Exploiting Sparsity in Hyperspectral Image Classification via Graphical Models

Umamahesh Srinivas, Yi Chen, Vishal Monga, Nasser M. Nasrabadi, and Trac D. Tran

***Abstract*—A significant recent advance in hyperspectral image classification relies on the observation that the spectral signature of a pixel can be represented by a *sparse* linear combination of training spectra which come from an over-complete dictionary. The sparse representation corresponding to a test pixel is obtained by solving a sparsity-constrained optimization problem, and has been shown to be discriminative while simultaneously enabling excellent noise robustness. A spatio-spectral notion of sparsity is further captured by developing a joint sparsity model according to which, spectral signatures of pixels in a local spatial neighborhood (of the pixel of interest) are constrained to be represented by a common collection of training spectra, albeit with different weights. A challenging open problem is to effectively capture the *class conditional correlations* between these multiple sparse representations corresponding to different pixels in the spatial neighborhood. In this letter, we propose a probabilistic graphical model framework to explicitly mine the conditional dependencies between these distinct sparse features. In particular, our probabilistic graphical models are synthesized using simple tree structures which can be discriminatively learnt (even under limited training) for the purpose of classification. Experimental results on benchmark hyperspectral image databases reveal significant practical improvements over competing approaches that are particularly pronounced in the low training regime.**

I. Introduction

Hyperspectral imaging sensors acquire digital images in hundreds of continuous narrow spectral bands spanning the visible to infrared spectrum [1]. A pixel in hyperspectral images (HSI) is typically a high-dimensional vector of intensities as a function of wavelength. The high spectral resolution of the HSI pixels facilitates superior discrimination of object types in a captured scene. HSI has varied applications including both commercial [2] and military domains [3].

Classification is an important research topic in HSI processing, wherein the class label of each pixel is determined based on its spectral characteristics given a representative training set from each class. The support vector machine (SVM) [4], which solves supervised binary classification problems by finding the optimal separating hyperplane between the two classes, has proved to be a powerful classifier for HSI classification tasks [5]. Classification performance can be further improved using variations of SVM-based classifiers such as the transductive SVM which exploits both labeled and unlabeled samples [6], and SVM with composite kernels which incorporates spatial information directly in the SVM kernels [7].

Recent work has highlighted the relevance of incorporating contextual information during HSI classification to improve performance [7]–[10], particularly because HSI pixels in a local neighborhood generally correspond to the same material and have similar spectral characteristics. A number of techniques have exploited this aspect, for example by including post-processing of individually-labeled samples [8], [9] and Markov random fields in Bayesian approaches [10]. The composite kernel approach [7] explicitly extracts spatial information for each spectral pixel and then combines the spectral and spatial information via kernel composition.

A seminal advance in efforts towards robust image classification is the recent sparse representation-based classification (SRC) technique for automatic face recognition [11]. Experiments have demonstrated the superior recognition performance and robustness of this approach to a variety of imaging distortion scenarios. This sparsity model has been adopted in HSI classification [12], relying on the observation that spectral signatures of the same material usually lie in a subspace whose dimensionality is much smaller than the number of spectral bands. An unknown pixel is then expressed as a sparse linear combination of a few training samples from a given dictionary and the underlying sparse representation vector implicitly encodes the class information. To exploit contextual (spatial) correlation, a joint sparsity model is employed in [12], where neighboring pixels are assumed to be represented by linear combinations of a few *common* training samples in order to enforce smoothness across these neighboring pixels.

The technique in [12] performs classification by using (spectral) reconstruction error computed over the pixel neighborhood. We propose a probabilistic graphical model framework to explicitly determine conditional dependencies between distinct sparse features obtained via the joint sparsity model. Specifically, a pair of discriminative tree graphs [13] is learnt for each distinct set of features, i.e. the sparse representation vectors of each pixel in the local spatial neighborhood of a central pixel. These features invariably exhibit class conditional correlations. To capture these correlations for classification, we thicken (i.e. introduce new edges) the individual (disjoint) graphs corresponding to each sparse feature set via a boosting approach [14]. Hence we learn a *discriminative* classifier that combines the distinct sparse features unlike the *reconstruction* residual in [12] which does not capture inter class information. Further, probabilistic graphical models as proposed in this work can be robustly learnt even under limited training.

