Rank Preserving Sparse Learning for Kinect Based Scene Classification

Dapeng Tao, Lianwen Jin, *Member, IEEE,* Zhao Yang, and Xuelong Li, *Fellow, IEEE*

***Abstract*—With the rapid development of the RGB-D sensors and the promptly growing population of the low-cost Microsoft Kinect sensor, scene classification, which is a hard, yet important, problem in computer vision, has gained a resurgence of interest recently. That is because the depth of information provided by the Kinect sensor opens an effective and innovative way for scene classification. In this paper, we propose a new scheme for scene classification, which applies locality-constrained linear coding (LLC) to local SIFT features for representing the RGB-D samples and classifies scenes through the cooperation between a new rank preserving sparse learning (RPSL) based dimension reduction and a simple classification method. RPSL considers four aspects: 1) it preserves the rank order information of the within-class samples in a local patch; 2) it maximizes the margin between the between-class samples on the local patch; 3) the L1-norm penalty is introduced to obtain the parsimony property; and 4) it models the classification error minimization by utilizing the least-squares error minimization. Experiments are conducted on the NYU Depth V1 dataset and demonstrate the robustness and effectiveness of RPSL for scene classification.**

***Index Terms*—Dimension reduction, Kinect sensor, rank preserving and sparse learning, RGB-D sensor, scene classification.**

I. Introduction

SCENE classification receives intensive attention as it benefits many practical applications, such as content-based image retrieval [8], [39], and [47], robotics path planning [55], [59], and image annotation [32], [44]. Tamura *et al*. [52] explained scene classification as a procedure that the basic local measurements, e.g., edge detection and region analysis, are successively integrated into representative patterns.

In general, scene classification can be considered a viewpoint-independent object recognition problem [20], [36], and [46] but a scene is constituted by a number of entities. For example, consider an indoor scene, which may contain chairs, desks, people, and bookshelves in an unpredictable fashion. Typically, it is accomplished by the following important steps. First, the visual features are computed from a collection of training images. Second, an efficient model of dimension reduction is trained based on training images to retain the most effective features for the subsequent scene classification. Third, a suitable classifier is selected for final classification.

Despite all recent efforts in computer vision and robotics, the scene classification problem remains largely unsolved. Scene classification is difficult because of sophisticated environments and the variations of the illumination conditions in real-world situations. Thus, many approaches have been developed to extract robust features to represent scene images.

Low-level visual features, such as color, texture, and shape, have gained prominence and shown many merits in scene classification. The HSV color histogram [49], color moment [31], and color coherence vector [38] are invariant to resolution and perspective changes and, thus, perform better than traditional texture and shape features. However, the color descriptors are sensitive to light conditions [41]. Although texture analysis is valuable for many tasks in computer vision [29], [34], empirical studies [35], [48], and [59] indicate detailed micro textural information is not helpful for scene classification. Shape features [35], [42], [58], and [59] have been validated to be effective for scene classification.

In contrast to low-level global visual features, local features encode features on interest points or regions [43]. Scale invariant feature transform (SIFT) [25] is popular for extracting local features. By utilizing integral images, speeded up robust features (SURF) [2] reduce the computations of local gradient histograms. Local binary patterns (LBP) [12] are proposed for texture classification originally. It estimates the local geometric structure of an image based on a non-parametric method and has been widely used in facial image description [19].

The success of Microsoft Kinect [16], [30] opens an innovative channel to add the effective depth information to the visual recognition in the 2-D space. Note that in contrast to conventional time-of-flight (ToF) and light detection and ranging (LIDAR) techniques, Microsoft Kinect, which equips consumer high-resolution depth and visual sensors, can

Manuscript received September 16, 2012; revised March 21, 2013; accepted May 13, 2013. Date of publication July 3, 2013; date of current version September 11, 2013. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2012CB316400, the National Natural Science Foundation of China under Grant 61125106, Grant 61075021, Grant 61201348, Grant 91120302, and Grant 61072093, the National Science and Technology Support Plan under Grant 2013BAH65F01 and Grant 2013BAH65F04, the Guangdong Natural Science Funds under Grant S2011020000541, the Guangdong Scientific and Technology Research Plan under Grant 2012A010701001, the Fundamental Research Funds for the Central Universities of China under Grant 2012ZP0002 and Grant D2116320, and the Shaanxi Key Innovation Team of Science and Technology under Grant 2012KCT-04. This paper was recommended by Associate Editor L. Shao.

D. Tao, L. Jin, and Z. Yang are with the School of Electronic and Information Engineering, South China University of Technology, GuangZhou 510640, China (e-mail: dapeng.tao@gmail.com; lianwen.jin@gmail.com; eezhao.yang@gmail.com).

X. Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xuelong_li@opt.ac.cn).

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/TCYB.2013.2264285

easily provide high quality synchronized videos of both color and depth. Microsoft Kinect enjoys widespread public acceptance that 10 million units have been sold by January 2012. Thus, Kinect sensor (RGB-D Sensor) has attracted a lot of attention in computer vision. More and more labeled datasets are publicly available. Janoch *et al*. [21] released Berkeley 3-D Object Dataset (B3DO), which contains RGB and depth image pairs gathered in real domestic and office environments. Lai *et al*. [22] built a large-scale, hierarchical multiview objects dataset of everyday objects collected using an RGB-D (Kinect style) camera. Silberman and Fergus [46] established a challenging indoor scene dataset NYU Depth. Each image of NYU Depth dataset has a preprocessing depth map with the corresponding manual labels. Remarkably, a number of effective methods have been proposed to utilize the information of depth to improve the performance of object recognition. Bo *et al*. [5] proposed a set of kernel descriptors to extract features from depth images. Janoch *et al*. [21] extracted the traditional histogram of oriented gradients (HOG) from the depth image. But experiments show that Depth Hog was suboptimal. Silberman and Fergus [46] utilized the spatial pyramid matching (SPM) [23] to represent the samples by using the local SIFT features extracted from both the RGB image and the corresponding depth image. The approach improves the scene classification performance effectively. SPM has been successfully utilized in the recent state-of-the-art image classification systems [5], [27], [59].

It is well known that the classical SPM scheme provides an effective solution while the classifier is constructed by Mercer kernels. This approach is computational expensive. Inspired by SPM, Yang *et al*. [60] proposed the spatial pyramid matching using sparse coding (ScSPM) feature representation scheme, in which sparse coding technique was used for nonlinear feature representation. ScSPM achieved top level performance for image classification. However, the computation of ScSPM is expensive. Thus, locality-constrained linear coding (LLC) [57] is more suitable, because LLC can use a linear SVM classifier to obtain good performance of object classification. LLC is based on local coordinate coding (LCC) [63] and explores the locally linear characteristic of the sample distribution. The effectiveness of LLC is ensured by the several attractive properties, i.e., better reconstruction, local smooth sparsity, and analytical solution.

Besides robust visual features for image representation, dimension reduction is essential because the dimension of the image represented by LLC is high. This high dimensionality limits applications of LLC in Kinect based scene classification due to limited computational resources. Dimension reduction results in a succinct yet effective representation of a high-dimensional sample. Over the past decades, although classical linear dimension reduction algorithms [14], [18], [50], and popular manifold learning algorithms [53] have been largely proposed to reduce the data dimensionality for classification tasks, there is big room to improve the efficiency and stability. First, the Euclidean metric has been generally considered to suffer from the concentration of measure phenomenon [4], [11]. Extensive experiments [16], [19], [29] confirmed the importance of the ranking of neighbors for characterizing the data distribution property. Thus, the preservation of the rank order in supervised manifold learning benefits to recover the intrinsic geometry of the data distribution. Second, there are considerable interests and successes on sparse learning algorithms to obtain the parsimony property. Third, it is necessary to minimize the classification error in dimension reduction. The design will improve the accuracy of the subsequent classification [61]. In this paper, we introduce the rank order information to improve sparse learning for Kinect based scene classification and present a new dimension reduction algorithm termed rank preserving sparse learning.

Based on the above descriptions, we conduct the Kinect based scene classification through the following stages: 1) using Kinect to record scene images; 2) applying locality-constrained linear coding (LLC) to local SIFT features to represent the RGB-D images; 3) training rank preserving sparse learning projection matrix by using labeled samples; and 4) classifying the RPSL projected samples. The main contribution of this paper is the newly developed RPSL for Kinect based scene classification. Given the limited page length, the other parts will not be detailed, because they are easy to implement based on the references cited therein.

