A GROUP-LASSO ACTIVE SET STRATEGY FOR MULTICLASS HYPERSPECTRAL IMAGE CLASSIFICATION

D. Tuia[a,*], N. Courty[b], R. Flamary[c]

[a] EPFL, Laboratory of Geographic Information Systems, Lausanne, Switzerland - devis.tuia@epfl.ch
[b] Université de Bretagne du Sud, IRISA, Vannes, France - nicolas.courty@irisa.fr
[c] Université de Nice Sophia-Antipolis, Lab. Lagrange, UMR CNRS 7293, Nice, France - remi.flamary@unice.fr

WG Commission III, WG VII

KEY WORDS: Algorithms, Learning, Feature extraction, Classification, High resolution, Hyper spectral

ABSTRACT:

Hyperspectral images have a strong potential for landcover/landuse classification, since the spectra of the pixels can highlight subtle differences between materials and provide information beyond the visible spectrum. Yet, a limitation of most current approaches is the hypothesis of spatial independence between samples: images are spatially correlated and the classification map should exhibit spatial regularity. One way of integrating spatial smoothness is to augment the input spectral space with filtered versions of the bands. However, open questions remain, such as the selection of the bands to be filtered, or the filterbank to be used. In this paper, we consider the entirety of the possible spatial filters by using an incremental feature learning strategy that assesses whether a candidate feature would improve the model if added to the current input space. Our approach is based on a multiclass logistic classifier with group-lasso regularization. The optimization of this classifier yields an optimality condition, that can easily be used to assess the interest of a candidate feature without retraining the model, thus allowing drastic savings in computational time. We apply the proposed method to three challenging hyperspectral classification scenarios, including agricultural and urban data, and study both the ability of the incremental setting to learn features that always improve the model and the nature of the features selected.

1. INTRODUCTION

Remote sensing technologies allow to observe the Earth from a distance. The use of satellite and aerial data allows to monitor the processes occurring at the surface in a non-extrusive way, both at the local and global scale [Lillesand et al., 2008, Richards and Jia, 2005]. The reduced revisit time of satellites, in conjunction with the potential for quick deployment of aerial and unmanned systems, made remote sensing systems more and more appealing and nowadays the use of satellite data has become a standard for researchers and public bodies.

In order to be usable by end-users and decision makers, remote sensing pixel information must be processed and converted into products depicting a particular facet of the processes occurring at the surface. Among the different products traditionally available, land cover maps issued from image classification[1] are the most common (and probably also the most used). Land cover maps can then be used for urban planning [Taubenboeck et al., 2012], agriculture surveys [Alcantara et al., 2012] or surveying of deforestation [Asner et al., 2005].

The quality of land cover maps is of prime importance. Therefore, a wide panel of research works consider image classification algorithms and their impact on the final maps [Plaza et al., 2009, Camps-Valls et al., 2011]. This challenge is not trivial, as remote sensing systems are often high dimensional (number of spectral bands acquired), spatially and spectrally correlated and affected by noise [Camps-Valls et al., 2014]. Moreover, temporal dependencies are also present, since a type of land cover may evolve throughout the year.

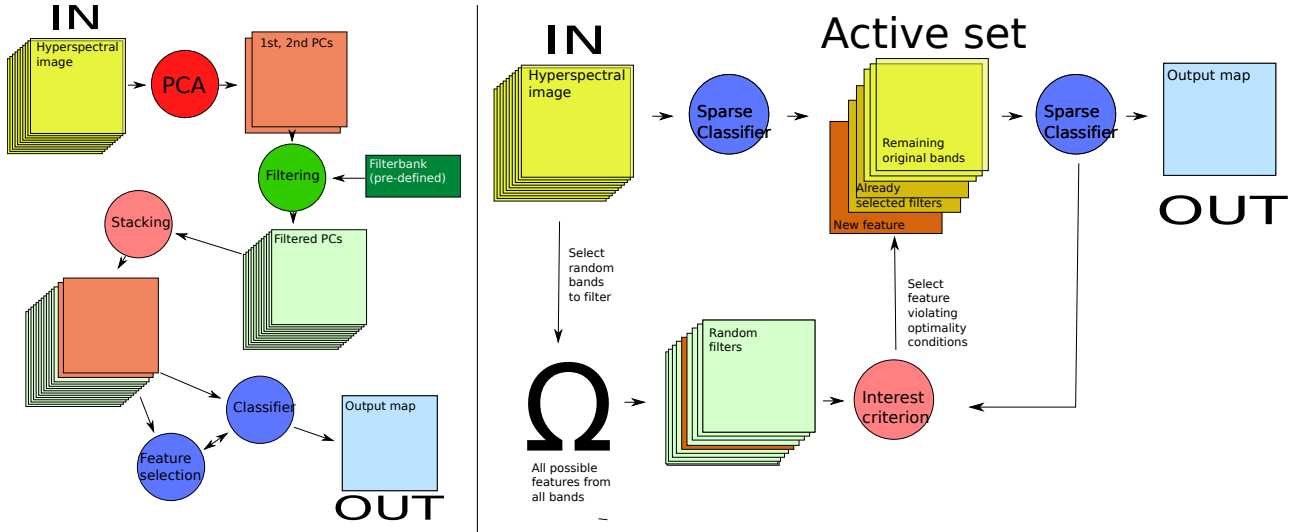
Among these aspects of the data, spatial relations have received particular attention [Fauvel et al., 2013]: the land cover maps are generally smooth, in the sense that neighboring pixels tend to belong to the same type of land cover [Schindler, 2012]. On the contrary, the spectral signatures of pixels of a same type of cover tend to become more and more variable, especially with the increase of spatial resolution. Therefore, we want to describe a smooth land cover random field using spectral information with a high within-class variability. Solutions to this problem have been proposed in the community and mostly recur to spatial filtering [Fauvel et al., 2013] – i.e., work at the level of the input vector – and Random Fields/graph cuts [Moser et al., 2013] – i.e. work within the optimization of a context-aware energy function.

