Joint Group Sparse PCA for Compressed Hyperspectral Imaging

Zohaib Khan, Faisal Shafait, and Ajmal Mian

***Abstract*— A sparse principal component analysis (PCA) seeks a sparse linear combination of input features (variables), so that the derived features still explain most of the variations in the data. A group sparse PCA introduces structural constraints on the features in seeking such a linear combination. Collectively, the derived principal components may still require measuring all the input features. We present a joint group sparse PCA (JGSPCA) algorithm, which forces the basic coefficients corresponding to a group of features to be jointly sparse. Joint sparsity ensures that the complete basis involves only a sparse set of input features, whereas the group sparsity ensures that the structural integrity of the features is maximally preserved. We evaluate the JGSPCA algorithm on the problems of compressed hyperspectral imaging and face recognition. Compressed sensing results show that the proposed method consistently outperforms sparse PCA and group sparse PCA in reconstructing the hyperspectral scenes of natural and man-made objects. The efficacy of the proposed compressed sensing method is further demonstrated in band selection for face recognition.**

***Index Terms*— Principal component analysis, compressed sensing, image reconstruction, hyperspectral imaging.**

I. Introduction

MULTIVARIATE Image Analysis (MIA) deals with the analysis of images with multiple measurements per pixel (such as RGB, multispectral or hyperspectral), generally by treating individual pixels as samples and the spectral measurements as the variables [1]. MIA can be useful for a variety of image analysis tasks such as image interpretation, visualization, and compression from the spatio-spectral perspective [2]. One of the most common statistical modeling methods associated with MIA is the Principal Component Analysis (PCA). PCA gives an orthogonal basis aligned with the directions of maximum variances of the data. It is useful for projecting the data onto a subspace defined by the most significant basis vectors. However, each principal component is a linear combination of *all* features which makes measurement of all features essential. In some applications (e.g. hyperspectral imaging), each sensed feature (band) may come with an additional cost of acquisition, processing and storage. Moreover, not all measured features may be important for the potential application. Therefore, it is desirable to select the most informative subset of the features, and hence eliminate the cost of sensing the additional, less informative features.

Lasso (least absolute shrinkage and selection operator) is a regularized regression method which penalizes the absolute sum of regression coefficients forcing some of the coefficients to be zero [3]. It results in feature selection by selecting features corresponding to the non-zero regression coefficients. Zou et al. [4] cast PCA as a regression type optimization problem and imposed lasso to approximate the data with a sparse linear combination of the input features. The result was Sparse PCA, which enforces sparsity on the linear combination of input features used to compute the PCA basis. Such an approach is good for interpretation of the data but requires measurement of all the input features. Other algorithms for computing Sparse PCA include the SCoTLASS algorithm [5] which aims at maximizing the Rayleigh quotient of the covariance matrix of the data using non-convex optimization, the DSPCA algorithm [6] which solves a convex relaxation of the sparse PCA problem, the low rank matrix approximation method with sparsity constraint [7], Sparse PCA with positivity constraints [8] and the generalized power method [9]. In these methods, the computation of each basis vector is dealt as an independent problem, the basis vectors are individually sparse but may not be jointly sparse.

Another aspect overlooked by Sparse PCA is the structure of the data in terms of groups of features [10]. For example, image pixels are organized on a rectangular grid exhibiting connectivity and neighborhood relationships. Similarly, gene expression data involves groups of genes corresponding to the same biological processes or sets of genes which are functional neighbors. It is sometimes desirable to encode such group relationships in Sparse PCA so that sparsity follows the group structure. Standard sparse solutions do not offer incorporation of feature groups.

A rather obvious extension of the lasso formulation in Sparse PCA to Group Sparse PCA is to introduce the group lasso penalty [11], [12]. Group lasso uses the ℓ_1/ℓ_2 mixed vector norm to shrink all features in predefined groups with small magnitudes to zero. Guo et al. [13] proposed Sparse Fused PCA which derives group structures from feature correlation. They augmented the Sparse PCA formulation [4] by an additional penalty term which encouraged the coefficients of highly correlated features to be similar and subsequently fused.

Manuscript received July 6, 2014; revised May 19, 2015 and July 6, 2015; accepted August 6, 2015. Date of publication August 24, 2015; date of current version September 18, 2015. This work was supported by the Australian Research Council under Grant DP110102399. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vladimir Stankovic.

Z. Khan and A. Mian are with the School of Computer Science and Software Engineering, The University of Western Australia, Crawley, WA 6009, Australia (e-mail: zohaib@csse.uwa.edu.au; ajmal.mian@uwa.edu.au).

F. Shafait is with the School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad 44000, Pakistan, and also with the School of Computer Science and Software Engineering, The University of Western Australia, Crawley, WA 6009, Australia (e-mail: faisal.shafait@uwa.edu.au).

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

Digital Object Identifier 10.1109/TIP.2015.2472280

However, their solution did not directly result in sparsity, but only forced the coefficients to be of similar value which may or may not be zero. Jenatton et al. [14] used the non-convex ℓ_α/ℓ_2 quasi-norm (where $\alpha \in (0,1)$) for structured sparse PCA. Rectangular patterns were rotated to obtain a larger set of convex patterns for group definition. They demonstrated the use of structured sparsity in image denoising and face recognition tasks. Grbovic et al. introduced two types of grouping constraints into the Sparse PCA problem to ensure reliability of the resulting groups [15]. Jacob et al. [16] proposed a new penalty function which allowed potentially overlapping groups, whereas, Huang et al. [10] generalized the group sparsity to accommodate arbitrary structures.

While group sparsity accounts for the data structure, it does not guarantee joint sparsity of the complete PCA basis with respect to the input features. We present Joint Group Sparse PCA (JGSPCA) which forces the basis coefficients corresponding to a group of features to be jointly sparse. Joint sparsity allows to reconstruct the complete data from only a sparse set of input features, whereas the group sparsity ensures that the structure of the correlated features is preserved. An important application of Sparse PCA and Group Sparse PCA is data interpretation through dimensionality reduction. However, the proposed Joint Group Sparse PCA (JGSPCA) can also be used for model based compressed sensing. Classical compressed sensing does not assume any prior model over the data and is based on the restricted isometry property (see [17] and the references therein). In other words, they are not learning based. On the other hand, the proposed JGSPCA algorithm is learning based and is closer to the model based compressive sensing theory introduced by Baranuik [18].

We validate the proposed JGSPCA algorithm on the problem of compressed hyperspectral imaging and recognition. A hyperspectral image is a data cube comprising two spatial and one spectral dimension. Since the spectra of natural objects are smooth, their variations can be approximated by a few basis vectors [19]. Besides, there is a high correlation among neighboring pixels in the spatial dimension. In a compact representation of such data, structure needs to be preserved in the spatial dimension, while sparsity is desirable in the spectral dimension. Pixels from local spatial neighborhood are grouped together, while sparsity is induced along the spectral dimension. This redundancy in the data makes hyperspectral images a good candidate for sparse representation [20] as well as compressed sensing [21]. We present a Joint Group Sparse PCA algorithm in Section II. Description of the experimental setup, evaluation protocol, and datasets used in the experiments are given in Section III. The results of compressed sensing and recognition experiments are presented in Section IV. The paper is concluded in Section V.