The rest of the paper is organized as follows. In Section II, we review related works on dimension reduction, which are important for scene classification and the experiment section. We detail the newly proposed rank preserving sparse learning in Section III. Section IV shows the experimental results on the NYU Depth V1 dataset [46]. Section V concludes this paper.

II. Related Work

In the previous section, we have quickly surveyed visual feature extraction and representation for scene images. Since the number of the original visual features is high and all visual features are not equally informative, dimension reduction can effectively reduce this problem. Over the past decades, many dimension reduction approaches have been proposed and applied to the feature selection in the transformed space. We simply grouped these algorithms into two categories and review them as follows.

A. Classical Dimension Reduction Algorithms

The most widely used dimension reduction methods in the visual object classification problems are principal component analysis (PCA) [18] and linear discriminant analysis (LDA) [14], [51]. PCA seeks the principal subspace and projects the data points onto the subspace, in which the variance of the projected points is maximized. Siagian *et al.* [45] used PCA to reduce the dimensionalities of raw gist features. It helps obtain a more practical number of dimensionality while at the same time preserving the variance in the dataset. LDA, which is supervised, aims to best separate the classes of objects when classes are sampled from Gaussians with equal covariance. Ye [61] studied the multiclass classification performance of LDA by utilizing the least squares error. This approach is developed for the homoscedastic Gaussian.

Manifold learning algorithms are developed to tackle high dimensional data. The most representative ones include locally linear embedding (LLE) [40], ISOMap [54] and Laplacian eigenmaps (LE) [3], as well as their linear approximations, such as locality preserving projections [17]. LLE is unsupervised and seeks a low-dimensional, neighborhood-preserving embeddings of the high-dimensional data. ISOMap is a variant of the multidimensional scaling by considering the geodesic distance between samples. It discovers the nonlinearlity of the high dimensional data. LE is a computationally efficient approach for reducing the data dimensionality. It preserves the local geometry of the data and constructs a compact representation of the data lying on the low dimensional manifold. Zhang *et al.* [64] proposed a framework to unify representative dimension reduction algorithms. Yu *et al.* [62] compared the performance of some manifold learning-based dimensionality reduction methods in the application of scene classification.

B. Sparse Learning Based Dimension Reduction

Sparse learning refers to a collection of variable selection algorithms [66], [67] and trades model fitting off the model complexity by adding a sparse penalty to the model. Since sparse learning obtains better interpretability and reduces the computational cost for the subsequent processing, it has become a powerful tool to obtain succinct models of high-dimensional data. Practically, the sparse learning algorithms are useful for understanding large collections of images by finding effective features. Naikal *et al.* [33] utilized sparse PCA (SPCA) [67] to select informative visual features. This approach can effectively eliminate the useless or even harmful terms in the feature representation. But it ignores the class labels that are important for classification. Clemmensen *et al.* [9] proposed sparse discriminant analysis by using a *lasso* penalty. Recently, Cai *et al.* [6] presented a framework to unify several manifold learning based dimension reduction algorithms and obtained their corresponding sparse solutions. Although sparse learning has many merits, it is difficult to find the optimal solution. Note that the least angle regression (LARS) [12] is an efficient and effective tool and can be used to seek a closed-form solution for the situation of the lasso penalty. Furthermore, the accuracy of the solution is high.

III. Rank Preserving Sparse Learning

In this section, we present a new supervised dimension reduction algorithm for scene classification, rank preserving sparse learning (RPSL).

In scene classification, we present the visual information of a scene image by using a group of robust features, i.e., $X = [x_1, x_2, \cdots, x_N] \in R^{D\times N}$ with a D-dimensional visual feature vector $x_i \in R^D$, and each sample has the corresponding class label $C_i \in Z^n$. The objective of dimension reduction is to find a projection matrix $U \in R^{D\times d}$ to linearly map samples from the high-dimensional space R^D to a low-dimensional subspace R^d, with $d < D$, i.e., $Y = U^T X = [y_1, y_2, \cdots y_N] \in R^{d\times N}$. By using Y, an improved classification result can be obtained.

Supervised manifold learning algorithms consider the data intrinsic structure and are dedicated to obtain a low-dimensional sub-manifold to encode the distribution of samples. However, the performance of popular manifold learning algorithms has drastically decreased due to the concentration of the measure phenomenon [4], [11]. Preserving the rank order information can be regarded as an effective and efficient solution in the process of dimension reduction [10], [24]

Sparse learning algorithms aim to find samples sparse representations via variable selection [66], [67]. Thus, by sparse learning, we can obtain an interpreted model and save the cost of computation. Furthermore, decreasing irrelevant features can contribute to the stability of classification. In addition, in order to improve the accuracy of classification, the minimization of classification error is important.

Thus, there are several critical factors to consider for the design of rank preserving sparse learning.

1) In a local patch, it preserves as much as possible the rank order information of the within-class samples and ignores the rank order information of the between-class samples simultaneously, considering the variations in the original distribution resulted by dimension reduction [24];
2) it maximizes the margin between the samples from different classes on a local patch;
3) the L1-norm penalty is introduced to achieve sparse representation; and
4) it models the classification error minimization by utilizing the least squares error minimization.

A. Rank Preserving and Discriminant Analysis

Patch alignment framework (PAF) [64] unifies popular dimension reduction algorithms, such as PCA [18], LDA [14], ISOMap [54], LLE [42], and LE [3], and provides useful understanding to these algorithms. In this paper, the process of the rank order information preserving is developed under the PAF. The development can be reasonably divided into part optimization and whole alignment two stages.

Given a labeled sample x_i, a local patch $X_i = [x_i, x_{i^1}, ...x_{i^{k_1}}, x_{i_1}, ..., x_{i_{k_2}}] \in R^{D\times(k_1+k_2+1)}$can be formed by its k_1 closest within-class samples $x_{i^1}, ..., x_{i^{k_1}}$ and k_2 closest between-class samples $x_{i_1}, ..., x_{i_{k_2}}$. Considering a linear projection mapping $f_i : X_i \mapsto Y_i$, the corresponding low-dimension representation of the local patch is $Y_i = [y_i, y_{i^1}, ...y_{i^{k_1}}, y_{i_1}, ..., y_{i_{k_2}}] \in R^{d\times(k_1+k_2+1)}$. The index set is defined as: $F_i = \left\{i, i^1, \cdots, i^{k_1}, i_1, \cdots i_{k_2}\right\}$.

In order to preserve the within-class rank order information as much as possible, we use the rank matrix R [24], in which each entry R_{ij} is the rank of the sample j with respect to the sample i. Note that the matrix R is not symmetric and cannot be used in PAF straightforwardly. Inspired by the nonlinear unsupervised learning algorithm DD-HDS [24] and the effective rank order information preservation by a sigmoid-like weighting function, we introduce a penalized factor as

$$(w_i)_j = \begin{cases} 1 - \int_{-\infty}^{\|x_i - x_{ij}\|} f(u\,|\mu, \sigma)\, du, & \text{if } x_{ij} \in N_{k_1}(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $f(u|\mu,\sigma)$ is a Gaussian variable probability density function parameterized in terms of a mean μ and a standard deviation σ, and N_{k_1} is the set of k_1 nearest within-class samples of x_i. In addition, Lespinats *et al.* [24] proposed an estimation method of the mean μ and the standard deviation σ, that is

$$\mu = \underset{1\le i<\mathrm{j}\le N}{\mathrm{mean}}\left(d_{ij}\right) - 2(1-\lambda)\underset{1\le i<\mathrm{j}\le N}{\mathrm{std}}\left(d_{ij}\right) \tag{2}$$

$$\sigma = 2\lambda\underset{1\le i<\mathrm{j}\le N}{\mathrm{std}}\left(d_{ij}\right) \tag{3}$$

where the mean and the std are operators to characterize the distribution of distances between all pairwise samples in the high-dimensional space (d_{ij}). The parameter λ affects the performance of dimension reduction and is selected in the range of $[0, 1]$. In a local patch, the penalized factor plays an essential role in the rank order information preservation. It emphasizes the distinction between large and small distances in the high-dimensional space by using small and large weightings, respectively. Thus, it well solves the problem arose by the concentration of the measure phenomenon. The strategy of within-class samples rank preserving can be written as

$$R(y_i) = \sum_{j=1}^{k_1}\|y_i - y_{i^j}\|^2(w_i)_j. \tag{4}$$

For supervised learning, we expect to maximize the margin. The margin can be defined as the sum of the distances between y_i and the k_2 between-class samples