In this paper, we consider the first family of methods, i.e. those based on the extraction of spatial filters. Methods proposed in remote sensing image classification tend to pre-compute a large quantity of spatial filters, related to the user's preference and knowledge of the problem: texture [Pacifici et al., 2009], Gabor [Li and Du, in press], morphological [Benediktsson et al., 2005, Tuia et al., 2009, Dalla Mura et al., 2010] or bilateral filters [Schindler, 2012] are among those used in recent literature.

Even if successful, these studies relied on the definition a-priori of a filterbank. As shown in Fig. 1a, the filter bank is applied to each band of the image, resulting into a $(f \times B)$-dimensional filter bank, where f is the number of filters and B the number of bands. This proved to be unfeasible for high dimensional datasets, such as hyperspectral data, for which the traditional way to deal with the problem is to perform a principal components analysis (PCA) and then extract the filters from the $p << B$ principal components related to maximal variance [Benediktsson et al., 2005]. In that case, the final input space is $(f \times p)$-dimensional. In all cases, the dimensionality of the final vector makes it necessary to run a feature selection step, to select the subset which is the most effective for the classification.

*Corresponding author.

[1] In this paper, we refer to classification as the process of attributing one land cover type (type of material) or land use (use of the land, e.g., road vs parking) type to each pixel in the scene.



(a) Pre-defined filterbank

(b) Proposed incremental strategy

Figure 1: Spatio-spectral classification with contextual filters. (a) Using pre-defined filterbanks, applied on the first principal component ($p = 1$). (b) The proposed incremental system based on active set and sparse SVM.

Proceeding this way is suboptimal in two senses: first, one forces to restrict the number and parameters of filters to be used to a subset, whose appropriateness only depends on the prior knowledge of the user. Second, generating hundreds (if not thousands) of spatial filters and use them in a classifier, that also might operate with a feature selection strategy, increases the computational cost significantly. Also note that, if the spatial filters considered bear continuous parameters (e.g. Gabor or angular features), there is theoretically an infinite number of feature candidates. An integrated approach, which would incrementally build the set of filters from an empty subset and select only the filters that improve class discrimination to the classifier is of great importance. Two approaches are of particular interest in this sense: Grafting [Perkins et al., 2003] and Group Feature Learning [Rakotomamonjy et al., 2013], which incrementally select the most promising feature among a batch of features extracted from the universe of all possible features admitted. Since this selection is based on a heuristic criterion ranking the features by their informativeness when added to the model, it may be seen as performing active learning [Crawford et al., 2013] in the space of possible features (in this case, the active learning oracle is replaced by an optimality condition, for which only the features improving the current classifier are selected).

In this paper, we extend the Group Feature Learning model [Rakotomamonjy et al., 2013] to multiclass logistic regression (also known as multinomial regression). The use of a group-lasso regularization [Yuan and Lin, 2007] allows to jointly select the relevant features and also to derive efficient conditions for evaluating the discriminative power of a new feature. In [Rakotomamonjy et al., 2013], authors propose to use group-lasso for multitask learning by allowing to use an additional sparse average classifier common to all tasks. Adapting their model in a multiclass classification setting leads to the use of the sole group-lasso regularization. Note that one could use a ℓ_1 support vector machine as in [Tuia et al., 2014] to select the relevant feature in a One-VS-All setting but this approach is particularly computationally intensive, as the incremental problem is solved for each class separately. This implies the generation of millions of features, that may be useful for more than one class at a time. The proposed group lasso regularization allows to select features useful for many classes, even if it does not show the highest score for a particular class. This means sharing information among the classes, similarly to what would happen in a multitask setting [Leiva-Murillo et al., 2013].
Moreover we propose in this work to learn a multiclass logistic classifier (MLC) with a softmax loss instead of a SVM classifier. This approach indeed allows to natively handle several classes without using the One-VS-All approach and has the advantage of providing probabilistic prediction scores that can more easily be used in any post-processing methods (such as Markov random fields).

We test the proposed method on two landcover classification tasks with hyperspectral images of agricultural areas and on one landuse classification example over an urban area exploiting jointly hyperspectral and LiDAR images. In all cases, the proposed feature learning method solves the classification tasks with at least state of the art numerical performances and returns compact models including only features that are discriminative for more than one class.

The remainder of this paper is as follows: Section 2. details the proposed method, as well as the multiclass feature selection using group-Lasso. In Section 3. we present the datasets and the experimental setup. In Section 4. we present and discuss the experimental results. Section 5. concludes the paper.

2. FEATURE LEARNING WITH MULTICLASS LOGISTIC CLASSIFICATION

The proposed methodology is described in this Section. We first present the multiclass logistic classification and then derive its optimality conditions, which are used in the active set algorithm.

2.1 Multiclass logistic classifier with group Lasso regularization

Consider an image composed of pixels $\mathbf{x}_i \in \mathbb{R}^B$. A subset of l_c pixels is labeled into one of C classes: $\{\mathbf{x}_i, y_i\}_{i=1}^{l_c}$, where y_i are integer values $\in \{1, \ldots, C\}$. We consider a (possibly infinite) set of θ-parametrized functions $\phi_\theta(\cdot)$ mapping each pixel in the image into the feature space of the filter defined by θ. As in [Tuia et al., 2014], we define as $\mathcal{F}$ the set of all possible finite subsets of features and φ as an element of $\mathcal{F}$ composed of d features $\varphi = \{\phi_{\theta_j}\}_{j=1}^{d}$. We also define $\Phi_\varphi(\mathbf{x}_i)$ as the stacked vector of

all the values obained by applying the filters φ to pixel $\mathbf{x}_i$ and $\mathbf{\Phi}_\varphi \in \mathbb{R}^{n_l \times d}$ the matrix containing the features in φ computed for all the labeled pixels.

