Hyperspectral Image Classification Using Gaussian Mixture Models and Markov Random Fields

Wei Li, *Member, IEEE,* Saurabh Prasad, *Member, IEEE,* and James E. Fowler, *Senior Member, IEEE*

Abstract—The Gaussian mixture model is a well-known classification tool that captures non-Gaussian statistics of multivariate data. However, the impractically large size of the resulting parameter space has hindered widespread adoption of Gaussian mixture models for hyperspectral imagery. To counter this parameter-space issue, dimensionality reduction targeting the preservation of multimodal structures is proposed. Specifically, locality-preserving nonnegative matrix factorization, as well as local Fisher's discriminant analysis, is deployed as preprocessing to reduce the dimensionality of data for the Gaussian-mixture-model classifier, while preserving multimodal structures within the data. In addition, the pixel-wise classification results from the Gaussian mixture model are combined with spatial-context information resulting from a Markov random field. Experimental results demonstrate that the proposed classification system significantly outperforms other approaches even under limited training data.

Index Terms—Gaussian mixture model (GMM), hyperspectral classification, Markov random field (MRF), nonnegative matrix factorization.

I. INTRODUCTION

MANY classification methods, such as support vector machines (SVMs) [1], employ the high-dimensional spectral signatures of hyperspectral data as input features. However, Gaussian mixture models (GMMs) [2], although a powerful supervised classification tool, have not been as popular within the hyperspectral community owing to the tremendous size of the resulting parameter space. For example, for d-dimensional data with a K-component GMM, the parameter space is on the order of $K(1 + d(d - 1)/2 + d)$. Even for a reasonably sized hyperspectral dataset with, say, $d = 100$ and $K = 5$, the resulting parameter space has a dimensionality of 25 255—clearly impractical when using limited ground-truth hyperspectral data. Popular dimensionality-reduction techniques, such as linear discriminant analysis (LDA), may destroy the underlying multimodal structure of hyperspectral data; consequently, GMM classifiers are

expected to be ineffective when such structure-destroying dimensionality reduction is used. In contrast, in this letter, we consider two alternatives for structure-preserving dimensionality reduction—locality-preserving nonnegative matrix factorization (LPNMF) [3] and local Fisher's discriminant analysis (LFDA) [4]—as a preprocessing for GMM classifiers.

Nonnegative matrix factorization (NMF)—recently popular for blind-source-separation hyperspectral unmixing [5] since it allows only additive operations—has received attention in the literature because of its physical parts-based representation. LPNMF combines the advantages of NMF with those of locality-preserving projection (LPP) [6], leading to an intuitive representation using only additive operations, all the while preserving the intrinsic geometric structure in the embedded subspace. On the other hand, LFDA—which has been introduced for hyperspectral data in our previous work [7]—is an extension of conventional LDA such that the class-distributions are no longer restricted to be unimodal Gaussian. LFDA obtains good between-class separation in the projection, while preserving the within-class local structure at the same time. Ultimately, both LPNMF and LFDA have been shown to be effective at preserving multimodal statistical distributions.

Most classifiers, such as SVMs and GMMs, do not consider the spatial correlation between neighboring pixels. Yet, for hyperspectral imagery (HSI), it is highly probable that two adjacent pixels belong to the same class. Recently, several spatial-spectral classifiers—which consider spatial features and spectral signatures simultaneously—have been developed [8], [9]. In particular, a Markov random field (MRF) [8], [9] is a common model for incorporating spatial-context information into image-classification problems. The MRF model has the advantage that it can work with many probability distributions in the sense that it can capture complex spatial interdependencies among neighboring pixels. In particular, [9] uses multinomial logistic regression (MLR) to learn posterior probability distributions from which posterior probabilities are combined with the MRF model to extract spatial correlation among pixels.

In this letter, we propose a GMM-MRF classification—specifically, a GMM classifier, based on a low-dimensional feature space induced by either LPNMF or LFDA, is first employed to learn posterior probabilities from the spectral signatures of the data. Afterward, an MRF model integrates the results from the classifier to exploit spatial-context information. Via appropriate preprocessing, a GMM classifier is able to accurately learn class-conditional statistics in the corresponding reduced-dimensional subspace. The posterior probabilities from this model are incorporated into the MRF

W. Li is with the University of California, Davis, CA 95616 USA (e-mail: liwei089@ieee.org).

S. Prasad is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77204-4005 USA (e-mail: saurabh.prasad@ieee.org).

J. E. Fowler is with the Department of Electrical and Computer Engineering and the Geosystems Research Institute, Mississippi State University, Starkville, MS 39762 USA (e-mail: fowler@ece.msstate.edu).

model, which is, in turn, expected to result in superior classification performance.