$$M(y_i) = \sum_{p=1}^{k_2}\left\|y_i - y_{i_p}\right\|^2. \tag{5}$$

Therefore the part optimization can be obtained by combining (8) and (9) via a trade-off parameter γ

$$\arg\min_{y_i}\left(\sum_{j=1}^{k_1}\|y_i - y_{i^j}\|^2(w_i)_j - \gamma\sum_{p=1}^{k_2}\left\|y_i - y_{i_p}\right\|^2\right) \tag{6}$$

where $\gamma \in [0, +\infty]$ is a trade-off parameter to integrate the contributions of the two parts. Equation (6) reduces to

$$\begin{aligned}
&\arg\min_{y_i}\sum_{j=1}^{k_1}\left\|y_i - y_{i_j}\right\|^2(w_i)_j - \gamma\sum_{p=1}^{k_2}\left\|y_i - y_{i_j}\right\|^2,\\
&= \arg\min_{y_i}\sum_{j=1}^{k_1+k_2}\left\|y_{F_i\{1\}} - y_{F_i\{j+1\}}\right\|^2(v_i)_j\\
&= \arg\min_{Y_i}\mathrm{tr}\left\{Y_i\begin{bmatrix}-e_{k_1+k_2}^T\\ I_{k_1+k_2}\end{bmatrix}\mathrm{diag}(v_i)\begin{bmatrix}-e_{k_1+k_2} & I_{k_1+k_2}\end{bmatrix}Y_i^T\right\}\\
&= \arg\min_{Y_i}\mathrm{tr}\left(Y_iL_iY_i^T\right)
\end{aligned} \tag{7}$$

where $\mathrm{tr}(\cdot)$ is the trace operator

$$e_{k_1+k_2} = [1,\cdots,1]^T \in R^{k_1+k_2},\ I_{k_1+k_2} = \mathrm{diag}\left(\overbrace{1,\cdots,1}^{k_1+k_2}\right),$$

$$v_i = \left[\overbrace{(w_i)_1,\cdots,(w_i)_{k_1}}^{k_1},\cdots,\overbrace{-\gamma_{k_1+1},,-\gamma_{k_2}}^{k_2}\right]$$

and $L_i = \begin{bmatrix}-e_{k_1+k_2}^T\\ I_{k_1+k_2}\end{bmatrix}\mathrm{diag}(w_i)\begin{bmatrix}-e_{k_1+k_2} & I_{k_1+k_2}\end{bmatrix}$.

Under PAF, we can align all local patches together into a consistent coordinate by utilizing the selection matrix $S_i \in R^{N\times(k_1+k_2+1)}$. The selection matrix is defined as

$$(S_i)_{pq} = \begin{cases}1, & \text{If } p = \mathrm{F_i}\{q\}\\ 0\text{else.}\end{cases}. \tag{8}$$

The coordinate of the low dimensional representation Y_i is then given by $Y = U^TX = [y_1, y_2, \cdots y_N] \in R^{d\times N}$, i.e.

$$Y_i = YS_i. \tag{9}$$

According to (9), the part optimization (12) can be rewritten as

$$\arg\min_{Y}\mathrm{tr}\left(YS_iL_iS_i^TY^T\right). \tag{10}$$

We sum over all the part optimizations defined in (10) over all samples to obtain the whole alignment objective function and then have

$$\begin{aligned}
&\arg\min_{Y}\sum_{i=1}^{N}\mathrm{tr}\left(YS_iL_iS_i^TY^T\right)\\
&= \arg\min_{Y}\mathrm{tr}\left(YLY^T\right)
\end{aligned} \tag{11}$$

where $L = \sum_{i=1}^{N}S_iL_iS_i^T \in R^{N\times N}$ is the alignment matrix. For linearization, we substitute $Y = U^TX$ into (11) and get

$$\arg\min_{U}\mathrm{tr}\left(U^TXLX^TU\right). \tag{12}$$

B. Sparsity Penalty Term

The sparsity of the projection matrix controls the number of nonzero entries. Although the L0-norm of the projection matrix can be directly used to model the sparsity, it results an NP-hard problem that is computationally intractable. In general, the L1-norm of the projection matrix is an alternative way for approximating the L0-norm. Therefore, the objective function can be written as

$$\arg\min_{U}\mathrm{tr}\left(U^TXLX^TU\right) + \lambda\|U\|_1. \tag{13}$$

C. Classification Error Minimization Penalty Term

In order to improve the performance of the subsequent classification, we expect that within-class samples can be mapped to the same point by using the dimension reduction algorithm. Similar to manifold elastic net [65], we utilize weighted PCA to estimate the class centers in the low-dimensional space. The specific procedure is described as follows.

Suppose there are N samples drawn from c classes, and there are n_i samples in the *ith* class. We can obtain the *ith* class center

$$o_i = \left(\frac{1}{n_i}\right)\sum_{j=1}^{n_i}x_j. \tag{14}$$

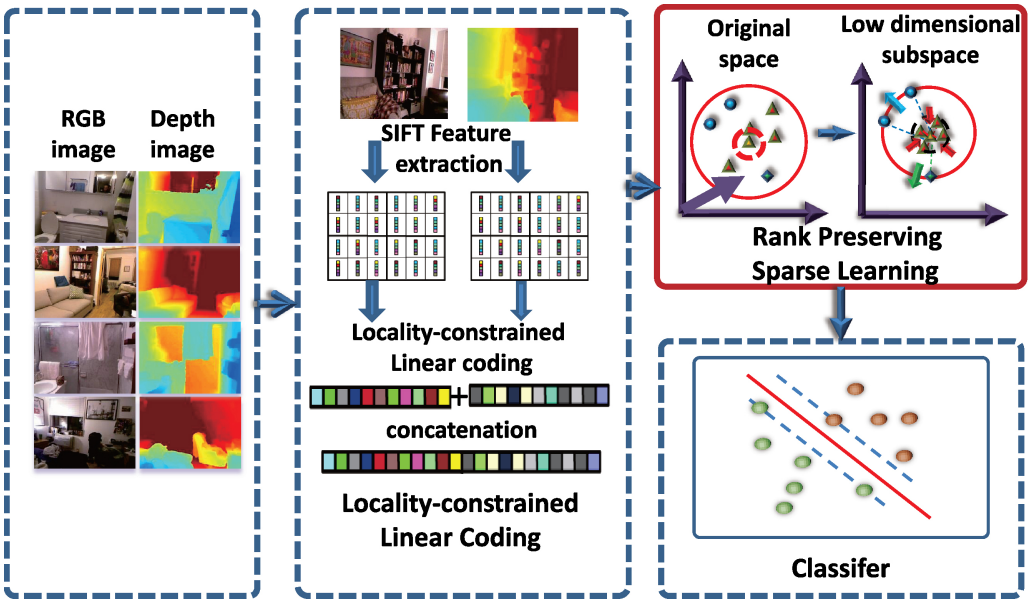


Fig. 1. Framework of rank preserving sparse learning. This scheme contains the following four components: 1) using Kinect to record RGB-D scene images; 2) using locality-constrained linear coding (LLC) to represent the samples by using local SIFT features from which extracted both the RGB image and the corresponding depth image; 3) training rank preserving sparse learning projection matrix by using labeled samples; and 4) classifying the RPSL projected samples.

The corresponding weighted covariance matrix of class centers can be written as

$$C = \sum_{i=1}^{C} \left(\frac{n_i}{N}\right) o_i o_i^T. \tag{15}$$

We conduct standard eigenvalue decomposition and achieve the d eigenvectors associated with the largest d eigenvalues $\zeta = [\zeta_1, \zeta_2, \cdots, \zeta_d]$. Thus, the class center in the d-dimensional subspace can be calculated by

$$o_i' = \zeta^T o_i. \tag{16}$$

The objective incorporated with the classification error minimization can be written as

$$\arg\min_{Y,U} \left\| Z - U^t x \right\|_2^2 + \eta \mathrm{tr}\left(U^T X L X^T U\right) + \lambda \left\| U \right\|_1 \tag{17}$$

where $Z = [z_1, z_2, \cdots z_n]$, and $z_j = o_i'$.

