# Supervised Hyperspectral Image Classification With Rejection

Filipe Condessa, *Student Member, IEEE*, José Bioucas-Dias, *Senior Member, IEEE*,
and Jelena Kovačević, *Fellow, IEEE*

*Abstract*—**Hyperspectral image classification is a challenging problem as obtaining complete and representative training sets is costly, pixels can belong to unknown classes, and it is generally an ill-posed problem. The need to achieve high classification accuracy may surpass the need to classify the entire image. To account for this scenario, we use classification with rejection by providing the classifier with an option not to classify a pixel and consequently reject it. We present and analyze two approaches for supervised hyperspectral image classification that combine the use of contextual priors with classification with rejection: 1) by jointly computing context and rejection and 2) by sequentially computing context and rejection. In the joint approach, rejection is introduced as an extra class that models the probability of classifier failure. In the sequential approach, rejection results from the hidden field associated with a marginal maximum *a posteriori* classification of the image. We validate both approaches on real hyperspectral data.**

*Index Terms*—**Classification with context, classification with rejection, hyperspectral image classification.**

## I. Introduction

**S**UPERVISED image classification is pivotal in a large number of hyperspectral image applications [3]. The problem of supervised hyperspectral image classification is generally ill-posed. Contextual information is used in image classification as a regularizer to impose desired characteristics in the resulting classification, e.g., through the use of multi-level logistic priors based on Markov random fields [4], widely used in hyperspectral image classification [5], or graph-based methods [6], [7]. Whereas there are alternatives to supervised hyperspectral image classification, such as curve fitting of absorption bands [8], the need for contextual information-based regularization is still present. By itself, however, contextual information does not totally remove the effects of classification errors associated with overlapping classes, small or incomplete training sets, and the existence of unknown classes.

Classification errors can be mitigated if we adapt the behavior of the classifier to avoid classifying samples (pixels in the case of images) with high potential for incorrect classifications. This can be achieved by equipping the classifier with rejection, thus obtaining an *increase in classification performance at the expense of not classifying the entire image*.

Classification with rejection was first analyzed in [9], where Chow's rule for optimum error-reject trade-off was presented. Given the knowledge of posterior probabilities and of the costs of erring and rejecting, Chow's rule allows the derivation of a rejection rule based on the thresholding of probabilities such that empirical classification risk is minimized. Extensive work exists on the design of systems for classification with rejection (see [10] and references therein); however, the application of pixelwise classification with rejection to images has been limited to medical image classification [11], [12].

In hyperspectral images, the acquisition of representative, nonoverlapping, and balanced pixelwise training sets is costly, pixels can belong to unknown classes, and the need for high accuracy may surpass the need to classify the entire image. These characteristics are shared among a class of image classification problems, e.g., the task of histopathology image classification [11], [12], where the combination of classification with context and classification with rejection has shown improved classification performances. We thus hypothesize that applying rejection to classification can be fruitful in hyperspectral image classification problems as well.

Classification with rejection can be conceptualized as a coupling of a classifier (that maps feature vectors into class labels) with a rejector (that maps class labels into a binary decision to reject or not). There is an interplay between the performance of the classifier and the required performance of the rejector: the higher the accuracy obtained by the classifier, the harder it becomes for the rejector not to reject correctly classified pixels. This means that performance improvements of combining rejection and context are clearer when the performance of the classifier is lower. We will show that by using classification with rejection systems (as schematized in Fig. 1), we are able, with small training sets, to achieve classification performances close to those obtained with larger training sets.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

2                    IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING
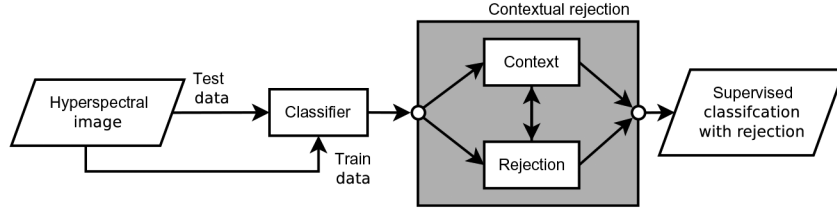


Fig. 1. General diagram of supervised hyperspectral image classification with rejection. The classification block corresponds to a supervised classifier trained with labeled training pixels and applied to unlabeled test pixels. The contextual rejection block combines the classification with rejection with the classification context. In Section III, two instantiations of contextual rejection are discussed.

In this paper, we combine classification with rejection with classification with context in two different ways, corresponding to two different instantiations of the general scheme in Fig. 1:

1) joint computation of context and rejection (JCR) as in [1], where rejection is considered as an extra class and computed alongside with context;

2) sequential computation of context and rejection (SCR) as in [2], where rejection is computed after the context by use of a rejection field.

We extend and compare these two different formulations for supervised hyperspectral image classification with rejection.

The contribution of this paper is two fold: 1) the application of classification with rejection to the hyperspectral image classification problem and 2) the development of algorithms for contextual classification with rejection for hyperspectral image classification.

The paper is organized as follows. Section II provides the background on the contextual classification techniques and performance measures for classification with rejection. Section III describes our classification method with rejection and context, with Section III-A corresponding to JCR and Section III-B to SCR. Section IV presents experimental results and Section V concludes the paper.

## II. BACKGROUND

We now introduce the necessary notation and background for the computation of classification with context by presenting the SegSALSA algorithm and for the evaluation of performance of classifiers with rejection.

Let $\mathbf{x} \in \mathbb{R}^{d \times n}$ denote an $n$-pixel hyperspectral image with $d$ spectral bands, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the vector of $d$ spectral values at pixel corresponding to the image pixel $i$. Let $\mathcal{S} = \{1, \ldots, n\}$ be the set that indexes the image pixels, $\mathcal{L} = \{1, \ldots, K\}$ be the set of possible $K$ labels, and $\mathbf{y} \in \mathcal{L}^n$ be a labeling of the image.

### A. Classification With Context—SegSALSA

The goal behind classification with context is to combine pixelwise classification results with a contextual (often spatial) prior. Desired properties, such as piecewise smooth labelings, are imposed on the labelings through the use of contextual priors. Classification with context is achieved by the SegSALSA algorithm [13], [14] that combines the idea of a hidden field driving a segmentation [15], with a vectorial total variation

prior [16], [17] in a convex segmentation formulation solved by the SALSA algorithm [18].

Adopting a Bayesian perspective, the maximum *a posteriori* (MAP) labeling $\widehat{\mathbf{y}}$ is given by

$$\widehat{\mathbf{y}} = \arg \max_{\mathbf{y} \in \mathcal{L}^n} p(\mathbf{y}|\mathbf{x}) = \arg \max_{\mathbf{y} \in \mathcal{L}^n} p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \qquad (1)$$

where $p(\mathbf{y}|\mathbf{x})$ denotes the posterior probability of the labeling $\mathbf{y}$ given the feature vectors $\mathbf{x}$, $p(\mathbf{x}|\mathbf{y})$ denotes the observation model, and $p(\mathbf{y})$ denotes the prior probability of the labeling. Assuming conditional independence of the features given the labels, we have

$$p(\mathbf{x}|\mathbf{y}) = \prod_{k \in \mathcal{L}} \prod_{i \in \mathcal{S}} p(\mathbf{x}_i|y_i = k).$$

To introduce the hidden field [15], let $\mathbf{z}$ be a $K \times n$ matrix containing a collection of hidden random vectors $\mathbf{z}_i \in \mathbb{R}^K$, for $i \in \mathcal{S}$. The joint probability of labels $\mathbf{y}$ and field $\mathbf{z}$ is defined as

$$p(\mathbf{y}, \mathbf{z}) = p(\mathbf{y}|\mathbf{z})p(\mathbf{z})$$

with the assumption of conditional independence of the labels $\mathbf{y}$ given the field $\mathbf{z}$

$$p(\mathbf{y}|\mathbf{z}) = \prod_{i \in \mathcal{S}} p(y_i|\mathbf{z}_i).$$

This allows us to express the joint probability of the features, labels and fields $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ as

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}|\mathbf{z})p(\mathbf{z}).$$

Armed with the hidden field $\mathbf{z}$ and the joint probabilities, we can now marginalize on the discrete labels

$$p(\mathbf{x}, \mathbf{z}) = \prod_{i \in \mathcal{S}} \left\{ \overbrace{\sum_{y_i \in \mathcal{L}} p(\mathbf{x}_i|y_i)p(y_i|\mathbf{z}_i)}^{\text{marginalization}} \right\} p(\mathbf{z}). \qquad (2)$$

The marginalization in (2) corresponds to the marginalization of the joint probability of the features $\mathbf{x}$, the labels $\mathbf{y}$, and the hidden field $\mathbf{z}$ across the discrete labels $\mathbf{y}$. This allows us to have the joint the probabilities of the features and the hidden field $p(\mathbf{x}, \mathbf{z})$, depending only on the continuous hidden field $\mathbf{z}$, as the features $\mathbf{x}$ are known. The marginal maximum *a posteriori* (MMAP) is then

$$\widehat{\mathbf{z}}_{\text{MMAP}} = \arg \max_{\mathbf{z} \in \mathbb{R}^{K \times n}} p(\mathbf{x}, \mathbf{z})$$

with the soft classification obtained by $p(\mathbf{y}|\widehat{\mathbf{z}}_{\text{MMAP}})$ and the labeling obtained by finding the labeling $\mathbf{y}$ that maximizes the soft classification.