The remainder of this letter is structured as follows. We first

U. Srinivas and V. Monga are with the Department of Electrical Engineering, Pennsylvania State University, University Park, PA, USA. Y. Chen and T. D. Tran are with Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA. N. M. Nasrabadi is with US Army Research Laboratory, Adelphi, MD, USA.

This work has been partially supported by NSF under Grants CCF-1117545 and ARO under Grant 60219-MA.

briefly review sparsity-based classification and probabilistic graphical models in Section II. Our main contribution is presented in Section III. The effectiveness of the proposed approach is demonstrated on several real HSI data sets in Section IV. Section V concludes the paper.

II. Background

A. Sparsity model for hyperspectral classification

As discussed earlier, the HSI sparsity model is an extension of the sparse representation-based framework first introduced for face recognition [11]. This model relies on the key observation that the spectral signatures of pixels approximately lie in a low-dimensional subspace spanned by representative training pixels from the same class. Consequently, for a test pixel whose class identity is unknown, there exists a sparse representation in terms of training samples from all classes. Let $\boldsymbol{y} \in \mathbb{R}^B$ be a pixel with B indicating the number of spectral bands, $\boldsymbol{D}_m \in \mathbb{R}^{B \times N_m}, m = 1, 2, \ldots, M$ be the sub-dictionary whose columns are the N_m training samples from the m-th class. The HSI pixel $\boldsymbol{y}$ can then be written as:

$$\boldsymbol{y} = \boldsymbol{D}_1\boldsymbol{\alpha}_1 + \cdots + \boldsymbol{D}_M\boldsymbol{\alpha}_M = \underbrace{\begin{bmatrix}\boldsymbol{D}_1 & \cdots & \boldsymbol{D}_M\end{bmatrix}}_{\boldsymbol{D}} \underbrace{\begin{bmatrix}\boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_M\end{bmatrix}}_{\boldsymbol{\alpha}} = \boldsymbol{D}\boldsymbol{\alpha}, \quad (1)$$

where $\boldsymbol{D} \in \mathbb{R}^{B \times N}$ with $N = \sum_{m=1}^{M} N_m$ is a structured dictionary consisting of training samples (referred to as atoms) from all classes, and $\boldsymbol{\alpha} \in \mathbb{R}^N$ is a sparse vector. Given the overcomplete dictionary $\boldsymbol{D}$, the sparse coefficient vector $\boldsymbol{\alpha}$ is obtained by solving the following optimization problem:

$$\hat{\boldsymbol{\alpha}} = \arg\min \|\boldsymbol{\alpha}\|_0 \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{\alpha}\|_2 \leq \varepsilon, \quad (2)$$

where ε is a suitably chosen reconstruction error tolerance. The sparse vector $\hat{\boldsymbol{\alpha}}$ can be recovered efficiently using many norm minimization techniques, including greedy algorithms or ℓ_1-norm relaxation [15]. The class label of $\boldsymbol{y}$ is finally determined by the minimal residual between $\boldsymbol{y}$ and its approximation from each class sub-dictionary:

$$\text{Class}(\boldsymbol{y}) = \arg \min_{m=1,\ldots,M} \|\boldsymbol{y} - \boldsymbol{D}_m\hat{\boldsymbol{\alpha}}_m\|_2, \quad (3)$$

where $\hat{\boldsymbol{\alpha}}_m$ is the collection of coefficients in $\hat{\boldsymbol{\alpha}}$ corresponding to the m-th class sub-dictionary.

