Learning Relevant Image Features With Multiple-Kernel Classification

Devis Tuia, *Member, IEEE*, Gustavo Camps-Valls, *Senior Member, IEEE*, Giona Matasci, and Mikhail Kanevski

***Abstract*—The increase in spatial and spectral resolution of the satellite sensors, along with the shortening of the time-revisiting periods, has provided high-quality data for remote sensing image classification. However, the high-dimensional feature space induced by using many heterogeneous information sources precludes the use of simple classifiers: thus, a proper feature selection is required for discarding irrelevant features and adapting the model to the specific problem. This paper proposes to classify the images and simultaneously to learn the relevant features in such high-dimensional scenarios. The proposed method is based on the automatic optimization of a linear combination of kernels dedicated to different meaningful sets of features. Such sets can be groups of bands, contextual or textural features, or bands acquired by different sensors. The combination of kernels is optimized through gradient descent on the support vector machine objective function. Even though the combination is linear, the ranked relevance takes into account the intrinsic nonlinearity of the data through kernels. Since a naive selection of the free parameters of the multiple-kernel method is computationally demanding, we propose an efficient model selection procedure based on the kernel alignment. The result is a weight (learned from the data) for each kernel where both relevant and meaningless image features automatically emerge after training the model. Experiments carried out in multi- and hyperspectral, contextual, and multisource remote sensing data classification confirm the capability of the method in ranking the relevant features and show the computational efficience of the proposed strategy.**

***Index Terms*—Feature selection, image classification, kernel alignment, multiple-kernel learning (MKL), ranking, recursive feature elimination (RFE), support vector machine (SVM).**

I. Introduction

In RECENT years, satellite sensors design has known an unprecedented development in terms of both radiometric and spatial resolution. Thus, the amount of information sensed has increased constantly, opening a wide area of interesting and challenging applications for remote sensing. However, in light of the mass of information sensed, the need for efficient data processing tools capable of treating the acquired images efficiently is even stronger than before.

Statistical models [1] have been proposed as tools to treat remote sensing data efficiently: neural networks and more recently, support vector machines (SVMs), [2]–[5] have been proven to be robust and efficient methods to solve remote sensing image classification problems. Nonetheless, these methods show two main problems. First, they depend on the size of the data set. Even if it is less prone to the curse of dimensionality [6], the number of the parameters of SVM still depends on the number of observations and features. When the ratio between the number of observations and the number of bands is low, SVM may fail to find the correct solution. In remote sensing, this situation happens in several typical settings, for instance: 1) when several sensors' bands are used simultaneously [7]–[9]; 2) when multitemporal images are needed [9]–[11]; and/or 3) when contextual information is included [12]–[14]; in these cases, the size of the feature space strongly increases and the risk of overfitting becomes real.

Secondly, SVMs work as a black box model and no insight about the importance of the distinct features can be obtained directly from the model's solution. Aside from the model's accuracy, insight about the relative importance of the single features/sensors can prove to be a valuable information for the end-user. Such knowledge can help the user understand some physical properties of the problem at hand, and focus the data acquisition/processing on specific sensors or types of features. For instance, hyperspectral features are strongly correlated, and valuable and independent information can often be summarized in a well-chosen subset. Such set represents physical properties of the objects of interest. As for contextual features, which can be generated at different scales, the selection of good scales and types of features is useful and crucial information for the user. For instance, in [15], authors identified a set of very informative textural features for urban land-cover classification. The results obtained for an independent test case indicated the value of reducing the input set down to only ten features and their ability to generalize to new scenes.

Both problems can be addressed by using feature selection [16] algorithms, which allow to select informative and relevant input features by analyzing their relevancy for a certain classification problem. Feature selection algorithms can be divided into three classes: 1) filters; 2) wrappers; and 3) embedded methods [16], [17]. Filters rank features according to either a similarity measure such as the correlation, or a measure of distance between distributions (for instance, the Kullback–Leibler divergence). Filters may be considered as a preprocessing step and are generally independent from the selection phase.

Manuscript received November 6, 2009; revised March 11, 2010. Date of publication June 28, 2010; date of current version September 24, 2010. This work was supported in part by the Swiss National Science Foundation under Grants 200021-126505 and PBLAP2-127713/1 and in part by the Spanish Ministry of Education and Science under projects TEC2006-13845, AYA2008-05965-C04-03, and CONSOLIDER/CSD2007-00018.

D. Tuia, G. Matasci, and M. Kanevski are with the Institute of Geomatics and Analysis of Risk, University of Lausanne, 1015 Lausanne, Switzerland (e-mail: devis.tuia@unil.ch; giona.matasci@unil.ch; mikhail.kanevski@unil.ch).

G. Camps-Valls is with the Image Processing Laboratory (IPL), Universitat de València, 46010 València, Spain (e-mail: gustavo.camps@uv.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2010.2049496

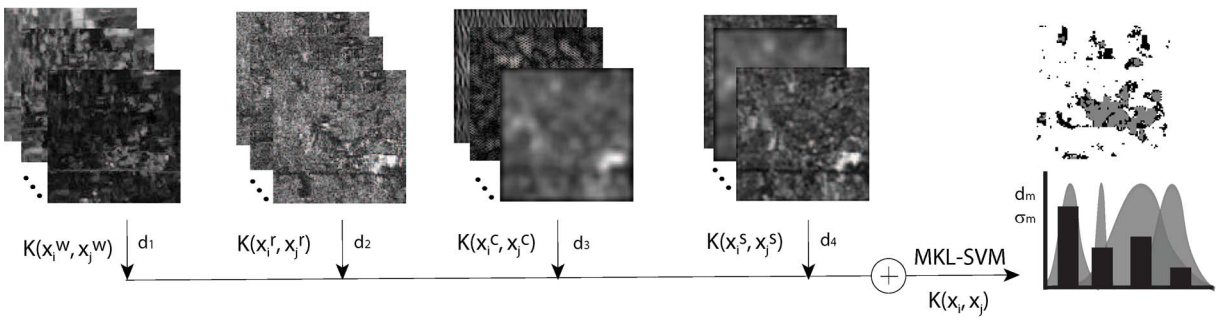


Fig. 1. Illustrative example of the multiple-kernel approach to information fusion for image classification. A set of different features (optical $[\mathbf{x}^w]$, radar $[\mathbf{x}^r]$, contextual $[\mathbf{x}^c]$, and textural $[\mathbf{x}^s]$) are usually considered for image classification. MKL builds a multiple kernel by a linear weighted combination of dedicated RBF kernels with respective weights d_m and kernel parameters σ_m. Using MKL, both parameters are optimized simultaneously.

Remote sensing applications of filters can be found in [18]–[23]. Wrappers use the model's accuracy to rank subsets of features according to their predictive power, but they imply a high computational cost. Embedded methods differ from wrappers because they allow interaction between the feature selection procedure and the model.

A well-known embedded backward selection algorithm is the recursive feature elimination (RFE), which uses the changes in the decision function of the SVM as a criterion for the selection [24], [25]. In [26], SVM-RFE is used to select subsets of contextual features. In [27], an embedded feature selection is proposed. The algorithm, based on boosting, finds an optimal weighting and eliminates the bands linked to small weights. In [28], a genetic algorithm is used to select an optimal subset of features for a successive SVM classification. Other genetic algorithm-based wrappers have recently been proposed in [29] for rapid disaster image retrieval and, in [30] and [31], for hyperspectral image classification, where sparsity of the selection must be enforced. A typical drawback of the SVM-RFE is the high computational load: at each stage of the selection and for ζ active variables, $\zeta + 1$ SVM models need to be evaluated, one for the model with the complete feature set and one for each subset composed by all the features minus the one considered. Moreover, no efficient stopping criterion has been proposed and the model should be run for all the features before the selection of the optimal set. Finally, the ranking of the features does not provide a feature weighting since all the features are given the same weight in the kernel function.

A possible answer to the needs of the SVM-based feature selection can be found in the framework of multiple kernel learning (MKL) [32]. In MKL, the SVM kernel function is defined as a weighted linear combination of M base kernels built using subsets of features. When using RBF kernels, each subkernel K_m owns a kernel parameter σ_m and a weight d_m (see Fig. 1). MKL works iteratively, optimizing both the d_m and the σ_m[33]. This way, the SVM model is optimized at the same time as the combination of base kernels. When constructing the base kernels with single features (or physically inspired groups of features), the optimization of the SVM kernel works as a feature selector providing a weighted ranking of the importance of its components. Since zero-weight kernels does not contribute to the final function, the features used to build the zero-weight kernels are automatically ignored in the final model.

The so-called composite kernels framework was developed to combine spectral with contextual [34]–[37] multitemporal and multisource [9] information for image classification. The models permits the study of the importance of the different sources of information. Besides, by optimizing a weight per kernel, useful information about the associated features is obtained. However, in these works, the tuning of the parameters in this framework is done through cross-validation strategies, which dramatically increase the computational cost when several kernels are combined. So far, the only application in remote sensing of strict MKL can be found in [38] and, taking advantage of a similar idea, spectrally weighted kernels are proposed in [39].