II. Joint Group Sparse PCA

Notations: In the following text, a lowercase letter (x) represents a scalar, a lowercase letter in bold font ($\mathbf{x}$) represents a vector, and an uppercase letter in bold font ($\mathbf{X}$) a matrix. All vectors are treated as column vectors. x_i is the i^{th} element of $\mathbf{x}$. $\mathbf{x}^i$ is the i^{th} row and $\mathbf{x}_j$ is the j^{th} column of a matrix. We use $\mathcal{G}$ to denote a set of integers. $\mathbf{x}_{\cdot\mathcal{G}}$ gives a sub-vector after indexing the vector $\mathbf{x}$ by the elements of $\mathcal{G}$. $\mathbf{X}_{\cdot\mathcal{G}}$ is the submatrix obtained by indexing the columns of $\mathbf{X}$ by the set $\mathcal{G}$. Similarly $\mathbf{X}^{\cdot\mathcal{G}}$ indexes the rows. The number of elements in a set is returned by $|\mathcal{G}|$.

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^\mathsf{T} \in \mathbb{R}^{n\times p}$ be a data matrix which comprises n observations $\mathbf{x}^i \in \mathbb{R}^p$, where p is the number of features. Assume that the sample mean $\bar{\mathbf{x}} \in \mathbb{R}^p$ has been subtracted from all n observations so that the rows of $\mathbf{X}$ are centered. Generally, a PCA basis can be computed by singular value decomposition of the data matrix.

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\mathsf{T} \tag{1}$$

where $\mathbf{V} \in \mathbb{R}^{p\times p}$ are the PCA *basis vectors* (loadings) and $\mathbf{S}$ is a diagonal matrix of *singular values*. $\mathbf{V}$ is an orthonormal basis such that $\mathbf{v}_i^\mathsf{T}\mathbf{v}_j = 0\ \forall\ i \neq j$ and $\mathbf{v}_i^\mathsf{T}\mathbf{v}_j = 1\ \forall\ i = j$. If $\mathbf{X}$ is low rank, it is possible to significantly reduce its dimensionality by using the k most significant basis vectors. The projection of data $\mathbf{X}$ upon the first k basis vectors of $\mathbf{V}$ gives the *principal components* (scores). An alternative formulation treats PCA as a regression type optimization problem

$$\begin{aligned} &\min_{\mathbf{A}}\ \|\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}^\mathsf{T}\|_F^2 \\ &\text{subject to } \mathbf{A}^\mathsf{T}\mathbf{A} = \mathbf{I}_k, \end{aligned} \tag{2}$$

where $\|.\|_F$ is the Frobenius norm, $\mathbf{A} \in \mathbb{R}^{p\times k}$ is a matrix whose columns form an orthonormal basis $\{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_k\}$. The columns of $\mathbf{A}$ which minimize (2) are referred to as the PCA basis $\mathbf{V}$. Each principal component is derived from a linear combination of all p features, consequently making $\boldsymbol{\alpha}$ non-sparse. In order to obtain a sparse PCA basis, a regularization term is usually included in the regression formulation (2). Inclusion of a sparse penalty reduces the number of features involved in each linear combination for obtaining the principal components. One way to obtain sparse basis vectors is by imposing the ℓ_0 constraint upon the regression coefficients (basis vectors) [4].

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B}}\ \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\mathsf{T}\|_F^2 + \lambda \sum_{j=1}^{k} \|\boldsymbol{\beta}_j\|_0 \\ &\text{subject to } \mathbf{A}^\mathsf{T}\mathbf{A} = \mathbf{I}_k, \end{aligned} \tag{3}$$

where $\mathbf{B} \in \mathbb{R}^{p\times k}$ corresponds to the required sparse basis $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_k\}$. The ℓ_0-norm regularization term penalizes the number of non-zero coefficients in $\boldsymbol{\beta}$, whereas the loss term simultaneously minimizes the reconstruction error $\|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\mathsf{T}\|_F^2$. If λ is zero, the problem reduces to finding the ordinary PCA basis vectors, equivalent to (2). When $\lambda > 0$, some coefficients of $\boldsymbol{\beta}_j$ are forced to zero due to ℓ_0 penalization, resulting in sparsity as shown in Figure 1(a).

The above formulation allows us to individually determine informative features. However, it may not account for the structural relationship among multiple features. It is sometimes desirable that the sparsity patterns in the computed basis be consistent over a group of features. The group sparsity should collectively improve the interpretation of the underlying sources. To address this issue, we reconsider the problem from the perspective of groups of features. The grouping of features can either be known a priori from the domain information, or

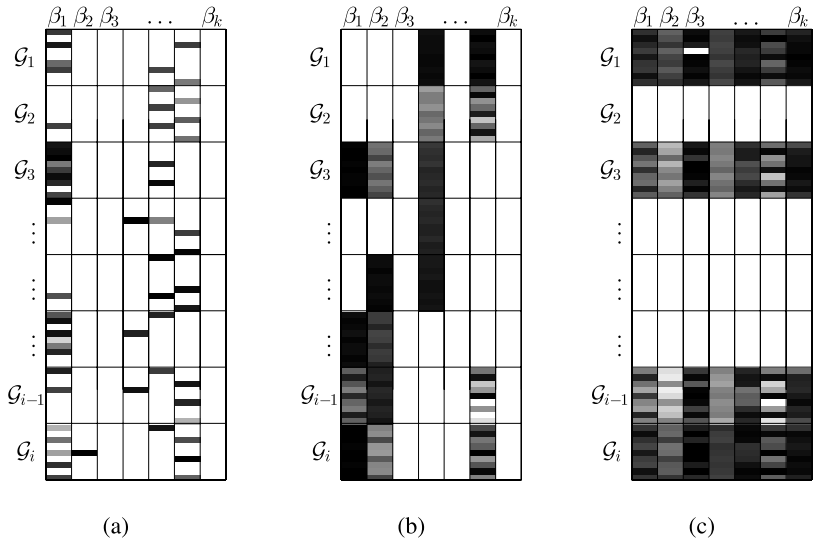


Fig. 1. This example illustrates the basis vectors ($\boldsymbol{\beta}_j$) computed from a data ($\mathbf{X}$) consisting of feature groups ($\mathcal{G}_i$) ($g = 8$, $p_i = 9, k = 7$). Dark cells denote the non-zero coefficients. The group sparsity applies across the groups of features in each basis vector, individually. The joint group sparsity ensures both sparsity among the groups and joint selection of the groups across the basis vectors. (a) SPCA Basis. (b) GSPCA Basis. (c) JGSPCA Basis.

computed directly from the data by utilizing cross-correlation. For instance, in image data, the information from region based segmentation of image pixels could be used to form spatial groups of features. Another possible approach for spatial grouping of assorted pixels is to use dense keypoint based feature correspondences. In general, a region grouping or partitioning scheme may be used to compute a group structure.