In this paper we consider the classification problem as a multiclass logistic regression problem. Learning such a classifier for a fixed amount of features φ corresponds to learning a weight matrix $\mathbf{W} \in \mathbb{R}^{d \times c}$ and the bias vector $\mathbf{b} \in \mathbb{R}^{1 \times C}$ using the softmax loss. In the following, we refer to $\mathbf{w}_c$ as the weights corresponding to class c, which corresponds to the c-th column of matrix $\mathbf{W}$. The k-th line of matrix $\mathbf{W}$ is denoted as $W_{k,.}$. The optimization problem for a fixed feature set φ is defined as:

$$\mathcal{L}(\varphi) = \min_{\mathbf{W},\mathbf{b}} \frac{1}{l_c} \sum_{i=1}^{l_c} \log\Big(\sum_{c=1}^{C} \exp\big((\mathbf{w}_c - \mathbf{w}_{y_i})^\top \Phi_\varphi(\mathbf{x}_i) + (b_c - b_{y_i})\big)\Big) + \lambda\Omega(\mathbf{W}) \quad (1)$$

where the first term corresponds to the soft-max loss and the second term is a group-lasso regularizer. In this paper, we use the mixed $\ell_1 - \ell_2$ norm:

$$\Omega(\mathbf{W}) = \sum_{k=1}^{d} ||W_{k,\cdot}||_2 \quad (2)$$

This regularization term promotes group sparsity, due to its non differentiability at zero. In this case we grouped the coefficients of $\mathbf{W}$ by lines, meaning that it will promote joint feature selection for all classes. Note that this approach can be seen as multi-task learning where the tasks corresponds to the classifier weights of each class [Obozinski et al., 2006, Rakotomamonjy et al., 2011]. As a result, if a variable (filter) is active, it will be active for all classes. In our opinion, this makes particular sense in a multiclass setting, because a feature that helps in detecting a given class also helps in "not detecting" the others in the $C-1$ other classifiers.

In order to discuss the proposed algorithm, we first have to derive the optimality conditions of the problem. To this end, we compute the sub-differential of the cost function defined in Eq. (1):

$$\partial\mathcal{L} = \mathbf{\Phi}_\varphi^\top \mathbf{R} + \lambda \partial\Omega(\mathbf{W}) \quad (3)$$

where, $\mathbf{R}$ is a $l_c \times C$ matrix that, for a given sample $i \in \{1, ., l_c\}$ and a class $c \in \{1, ., C\}$, equals:

$$R_{i,c} = \frac{\exp(M_{i,c} - M_{i,y_i}) - \delta_{\{y_i - c\}} \sum_{k=1}^{C} \exp(M_{i,k} - M_{i,y_i})}{l_c \sum_{k=1}^{C} \exp(M_{i,k} - M_{i,y_i})} \quad (4)$$

where $\mathbf{M} = \mathbf{\Phi}_\varphi \mathbf{W} + \mathbb{1}\mathbf{b}$ and $\delta_{\{y_i - c\}} = 1$ if $c = y_i$ and 0 otherwise. In the following, we set $\mathbf{G} = \mathbf{\Phi}_\varphi^\top \mathbf{R}$ a $d \times C$ matrix corresponding to the gradient of the data fitting term *w.r.t* $\mathbf{W}$. Note that this gradient is simply computed by multiple scalar product between the features $\mathbf{\Phi}_\varphi$ and the multiclass residual $\mathbf{R}$. The optimality conditions can be obtained separately for each group j, *i.e.* for each line j of the $\mathbf{W}$ matrix. $\Omega(\mathbf{W})$ is a non differentiable norm-based regularization [Bach et al., 2011]. The optimality condition for an Euclidean norm regularization consists in a constraint with its dual norm (namely itself):

$$||G_{j,\cdot}||_2 \leq \lambda \qquad \forall j \in \varphi \quad (5)$$

which in turn breaks down to:

$$\begin{cases} ||G_{j,\cdot}||_2 = \lambda & \text{if} \quad W_{j,\cdot} \neq \mathbf{0} \\ ||G_{j,\cdot}||_2 \leq \lambda & \text{if} \quad W_{j,\cdot} = \mathbf{0} \end{cases} \quad (6)$$

These optimality conditions suggest the use of an active set algorithm. Indeed, if the norm of correlation of a feature with the residual matrix is below λ, it means that this feature is not useful and its weight will be set to 0 for all the classes.

Algorithm 1 Multiclass active set selection for MLC

Inputs

- Initial active set φ_0

repeat
  Solve a MLC with current active set φ
  Generate a minibatch $\{\phi_{\theta_j}\}_{j=1}^{p} \notin \varphi$
  Compute G as in (7) $\forall j = 1 \ldots p$
  Find feature $\phi^*_{\theta_j}$ maximizing $||r_{\theta_j}||_2$
  if $||G_{\theta^*_j,\cdot}||_2 > \lambda + \epsilon$ **then**
    $\varphi = \phi^*_{\theta_j} \cup \varphi$
  end if
until stopping criterion is met

2.2 Proposed active set criterion

In this paper, we want to learn jointly the best set of filters $\varphi^* \in \mathcal{F}$ and the corresponding MLC classifier. This is achieved by minimizing Eq. (1) jointly on φ and $\mathbf{W}, \mathbf{b}$, respectively. As in [Rakotomamonjy et al., 2013], we can extend the optimality conditions in (6) to all filters with zero weights that are *not* included in the current active set φ:

$$||G_{\phi_\theta,\cdot}||_2 \leq \lambda \qquad \forall \phi_\theta \notin \varphi \quad (7)$$

Indeed, if this constraint holds for a given feature not in the current active set, then adding this feature to the optimization problem will lead to a row of zero weights $W_{(d+1),\cdot}$ for this feature. This also means that if we find a feature that violates Eq. (7), its inclusion in φ will i) make the global MLC cost decrease and ii) provide a feature with non-zero coefficients for all classes after reoptimization.

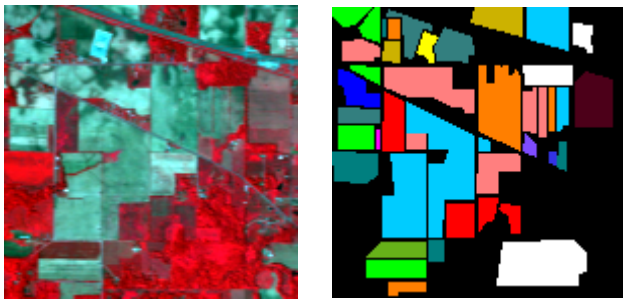
The proposed algorithm is illustrated in Fig. 1b and pseudocode is given in Algorithm 1: we initialize the active set φ_0 with the spectral bands and run a first MLC minimizing Eq. (1). Then we generate a random minibatch of candidate features, Φ_{θ_j}, involving spatial filters with random types and parameters. We then assess the optimality conditions with (7): if the feature $\phi^*_{\theta_j}$ with maximal $||G_{\theta_j,\cdot}||_2$ is greater than $\lambda + \epsilon$, it is selected and added to the current active set $[\phi^*_{\theta_j} \cup \varphi]$.