B. Joint sparsity model

Hyperspectral images are usually smooth in the sense that pixels within a small neighborhood usually represent the same material, and thus their spectral characteristics are highly correlated. In order to incorporate this spatial correlation information, the joint sparsity model [16] is employed in HSI classification in [12] by assuming that the sparse vectors associated with pixels in a local spatial neighborhood share a common sparsity pattern. Specifically, let $\{\boldsymbol{y}_t\}_{t=1,\ldots,T}$ be T pixels in a spatial neighborhood centered at $\boldsymbol{y}_1$. These neighboring pixels can be expressed as:

$$\begin{aligned}\boldsymbol{Y} &= \begin{bmatrix}\boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_T\end{bmatrix} = \begin{bmatrix}\boldsymbol{D}\boldsymbol{\alpha}_1 & \boldsymbol{D}\boldsymbol{\alpha}_2 & \cdots & \boldsymbol{D}\boldsymbol{\alpha}_T\end{bmatrix} \\ &= \boldsymbol{D}\underbrace{\begin{bmatrix}\boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \cdots & \boldsymbol{\alpha}_T\end{bmatrix}}_{\boldsymbol{S}} = \boldsymbol{D}\boldsymbol{S}. \end{aligned} \quad (4)$$

The sparse vectors $\{\boldsymbol{\alpha}_t\}_{t=1,\ldots,T}$ share the same support, i.e. they are linear combination of the same collection of atoms from $\boldsymbol{D}$, but with possibly different weights assigned to each atom. As a result, $\boldsymbol{S}$ is a sparse matrix with only a few nonzero rows. This row-sparse matrix $\boldsymbol{S}$ can be recovered by solving the following constrained optimization problem:

$$\hat{\boldsymbol{S}} = \arg\min \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{S}\|_F \quad \text{subject to} \quad \|\boldsymbol{S}\|_{\text{row},0} \leq K_0, \quad (5)$$

where $\|\boldsymbol{S}\|_{\text{row},0}$ denotes the number of non-zero rows of $\boldsymbol{S}$ and $\|\cdot\|_F$ is the Frobenius norm. The problem in (5) can be approximately solved by the greedy Simultaneous Orthogonal Matching Pursuit (SOMP) algorithm [16]. The identity of $\boldsymbol{y}_1$ is then determined by the minimal total residual:

$$\text{Class}(\boldsymbol{y}_1) = \arg \min_{m=1,\ldots,M} \left\|\boldsymbol{Y} - \boldsymbol{D}_m\hat{\boldsymbol{S}}_m\right\|_F, \quad (6)$$

where $\hat{\boldsymbol{S}}_m$ contains the rows of $\hat{\boldsymbol{S}}$ associated with the m-th class sub-dictionary $\boldsymbol{D}_m$.

C. Probabilistic graphical models

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a collection of nodes $\mathcal{V} = \{v_1, \ldots, v_r\}$ and a set of (undirected) edges $\mathcal{E} \subset \binom{\mathcal{V}}{2}$. A probabilistic graphical model describes the joint distribution of a random vector such that each node represent one (or a group of) random variables whose conditional dependencies are indicated by the presence of the connecting edges. The graph structure defines a particular factorization of the joint probability distribution of the random vector in terms of marginal and pairwise statistics. The use of graphical models imparts robustness to the process of learning models for high-dimensional data using limited training (which is usually the case in many practical HSI applications) under moderate computational complexity.

Graphical models can be learnt either generatively or discriminatively. In the generative setting, a single graph which approximates a given distribution is learnt by minimizing the approximation error. The seminal contribution in this area is due to Chow and Liu [17], who obtained the optimal tree approximation $\hat{p}$ of a multivariate distribution p by minimizing the Kullback-Leibler (KL) distance $D(p||\hat{p}) = E_p[\log(p/\hat{p})]$ using first- and second-order statistics, via a maximum-weight spanning tree (MWST) problem. In discriminative learning, a pair of graphs is jointly learnt by minimizing the classification error. Recently, a discriminative learning framework has been proposed [13] by maximizing the tree-approximate J-divergence (a symmetric extension of the KL distance):

$$\hat{J}(\hat{p}, \hat{q}; p, q) = \int (p(x) - q(x)) \log\left[\frac{\hat{p}(x)}{\hat{q}(x)}\right] dx. \quad (7)$$

