ADAPTIVE COMPRESSED CLASSIFICATION FOR HYPERSPECTRAL IMAGERY

Jürgen Hahn *Simon Rosenkranz* *Abdelhak M. Zoubir*

Signal Processing Group
Technische Universität Darmstadt
Merckstraße 25, 64283 Darmstadt, Germany
E-mail: {jhahn, srosenkranz, zoubir}@spg.tu-darmstadt.de

ABSTRACT

Hyperspectral imaging (HSI) is a useful tool for the classification of vast areas. High accuracy is achieved by means of spectral information for each pixel, which inherently leads to a huge amount of data and, thus, requires costly processing. We present an Adaptive Compressed Classification (ACC) framework for HSI that allows a compressive acquisition of the scene of interest. Since classification is performed in the compressive domain, expensive reconstruction is avoided, significantly reducing computational requirements. For ACC, we propose an adaptive probabilistic approach to optimize the measurement and basis matrices. Based on real data sets, we show that Compressed Classification yields high classification accuracy close to results obtained for the complete data. Using the proposed adaptive approach, even higher accuracies are achieved in all tested cases.

Index Terms— Compressed Sensing, classification, hyperspectral imaging

1. INTRODUCTION

Hyperspectral imaging (HSI) is an emerging optical sensing technique that allows to simultaneously capture hundreds of bands of the visible and infrared light range. Since each material possesses a characteristic spectrum or signature, scene analysis and classification based on HSI becomes fairly easy [1]. Today, HSI is commonly used in remote sensing, e.g. in agriculture, urban mapping, or security applications [2]. Thanks to efficient algorithms and advanced hardware, HSI finds new applications in many other fields, such as e.g. in nutrition analysis [3].

Due to the high number of bands, HSI is still considered to be computationally expensive. Besides high storage demands, post-processing and analysis are often time-consuming tasks, requiring costly computation resources. In order to solve this problem, there has been a lot of research in the field of dimensionality reduction, also known as band selection. Most algorithms exploit the correlation between the bands to select most important candidates [4]. Recent research is concerned with the utilization of clustering algorithms for this purpose, yielding higher classification accuracy with less bands [5, 6]. A rather new approach is to directly capture the hypespectral image in a lower dimensional representation. This can be achieved by Compressed Sensing (CS), as shown in [7, 8, 9]. In this work, we will build on [7], where CS is used to efficiently capture hyperspectral images. Based on [10], we will introduce a Compressed Classification (CC) framework for HSI that allows the capture of samples in the Compressed domain, in which they are directly classified. The advantage of this method is that a costly reconstruction of the HSI at the base station can be skipped, remarkably reducing computational cost. Further, approximation errors due to a poor reconstruction are avoided.

In this contribution, we will introduce (i) a system design for pixelwise Compressed Classification in hyperspectral imaging. Furthermore, we will investigate how the (ii) basis transform matrix and the (iii) sensing matrices required for the capturing process can be optimized with respect to the training samples.

In Section 2, we briefly revisit the idea of Compressed Classification. In Section 3, a theoretical system for the compressed acquisition of the images is proposed. In Section 4, we explain how to optimize the measurement and basis matrices with respect to the training samples. Finally, results based on two widely used data sets are presented in Section 5 and a short conclusion is drawn in Section 6.

2. COMPRESSED CLASSIFICATION BASICS

Compressed Sensing (CS) is a technique used to capture the information of a signal $\mathbf{x} \in \mathbb{R}^D$ by taking only $M < D$ linear measurements [11, 12, 13],

$$\mathbf{x_c} = \mathbf{\Phi x}, \tag{1}$$

where $\mathbf{\Phi}$ is our $M \times D$ sensing or measurement matrix and $\mathbf{x_c} \in \mathbb{R}^M$ is the measurement vector. To recover $\mathbf{x}$ from only M measurements, $\mathbf{x}$ is assumed to be sparse in some domain, i.e. $\mathbf{x} = \mathbf{\Psi s}$ with basis $\mathbf{\Psi} \in \mathbb{R}^{D \times D}$ and sparse coefficient vector $\mathbf{s} \in \mathbb{R}^D$. The reconstruction is formulated as an optimization problem where the sparsity is to be maximized.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Usually, this problem cannot be solved directly, meaning that time-consuming iterative algorithms are required.
In hyperspectral imaging, we are mainly interested in the content of the image, i.e. not in the image itself. Thus, it is desirable to skip the step of reconstructing the image and directly classify in the measurement domain. For this purpose, Davenport *et al.* presented the Smashed filter [10]. The idea is to basically transform the training data into the measurement domain and use a matched filter for the classification.