D. Solution for Rank Preserving Sparse Learning

Up to now, the RPSL objective function (15) for dimension reduction has been obtained. It is known that the least angle regression (LARS) can be used to solve the lasso penalized problem effectively, we transform (17) to the style of a quadratic form with the L1-norm penalty

$$\begin{aligned}
&\arg\min_{U} \left\| Z - U^T X \right\|_2^2 + \eta \mathrm{tr}\left(U^T X L X^T U\right) + \lambda \left\| U \right\|_1 \\
&= \arg\min_{U} \mathrm{tr}\left(\left(Z - U^T X\right)\left(Z - U^T X\right)^T\right) \\
&\quad + \eta \mathrm{tr}\left(U^T X L X^T U\right) + \lambda \left\| U \right\|_1 \\
&= \arg\min_{U} \mathrm{tr}\left(U^T X \left(\eta L + I\right) X^T U \right. \\
&\quad \left. - Z X^T U - U^T X Z^T\right) + \lambda \left\| U \right\|_1 \\
&= \arg\min_{U} \mathrm{tr} A + \lambda \left\| U \right\|_1
\end{aligned} \tag{18}$$

$$A = U^T X \left(\eta L + I\right) X^T U - Z X^T U - U^T X Z^T. \tag{19}$$

Since the alignment matrix L is symmetric, we conduct standard eigenvalue decomposition on $\eta L + I$, and get

$$\begin{aligned}
\eta L + I &= B \mathrm{diag}\left(\Lambda_i\right) B^T, \\
&= B \Lambda B^T
\end{aligned} \tag{20}$$

where B is the eigenvector matrix, $\Lambda_i$is the *ith* eigenvalue, and $\Lambda = \mathrm{diag}(\Lambda_i)$ is the diagonal eigenvalue matrix.

Substituting (20) back into (19), we get

$$\begin{aligned}
&U^t X \left(\eta L + I\right) X^T U - Z X^T U - U^T X Z^T \\
&= U^T X \left(B \Lambda^{1/2}\right)\left(\Lambda^{1/2} B^T\right) X^T U - Z \left(B \Lambda^{1/2}\right)\left(B \Lambda^{1/2}\right)^{-1} X^T U \\
&\quad - U^T X \left(B \Lambda^{1/2}\right)\left(B \lambda^{1/2}\right)^{-1} Z^T.
\end{aligned} \tag{21}$$

This implies (18) can be transformed to (22)

$$\begin{aligned}
&\arg\min_{U} \mathrm{tr} A + \lambda \left\| U \right\|_1 \\
&= \arg\min_{U} \left\| \left(B \Lambda^{1/2}\right)^{-1} Z^T - \left(\Lambda^{1/2} B^T\right) X^T U \right\|^2 \\
&\quad - \mathrm{tr}\left(Z \left(B \Lambda B^T\right)^{-1} Z^T\right) + \lambda \left\| U \right\|_1.
\end{aligned} \tag{22}$$

Since $\mathrm{tr}\left(Z \left(B \Lambda B^T\right)^{-1} Z^T\right)$is a constant item, we can ignore it and get a new objective function (23)

$$\begin{aligned}
&\arg\min_{U} \left\| \left(B \Lambda^{1/2}\right)^{-1} Z^T - \left(\Lambda^{1/2} B^T\right) X^T U \right\|^2 + \lambda \left\| U \right\|_1 \\
&= \arg\min_{U} \left\| \hat{Z} - \hat{X} U \right\|^2 + \lambda \left\| U \right\|_1
\end{aligned} \tag{23}$$

where

$$\begin{aligned}
&\hat{Z} = \left[\left(B \Lambda^{1/2}\right)^{-1} Z^T\right] = [z_1, z_2, \cdots, z_d] \in R^{n \times d}, \hat{X} \\
&= \left(\Lambda^{1/2} B^T\right) X^T \in R^{n \times D}, \text{ and } U = [u_1, u_2, \ldots u_d] \in R^{D \times d}.
\end{aligned}$$

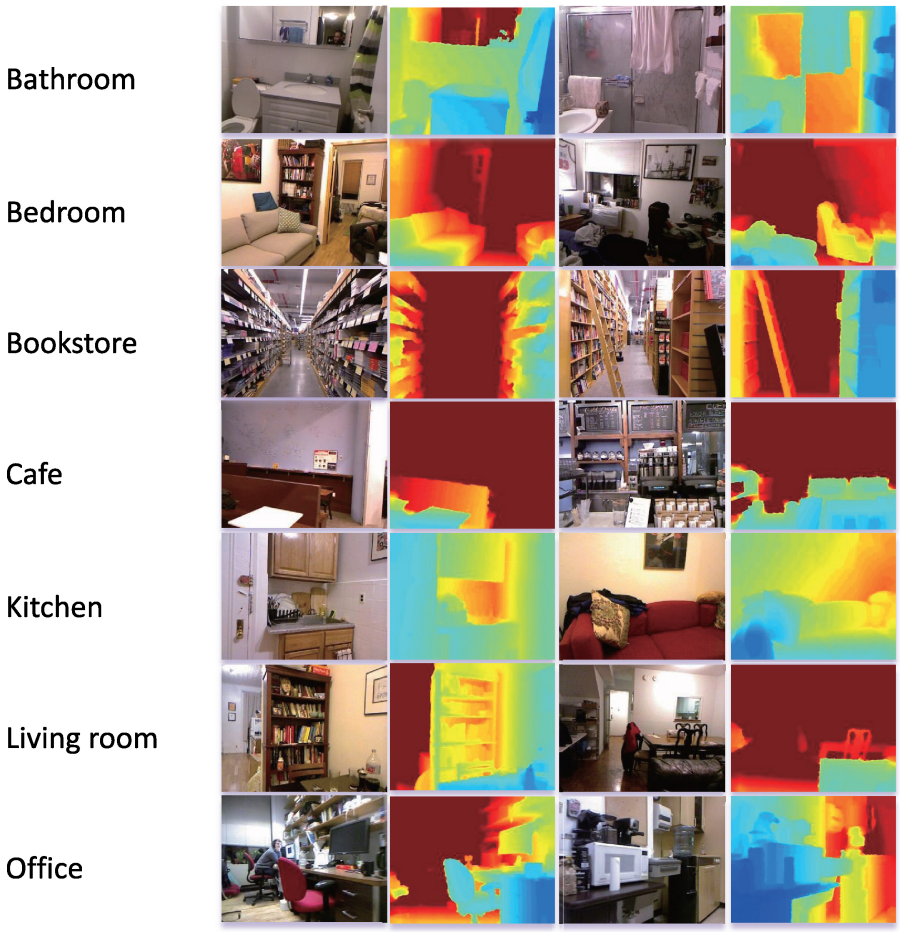


Fig. 2. Example images in the NYU Depth V1 dataset, which contains seven object classes (bathroom, bedroom, bookstore, cafe, kitchen, living room, and office). There are 14 paired samples shown in this figure. For each pair, the RGB color image is shown on the right and the corresponding depth image is on the left (blue = close, red = far).

We can rewrite (23) to

$$\begin{aligned}&\arg\min_{U} \left\|\hat{Z} - \hat{X}U\right\|^2 + \lambda \|U\|_1 \\ &= \sum_{i=1}^{d} \left(\arg\min_{u_i} \left(\left\|\hat{z}_i - \hat{X}u_i\right\|^2 + \lambda \|u_i\|_1 \right) \right). \end{aligned} \tag{24}$$

Given (24), it is straightforward to use LARS to obtain the optimal solution of u_i.

IV. Experimental Results

In this section, we conduct the experiments of scene classification on the NYU Depth V1 dataset [46] to demonstrate the effectiveness of the proposed RPSL. The dataset contains 2,284 samples belonging to seven scene categories. For each image, the SIFT features are extracted from salient regions and locality-constrained linear coding (LLC) are used for feature representation. We compare LLC with ScSPM for feature representation on the NYU Depth V1 dataset, to demonstrate LLC is more suitable than ScSPM for Kinect based scene classification. We measure the performance of our method by using the average accuracy for each scene category. In addition, the confusion matrix, in which each column represents the most likely inferred label information while each row represents the ground truth label information, is used for better understanding of where the approach fails. In addition, we conduct the experiments on the popular fifteen scene categories dataset [1] to further verify the proposed algorithm. The dataset contains 4,485 samples belonging to 15 scene categories. Details of the experimental setup and baseline models are given below.

A. Dataset

The NYU Depth V1 dataset is collected by New York University. It is different from most works applying the depth signal of a scene, and it is worthwhile to highlight the following few points.

1) The scene images have been captured by the Microsoft Kinect and the accurate depth maps can be achieved by utilizing a certain correction technique.
2) There are 64 different indoor environments spread over and all images in the dataset can be grouped into seven categories, including bathroom, bedroom, bookstore, cafe, kitchen, living room, and office.
3) The cross-bilateral filter [37] is a good solution to remove the depth shadow regions.