Based on the observation that maximizing the J-divergence minimizes the upper bound on the probability of classification error, the discriminative learning problem then becomes:

$$(\hat{p}, \hat{q}) = \arg \max_{\hat{p},\hat{q} \text{ are trees}} \hat{J}(\hat{p}, \hat{q}; \tilde{p}, \tilde{q}), \quad (8)$$

where $\tilde{p}$ and $\tilde{q}$ are the empirical estimates. The problem in (8) is shown to decouple into two MWST problems [13]:

$$\begin{aligned}\hat{p} &= \arg \min_{\hat{p} \text{ is a tree}} D(\tilde{p}||\hat{p}) - D(\tilde{q}||\hat{p}) \\ \hat{q} &= \arg \min_{\hat{q} \text{ is a tree}} D(\tilde{q}||\hat{q}) - D(\tilde{p}||\hat{q}).\end{aligned} \quad (9)$$

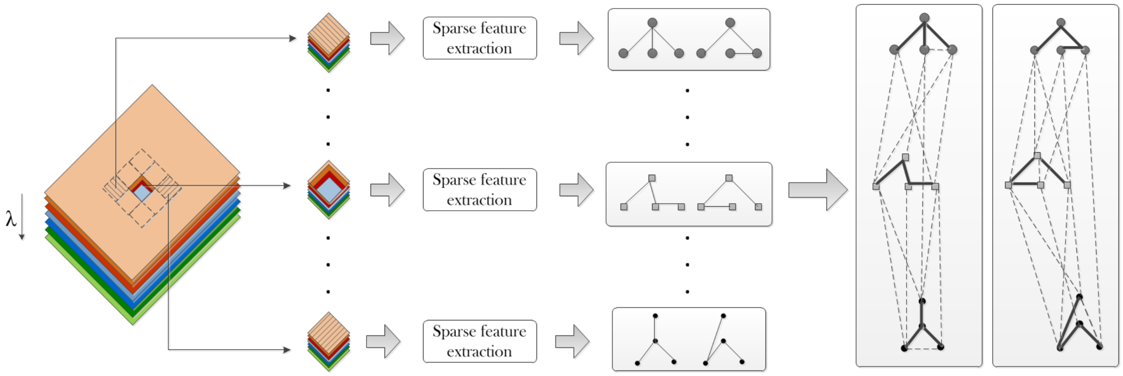

λ
Sparse feature extraction
Sparse feature extraction
Sparse feature extraction

Fig. 1. Hyperspectral image classification using discriminative graphical models on sparse feature representations obtained from local pixel neighborhoods.

Algorithm 1 LSGM (Steps 1-4 offline)

```
1: **Feature extraction (training):** Compute sparse representations
   $\boldsymbol{\alpha}_l, l = 1, \ldots, T$ for neighboring pixels of the training data
2: **Initial disjoint graphs:**
   Discriminatively learn $T$ pairs of $N$-node tree graphs $\mathcal{G}_l^p$ and $\mathcal{G}_l^q$
   on $\{\boldsymbol{\alpha}_l\}$, for $l = 1, \ldots, T$, obtained from training data
3: Separately concatenate nodes corresponding to the two classes,
   to generate initial graphs
4: **Boosting on disjoint graphs**: Iteratively thicken initial disjoint
   graphs via boosting to obtain final graphs $\mathcal{G}^p$ and $\mathcal{G}^q$
   {**Online process**}
5: **Feature extraction (test):** Obtain sparse representations $\boldsymbol{\alpha}_l, l =$
   $1, \ldots, T$ in $\mathbb{R}^N$ from test image
6: **Inference**: Classify based on output of the resulting classifier
   using (10).
```

It is clear from (9) that the optimal choice of $\hat{p}$ ($\hat{q}$) simultaneously minimizes its distance to $\tilde{p}$ ($\tilde{q}$) and maximize its distance from $\tilde{q}$ ($\tilde{p}$). The trade-off between generalization and performance inherent to graphical models has been overcome by iteratively thickening the initial graph with more edges via boosting [14] to learn a richer structure.