3. DATA AND SETUP

3.1 Datasets

We tested the proposed active set method on three hyperspectral classification tasks:

a) Indian Pines 1992 (AVIRIS spectrometer, HS): the first dataset is a 20-m resolution image taken over the Indian Pines (IN) test site in June 1992 (see Fig. 2). The image is 145×145 pixels and contains 220 spectral bands. A ground survey of 10366 pixels, distributed in 16 crop types classes, is available. The classes are unevenly sampled (see Table 1). This dataset is a classical benchmark to validate model accuracy and is known to be very challenging because of the strong mixture of the classes' signatures, since the image has been acquired shortly after the crops were planted. As a consequence, all signatures are contaminated by soil signature, making thus a spectral-spatial processing compulsory to solve the classification problem. As preprocessing, 20 noisy bands covering the region of water absorption have been removed.



(a) (b)

Figure 2: Indian Pines 1992 AVIRIS data.(a) False color composition and (b) Ground truth (for color legend, see Tab. 1). Unlabeled samples are in black.

Table 1: Classes and Samples (n_l^c) of the ground truth of the Indian Pines 1992 dataset (cf. Fig. 2).

	Class	n_l^c		Class	n_l^c
	Alfalfa	54		Oats	20
	Corn-notill	1434		Soybeans-notill	968
	Corn-min	834		Soybeans-min	2468
	Corn	234		Soybeans-clean	614
	Grass/Pasture	497		Wheat	212
	Grass/Trees	747		Woods	1294
	Grass/Past.-mowed	26		Towers	95
	Hay-windrowed	489		Other	380
				Total	10366

(a) (b)

Figure 3: Indian Pines 2010 SpecTIR data.(a) RGB composition and (b) Ground truth (for color legend, see Tab. 2). Unlabeled samples are in black.

b) Indian Pines 2010 (ProSpecTIR spectrometer, VHR HS): the ProSpecTIR system acquired multiple flightlines near Purdue University, Indiana, on May 24-25, 2010 (Fig. 3). The image subset analyzed in this study contains 445×750 pixels at 2*m* spatial resolution, with 360 spectral bands of 5*nm* width. Sixteen land cover classes were identified by field surveys, which included fields of different crop residue, vegetated areas, and man-made structures. Many classes have regular geometry associated with fields, while others are related with roads and isolated man-made structures. Table 2 shows class labels and number of training samples per class.

Table 2: Classes and Samples (n_l^c) of the ground truth of the Indian Pines 2010 dataset (cf. Fig. 3).

	Class	n_l^c		Class	n_l^c
	Corn-high	3387		Hay	50045
	Corn-mid	1740		Grass/Pasture	5544
	Corn-low	356		Cover crop 1	2746
	Soy-bean-high	1365		Cover crop 2	2164
	Soy-bean-mid	37865		Woodlands	48559
	Soy-bean-low	29210		Highway	4863
	Residues	5795		Local road	502
	Wheat	3387		Buildings	546
				Total	198074

Table 3: Classes and Samples (n_l^c) of the ground truth of the Houston 2013 dataset (cf. Fig. 4).

	Class	n_l^c		Class	n_l^c
	Healthy grass	1231		Road	1219
	Stressed grass	1196		Highway	1224
	Synthetic grass	697		Railway	1162
	Trees	1239		Parking Lot 1	1233
	Soil	1152		Parking Lot 2	458
	Water	325		Tennis Court	428
	Residential	1260		Running Track	660
	Commercial	1219		Total	14703

c) Houston 2013 (CASI spectrometer VHR HS + LiDAR data). The third dataset depicts an urban area nearby the campus of the University of Houston (see Fig. 4). The dataset was proposed as the challenge of the IEEE IADF Data Fusion Contest 2013 [Pacifici et al., 2013]. The area has been imaged by the CASI hyperspectral sensor (144 spectral bands at 2.5m resolution) and scanned by a LiDAR. From the latter, a digital surface model (DSM) at the same resolution has been extracted. Both imaging sources have been coregistered. 15 urban land-use classes are to be classified (Tab. 3). Two preprocessing steps have been performed: 1) histogram matching has been applied to the large shadowed area in the right part of the image (cf. Fig 4): the shadowed area has been extracted by segmenting a near-infrared band and the matching with the rest of the image has been applied. 2) A height trend has been removed from the DSM, by applying a linear detrending of 3m from the West along the x-axis.

3.2 Setup of experiments

For all dataset, all the features have been mean-centered and normalized to unit norm. This normalization is mandatory due to the optimality conditions, which is based on a scalar product. In order to compare fairly the alignment of all the candidate features to the residual, all feature must have the same norm (See Eq. (7).).

In all experiments, we use the multiclass logistic classifier with $\ell_1 - \ell_2$ norm implemented in the SPAMS package[2]. We start by training a model with all available bands (plus the DSM in the HOUSTON2013 case) and use its result as the first active set. Regarding the active set, we used the following parameters:

- The stopping criterion is a number of iterations (150).
- A minibatch is composed of filters extracted from 30 bands, randomly selected. In the HOUSTON 2013 case, the DSM is added to each minibatch.