As the $k$th component of the $i$th random vector $[\mathbf{z}_i]_k$ is modeled by the conditional probability $p(y_i = k|\mathbf{z}_i)$, two constraints are introduced in the hidden field $\mathbf{z}$ as a result: nonnegativity $[\mathbf{z}_i]_k \geq 0$ and sum-to-one $\mathbf{1}_K^T \mathbf{z}_i = 1$.

The conditional probabilities $p(y_i|\mathbf{x}_i)$, collected in the vector $\mathbf{p}_i = [p(y_i = 1|\mathbf{x}_i), \ldots, p(y_i = K|\mathbf{x}_i)]$, are modeled with a sparse multinomial logistic (MLR) with the Logistic Regression via Split Augmented Lagrangian (LORSAL) algorithm [19] as follows. Let $k(\mathbf{x}_i)$ be a kernel function computed between the spectra of the $i$th pixel $x_i$ and the spectra of the pixels belonging to the training set, we have that the *a posterior* probabilities can be modeled as

$$p(y_i = \ell|\mathbf{x}_i, [\mathbf{w}_1, \ldots, \mathbf{w}_K]) = \frac{e^{\mathbf{w}_\ell^T k(\mathbf{x}_i)}}{\sum_{j=1}^K e^{\mathbf{w}_j^T k(\mathbf{x}_i)}}$$

where $\mathbf{w}_j$ is the regression vector for the $j$th class. We learn the regression vectors using the LORSAL algorithm, with an element-wise independent Laplacian prior for the regression vectors, and computing the maximum *a posteriori* estimate of $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$ by solving the following decoupled optimization problem:

$$\arg\max_{\mathbf{W},\mathbf{\Omega}} l(\mathbf{W}) + \log p(\mathbf{\Omega}), \text{ subject to } \mathbf{W} = \mathbf{\Omega}$$

where $l(\mathbf{W})$ demotes the log-likelihood of $\mathbf{W}$ and $p(\mathbf{\Omega}) \propto e^{-\lambda\|\mathbf{\Omega}\|_1}$ promotes the sparsity of the regression vectors.

As we deal now with the MMAP instead of the MAP, the prior is no longer applied on the discrete labels $\mathbf{y}$ but on the continuous hidden field $\mathbf{z}$. We adopt a vectorial total variation (VTV) prior [16], [17] for the hidden field $\mathbf{z}$ as it promotes piecewise smoothness of the field, preservation and alignment of the discontinuities across the classes, and it is convex. The prior $p(\mathbf{z})$ is defined such that

$$-\ln p(\mathbf{z}) \equiv \lambda_{\text{TV}} \sum_{i \in \mathcal{S}} \sqrt{\|D_h \mathbf{z}_i\|^2 + \|D_v \mathbf{z}_i\|^2} \quad (3)$$

where $D_h$ is the horizontal difference operator, $D_v$ the vertical difference operator, and $\lambda_{\text{TV}}$ a regularization parameter. The regularization parameter $\lambda_{\text{TV}}$ controls the relative weight of the vectorial total variation prior compared to the connection of the hidden field to the class probabilities, thus the value of $\lambda_{\text{TV}}$ affects the piecewise smoothness of classification. A larger value of $\lambda_{\text{TV}}$ results in smoother segmentations whereas a smaller value of $\lambda_{\text{TV}}$ results in segmentations with speckles.

From the initial integer optimization problem in (1), the contextual classification problem is then formulated as a convex optimization problem

$$\widehat{\mathbf{z}}_{MMAP} = \arg\min_{\mathbf{z}\in\mathbb{R}^{K\times n}} -\sum_{i\in\mathcal{S}} \left[\ln\left(\mathbf{p}_i^T \mathbf{z}_i\right)\right] - \ln p(\mathbf{z})$$

$$\text{subject to: } \mathbf{z} \geq 0, \quad \mathbf{1}_K^T \mathbf{z} = \mathbf{1}_n^T. \quad (4)$$
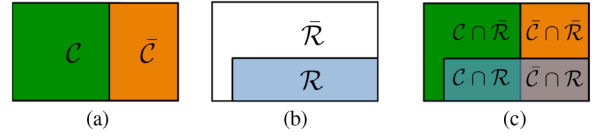
Fig. 2. Example of classification with rejection of image. Classification (a) with correctly classified pixels in green and incorrectly classified pixels in orange. Rejection (b) with rejected pixels in gray and nonrejected pixels in white. Classification with rejection (c) with intersection of rejected/nonrejected and correctly/incorrectly classified pixels.

This problem can be solved efficiently with SALSA (constraint split augmented Lagrangian shrinkage algorithm) [18], an instance of the alternating direction method of multipliers. The SALSA algorithm allows us to solve a convex optimization problem with an arbitrary number of terms, such as (4), using a flexible variable splitting mechanism without the incurring on the computational costs associated with double loops present in the Douglas–Rachford Splitting algorithm in [20]. The problem (4) can then be solved with a complexity of $O(Kn \log n)$. We point the interested reader to [13] for the formulation of SegSALSA as an instance of the SALSA algorithm.

### B. Performance Measures for Classification With Rejection

The performance of classification with rejection is frequently assessed by comparing the accuracy of the subset of nonrejected samples (in our case pixels), the *nonrejected accuracy A*, with the the *fraction of rejected pixels r* (as in [21]–[24]). Let $\mathcal{R}$ be the set of rejected pixels ($\bar{\mathcal{R}}$ denotes the set of nonrejected pixels) and $\mathcal{C}$ the set of correctly classified pixels ($\bar{\mathcal{C}}$ denotes the set of incorrectly classified pixels) with $\mathcal{S} = \mathcal{R} \cup \bar{\mathcal{R}} = \mathcal{C} \cup \bar{\mathcal{C}}$, as illustrated in Fig. 2.

We represent the rejected fraction $r$ as

$$r = \frac{|\mathcal{R}|}{|\mathcal{R} \cup \bar{\mathcal{R}}|} = \frac{|\mathcal{R}|}{|\mathcal{S}|} = \quad (5)$$

corresponding to the fraction of pixels that are rejected. The nonrejected accuracy $A$ can be represented as

$$A = \frac{|\mathcal{C} \cap \bar{\mathcal{R}}|}{|\bar{\mathcal{R}}|} = $$

with

$$A = \frac{|\mathcal{C} \cap \bar{\mathcal{R}}|}{|\mathcal{S}|} \frac{1}{1-r} = A(r) \quad (6)$$

corresponding to the fraction of nonrejected pixels that are correctly classified. We note that $A(0)$ corresponds to the total accuracy of the classifier, with no rejection in place.

The nonrejected accuracy, combined with the rejected fraction, is unable to compare directly the behavior of two classifiers with rejection working at different rejected fractions. A clear example of this inability is the following. Let us consider three cases with a classifier that classifies the same set of pixels as follows:

|  | Case 1 | Case 2 | Case 3 |
|---|---|---|---|
| Nonrejected fraction $1 - r$ | 79.00% | 80.00% | 81.01% |
| Nonrejected accuracy $A(r)$ | 81.01% | 80.00% | 79.00% |

In all three cases, 64% of the pixels are correctly classified; however, the pair rejected fraction/nonrejected accuracy will not be able to compare the three cases working at different rejected fractions.

To account for this, we extend the concept of nonrejected accuracy to the *classification quality Q* [25].

*Definition 1:* Classification quality: given a classifier defined by the pair correctly classified and incorrectly classified pixels $(\mathcal{C}, \bar{\mathcal{C}})$ and a rejector defined by pair rejected and not rejected pixels $(\mathcal{R}, \bar{\mathcal{R}})$, the classification quality $Q$ measures the proportion of pixels that are *either* correctly classified and not rejected *or* incorrectly classified and rejected, relative to the total number of pixels

$$Q = \frac{|\mathcal{C} \cap \bar{\mathcal{R}}| + |\bar{\mathcal{C}} \cap \mathcal{R}|}{|\mathcal{R} \cup \bar{\mathcal{R}}|} = \frac{|\mathcal{C} \cap \bar{\mathcal{R}}|}{|\mathcal{S}|} + \frac{|\bar{\mathcal{C}} \cap \mathcal{R}|}{|\mathcal{S}|} = \blacksquare . \tag{7}$$

The classification quality combines the performance of the classifier on the subset of nonrejected pixels with the performance of the rejector on the subset of misclassified pixels. The maximum value of classification quality is 100% and it is achieved when $\bar{\mathcal{R}} = \mathcal{C}$ and $\mathcal{R} = \bar{\mathcal{C}}$, i.e., if all the correctly classified samples are not rejected and if all the incorrectly classified samples are rejected, corresponding to a perfect rejector that achieves a nonrejected accuracy of 100% with the minimum number of pixels rejected. When the opposite occurs, the classification quality achieves its minimum value of 0%.