The main problem of MKL is that, even though it is a convex problem, the minimization is not smooth [40]. Very recently, other algorithmical approaches have reformulated the original version of the MKL algorithm [33], [41]. In [42], the different optimization problems are summarized and compared. The computational cost involved and the nonsmoothness issue are critical constraints of the different methods that are, actually, not applicable for remote sensing data. In this paper, the application of MKL to remote sensing data is studied by applying a recently proposed method called SimpleMKL [43]. This method solves the MKL problem by gradient descent over the SVM decision function and converges to the same solution as the other MKL formulations mentioned above. Computational issues are considered, addressing the problems related to the model selection strategy of [43], which is performed by stacking several kernels sharing the same features, but using different kernel parameters. To solve this problem, a model selection strategy based on kernel alignment [44] is proposed, similarly to [45], and analyzed in detail. The strategy proposed in this paper decreases the computational cost of MKL and the memory requirements in a straightforward yet very effective way.

This paper also considers challenging experimental settings with strongly unbalanced pixels/features ratios. Three typical remote sensing image classification scenarios are studied: 1) mixed spectral/contextual very high-resolution classification; 2) hyperspectral data classification; and 3) multisource image classification. Results are discussed both in terms of the model accuracy and of the image properties discovered, showing that the MKL provides nondegenerated SVM solutions, and bringing additional information about the image properties. Even

though a marginal decrease in performance can sometimes be observed when using MKL with respect to the standard SVM [46], the loss is compensated both by the benefits of the weighted feature ranking (which is consistent with the result obtained with other feature selection methods) and by the compactness of the resulting solution.

For all these reasons, the novelty of the paper can be found in the following: 1) the evolution from composite to multiple kernels, which has never been studied intensively in the remote sensing literature; 2) the kernel alignment-based model selection, which addresses one of the main issues of SimpleMKL; and 3) the study of the image properties obtained from MKL results, that allow to understand the role of the different features in the classification process.

The remainder of the paper is organized as follows. Section II reviews both the framework of kernel methods, paying special attention to their general properties, and the formulation of the standard SVM classifier. Section III revises the MKL framework. Since the definition of the family of "kernels on features" is critical, a detailed discussion of model selection and design is given in Section III-C. Section IV describes the data and the experimental setup of the experiments presented in Section V. Finally, Section VI concludes the paper.

II. Kernel Methods and Data Classification

Kernel methods offer a general framework for machine learning problems (classification, clustering, regression, density estimation, and visualization) with heterogeneous types of data, such as the following: 1) time series; 2) images; 3) strings; or 4) objects [47], [48]. In this section, we briefly review the main properties of Mercer's kernels, and the standard formulation for the binary SVM [49] used in this paper.

A. Background on Kernel Methods

When using linear algorithms, a well-established theory and efficient methods are available. Kernel methods exploit this fact by embedding a data set $S = \{\mathbf{x}_i\}_{i=1}^n$ defined over the input or attribute space $\mathcal{X}$ ($S \subseteq \mathcal{X}$) into a higher (possibly infinite) dimensional Hilbert space $\mathcal{H}$, or feature space, and then they build a linear algorithm therein, resulting in an algorithm which is nonlinear with respect to the input data space. The mapping function is denoted as $\mathbf{\Phi} : \mathcal{X} \to \mathcal{H}$. Though linear algorithms will benefit from this mapping because of the higher dimensionality of the feature space (see Cover's theorem, which guarantees that the transformed samples are more likely to be linearly separable [50]), the computational load would dramatically increase because we should compute sample coordinates in that high-dimensional space. Such a computation is avoided through the use of the kernel trick: if an algorithm can be expressed with dot products in the input space, its (nonlinear) kernel version only needs the dot products among mapped samples. Kernel methods compute the similarity between the training examples using pairwise inner products between mapped samples, and therefore the so-called kernel matrices contain all the necessary information to perform many classical linear algorithms in the feature space.

B. Support Vector Machine (SVM)

Given a labeled training data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, and $y_i \in \{-1, +1\}$, and given a nonlinear mapping $\boldsymbol{\phi}(\cdot)$, the SVM method solves [49]

$$\min_{\mathbf{w},\xi_i,b} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \right\} \tag{1}$$

constrained to

$$y_i \left(\langle \boldsymbol{\phi}(\mathbf{w}), \mathbf{x}_i \rangle + b \right) \geq 1 - \xi_i \quad \forall i = 1, \ldots, n \tag{2}$$

$$\xi_i \geq 0 \quad \forall i = 1, \ldots, n \tag{3}$$

where $\mathbf{w}$ is the vector of parameters defining the optimal decision hyperplane $\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}) \rangle + b = 0$, and b represents the bias. These parameters define a linear classifier in the Hilbert space $\mathcal{H}$

$$\hat{y}_* = f(\mathbf{x}_*) = \operatorname{sgn}\left(\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_*) \rangle + b \right). \tag{4}$$

The regularization parameter C controls the generalization capabilities of the classifier and must be selected by the user. Positive slack variables ξ_i allow to deal with the permitted errors.

The primal problem (1) is solved by maximizing its dual counterpart [47]

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle \right\} \tag{5}$$

constrained to $0 \leq \alpha_i \leq C$ and $\sum_i \alpha_i y_i = 0$, $\forall i = 1, \ldots, n$, where auxiliary variables α_i are Lagrange multipliers corresponding to restrictions (2) and (3). This way, the explicit estimation of the very high-dimensional vector $\mathbf{w}$ is avoided by estimating a 1-D parameter vector $\boldsymbol{\alpha}$. Finally, the decision function for any test vector $\mathbf{x}_*$ is given by

$$f(\mathbf{x}_*) = \operatorname{sgn}\left(\sum_{i=1}^{n} y_i \alpha_i \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_*) \rangle + b \right) \tag{6}$$

where b can be easily computed from a few support vectors (SVs), which are those training samples $\mathbf{x}_i$ with $\alpha_i \neq 0$ [47].

It is worth noting that all $\boldsymbol{\phi}$ mappings used in the SVM learning [(5)] and prediction [(6)] occur in the form of inner products. As mentioned above, this allows to replace such products with a kernel function K

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle \tag{7}$$

and then, without considering the mapping $\boldsymbol{\phi}$ explicitly, a nonlinear SVM can be defined.

C. Kernel Functions and Basic Properties

The bottleneck for any kernel method is the definition of a kernel K that accurately reflects the similarity among samples. However, not all metric distances are permitted. In fact, valid kernels are only those fulfilling the Mercer's Theorem [51] and the most common ones are the following: 1) the linear

$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$; 2) the polynomial $K(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1)^d$, $d \in \mathbb{N}$; and 3) the radial basis function (RBF), $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2)$, $\sigma \in \mathbb{R}^+$.

In this paper, we will take advantage of the properties of Mercer's kernels stated below.

Property 1: Be K_1 and K_2, two Mercer's kernels on $S \times S$ and $\mu \geq 0$. Then, the following kernels:

$$K(\mathbf{x}_i, \mathbf{x}_j) = K_1(\mathbf{x}_i, \mathbf{x}_j) + K_2(\mathbf{x}_i, \mathbf{x}_j) \tag{8}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mu K_1(\mathbf{x}_i, \mathbf{x}_j) \tag{9}$$

are valid Mercer's kernels.

Therefore, one can design kernels by summing up (weighted) valid kernels. This intuitive idea is formally extended for optimizing linearly weighted combinations of kernels in the following sections.

III. Efficient Multiple-Kernel Learning

This section reviews the main formulation of MKL [32]. Then, a recently proposed method for solving the MKL problem, named SimpleMKL [43], is described in detail. Finally, we present a straightforward yet efficient way to reduce the computational cost when using this method for joint feature selection and classification.

A. Multiple-Kernel Learning: Formulation and Problems

As mentioned above, the success of kernel methods depends strongly on the data representation encoded into the kernel function. In the case of SVM, by plugging (7) into (6), the decision function takes the form

$$f(\mathbf{x}_*) = \operatorname{sgn}\left(\sum_{i=1}^{n} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}_*) + b \right). \tag{10}$$

Common kernel functions, like the polynomial or RBF, are rigid representations of the data, that may be replaced by more flexible and data-adapted kernels. In the multiple-kernel framework, the optimal kernel is learned from data by building a weighted linear combination of M base kernels. Each kernel in the mixture may account for different features or set of features.

