SPARSE REPRESENTATION WAVELET BASED CLASSIFICATION

Long H. Ngo[1], Marie Luong[1], Nikolay M. Sirakov[2], Thuong Le-Tien[3], Sébastien Guérif[4], Emmanuel Viennet[1]

[1]L2TI, Université Paris 13, Sorbonne Paris Cité, France
[2] Dept. of Mathematics and Computer Science, Texas A&M Uni.-Commerce, USA
[3] Dept. of Telecommunications, HCM City Uni. of Technology, Vietnam
[4] LIPN, CNRS UMR 7030, Université Paris 13, Sorbonne Paris Cité, France

***Abstract*—This paper improves the conventional sparse representation based classification (SRC) method, through incorporating wavelet coefficients. For this reason, the proposed method is called Sparse Representation Wavelet based Classification (SRWC). In the present study, we fuse the image features described by the complementary information from the low sub-band of the wavelet coefficients and sparse representation to outperform the conventional SRC according to accuracy. This holds because the wavelets promote sparsity and provide structural information about the image, which increases the accuracy of classification. To validate the capabilities and underline the advantages of the novel SRWC, we conducted an extensive number of experiments using publicly available datasets and compared our results with contemporary methods.**

***Keywords*—face/object classification, sparse representation, sparse dictionary, wavelet, low-pass subband.**

I. INTRODUCTION

In recent years, there was an explosion of interest toward machine and deep learning, which helps computers learn from data without being explicitly programmed for that. Over the last decades, sparse representation (SR) modeling has undergone expansion with applications in image processing, machine learning, statistics and computer vision. SR also outperforms many state-of-the-art frameworks in both theoretical research and practical applications in the field of image processing, such as denoising, inpainting, super-resolution [1], segmentation and more recently, in signal classification [2]. Mathematically, SR consists of finding the sparsest solution to an underdetermined linear system. In other words, SR locates the solution with the fewest nonzero entries. Based on the observation that small-scale structures are inclined to repeat themselves in a single image or a group of similar images [3]. Therefore, an image can be sparsely represented over some well-chosen redundant basis.

Classification is a typical task in supervised learning. A fundamental problem in supervised classification is to use labeled training samples from k distinct object classes to predict the category which a new observation belongs to. In this paper, we propose to apply sparse representation based classification in the wavelet domain. The advantage is that we build an overcomplete dictionary, which allows for representing a test sample from a given dataset, by transforming the training samples into the wavelet domain. Hence, the test samples are considered as a linear combination of the transformed training samples into the wavelet domain. This representation is naturally sparse, and help to reject test samples, which do not belong to the dataset [4].

The main contribution of this paper is the fusion of features described by the complementary information from the low sub-band of the wavelet coefficients and sparse coding. It boosts the classification capabilities and enhances the classification accuracy. An extensive number of experimental results on face and object classification demonstrated that the proposed approach outperforms state-of-the-art methods.

The rest of the paper is organized as follow: Section II introduces the related works; Section III presents the wavelet transform and develops the novel method, SRWC; In section IV, SRWC is validated on commonly used datasets, and compared with several contemporary methods; Section V concludes the paper while listing the contributions.

II. RELATED WORK

SRC method [4] assumes that a test sample can be represented as a linear combination of a few basis vectors taken from a dictionary whose base elements are the training samples. More specifically, SRC exploits the linear combination of training samples to represent the test sample by computing the sparse codes of the test sample on the dictionary basis. The reconstruction residuals of each class are computed through the SR coefficients and training samples. The membership of a test sample is determined by the minimum residual. In [4], it is shown that corrupted face images could be recognized by the SRC algorithm, developed for robust face detection. Later, SRC was adapted to numerous image classification problems, such as hyperspectral SRC [5].