II. BACKGROUND

Consider the HSI with pixels $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ in $\mathbb{R}^d$ (d-dimensional feature space) and class labels $\boldsymbol{\omega} = \{\omega_i\}_{i=1}^n \in \{1, 2, \ldots, C\}$, where C is the number of classes, and n is the total number of pixels. According to the Bayes rule, the posterior probability of class labels $\boldsymbol{\omega}$ given data $\mathbf{X}$ is

$$\Pr(\boldsymbol{\omega}|\mathbf{X}) = \frac{\Pr(\mathbf{X}|\boldsymbol{\omega})\Pr(\boldsymbol{\omega})}{\Pr(\mathbf{X})} \quad (1)$$

where $\Pr(\mathbf{X})$ is a constant that does not affect the final result, $\Pr(\boldsymbol{\omega})$ can be viewed as the spatial prior, and $\Pr(\mathbf{X}|\boldsymbol{\omega})$ is the likelihood function that is estimated from the data distribution. The $\boldsymbol{\omega}$ that maximizes $\Pr(\boldsymbol{\omega}|\mathbf{X})$ in (1) is referred to as the maximum *a posteriori* (MAP) estimation of $\boldsymbol{\omega}$

$$\hat{\boldsymbol{\omega}} = \arg \max_{\omega} \sum_{i=1}^n \left(\log \Pr(\mathbf{x}_i|\omega_i) + \log \Pr(\boldsymbol{\omega}) \right). \quad (2)$$

A. GMMs

In this letter, we model the likelihood function $\Pr(\mathbf{X}|\boldsymbol{\omega})$ as a Gaussian mixture. The spectral response of HSI can be affected by many factors, such as differences in illumination conditions, geometric features of material surfaces, and atmospheric effects. As a consequence, it is reasonable to expect that statistical distributions of HSI classes pose a complicated multimodal structure. The GMM model is hence a natural choice for this letter.

For GMMs, a probability density function for the likelihood function is written as the sum of K Gaussian components or modes, $\Pr(\mathbf{X}|\boldsymbol{\omega}) = \sum_{k=1}^K \alpha_k \mathcal{N}(\mathbf{X}, \mu_k, \Sigma_k)$, where $\mathcal{N}(\mathbf{X}, \mu_k, \Sigma_k)$ represents the kth Gaussian component of the mixture, K is the number of mixture components, and α_k, μ_k, and Σ_k are the mixing weight, mean, and covariance matrix of the kth component, respectively. The parameters for the mixture model can be estimated by the expectation-maximization algorithm. Estimating an appropriate number of modes, K, is an important aspect of designing a GMM classifier, and the Bayes information criterion (BIC) [10] is commonly used for this task.[1] A detailed description of our approach to GMM classifiers for HSI can be found in [7].

B. MRFs

MRFs have been applied recently [8], [9], [11] to extract spatial-context information for HSI, in which spatially adjacent (homogeneous) objects tend to be of the same class. In [9], the spatial prior is defined as $\Pr(\boldsymbol{\omega}) = Z^{-1} \exp(U(\boldsymbol{\omega}))$, where Z is a normalizing constant, $U(\boldsymbol{\omega}) = \nu \sum_{(i,j) \in \Xi} \delta(\omega_i - \omega_j)$ is an energy function that is a sum of clique potentials $\delta(\omega_i - \omega_j)$ over all possible cliques Ξ, and ν is a parameter of smoothness. In an MRF, a clique is defined as a set of pixels in which distinct sites are neighbors. In addition, the property of a pixel is determined by only its neighborhoods, while the associated density is a Gibbs distribution. In this letter, we use the MRF model from [9]; thus, based on a multilevel logistic prior, the MRF computes the MAP in (2) via a graph-cut algorithm. A detailed description can be found in [9].