The use of multiple kernels can enhance the performance of the model and, more importantly, the interpretability of the results. Let $\mathbf{d} = [d_1, \ldots, d_m, \ldots, d_M]^\top$ be a vector of weights for the mixture of kernels. A mutiple kernel is the combination of the M basis kernels K_m

$$\begin{aligned} &K(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{M} d_m K_m(\mathbf{x}_i, \mathbf{x}_j) \\ &\text{s.t. } d_m \geq 0 \\ &\sum_{m=1}^{M} d_m = 1. \end{aligned} \tag{11}$$

Multiple-kernel learning aims at simultaneously optimizing α_i in (10) and d_m in (11), subject to positivity and sum-to-one constraints. In our case, the learned (optimized) weights d_m will directly give a ranked relevance of each set of features, e.g., groups of bands or contextual and spectral features.

Even being a very attractive formulation, solving the MKL problem becomes rapidly intractable with the increase of training examples and number of kernels. For example, in binary classification, the MKL results in a quadratically constrained quadratic programming problem, which is computationally unaffordable when a high number of samples or kernels is used [32]. The main problem is that, even being convex, the minimization is not smooth [40]. Very recently, other algorithmical approaches have reformulated the original version of the MKL [33], [41]. Again, the computational cost involved and the nonsmoothness issue are the critical constraints of the methods. In the next section, we summarize a simple MKL formulation [43] that solves the aforementioned problems and converges to the same solution of the MKL formulations in [33], [40], and [41].

B. Simple Multiple-Kernel Learning

SimpleMKL is an efficient algorithm to solve the MKL problem [43]. Similarly to [33], SimpleMKL wraps a SVM solver with a single kernel, which is already the linear combination in (11). Essentially, the algorithm is based on a gradient descent on the SVM objective value. Note that by using (11) in (7), and then plugging it into (5), the following multiple-kernel dual problem is obtained:

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \sum_{m=1}^{M} d_m K_m(\mathbf{x}_i, \mathbf{x}_j) \right\} \tag{12}$$

constrained to $0 \leq \alpha_i \leq C$, $\sum_i \alpha_i y_i = 0$, $\forall i = 1, \ldots, n$, $\sum_m d_m = 1$ and $d_m \geq 0$.

At this point, one can show (see [43]) that maximizing the dual problem in (12) is equivalent to solving the problem

$$\min_{\mathbf{d}} J(\mathbf{d}) \quad \text{such that} \quad \sum_{m=1}^{M} d_m = 1, \quad d_m \geq 0 \tag{13}$$

where

$$J(\mathbf{d}) = \begin{cases} \min_{\mathbf{w}, b, \xi}, & \frac{1}{2} \sum_{m=1}^{M} \frac{1}{d_m} \|\mathbf{w}_m\|^2 + C \sum_{i=1}^{n} \xi_i \\ \text{s.t.}, & y_i \left(\sum_{m=1}^{M} \langle \mathbf{w}_m, \boldsymbol{\phi}_m(\mathbf{x}_i) \rangle + b \right) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{cases} \tag{14}$$

and $\mathbf{w}_m$ represents the weights of the partial decision function of the subproblem m with associated kernel mapping $\boldsymbol{\phi}_m(\mathbf{x}_i)$.

Algorithm 1 SimpleMKL (adapted from [43])

1: initialize the weights $d_m = (1/M), \quad m = \{1, \cdots, M\}$
2: compute the objective value $J(\mathbf{d})$ according to Eq. (14).
3: **repeat**
4: compute the reduced gradient and find the descent direction $\mathbf{D}$. Set $\mu = \arg\max\{d_m\}$
5: **repeat** {Descent direction update}
6: find the component $\nu = \arg\min\{-d_m / D_m\}$
7: find maximum admissible step size $\gamma_{\max} = -d_\nu / D_\nu$
8: update $\mathbf{d} = \mathbf{d} + \gamma_{\max} \mathbf{D}$, set $D_\mu = D_\mu - D_\nu$, $D_\nu = 0$ and normalize $\mathbf{D}$

9: compute the new $J(\mathbf{d})$
10: **until** $J(\mathbf{d})$ stops decreasing or when it reaches a stopping criterion
11: line search along $\mathbf{D}$ to find the optimal γ
12: **until** a stopping criterion is met.

C. Efficient Model Selection via Kernel Alignment

In the original implementation of SimpleMKL [43], model selection was performed by building several kernels with different kernel parameters. Rather than optimizing the σ values for each kernel, several kernels sharing the features, but with different parameters, were built. This way, the optimization of $\mathbf{d}$ allowed to find automatically the best combination of kernels by finding the non-zero weights. Strictly speaking, no model selection is performed this way. This strategy is clearly heavy in terms of the computational cost, but has the advantage of building multiscale solutions, i.e., solutions where more than a single kernel is selected for a single feature. In this case, two (or more) kernels in the final combination of (11) encode signal similarities for the feature at different ranges (or scales).

When dedicating separate kernels to each variable, MKL can be used for direct feature ranking. Moreover, since the MKL result is sparse, i.e., most of the kernel weights in $\mathbf{d}$ are zero (see [43]), the variables associated to kernels with zero weights are not accounted into the final mixture of kernels. In this sense, MKL can be used as an algorithm of feature selection. Nonetheless, a model selection strategy as the one proposed in [43] is not efficient for this purpose: when confronted to data sets carrying more than 100 bands, such a strategy would imply the creation (and storage in memory) of a kernel of size $(n \times n \times (M \cdot l))$, where n is the number of training pixels, M the number of kernels (using single or groups of features) and l is the number of kernel parameters to search upon. In particular, when n increases, the storage needs become intractable for current computers. Therefore, these facts preclude the use of the original formulation of SimpleMKL for problems including thousands of training pixels or hundreds of features.

To overcome these problems, we propose a technique to estimate reasonable values of kernel-free parameters before training the model. To do this, we use the training labels $\mathbf{y}$ to build an ideal kernel, $\mathbf{K}_{ideal} = \mathbf{y}\mathbf{y}^\top$, and the training samples $\mathbf{x}_i$ to build a set of kernels by varying the free parameter. If the RBF is adopted, $K_\sigma = K(\mathbf{x}_i, \mathbf{x}_j|\sigma)$ for a given σ value. The idea is to evaluate the distance of the different parametrizations of $\mathbf{K}_\sigma$ to the ideal one encoded in the data. For such purpose, we take advantage of the concept of kernel alignment [44], a measure of similarity between matrices. Given a kernel matrix $\mathbf{K}_\sigma$, and a vector of labels $\mathbf{y} \in \{-1, 1\}$, the alignment between them can be written as

$$A(\mathbf{K}_{ideal}, \mathbf{K}_\sigma) = \frac{\langle \mathbf{K}_\sigma, \mathbf{y}\mathbf{y}^\top \rangle_F}{\sqrt{\langle \mathbf{K}_\sigma, \mathbf{K}_\sigma \rangle_F \langle \mathbf{y}\mathbf{y}^\top, \mathbf{y}\mathbf{y}^\top \rangle_F}} \tag{15}$$

where $\langle \cdot, \cdot \rangle_F$ stands for the Frobenius distance between matrices, that is $\langle \mathbf{U}, \mathbf{V} \rangle_F = \sum_{i,j} u_{ij} v_{ij}$. Since the kernel shows high values for similar points, the alignment can be seen as a correlation coefficient between the kernel values and the correct label assignments, and it can take values in the range $[-1, 1]$. In the case of the multiclass classification, the ideal kernel $\mathbf{y}\mathbf{y}^\top$ must be replaced by a kernel returning the value 1 if the considered pixels belong to the same class and 0 otherwise. The algorithmical advantage of this solution is clear: the size of the kernel to store is reduced to $(n \times n \times M)$.

In [44], kernel alignment was used to evaluate combinations of kernels. If two kernels are aligned with the labels vector and not aligned with each other, their combination will be valuable to solve the problem because both kernels contain independent information. In our setting, we select the best candidates for SimpleMKL by maximizing the alignment of each feature's kernels K_m with the output vector. Then, SimpleMKL selects the best combination to solve the problem. Algorithm 2 summarizes the procedure.

Algorithm 2 Parameter optimization with kernel alignment
$M =$ number of kernels;
$\boldsymbol{\sigma} =$ vector of σ to search upon.
1: Compute the ideal kernel $\mathbf{K}_{ideal} = \mathbf{y}\mathbf{y}^\top$
2: **for** m in 1: M **do**
3: **for** σ_j in $\boldsymbol{\sigma}$ **do**
4: Compute kernel alignment $A(\mathbf{K}_{ideal}, \mathbf{K}_{m,\sigma_j})$ using Eq. (15).
5: **end for**
6: Select $\sigma_m = \arg\max_{\sigma_j} \{A(\mathbf{K}_{ideal}, \mathbf{K}_{m,\sigma_j})\}$
7: **end for**

IV. Data Sets Considered and Experimental Setting

This section specifies the data sets studied and the setting of the experiments.