Calderbank *et al.* [14] show how to use Support Vector Machines [15, 16] in the measurement domain. Therefore, a discriminant function $y_c(\mathbf{x_c})$,

$$y_c(\mathbf{x_c}) = \mathbf{w_c}\mathbf{x_c} + b_c, \tag{2}$$

which is able to separate two classes, is trained on Compressed training samples. The parameters $\mathbf{w_c} \in \mathbb{R}^M$ and b_c are estimated during training by maximizing the margin between the decision boundary and the closest training samples and simultaneously minimizing the empirical hinge loss on the Compressed training samples. The sign of y_c indicates the predicted class label of $\mathbf{x}_c$. Note that there are techniques that allow discrimination between more than two classes, e.g. one-vs-one classification, which can also be used here.

3. SYSTEM DESIGN

Let $\mathbf{X} \in \mathbb{R}^{N_x \times N_y \times D}$ denote the hyperspectral image of size $N_x \times N_y \times D$ where N_x is the number of pixels in cross-track and N_y in track direction. The number of bands is denoted by D. A pixelwise Compressed Classification requires an imaging system that preserves the spatial domain of the scene. The acquisition is performed in the spectral domain by taking M measurements of all pixels $\mathbf{x}_{i,j} \in \mathbb{R}^D$ with $1 \leq i \leq N_x, 1 \leq j \leq N_y$, i.e. we obtain an $N_x \times N_y \times M$ image.

An imaging system based on Compressed Classification could be designed as follows. The scene of interest is measured in scan-lines similar to the push-broom technique known from HSI in remote sensing [17]. For remote sensing, this system could be mounted on an airborne device. As shown in Fig. 1, the incident light of a scan-line is split, using a dispersive element. Similar to [18], the light components are reflected from a Digital Mirror Device (DMD) with $D \times N_x$ elements and are detected by a single sensor. Note that we cannot compressively capture all pixels of the scan line at once as the spatial information would be mixed. A possibility to preserve the cross-track resolution is to spread the spatial information into the time domain. For this purpose, only one DMD line is activated at a time, meaning that the other lines are set to zero. Hence, a measurement can be considered as the inner product of the pattern represented by a DMD line and the signature of the pixel. Having received a measurement, we permute the active line with the remaining zero-lines to capture the next pixel. For each scan-line, this

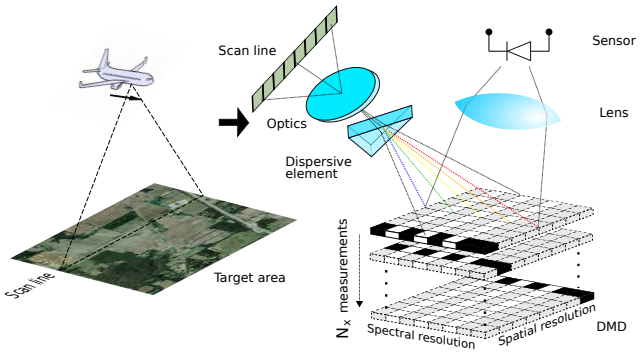


Fig. 1: Design of a hyperspectral imaging system for the use of Compressed Classification

procedure is continued N_x times. Additionally, M measurements of each pixel are obtained by choosing M different DMD patterns in each time step.

4. OPTIMIZATION OF THE ACQUISITION PROCESS

In Compressed Classification, training samples suitable for the scene of interest are required. Thus, those samples can also be used to optimize the sensing matrix $\mathbf{\Phi}$ and basis matrix $\mathbf{\Psi}$. In the following, we explain how to adapt those matrices. According to Compressed Sensing theory, $\mathbf{\Phi}$ is designed such that it is approximately incoherent to the basis matrix $\mathbf{\Psi}$ of the data [11, 12, 13].

We propose to split the adaption process into two steps. First, an underlying basis matrix $\hat{\mathbf{\Psi}}$ is estimated from the training data. Second, we adapt a sensing matrix $\hat{\mathbf{\Phi}}$ to the estimated basis $\hat{\mathbf{\Psi}}$. In the sequel, $\mathbf{a}_i$ represents the i^{th} column of a matrix $\mathbf{A}$.

