sensing **46** (6): 1822–1835.

Fauvel, M., Chanussot, J. & Benediktsson, J.A., 2006: Evaluation of kernels for multiclass classification of hyperspectral remote sensing data. – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) **2006**.

Foody, G.M., 2004: Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. – Photogrammetric Engineering and Remote Sensing **70** (5): 627–634.

Halldorsson, G.H., Benediktsson, J.A. & Sveinsson, J.R., 2003: Support vector machines in multisource classification. – IEEE International Geoscience and Remote Sensing Symposium (IGARSS) **2003**.

Honeine, P. & Richard, C., 2010: The angular kernel in machine learning for hyperspectral data classification. – 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS) **2010**.

Huang, Z. & Lees, B.G., 2004: Combining nonparametric models for multisource predictive forest mapping. – Photogrammetric Engineering and Remote Sensing **70** (4): 415–426.

Hughes, G., 1968: On the mean accuracy of statistical pattern recognizers. – IEEE Transactions on Information Theory **14** (1): 55–63.

Jones, T.G., Coops, N.C. & Sharma, T., 2010: Assessing the utility of airborne hyperspectral and LiDAR data for species distribution mapping in the coastal Pacific Northwest, Canada. – Remote Sensing of Environment **114** (5): 2841–2852.

Landgrebe, D., 1997: On information extraction principles for hyperspectral data. – Purdue University, West Lafayette, IN, USA.

Lemp, D. & Weidner, U., 2004: Use of Hyperspectral and Laser Scanning Data for the Characterization of Surfaces in Urban Areas. – IAPRSIS **XXXV** (B/7): CD-ROM.

Mercer, J., 1909: Functions of positive and negative type, and their connection with the theory of integral equations. – Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character **209:** 415–446.

Mercier, G. & Lennon, M., 2003: Support Vector Machines for Hyperspectral Image Classification with Spectral-based kernels. –IEEE International Geoscience and Remote Sensing Symposium (IGARSS'03) **2003**.

Paola, J.D. & Schowengerdt, R.A., 1995: A review and analysis of backpropagation neural networks for classification of remotely-sensed multi-spectral imagery. – International Journal of Remote Sensing **16** (16): 3033–3058.

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:** 110–122.

Rong, H., Zhang, G. & Jin, W., 2006: Selection of Kernel Functions and Parameters for Support Vector Machines in System Identification. – Journal of System Simulation **11:** 179–189.

Schölkopf, B., Smola, A., Williamson, R. & Bartlett, P., 2000: New Support Vector Algorithms. – Neural Computation **12:** 1207–1245.

Solberg, A.H.S., Taxt, T. & Jain, A.K., 1996: A Markov random field model for classification of multisource satellite imagery. – IEEE Transactions on Geoscience and Remote Sensing **34** (1): 100–113.

Voss, M. & Sugumaran, R., 2008: Seasonal effect on tree species classification in an urban environment using hyperspectral data, LiDAR, and an object-oriented approach. – Sensors **45** (12): 3020–3026.

Waske, B. & Benediktsson, J.A., 2007: Fusion of support vector machines for classification of multisensor data. – IEEE Transactions on Geoscience and Remote Sensing **8** (5): 3858–3866.

Watanachaturaporn, P., Arora, M.K. & Varshney, P.K., 2008: Multisource classification using support vector machines: An empirical comparison with decision tree and neural network classifiers. – Photogrammetric Engineering and Remote Sensing **74** (2): 239–246.

Address of the Authors:

Andreas Ch. Braun, Uwe Weidner, Boris Jutzi & Stefan Hinz, Karlsruher Institut für Technologie – KIT, Institut für Photogrammetrie und Fernerkundung, D-76131 Karlsruhe, Tel.: +49-721-608-4-2315, Fax: +49-721-608-4-8450, e-mail: {andreas.ch.braun} {uwe.weidner} {boris.jutzi} {stefan.hinz}@kit.edu

Manuskript eingereicht: Oktober 2011
Angenommen: März 2012