A. Data

Multiple-kernel learning is applied to three challenging remote sensing image classification tasks, accounting for both very high-resolution and multisource problems:

1) The first image considered is a 0.6-m-resolution multispectral scene taken in 2006 by the QuickBird sensor over a part of the city of Zurich, Switzerland. Five classes of interest are considered: 1) "Buildings;" 2) "Roads;" 3) "Forest;" 4) "Shadows;" and 5) "Water." Four multispectral bands, accounting for RGB and near-infrared (NIR) channels along with 18 spatial features extracted using opening and closing by reconstruction morphological filters are used as features for classification. The goal here is not only to obtain good classification maps but also, and more importantly, to discern the relative relevance of the different spatial and spectral features. Note that, since we use morphological operators on some bands, ranking features provides some insight on the spatial scale of objects in the scene.
2) The second case study is an image acquired in 1999 by the HyMap airborne spectrometer over Barrax

TABLE I
REPRESENTATIVE BANDS FOR CLASSIFICATION OBTAINED IN [52]

Bands	λ [μm]	Characteristics
6	0.5030	Leaf pigments (carotenes and chlorophylls).
17	0.6710	Chlorophylla-a maximum absorption.
22	0.7470	Red edge (change Visible-Near Infrared). Leaf Area Index.
24	0.7770	Beginning of Near InfraRed (NIR) with high reflectance and low absorbance. Leaf biomass and structure.
99	1.9860	Water absorption. Soil moisture and leaf water content.
118	2.3210	Water absorption. Dry matter and soil minerals.

(Spain) during the DAISEX99 campaign. The image has 128 bands in the region 0.4 μm–2.5 μm and a spatial resolution of 5 m. The six classes of interest are: 1) "Corn;" 2) "Sugar beets;" 3) "Barley;" 4) "Wheat;" 5) "Alfalfa;" and 6) "Soil" [52]. The main goal in this application is concerned with evaluating the relevance of the different spectral bands for classification. Important gains in accuracy are not expected, since the classification problem is easy due to the high quality of the image acquisition along with the atmospheric and geometric corrections applied, as shown by the good results reported in [52]. In this previous work, the most important features were also highlighted under physical criteria (cf. Table I). The aim is to assess if MKL will end up finding the same (or a similar) set of features.

3) The third and last case study considers multisource information. The scene is a set of images of the city of Naples, Italy, used to detect urban areas. It is a binary detection problem with classes "Urban" and "Not urban." Images from ERS2 SAR and Landsat TM sensors were acquired. In the case of the optical images, the seven Landsat TM spectral bands (containing three RGB, one NIR, two short-wave IR, and one thermal IR bands) were directly used. For the SAR image, two backscattering intensities were available. Using them, the interferometric coherence was extracted and added as a third variable [53]. However, since speckle disturbs image interpretation, a multistage spatial filtering approach over coherence images was followed to increase the urban areas discrimination [54], which yielded the fourth radar input feature, that should be the most useful to solve the problem.

B. Experimental Setting

For each image classification problem, two types of experiments have been done:

- First, MKL setups using a single feature in each kernel have been considered. These experiments, represented by the abbreviation "**S**," result in single features ranking, and can be used for feature selection to interpret different properties of the images, either signal properties (spectral) or objects scales (contextual features).
- Secondly, MKL setups employing kernels built using groups of features have been studied. These experiments, represented by the abbreviation "**G**," allow us to estimate the relative importance of parts of the signal or types of contextual features. Since they do not give insight on single features importance, the "**G**" setting cannot be considered for *stricto sensu* feature selection. For the three images studied, the groups considered are the following:
  1) QuickBird (Zürich): types of information, divided in multispectral (four bands), opening by reconstruction (nine features), closing by reconstruction (nine features);
  2) DAISEX (Barrax): main physical properties of the spectrum [52]. Leaf pigments (bands 1–23), cell structure (24–59), leaf water content (60–128);
  3) Multisource (Naples): type of sensors, considering Landsat TM (7 bands) and ERS2 SAR (4 features).

Regarding model selection, the two strategies described in Section III-C have been considered in the experiments. In the following, the strategy stacking multiple kernels with different parameters is abbreviated by "**M**") and the one using kernel alignment by "**A**."

The combination of the two types of experiments with the two strategies of model selection described in Section III-C provides the four experiments detailed in Table II. All the models have been run using the MKL toolbox of [43] under MATLAB 7. As a comparison, a standard SVM model using a single kernel for all the features has been run using the same toolbox.

Regarding the single experiments, each one has been run ten times with different starting training points and using increasing number of labeled pixels per class, randomly picked from the ground survey pixels. The results reported are mean values for the experiments and their standard deviation provides an insight about the dependence of the model to the initial conditions. The scenario using 5 pixels per class is very challenging, and is here studied to assess the stability of MKL in settings with a small ratio between the number of labeled pixels and the number of parameters. Testing is performed using separate data sets of size 97 000 (QuickBird Zürich), 5000 (DAISEX Barrax), and 140 806 (multisource Naples) pixels.

All the kernels used are RBF with associated width σ. For the "**M**" experiments, four RBF kernels per (group of) feature(s) using different σ have been used. The number of σ parameters is kept small to limit the computational complexity and memory requirements; as mentioned, this model selection strategy builds (and stores) a matrix of size $(n \times n \times s \cdot l)$, where n is the number of training pixels, s is the number of sources (single or groups of features), and l is the number of parameters to test. In the settings considered, the values $\sigma = [0.1, 0.25, 0.35, 0.5]$ have been used. These values have been highlighted as reasonable after the study of different combinations of parameters. The need for such a previous study is a drawback for the "**M**" strategy; however, the study of larger number of σ values heavily impacts the computational efficiency of the algorithm. For the standard SVM and the "**A**" strategy, a grid search has been run for the kernel(s) parameters, with $\sigma \in [0.01, \ldots, 3]$. In this case, only a $(n \times n \times s)$ matrix is stored. For all the

TABLE II
SUMMARY OF THE SETUP FOR EACH EXPERIMENT CONSIDERED IN SECTION V

	Experiment 1 Context-based Multispectral		Experiment 2 Hyperspectral		Experiment 3 Multi-source			
Features	Spectral and morphological		Spectral bands		Landsat only		Landsat + SAR	
Exp.	Feat. per kernel	# kernels	Feat. per kernel	# kernels	Feat. per kernel	# kernels	Feat. per kernel	# kernels
SA SM	Single	22 $22 \cdot 4 = 88$	Single	128 $128 \cdot 4 = 512$	Single	7 $7 \cdot 4 = 28$	Single	11 $11 \cdot 4 = 44$
GA GM	Types of information	3 $3 \cdot 4 = 12$	Physically inspired	3 $3 \cdot 4 = 12$	–	–	Type of sensor	2 $2 \cdot 4 = 8$
SVM	All	1	All	1	All	1	All	1

experiments, SVM regularization parameter has been optimized by cross-validation in the interval $C \in [100, \ldots, 10000]$.

V. EXPERIMENTAL RESULTS

This section describes the results for the three classification scenarios: 1) spectral/contextual image classification; 2) hyperspectral pixel-based image classification; and 3) multisource remote sensing data classification. In all cases, the goal is twofold: 1) improve (or at least conserve) classification performance in a supervised way; and 2) to extract knowledge from the built classifier with regard to the rank of the features. For the experiments, images from four different sensors have been considered: 1) QuickBird; 2) HyMap; 3) Landsat TM; and 4) ERS2 SAR. We compare the presented method to the standard SVM in terms of the accuracy and robustness to the number of training samples. Moreover, the analysis of the weights d_m of the obtained model is carried out. Finally, in Section V-D, the proposed method is compared with the standard feature selection methods.

A. Experiment 1: Contextual-Based Multispectral Image Classification

The first experiment deals with a common problem in remote sensing of classification of very high- resolution images combining spectral bands and contextual features extracted from the panchromatic image.