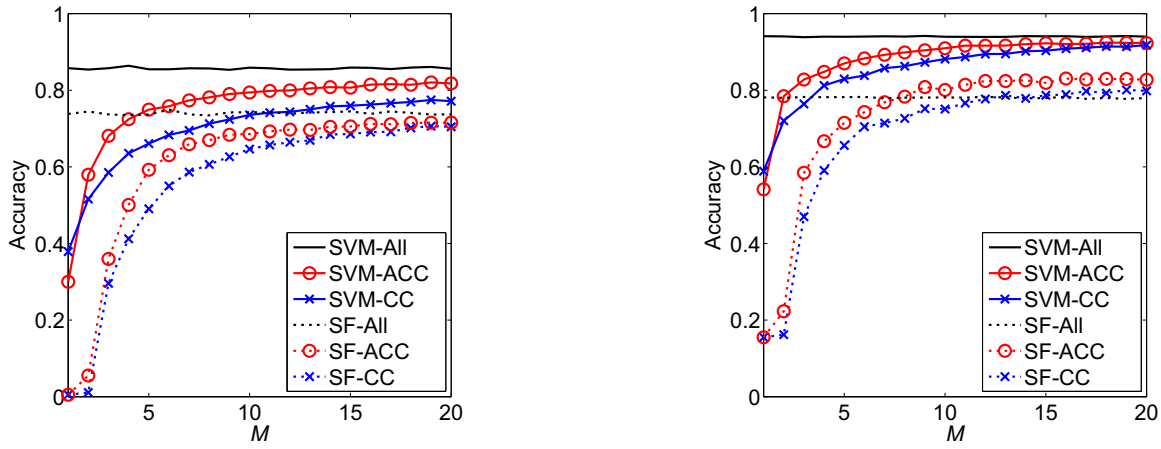
4.1. Learning the basis

Suppose we are given N training samples $\mathbf{x}_{\mathrm{T},n} \in \mathbb{R}^D$ with $n = 1, ..., N$ that we stack into an $D \times N$ matrix $\mathbf{X}_\mathrm{T} = \left[\mathbf{x}_{\mathrm{T},1}, \ldots, \mathbf{x}_{\mathrm{T},N}\right]$ where the corresponding (unknown) sparse coefficients are denoted by $\mathbf{S} \in \mathbb{R}^{D \times N}$. In order to estimate an underlying basis $\hat{\mathbf{\Psi}} \in \mathbb{R}^{D \times J}$ from the training data, with $J \leq D$ denoting the number of retained basis vectors, we propose a probabilistic modeling of the quantities so that a maximum a posteriori estimate (MAP) can be obtained. Thus, the basis $\mathbf{\Psi}$ is estimated by

$$\hat{\mathbf{\Psi}} = \arg\max_{\mathbf{\Psi},\mathbf{S}} p(\mathbf{\Psi}, \mathbf{S}|\mathbf{X}_\mathrm{T}) \tag{3}$$

$$= \arg\max_{\mathbf{\Psi},\mathbf{S}} p(\mathbf{X}_\mathrm{T}|\mathbf{\Psi}, \mathbf{S}) p(\mathbf{S}) p(\mathbf{\Psi}) \tag{4}$$

where we assume $\mathbf{\Psi}$ and $\mathbf{S}$ to be independent. In the sequel, we make two further assumptions. First, we assume that the columns of $\mathbf{X}_\mathrm{T}$ and $\mathbf{S}$ are independent, respectively, allowing to model each column separately. Second, we assume



(a) Indian Pines (b) University of Pavia

Fig. 2: Classification results for (a) Indian Pines and (b) University of Pavia

that each sample $\mathbf{x}_{\mathrm{T},n}$ suffers from additive white Gaussian noise, i.e. $p(\mathbf{x}_{\mathrm{T},n}|\mathbf{\Psi},\mathbf{s}_n) = \mathcal{N}(\mathbf{\Psi}\mathbf{s}_n, \sigma^2\mathbf{I})$ with mean $\mathbf{\Psi}\mathbf{s}_n$ and variance $\sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