Typically in image classification, a variety of features with complementary benefits are employed and the individual classification decisions resulting from each such feature set can be fused intelligently to enhance classification performance. We recently proposed a principled framework to exploit this complementary yet correlated information using probabilistic graphs in [18]. The next Section presents an instantiation of this framework for HSI classification.

III. Exploiting Joint Sparsity via Probabilistic Graphical Models

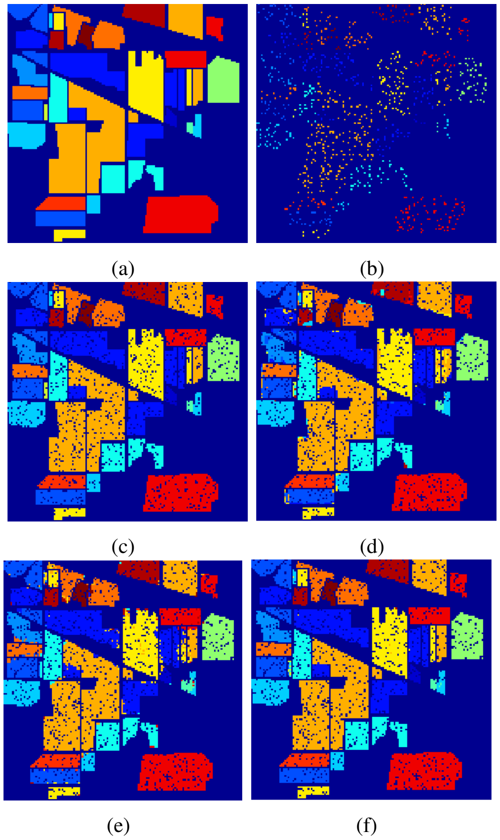
In this section, we introduce our proposed approach for joint sparsity and graphical model-based HSI classification. The proposed Local-Sparsity-Graphical-Model (LSGM) algorithm, summarized in Algorithm 1, consists of an *offline* training stage (Steps 1-4) and an *online* classification stage (Steps 5-6). The local sparsity in the name is indicative of the underlying joint sparsity model to obtain the local sparse features.

An illustration of the overall framework is shown in Fig. 1. The discriminative graphs are learnt in the training stage. Note that the process described here is for binary classification. The approach extends to multi-class problems by learning graphs in a one-against-all manner. That is, for an M-class classification problem, we learn M pairs of discriminative graphs that represent the class conditional p.d.fs $f(\boldsymbol{\alpha}|C_m)$ and $f(\boldsymbol{\alpha}|\tilde{C}_m)$ for $m = 1, 2, \ldots, M$, where C_m denotes the m-th class and $\tilde{C}_m$ denotes the complement of C_m (i.e., $\tilde{C}_m = \bigcup_{k=1,\ldots,M, k \neq m} C_k$).

We first obtain the feature vectors (i.e., sparse vectors with respect to a given training dictionary $\boldsymbol{D}$) of training samples and their neighboring pixels by solving the joint sparse recovery problem in (5). Let T be the size of the neighborhood. The extraction of sparse features may be viewed as a projection $\mathcal{P}_l : \mathbb{R}^B \mapsto \mathbb{R}^N$, and there are T such distinct projections $\mathcal{P}_l, l = 1, 2, \ldots, T$. For every pixel $\boldsymbol{y} \in \mathbb{R}^B$, T different features $\boldsymbol{\alpha}_l \in \mathbb{R}^N, l = 1, 2, \ldots, T$ are obtained, as illustrated in Fig. 1 for a 3×3 neighborhood with $T = 9$ (only three features are displayed). For each projection, training features for class C_m correspond to pixels in a neighborhood of training samples known to belong to class C_m. Features for $\tilde{C}_m$ are the sparse vectors associated with neighbors of representative training