[2] `http://spams-devel.gforge.inria.fr/`

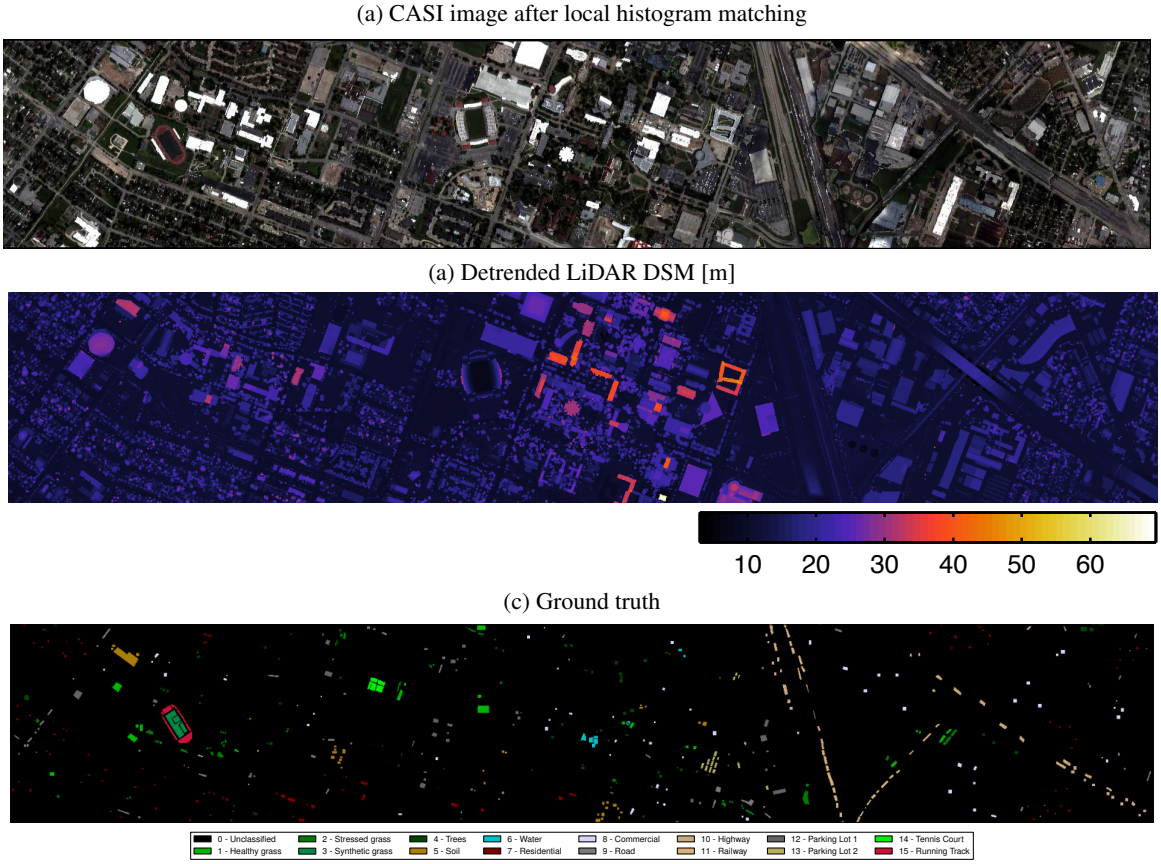


Figure 4: Houston 2013.(a) RGB composition of the CASI data, (b) DSM issued from the LiDAR point cloud and (c) ground truth. (for color legend, see Tab. 2). Unlabeled samples are in black.

- The possible filters are listed in Tab. 4. Structuring elements (SE) can be disks, diamonds, squares or lines. If a linear structuring elements is selected, an additional orientation parameter is also generated ($\alpha \in [-\pi/2, \dots \pi/2]$). These filters are those generally used in remote sensing hyperspectral classification literature [Fauvel et al., 2013], but any type of spatial filter can be used in the process.

- A single minibatch can be used twice (i.e. once a first filter has been selected, it is removed and Eq. (7) is re-evaluated on the remaining filters).

In each experiment, we start by selecting an equal number of labeled pixels per class l_c: we extracted 30 random pixels per class in the INDIAN PINES 1992 case, 60 in the INDIAN PINES 2010 and in the HOUSTON 2013 case. The difference in the amount of labeled pixels per class is related to i) the amount of labeled pixels available per task and ii) the complexity of the problem at hand. As test set, we considered all remaining labeled pixels, but disregard those in the spatial vicinity of the pixels used for training. In the INDIAN PINES 1992 case, we consider all labeled pixels out of a 3×3 window around the training pixels, in the INDIAN PINES 2010 case a 7×7 window. The difference is basically related to the image resolutions. In the HOUSTON 2013 case, a spatially disjoint test set was provided in a separate file and was used for testing purposes.

Each experiment was repeated 10 times, by varying the initial training set (the test set also varies, since it depends on the specific location of the training samples). Average performances, along with their standard deviation, are reported.

Table 4: Filters considered in the experiments (B_i, B_j: input bands indices ($i, j \in [1, \dots b]$); s: size of moving window, SE : type of structuring element; α: angle).

	Filter	θ
Morphological	Opening / closing	B_i, s, α
	Top-hat opening / closing	B_i, s, SE, α
	Opening / closing by reconstruction	B_i, s, SE, α
	Opening / closing by reconstruction top-hat	B_i, s, SE, α
Texture	Average	B_i, s
	Entropy	B_i, s
	Standard deviation	B_i, s
	Range	B_i, s
Attribute	Area	B_i, Area threshold
	Bounding box diagonal	B_i, Diagonal threshold
Ratios	Simple	B_i/B_j
	Normalized	$(B_i - B_j)/(B_i + B_j)$

4. EXPERIMENTAL RESULTS

Performances along the iterations. Numerical results for the three datasets are provided in Fig. 5: the left column illustrates the evolution of the Kappa statistic [Foody, 2004] along the iterations and for three levels of $\ell_1 - \ell_2$ regularization λ. The right column shows the evolution of the number of features in the active set.