Consider that the p features belong to g mutually exclusive groups. Let $\mathcal{G}_i$ be the set of indices of features corresponding to the i^{th} group. The number of features in the i^{th} group is $p_i = |\mathcal{G}_i|$ such that the total number of features $p = \sum_{i=1}^{g} p_i$. Hence, $\mathbf{X}$ can be considered a horizontal concatenation of g sub-matrices $[\mathbf{X}_{\bullet\mathcal{G}_1}, \mathbf{X}_{\bullet\mathcal{G}_2}, \ldots, \mathbf{X}_{\bullet\mathcal{G}_g}]$. Each $\mathbf{X}_{\bullet\mathcal{G}_i} \in \mathbb{R}^{n\times p_i}$ contains data (columns of $\mathbf{X}$) corresponding to the features of the i^{th} group. The group lasso regularization penalizes ℓ_2-norm of the coefficients corresponding to a *feature group* [11]. It enforces sparsity on a group of coefficients, instead of individual coefficients. The group lasso constraint can be incorporated into (3), to achieve the Group Sparse PCA (GSPCA) criterion

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B}} \|\mathbf{X} - \sum_{i=1}^{g} \mathbf{X}_{\bullet\mathcal{G}_i}\mathbf{B}^{\bullet\mathcal{G}_i}\mathbf{A}^{\mathsf{T}}\|_F^2 + \lambda \sum_{j=1}^{k}\sum_{i=1}^{g} \eta_i \|\boldsymbol{\beta}_j^{\bullet\mathcal{G}_i}\|_2 \\ &\text{subject to } \mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{I}_k, \end{aligned} \tag{4}$$

where $\|.\|_2$ is the Euclidean norm and η_i is the weight of the i^{th} group. $\mathbf{B}^{\bullet\mathcal{G}_i} \in \mathbb{R}^{p_i\times k}$ denotes the sub-matrix corresponding to the i^{th} group of features in $\mathbf{B}$. The group lasso penalty $\sum_{i=1}^{g} \eta_i \|\boldsymbol{\beta}^{\bullet\mathcal{G}_i}\|_2$ induces sparsity at the group level, i.e. if the coefficients of the i^{th} group are non-zero, the entire p_i features of the group will be selected and vice versa [12]. It is important to note that the factor η_i will only affect the regularization penalty for different sized groups (typically $\eta_i = \sqrt{p_i}$). In the case of equal sized groups, this factor can be ignored altogether (or assumed $\eta_i = 1$).

Notice that the ℓ_0 penalty in (3) has been replaced with an $\ell_{2,1}$ penalty in (4). This formulation can be considered to be a generalized form for group and non-group structured data. A group may even consist of a single feature, if it is not highly correlated with other features. Hence, in the extreme case of uncorrelated data, each group will contain a single feature, i.e. $g = p$.

Equation (4) gives a sparse basis which accounts for the group structure of the data. When the grouping constraint is enforced, the basis coefficients become sparse in a group-wise manner. Imposing the additional group constraint generally results in reduced sparsity within the feature groups. This phenomenon is illustrated for an example basis in Figure 1. Figure 1(a) depicts a sparse basis obtained by the SPCA criterion (3) which does not take the group structure into account. Figure 1(b) shows a group sparse basis obtained by the GSPCA criterion (4) for the same data. Consider for instance the null coefficients within the groups $\mathcal{G}_i$ of an SPCA basis vector $\boldsymbol{\beta}_j$. As a consequence of enforcing the grouping constraint, some of the coefficients that were null in the SPCA basis within the groups become non-zero in the GSPCA basis. Since the group sparsity is independently achieved in the basis vectors, each vector is sparse for a different group of features and the complete basis may still end up using all groups of features.

In several applications, it is desirable to perform feature selection such that the selected features explain the major variation of the data. This is particularly true for data consisting of a large number of redundant features, or where measurement of features is expensive. To achieve this goal, we expect all basis vectors $\boldsymbol{\beta}_j$ to end up using the same *groups of features*. This kind of sparsity is called *joint sparsity* [22], [23]. Joint sparsity is neither considered by SPCA nor GSPCA, since they solve (3) and (4) for individual basis vectors $\boldsymbol{\beta}_j$. We propose to directly optimize for $\mathbf{B}$ to ensure joint sparsity

while simultaneously achieving group sparsity. In other words, the coefficients corresponding to some groups of rows of $\mathbf{B}$ are forced to be null, as shown in Figure 1(c). Our proposed joint group sparsity can be obtained by imposing the following regularization penalty

$$\ell_{F_g,1}(\mathbf{B}) = \sum_{i=1}^{g} \eta_i \|\mathbf{B}^{\cdot\mathcal{G}_i}\|_F. \tag{5}$$

The ℓ_1 penalty on $\|\mathbf{B}^{\cdot\mathcal{G}_i}\|_F$ forces some of the sub-basis groups $\mathbf{B}^{\cdot\mathcal{G}_i}$ to be null. This results in joint group sparsity over the complete basis. The nullified groups directly correspond to the feature groups of $\mathbf{X}$ with minimum contribution in explaining the data. By including the joint group sparse regularization penalty (5) in (3), the proposed Joint Group Sparse PCA criterion is obtained:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B}} \|\mathbf{X} - \sum_{i=1}^{g} \mathbf{X}_{\cdot\mathcal{G}_i}\mathbf{B}^{\cdot\mathcal{G}_i}\mathbf{A}^\intercal\|_F^2 + \lambda \sum_{i=1}^{g} \eta_i \|\mathbf{B}^{\cdot\mathcal{G}_i}\|_F \\ &\text{subject to } \mathbf{A}^\intercal\mathbf{A} = \mathbf{I}_k \ . \end{aligned} \tag{6}$$

For sufficiently large values of λ, some group of rows of $\mathbf{B}$ vanish, resulting in a joint group sparse basis.

Although, the above formulation ensures a joint group sparse basis, simultaneous minimization for $\mathbf{A}$ and $\mathbf{B}$ makes the problem non-convex. If one of the two matrices is known, the problem becomes convex over the second unknown matrix. Hence, a locally convex solution of (6) can be obtained by iteratively minimizing $\mathbf{A}$ and $\mathbf{B}$. Therefore, the joint group sparse PCA formulation in (6) is decomposed into two independent optimization problems. In the first optimization problem, $\mathbf{A}$ is initialized with $\mathbf{V}$ obtained from (1) and the minimization under the joint group sparsity constraint on $\mathbf{B}$ is formulated as

$$\min_{\mathbf{B}} \|\mathbf{XA} - \mathbf{XB}\|_F^2 + \lambda \sum_{i=1}^{g} \eta_i \|\mathbf{B}^{\cdot\mathcal{G}_i}\|_F. \tag{7}$$

The loss term $\|\mathbf{X} - \sum_{i=1}^{g} \mathbf{X}_{\cdot\mathcal{G}_i}\mathbf{B}^{\cdot\mathcal{G}_i}\mathbf{A}^\intercal\|_F^2$ in (6) is equivalent to $\|\mathbf{XA} - \mathbf{XB}\|_F^2$ in (7) given $\mathbf{A}^\intercal\mathbf{A} = \mathbf{I}_k$, and non-overlapping feature groups, $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset \ \forall i \neq j$, so that $\sum_{i=1}^{g} \mathbf{X}_{\cdot\mathcal{G}_i}\mathbf{B}^{\cdot\mathcal{G}_i} = \mathbf{XB}$. The above formulation is similar to a multi-task regularized regression problem [24] with grouping constraints

$$\min_{\mathbf{W}} \|\mathbf{Q} - \mathbf{XW}\|_F^2 + \psi(\mathbf{W}), \tag{8}$$

where $\mathbf{Q} = \mathbf{XA}$ is the response matrix, $\mathbf{W} = \mathbf{B}$ is the matrix of regression coefficients, and ψ is any convex norm defined on the matrix. An optimization problem of the form of (8) can be efficiently solved by proximal programming methods [25].