For each of the T projections $\mathcal{P}_l$, a pair of N-node discriminative tree graphs $\mathcal{G}_l^p$ and $\mathcal{G}_l^q$, which respectively approximate the class distributions $f(\boldsymbol{\alpha}_l|C_m)$ and $f(\boldsymbol{\alpha}_l|\tilde{C}_m)$, are simultaneously learnt by solving the decoupled MWST problems in (9). The initial disjoint graphs with TN nodes representing the class distribution corresponding to C_m and $\tilde{C}_m$ are then generated by separately concatenating the nodes of $\mathcal{G}_l^p, l = 1, \ldots, T$ and $\mathcal{G}_l^q, l = 1, \ldots, T$, respectively. These graphs with sparse edge structure are then iteratively thickened via boosting [18]. Different pairs of discriminative graphs over the same sets of nodes with different weights are learnt in different iterations, and the newly-learnt edges are used to augment the graphs. The final "thickened" graphs $\mathcal{G}^p$ and $\mathcal{G}^q$ are shown in Fig. 1 (right side).

The process described above (Steps 1-4 in Algorithm 1) is performed offline, and M pairs of discriminative graphs are learnt for the M binary classification problems in a one-against-all manner. The classification of a new test sample is then performed online. Features $\boldsymbol{\alpha}$ are extracted from the test sample $\boldsymbol{y}$ by solving the sparse recovery problem in (5) for the T pixels in the neighborhood centered at $\boldsymbol{y}$. Let $\hat{f}(\boldsymbol{\alpha}|C_m)$ and $\hat{f}(\boldsymbol{\alpha}|\tilde{C}_m)$ denoted the final graphs learnt for C_m and $\tilde{C}_m$


(a) (b)

(c) (d)

(e) (f)

Fig. 2. Classification maps for the AVIRIS Indian Pine data set: (a) Ground truth map. (b) Training set. (c) Test set. (d) Composite kernel SVM (SVM-CK) [7]. (e) Simultaneous Orthogonal Matching Pursuit (SOMP) [12]. (f) Proposed LSGM approach.

respectively. The class label of $\boldsymbol{y}$ is determined as follows:

$$\text{Class}(\boldsymbol{y}) = \arg \max_{m \in \{1,\dots,M\}} \log \left(\frac{\hat{f}(\boldsymbol{\alpha}|C_m)}{\hat{f}(\boldsymbol{\alpha}|\tilde{C}_m)} \right). \tag{10}$$

IV. Experiments and Results

We compare our proposed LSGM approach with two other state-of-the-art approaches using support vector machines (SVM-CK) [7] and the joint sparsity model (SOMP) [12]. The SVM-CK approach, which develops a family of composite kernels that effectively combine both spectral and contextual spatial information for classification using SVM, has been shown [7] to perform better than many competing approaches. Experiments are performed on three benchmark real-world hyperspectral images and classification rates are reported for each class along with overall classification maps. We further investigate the performance as a function of training size.

A. AVIRIS Data Set: Indian Pines

TABLE I
Classification rates for the AVIRIS Indian Pines test set.

Class type	Training	Test	SVM-CK	SOMP	LSGM
Alfalfa	6	48	95.83	87.50	89.58
Corn-notill	144	1290	96.67	95.04	95.27
Corn-min	84	750	90.93	94.67	94.67
Corn	24	210	85.71	92.86	94.76
Grass/pasture	50	447	93.74	89.72	90.60
Grass/trees	75	672	97.32	98.81	99.40
Pasture-mowed	3	23	69.57	91.11	91.11
Hay-windrowed	49	440	98.41	99.13	99.55
Oats	2	18	55.56	0	38.89
Soybeans-notill	97	871	93.80	89.76	90.70
Soybeans-min	247	2221	94.37	96.96	97.43
Soybeans-clean	62	552	93.66	87.93	92.03
Wheat	22	190	99.47	100	100
Woods	130	1164	99.14	99.62	99.62
Building-trees	38	342	87.43	99.47	99.71
Stone-steel	10	85	100	97.65	98.82
Overall	1043	9323	94.86	95.34	96.07