According to recent studies [6]–[8], instead of using all the training samples as a dictionary, learning a dictionary from them could effectively improve the SRC performance. Based on the theory of K-SVD model [9], discriminative

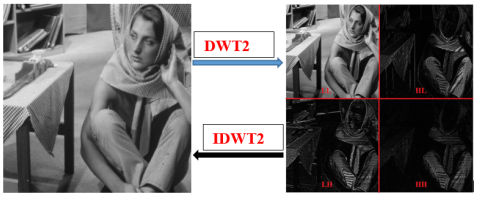


Fig. 1: Barbara image decomposed by DWT2. Left: Original image. Right: Visualization of each sub-band of wavelet coefficients.

K-SVD (D-KSVD) [6] and label consistent K-SVD (LC-KSVD) [7] are constructed to learn a discriminative dictionary, where the sparse codes are projected to be sparse enough. In [6], the authors differentiated LC-KSVD2 from LC-KSVD1 by including the classification error term in the objective function for dictionary learning, which makes the dictionary optimal for the classification task. In [8], the FDDL algorithm employs Fisher discrimination criterion to construct dictionaries and sparse codes.

The SRC methods mentioned above are applied to the spatial domain, to construct a dictionary used in classification. In our approach, we explore the advantages of utilizing wavelet coefficients, which are naturally sparse, for the classification task. The fusion of the wavelet coefficients and SR helps to enhance the classification accuracy.

III. Proposed Method: Sparse Representation Wavelet Based Classification (SRWC)

As proposed in [10], image features can be underlined by projecting the distribution of wavelet coefficients onto the x and y-axes. These projections can be represented by histograms with eight bins in both the x and y-axes. It is also shown that features described by wavelet coefficients can significantly improve the image classification. However, [10] proves that the histograms in high-pass bands are similar, which does not benefit the classification. On the other hand, the histograms in low-pass bands are different. In [10], 16 bins of histograms of projection of wavelet coefficients are exploited as an input for a neural network. Follows that, utilizing the wavelet coefficients in the low-pass band and sparse representation framework for classification would make the results more reliable as shown in Algorithm 1.

A. Single-level Discrete 2D Wavelet Transformation (DWT2)

Consider $\boldsymbol{I}$ as a 2D discrete-space signal (image), where $\boldsymbol{I}(u,v)$ denotes the pixel value. The 2D signal $\boldsymbol{I}(u,v)$ can be treated as 1D signals among the columns $\boldsymbol{I}(u,:)$ at a fixed $u-th$ row and among the rows $\boldsymbol{I}(:,v)$ at a fixed $v-th$ column. A single level 2D wavelet transform of an image can be captured by following the procedure in [11] using Haar kernels.

As illustrated in [11], discrete-time signals $G_L(n)$ and $G_H(n)$ are half-band low-pass and high-pass filters, respectively, defined in the spatial domain as Haar wavelet:

$$G_H(n) = \begin{cases} 1, & 0 \leqslant n < 1/2 \\ -1, & 1/2 \leqslant n < 1 \\ 0, & otherwise \end{cases}; \quad G_L(n) = \begin{cases} 1, & 0 \leqslant n < 1 \\ 0, & otherwise \end{cases} \tag{1}$$

where n denotes the $n-th$ sample of the discrete-time signal.

In wavelet decomposition, the filter $G_L(n)$ is an "averaging" filter while $G_H(n)$ describes details. The 2D Discrete Wavelet Transform (DWT2) [12] decomposes an image into four sub-bands: average (LL), vertical (HL), horizontal (LH) and diagonal (HH) information (Fig. 1). The details of an image, such as object's edges, are represented in the high-pass bands (LH, HL, and HH). While the primary energy of the image is represented in the low-pass band (LL).

B. Training Procedure

The training samples are first passed through DWT2 with Haar wavelet kernel to produce 4 wavelet Sub-Bands (SB):

$$SB = \{LL, HL, LH, HH\} := DWT2(\boldsymbol{I}), \tag{2}$$

where the LL, HL, LH, and HH are sub-bands containing wavelet coefficients for average, vertical, horizontal and diagonal details of the input image.