In (7), the classification performance on the subset of nonrejected pixels can be obtained as

$$\frac{|\mathcal{C} \cap \bar{\mathcal{R}}|}{|\mathcal{S}|} = A(r)(1 - r) \tag{8}$$

and the rejector performance on the subset of misclassified pixels obtained as

$$\frac{|\bar{\mathcal{C}} \cap \mathcal{R}|}{|\mathcal{S}|} = (1 - A(0)) - \frac{|\bar{\mathcal{C}} \cap \bar{\mathcal{R}}|}{|\mathcal{S}|}$$
$$\frac{|\bar{\mathcal{C}} \cap \mathcal{R}|}{|\mathcal{S}|} = \frac{|\mathcal{R}|}{|\mathcal{S}|} - \frac{|\mathcal{C} \cap \bar{\mathcal{R}}|}{|\mathcal{S}|} = (1 - r) - A(r)(1 - r)$$
$$\frac{|\bar{\mathcal{C}} \cap \mathcal{R}|}{|\mathcal{S}|} = -A(0) + r + A(r)(1 - r). \tag{9}$$

By combining (8) and (9), we are able to represent the classification quality as

$$Q(r) = 2A(r)(1 - r) + r - A(0) \tag{10}$$

with $Q(0) = A(0)$.

The value of $Q$ amounts to the proportion of correct decisions that the ensemble classifier and rejector performs. This means that a classifier with rejection that rejects a fraction $r$ of pixels with a value of classification quality of $Q$ is equivalent, *in terms of correct decisions made*, to a classifier with no

rejection $r = 0$ and accuracy numerically equal to $A(0) = Q$. The classification quality allows us to directly compare the performance of classification systems with rejection working at different rejected fractions.

We denote $Q(r) + A(0)$ as classification quality with offset, allowing us to compare classification qualities of different operating points of a classifier with rejection when the total accuracy without rejection is unknown but equal across different operating points. This is the situation we have in the three cases presented.

Armed with the classification quality, we can now compare the performance of the three aforementioned cases of a classifier with rejection. Under the assumption of unknown and equal total accuracy with no rejection $A(0)$, we can obtain the following values of classification quality with offset.

|  | Case 1 | Case 2 | Case 3 |
|---|---|---|---|
| Nonrejected fraction $1 - r$ | 79.00% | 80.00% | 81.01% |
| Nonrejected accuracy $A(r)$ | 81.01% | 80.00% | 79.00% |
| Classification quality with offset $Q(r) + A(0)$ | 149.00% | 148.00% | 146.99% |

This example shows that the operating point for case 1, where the 79% of the data are classified with 81% of nonrejected accuracy, achieves a higher number of correct joint decisions, independently of the total accuracy without rejection, than the other two cases. Thus, the use of classification quality allows a clear discrimination of the performance of the classification system with rejection at different operating points.

## III. REJECTION AND CONTEXT

With the background for the classification with context established, we now approach the problem of classification with rejection applied to our supervised classification problem.

Classification with rejection can be achieved based on the existence of two simple mechanisms as follows:

1) an implicit ordering of the pixels according to their potential to be rejected;
2) a concept of a threshold that controls the amount of pixels that are rejected.

This can be easily achieved by considering an extension of Chow's rule for two class classification with rejection, i.e., the derivation of a probability threshold for a binary classification problem that minimizes the empirical risk given a cost matrix and the posterior probabilities [9]. Let us consider an image with $K$ nonrejection classes, and a $K + 1$ class that corresponds to rejection. The pixelwise MAP classification of the $i$th pixel is

$$\widehat{y_i} = \arg\max_{y_i \in \mathcal{L} \cup \{K+1\}} p(y_i | \mathbf{x}_i) \tag{11}$$

where $p(y_i = K + 1 | \mathbf{x}_i) = \gamma$ represents the probability of rejection. The maximum probability of the $K$ nonrejected classes of each pixel imposes an implicit ordering of the pixels (higher probability leading to lower potential to be rejected) and the amount of rejection is controlled by probability of rejection $\gamma$, the threshold.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

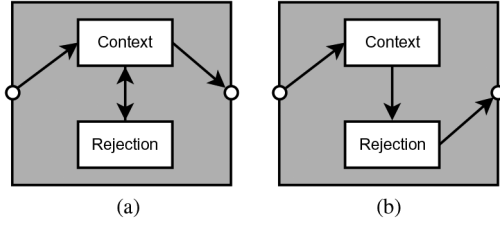CONDESSA *et al.*: SUPERVISED HYPERSPECTRAL IMAGE CLASSIFICATION WITH REJECTION

5

Fig. 3. Schemes for computation of context and rejection. (a) JCR—joint computation of context and rejection. (b) SCR—sequential computation of context and rejection.

The simple rejection scheme in (11) is limited by its pixel-based behavior. There is no awareness of context. In image classification, the use of context is of paramount importance as neighboring pixels are likely to belong to the same class. The same reasoning applies to the rejection. The potential for a pixel to be rejected should not be independent of whether the pixel is surrounded by other pixels that are rejected or surrounded by pixels that are not rejected. As discussed in Section I, the use of context in image classification, namely in hyperspectral image classification, is responsible for significant increase in performance.

To solve the need for contextual awareness of rejection, we combine rejection and context. We consider two different ways to combine classification with rejection with classification with context. We can jointly compute context and rejection—JCR [as seen in Fig. 3(a)]—by considering rejection to be an extra class, subject to the same contextual cues that the other classes are. This is explored in Section III-A, where we instantiate JCR with the the SegSALSA algorithm applied to an extended set of probabilities, containing rejection as a $K + 1$ class. On the other hand, we can harness the potential of the SegSALSA algorithm to provide a hidden field that provides us with an implicit ordering of the pixels according to their potential to be rejected—the maximum value of the hidden field for each pixel—that takes into account the contextual cues. This allows us to compute sequentially the rejection after the context—SCR [as seen in Fig. 3(b)]. We follow this approach in Section III-B, where we instantiate SCR with the rejection computed from the hidden field resulting from the SegSALSA algorithm with $K$ classes through the computation of a rejection field.

### A. Joint Computation of Context and Rejection

To compute jointly context and rejection, we consider rejection as an extra class. Rejection is conceptualized as an extra class that should be selected when there is evidence of probable misclassification by the classifier. In this formulation, the threshold $\gamma$ in (11) is connected to the probability of misclassification by the classifier.

Let $p_i^r$ denote the probability of the classifier misclassifying the $i$th pixel; we can easily extend the set of labels $\mathcal{L} = \{1, \ldots, K\}$ to include the extra class $K + 1$ corresponding to rejection $\mathcal{L}' = \{1, \ldots, K, K + 1\}$. With the new rejection class in place, we need to normalize the probabilities. The new class probabilities $p'$ become

$$p'(y_i|\mathbf{x}_i) = \begin{cases} p_i^r, & \text{if } y_i = K + 1 \\ (1 - p_i^r)p(y_i|\mathbf{x}_i), & \text{otherwise.} \end{cases} \quad (12)$$

*SegSALSA-JCR:* The JCR leads to an extended SegSALSA formulation of (4), where the hidden field is now of dimension $\mathbf{z} \in \mathbb{R}^{(K+1)\times n}$ and the probability vector $\mathbf{p}_i$ becomes $\mathbf{p}'_i$

$$\widehat{\mathbf{z}}_{MMAP} = \arg \min_{\mathbf{z} \in \mathbb{R}^{(K+1)\times n}} -\sum_{i \in \mathcal{S}} \left( \ln \left( \mathbf{p}'^T_i \mathbf{z}_i \right) \right) - \ln p(\mathbf{z})$$

$$\text{subject to: } \mathbf{z} \geq 0, \quad \mathbf{1}_K^T \mathbf{z} = \mathbf{1}_n^T.$$

The rejection extra class is subject to the same vectorial total variation prior as the other classes. By considering rejection as an extra class, we are able to seamlessly combine classification with context with classification with rejection in the SegSALSA formulation.

The basic assumption for the JCR is that of rejection as an extra class with a probability associated to classifier failure. A scaling parameter $\gamma$ controls the relative weight of the probability of classifier misclassification with regard to the probability of the other classes. By varying the value of $\gamma$ we are able to vary the amount of rejection obtained, with larger values of $\gamma$ corresponding to larger values of the rejected fraction. We now present two different rejection schemes based on two different models for classifier:

1) uniform probability of classifier failure—classifier failure is equiprobable across all the pixels;
2) entropy-weighted probability of classifier failure—classifier failure is more likely in pixels with higher entropy associated to their classification.

*1) JCR-U—Uniform Probability of Classifier Failure:* This uniformly weighted model assumes that, regardless of the probability distribution for each of the labels on a pixel, there is a constant probability of failure of the classifier, i.e., for all the pixels, the probability of misclassification, and thus rejection, is constant. The rejection depends only on the scaling parameter $\gamma$ that defines how frequently misclassification is assumed

$$p_i^r = \gamma.$$

*a) Class probabilities of extended set of labels:* The class probabilities for the extended set of labels $\mathcal{L}'$ are

$$p'(y_i|\mathbf{x}_i) = \begin{cases} \gamma, & \text{if } y_i = K + 1 \\ (1 - \gamma)p(y_i|\mathbf{x}_i), & \text{otherwise.} \end{cases} \quad (13)$$

In this model, misclassifications are assumed to be equiprobable across the entire image.

*2) JCR-E—Entropy Weighted Probability of Classifier Failure:* This entropy-weighted model assumes that the probability of failure of the classifier scales with the entropy associated with the probability vector from the classification, i.e., pixels with higher entropy are more likely to be misclassified, and thus rejected. The rejection depends both from the scaling parameter $\gamma$ that defines how frequent the misclassification is assumed, and from the uncertainty associated with the classification modeled by the entropy weighting

$$p_i^r = \gamma H(\mathbf{p}_i)$$

where $H(\mathbf{p}_i)$ denotes the entropy of the probability distribution $\mathbf{p}_i = [p(y_i = 1|\mathbf{x}_i) \ldots p(y_i = K|\mathbf{x}_i)]$.