According to CS theory, we choose a sparsity-inducing prior for $\mathbf{s}_n$, e.g $\mathbf{s}_n \sim \text{Laplace}(0, b_1)$, with zero mean and parameter b_1. If we additionally assume a flat prior for $\mathbf{\Psi}$, we obtain the following objective function $E_{\text{sparse}}(\mathbf{\Psi},\mathbf{S})$ that needs to be minimized with respect to $\mathbf{\Psi}$ and $\mathbf{S}$:

$$E_{\text{sparse}}(\mathbf{\Psi},\mathbf{S}) = ||\mathbf{X}_{\mathrm{T}} - \mathbf{\Psi}\mathbf{S}||_F^2 + b_1 \sum_{n=1}^{N} ||\mathbf{s}_n||_1, \quad (5)$$

where $||\cdot||_F$ denotes Frobenious' norm. When minimizing Eq. (5), a scaling problem can be observed: If we upscale the columns of $\mathbf{\Psi}$ and downscale the corresponding columns of $\mathbf{S}$, the value of $||\mathbf{X}_T - \mathbf{\Psi}\mathbf{S}||_F^2$ might not change, but the sparsity term $b_1 \sum_{n=1}^{N} ||\mathbf{s}_n||_1$ may decrease to zero and hence, no sparsity is achieved. A convenient way to overcome this problem is to normalize each basis vector $\mathbf{\Psi}_n$, i.e. $\tilde{\mathbf{\Psi}}_n := \frac{\mathbf{\Psi}_n}{||\mathbf{\Psi}_n||_2}$ [19].

As $\mathbf{\Psi}$ is supposed to be a basis matrix, we additionally require (approximately) orthogonal columns, i.e. $\mathbf{\Psi}^T\mathbf{\Psi} \approx \mathbf{D}$, where $\mathbf{D}$ is a $J \times J$ diagonal matrix. Therefore, we construct the following energy function $E_{\text{orth}}(\mathbf{\Psi})$ that needs to be minimized with respect to $\mathbf{\Psi}$:

$$E_{\text{orth}}(\mathbf{\Psi}) = ||\mathbf{\Psi}^T\mathbf{\Psi} - \mathbf{D}||_F^2. \quad (6)$$

Combining Eq. (5) and Eq. (6) results in

$$\begin{aligned} E(\mathbf{\Psi},\mathbf{S}) =& ||\mathbf{X}_{\mathrm{T}} - \mathbf{\Psi}\mathbf{S}||_F^2 + b_1 \sum_{n=1}^{N} ||\mathbf{s}_n||_1 \\ &+ b_2 ||\mathbf{\Psi}^T\mathbf{\Psi} - \mathbf{D}||_F^2, \end{aligned} \quad (7)$$

where b_2 is a regularization parameter to control the orthogonality strength.

Finally, Eq. (7) is minimized with respect to $\mathbf{\Psi}$ and $\mathbf{S}$. For this purpose, a gradient descent method is chosen where $\mathbf{\Psi}$ and $\mathbf{S}$ are iteratively updated in turns. In particular, Lee and Seung present in [19] a multiplicative update rule:

$$\mathbf{\Psi} \leftarrow \mathbf{\Psi} \circ \frac{[\partial E/\partial \mathbf{\Psi}]^-}{[\partial E/\partial \mathbf{\Psi}]^+} \quad \text{and} \quad \mathbf{S} \leftarrow \mathbf{S} \circ \frac{[\partial E/\partial \mathbf{S}]^-}{[\partial E/\partial \mathbf{S}]^+}, \quad (8)$$

where $[\cdot]^+$ and $[\cdot]^-$ are positive and negative parts of the partial derivatives and $\circ$ denotes the Hadamard product. Using this optimization strategy, $\mathbf{\Psi}$ and $\mathbf{S}$ are additionally constrained to be non-negative, which is related to Non-Negative Matrix Factorization (NMF) [19, 20]. Though this constraint is not required, this method has the advantage of choosing the step width automatically as opposed to many other gradient based optimization algorithms.

Note that the choice of parameters b_1, b_2 and J in Eq. (6) has a significant influence on the result. Since we are interested in maximizing the classification accuracy, we will select those in a cross-validation step with respect to the classification accuracy.

4.2. Learning the sensing matrix

Given a basis $\mathbf{\Psi} \in \mathbb{R}^{D\times J}$, we suggest to adapt the sensing matrix $\mathbf{\Phi} \in \mathbb{R}^{M\times D}$ such that the correlation of the columns of $\mathbf{\Phi\Psi}$ is minimized. As in Section 4.1, we solve $(\mathbf{\Phi\Psi})^T\mathbf{\Phi\Psi} \approx \mathbf{R}$, where $\mathbf{R}$ is a $J \times J$ diagonal matrix. For simplicity, let us assume that the columns of $\mathbf{\Phi\Psi}$ are normalized so that we can assume

$$(\mathbf{\Phi\Psi})^T\mathbf{\Phi\Psi} \approx \mathbf{I} \quad (9)$$

$$(\mathbf{\Phi\Psi\Psi}^T)^T\mathbf{\Phi\Psi\Psi}^T \approx \mathbf{\Psi\Psi}^T, \quad (10)$$

where $\mathbf{I}$ is a $J \times J$ identity matrix. Furthermore, assuming $J = D$, we can easily find an optimal solution for $\mathbf{\Phi}$ by substituting $\mathbf{\Psi\Psi}^T$ by its eigenvalue decomposition, i.e.