Once a solution for $\mathbf{B}$ is found via (7), the next step is to solve the optimization with respect to $\mathbf{A}$. For a known $\mathbf{B}$, the regularization penalty in (6) becomes irrelevant for the optimization with respect to $\mathbf{A}$. Therefore, the following objective function is minimized

$$\begin{aligned} &\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{XBA}^\intercal\|_F^2 \\ &\text{subject to } \mathbf{A}^\intercal\mathbf{A} = \mathbf{I}_k. \end{aligned} \tag{9}$$

Algorithm 1 Joint Group Sparse PCA

Input: $\mathbf{X} \in \mathbb{R}^{n\times p}, \{\mathcal{G}_i\}_{i=1}^{g}, \eta_i, \lambda, j_{\max}$
  Initialize: $j \leftarrow 1$, converge $\leftarrow$ false
  $\mathbf{USV}^\intercal \leftarrow \mathbf{X}$
  $\mathbf{A} \leftarrow \mathbf{V}_{\cdot\{1:k\}}$
  while $j \leq j_{\max} \wedge \neg$converge **do**
    $\hat{\mathbf{B}} \leftarrow \arg\min_{\mathbf{B}} \|\mathbf{XA} - \mathbf{XB}\|_F^2 + \lambda\ell_{F_g,1}(\mathbf{B})$
    $\dot{\mathbf{U}}\dot{\mathbf{S}}\dot{\mathbf{V}}^\intercal \leftarrow \mathbf{X}^\intercal\mathbf{X}\hat{\mathbf{B}}$
    $\hat{\mathbf{A}} \leftarrow \dot{\mathbf{U}}\dot{\mathbf{V}}^\intercal$
    if $\|\mathbf{B} - \hat{\mathbf{B}}\|_F < \epsilon$ **then**
      converge $\leftarrow$ true
    else
      $\mathbf{B} \leftarrow \hat{\mathbf{B}}, \mathbf{A} \leftarrow \hat{\mathbf{A}}$
      $j \leftarrow j + 1$
    end if
  end while
Output: $\hat{\mathbf{A}}, \hat{\mathbf{B}}$

A closed form solution to (9) can be obtained as the generalized solution to the orthogonal Procrustes problem [26], [27]. For minimizing (9), the solution proceeds by finding the nearest orthogonal matrix which maps $\mathbf{XB}$ to $\mathbf{X}$. This is done using Singular Value Decomposition, $\mathbf{X}^\intercal\mathbf{XB} = \dot{\mathbf{U}}\dot{\mathbf{S}}\dot{\mathbf{V}}^\intercal$. Then, $\dot{\mathbf{U}}\dot{\mathbf{V}}^\intercal$ is the nearest orthogonal matrix which minimizes (9).

The minimization process alternating between (7) and (9) continues either until convergence, or until a specified number of iterations are completed. Algorithm 1 summarizes the proposed Joint Group Sparse PCA method.

A. Implementation

We used the Fast-Iterative Shrinkage and Thresholding Algorithm (FISTA) library [25] to solve the optimization problems in (3), (4) and (6). Note that we modified the FISTA library to solve for the $\ell_{F_g,1}$ joint group regularization penalty in (5). All source codes for the algorithms in this paper (including the modified FISTA library) are publicly available.[1]

III. Experiments

A. Evaluation Criteria

The data matrix $\mathbf{X}$ is created by sampling non-overlapping spatio-spectral volumes of dimension $\sqrt{p_i} \times \sqrt{p_i} \times g$, $p_i = 9 \ \forall \ i$ (after vectorizing) from all training hyperspectral images. A model is learned from the training data using Algorithm 1. At most g models are learned with each algorithm, one for each $r = 1, 2, \ldots, g$ number of bands, where r is defined as the feature group cardinality

$$r = |\{i \mid \|\mathbf{B}^{\cdot\mathcal{G}_i}\|_{F,1} \neq 0\}|. \tag{10}$$

To evaluate compressive sensing performance of the r^{th} learned model with orthonormal basis $\mathbf{A}$ and the corresponding

[1] http://www.sites.google.com/site/zohaibnet/Home/codes.html

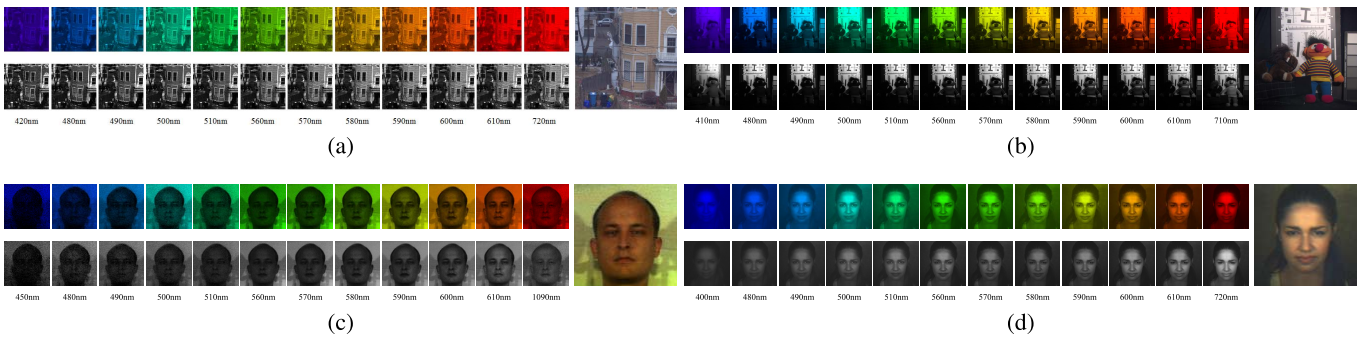


Fig. 2. Sample hyperspectral images in different datasets. Each image is shown as a series of bands in pseudo color and grayscale (only a subset of bands is shown here). Also shown are their corresponding RGB rendered images. (a) Harvard Data. (b) CAVE Data. (c) CMU Data. (d) UWA Data.