Numerical results illustrating the kappa statistic (κ) curves in test for this experiment are reported in Fig. 2. All the experiments with the SimpleMKL method clearly outperform SVM using the same number of training pixels; an excellent gain between 0.1 and 0.15 is achieved across for all the experiments. These results demonstrate that, by properly weighting the importance of the features, we can construct efficient kernel machines encoding the relationships in the observed data. Also note that the differences are consistent for different realizations; when using more than ten samples/class bars of MKL and SVM do not overlap. This goes in line with the fact that a higher number of kernels needs a larger number of samples for proper estimation of the sample similarity, and when this is achieved, MKL boosts its results. We also performed a statistical analysis of the differences between the SVM and the MKL results for the GM case. The McNemar's test [55] was used for assessing the statistical differences between MKL and SVM. The difference in accuracy between the two classifiers is said to be statistically significant if $|z| > 1.96$, and the sign of z indicates which is the most accurate classifier. In this experiment, we obtained

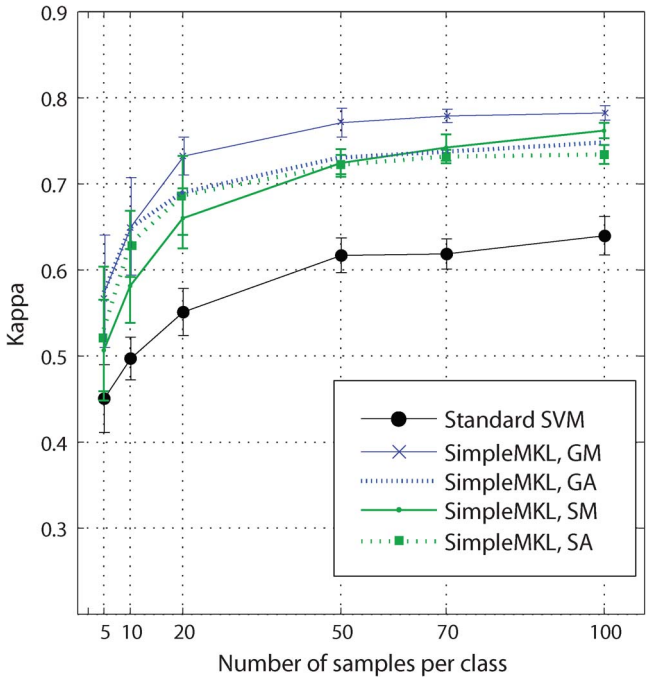


Fig. 2. Kappa statistic for the QuickBird image over Zurich (Switzerland).

$z = 70.13$ when using 20 pixels per class, and $z = 53.61$ when using 100 pixels per class. Therefore, in both situations, MKL is significantly different (and better) than SVM, but as the number of samples increases, the difference is lower, which matches the quantitative results. Both SVM and MKL saturate the performance for training sets using more than 50 samples/class (see Fig. 2).

In terms of kappa statistics, the GM experiment shows the best results. The simplicity of this solution confirms the intuition that each type of information may be related to different model parameters. On average, the optimal weights $\mathbf{d}$ are $\{K_{MS} = 0.7, K_{Or} = 0.1, K_{Cr} = 0.2\}$, matching the results found in [37]. The SA experiment shows good performances, yet inferior to those in the GM; the precomputation of the alignment avoids optimizing an 88-dimensional vector and the benefits of using such methods can be observed when sufficient examples are available. Nonetheless, the SM and SA experiment allow to visualize the chosen features (see Fig. 3 for an illustration of the iterative optimization of the weights using SM), starting by a uniform configuration of weights $(d_m = 1/M = 0.011, \forall m \in M)$, the NIR band is given a strong weight after five iterations. The blue and green bands are also selected in the following steps. This way of selecting features

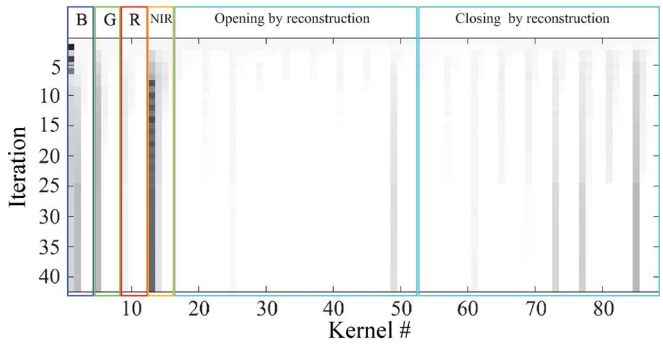


Fig. 3. Optimization of the weight vector $\mathbf{d}$: each line corresponds to an iteration of MKL for the QuickBird image (SM experiment, run #4, 88 kernels). Dark values indicate higher relevance. Since at the beginning $d_m = 1/M, \forall m$, first line shows equal light tones.

matches the classification results, while NIR bands serve to discriminate man-made from natural classes, the proper assignment to the green band (suited for forests) and blue band (water class) largely improves results. Regarding the morphological features, closing features related to the large structuring elements are retained and all the openings are concentrated into a single kernel, the #49, corresponding to a feature extracted with a large structuring element and a kernel built with a small σ; such kernel encodes local neighboring information and smooths the spatial details of the scene. Note that the solution shown in Fig. 3 has a multiscale nature in the sense that two kernels using different σ values are retained into the final solution for the blue or NIR bands. These kernels encode short- and middle-signal relationship between training data, which matches the wide range of object sizes in the image.

B. Experiment 2: Hyperspectral Image Classification

This experiment deals with the pixel-based hyperspectral image analysis. Here, the interest is to evaluate the relevance of the different spectral bands for image classification. Because of the relative simplicity of the classification problem, improvements of the standard SVM accuracy are not expected. Nevertheless, since a list of useful features for the classification is available (cf. Table I), the aim of this study is to evaluate if MKL automatically builds a kernel using these useful features.

Overall, estimated kappa statistic curves for the HyMap experiment are reported in Fig. 4; as expected in this case, MKL models may perform as well as SVM, but no clear improvement is observed. Considering the model comparison shown in Fig. 4, one can remark that no relevant numerical differences exist in the average accuracies of the GM, GA, SA, and standard SVM experiments for training sets using more than ten pixels per class. Additionally, in all cases, standard deviation overlaps, thus suggesting no statistically significant differences among methods. Nevertheless, an important issue is that an insight on the relevant features (see Fig. 5) can be automatically obtained using the MKL approach (SA experiment) without loss in the quality of the final result. It should be mentioned that the SM experiment shows a deterioration of the performance with respect to the standard SVM (between 0.05 and 0.1 in kappa) because, when dealing with hyperspectral images, such a model selection strategy the model overfits the

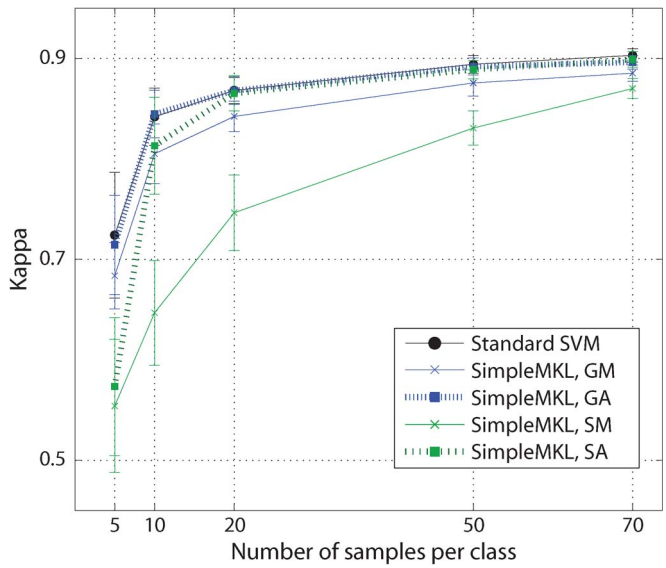


Fig. 4. Kappa statistic for the HyMap image over Barrax (Spain).

data (512 kernels and few training samples are used). Fig. 5(a) illustrates how the gradient algorithm iteratively optimizes the dedicated kernel weights assigning the proper relevance to the 128 spectral channels of the image. It is worth noting that, in this problem, the selected features by the method [Fig. 5(b)] are essentially the six most important spectral bands identified in [52] through careful heuristic analysis of the main nodes and surrogates of a classification tree. Along them, some other bands representing other discriminant regions of the spectrum are selected (bands 49 and 64). This fact not only confirms the correctness of the proposed model selection, but also offers a way to obtain trustable insight on the model (see Section V-D for a comparison with standard feature selection methods).

C. Experiment 3: Multisource Image Classification

Numerical results for the multisource image are reported in Fig. 6. First, the Landsat image is used alone to detect the urban areas [Fig. 6(a)]. The problem is complex and the standard SVM cannot treat the problem effectively; its performances saturate around a kappa index of about 0.55. On the contrary, SimpleMKL results in better performances, reaching values of kappa of 0.6–0.65. As observed earlier, MKL overfits when only few training samples are available (five in this case), but starting from ten training examples per class, the model converges to an optimal solution and outperforms the standard SVM.

Fig. 6(b) illustrates the second experiment, exploiting multisource information. The problem being easier, the standard SVM provides a correct and stable solution saturating around 0.8. Again, MKL cannot model properly the patterns where too few training examples are available, but from 20 examples per class on, SimpleMKL converges into an equivalent (GM) or better (SA, SM, GA) solution saturating around kappa values of 0.85.