[1]BIC is not the only method for estimating the number of modes for GMM. A number of other methods can be found in the literature.
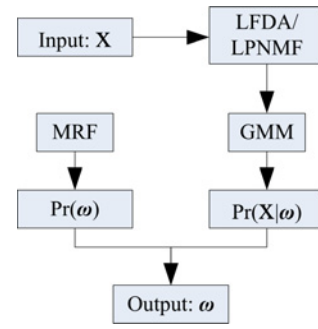

Fig. 1. Proposed GMM-MRF classification algorithm.

C. LFDA

LFDA [4] provides supervised dimensionality reduction designed to handle multimodal, non-Gaussian distributions. LFDA combines the properties of LDA and LPP [6] such that the *local* between-class $S^{(\mathrm{lb})}$ and within-class $S^{(\mathrm{lw})}$ scatter matrices are

$$S^{(\mathrm{lb})} = \frac{1}{2} \sum_{i,j=1}^n W_{i,j}^{(\mathrm{lb})} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \quad (3)$$

$$S^{(\mathrm{lw})} = \frac{1}{2} \sum_{i,j=1}^n W_{i,j}^{(\mathrm{lw})} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \quad (4)$$

where $W^{(\mathrm{lb})}$ and $W^{(\mathrm{lw})}$ are $n \times n$ matrices defined as

$$W_{i,j}^{(\mathrm{lb})} = \begin{cases} A_{i,j}(1/n - 1/n_l), & \text{if } y_i = y_j = l \\ 1/n, & \text{if } y_i \neq y_j \end{cases} \quad (5)$$

$$W_{i,j}^{(\mathrm{lw})} = \begin{cases} A_{i,j}/n_l, & \text{if } y_i = y_j = l \\ 0, & \text{if } y_i \neq y_j \end{cases} \quad (6)$$

where n_l is the number of available training samples for the lth class, $\sum_{l=1}^C n_l = n$, and A is defined as the affinity matrix. Maximizing Fisher's ratio as defined using the local scatter matrices, we have that the transformation matrix

$$\Phi_{\mathrm{LFDA}} = \arg \max_{\Phi_{\mathrm{LFDA}}} \mathrm{tr} \left[\left(\Phi_{\mathrm{LFDA}}^\top S^{(\mathrm{lw})} \Phi_{\mathrm{LFDA}} \right)^{-1} \Phi_{\mathrm{LFDA}}^\top S^{(\mathrm{lb})} \Phi_{\mathrm{LFDA}} \right] \quad (7)$$

is given by the generalized eigenvalue problem $S^{(\mathrm{lb})} \Phi_{\mathrm{LFDA}} = \Lambda S^{(\mathrm{lw})} \Phi_{\mathrm{LFDA}}$, where Λ is the diagonal eigenvalue matrix. LFDA is a localized variant of LDA since it does not force distant data pairs in the same class to be close, thereby preserving the local neighborhood even when the data distribution is complex. LFDA has been shown to yield performance better than that of LDA and LPP for dimensionality reduction for non-Gaussian or multimodal data.

D. LPNMF

LPNMF [3] was originally proposed for document clustering. The objective function is

$$\mathcal{O} = \sum_{i=1}^d \sum_{j=1}^n \left(x_{i,j} \log \frac{x_{i,j}}{y_{i,j}} - x_{i,j} + y_{i,j} \right) + \lambda \mathcal{R} \quad (8)$$

where $\mathbf{Y} = \{y_{i,j}\} = \mathbf{U}\mathbf{V}^\top$, and λ is a regularization parameter. The first part of (8) is the usual objective function of NMF that uses Kullback–Leibler divergence [3] between $\mathbf{X}$ and $\mathbf{Y}$.