TABLE I

AN OVERVIEW OF HYPERSPECTRAL IMAGE DATABASES USED IN THE EXPERIMENTS. OUR NEWLY DEVELOPED UWA FACE DATABASE IS A LOW NOISE HYPERSPECTRAL FACE DATABASE IN THE VISIBLE RANGE

Database	**Harvard**	**CAVE**	**CMU**	**UWA**
Spectral Range (nm)	420 - 720	410 - 710	450 - 1090	400 - 720
Number of Bands	31 (VIS)	31 (VIS)	65 (VIS-NIR)	33 (VIS)
Spatial Resolution	1392×1040	512×512	640×480	1024×1024
Images/Subjects	50	32	48	70
Acquisition Time	60 sec	-	8 sec	6 sec
Noise Grade	Low	Low	High	Low

sparse basis $\mathbf{B}$, the reconstruction error is computed as

$$e_r = \frac{\|\mathbf{X}\mathbf{V}_{\cdot\{1:k\}}\mathbf{V}_{\cdot\{1:k\}}^{\mathsf{T}} - \mathbf{X}\mathbf{B}\mathbf{A}^{\mathsf{T}}\|_F}{\|\mathbf{X}\mathbf{V}_{\cdot\{1:k\}}\mathbf{V}_{\cdot\{1:k\}}^{\mathsf{T}}\|_F} \tag{11}$$

where k is the number of basis vectors. The fraction of variance $\nu \in (0,1)$ explained by the first k basis vectors determines the number of principal components in a model

$$\arg\min_k \frac{\sum_{i=1}^{k}\Lambda_i}{\sum\Lambda} - \nu, \tag{12}$$

where $\Lambda = \frac{\text{diag}(\mathbf{S})^2}{N-1}$ is a vector of the principal component variances arranged in descending order. A common practice is to use $\nu = 0.9$. It should be noted that sparse PCA variants have the intrinsic ability to determine the extent to which a principal component is used, which is determined by the values of the basis coefficients. Due to the optimization criterion (minimizing reconstruction error), the less informative principal components will automatically have very small coefficients corresponding to their basis vectors, effectively reducing their contribution to the reconstruction of the data.

B. Databases

We perform compressed sensing experiments on four hyperspectral image databases. These hyperspectral images were acquired by sequentially tuning a filter through the spectrum and capturing image with a monochrome camera. A sample image from each dataset is shown in Figure 2. All images were downsampled using bilinear interpolation. A summary of specifications for all hyperspectral datasets used in the experiments is provided in Table I. A brief description of each database is as follows:

1) Harvard Scene Dataset: The dataset contains hyperspectral images of 50 indoor and outdoor scenes under daylight illumination [20]. The images were captured using a commercial grade hyperspectral camera (Nuance FX, CRI Inc.), which is based on a liquid crystal tunable filter design. The dataset consists of a diverse range of objects, materials and structures and is a good representative of real world spatio-spectral interactions. The training and testing datasets consist of 10 and 40 images, respectively. All images were spatially resized to 105×141 pixels.

2) CAVE Scene Dataset: The CAVE multispectral image database contains true spectral reflectance images of 32 scenes consisting of a variety of objects in an indoor setup [28]. It has 31 band hyperspectral images (420-720nm, spaced 10nm apart) at a resolution of 512×512 pixels. All images contain the true spectral reflectance of a scene i.e., they are corrected for ambient illumination. We used 10 images for training and 22 images for testing. Each band was spatially resized to 120×120 pixels.

3) CMU Face Dataset: The CMU hyperspectral face database [29] contains facial images of 48 subjects captured in multiple sessions over a period of about two months. The images cover both visible and near infrared spectral range (450nm to 1090nm, spaced 10nm apart). The data was obtained using a prototype spectro-polarimetric camera mainly comprising of an Acousto Optical Tunable Filter. For experiments, a single sample per subject is used for training and the remaining samples make the test set. Specifically, 48 samples were used for training and 103 for testing. All faces were spatially resized to 24×21 pixels after normalization.

4) UWA Face Dataset: The hyperspectral face database collected in our lab comprises 110 hyperspectral images

of 70 subjects of different ethnicity, gender and age. Each subject was imaged in different sessions, one week to two months apart. The system consists of a monochrome machine vision camera with a focusing lens (1:1.4/25mm) followed by a Liquid Crystal Tunable Filter (LCTF) which is tunable in the range of 400-720 nm. The average tuning time of the filter is 50 ms. The filter bandwidth, measured in terms of the *Full Width at Half Maximum (FWHM)* is 7 to 20nm which varies with the center wavelength. The scene was illuminated by twin-halogen lamps on both sides of the subject. The illumination was left partially uncontrolled as it was mixed with indoor lights and occasionally daylight, varying with the time of image capture. For spectral response calibration, the white patch from a standard 24 patch color checker was utilized.

The training and testing sets consist of 70 and 40 images, respectively. The database has been made publicly available for research.[2]

IV. Results and Discussion

A. Compressed Hyperspectral Imaging

In the first experiment, we examine the compressive sensing performance of all algorithms (SPCA, GSPCA and JGSPCA) in terms of reconstruction error. In the following text, feature refers to the *band* of a hyperspectral image. When the number of bands in a model increases, the reconstruction error should decrease. However, an attempt to reconstruct a band that is highly corrupted by noise may result in an increased reconstruction error. Therefore, it is important for a method to select bands that are most informative for the representation of the data. An algorithm is expected to be relatively better for compressed sensing if it achieves a lower reconstruction error with fewer bands.

The reconstruction error on the test data with different algorithms is provided in Figure 3. Interesting results are obtained for the reconstruction errors on the Harvard and CAVE scene datasets. The first few bands similarly explain the data with either SPCA or GSPCA. When more bands are added into the model, significant improvement in the reconstruction error is achieved with JGSPCA. We observe that GSPCA alone is only slightly better than SPCA, whereas the JGSPCA consistently achieves lower reconstruction error and outperforms both SPCA and GSPCA. The results on hyperspectral face datasets are slightly different from the scenes datasets. The JGSPCA consistently outperforms SPCA and GSPCA on both CMU and UWA datasets. It reconstructs the data with lower error from the first band up until the last band on both databases.

It is important to note that, in some cases, the reconstruction errors are similar regardless of the type of sparsity. Thus, if similar bands are selected, the reconstruction error using those bands may be similar as well. Beyond the first few bands, the proposed JGSPCA is able to identify and select the most informative bands earlier than the other algorithms and hence results in lower reconstruction errors. For instance,

[2]UWA Hyperspectral Face Database
http://www.sites.google.com/site/zohaibnet/Home/databases

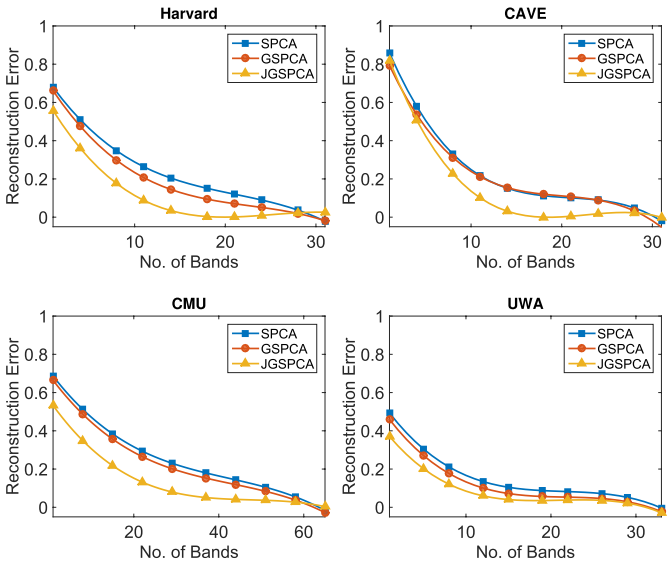


Fig. 3. Reconstruction errors (e_r) on Harvard, CAVE, CMU and UWA datasets. The smooth curves are the result of fitting multiple linear regression model with methods as categorical variables. In all instances, the p-value of the interaction term with the method JGSPCA is less than the common alpha level of 0.05, which proves its statistical significance.