Regarding the ranking of (groups of) features, Fig. 7 shows the results for the two experimental settings considered. First, the SA experiment using the Landsat bands only [Fig. 7(a)]

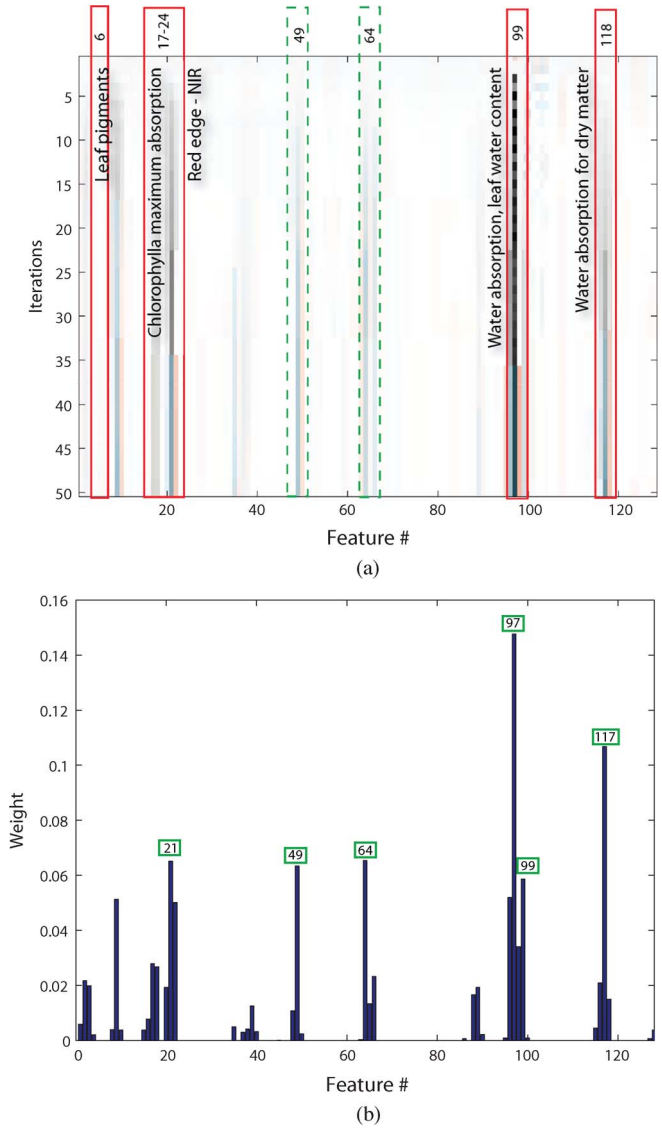


Fig. 5. HyMap image of Barrax, Spain. (a) Optimization of the weight vector $\mathbf{d}$ for the run #9 of the SA experiment (128 kernels) using 70 samples per class: each line corresponds to an iteration of SimpleMKL. Since at the beginning $d_m = 1/M, \forall m$, first line shows equal light tones. Boxes correspond to the features highlighted in Table I. With the exception of the feature #6, all the relevant features receive a non-zero weight. Dashed boxes correspond to additional features selected by SimpleMKL. (b) Average values of the weight vector $\mathbf{d}$ using 70 samples per class (on the ten runs of the SA experiment). The six most relevant bands are highlighted in green.

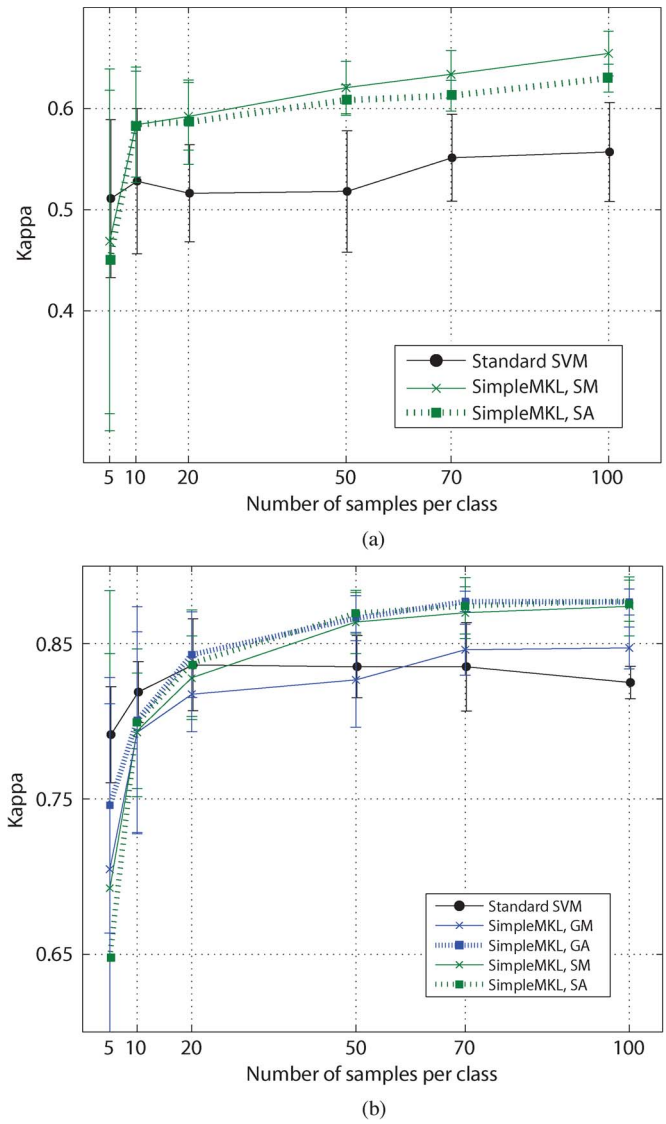


Fig. 6. Kappa statistics for the images image of Naples (Italy). (a) Using only Landsat bands ("L" series). (b) Using both Landsat and radar features ("LR" series).

selects principally the blue and NIR bands, that are valuable for the discrimination of water and built environment. The other bands are the most often ignored in the final solution. In the multisource experiments, the importance of the radar features comes out clearly either in the SA setting [Fig. 7(b)] or when considering a dedicated kernel to each source of information [Fig. 7(c) and (d)]. The radar features take a greater importance than the optical ones, showing the stronger impact they do have on the final solution. In the SA experiment [Fig. 7(b)], we can observe the importance of the third and fourth radar features, which are the original and contextually filtered coherence, while in the GM experiment, we can appreciate the role of a multiscale solution, retaining the radar kernel with small- and medium-scale σ.

D. On Computational and Selection Efficiency

Throughout the experiments, the capability of SimpleMKL to rank features (or meaningful groups of features) was studied and the features selected were interpreted in terms of image properties (see Figs. 3, 5 and 7 and respective discussion for details). However, it is also important to compare the features obtaining maximal rank with the features selected by other feature selection methods. In this section, we first compare SimpleMKL with a filter based on the correlation between the single features and the output class label, and then we compare it with the more advanced SVM-RFE [24] method. The issues of both the selection of useful features and the computational cost are reported below.

Fig. 8 shows the feature ranking for the three methods in the HyMap image. The correlation filter [Fig. 8(a)] correctly selects the regions of the spectral around features 99 and 118, but generally misses the relevant features around feature channel 20 (cf. Table I); moreover, the filter has the tendency to select

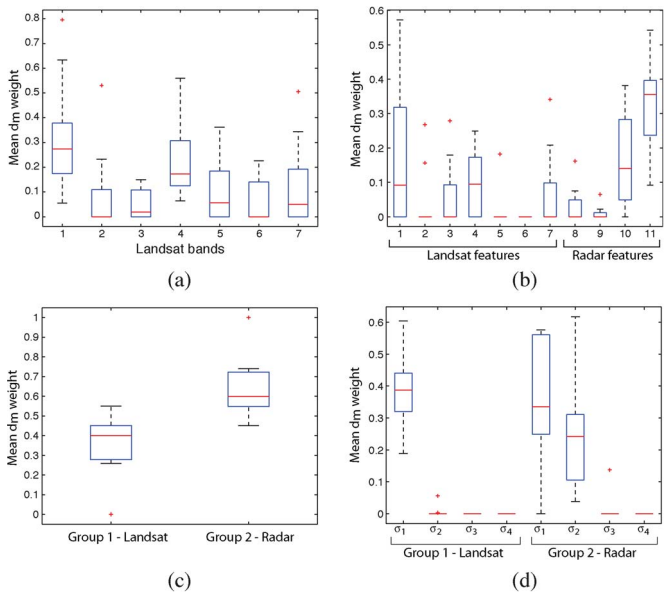


Fig. 7. Box plots of the feature weights d_m for some of the experiments considered in the multisource setting: (a) SA-L; (b) SA-LR; (c) GA-LR; (d) GM-LR. Each box represents distribution of values for ten experiments using 20 pixels per class.

features strongly correlated with the output, but also with each other. SVM-RFE [Fig. 8(b)] results in a more desirable solution, where all the features of Table I are selected by most of the runs of the model. Similarly, SimpleMKL selects sets of features corresponding to the meaningful features for solving the problem [Fig. 8(c)]. The difference of representation between the two previous methods and MKL is explained by the different rankings provided. Correlation-based filter and SVM-RFE provide a full ranking, where a single feature is excluded at each iteration; on the contrary, MKL optimizes a vector of weights and the selection is done implicitly by the exclusion of the kernels receiving a zero weight (in white in the figure).