*a) Class probabilities of extended set of labels:* The class probabilities for the extended set of labels $\mathcal{L}'$ are

$$p'(y_i|\mathbf{x}_i) = \begin{cases} \gamma H(\mathbf{p}_i), & \text{if } y_i = K+1 \\ (1 - \gamma H(\mathbf{p}_i))p(y_i|\mathbf{x}_i), & \text{otherwise.} \end{cases} \quad (14)$$

In this model, misclassifications are assumed to be more probable in pixels with higher entropy.

*3) Limitations of JCR:* A major limitation of considering rejection as an extra class modeling classifier failure is the inability to define *a priori* the amount of rejection obtained. Whereas $\gamma$ in (13) and (14) corresponds to the scaling factor associated with the probability of classifier failure, the use of context through SegSALSA makes it impossible to predict the rejected fraction before the computation of SegSALSA. This means that, given an ordering of the pixels according to their potential to be rejected before the computation of context, there is no guarantee that the ordering of the pixels will be the same after the computation of context.

### B. SCR—Sequential Computation of Context and Rejection

To mitigate the aforementioned limitations associated with the JCR, we consider a second approach where rejection is computed after the context, i.e., a sequential approach. We start by noting that by using SegSALSA to compute the context, in addition to the labeling $\widehat{\mathbf{y}}$, we have the hidden field $\widehat{\mathbf{z}}_{\text{MMAP}}$ resulting from the optimization problem (4) from where the labeling is computed.

*SegSALSA-SCR:* This hidden field $\mathbf{z}$ provides an indication of the degree of confidence associated with the label of each pixel. If $[\mathbf{z}_i]_k > [\mathbf{z}_j]_l$, i.e., if the $k$th component of the hidden vector associated with the $i$th pixel $[\mathbf{z}_i]_k$ has a larger value than the $l$th component of the hidden vector associated with $j$th pixel $[\mathbf{z}_j]_l$, then we are led to believe that assigning the label $l$ in the $j$th pixel corresponds to a lower degree of confidence than assigning the label $k$ in the $i$th pixel.

Let us consider the labeling $\widehat{\mathbf{y}}$

$$\widehat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{L}^n} p(\mathbf{y}|\widehat{\mathbf{z}}_{\text{MMAP}})$$

and the associated maximum probabilities of the labeling $\mathbf{z}_{\widehat{\mathbf{y}}}$, such that

$$\mathbf{z}_{\widehat{y}_i} = p(\widehat{y}_i|\widehat{\mathbf{z}}_{\text{MMAP}}). \quad (15)$$

If $[\mathbf{z}_i]_{\widehat{y}_i} > [\mathbf{z}_j]_{\widehat{y}_j}$, there is strong evidence that a higher degree of confidence exists in the labeling of the $i$th pixel as $\widehat{y}_i$ than in the labeling of the $j$th pixel as $\widehat{\mathbf{y}}_j$. We denote the resulting field $\mathbf{z}_{\widehat{\mathbf{y}}}$ as *rejection field*.

By sorting $\mathbf{z}_{\widehat{\mathbf{y}}}$, we obtain an ordering of the pixels according to their relative confidence. Thus, from the hidden field $\mathbf{z}$ and the resulting rejection field $\mathbf{z}_{\widehat{\mathbf{y}}}$, we obtain an implicit ordering of the pixels according to their potential to be rejected. The selection of a fraction of the lowest confidence pixels to be rejected yields a simple, yet effective scheme for rejection. This method allows one not only to define arbitrary values of the rejected fraction, but also to change the values on the fly, without the need to re-solve any contextual problem.
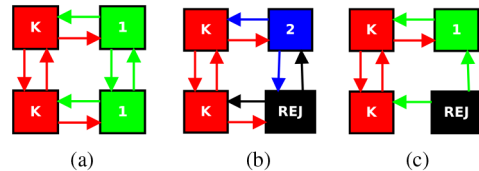


Fig. 4. Approximation effects of joint versus sequential context and rejection. (a) Classification. (b) JCR. (c) SCR.

By promoting preservation and alignment of the discontinuities across the classes, the vectorial total variation prior (3), when applied to the hidden field $\mathbf{z}$, influences the behavior of the rejection field $\mathbf{z}_{\widehat{\mathbf{y}}}$. This results on an emergent prior behavior on the rejection field. The preservation and alignment of the discontinuities is thus imposed on the rejection field.

*1) Approximation Effects of Rejection by Rejection Field:* By obtaining rejection through the use of a rejection field computed from the hidden field, pixels cannot switch label as a result of the introduction of rejection. The influence that a pixel and its assigned label (before rejection) exerts on neighboring pixels does not disappear when the pixel is rejected. This is illustrated in Fig. 4, where the top-right pixel, assigned to class 1 (green), is influenced by the green label assigned to the bottom-right pixel, otherwise being assigned to the class 2 (blue). If the rejection is computed jointly with context (rejection as an extra class), the rejection of the bottom-right pixel stops the interaction (green) that forces the top-right pixel to belong to class 1 (green), and consequently the top-right pixel switches to class 2 (blue). If the rejection is computed from a rejection field, the interaction (green) that forces the top-right pixel to belong to class 1 (green) persists even though the bottom-right pixel is rejected. In practice, the changes caused by these effects apply only to a very small portion of the data.

## IV. EXPERIMENTAL RESULTS

To evaluate the proposed methodologies of joint and sequential computation of context and rejection, we apply them to the task of supervised hyperspectral image classification of two well-known hyperspectral scenes: 1) AVIRIS Indian Pines and 2) ROSIS Pavia University scene. In both scenes, the labeled ground truth is only available for a portion of the image. We apply the methodologies on the entire image and assess the performance on the subset of pixels that belongs to the labeled ground truth. We aim to show the following characteristics of supervised hyperspectral image classification with rejection:

1) classification with context and rejection can outperform classification with context only;
2) classification with rejection does not affect all the classes equally;
3) by using classification with context and rejection with small training sets, we are able to achieve performances comparable to context only with larger training sets.

This is achieved by assessing the performance of the joint (SegSALSA-JCR-U and SegSALSA-JCR-E) and sequential (SegSALSA-SCR) schemes for context and rejection using SegSALSA to compute the context. The multinomial logistic

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

CONDESSA *et al.*: SUPERVISED HYPERSPECTRAL IMAGE CLASSIFICATION WITH REJECTION

7

regression (MLR) weights are modeled with LORSAL [19], thus obtaining the LORSAL-SegSALSA-JCR-U, LORSAL-SegSALSA-JCR-E, and LORSAL-SegSALSA-SCR methods for image classification with context and rejection. The SegSALSA algorithm requires the existence of class probabilities, which restrict us to the use of classifiers that output probabilities. The use of an MLR modeled with LORSAL can be easily replaced by the use of a probabilistic extension to support vector machines (SVM), such as relevance vector machines [26]. The LORSAL parameters used are $\lambda = 0.01$, $\theta = 0.001$ with radial basis function (RBF) kernels with a width of 1. For the SegSALSA algorithm, the value of $\lambda_{\text{TV}}$ is 2. Computational complexities of both LORSAL-SegSALSA-SCR and LORSAL-SegSALSA-JCR approaches are dominated by the SegSALSA, which is $O(Kn \log n)$, with $K$ the number of classes and $n$ the number of image pixels. This means that computing LORSAL-SegSALSA-SCR has complexity of $O(Kn \log n)$ and computing LORSAL-SegSALSA-JCR has complexity of $O((K+1)n \log n)$. In LORSAL-SegSALSA-JCR-U and LORSAL-SegSALSA-JCR-E, a sweep on the scaling parameter of $\gamma$ from 0 to 1 is performed to observe the joint variation of nonrejected accuracy, classification quality, and fraction of rejected pixels.

### A. Indian Pine

The AVIRIS Indian Pine scene (Fig. 5) was acquired by the AVIRIS sensor in NorthWest Indiana, USA. The scene consists of $145 \times 145$ pixel section with 200 spectral bands (with water absorption bands already purged) and contains 16 mutually nonexclussive classes, with the classification accuracy and classification quality being measured on those 16 classes.

The classification maps present in Fig. 5 show clearly the effects of classification with context and rejection: a significant number of misclassified pixels are rejected, thus increasing classification performance. We start from an accuracy of 51.39% with the MAP classification (with the training set composed of 10 pixels randomly selected per class, roughly 1.6% of the entire labeled data set) in Fig. 5(b), and by computing the context alone with LORSAL-SegSALSA achieve an accuracy of 69.55% in Fig. 5(c).