TABLE II
The Number of Bands Required to Achieve a Specific Reconstruction Error. Lower Number of Bands Indicates the Superiority of a Method in Delivering a Linear Combination of Informative Bands

	Harvard			CAVE		
e_r (%)	30%	20%	10%	30%	20%	10%
SPCA	11	14	21	9	12	20
GSPCA	9	12	17	9	14	19
JGSPCA	4	6	9	6	7	10

	CMU			UWA		
e_r (%)	30%	20%	10%	30%	20%	10%
SPCA	22	33	52	4	8	18
GSPCA	21	30	45	3	8	13
JGSPCA	8	16	28	2	4	9

the reconstruction error curves on CMU dataset suggest that crucial bands are selected by JGSPCA when the number of bands is increased from 1 to 30 which is illustrated by a steep drop in e_r down to 15%. To reach the same level of e_r, GSPCA and SPCA require 49 and 56 bands, respectively.

The overall trend of reconstruction errors is also related to the variety of objects, and the number of samples used for training in each database. It is difficult to model spatio-spectral variation of complex objects (such as those in the CAVE database) with a few bands and limited training data. On the other hand, faces are a particular class of objects and can be reconstructed by only a few bands. Moreover, because the image noise is not modeled, it is highly unlikely to achieve zero reconstruction error, which is in turn a benefit of sparse modeling techniques.

Table II provides the number of bands required by a model to limit the reconstruction error within an upper bound.

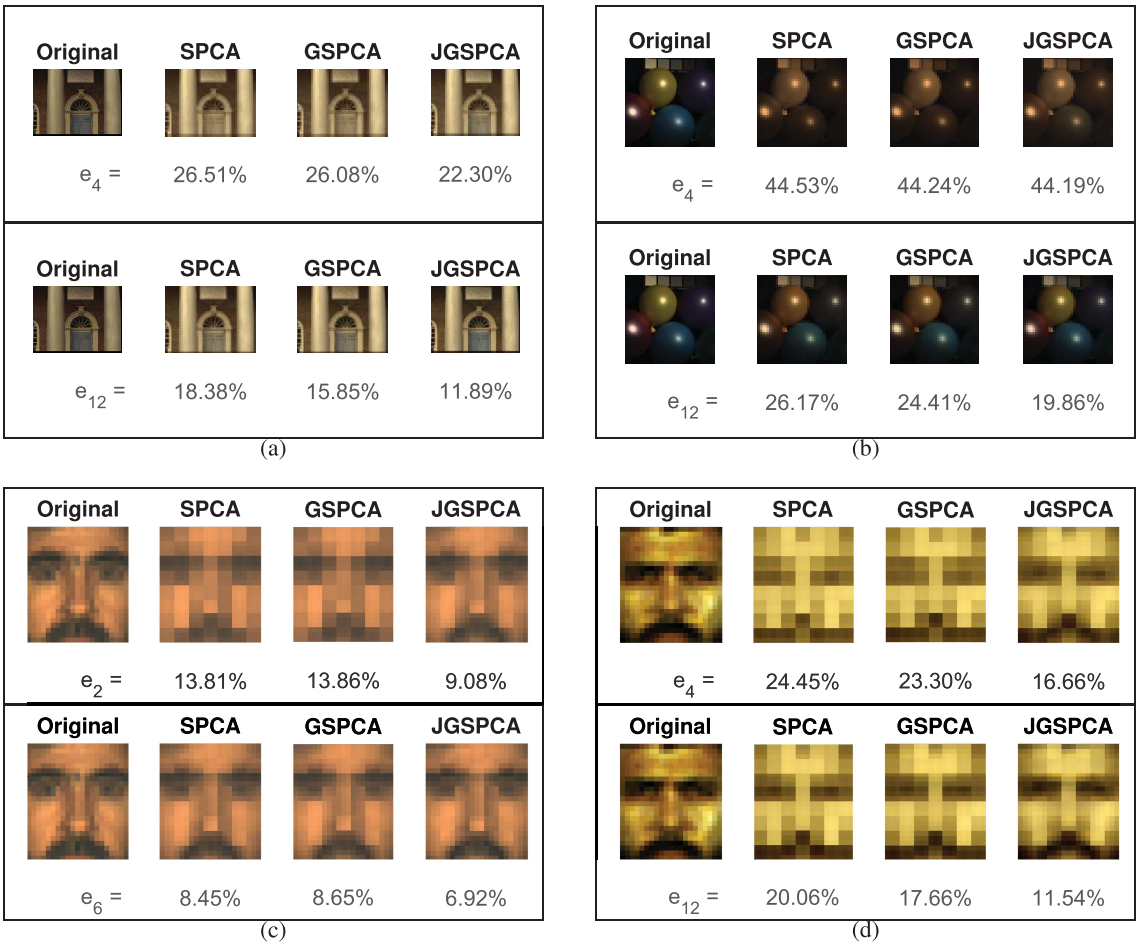


Fig. 4. Compressed sensing results of hyperspectral images (rendered as RGB). The results are shown for the same number of bands used for reconstruction of the hyperspectral image using SPCA, GSPCA and JGSPCA. The original images are rendered using all bands of the hyperspectral images. The differences are numerically and visually appreciable in all examples. (a) A building scene from the Harvard dataset. (b) A balloons scene from the CAVE dataset. (c) A face image from the UWA dataset. (d) A face image from the CMU dataset.

As we are interested in achieving lowest reconstruction errors, we restrict to 30%, 20% and 10% error marks. With a small number of bands, the reconstruction error of the hyperspectral data is too high. When more bands are added, JGSPCA selects a fewer number of bands to achieve the same reconstruction error as the SPCA and GSPCA. For low reconstruction errors, all methods require a relatively higher number of bands, whereas JGSPCA still requires comparatively fewer bands. Figure 4 shows compressed sensing of four example images using SPCA, GSPCA and JGSPCA methods. The proposed JGSPCA exhibits significantly lower reconstruction errors which can also be visually appreciated. The difference is more obvious when using a small number of bands for compressed sensing. Overall, JGSPCA performs the best in compressed hyperspectral imaging, followed by GSPCA and SPCA.