TABLE I
CLASSIFICATION ACCURACY (%) FOR THE UNIVERSITY OF PAVIA DATASET

| Class	# samples		Classification algorithms						
	Train	Test	SVM	SVM-MRF	MLR-MRF	LFDA-GMM	LFDA-GMM-MRF	LPNMF-GMM	LPNMF-GMM-MRF
Asphalt	60	6631	81.48	91.27	97.52	85.21	96.71	87.34	98.23
Bare soil	60	18 649	82.97	89.59	93.98	78.73	88.21	85.86	93.74
Bitumen	60	2099	87.18	96.09	58.54	73.20	83.11	78.77	86.89
Bricks	60	3064	95.89	98.40	72.94	95.90	97.55	96.86	98.60
Gravel	60	1345	99.55	99.63	99.90	100	99.91	100	99.95
Meadows	60	5029	88.43	95.82	97.42	85.28	98.80	88.19	99.44
Metal sheets	60	1330	92.78	96.02	82.55	69.08	65.72	82.86	88.26
Shadows	60	3682	78.14	88.24	87.44	71.05	86.15	74.51	90.60
Trees	60	947	100	100	99.23	96.89	97.43	97.22	97.67
Overall accuracy			85.30	92.16	86.36	81.54	90.88	86.41	**94.96**
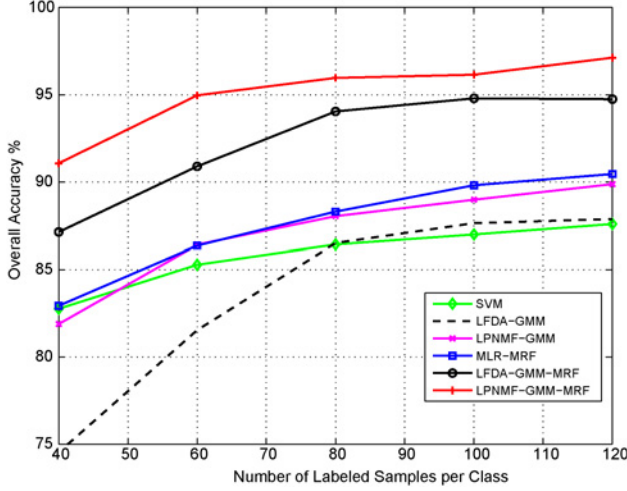
Fig. 2. Overall accuracy as a function of the number of labeled training samples per class for the University of Pavia dataset. Standard deviations of the accuracies (%) for 40, 60, 80, 100, and 120 labeled samples for LFDA-GMM-MRF are 0.1357, 0.1828, 0.0171, 0.0466, and 0.0649, respectively; the corresponding standard deviations for LPNMF-GMM-MRF are 0.2243, 0.0215, 0.0142, 0.0265, and 0.0037, respectively.
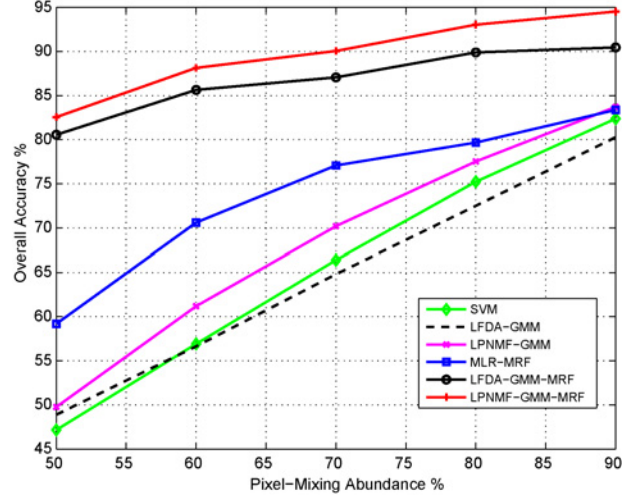

Fig. 3. Overall accuracy as a function of the pixel-mixing abundance for the University of Pavia dataset using 60 labeled training samples per class.

On the other hand, $\mathcal{R}$ is used to force a geometric locality constraint among points in the reduced-dimensional subspace $\mathbf{V}$ ($t < d$), i.e.,

$$\mathcal{R} = \frac{1}{2} \sum_{j,s=1}^{n} \sum_{q=1}^{t} \left(v_{j,q} \log \frac{v_{j,q}}{v_{s,q}} - v_{s,q} + v_{s,q} \log \frac{v_{s,q}}{v_{j,q}} \right) \mathbf{W}_{j,s} \quad (9)$$

where $\mathbf{W}$ is an edge-weight matrix measuring the distance between points in the original space $\mathbf{X}$. This matrix, defined from the theory of LPP [6], is employed to preserve the intrinsic geometry of the data distribution. The following multiplicative rules are used to minimize the function $\mathcal{O}$ and to estimate the matrices $\mathbf{U}$ and $\mathbf{V}$:

$$u_{i,q} \leftarrow u_{i,q} \frac{\sum_j (x_{i,j} v_{j,q}) / \sum_q (u_{i,q} v_{j,q})}{\sum_j v_{j,q}} \quad (10)$$