In Fig. 5(d)–(f), we show the classification maps for the rejected fraction that corresponds to the maximum classification quality. This means that starting from the 69.55% accuracy of LORSAL-SegSALSA, the value of rejected fraction is selected such that the number of correct decisions (rejected the pixel when incorrectly classified, and not reject the pixel when correctly classified) is maximized. For LORSAL-SegSALSA-JCR-U, we achieve a nonrejected accuracy of 80.31% at a rejected fraction of 20.65% leading to a classification quality of 78.56%. This means that by not classifying the entire image, we depart from an accuracy of 69.55% on the entire image to an accuracy of 80.31% on 79.35% of the image, with 78.56% of the pixels either correctly classified and not rejected, or incorrectly classified and rejected. For LORSAL-SegSALSA-JCR-E, we achieve a nonrejected accuracy of 76.01% at a rejected fraction of 15.85% leading to a classification quality of 74.23%. For LORSAL-SegSALSA-SCR, we achieve 79.97%
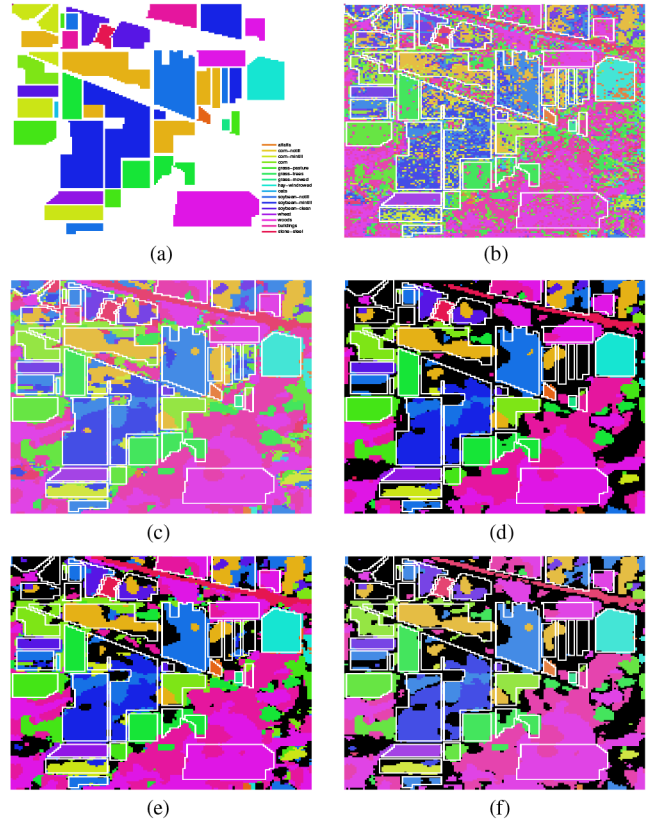


Fig. 5. Classification results for Indian Pines (10 pixels per class as training set), with rejection in black. (a) Ground truth. (b) MAP classification using LORSAL and (c) classification with context—LORSAL-SegSALSA. Classification with context and rejection with maximum classification quality for (d) LORSAL-SegSALSA-JCR-U, (e) LORSAL-SegSALSA-JCR-E, and (f) LORSAL-SegSALSA-SCR. Overall and classwise nonrejected accuracy, rejected fraction, and classification quality in Table I.

of nonrejected accuracy at a rejected fraction of 23.75% and a classification quality of 76.16%.

The introduction of rejection does not affect all the classes equally. Some classes are more positively affected by rejection, whereas the classification performance of other classes suffers. The classwise classification performances are shown in Table I, with classwise performance improvement highlighted in green and classwise performance decrease highlighted in red.

In Fig. 6, we illustrate the variation of the performance measures for classification with rejection as a function of the rejected fraction. It is clear, there is a steady increase of nonrejected accuracy by increasing the amount of the image rejected. On the other hand, by using the classification quality, we can compare the number of correct decisions made as we change the rejected fraction. From not rejecting any portion of the image, leading to a classification quality equal to the accuracy of the LORSAL-SegSALSA, we are able to increase the performance until it peaks, corresponding to a higher accuracy on the nonrejected pixels without rejecting too much of the image. We note the close position of peaks of the classification qualities for the LORSAL-SegSALSA-JCR-U and the LORSAL-SegSALSA-SCR approaches.

To compare the approaches of classification with rejection with the state of the art methods, we need to consider an

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

8                                   IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING

TABLE I

PERFORMANCE OF CLASSIFICATION WITH REJECTION FOR INDIAN PINE (10 PIXELS PER CLASS AS TRAINING SET). OVERALL AND CLASSWISE NONREJECTED ACCURACY, REJECTED FRACTION, AND CLASSIFICATION QUALITY CORRESPONDING TO MAXIMUM OVERALL CLASSIFICATION. INCREASE IN PERFORMANCE (GREEN) AND DECREASE IN PERFORMANCE (RED). BEST CLASSWISE CLASSIFICATION PERFORMANCE IN BOLD TYPESET.

| Class | Number pixels | No rejection — Initial accuracy (%) | LORSAL-SegSALSA-JCR-U — Nonrejected accuracy (%) | Rejected fraction (%) | Classification quality (%) | LORSAL-SegSALSA-JCR-E — Nonrejected accuracy (%) | Rejected fraction (%) | Classification quality (%) | LORSAL-SegSALSA-SCR — Nonrejected accuracy (%) | Rejected fraction (%) | Classification quality (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Alfalfa | 46 | 95.65 | 100.00 | 6.52 | 97.83 | 100.00 | 4.35 | **100.00** | 100.00 | 4.35 | **100.00** |
| Corn-notill | 1428 | 54.55 | 63.66 | 32.35 | **63.94** | 57.45 | 29.06 | 56.02 | 63.15 | 29.69 | **63.94** |
| Corn-mintill | 830 | 25.66 | 44.54 | 55.90 | **69.52** | 38.62 | 42.29 | 61.20 | 40.59 | 51.33 | 65.18 |
| Corn | 237 | **99.16** | 100.00 | 2.53 | 98.31 | 99.16 | 0.00 | **99.16** | 100.00 | 5.06 | 95.78 |
| Grass-pasture | 483 | 82.82 | 88.01 | 8.49 | **86.75** | 84.60 | 4.55 | 83.23 | 87.33 | 8.49 | 85.51 |
| Grass-trees | 730 | **96.85** | 97.41 | 4.66 | 93.56 | 97.34 | 7.40 | 90.82 | 97.54 | 5.48 | 93.01 |
| Grass-mowed | 28 | **100.00** | 100.00 | 14.29 | 85.71 | 100.00 | 0.00 | **100.00** | 100.00 | 14.29 | 85.71 |
| Hay-windrowed | 478 | 99.37 | 100.00 | 0.42 | **100.00** | 99.37 | 0.00 | 99.37 | 99.37 | 0.42 | 98.95 |
| Oats | 20 | **95.00** | NaN[1] | 100.00 | 5.00 | 95.00 | 0.00 | **95.00** | 100.00 | 60.00 | 45.00 |
| Soybean-notill | 972 | 86.42 | 98.44 | 20.99 | **90.12** | 93.94 | 16.87 | 86.63 | 94.88 | 19.55 | 85.80 |
| Soybean-mintill | 2455 | 52.75 | 62.60 | 24.20 | **66.35** | 58.15 | 18.74 | 60.49 | 59.46 | 24.44 | 61.55 |
| Soybean-clean | 593 | 72.68 | 85.62 | 21.42 | **83.31** | 78.85 | 14.67 | 76.56 | 83.70 | 23.44 | 78.92 |
| Wheat | 205 | **99.51** | 100.00 | 1.46 | 99.02 | 99.51 | 0.00 | **99.51** | 100.00 | 1.46 | 99.02 |
| Woods | 1265 | 89.09 | 91.88 | 2.69 | **92.41** | 88.82 | 0.32 | 88.30 | 90.79 | 3.00 | 90.04 |
| Buildings | 386 | **63.21** | 63.14 | 29.02 | 55.44 | 65.73 | 16.84 | 62.95 | 61.59 | 28.50 | 53.37 |
| Stone-steel | 93 | 93.55 | 100.00 | 6.45 | **100.00** | 93.55 | 0.00 | 93.55 | 100.00 | 6.45 | **100.00** |
| All | 10249 | 69.55 | 80.31 | 20.65 | 78.56 | 76.01 | 15.85 | 74.23 | 79.97 | 23.75 | 76.16 |

[1] As all the pixels corresponding to the *oats* class are rejected, it is not possible to compute the nonrejected accuracy.
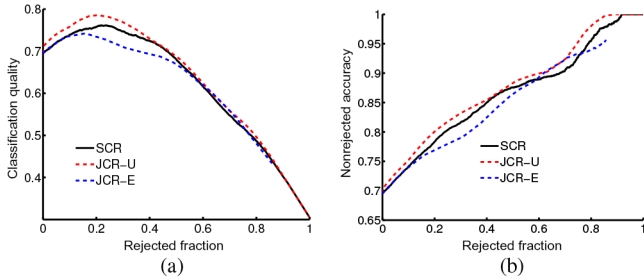


Fig. 6. Performance for classification with rejection of the Indian Pine scene (10 pixels per class as training set). Classification with LORSAL-SegSALSA-SCR (black), and by LORSAL-SegSALSA-JCR-U (red) and LORSAL-SegSALSA-JCR-E (blue). (a) Classification quality. (b) Nonrejected accuracy.

increase of the training set dimension. In Table II, we compare the performance of our methods with the results available in [27] for multiple classifiers with large training sets (10% of the pixels as training set): classifiers without context, classifiers with context, and classifiers with context based on superpixelization (where an unsupervised segmentation produces an oversegmented partitioning of the image and forces pixels belonging to the same partition element to belong to the same class). We compare the performance of our methods with equivalent and smaller training sets (10% and 5% of pixels randomly selected as training set, respectively).