Table 1: Classification results (in [%]) of the data sets with 10% of the samples

	Indian Pines				University of Pavia			
	SF		SVM		SF		SVM	
	OA	CA	OA	CA	OA	CA	OA	CA
All	73.725	75.093	85.766	84.892	77.994	73.957	94.034	92.226
ACC	71.565	72.831	81.801	81.812	82.772	79.442	90.981	88.430
CC	70.454	71.131	77.206	77.529	79.938	76.895	88.167	85.107

$\mathbf{\Psi}\mathbf{\Psi}^T = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, with $\mathbf{V} \in \mathbb{R}^{D\times D}$ containing the eigenvectors and the diagonal matrix $\mathbf{\Lambda} \in \mathbb{R}^{D\times D}$ the eigenvalues on the main diagonal. Clearly, the optimal solution is then given by

$$\mathbf{\Phi}_{\text{opt}} = \mathbf{\Lambda}^{-1/2}\mathbf{V}^T \tag{11}$$

which still approximately holds for $J < D$. Based on this result, we reformulate the problem in Eq. (10) and solve the optimization problem

$$\hat{\mathbf{\Phi}}' = \arg\min_{\mathbf{\Phi}'} ||\mathbf{\Phi}' - \mathbf{\Phi}_{\text{opt}}||_F^2 \quad \text{s.t.} \quad \text{rank}(\mathbf{\Phi}') = M \tag{12}$$

where $\mathbf{\Phi}'$ is a $D \times D$ matrix of rank M. Eckart and Young [21] provide a closed form solution for Eq. (12), known as the *Eckart-Young-Theorem*. In fact, we use only the top M eigenvalues of $\mathbf{\Phi}_{\text{opt}}$ to estimate $\mathbf{\Phi}'$, i.e.

$$\hat{\mathbf{\Phi}}' = \lceil \mathbf{\Lambda}^{-1/2} \rceil^M \mathbf{V}^T, \tag{13}$$

where $\lceil\cdot\rceil^M$ choses only the top M values and replaces the remaining elements with zero. Subsequently, we remove $D - M$ rows in $\hat{\mathbf{\Phi}}'$ to obtain the estimate $\hat{\mathbf{\Phi}}$ of the $M \times D$ sensing matrix.

5. RESULTS & DISCUSSION

We evaluated the proposed methods using Indiana's Indian Pines (IP) [22] and the University of Pavia (UP) [23] data sets based on the overall accuracy (OA) and class accuracy (CA) [6]. The Indian Pines data set has a size of 145×145 pixels and 220 bands in total (including noisy bands) mainly showing 16 different classes of vegetation. The University of Pavia image shows an urban scene comprised of 9 different classes. It has a size of 610×340 pixels and 107 bands.

In order the show the performance of the presented methods, two classifiers are used: Matched or Smashed filter (SF) [10] and Support Vector Machines (SVM) [24]. In the sequel, we show the overall classification accuracy of Compressed Classification (CC), Adaptive Compressed Classification (ACC), and conventional classification, using all available data (All) for different subsampling rates, i.e. reduced numbers of bands M. Note that All is clearly independent of M since all bands are used and, thus, can be considered as baseline. In each setup we compute the correct classification accuracies by randomly splitting the data set into 10% training and 90% test data. The results are averaged over 50 repetitions, using different test and training sets and sensing and basis matrix in each run. Note that this leads to some fluctuations in the results for All. In order to estimate the parameters b_1, b_2 and J for ACC, cross-validation is used (IP: $b_1 = 0.75, b_2 = 0.5, J = 100$; UP: $b_1 = 0.25, b_2 = 0.25, J = 50$). The results for the data sets are shown in Fig. 2. Most important, we observe that even for a low number of measurements, CC as well as ACC yield results close to the conventional classification accuracy. Capturing only 10% of the data gives in average an accuracy of 77.15% for the Indian Pines and 88.17% for the University of Pavia using SVMs. In case of SF applied to IP, CC and ACC provide even more accurate results than All. Using the proposed adaptation of the basis and measurement matrices, the results are in average improved by 3.5% to 7% . This effect becomes more apparent if we decrease the number of measurements to e.g. only three measurements, which leads to an improvement of more than 6.5% (UP) and 9.3% (IP). Tab. 1 shows more detailed results for IP and UP where only 10% of the original data is used for classification.