$$\mathbf{v}_q \leftarrow \left[\sum_i u_{i,q} \mathbf{I} + \lambda \mathbf{L} \right]^{-1} \begin{bmatrix} v_{1,q} \sum_i \frac{x_{i,1} u_{i,q}}{\sum_q (u_{i,q} v_{1,q})} \\ v_{2,q} \sum_i \frac{x_{i,2} u_{i,q}}{\sum_q (u_{i,q} v_{2,q})} \\ \vdots \\ v_{n,q} \sum_i \frac{x_{i,n} u_{i,q}}{\sum_q (u_{i,q} v_{n,q})} \end{bmatrix} \quad (11)$$

where $\mathbf{v}_k$ is the kth column of $\mathbf{V}$, $\mathbf{I}$ is the $n \times n$ identity matrix, and $\mathbf{L}$ is the graph Laplacian of $\mathbf{W}$. A more detailed description of this technique can be found in [3].

Compared to NMF, the locality-preserving constraint of LP-NMF guarantees that the local manifold structure is maintained from the original high-dimensional space. The constraint measures the similarity among local neighborhoods via a graph Laplacian that describes the latent distribution structures. For HSI, the statistical (geometrical) structure is complex, and, in [7] and [12], we have demonstrated that imposing an additional locality-preserving constraint is important for feature extraction in hyperspectral applications, providing a tighter distribution and smaller Bayes error in the induced feature space. Both LFDA and LPNMF have been shown to be effective at preserving multimodal statistical distributions for the GMM classifier.

III. GMM-MRF CLASSIFICATION

Fig. 1 depicts a flowchart for the proposed GMM-MRF classification.

1) Supervised LFDA combined with a GMM classifier (LFDA-GMM) is employed to learn the posterior probability distribution $\Pr(\mathbf{X}|\boldsymbol{\omega})$ from spectral information. Spatial prior $\Pr(\boldsymbol{\omega})$ is calculated by MRF models using results from LFDA-GMM. The final classification label
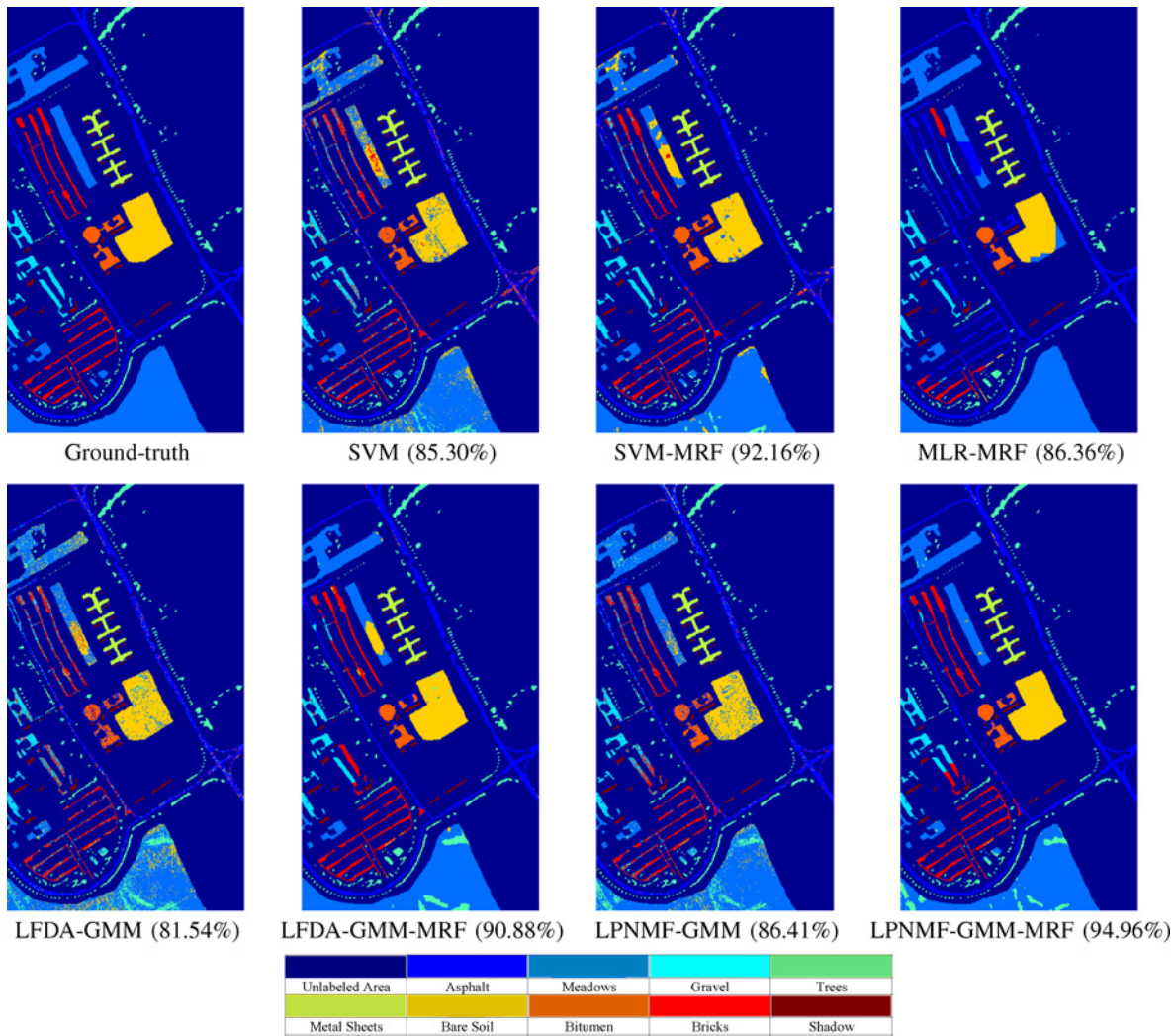
Fig. 4.   Classification maps for the University of Pavia dataset.