Above, we have proven that MKL selects the relevant features of the image. However, to prove the competitiveness of the method with respect to SVM-RFE, the study of the computational load is compulsory, since the optimization of the combination of $(s \cdot l)$ kernels is needed, the cost of the (efficient) SimpleMKL becomes a crucial issue. The Hymap image, which is the largest one studied in this paper, is considered in the experiments reported in Fig. 9. Compared to the standard SVM (and thus to the correlation filter), both methods require a larger computational effort, but MKL is more efficient than SVM-RFE: particularly for mid-sized data sets (20 pixels per class), and the difference in speed is significant. When using 50 pixels per class, the SM experiment (that optimizes 512 kernel weights) results into the highest computational load, showing that such a strategy of model selection is not desirable for high-dimensional problems and supports even further the kernel alignment model selection strategy proposed in this paper.

VI. Conclusion

In this paper, we presented a kernel framework for combining and assessing the relevance of different sources of information in multiple-kernel SVM image classification. The methodology

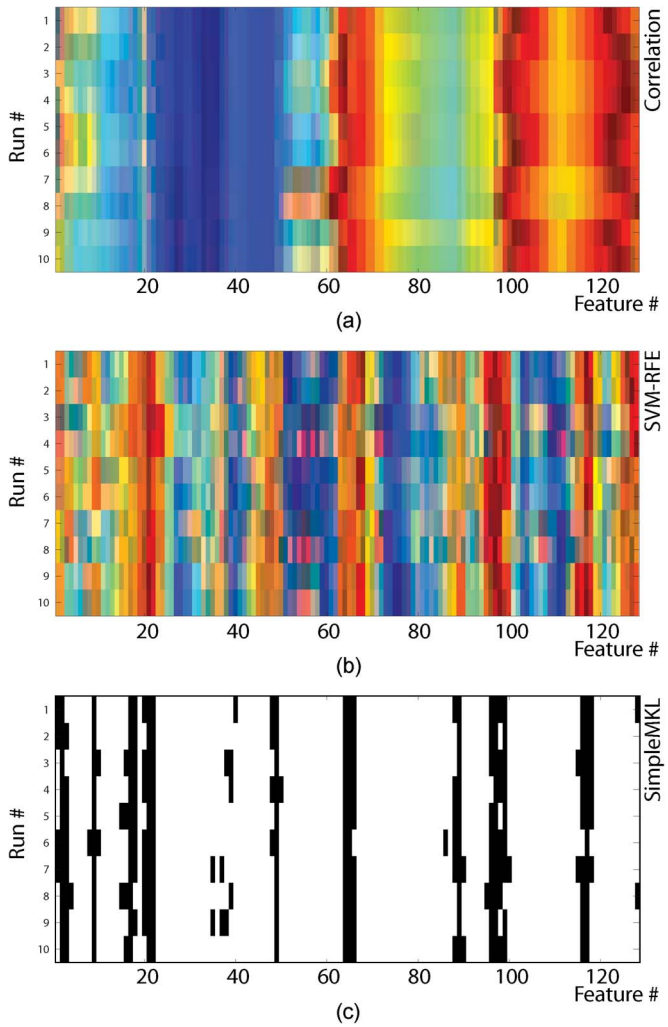


Fig. 8. Comparison of feature selection/ranking methods for ten experiments using 50 pixels per class (HyMap image): (a) correlation filter; (b) SVM-RFE; (c) the proposed MKL. In subfigures (a) and (b) position in the ranking is reported for each feature, going from blue (least important) to red (most important). In (c), the features receiving a non-zero weight after MKL optimization are reported in black.

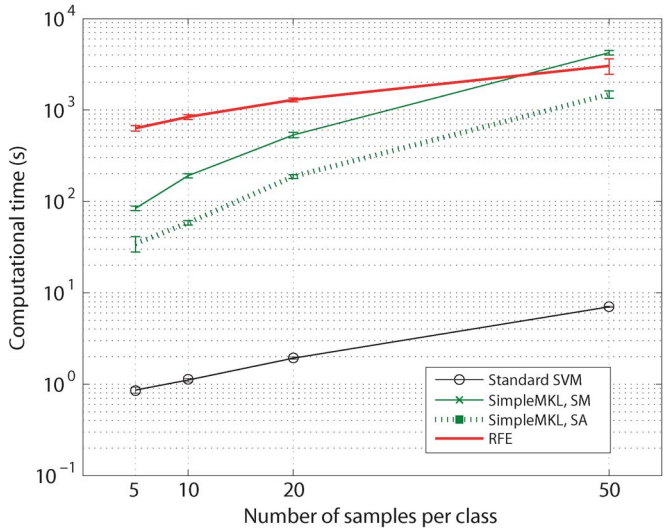


Fig. 9. Computational load of SVM, MKL, and RFE for data sets of varying size (HyMap image).

yields two main advantages: 1) it allows to automatically optimize different kernel parameters and weights per dedicated kernel; and 2) particularly when using kernel alignment for free parameters estimation, the computational cost is affordable for large-scale remote sensing applications. Besides, and more importantly, it is the fact that a single SVM is obtained for testing and a direct feature ranking along it. We have illustrated the good performance of the method in image classification, when multispectral, hyperspectral, contextual or multisensor information is used. The obtained models are competitive with respect to the standard SVM in terms of classification accuracy. Furthermore, we validated the correctness of the resulting feature ranking by comparing it to the outcome of classical feature selection algorithms. Also, we have highlighted the interest of the SimpleMKL technique for the remote sensing community by demonstrating that the opportunity to analyze the obtained feature weights and scales under physical criteria provides understandable and interpretable models for image classification.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York: Springer-Verlag, 2001, ser. Series in Statistics.

[2] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[3] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.

[4] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[5] G. Camps-Valls and L. Bruzzone, Eds., *Kernel Methods in Remote Sensing Image Processing*. Hoboken, NJ: Wiley, 2009.

[6] G. F. Hughes, "On the mean accuracy of statistical pattern recognition," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

[7] L. Bruzzone, R. Cossu, and G. Vernazza, "Detection of land-cover transitions by combining multidate classifiers," *Pattern Recognit. Lett.*, vol. 25, no. 13, pp. 1491–1500, Oct. 2004.

[8] B. Waske and J. A. Benediktsson, "Fusion of support vector machines for classification of multisensor data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3858–3866, Dec. 2007.

[9] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. L. Rojo-Álvarez, and M. Martínez-Ramón, "Kernel-based framework for multi-temporal and multi-source remote sensing data classification and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1822–1835, Jun. 2008.

[10] L. Bruzzone and S. B. Serpico, "An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 35, no. 4, pp. 858–867, Jul. 1997.

[11] L. Bruzzone and D. Fernández Prieto, "A partially unsupervised cascade classifier for the analysis of multitemporal remote-sensing images," *Pattern Recognit. Lett.*, vol. 23, no. 9, pp. 1063–1071, Jul. 2002.

[12] D. Clausi, "Comparison and fusion of co-occurrence, Gabor, and MRF texture features for classification of SAR sea ice imagery," *Atmos. Oceans*, vol. 39, no. 4, pp. 183–194, 2001.

[13] J. A. Benediktsson, M. Pesaresi, and K. Arnason, "Classification and feature extraction for remote sensing images from urban areas based on morphological transformations," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 9, pp. 1940–1949, Sep. 2003.

[14] A. Plaza, P. Martínez, J. Plaza, and R. Pérez, "Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 466–479, Mar. 2005.

[15] F. Pacifici, M. Chini, and W. Emery, "A neural network approach using multi-scale textural metrics from very high resolution panchromatic imagery for urban land-use classification," *Remote Sens. Environ.*, vol. 113, no. 6, pp. 1276–1292, Jun. 2009.

[16] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications*. New York: Springer-Verlag, 2006.

[17] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, no. 7/8, pp. 1157–1182, Mar. 2003.

[18] T. A. Warner, K. Steinmaus, and H. Foote, "An evaluation of spatial autocorrelation feature selection," *Int. J. Remote Sens.*, vol. 20, no. 8, pp. 1601–1616, May 1999.

[19] J. S. Borak, "Feature selection and land cover classification of a MODIS-like data set for a semiarid environments," *Int. J. Remote Sens.*, vol. 20, no. 5, pp. 919–938, Mar. 1999.

[20] L. Bruzzone and S. Serpico, "A technique for features selection in multiclass problems," *Int. J. Remote Sens.*, vol. 21, no. 3, pp. 549–563, Feb. 2000.

[21] T. Kavzoglu and P. M. Mather, "The role of feature selection in artificial neural network applications," *Int. J. Remote Sens.*, vol. 23, no. 15, pp. 2919–2937, Aug. 2002.

[22] B. Demir and S. Ertürk, "Phase correlation based redundancy removal in feature weighting band selection for hyperspectral images," *Int. J. Remote Sens.*, vol. 29, no. 6, pp. 1801–1807, Mar. 2008.