6. CONCLUSION

We extended the work on Compressed Sensing in hyperspectral imaging to a Compressed Classification framework. The benefits of this new approach are numerous: (i) a computationally expensive reconstruction of the data is not required and (ii) the measurement and transform matrices can be optimized using the available training data. The results based on real data show that the presented approach works well and easily outperforms conventional Compressed Sensing and Classification methods in hyperspectral imaging. Future research involves the adaptation of other classifiers and the exploitation of the class information during optimization of the matrices.

7. ACKNOWLEDGMENT

The authors would like to thank Prof. D. Landgrebe from Purdue University, U.S., for kindly providing the Indiana's Indian Pines image and Prof. P. Gamba from University of Pavia, Italy, for providing the University of Pavia image.

8. REFERENCES

[1] G. Shaw and D. Manolakis, "Signal processing for hyperspectral image exploitation," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 12–16, 2002.

[2] R.A. Schowengerdt, *Remote Sensing: Models and Methods for Image Processing*, Elsevier, 3rd edition, 2007.

[3] G. Elmasry, M. Kamruzzaman, D.W. Sun, and P. Allen, "Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: a review," *Critical Reviews in Food Science and Nutrition*, vol. 52, no. 11, pp. 999–1023, 2012.

[4] C.-I Chang and Q. Du, "Estimation of number of spectrally distinct signal sources in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 3, pp. 608–619, 2004.

[5] A. Martinez-Uso, F. Pla, J.M. Sotoca, and P. Garcia-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 4158 –4171, 2007.

[6] V. Kumar, J. Hahn, and A.M. Zoubir, "Band selection for hyperspectral images based on self-tuning spectral clustering," in *Proceedings of the European Signal Processing Conference (EUSIPCO 2013)*, September 2013.

[7] M. Golbabaee, S. Arberet, and P. Vandergheynst, "Multichannel compressed sensing via source separation for hyperspectral images," in *Proceedings of the European Signal Processing Conference (EUSIPCO 2010)*, 2010.

[8] C. Li, T. Sun, K.F. Kelly, and Y. Zhang, "A compressive sensing and unmixing scheme for hyperspectral data processing," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1200–1210, 2012.

[9] J. Hahn, C. Debes, M. Leigsnering, and A. M. Zoubir, "Compressive sensing and adaptive direct sampling in hyperspectral imaging," *Digital Signal Processing*, vol. 26, no. 0, pp. 113–126, 2014.

[10] M.A. Davenport, M.F. Duarte, M. Wakin, J. Laska, D. Takhar, K. Kelly, and R. Baraniuk, "The smashed filter for compressive classification and target recognition," in *Proceedings of SPIE*, 2007, vol. 6498.

[11] E.J. Candès and M.B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[12] D.L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[13] R. G. Baraniuk, "Compressive sensing [lecture notes]," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–121, 2007.

[14] R. Calderbank, S. Jafarpour, and R. Schapire, "Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain," Tech. Rep., Computer Science, Princeton University, 2009.

[15] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[16] M. Fauvel, J.A. Benediktsson, J. Chanussot, and J.R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 11, pp. 3804–3814, 2008.

[17] Gary A. Shaw and Hsiao-hua K. Burke, "Spectral imaging for remote sensing," *Lincoln Laboratory Journal*, vol. 14, no. 1, pp. 3–28, 2003.

[18] D. Takhar, J.N. Laska, M.B. Wakin, M.F. Duarte, D. Baron, S. Sarvotham, K.F. Kelly, and R.G. Baraniuk, "A new compressive imaging camera architecture using optical-domain compression," in *Electronic Imaging 2006*. International Society for Optics and Photonics, 2006, pp. 606509–606509.

[19] D.D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[20] D.D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556–562, 2001.

[21] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[22] S. Tadjudin and D.A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 4, pp. 2113–2118, 1999.

[23] P. Gamba, "A collection of data for urban area characterization," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2004, vol. 1.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at *http://www.csie.ntu.edu.tw/∼cjlin/libsvm*.