is obtained via (2). This algorithm is called LFDA-GMM-MRF.

2) As in LFDA-GMM-MRF, unsupervised LPNMF combined with a GMM classifier (LPNMF-GMM) is applied for learning $\Pr(\mathbf{X}|\boldsymbol{\omega})$ from spectral information. The MRF model integrates the results to produce spatial prior $\Pr(\boldsymbol{\omega})$. The final label is assigned via (2). This algorithm is called LPNMF-GMM-MRF.

IV. EXPERIMENTS AND ANALYSIS

In this letter, the experimental hyperspectral dataset employed was collected by the reflective optics system imaging spectrometer (ROSIS) sensor. The image, covering the city of Pavia, Italy, was collected under the HySens project managed by the German Aerospace Agency. The data have a spectral coverage from 0.43 to 0.86 μm, and a spatial resolution of 1.3 m. The scene used in our experiment is the university area that has 103 spectral bands with a spatial coverage of 610×340 pixels. There are nine classes for the dataset, and the number of training and testing samples are shown in Table I.

We evaluate the efficacy of the proposed GMM-MRF classification scheme for hyperspectral images. The proposed approaches—LFDA-GMM-MRF and LPNMF-GMM-MRF—are compared with conventional algorithms, including a prob-

ability SVM classifier (LIBSVM tools[2]) SVM-MRF, MLR-MRF,[3] and our previously proposed LFDA-GMM [7]. We have developed training and testing data subsets for parameter tuning. All classification techniques are implemented using optimal parameters, such as the Gaussian kernel parameter for the SVM (set here to 0.6).

LFDA and LPNMF can be viewed as preprocessing for the proposed methods, which implies that reduced dimensionality is an important parameter for LFDA-GMM-MRF and LPNMF-GMM-MRF classification systems. In this letter, the optimal dimensionality of LFDA and LPNMF is found to be 11 and 17, respectively. We use implementations of LFDA and LPNMF from the authors of [3] and [4], respectively.[4,5]

Classification accuracies of the proposed approaches, as well as those for current state-of-the-art techniques, are shown in Table I for the experimental dataset for each class. The number of available training samples is often insufficient for hyperspectral images. Thus, we report the classification accuracy of different techniques as a function of the training-sample size in Fig. 2 (the whole labeled scene is used as

testing). To avoid any bias, the experiments are repeated 20 times, and we report the average classification accuracy. From Fig. 2, it can be seen that the proposed LFDA-GMM-MRF and LPNMF-GMM-MRF significantly outperform other classification techniques. For further purposes of illustration, we show the classification maps resulting from the various approaches in Fig. 4.