For the classifiers without context, we consider SVM [28] and LORSAL [19]. For the classifiers with context, we consider SVM with composite kernels (SVM-CK) [29], LORSAL with multilevel logistic Markov random field priors (LORSAL-MLL) [19], sparse representation-based classification (SRC) [30], multinomial logistic regression with generalized composite kernel (MLR-GCK) [31], and LORSAL-SegSALSA [13]. For the classifiers with context based on superpixelization, we consider the superpixel-based classification via multiple kernels (SC-MK) [27] and its simplified version (INTRASC-MK) [27].

TABLE II

COMPARISON OF CLASSIFICATION PERFORMANCE FOR INDIAN PINE. OVERALL ACCURACY (WITH NO REJECTION) FOR MULTIPLE CLASSIFIERS WITH 10% OF PIXELS AS TRAINING SET. COMPARISON WITH SCR AND JCR FOR 5% AND 10% OF PIXELS AS TRAINING SET FOR DIFFERENT REJECTED FRACTIONS. COMPARABLE NONREJECTED ACCURACIES FOR CLASSIFICATION WITH CONTEXT ONLY (MAGENTA) AND FOR CLASSIFICATION WITH CONTEXT ONLY BASED ON SUPERPIXELS (CYAN).

| Classifier | Training set size (%) | Nonrejected accuracy $A(r)$ (%) | Rejected fraction $r$ (%) | Classification quality $Q(r)$ (%) |
|---|---|---|---|---|
| SVM [28] | 10 | 79.53 | 0.00 | 79.53 |
| SVM-CK [29] | 10 | 91.51 | 0.00 | 91.51 |
| LORSAL-MLL [19] | 10 | 94.73 | 0.00 | 94.73 |
| SRC [30] | 10 | 94.66 | 0.00 | 94.66 |
| MLR-CGK [31] | 10 | 96.29 | 0.00 | 96.29 |
| INTRASC-MK [27] | 10 | 97.53 | 0.00 | 97.53 |
| SC-MK [27] | 10 | 98.06 | 0.00 | 98.06 |
| LORSAL [19] | 5 | 72.72 | 0.00 | 72.72 |
| LORSAL-SegSALSA [25] | 5 | 86.01 | 0.00 | 86.01 |
| LORSAL-SegSALSA-SCR | 5 | 88.97 | 5.00 | 88.03 |
|  |  | 91.36 | 10.00 | 88.44 |
|  |  | 93.63 | 15.00 | 88.15 |
|  |  | 95.16 | 20.00 | 86.24 |
| LORSAL-SegSALSA-JCR-U | 5 | 89.52 | 5.03 | 88.93 |
|  |  | 92.06 | 10.05 | 89.53 |
|  |  | 93.74 | 15.00 | 88.23 |
|  |  | 95.16 | 19.92 | 86.19 |
| LORSAL-SegSALSA-JCR-E | 5 | 89.43 | 5.01 | 88.78 |
|  |  | 91.35 | 10.04 | 88,25 |
|  |  | 92.73 | 14.99 | 86.53 |
|  |  | 93.87 | 20.02 | 84.01 |
| LORSAL [19] | 10 | 78.57 | 0.00 | 78.57 |
| LORSAL-SegSALSA [25] | 10 | 92.26 | 0.00 | 92.26 |
| LORSAL-SegSALSA-SCR | 10 | 94.72 | 5.00 | 92.70 |
|  |  | 96.38 | 10.00 | 91.22 |
|  |  | 97.64 | 15.00 | 88.71 |
|  |  | 98.66 | 20.00 | 85.59 |
| LORSAL-SegSALSA-JCR-U | 10 | 95.55 | 5.03 | 94.02 |
|  |  | 96.61 | 10.00 | 91.39 |
|  |  | 97.52 | 15.02 | 88.27 |
|  |  | 98.63 | 20.05 | 85.27 |
| LORSAL-SegSALSA-JCR-E | 10 | 94.38 | 5.02 | 91.81 |
|  |  | 95.60 | 10.00 | 89.58 |
|  |  | 96.21 | 15.00 | 86.07 |
|  |  | 97.19 | 20.04 | 82.97 |

The use of classification with context and rejection is able to obtain significant performance improvements. We note that, by using classification with context and rejection with smaller training sets, both in sequential (SCR) and joint (JCR) approaches, we are able to achieve performances on the nonrejected data equivalent to those achieved by using classification with context only in larger training sets (highlighted in magenta in Table II) not considering the superpixel-based
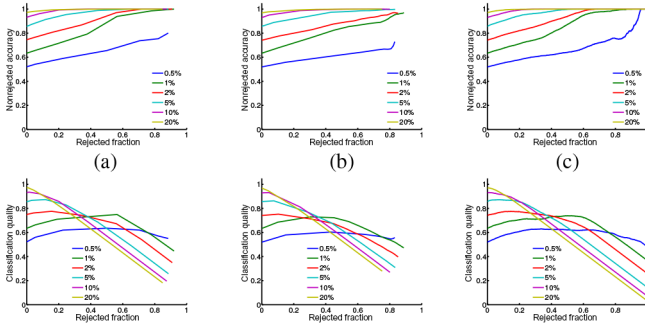
Fig. 7. Effect of weak versus strong classifiers in classification with rejection. SegSALSA-JCR and SegSALSA-SCR approaches with increasing training size. Stronger classifiers (larger training sets) achieve peak classification quality with smaller values of rejected fraction than weak classifiers (smaller training sets). (a) SegSALSA-JCR-U. (b) SegSALSA-JCR-E. (c) SegSALSA-SCR.

methods. For example, with 5% of the pixels as training set, and while rejecting close to 15% of the pixels, we are able to achieve performances close to the ones achieved by context only with 10% of the pixels as training set, such as LORSAL-MLL, SRC, MLR-CGK, and SegSALSA.

On the other hand, we can achieve accuracies equivalent to the accuracies of superpixel-based methods (highlighted in cyan in Table II), with equivalent training set size, by using rejection. By using context and rejection, we are able to close the gap between the state of the art methods using superpixels (98.06% overall accuracy) and SegSALSA (92.26% overall accuracy). The rejection of 15% of the pixels in SCR allows us to attain values of nonrejected accuracy (97.64%) comparable to the state of the art.

As pointed in Section I, the performance improvements resulting from the combination of rejection and context are more significant for weaker classifiers with lower performance. This is illustrated in Fig. 7, where the strength of the classifier is a result of the training set size (from 0.5% to 20% of the labeled pixels used as training set). It is interesting to note that the shift of the peak of classification quality to lower values of rejected fraction as the classification problem gets easier and the classifier gets more accurate. There is an increased dependency on the rejector as the classifier gets weaker.

### B. Pavia University

The Pavia University scene (Fig. 8) was acquired with the ROSIS sensor in Pavia, Italy. The scene consists of a $610 \times 340$-pixel hyperspectral image with 103 spectral bands containing nine not mutually exclusive classes, with the classification accuracy and classification quality being measured on those nine classes.

The classification maps in Fig. 8 show an easier problem for the LORSAL and LORSAL-SegSALSA, with higher classification performances with context only (shown in Table III) when compared to the Indian Pine scene. The rejector will have a harder task to improve the performance, leading to maximum classification qualities with smaller respective rejected fractions, i.e., a larger proportion of correct decisions is achieved by rejecting less.
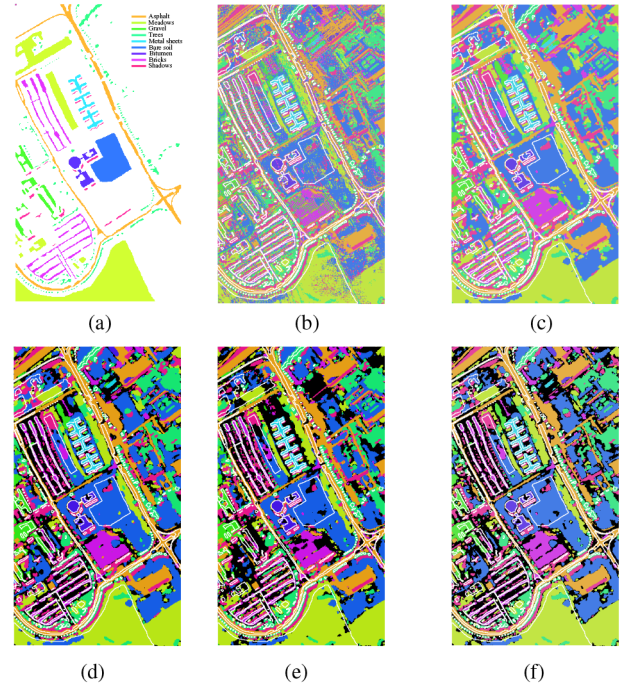


Fig. 8. Classification results for Pavia University (rejection in black). (a) Ground truth. (b) MAP classification using LORSAL and (c) classification with context—LORSAL-SegSALSA. Classification with context and rejection with maximum classification quality for (d) LORSAL-SegSALSA-JCR-U, (e) LORSAL-SegSALSA-JCR-E, and (f) LORSAL-SegSALSA-SCR. Overall and classwise nonrejected accuracy, rejected fraction, and classification quality in Table III.