B. Hyperspectral Face Recognition

In this experiment, we compare the compressive sensing of hyperspectral images using different algorithms in the context of a recognition task. We expect a compressive sensing algorithm to achieve a high recognition accuracy while sensing only a small number of bands. We evaluate our proposed JGSPCA algorithm for band selection in hyperspectral face recognition and compare it to SPCA and GSPCA. In order to understand the purpose of this experiment, the following points need due consideration

1) We use several widely accepted classification methods to evaluate the trend of recognition accuracy against compressive sensing of hyperspectral face images. Any other state-of-the-art algorithm may perform better than the chosen baseline algorithms, however the trend is expected to be similar.
2) We assume that the bands that are informative for class separation are the bands that are informative for explanation of the data, which is the default criterion in PCA. A discriminant criterion [30], [31] is expected to adequately satisfy this assumption. A more attractive approach could be to learn a model which maximizes the covariance between hyperspectral bands and the classes to separate. Partial Least Squares (PLS) Discriminant Analysis can take into account class separation without any further assumption about the variance or covariance structure of the data as opposed to PCA. It can potentially model the data from the perspective of regression (numerical responses) or classification (categorical responses), a direction worth exploring in the future.

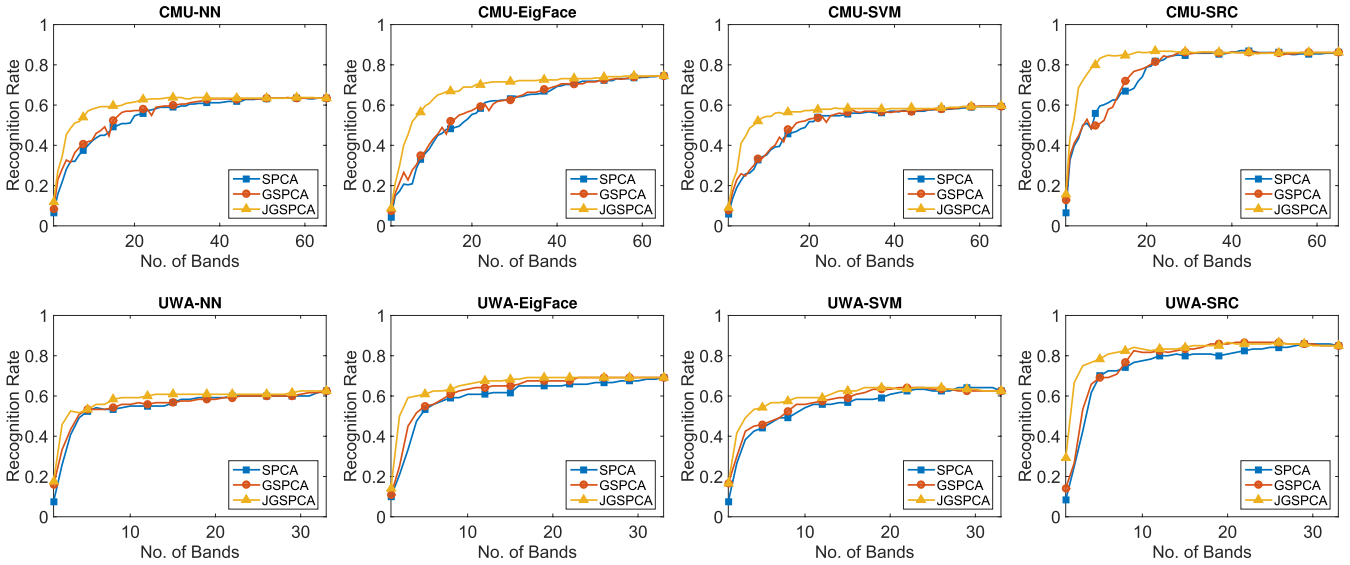


Fig. 5. Recognition accuracy (a_r) versus the number of selected bands on CMU and UWA face datasets. The JGSPCA demonstrates consistently higher recognition accuracy compared to SPCA and GSPCA.

TABLE III
THE NUMBER OF BANDS TO ACHIEVE A SPECIFIC RECOGNITION ACCURACY. LOWER NUMBER OF BANDS INDICATES THE SUPERIORITY OF A METHOD IN SELECTING INFORMATIVE BANDS

NN	CMU			UWA		
a_r (%)	50%	55%	60%	50%	55%	60%
SPCA	17	21	33	5	15	32
GSPCA	15	17	30	4	9	30
JGSPCA	6	9	18	3	6	12
EigFace	CMU			UWA		
a_r (%)	50%	60%	70%	50%	55%	60%
SPCA	18	20	23	5	6	10
GSPCA	15	17	23	4	6	8
JGSPCA	6	8	10	3	3	5
SVM	CMU			UWA		
a_r (%)	45%	50%	55%	45%	50%	55%
SPCA	15	20	28	6	9	11
GSPCA	15	17	26	5	8	9
JGSPCA	6	7	12	3	4	6
SRC	CMU			UWA		
a_r (%)	60%	70%	80%	60%	70%	80%
SPCA	10	18	22	4	5	14
GSPCA	13	15	21	4	7	9
JGSPCA	4	5	9	2	3	6

A model is learned using a single hyperspectral image per subject in the training set which makes the gallery. All remaining hyperspectral images which comprise the test set, serve as the probes. A test hyperspectral image cube is compressively sensed (reconstructed by learned model) and used for classification. Consider a training set $\mathbf{X}$ and test set $\mathbf{Z}$, where each row is a hyperspectral face image. The compressive sensing performance of the r^{th} learned model in terms of recognition accuracy is computed as

$$a_r = \texttt{classify}(\{\mathbf{A}, \mathbf{B}\}, \mathbf{X}, \mathbf{Z}), \tag{13}$$

where the operator `classify` is a classification method such as Nearest Neighbor (NN), EigenFaces [32], Support Vector Machine (SVM) [33] or Sparse Representation-based Classification (SRC) [34]. The recognition accuracies from each algorithm are averaged over three folds.

Figure 5 shows the recognition accuracy against the number of bands used for reconstruction of test hyperspectral images. It can be easily observed that JGSPCA consistently achieves higher recognition accuracy with fewer bands compared to SPCA and GSPCA on both databases. The consistency of the trend can be observed across different recognition algorithms. In order to numerically analyze the recognition performance through compressed sensing, we tabulate the number of bands required to achieve a certain recognition accuracy mark. In Table III, we are interested in achieving higher recognition accuracies with small number of bands. It can be observed that the proposed JGSPCA algorithm achieves higher recognition accuracy by sensing only a few bands compared to SPCA and GSPCA. This implicitly indicates the ability of JGSPCA to select more informative bands for a recognition task.

V. CONCLUSION

In this paper, we presented a Joint Group Sparse PCA algorithm which addresses the problem of finding a few *groups of features* that *jointly* capture most of the variation in the data. Unlike other sparse formulations of PCA, for which all features might still be needed for reconstructing the data, the presented approach requires only a few features to represent the whole data. This property makes the presented formulation most suitable for compressed sensing, in which the main goal is to measure only a few features that capture most significant information. The efficacy of our approach has been demonstrated by experiments on several real-world datasets of hyperspectral images. The results show that our presented approach outperforms Sparse PCA and Group Sparse PCA algorithms when applied to compressed hyperspectral imaging and hyperspectral face recognition. The proposed methodology

is well adaptable to scenarios where the features can be implicitly or explicitly categorized into groups.