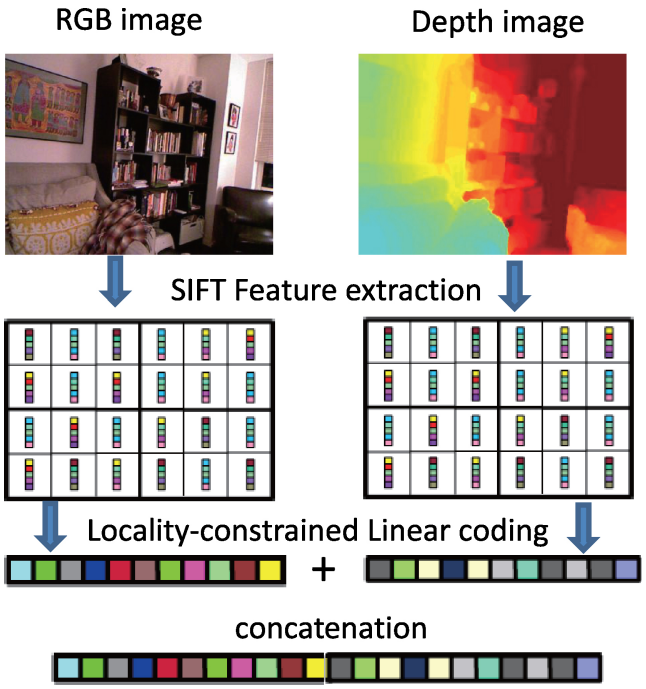


Fig. 3. Locality-constrained linear coding for RGBD-LLC feature.

TABLE I
STATISTICS OF SAMPLES

Scene Classes	Scene	Sample size
Bathroom	6	70
Bedroom	17	463
Bookstore	3	781
Cafe	1	47
Kitchen	10	276
Living Room	13	342
Office	14	305

4) By using 3-D accelerometer in Kinect, the effects of pitch and roll are eliminated in the process of samples collection. In our experiments, we use the scene name information given by the annotation file during the training stage. Example images are given in Fig. 2. There are 14 paired samples that color image is on the right and contrasting depth image is on the left (blue means close and red indicates far). Table I shows the distribution of samples. Note that the data acquisition subjects to the limitation of the Kinect.

The 15 scene categories dataset is expanded upon in [2] and [26]. The dataset consists of 4485 samples collected from 15 scenes. The number of samples of each scene varies from 215 to 410. The samples in the dataset come from a broader range of sources, including personal photographs, the COREL collection, and Google search. Thus, it is suitable for evaluating the scene classification schemes.

In particular, we randomly select $p = 30$ samples per scene category for training, while the remaining samples are used as the test data. The training set was used to learn the orthogonal projection matrix. The test set was used for performance evaluation. We conduct all experiments ten times, and the average recognition rates were calculated for comparing different methods.

B. Feature Descriptor

RGB images and depth images are processed as follows to obtain the feature descriptors. First, all images are transformed into gray scale and resized to be no larger than 300×300 pixels with fixed ratio. Second, the SIFT features of 16×16 pixel patches are extracted over a grid with spacing 8 pixels both in RGB and depth images. The dimension of the SIFT descriptor is 128. Third, we adopt LLC and ScSPM, respectively, in the step of computing feature representation. Parameters are chosen through the empirical evaluations. We perform k-means clustering on all patches from the whole dataset to form a codebook (dictionary). Note that the k-means algorithm selects the initial cluster centers randomly and the termination criterion for iteration is judged by the center and varies in a small range. The vocabulary size of the codebook in our experiment is $M = 1024$. Max pooling is conducted on a 3-level spatial pyramid, partitioned into 1×1, 2×2, and 4×4 subregions. For an RGB and depth image pair, lengths of LLC and ScSPM representations are both $(1 + 4 + 16) \times 1024 = 21{,}504$. Finally, RGB and depth representation are concatenated as one feature vector. Note that the samples of fifteen scene categories dataset only have RGB representation by adopting LLC.

C. Baselines and Performance Evaluation

In this section, we conduct experiments to evaluate the performance of RPSL by comparing it with five representative algorithms, including PCA, SPCA, LDA, supervised LPP (SLPP), and discriminative locality alignment (DLA). Each algorithm has its own merits. PCA and SPCA are unsupervised algorithms. LDA, SLPP, and DLA are supervised algorithms. Before we conduct LDA, SLPP, DLA, and RPSL, the first stage is the PCA projection. In the PCA stage, because the number of the original features is much larger than the number of training samples, N-C dimensions are retained to ensure that within-scatter matrix S_w in LDA [28] is non-singular, where N is the size of samples and C is the number of class. In order to accelerate the learning process, we also conduct PCA step to retain N-1 dimensions in SLPP, DLA and RPSL.

Because the nearest neighbor (NN) classifier does not need to train a model, we use it in the classification stage. We repeat all experiments ten times and the performance is measured by the average accuracy for each class. In addition, to further improve the recognition performance of the proposed dimension reduction algorithm, we use nonlinear support vector machines (SVMs) [15], [56] to replace the NN rule for scene classification. In our experiments, we use LIBSVM [7] to conduct the SVM classification experiments.

To inspect that the depth information can be used to improve the accuracy of the scene recognition, we conduct all experiments on six different feature datasets by utilizing the LLC and ScSPM to represent the SIFT feature extracted from the RGB image, the depth image, and both images (the methods are explained in Section IV-B), respectively. For convenience, descriptors are defined as follows.

1) RGB-LLC: Feature set extracted from RGB images and using the LLC representation.

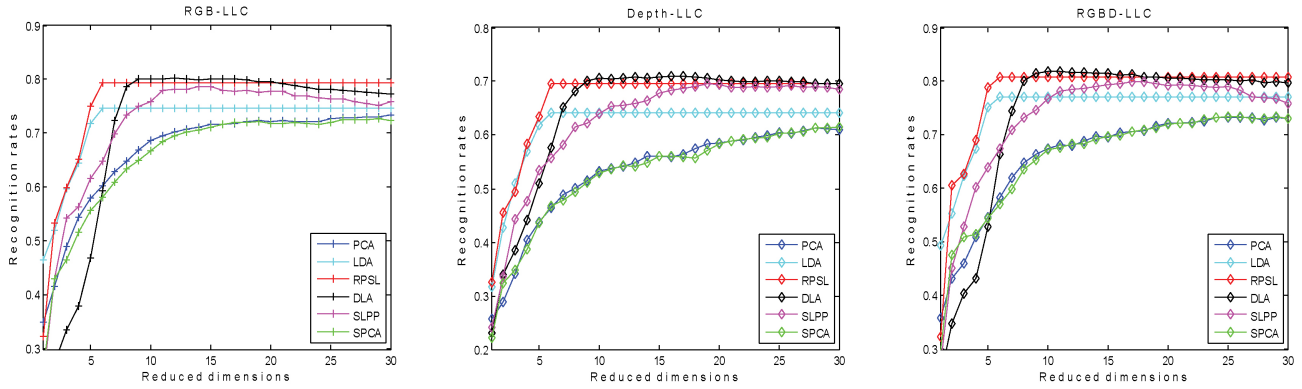


Fig. 4. We compare RPSL with PCA, SPCA, LDA, SLPP, and DLA on three different feature datasets, i.e., RGB-LLC, Depth-LLC and RGBD-LLC. NN classifier is used for recognition. In each subfigure, the x-coordinate is the number of the dimension of all the algorithms on the test set and the y-coordinate is the average recognition.

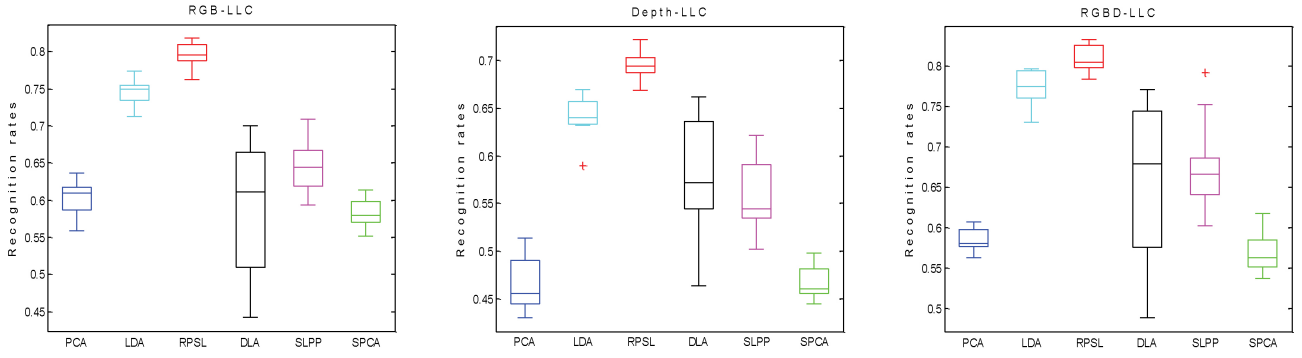


Fig. 5. Box plots of different methods for NYU Depth V1 dataset using LLC. NN classifier is used for recognition. There are three subfigures, each of which corresponds to the performance obtained from a particular feature dataset. For all subfigures, we set the number of dimensionalities of all the algorithms on the test set to 6.