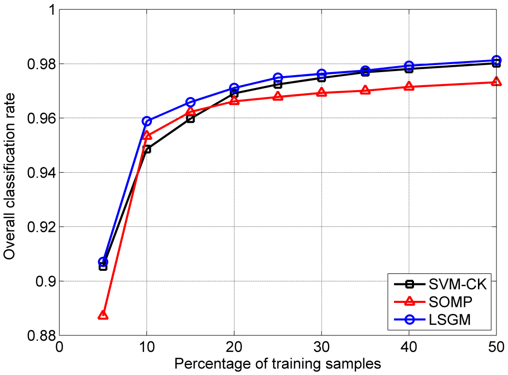


Fig. 3. AVIRIS Indian Pines test set: Performance of different approaches as a function of amount of training provided.

The first hyperspectral image in our experiments is the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Indian Pines image [19]. The AVIRIS sensor generates 220 bands across the spectral range from 0.2 to 2.4 μm, of which only 200 bands are considered by removing 20 water absorption bands [20]. This image has spatial resolution of 20m per pixel and spatial dimension 145 × 145. There are 16 ground-truth classes, listed in Table I. For each class, we randomly choose around 10% of the labeled samples for training and use the remaining 90% for testing. The training and test sets are visually shown in Fig. 2(b) and 2(c) respectively. Classification rates for each class as well as overall accuracy are shown in Table I for the different classifiers. Our LSGM approach outperforms the competing approaches in terms overall classification performance. The improvement over SOMP indicates the benefits of using a discriminative classifier instead of reconstruction residuals for class assignment. For the class of pixels corresponding to Oats, the SOMP approach performs very poorly because a large local neighborhood (9×9) is chosen while the actual class spans only 20 pixels. The proposed LSGM approach performs slightly better owing to the choice of a smaller local neighborhood (3×3).

Fig. 3 shows the variation in overall classification rates as a function of number of training samples provided. As expected, the classification accuracy decreases as training is reduced and for each of the three techniques. That said, the LSGM approach offers a more graceful degradation when compared to the SVM-CK and SOMP algorithms as training is varied from high (half the available pixels) to very low (a few pixels).

B. ROSIS Urban Data Over Pavia, Italy

The next two hyperspectral images, University of Pavia and Center of Pavia, are urban images acquired by the Reflective

TABLE II
CLASSIFICATION RATES FOR THE UNIVERSITY OF PAVIA TEST SET.

Class type	Training	Test	SVM-CK	SOMP	LSGM
Asphalt	548	6304	79.89	59.42	66.56
Meadows	540	18146	84.88	78.25	85.95
Gravel	392	1815	82.26	83.91	86.83
Trees	524	2912	95.09	96.36	96.88
Metal sheets	265	1113	99.82	87.87	98.74
Bare soil	532	4572	93.13	77.56	94.34
Bitumen	375	981	90.21	98.78	99.29
Bricks	514	3364	93.01	89.15	94.50
Shadows	231	795	95.72	92.20	95.72
Overall	3921	40002	87.11	78.74	86.28

TABLE III
CLASSIFICATION RATES FOR THE CENTER OF PAVIA TEST SET.