For all the datasets, the iterative feature learning corresponds to a continuous, almost monotonic, increase of the performance. This is related to the optimality conditions of Eq. (1): each time the

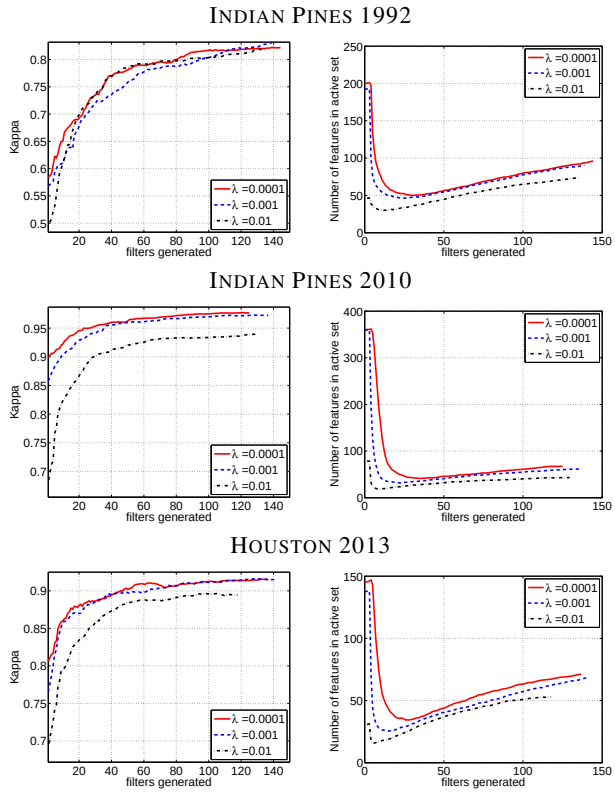


Figure 5: Left: numerical performance (Kappa statistic) for different degrees of regularization λ and filtering the original bands. Right: number of active features during the iterations.

model adds one filter $\phi_{\theta_j^*}$ to φ, the MLC cost function decreases while the classifier performances raises. Overfitting is prevented by the group-lasso regularization: on the one hand this regularizer promotes sparsity through the ℓ_1 norm, while on the other hand it limits the magnitude of the weight coefficients $\mathbf{W}$ and promotes smoothness of the decision function by the use of the ℓ_2 norm. Note that for the HOUSTON 2013 dataset, the final classification performance is at the same level as the one of the winners of the contest, thus showing the credibility of our approach.

For each case study, the model with the lowest sparsity ($\lambda = 0.0001$) shows the initial best performance (it utilizes more features, as shown in the right column) and then keeps providing the best performances. However, the model with $\lambda = 0.001$ has an initial sparser solution and shows a steeper increase of the curve in the first iterations. When both models provide similar performance, they are actually using the same number of features in all cases. The sparsest model ($\lambda = 0.01$) shows the worst results in two out of the three datasets and in general is related to less features selected: our interpretation is that the regularization ($\lambda = 0.01$) is too strong, leading to a model that discards relevant features and is too biased for a good prediction (even when more features are added). As a consequence, the learning rate is surely steeper than for the other models, but the model does not converge to an optimal solution.

Numerical performances at the end of the feature learning. Comparisons with competing strategies where the MLC classifier is learned on pre-defined feature sets are reported in Table 5. First we discuss the performance of our active set approach when learning the filters applied on the original bands (AS-BANDS): in the INDIAN PINES 1992 case, the AS-BANDS method obtains an average Kappa of 0.83 using 96 features. This is a good result if compared to the upper bound of 0.86 obtained by a classifier

Table 5: Results by MLC classifiers trained with the spectral bands (ω), with spatial features extracted from the three first PCAs (s, including morphological and attribute filters) or with the proposed active set (AS-). In the HOUSTON 2013 case, features extracted from the DSM have been added to the input space.

Method	Ω	PINES 1992	PINES 2010	HOUSTON 2013
MLC-ω	ℓ_1	0.42 ± 0.02	0.58 ± 0.01	0.61 ± 0.01
# features		60 ± 3	107 ± 9	54 ± 3
MLC-ω	ℓ_2	0.59 ± 0.03	0.90 ± 0.01	0.80 ± 0.01
# features		200	360	145
AS-BANDS	$\ell_1\ell_2$	0.83 ± 0.02	$\mathbf{0.98 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$
# features		96 ± 5	68 ± 5	71 ± 3
MLC-s	ℓ_1	0.85 ± 0.02	0.84 ± 0.01	0.76 ± 0.01
# features		85 ± 7	64.2 ± 3	82 ± 5
MLC-s	ℓ_2	0.85 ± 0.01	0.96 ± 0.01	0.88 ± 0.01
# features		217	228	303
AS-PCAS	$\ell_1\ell_2$	$\mathbf{0.89 \pm 0.03}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$
# features		82 ± 4	83 ± 8	64 ± 4

using the complete set of 14'627 morphological and attribute features extracted from each spectral band (result not reported in the table)[3]. On the two other datasets, the AS-BANDS method provided average Kappa of 0.98 and 0.92, respectively.

We compared these results to those obtained by classifiers trained on fixed raw bands (MLC$-\omega$) or on sets of morphological and attribute filters extracted form the three first principal components (MLC-s). We followed the generally admitted hypothesis that the first(s) PCA(s) contain most of the relevant information in hyperspectral images [Benediktsson et al., 2005]. On all the datasets, the proposed AS-BANDS method performs remarkably well compared with models using only the spectral information (MLC-ω) and compares at worse equivalently (and significantly better in the INDIAN PINES 2010 and HOUSTON 2013 cases) with models using ℓ_2 classifiers (thus without sparsity) and three to four times more features including spatial information (MLC-s). The good performance of the ℓ_2 method on the INDIAN PINES 1992 dataset (Kappa observed of 0.85) is probably due to the application of the PCA transform prior to classification, which, besides allowing to decrease the dimensionality of the data, also decorrelates the signals and isolates the bare soil reflectance, which is present for almost all classes (cf. the data description). For this reason, we also investigated a variant of our approach where, instead of working on the spectral space, we filtered the PCA components extracted from the original data (AS-PCAS). In the INDIAN PINES 1992 case, the increase in performance is striking, with a final Kappa of 0.89. For the two other datasets, the results remain in the same range as for the AS-BANDS results.

Selected features: for the three images, the active set models end up with a maximum of $70 - 100$ features, shared by all classes. This model is very compact, since it corresponds to only $30 - 50\%$ of the initial dimensionality of the spectra. Due to the optimization problem, the features selected are active for several classes simultaneously, as shown in Fig. 6, which illustrates the weights matrix $\mathbf{W}^\top$ for the INDIAN PINES 2010 and HOUSTON 2013 experiments at the end of the feature learning for one specific run with $\lambda = 0.0001$. Each column corresponds to a feature selected by the proposed algorithm and each row to a class; the color corresponds to the strength of the weight (positive or negative). One can appreciate that the selected features (columns) have large coefficients – corresponding to strong green or brown tones in the figures – for more than one class (the rows).