[23] B.-C. Kuo, C.-H. Li, and J.-M. Yang, "Kernel nonparametric weighted feature extraction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp. 1139–1155, Apr. 2009.

[24] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, Jan. 2002.

[25] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero-norm with linear models and kernel methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1439–1461, Mar. 2003.

[26] D. Tuia, F. Pacifici, M. Kanevski, and W. Emery, "Classification of very high spatial resolution imagery using mathematical morphology and support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3866–3879, Nov. 2009.

[27] R. Archibald and G. Fann, "Feature selection and classification of hyperspectral images with support vector machines," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 4, pp. 674–677, Oct. 2007.

[28] Y. Bazi and F. Melgani, "Toward an optimal SVM classification system for hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3374–3385, Nov. 2006.

[29] S. S. Durbha, R. L. King, and N. H. Younan, "Wrapper-based feature subset selection for rapid image information mining," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 43–47, Jan. 2010.

[30] L. Zhang, Y. Zhong, B. Huang, and P. Li, "Dimensionality reduction based on clonal selection for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4172–4186, Dec. 2007.

[31] P. Zhong, P. Zhang, and R. Wang, "Dynamic learning of SMLR for feature selection and classification of hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 2, pp. 280–284, Apr. 2008.

[32] G. Lanckriet, T. De Bie, N. Cristianini, M. Jordan, and W. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol. 20, no. 16, pp. 2626–2635, Nov. 2004.

[33] S. Sonnenburg, C. S. G. Rätsch, and B. Schölkopf, "Large scale multiple kernel learning," *J. Mach. Learn. Res.*, vol. 7, pp. 1531–1565, 2006.

[34] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[35] M. Marconcini, G. Camps-Valls, and L. Bruzzone, "A composite semisupervised SVM for classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 234–238, Apr. 2009.

[36] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "A joint spatial and spectral SVM's classification of panchromatic images," in *Proc. IEEE IGARSS*, Barcelona, Spain, Jul. 2007, pp. 1497–1500.

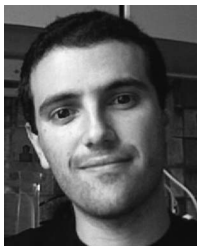
[37] D. Tuia, F. Ratle, A. Pozdnoukhov, and G. Camps-Valls, "Multi-source composite kernels for urban image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 88–92, Jan. 2010.

[38] A. Villa, M. Fauvel, J. Chanussot, P. Gamba, and J. A. Benediktsson, "Gradient optimization for multiple kernel parameters in support vector machines classification," in *Proc. IEEE IGARSS*, 2008, pp. IV-224–IV-227.

[39] B. Guo, S. Gunn, R. I. Damper, and J. D. B. Nelson, "Customizing kernel functions for SVM-based hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 622–629, Apr. 2008.

[40] F. Bach, G. Lanckriet, and M. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. Int. Conf. Mach. Learn.*, 2004, pp. 41–48.

[41] S. V. N. Vishwanathan, A. J. Smola, and M. Murty, "Simple SVM," in *Proc. Int. Conf. Mach. Learn.*, 2003, vol. 13, pp. 682–688.
[42] A. Zien and C. S. Ong, "Multiclass multiple kernel learning," in *Proc. ICML*, Corvallis, OR, 2007, vol. 227, pp. 1191–1198.
[43] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "Simple MKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, 2008.
[44] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor, "On kernel target alignment," Roy. Holloway College, Univ. London, London, U.K., NeuroCOLT, Tech. Rep. 2001-087, 2001.
[45] C. Igel, T. Glasmachers, B. Mersch, N. Pfeifer, and P. Meinicke, "Gradient-based optimization of kernel-target alignment for sequence kernels applied to bacterial gene start detection," *IEEE Trans. Comput. Biol. Bioinformatics*, vol. 4, no. 2, pp. 216–226, Apr.–Jun. 2007.
[46] A. Kumar and C. Sminchisescu, "Support kernel machines for object recognition," in *Proc. IEEE ICCV*, 2007, pp. 1–8.
[47] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA: MIT Press, 2002.
[48] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[49] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th ACM Workshop Comput. Learn. Theory*, Pittsburgh, PA, 1992, pp. 144–152.
[50] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with application in pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-14, no. 3, pp. 326–334, Jun. 1965, reprinted in Artificial Neural Networks: Concepts and Theory, IEEE Computer Society Press, Los Alamitos, California, 1992, Eds. P. Mehra and B. Wah.
[51] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Philos. Trans. R. Soc. Lond. Ser. A*, vol. CCIX, no. A456, pp. 215–228, May 1905.
[52] G. Camps-Valls, L. Gómez-Chova, J. Calpe, and E. Soria, "Robust support vector method for hyperspectral data classification and knowledge discovery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1530–1542, Jul. 2004.
[53] A. Fanelli, M. Santoro, A. Vitale, P. Murino, and J. Askne, "Understanding ERS coherence over urban areas," in *Proc. ERS-Envisat Symp.—Looking Down to Earth in the New Millennium*, Gothenburg, Sweden, Oct. 16–20, 2000. [CD-ROM].
[54] L. Gómez-Chova, D. Fernández-Prieto, J. Calpe, E. Soria, J. Vila, and G. Camps-Valls, "Urban monitoring using multitemporal SAR and multispectral data," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 234–243, Mar. 2006.
[55] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 5, pp. 627–633, May 2004.



Devis Tuia (SM'07–M'09) was born in Mendrisio, Switzerland, in 1980. He received the diploma degree in geography at the University of Lausanne, Lausanne, in 2004, the Master degree of Advanced Studies in environmental engineering at the Federal Institute of Technology of Lausanne (EPFL), in 2005 and the Ph.D. degree in environmental sciences at the University of Lausanne, in 2009.

He is currently a postdoc researcher at both the Univeristy of València, València, Spain, and the Univeristy of Colorado at Boulder, under a Swiss National Foundation program. His research interests include the development of algorithms for feature selection and classification of very high-resolution remote sensing images and socio-economic data using kernel methods. Particularly, his studies focused on the use of unlabeled samples and on the interaction between the user and the machine through active and semisupervised learning.

Dr. Tuia was one of the winners of the IEEE GEOSCIENCE AND REMOTE SENSING Data Fusion Contest, in 2008. In 2009, he ranked second place at the student paper competition of the Joint Urban Remote Sensing Event (JURSE) and third place at the student competition of the Geoscience and Remote Sensing Symposium (IGARSS).

Gustavo Camps-Valls (M'04–SM'07) was born in València, Spain, in 1972, and received the B.Sc. degree in physics, the B.Sc. degree in electronics engineering, and the Ph.D. degree in physics from the Universitat de València, in 1996, 1998, and 2002, respectively.

He is currently an associate professor in the Department of Electronics Engineering at the Universitat de València, where he teaches electronics, advanced time series processing, and machine learning for remote sensing. His research interests are tied to the development of machine learning algorithms for signal and image processing with special focus on remote sensing data analysis. He conducts and supervises research within the frameworks of several national and international projects, and he is an Evaluator of project proposals and scientific organizations. He is the author (or coauthor) of 70 international journal papers, more than 80 international conference papers, several international book chapters, and editor of the books *Kernel Methods in Bioengineering, Signal and Image Processing* (IGI, 2007) and *Kernel Methods for Remote Sensing Data Analysis* (Wiley, 2009). He is a referee of many international journals and conferences, and currently serves on the Program Committees of several International Conferences.

Dr. Camps-Valls is member of the Data Fusion technical committee of the IEEE Geoscience and Remote Sensing Society, since 2007, and since 2009 he is member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society.

Giona Matasci was born in Locarno, Switzerland, in 1985. He received the M.Sc. degree in environmental sciences from the University of Lausanne, Lausanne, in 2009. He is currently working toward the Ph.D. degree at IGAR (University of Lausanne) in the field of machine learning and its applications to remote sensing and environmental data analysis.

Mikhail Kanevski received the Ph.D. degree in plasma physics from the Moscow State University, Moscow, Russia, in 1984 and Doctoral theses in computer science form the Institute of Nuclear Safety (IBRAE) of Russian Academy of Science, in 1996.

Until 2000, he was a Professor at Moscow Physico-Technical Institute (Technical University) and head of laboratory at the Moscow Institute of Nuclear Safety, Russian Academy of Sciences. Since 2004, he is a professor at the Institute of Geomatics and Analysis of Risk (IGAR) of the University of Lausanne, Lausanne, Switzerland. He is a principal investigator of several national and international grants. His research interests include geostatistics for spatio-temporal data analysis, environmental modeling, computer science, numerical simulations, and machine learning algorithms. Remote sensing image classification, natural hazards assessments (forest fires, avalanches, landslides) and time series predictions are the main applications considered at his laboratory.