REFERENCES

[1] J. M. Prats-Montalbán, A. de Juan, and A. Ferrer, "Multivariate image analysis: A review with applications," *Chemometrics Intell. Lab. Syst.*, vol. 107, no. 1, pp. 1–23, May 2011.

[2] A. de Juan, M. Maeder, T. Hancewicz, T. R. Duponchel, and R. Tauler, "Chemometric tools for image analysis," in *Infrared and Raman Spectroscopic Imaging*. New York, NY, USA: Wiley, 2009, pp. 65–109.

[3] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[4] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Statist.*, vol. 15, no. 2, pp. 265–286, 2006.

[5] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, "A modified principal component technique based on the LASSO," *J. Comput. Graph. Statist.*, vol. 12, no. 3, pp. 531–547, 2003.

[6] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *SIAM Rev.*, vol. 49, no. 3, pp. 434–448, 2007.

[7] H. Shen and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *J. Multivariate Anal.*, vol. 99, no. 6, pp. 1015–1034, Jul. 2008.

[8] R. Zass and A. Shashua, "Nonnegative sparse PCA," in *Proc. Neural Inf. Process. Syst.*, 2006, pp. 1561–1568.

[9] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 517–553, Jan. 2010.

[10] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," *J. Mach. Learn. Res.*, vol. 12, pp. 3371–3412, Jan. 2011.

[11] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. B*, vol. 68, no. 1, pp. 49–67, Feb. 2006.

[12] J. Friedman, T. Hastie, and R. Tibshirani. (2010). "A note on the group lasso and a sparse group lasso." [Online]. Available: http://arxiv.org/abs/1001.0736

[13] J. Guo, G. James, E. Levina, G. Michailidis, and J. Zhu, "Principal component analysis with sparse fused loadings," *J. Comput. Graph. Statist.*, vol. 19, no. 4, pp. 930–946, Sep. 2010.

[14] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2010. [Online]. Available: http://arXiv:0909.1440

[15] M. Grbovic, C. R. Dance, and S. Vucetic, "Sparse principal component analysis with constraints," in *Proc. AAAI*, Jul. 2012, pp. 935–941.

[16] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 433–440.

[17] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[18] R. G. Baraniuk, "Compressive sensing [lecture notes]," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 118–121, Jul. 2007.

[19] J. P. S. Parkkinen, J. Hallikainen, and T. Jaaskelainen, "Characteristic spectra of Munsell colors," *J. Opt. Soc. Amer.*, vol. 6, no. 2, pp. 318–322, 1989.

[20] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 193–200.

[21] M. Golbabaee and P. Vandergheynst, "Hyperspectral image compressed sensing via low-rank and joint-sparse matrix recovery," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Mar. 2012, pp. 2741–2744.

[22] M. F. Duarte, S. Sarvotham, M. B. Wakin, D. Baron, and R. G. Baraniuk, "Distributed compressed sensing of jointly sparse signals," in *Proc. Workshop Signal Process. Adaptative Sparse Struct. Represent.*, Oct./Nov. 2005, pp. 1537–1541.

[23] K. Lee, Y. Bresler, and M. Junge, "Subspace methods for joint sparse recovery," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3613–3641, Jun. 2012.

[24] J. Mairal, R. Jenatton, F. R. Bach, and G. R. Obozinski, "Network flow algorithms for structured sparsity," in *Proc. Neural Inf. Process. Syst.*, 2010, pp. 1558–1566.

[25] R. Jenatton, J. Mairal, G. Obozinski, and F. R. Bach, "Proximal methods for sparse hierarchical dictionary learning," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 487–494.

[26] P. H. Schönemann, "A generalized solution of the orthogonal Procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, Mar. 1966.

[27] J. C. Gower and G. B. Dijksterhuis, *Procrustes Problems*, vol. 3. London, U.K.: Oxford Univ. Press, 2004.

[28] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.

[29] L. J. Denes, P. Metes, and Y. Liu, "Hyperspectral face database," Dept. Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-RI-TR-02-25, Oct. 2002.

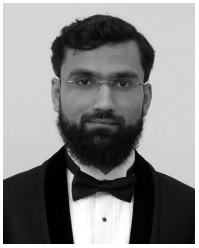
[30] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[31] L. F. S. Merchante, Y. Grandvalet, and G. Govaert, "An efficient approach to sparse linear discriminant analysis," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 1167–1174.

[32] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

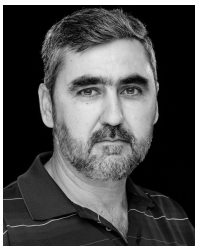
[33] G. Guo, S. Z. Li, and K. Chan, "Face recognition by support vector machines," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 196–201.

[34] J. Wright, A. Y. Yang, A. Ganesh, S. 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.



Zohaib Khan received the B.E. (Hons.) degree in avionics from the National University of Sciences and Technology, Pakistan, in 2008, and won the award for best performance in academics and the gold medal for best project in engineering, and the Ph.D. degree in computer science from The University of Western Australia, Australia, in 2014, as a recipient of the prestigious International Postgraduate Research Scholarship (2010-2014). He is currently a Research Associate with the School of Computer Science and Software Engineering, The University of Western Australia. His research interests mainly include image processing and pattern recognition with a focus on multimodal imaging.

Faisal Shafait received the Ph.D. (Hons.) degree in computer engineering from TU Kaiserslautern, Germany, in 2008. He is currently an Associate Professor with the School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan. Besides, he is an Adjunct Senior Lecturer with the School of Computer Science and Software Engineering, The University of Western Australia, where he formerly held an Assistant Professor position. He has worked for a number of years as a Senior Researcher with the German Research Center for Artificial Intelligence, Germany and a Visiting Researcher with Google, CA. He has co-authored over 100 publications in international peer-reviewed conferences and journals in this area. His research interests include machine learning and computer vision with a special emphasis on applications in document image analysis and recognition. He is an Editorial Board member of the *International Journal on Document Analysis and Recognition*, and a Program Committee Member of leading document analysis conferences, including ICDAR, DAS, and ICFHR. He is also serving on the Leadership Board of IAPR's Technical Committee on Computational Forensics (TC-6).



Ajmal Mian received the Ph.D. (Hons.) degree from The University of Western Australia (UWA), in 2006, and received the Australasian Distinguished Doctoral Dissertation Award from the Computing Research and Education Association of Australasia. He is currently with the School of Computer Science and Software Engineering, The University of Western Australia. His research interests include computer vision, action recognition, 3D shape analysis, hyperspectral image analysis, machine learning, and multimodal biometrics. He received two prestigious nationally competitive fellowships namely the Australian Post-Doctoral Fellowship in 2008 and the Australian Research Fellowship in 2011. He received the UWA Outstanding Young Investigator Award in 2011, the West Australian Early Career Scientist of the Year Award in 2012 and the Vice-Chancellor's Mid-Career Research Award in 2014. He has secured five Australian Research Council grants worth over $2.3 Million.