[3] Only squared structuring elements were used and the filter size range was pre-defined by expert knowledge.

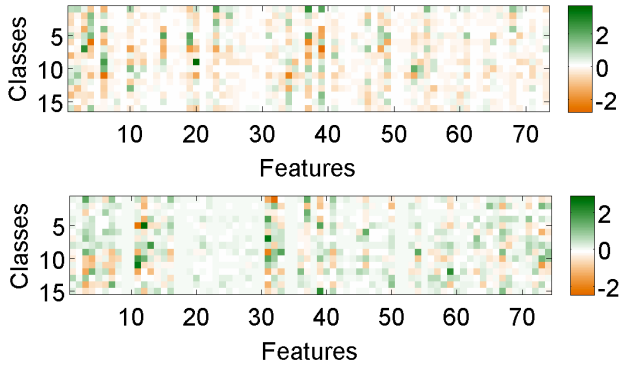


Figure 6: Final weight matrix for a run of the INDIAN PINES 2010 (top) and HOUSTON 2013 (bottom) experiments.

Finally, Fig. 7 illustrates some of the features selected in the HOUSTON 2013 case. Each column corresponds to a different zoom in the area and highlights a specific class. We visualized the features of the same run as the bottom row of Fig. 6 and visualized the six features with highest $||W_{j,\cdot}||_2$, corresponding to those active for most classes with the highest squared weights. By analysis of the features learned, one can appreciate that they clearly are discriminative for the specific classification problem: this shows that, by decreasing the overall loss, adding these features to the active set really improves class discrimination.

5. CONCLUDING REMARKS

In this paper we proposed an active set algorithm for the automatic selection of contextual features in hyperspectral image classification. The proposed method uses an optimality criterion based on the mixed-norm group lasso and the multiclass logistic regression classifier. For a minibatch of candidate features, it selects those that lead to a non-null coefficient in all the classes and therefore yields a decrease in the cost function. Compared to other existing active set algorithms, our approach integrates all the classes information simultaneously through the multiclass soft-max loss function, avoids the computation of irrelevant features with respect to the multi class classification problem and is much faster, since it solves a single active set problem instead of one per class. Experiments on three benchmark hyperspectral images illustrated the benefits of the approach, which reaches at least state of the art performances with a reduced set of features, and without the need of defining them by prior/expert knowledge. Extension to contextual classifier based on spatial priors (MRF, CRF) is the logical next step.

REFERENCES

Alcantara, C., Kuemmerle, T., Prishchepov, A. V. and Radeloff, V. C., 2012. Mapping abandoned agriculture with multi-temporal MODIS satellite data. Remote Sens. Enviro. 124, pp. 334–347.

Asner, G. P., Knapp, D., Broadbent, E., Oliveira, P., Keller, M. and Silva, J., 2005. Ecology: Selective logging in the Brazilian Amazon. Science 310, pp. 480–482.

Bach, F., Jenatton, R., Mairal, J. and Obozinski, G., 2011. Convex optimization with sparsity-inducing norms. In: Optimization for Machine Learning, MIT Press.

Benediktsson, J., Palmason, J. A. and Sveinsson, J. R., 2005. Classification of hyperspectral data from urban areas based on extended morphological profiles. IEEE Trans. Geosci. Remote Sens. 43(3), pp. 480–490.

Camps-Valls, G., Tuia, D., Bruzzone, L. and Benediktsson, J. A., 2014. Advances in hyperspectral image classification. IEEE Signal Proc. Mag. 31, pp. 45–54.

Camps-Valls, G., Tuia, D., Gómez-Chova, L., Jimenez, S. and Malo, J., 2011. Remote Sensing Image Processing. Synthesis Lectures on Image, Video, and Multimedia Processing, Morgan and Claypool.

Crawford, M. M., Tuia, D. and Hyang, L. H., 2013. Active learning: Any value for classification of remotely sensed data? Proc. IEEE 101(3), pp. 593–608.

Dalla Mura, M., Atli Benediktsson, J., Waske, B. and Bruzzone, L., 2010. Morphological attribute profiles for the analysis of very high resolution images. IEEE Trans. Geosci. Remote Sens. 48(10), pp. 3747–3762.

Fauvel, M., Tarabalka, Y., Benediktsson, J. A., Chanussot, J. and Tilton, J. C., 2013. Advances in spectral-spatial classification of hyperspectral images. Proc. IEEE 101(3), pp. 652 – 675.

Foody, G. M., 2004. Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. Photogramm. Eng. Rem. S. 50(5), pp. 627–633.

Leiva-Murillo, J., Gomez-Chova, L. and Camps-Valls, G., 2013. Multitask remote sensing data classification. IEEE Trans. Geosci. Remote Sens. 51(1), pp. 151 –161.

Li, W. and Du, Q., in press. Gabor-filtering based nearest regularized subspace for hyperspectral image classification. IEEE J. Sel. Topics Appl. Earth Observ.

Lillesand, T. M., Kiefer, R. W. and Chipman, J., 2008. Remote Sensing and Image Interpretation. J. Wiley & Sons, NJ, USA.

Moser, G., Serpico, S. B. and Benediktsson, J. A., 2013. Land-cover mapping by Markov modeling of spatial-contextual information. Proc. IEEE 101(3), pp. 631–651.

Obozinski, G., Taskar, B. and Jordan, M., 2006. Multi-task feature selection. Statistics Department, UC Berkeley, Tech. Rep.

Pacifici, F., Chini, M. and Emery, W., 2009. A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. Remote Sens. Environ. 113(6), pp. 1276–1292.

Pacifici, F., Du, Q. and Prasad, S., 2013. Report on the 2013 IEEE GRSS data fusion contest: Fusion of hyperspectral and LiDAR data. IEEE Remote Sens. Mag. 1(3), pp. 36–38.