As stated above, we will only use the LL wavelet coefficients to increase the classification accuracy. Then, principal component analysis (PCA) [13] is employed to reduce the dimension of each vectorized component, which we call an *atom*. We arrange the n_c atoms from the $c-th$ class after being processed via DWT2 and PCA as columns of a matrix $\boldsymbol{D}_c = [\boldsymbol{d}_{c,1}, \boldsymbol{d}_{c,2}, ..., \boldsymbol{d}_{c,n_c}] \in \mathbb{R}^{m \times n_c}$, where m is the dimension of the column vector $\boldsymbol{d}_{c,j},\ j = 1, 2, ..., n_c$. Further, we define a new matrix $\boldsymbol{D}$ to describe the relations between the n atoms from all k categories ($n = \sum_{c=1 \div k} n_c$):

$$\boldsymbol{D} = [\boldsymbol{D}_1, \boldsymbol{D}_2, ..., \boldsymbol{D}_k] = [\boldsymbol{d}_{1,1}, ..., \boldsymbol{d}_{1,n_1}, ..., \boldsymbol{d}_{k,1}, ..., \boldsymbol{d}_{k,n_k}] \tag{3}$$

C. Single Test Sample Detection

Some discriminative models are proposed to exploit the structure of $\boldsymbol{D}_i$ for classification purposes [4]. An approach is considered simple and efficient if it can model the images from a single class as lying on a linear subspace [14]. Subspace models are flexible enough to capture much of the variation in real datasets. It has been observed that the images of faces under varying illuminations and expressions lie on a unique low-dimensional subspace [14]. For ease of presentation, we assume that the training samples from a single class do lie on a subspace, which is only the knowledge our method will use.

Given n_c atoms of the $c-th$ category $\boldsymbol{D}_c = [\boldsymbol{d}_{c,1}, \boldsymbol{d}_{c,2}, ..., \boldsymbol{d}_{c,n_c}] \in \mathbb{R}^{m \times n_c}$, any new sample $\boldsymbol{y} \in \mathbb{R}^{m \times 1}$ (or $\boldsymbol{y} \in \mathbb{R}^m$) from the same class will lie in the linear span of the atoms associated with class c as below:

$$\boldsymbol{y} = \boldsymbol{x}_{c,1}\boldsymbol{d}_{c,1} + \boldsymbol{x}_{c,2}\boldsymbol{d}_{c,2} + ... + \boldsymbol{x}_{c,n_c}\boldsymbol{d}_{c,n_c}, \tag{4}$$

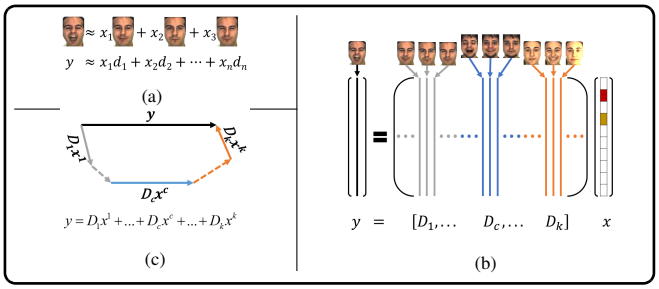


Fig. 2: SRWC main idea: A sample is a linear combination of the other samples from the same class with a sparse vector $\boldsymbol{x}$.

where $\boldsymbol{x}_{c,j} \in \mathbb{R},\ j = 1, 2, ..., n_c$. Then $\boldsymbol{y}$ can be rewritten as the linear combination of the entire set of atoms as below:

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x} \quad \in \mathbb{R}^m, \tag{5}$$

where, ideally, $\boldsymbol{x} = [0, ..., 0, \boldsymbol{x}_{c,1}, \boldsymbol{x}_{c,2}, ..., \boldsymbol{x}_{c,n_c}, 0, ...0]^T \in \mathbb{R}^n$ is a sparse approximation vector whose non zero entries are those associated with the $c-th$ class. Since the entries of the vector $\boldsymbol{x}$ are related to the identity of the test sample $\boldsymbol{y}$, we are able to obtain $\boldsymbol{x}$ by solving the linear system of equation $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}$. When in Eq. (5) $m < n$, the system of equations $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}$ is underdetermined, and $\boldsymbol{x}$ cannot be found in a unique way. Further, this difficulty is resolved by taking the minimum l^2-norm solution:

$$(l^2): \quad \widehat{\boldsymbol{x}_2} = \arg\min \|\boldsymbol{x}\|_2 \quad subject\ to\ \boldsymbol{y} = \boldsymbol{D}\boldsymbol{x} \tag{6}$$

Note that the solution $\widehat{\boldsymbol{x}_2}$ from (6) is not instructive for recognizing the test sample $\boldsymbol{y}$ because $\widehat{\boldsymbol{x}_2}$ has a large number of nonzero entries corresponding to atoms from various classes. To resolve this difficulty in recognition, the vector $\boldsymbol{y}$ can be represented by only the atoms from a single class. On the other hand it is known from [4] that the sparser the code $\boldsymbol{x}$ the higher is the accuracy of classification. Therefore, large number of classes k is needed to make the representation of $\boldsymbol{y}$ sufficiently sparse to provide high accuracy of classification. This leads to the requirement to find the sparsest solution to $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}$ by solving the following optimization problem:

$$(l^0): \quad \widehat{\boldsymbol{x}_0} = \arg\min \|\boldsymbol{x}\|_0 \quad subject\ to\ \boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}. \tag{7}$$

In Eq. (7), $\|.\|_0$ denotes the l^0-norm, which counts the number of nonzero elements in a vector. Nevertheless, it is NP-hard to find the sparsest solution of an underdetermined system of linear equations. Fortunately, if the solution $\boldsymbol{x}_0$ is sparse enough, the solution of the l^0-minimization problem is equivalent to the solution to the l^1-minimization problem as follows [4]:

$$(l^1): \quad \widehat{\boldsymbol{x}_1} = \arg\min \|\boldsymbol{x}\|_1 \quad subject\ to\ \boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}, \tag{8}$$

which can be calculated in polynomial time [15].

Fig. 2 illustrates the main idea of SRWC that one sample is a linear combination of other samples from the same class with sparse $\boldsymbol{x}$.

D. Recognition Results Generation

Given a new observation $\boldsymbol{y}$, we first estimate its sparse representation $\widehat{\boldsymbol{x}_1}$ via solving the problem in Eq. (8). In the perfect case, the nonzero entries in the estimate $\widehat{\boldsymbol{x}_1}$ will be associated with the basis of the dictionary from a single class c; then we can determine the class which $\boldsymbol{y}$ belongs to. Nevertheless, there may be some nonzero entries associated with other categories due to noise and modeling error. To resolve this problem, $\boldsymbol{y}$ can be classified based on how well the coefficients in Eq. (4) are associated with the atoms of each object in the reconstruction of the observation $\boldsymbol{y}$.

For each class c, let δ_c be the characteristic function that selects the coefficients (from $\boldsymbol{x}$) associated only with the $c-th$ class. For $\boldsymbol{x} \in \mathbb{R}^n$, $\delta_c(\boldsymbol{x}) \in \mathbb{R}^n$ is a new vector whose only nonzero entries are the entries in $\boldsymbol{x}$ associated with class c. Using the nonzero entries one can approximate $\boldsymbol{y}$ as $\widehat{\boldsymbol{y}_c} = \boldsymbol{D}\delta_c(\widehat{\boldsymbol{x}_1})$, which is then classified as follow:

$$\min_c\ r_c(\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{D}\delta_c(\widehat{\boldsymbol{x}_1})\|_2. \tag{9}$$

Algorithm 1 summarizes the recognition procedure.