In many situations, the spatial resolution of HSI is poor, and pixels are mixed from multiple classes. Thus, we also compose artificially mixed hyperspectral pixels according to a linear mixing model from background classes. For example, a pixel-mixing abundance of 70% indicates that 30% of background signatures were mixed linearly with 70% of the target class. An abundance of 100% implies that pure pixels are employed without any mixing. Here, target class simply refers to the true class of the pixel currently being classified/tested. The background signatures/pixels used for mixing the target class are gathered (with uniform weights) from across all the other classes. Fig. 3 illustrates the accuracy as a function of pixel-mixing abundance for the experimental dataset (we implement pixel mixing for both unlabeled and labeled areas for the scene—unlabeled data are viewed as a new class mixed with other labeled classes). From Fig. 3, it can be seen that LPNMF-GMM-MRF has the best performance compared to others, even under severe pixel-mixing conditions.

Finally, we compare the computational cost of the classification methods using the same experimental data. All experiments are carried out using MATLAB on a 3.2-GHz machine with 5.8 GB of RAM. The execution times (in seconds) are 13 for SVM, 9 for MLR-MRF, 491 for LFDA-GMM, 494 for LFDA-GMM-MRF, 466 for LPNMF-GMM, and 469 for LPNMF-GMM-MRF. As we can see, the MRF post-processing amounts to only an extra 3 s for the experimental data.

V. CONCLUSION

In this letter, we considered a GMM classifier based on an LFDA- and LPNMF-induced feature subspace for HSI classification. The LFDA and LPNMF dimensionality-reduction techniques have superior locality-preserving properties and preserve the local manifold structure for hyperspectral data with complex distributions, from which the GMM classifier is able to accurately learn the class-conditional statistics. The GMM-based pixel-wise classification results were subsequently followed by an MRF-based technique that integrated the spatial-context information for the image. Experimental results showed that LFDA and LPNMF provide effective dimensionality reduction, and LPNMF performed especially well under a pixel-mixing scenario. The proposed GMM-MRF classification strategy outperformed other traditional techniques over a wide range of operating conditions. GMM-based probabilistic pixel-wise classification works effectively when used in an MRF model. As a final remark, we noted that, in [13], spectral-spatial information was investigated for LFDA-GMM. Consequently, we planned to study the benefit of such information for preserving multimodal statistical distributions for GMM classifiers in subsequent work.

REFERENCES

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

[2] A. Berge and A. H. S. Solberg, "Structured Gaussian components for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3386–3396, Nov. 2006.

[3] D. Cai, X. He, X. Wang, H. Bao, and J. Han, "Locality preserving nonnegative matrix factorization," in *Proc. Int. Joint Conf. Artif. Intell.*, Pasadena, CA, USA, Jul. 2009, pp. 1010–1015.

[4] M. Sugiyama, "Local Fisher discriminant analysis for supervised dimensionality reduction," in *Proc. Int. Conf. Mach. Learning*, Pittsburgh, PA, USA, Jun. 2006, pp. 905–912.

[5] S. Jia and Y. Qian, "Constrained nonnegative matrix factorization for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 1, pp. 161–173, Jan. 2009.

[6] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing System*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA, USA: MIT Press, 2004.

[7] W. Li, S. Prasad, J. E. Fowler, and L. M. Bruce, "Locality-preserving dimensionality reduction and classification for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1185–1198, Apr. 2012.

[8] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, "SVM- and MRF-based method for accurate classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 736–740, Oct. 2010.

[9] 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.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.

[10] G. Schwarz, "Estimation the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[11] X. Huang and Z. Liao, "A labeling scheme based on Markov random fields and Gaussian mixture models for hyperspectral images," in *Proc. 7th Int. Conf. Mach. Learning Cybern.*, Kunming, China, Jul. 2008, pp. 3619–3624.

[12] W. Li, S. Prasad, J. E. Fowler, and L. M. Bruce, "Locality-preserving discriminant analysis in kernel-induced feature spaces for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 5, pp. 894–898, Sep. 2011.

[13] W. Li, S. Prasad, Z. Ye, J. E. Fowler, and M. Cui, "Locality-preserving discriminant analysis for hyperspectral image classification using local spatial information," in *Proc. Int. Geosci. Remote Sens. Symp.*, Munich, Germany, Jul. 2012, pp. 4134–4137.