We start from an accuracy of 70.13% with the MAP classification (with the training set composed of 10 pixels randomly selected per class, roughly 0.2% of the entire labeled data set) in Fig. 8(b) and by computing the context alone with SegSALSA achieve an accuracy of 80.67% in Fig. 8(c).

In Fig. 8(d)–(f), we show the classification maps that correspond to the maximum classification quality. This means that starting from the 80.67% accuracy of LORSAL-SegSALSA, we reject such that the number of correct decisions is maximized. For LORSAL-SegSALSA-JCR-U, we achieve a nonrejected accuracy of 82.25% at a rejected fraction of 3.12% leading to a classification quality of 81.81%. For LORSAL-SegSALSA-JCR-E, we achieve a nonrejected accuracy of 86.45% at a rejected fraction of 12.75% leading to a classification quality of 82.93%. This means that by not classifying the entire image, we depart from an accuracy of 80.67% on the entire image to an accuracy of 86.45% on 86.25% of the image, with 82.93% of the pixels either correctly classified and not rejected, or incorrectly classified and rejected. For LORSAL-SegSALSA-SCR, we achieve 84.54% nonrejected accuracy at a rejected fraction of 9.16% and a classification quality of 82.08%.

The classwise classification performances are shown in Table III. Taking the example of the LORSAL-SegSALSA-JCR-E results, only the classification performance of the *meadows* class is increased, with the performance of the other classes decreasing slightly. However, the abundance of the *meadows*

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

10 IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING

TABLE III
PERFORMANCE OF CLASSIFICATION WITH REJECTION FOR PAVIA UNIVERSITY. OVERALL AND CLASSWISE NONREJECTED ACCURACY, REJECTED FRACTION, AND CLASSIFICATION QUALITY CORRESPONDING TO MAXIMUM OVERALL CLASSIFICATION QUALITY. INCREASE IN PERFORMANCE (GREEN) AND DECREASE IN PERFORMANCE (RED). BEST CLASSWISE CLASSIFICATION PERFORMANCE IN BOLD TYPESET.

| Class | Number pixels | No rejection Initial accuracy (%) | LORSAL-SegSALSA-JCR-U Nonrejected accuracy (%) | Rejected fraction (%) | Classification quality (%) | LORSAL-SegSALSA-JCR-E Nonrejected accuracy (%) | Rejected fraction (%) | Classification quality (%) | LORSAL-SegSALSA-SCR Nonrejected accuracy (%) | Rejected fraction (%) | Classification quality (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Asphalt | 6631 | **96.80** | 97.94 | 4.13 | 95.11 | 97.50 | 9.03 | 89.61 | 98.19 | 8.11 | 91.77 |
| Meadows | 18649 | 65.88 | 67.15 | 1.86 | 67.79 | 74.63 | 17.32 | **74.85** | 70.47 | 10.11 | 70.92 |
| Gravel | 2099 | 71.56 | 74.57 | 8.00 | **73.65** | 77.15 | 20.77 | 71.46 | 76.35 | 15.01 | 73.23 |
| Trees | 3064 | 88.41 | 90.66 | 4.28 | **89.43** | 90.27 | 5.42 | 87.76 | 91.76 | 9.73 | 86.98 |
| Metal sheets | 1345 | **100.00** | 100.00 | 0.00 | **100.00** | 100.00 | 0.00 | **100.00** | 100.00 | 0.00 | **100.00** |
| Bare soil | 5029 | 92.19 | 94.60 | 2.03 | **95.21** | 96.08 | 10.10 | 90.67 | 95.72 | 7.62 | 92.28 |
| Bitumen | 1330 | 95.11 | 95.78 | 0.30 | **96.17** | 95.11 | 0.23 | 94.89 | 96.89 | 3.16 | 95.71 |
| Bricks | 3682 | **92.40** | 96.24 | 8.31 | **92.40** | 97.20 | 13.82 | 88.95 | 97.71 | 12.33 | 91.25 |
| Shadows | 947 | **99.89** | 99.89 | 0.21 | 99.68 | 99.89 | 0.11 | 99.79 | 99.89 | 0.21 | 99.68 |
| All | 42776 | 80.67 | 82.25 | 3.12 | 81.81 | 86.45 | 12.75 | 82.93 | 84.54 | 9.16 | 82.08 |
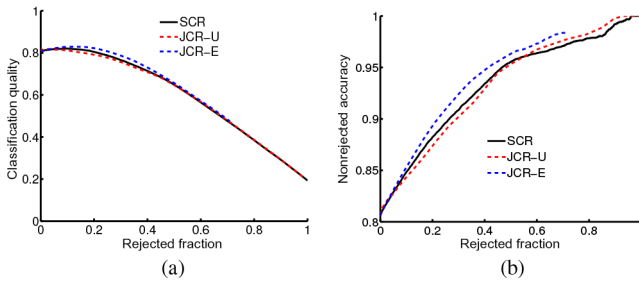


Fig. 9. Performance for classification with rejection of the Pavia University scene. Classification with rejection by LORSAL-SegSALSA-SCR (black), and LORSAL-SegSALSA-JCR-U (red) and LORSAL-SegSALSA-JCR-E (blue) cost. (a) Classification quality. (b) Nonrejected accuracy.
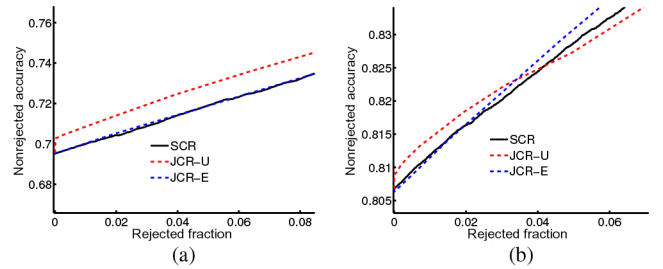


Fig. 10. Approximation effects of SCR versus JCR. Detail of nonrejected accuracy–rejection curve. Classification with LORSAL-SegSALSA-SCR (black), and with LORSAL-SegSALSA-JCR-U (red) and LORSAL-SegSALSA-JCR-E (blue). Increase of accuracy in the joint approaches due to the introduction of the rejected option. (a) Indian Pine. (b) Pavia University.

class compensates the results, with a resulting increase in overall classifier performance. There is no decrease on nonrejected accuracies, in LORSAL-SegSALSA-JCR-E a large portion of correctly classified samples are being rejected across all the classes with exception of the *meadows* class.

In Fig. 9, we illustrate the variation of the performance measures for classification with rejection as a function of the rejected fraction. The peak in classification quality is achieved for values of rejected fraction smaller than the ones in the Indian Pine case. This is a result of an easier classification problem: the high performances achieved by the classifier lead to a low impact of the rejector. As most of the data are correctly classified, it is harder for the rejector to correctly reject pixels. This means that the rejected fraction that optimizes the classification quality, the number of correct decisions made, is much smaller than in the Indian Pine case.

### C. Approximation Effects

Whereas the JCR approaches, with LORSAL-SegSALSA-JCR-U in Indian Pine and LORSAL-SegSALSA-JCR-E in Pavia University, achieve higher performance than the SCR approach for smaller training sets (10 pixels per class), they are computationally more expensive. First, there is not a clear direct connection between the value of $\gamma$ and the rejected fraction; this connection is largely affected by the computation of the context. Whereas, an increase of the value of $\gamma$ can lead to larger rejected fractions, it is not possible to predict how much is rejected by

the joint context and rejection. This is clear in Table II, where we are able to precisely define *a priori* the rejected fraction for the LORSAL-SegSALSA-SCR approaches, but not able to do so for the LORSAL-SegSALSA-JCR approaches.

Second, obtaining the results for joint context and rejection requires a parameter sweep on the value of $\gamma$. This implies, *for each* value of $\gamma$, the computation the SegSALSA algorithm, or any other context computing algorithm, with $K + 1$ classes. For the SCR approach, the rejection is computed after the context, allowing us to obtain all possible values of the rejected fraction in a single computation of the SegSALSA algorithm, or any other context computing algorithm that provides a rejection field. However, the sequential approach is subject to the approximation effect described in Fig. 4. This is clear when we observe in detail the accuracy rejection curves, both for Indian Pine and Pavia University (Fig. 10). For the LORSAL-SegSALSA-JCR-U (in red), there is an increase of classification accuracy with no rejection happening. This corresponds to a change on the labeling simply by the inclusion of the rejection class, as illustrated in Fig. 4. The effect of the alteration of the labeling by introduction of the rejection class cannot be captured in any SCR approach, as the only change on the labeling allowed is for a pixel to be rejected.

### V. CONCLUDING REMARK

In this paper, we introduced classification with rejection in hyperspectral image classification problem as a way to cope

with classification errors arising from known and unknown sources. We presented two different approaches for achieving classification with rejection using context based on *joint* and *sequential* computations of context and rejection. We present experimental results of the methods for supervised hyperspectral image classification with rejection, with context computed using the SegSALSA algorithm. By classifying with rejection, not only we are able to deal with imperfect knowledge in the training set and with smaller training sets, but also able to attain performance gains equivalent to increasing the training set size.

## References

[1] F. Condessa, J. Bioucas-Dias, and J. Kovačević, "Robust hyperspectral image classification with rejection fields," in *Proc. IEEE GRSS Workshop Hyperspectral Image Signal Process.: Evol. Remote Sens. (WHISPERS'15)*, Jun. 2015.