Algorithm 1: SRWC

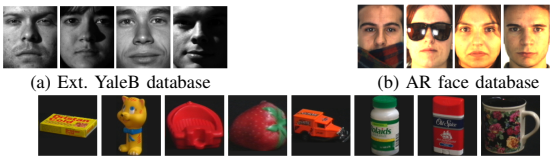
Input : a matrix of entire set of *atoms*, dictionary $\boldsymbol{D} = [\boldsymbol{D}_1, \boldsymbol{D}_2, ..., \boldsymbol{D}_k] \in \mathbb{R}^{m \times n}$ for k classes, and a test sample $\boldsymbol{y} \in \mathbb{R}^m$ in wavelet domain, and an error tolerance $\varepsilon > 0$

1. Normalize the columns of $\boldsymbol{D}$ to have unit l^2-norm.
2. Solve the l^1-minimization problem Eq. (8) in the form $\widehat{\boldsymbol{x}_1} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1\ subject\ to\ \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2 \leqslant \varepsilon$.
3. Compute the residuals $r_c(\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{D}\delta_c(\widehat{\boldsymbol{x}_1})\|_2\ for\ c = 1, ..., k$

Output: $identity(\boldsymbol{y}) = \arg\min_c r_c(\boldsymbol{y})$

IV. Experimental Results

We applied the new SRWC to three public databases, examples of which are shown in Fig. 3. The source codes of the LC-KSVD, FDDL methods are provided by the authors of the papers [7], [8]. Feature extraction helps to reduce data dimension and computational cost. For raw images, the corresponding linear system $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}$ is too large to allow robust and fast object recognition. As in [3], the dimensions of the feature space extracted are sufficiently large to correctly compute the sparse representation. For SRWC, features are extracted by following the procedure stated in Section III. For other methods (SRC, LC-KSVD, and FDDL), face feature descriptor is a random face, made by projecting face images onto random vectors using a random projection matrix [4].



(a) Ext. YaleB database (b) AR face database

(c) COIL-100 database

Fig. 3: Examples from three datasets.

A. *Image Databases Preparation*

The Extended YaleB database has 2,414 frontal-face images of 38 people (∼64 images per person) [16]. The images are captured under various laboratory-controlled luminance states; then the images are cropped and normalized to 192x168 pixels. As in [6], we randomly select 30 images from each person for training and the rest (∼34 images) for testing. The dimension of a random-face feature vector is $d = 504$.

The AR face database includes over 4,000 frontal images for 126 individuals. Each subject has 26 pictures [17]. The images are cropped to 165x120 pixels. In the experiment, a subset including 2600 images from 100 classes (50 male and 50 female) is chosen. We randomly select 20 images from every subject for training and the remaining 6 for testing. The dimension of a random-face feature vector is $d = 540$ [6].

The COIL-100 database contains 7200 color images of 100 objects captured with a black background and different lighting conditions. Analogously to [18], ten images of each object are chosen randomly for training, and the rest 62 images are used for testing in our experiment. To obtain the feature vector of each image, we first convert it to grayscale, resize the image to 32x32 pixels, and present it as a 1024-dimensional vector normalized to have unit norm.

All images used in our experiments are converted to the grayscale. The feature vectors are all normalized to have unit norm.

B. *Results*

The overall recognition rates for the Extended YaleB, AR face and COIL-100 datasets are presented in Fig. 4. Our method is coded in Matlab environment and is repeated ten times for each dataset, and the average recognition rates for each method are then reported in Table I. In real-world classification tasks, we often have to deal with lack of large training sets. Fig. 4 illustrates the accuracy of classification according to the number of training samples per class. As one can tell, the accuracy increases along with the gradually growing number of training samples per class and the novel SRWC outperforms the classical SRC [2]. The proposed method is an improvement on the SRC approach, which we are comparing with. Further, we determined throughout experiments that in order to receive accuracy over 80%, the number of training images should be over 10% or 15% of the size of the entire dataset.

TABLE I: Mean accuracy of SRWC and SRC methods. Numbers in parentheses show the training set size per class.