TABLE II

BASELINE OF SCENE CLASSIFICATION RESULTS ON THE NYU DEPTH V1 (%)

Classifier	Linear Support Vector Machines (SVM)											
Feature set	RGB-LLC		RGBD-LLC		Depth-LLC		RGB-ScSPM		RGBD-ScSPM		Depth-ScSPM	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Average accuracy	78.1	1.7	79.9	1.5	68.5	1.5	77.5	1.4	79.0	1.6	68.0	2.4

2) Depth-LLC: Feature set extracted from depth images and using the LLC representation.
3) RGBD-LLC: The concatenation of RGB-LLC and Depth-LLC.
4) RGB-ScSPM: Feature set extracted from RGB images and using the ScSPM representation.
5) Depth-ScSPM: Feature set extracted from depth images and using the ScSPM representation.
6) RGBD-ScSPM: The concatenation of RGB-ScSPM and Depth-ScSPM.

To better illustrate the classification performance of RPSL, the confusion matrix between the ground truth class label and the most likely inferred label is reported.

D. Experimental Results and Analysis

To better demonstrate the effectiveness of our scheme, we consider Linear SVM [13] to classify the feature dataset represented by LLC or ScSPM as a baseline. Table II reports the average accuracy and the standard deviation of the baseline on different feature datasets.

Figs. 4 and 6 compare the proposed RPSL with PCA, SPCA, LDA, SLPP, and DLA on the NYU Depth V1 dataset. The average recognition rate is computed on six different feature datasets and varied with the number of the reduced dimensionalities. The dimension of the RPSL and LDA subspace is C-1, and the dimension of other algorithms subspace is 30. It can be observed that RPSL and DLA are comparable to each other and outperform the others in terms of average recognition rates. This is because RPSL and DLA not only preserve the local geometry of within-class samples but also introduce margin maximization to properly model the discrimination of between-class samples. In particular, in Fig. 4, DLA achieves its highest accuracy 81.87% at the dimensionality 11, while RPSL achieves its highest accuracy 80.82% at the dimensionality 6. In Fig. 6, DLA achieves its highest accuracy 81.55% at the dimensionality 9, while RPSL achieves its highest accuracy 80.28% at the dimensionality 6. Note that in con-

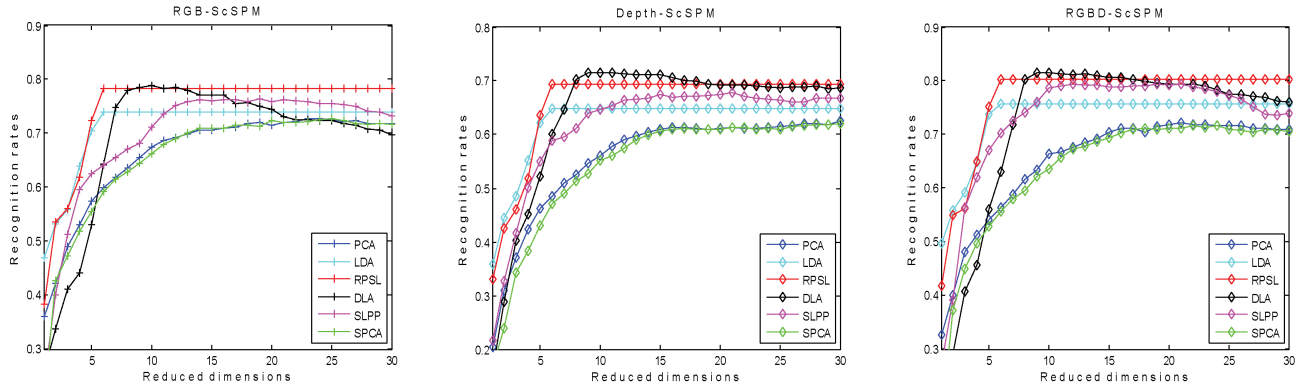


Fig. 6. We compare RPSL with PCA, SPCA, LDA, SLPP and DLA on three different feature datasets, i.e., RGB-ScSPM, Depth-ScSPM, and RGBD-ScSPM. NN classifier is used for recognition. In each subfigure, the x-coordinate is the number of the dimension of all the algorithms on the test set and the y-coordinate is the average recognition.

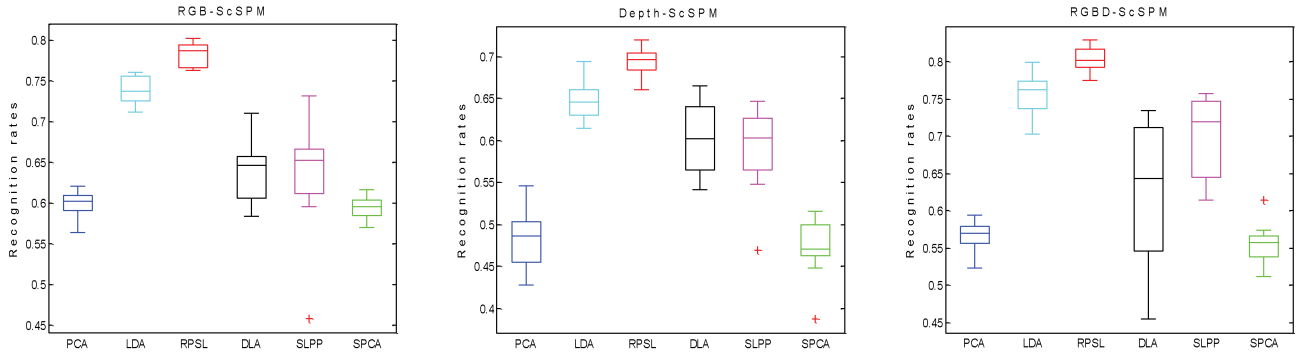


Fig. 7. Box plots of different methods for NYU Depth V1 dataset using ScSPM. NN classifier is used for recognition. There are three subfigures, each of which corresponds to the performance obtained from a particular feature dataset. For all subfigures, we set the number of dimensionalities of all the algorithms on the test set to 6.

trast to other algorithms, RPSL is more efficient and obtains a more compact representation than other dimension reduction algorithms. This is because the lasso penalty effectively helps select the most valuable features for the subsequent classification.

Figs. 5 and 7 show the box-and-whisker plots of different methods. There are six subfigures, each of which corresponds to the performance obtained from a particular feature dataset. For all subfigures, we set the number of dimensionalities of all the algorithms on the test set to 6. It can be observed that RPSL achieves the most robust recognition performance, because it simultaneously considers the within-class nearest neighborhood ranks, between-class nearest neighborhood ranks, classification error minimization of the between-class samples and sparsity.

Fig. 8 shows RPSL classification confusion matrix for one test split. Average classification rates for individual classes are listed along the diagonal. It is not surprising that confusions occur between bedroom and living room, because there are very similar things in these two scenes, such as chair and desk. Also it can be seen from example images in Fig. 2 that there is confusion between bedroom and living room. From the confusion matrix we can also find that Office is confused with Kitchen and Bedroom.

Fig. 9 shows the coefficient path of RPSL obtained by LARS on one training split. According to the proposed RPSL algorithm, all entries of a column of the project matrix are

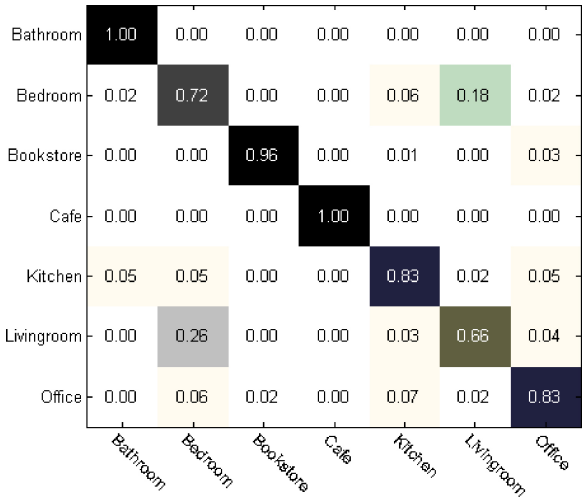


Fig. 8. RPSL classification confusion matrix for one test split. The NN classifier is used for recognition.

set to zero in the initial stage. LARS iteratively finds the most correlated entry and adds it into the active set. In this procedure, we observe that each coefficient path changes its direction when a new variable is added into the active set. These tracks are called coefficient path in LARS and interpret that RPSL selects valuable features iteratively.