Class type	Training	Test	SVM-CK	SOMP	LSGM
Water	745	64533	97.46	99.32	99.44
Trees	785	5722	93.08	92.38	92.96
Meadow	797	2094	97.09	95.46	96.99
Brick	485	1667	77.02	85.66	87.46
Soil	820	5729	98.39	96.37	97.59
Asphalt	678	6847	94.32	93.81	94.51
Bitumen	808	6479	97.50	94.68	97.05
Tile	223	2899	99.83	99.69	99.90
Shadow	195	1970	99.95	98.68	99.19
Overall	5536	97940	96.93	97.81	98.20

Optics System Imaging Spectrometer (ROSIS). The ROSIS sensor generates 115 spectral bands ranging from 0.43 to 0.86 m and has a spatial resolution of 1.3 m per pixel. The University of Pavia image consists of 610×340 pixels, each having 103 bands with the 12 noisiest bands removed. About 9% of all labeled data are used as training and the rest are used for testing. The third image, Center of Pavia, is the other urban image collected by the ROSIS sensor over the center of Pavia City. This image consists of 1096×492 pixels, each having 102 spectral bands after 13 noisy bands are removed. For this image, about 5% of the labeled data are used as training samples. Classification rates for the two ROSIS images are provided in Tables II and III respectively. In Table II the SVM-CK technique performs marginally better than LSGM in the sense of overall classification accuracy. However, for most individual classes LSGM does better and particularly in cases where training sample size is smaller. In Table III, LSGM performs better than SOMP as well as SVM-CK.

V. CONCLUSION

In this letter, we propose a principled graphical model-based framework to exploit contextual correlation information for hyperspectral image classification. Our approach extends recent work in the area of sparsity-based HSI classification, wherein the spectral signature of each pixel can be approximately represented by a sparse linear combination of training pixels from all available classes. Sparse feature vectors corresponding to pixels in a local spatial neighborhood are obtained by solving a joint sparsity recovery problem. The statistical correlations between these local sparse features are then explicitly learnt in a discriminative setting via probabilistic graphical models. Experiments on benchmark hyperspectral images reveal the benefits of our proposed approach over competing state-of-the-art schemes, and over a range of scenarios corresponding to varying training set sizes.

ACKNOWLEDGMENT

The authors would like to thank the University of Pavia and the HySenS project for kindly providing the ROSIS images of University of Pavia and Center of Pavia.

REFERENCES

[1] M. Borengasser, W. S. Hungate, and R. Watkins, *Hyperspectral Remote Sensing - Principles and Applications*. Boca Raton, FL, USA: CRC Press, 2008.

[2] B. Datt, T. R. McVicar, T. G. V. Niel, D. L. B. Jupp, and J. S. Pearlman, "Preprocessing EO-1 hyperion hyperspectral data to support the application of agricultural indexes," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 6, pp. 1246–1259, Jun. 2003.

[3] D. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 29–43, Jan. 2002.

[4] V. N. Vapnik, *The nature of statistical learning theory*. Springer, 1995.

[5] 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.

[6] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVM for the semisupervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3363–3373, Nov. 2006.

[7] G. Camps-Valls, L. Gomez-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.

[8] F. Bovolo, L. Bruzzone, and M. Marconcini, "A novel context-sensitive SVM for classification of remote sensing images," in *Proc. of IEEE International Geoscience and Remote Sensing Symposium*, Denver, Colorado, Jul. 2006, pp. 2498–2501.

[9] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, Aug. 2009.

[10] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, to appear.

[11] J. Wright, A. Y. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

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

[13] V. Y. F. Tan, S. Sanghavi, J. W. Fisher, and A. S. Willsky, "Learning graphical models for hypothesis testing and classification," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5481–5495, Nov. 2010.

[14] Y. Freund and R. E. Schapire, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, Sep. 1999.

[15] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[16] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing*, vol. 86, pp. 572–588, Mar. 2006.

[17] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, Mar. 1968.

[18] U. Srinivas, V. Monga, and R. G. Raj, "Automatic target recognition using discriminative graphical models," in *Proc. IEEE Intl. Conf. Image Processing*, 2011, pp. 33–36.

[19] "AVIRIS NW Indiana's Indian Pines 1992 Data Set." [Online]. Available: http://cobweb.ecn.purdue.edu/biehl/MultiSpec/documentation.html

[20] J. A. Gualtieri and R. F. Cromp, "Support vector machines for hyperspectral remote sensing classification," *Proc. SPIE*, vol. 3584, pp. 221–232, Jan. 1998.