Perkins, S., Lacker, K. and Theiler, J., 2003. Grafting: Fast, incremental feature selection by gradient descent in function space. J. Mach. Learn. Res. 3, pp. 1333–1356.

Plaza, A., Benediktsson, J. A., Boardman, J., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Marconcini, M., Tilton, J. and Trianni, G., 2009. Recent advances in techniques for hyperspectral image processing. Remote Sens. Environ. 113(Supplement 1), pp. S110–S122.

Rakotomamonjy, A., Flamary, R. and Yger, F., 2013. Learning with infinitely many features. Machine Learning 91(1), pp. 43–66.

Rakotomamonjy, A., Flamary, R., Gasso, G. and Canu, S., 2011. lp-lq penalty for sparse linear and sparse multiple kernel multitask learning. Neural Networks, IEEE Transactions on 22(8), pp. 1307–1320.

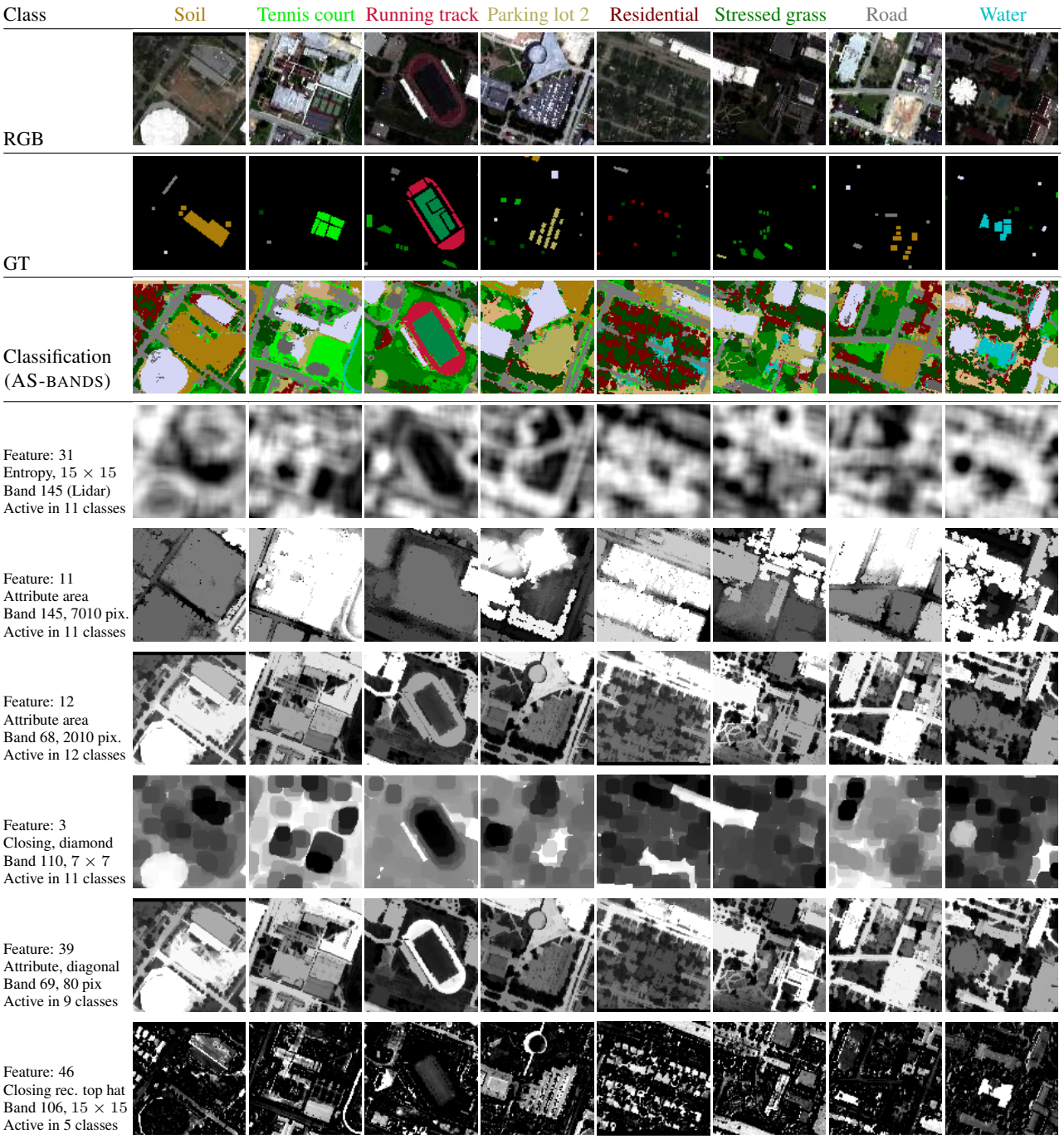


Figure 7: Visualization of the features with highest $||W_{j,\cdot}||_2$ for one run of the HOUSTON 2013 results (cf. bottom matrix of Fig. 6). First row: RGB subsets; Second row: ground truth; Third row: output of the classification with the proposed approach; Fourth row to end: visualization of the six features with highest squared weights.

Richards, J. A. and Jia, X., 2005. Remote Sensing Digital Image Analysis: An Introduction. 4th edn, Springer, Berlin, Germany.

Schindler, K., 2012. An overview and comparison of smooth labeling methods for land-cover classification. IEEE Trans. Geosci. Remote Sens. 50(11), pp. 4534–4545.

Taubenboeck, H., Esch, T., Wiesner, M., Roth, A. and Dech, S., 2012. Monitoring urbanization in mega cities from space. Remote Sens. Enviro. 117, pp. 162–176.

Tuia, D., Pacifici, F., Kanevski, M. and Emery, W., 2009. Classification of very high spatial resolution imagery using mathematical morphology and support vector machines. IEEE Trans. Geosci. Remote Sens. 47(11), pp. 3866–3879.

Tuia, D., Volpi, M., dalla Mura, M., Rakotomamonjy, A. and Flamary, R., 2014. Automatic feature learning for spatio-spectral image classification with sparse SVM. IEEE Trans. Geosci. Remote Sens. 52(10), pp. 6062–6074.

Yuan, M. and Lin, Y., 2007. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68(1), pp. 49–67.