TABLE III
SCENE CLASSIFICATION RESULTS ON THE NYU DEPTH V1 USING LLC (%)

Classifier	Nearest Neighbor classifier (NN)						Support Vector Machines (SVM)					
Feature set	RGB-LLC		RGBD-LLC		Depth-LLC		RGB-LLC		RGBD-LLC		Depth-LLC	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
PCA	60.2	2.5	58.3	1.5	46.4	2.7	59.1	1.7	59.0	3.2	48.3	3.8
LDA	74.6	1.9	77.1	2.5	64.2	2.2	77.4	1.3	78.6	2.2	67.3	2.8
RPSL	79.5	1.9	80.8	1.7	69.5	1.7	79.4	1.9	80.8	1.7	69.6	1.7
DLA	59.4	8.7	66.4	9.5	57.6	6.8	54.3	11.9	64.4	12.1	60.2	6.3
SLPP	64.8	3.5	67.5	5.9	55.7	3.8	64.4	3.8	68.2	5.1	56.9	4.0
SPCA	58.2	1.9	57.1	2.4	46.7	1.9	57.6	3.5	57.0	2.3	47.5	3.8

TABLE IV
SCENE CLASSIFICATION RESULTS ON THE NYU DEPTH V1 USING SCSPM (%)

Classifier	Nearest Neighbor classifier (NN)						Support Vector Machines (SVM)					
Feature set	RGB-ScSPM		RGBD- ScSPM		Depth- ScSPM		RGB- ScSPM		RGBD- ScSPM		Depth- ScSPM	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
PCA	59.9	1.7	56.4	2.3	48.6	3.5	58.2	2.4	61.1	2.1	50.5	2.5
LDA	73.9	1.6	75.7	2.8	64.8	2.4	74.9	2.3	78.2	2.2	68.0	1.6
RPSL	78.4	1.5	80.3	1.8	69.3	1.8	78.5	1.6	80.5	1.6	69.6	1.6
DLA	64.4	4.2	63.1	9.8	60.1	4.4	61.5	5.1	64.0	8.9	66.8	2.3
SLPP	64.0	7.7	70.2	5.5	58.8	5.2	63.8	7.4	66.1	4.7	54.9	7.1
SPCA	59.3	1.5	55.7	2.7	47.0	3.6	57.6	3.1	58.6	2.7	48.6	3.6

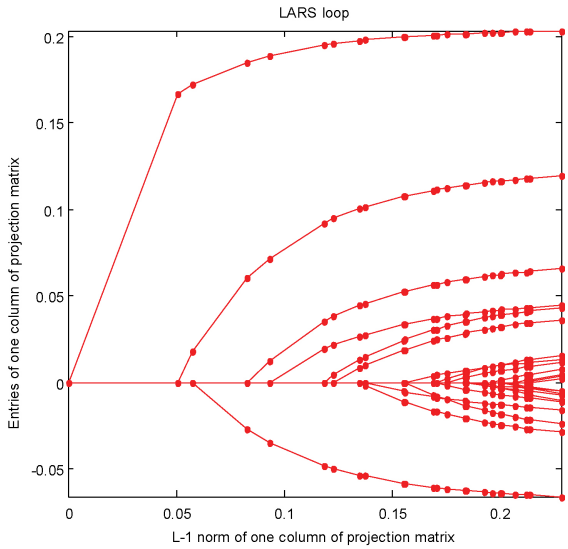


Fig. 9. Coefficient path of LARS for RPSL. This figure shows the entries of the first column of the projection matrix vs. the L1-norm in RPSL obtained by LARS.

TABLE V
SCENE CLASSIFICATION RESULTS ON THE SCENE 15 USING LLC (%)

d	PCA	LDA	RPSL	DLA	SLPP	SPCA
1	16.7	22.6	24.5	16.6	20.8	10.8
2	29.6	37.9	37.1	29.3	31.4	20.4
3	37.9	46.1	49.2	35.1	38.6	34.1
4	43.2	51.6	56.7	40.2	44.2	40.9
5	47.7	55.4	60.4	44.9	49.0	43.4
6	50.2	58.5	61.2	49.0	51.2	48.2
7	52.4	61.9	61.8	51.1	53.8	50.4
8	53.4	64.2	63.9	53.4	56.9	52.8
9	54.0	66.3	66.6	57.4	59.9	53.8
10	55.2	67.9	68.3	59.6	61.1	54.3
11	56.0	68.9	67.2	62.1	63.6	55.2
12	56.3	70.0	68.3	67.0	64.5	55.8
13	56.7	71.2	71.5	71.0	65.9	56.3
14	56.6	72.0	76.1	75.2	66.5	56.5

d is the reduced dimensions.

In addition, by utilizing LIBSVM, we conduct the SVM (with Gaussian kernel) classification experiments to investigate the influence of classifier. Table III and IV report the average accuracy and standard deviation of all the algorithms on the test sets of all splits. The number of dimensionalities of all the algorithms is set to six.

In Table V, we compare the proposed RPSL with PCA, SPCA, LDA, SLPP and DLA on the fifteen scene categories. To be clear, under the same experimental setting, we have tested the Linear SVM classifier on the LLC represent feature dataset. The average accuracy of scene classification is 74.2%.

The main observations from the recognition accuracy comparisons are as follows.

1) These algorithms can be grouped into three levels according to their average recognition rate. PCA and SPCA are not promising, because they ignore the class label information. DLA and SLPP are at the middle level. Because the class label information is considered, their performance is superior to unsupervised algorithms. RPSL and LDA are at the top level. Since RPSL preserves the rank order information in a local patch, it outperforms LDA.

2) We compare the proposed scheme with baselines, in terms of average accuracy in Tables II–IV. It can be observed that the proposed RPSL are helpful for enhancing the classification performance for LLC based scene classification. This is because the sparsity of RPSL decreases non-discriminative features significantly. When the number of original features is much larger than the number of classes, the sparsity is important for estimating a robust projection matrix for reducing the dimensionality.
3) If we ignore the dimensionality of reduced feature space, DLA achieves the highest accuracy in most situations. However, in contrast to conventional dense projection matrices, a sparse projection matrix interprets the generated low-dimensional representation by linking it with a small number of original features and reduces the computational cost in the testing stage. In addition, the sparse projection matrix has an advantage in psychological interpretations.
4) The NN classifier and SVM perform similarly after dimension reduction.
5) RPSL is a general sparse learning algorithm and can be applied to scene classification without depth information.
6) LLC is more robust than ScSPM on Kinect based scene classification.

V. Conclusion

In this paper, we presented a new dimension reduction method termed RPSL for scene classification. RPSL preserved the rank order information and obtained a sparse projection matrix, so it reduced the concentration of the measure phenomenon and obtained the parsimony in computation. In addition, the minimization of classification error was considered to facilitate classification. By utilizing a series of equivalent transformations, we can transform the objective function of RPSL into a lasso penalized least-squares problem.

Compared to the classical dimension reduction algorithms, such as principal component analysis, linear discriminant analysis, discriminative locality alignment, supervised locality preserving projections, and sparse principal component analysis, RPSL showed many competitive and attractive properties for Kinect-based scene classification.

References

[1] C. C. Aggarwal, A. Hinneburg, and D.A. Keim, "On the surprising behavior of distance metrics in high-dimensional space," in *Lecture Notes in Computer Science*, vol. 1973, J.V. Bussche, V. Vianu Eds. Berlin, Germany: Springer, 2001 pp. 420–434.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," *Comput. Vision Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[3] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Adv. Neural Inf. Process. Syst.*, vol. 14, pp. 585–591, Dec. 2001.

[4] R. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ, USA: Princeton Univ. Press, 1961.

[5] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," *in Proc. Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 821–826.

[6] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. IEEE Int. Conf. Comput. Vision*, Oct. 2007, pp. 1–8.

[7] C.-C. Chang and C.-J. Lin. (2001). *LIBSVM: A Library for Support Vector Machines* [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm, 2001.