[2] F. Condessa, J. Bioucas-Dias, and J. Kovačević, "Supervised hyperspectral image classification with rejection," in *Proc. IEEE Geosci. Remote Sens. Symp. (IGARSS'15)*, Jul. 2015, pp. 2600–2603.

[3] J. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, Jun. 2013.

[4] S. Li, *Markov Random Field Modeling in Computer Vision*. Berlin, Germany: Springer-Verlag, 1995.

[5] J. Li, J. Bioucas-Dias, and A. Plaza, "Spectral-spatial hyperspectral image segmentation using subpsace mnultinomial logistic regression and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.

[6] X. Bai, H. Zhang, and J. Zhuo, "VHR object detection based on structural feature extraction and query expansion," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6508–6520, Oct. 2014.

[7] X. Bai, Z. Guo, Y. Wang, Z. Zhang, and J. Zhuo, "Semi-supervised hyperspectral band selection via spectral-spatial hypergraph model," *IEEE J. Sel. Topics App. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2774–2783, Jun. 2015.

[8] A. Brown, "Spectral curve fitting for automatic hyperspectral data analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1601–1608, Jun. 2006.

[9] C. K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inf. Theory*, vol. 16, no. 1, pp. 41–46, Jan. 1970.

[10] I. Pillai, G. Fumera, and F. Roli. "Multi-label classification with a reject option," *Pattern Recognit.*, vol. 46, no. 8, pp. 2256–2266, 2013.

[11] F. Condessa, J. Bioucas-Dias, C. Castro, J. Ozolek, and J. Kovačević, "Classification with rejection option using contextual information," in *Proc. IEEE Int. Symp. Biomed. Imag.*, San Francisco, CA, USA, Apr. 2013, pp. 1340–1343.

[12] F. Condessa, C. Castro, J. Ozolek, J. Bioucas-Dias, and J. Kovačević, "Image classification with rejection using contextual information," *IEEE Trans. Med. Imaging*, 2016, http://arxiv.org/abs/1509.01287, to be published.

[13] J. Bioucas-Dias, F. Condessa, and J. Kovačević, "Alternating direction optimization for image segmentation using hidden Markov measure field models," in *Proc. SPIE 9019, Image Process. Algorithms Systems XII*, San Francisco, CA, USA, Feb. 25, 2014, 90190 p., doi: 10.1117/12.2047707.

[14] F. Condessa, J. Bioucas-Dias, and J. Kovačević, "Supervised hyperspectral image segmentation: A convex formulation using hidden fields," in *Proc. IEEE GRSS Workshop Hyperspectral Image Signal Process.: Evol. Remote Sens. (WHISPERS'14)*, Lausanne, Switzerland, Jun. 2014.

[15] J. Marroquin, E. Santana, and S. Botello, "Hidden Markov measure field models for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 11, pp. 1380–1387, Nov. 2003.

[16] B. Goldluecke, E. Strekalovskiy, and D. Cremers, "The natural vectorial total variation which arises from geometric measure theory," *SIAM J. Imag. Sci.*, vol. 5, no. 2, pp. 537–563, 2012.

[17] L. Sun, Z. Wu, J. Liu, and Z. Wei, "Supervised hyperspectral image classification using sparse logistic regression and spatial-TV regularization," in *Proc. IEEE Geosci. Remote Sens. Symp. (IGARSS'13)*, Melbourne, Australia, 2013, pp. 1019–1022.

[18] M. Afonso, J. Bioucas-Dias, and M. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–695, Mar. 2011.

[19] J. Li, J. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new Bayesian approach with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, 10, pp. 3947–3960, Oct. 2011.

[20] J. Lellmann, J. Kappes, J. Yuan, F. Becker, and C. Schnörr, "Convex multi-class image labeling by simplex-constrained total variation," in *Scale Space and Variational Methods in Computer Vision*, X.-C. Tai, K. Morken, M. Lysaker, and K.-A. Lie, Eds. New York, NY, USA: Springer, 2009, vol. 5567 pp. 150–162.

[21] G. Fumera, F. Roli, and G. Giacinto, "Reject option with multiple thresholds," *Pattern Recognit.*, vol. 33, no. 12, pp. 2099–2101, Dec. 2000.

[22] G. Fumera and F. Roli, "Support vector machines with embedded reject option," in *Proc. Int. Workshop Pattern Recognit. Support Vector Mach. (SVM'02)*, Aug. 2002, pp. 68–82.

[23] G. Fumera and F. Roli, "Analysis of error-reject trade-off in linearly combined multiple classifiers," *Pattern Recognit.*, vol. 37, no. 6, pp. 1245–1265, 2004.

[24] R. Sousa and J. Cardoso, "The data replication method for the classification with reject option," *AI Commun.*, vol. 26, no. 3, pp. 281–302, 2013.

[25] F. Condessa, J. Bioucas-Dias, and J. Kovačević, "Performance measures for classification systems with rejection," *Pattern Recognit.*, 2016, [Online]. Available: http://arxiv.org/abs/1504.02763, to be published.

[26] B. Demir and S. Ertürk, "Hyperspectral image classification using relevance vector machines," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 4, pp. 586–590, Oct. 2007.

[27] L. Fang, S. Li, W. Duan, J. Ren, and J. Benediktsson, "Classification of hyperspectral images by exploiting spectral-spatial information of superpixel via multiple kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6663–6674, Dec. 2015.

[28] 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.

[29] G. Camps-Valls, L. Gomez-Chova, J. Mu noz 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.

[30] Y. Chen, N. Nasrabadi, and T. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.

[31] J. Li, P. Marpu, A. Plaza, J. Bioucas-Dias, and J. Benediktsson, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4816–4829, Sep. 2013.

**Filipe Condessa** (S'13) received the B.Sc. and M.Sc. degrees in biomedical engineering from the Instituto Superior Técnico (IST), Technical University of Lisbon (TULisbon, now University of Lisbon), Lisbon, Portugal, in 2009 and 2011, respectively. Currently, he is pursuing the Ph.D. degree in electrical and computer engineering at Carnegie Mellon University, Pittsburgh, PA, USA, and at IST.

His research interests include hyperspectral image classification and biomedical image classification, with the current focus on the design of robust classification techniques combining rejection and context.

Mr. Condessa serves as a Reviewer for the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, and the IEEE Transactions on Medical Imaging. He was the recipient of the Mikio Takagi Student Prize at the 2015 IEEE Geoscience and Remote Sensing Symposium (IGARSS).

**José Bioucas-Dias** (S'87–M'95–SM'15) received the E.E., M.Sc., Ph.D., and Agregado degrees from the Instituto Superior Técnico (IST), Technical University of Lisbon (TULisbon, now University of Lisbon), Lisbon, Portugal, in 1985, 1991, 1995, and 2007, respectively, all in electrical and computer engineering.

Since 1995, he has been with the Department of Electrical and Computer Engineering, IST, where he was an Assistant Professor from 1995 to 2007 and has been an Associate Professor since 2007. Since 1993, he has been a Senior Researcher with the Pattern and Image Analysis Group, Instituto de Telecomunicações, Lisbon, Portugal, which is a private nonprofit research institution. He has authored or coauthored more than 250 scientific publications including more than 70 journal papers (48 of which published in IEEE journals) and 180 peer-reviewed international conference papers and book chapters. His research interests include inverse problems, signal and image processing, pattern recognition, optimization, and remote sensing.

**Jelena Kovačević** (S'88–M'91–SM'96–F'02) received the Dipl.Electr.Eng. degree from the EE Department, University of Belgrade, Beograd, Serbia, in 1986, and the M.S. and Ph.D. degrees from Columbia University, New York, NY, USA, in 1988 and 1991, respectively.

From 1991 to 2002, she was with Bell Labs, Murray Hill, NJ, USA. She was a Co-Founder and Technical VP with xWaveforms, New York City, NY, USA, and an Adjunct Professor with Columbia University. In 2003, she was a Schramm Professor and Head of Electrical and Computer Engineering, Professor of Biomedical Engineering with Carnegie Mellon University, Pittsburgh, PA, USA. She was the Director of the Center for Bioimage Informatics, Carnegie Mellon University. She was a Plenary/keynote Speaker at a number of international conferences and meetings. Her research interests include wavelets, frames, graphs, and applications to bioimaging and smart infrastructure.

Dr. Kovačević is a past Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING, served as a Guest Co-Editor on a number of special issues and is/was on the editorial boards of several journals. She was a regular member of the NIH Microscopic Imaging Study Section and served as a Member-at-Large of the IEEE Signal Processing Society Board of Governors. She is a past Chair of the IEEE Signal Processing Society Bio Imaging and Signal Processing Technical Committee. She has coauthored the books *Wavelets and Subband Coding* (Prentice Hall, 1995) and *Foundations of Signal Processing* (Cambridge University Press, 2014), a top-10 cited paper in the *Journal of Applied and Computational Harmonic Analysis*, and the paper for which he was a recipient of the Young Author Best Paper Award. Her paper on multidimensional filter banks and wavelets was selected as one of the Fundamental Papers in Wavelet Theory. She was the recipient of the Belgrade October Prize in 1986, the E.I. Jury Award at Columbia University in 1991, and the 2010 CIT Philip L. Dowd Fellowship Award from the College of Engineering, Carnegie Mellon University. She has been involved in organizing numerous conferences.