[8] Y. Chen, J. Z. Wang, and R. Krovetz, "Content-based image retrieval by clustering," in *Proc. ACM Int. Conf. Multimedia Information Retrieval*, 2003, pp. 193–200.

[9] L. Clemmensen, T. Hastie, and B. Ersbøll, "Sparse discriminant analysis," Technical Univ. Denmark and Stanford Univ, Tech. Rep., 2008.

[10] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil, "Genomic signature: Characterization and classification of species assessed by chaos game representation of sequences," *Mol. Biol. Evol.*, vol. 16, no. 10, pp. 1391–1399, Oct. 1999.

[11] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," in *Proc.* Amer. Math. Soc. Lecture, Los Angeles, CA, USA, Aug. 2000.

[12] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Stat*, vol. 32, no. 2, pp. 407–499, 2004.

[13] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Machine Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

[14] R.A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[15] B. Geng, D. Tao, C. Xu, L. Yang, and X. Hua, "Ensemble manifold regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1227-1233, Jun. 2012.

[16] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft Kinect sensor: A review," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1318–1334, Oct. 2013.

[17] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Advan. Neural. Inform. Proc. System*, vol. 16. Dec. 2004, pp. 153–160.

[18] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, 1933.

[19] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," *IEEE Trans. Syst., Man, Cybern., C*, vol. 41, no. 6, pp. 765–781, Nov. 2011.

[20] K. Huang, D. Tao, Y. Tang, X. Li, and T. Tan, "Biologically inspired features for scene classification in video surveillance," *IEEE Trans. Syst., Man, Cybern., B*, vol. 41, no. 1, pp. 307-313, Feb. 2011.

[21] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3-D object dataset: Putting the Kinect to work," in *Proc. ICCV Workshop Consumer Depth Cameras Comput. Vision*, 2011.

[22] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multiview RGB-D object dataset," in *Proc. ICRA*, May 2011, pp. 1817–1824.

[23] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recogn.*, Jun. 2006, pp. 2167–2178.

[24] S. Lespinats, M. Verleysen, A. Giron, and B. Fertil, "DD-HDS: A tool for visualization and exploration of high dimensional data," *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1265–1279, Sep. 2007.

[25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[26] F.-F. Li, and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *IEEE Conf. Comput Vision Pattern Recogn.*, vol. 2. Jun. 2005, pp. 524–531.

[27] L. Liu, L. Shao, and P. Rockett, "Human action recognition based on boosted feature selection and naïve Bayes nearest-neighbor classification," *Signal Process.*, vol.93, no. 6, pp. 1521–1530, 2013.

[28] J. Lu, K. Plataniotis, and A. Venetsanopoulos, "Face recognition using LDA-based algorithms," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 195–200, Jan. 2003.

[29] J. Mao and A. K. Jain, "Texture classification and segmentation using multiresolution simultaneous autoregressive models," *Pattern Recogn.*, vol. 25, no. 2, pp. 173–188, 1992.

[30] Microsoft Kinect [Online]. Available: http://www.xbox.com/en-us/kinect

[31] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons, "Moment invariants for recognition under changing viewpoint and illumination," in *Proc. Comput. Vision Image Understand.*, vol. 94, no. 1–3. 2004, pp. 3–27.

[32] F. Monay, and D. Gatica-Perez, "On image auto-annotation with latent space models," in *Proc. ACM Int. Conf. Multimedia, 2003*, pp. 275–278.

[33] N. Naikal, A. Yang, and S. S. Sastry, "Informative feature selection for object recognition via sparse PCA," Electrical Engineering and Computer Sciences, Tech. Rep., UCB, No. UCB/EECS-2011-27 [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2011/EECS-2011-27.html
[34] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[35] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, 2001.
[36] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer, "Weak hypotheses and boosting for generic object detection and recognition," in *Proc. IEEE Eur. Conf. Comput. Vision*, May 2004, pp. 71–84.
[37] S. Paris and F. Durand, "A fast approximation of the bilateral filter using a signal processing approach," *Proc. IEEE Eur. Conf. Comput. Vision*, May 2006, pp. 568–580.
[38] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *Proc. ACM Int. Conf. Multimedia*, 1996, pp. 65–73.
[39] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, pp. 1575–1589, Sep.2007.
[40] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[41] K. Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no.9, pp. 1582–1596, Sep. 2010.
[42] L. Shao, L. Ji, Y. Liu, and J. Zhang, "Human action segmentation and recognition via motion and shape analysis," *Pattern Recogn. Lett.*, vol. 33, no. 4, pp. 438–445, Mar. 2012.
[43] L. Shao and R. Mattivi, "Feature detector and descriptor evaluation in human action recognition," in *Proc. ACM Int. Conf. Image Video Retrieval*, Jul. 2010, pp. 477–484.
[44] Y. Shao, Y. Zhou, X. He, D. Cai, and H. Bao, "Semi-supervised topic modeling for image annotation," in *Proc. ACM Int. Conf. Multimedia,* 2009, pp. 521–524.
[45] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 300–312, Feb. 2007.
[46] N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," in *Proc. ICCV Workshop 3-D Representation Recogn.,* Nov. 2011, pp. 601–608.
[47] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.
[48] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 174–184, Jan. 2010.
[49] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.
[50] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for Gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.
[51] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.
[52] H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *IEEE Trans. Syst., Man, Cybern.*, vol. 8, no. 6, pp. 460–473, Jun. 1979.
[53] D. Tao and L. Jin, "Discriminative information preservation for face recognition," *Neurocomputing*, vol. 91, pp. 11–20, Aug. 2012.
[54] J. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, Dec. 2000.
[55] A. Ude and R. Dillmann, "Vision-based robot path planning," *Adv. Robot Kinematics Comput. Geometry*, pp. 505–512, 1994.
[56] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.
[57] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recogn.*, Jun. 2010, pp. 3360–3367.
[58] D. Wu and L. Shao, "Silhouette analysis based action recognition via exploiting human poses," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 2, pp. 236–243, Feb. 2012.
[59] J. Wu, and M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, Aug. 2011.
[60] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2009, pp. 1794–1801.
[61] J. Ye, "Least squares linear discriminant analysis," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 1087–1093.
[62] J. Yu, D. Tao, Y. Rui, and J. Cheng, "Pairwise constraints based multiview features fusion for scene classification," *Pattern Recogn.*, vol. 46, no. 2, pp. 483–496, Feb. 2012.
[63] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," *Adv. Neural Inform. Process. Syst.*, vol. 22, pp. 2223–2231, Dec. 2009.
[64] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1299–1313, Sep. 2009.
[65] T. Zhou, D. Tao, and X. Wu, "Manifold elastic net: A unified framework for sparse dimension reduction," *Data Mining Knowl. Discov.* vol. 22, no. 3, pp. 340–371, May 2011.
[66] H. Zou, "The adaptive Lasso and its Oracle properties," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.
[67] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 262–286, 2006.

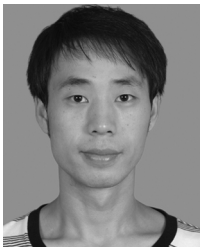
Dapeng Tao received the B.S. degree in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China. He is currently pursuing the Ph.D. degree in information and communication engineering at the South China University of Technology, Guangzhou, China.

His current research interests include machine learning, computer vision, and cloud computing.

Lianwen Jin (M'98) received the B.S. degree from the University of Science and Technology of China, Anhui, China, and the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 1991 and 1996, respectively.

He is currently a Professor with the School of Electronic and Information Engineering, South China University of Technology. He is the author of more than 100 scientific papers. His current research interests include image processing, handwriting analysis and recognition, machine learning, cloud computing, and intelligent systems.

Dr. Jin is a member of the China Image and Graphics Society and the Cloud Computing Experts Committee of the China Institute of Communications. He was a recipient of the award of New Century Excellent Talent Program of MOE in 2006 and the Guangdong Pearl River Distinguished Professor Award in 2011. He served as a Program Committee member for a number of international conferences, including ICMLC2007∼2011, ICFHR2008-2012, ICDAR2009, ICDAR2013, ICPR2010, ICPR2012, ICMLA2012, etc..



Zhao Yang received the B.E. degree in communication engineering from Hubei University, Hubei, China, in 2008. He joined the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China, as a master's student, in 2009 and is currently pursuing the Ph.D. degree in machine learning and computer vision.

Xuelong Li (M'02–SM'07–F'12) is currently a Full Professor with the Center for OPTical IMagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi, China.