	Ext. YaleB (30)	AR (20)	COIL (10)
SRC [4]	97.54	97.61	81.16
LC-KSVD1 [7]	97.09	97.78	81.37
LC-KSVD2 [7]	97.80	97.70	**81.42**
FDDL [8]	97.52	96.16	77.45
SRWC	**98.06**	**98.39**	81.29

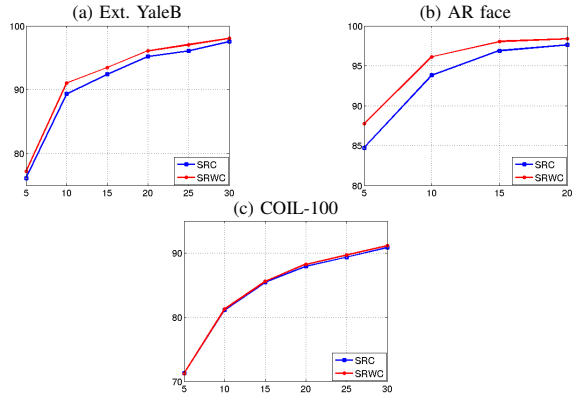


Fig. 4: Comparison on three databases, with classification accuracy (%) as a function of the number of training samples per class.

V. Discussion and Conclusion

The main contribution of the present study is the fusion of the low-band wavelet coefficients and the SRC approach. This fusion boosted the classification capabilities and led to an increase in the classification accuracy of the databases containing images of the same size.

One may tell from Fig. 4, and Table I, SRWC method outperformed the methods SRC, LC-KSVD, FDDL on Extended YaleB and AR, and is very close (0.13%) to the highest accuracy for COIL-100. The recognition rates obtained by SRWC on the two face datasets are promising, compared to LC-KSVD1 [7], LC-KSVD2 [7], and FDDL [8] (Table I). One may derive from Fig. 4 that the higher the number of atoms used, the higher is the recognition rate. Thus, the highest recognition rate of the proposed SRWC is 98.06%, achieved for 30 atoms per class. Considering the AR (20) column in Table I, one may also notice the superiority of the newly proposed SRWC over the conventional SRC methods. A comparison on AR (30) has not been conducted because in AR there are not available 30 training samples per class [17].

Note that the image bases reported in Table I consists of images of the same size. We applied SRWC on Caltech101, Caltech256 and Oxford Flower datasets, where the images are of different size. This led to problems in designing training dictionary. In our future study, we will apply Clifford Algebras, multivectors and patch-wise approach to overcome the problem.

References

[1] M. Elad, M. A. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 972–982, 2010.

[2] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.
[3] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, 1st ed. Springer Publishing Company, 2010.
[4] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE TPAMI*, vol. 31, no. 2, pp. 210–227, 2009.
[5] X. Sun, N. M. Nasrabadi, and T. D. Tran, "Task-driven dictionary learning for hyperspectral image classification with structured sparsity constraints," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 8, pp. 4457–4471, 2015.
[6] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010*. IEEE, 2010, pp. 2691–2698.
[7] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *IEEE TPAMI*, vol. 35, no. 11, pp. 2651–2664, 2013.
[8] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *Inter. J. of Computer Vision*, vol. 109, no. 3, pp. 209–232, 2014.
[9] M. Aharon, M. Elad, and A. Bruckstein, "*k*-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE TSP*, vol. 54, no. 11, pp. 4311–4322, 2006.
[10] W. Zou and Y. Li, "Image classification using wavelet coefficients in low-pass bands," in *International Joint Conference on Neural Networks, 2007. IJCNN 2007.* IEEE, 2007, pp. 114–118.
[11] T. Guo, H. S. Mousavi, T. H. Vu, and V. Monga, "Deep wavelet prediction for image super-resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.
[12] S. Mallat, *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.
[13] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal component analysis*. Springer, 1986, pp. 115–128.
[14] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE TPAMI*, vol. 25, no. 2, pp. 218–233, 2003.
[15] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011*. IEEE, 2011, pp. 1873–1879.
[16] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE TPAMI*, vol. 23, no. 6, pp. 643–660, 2001.
[17] A. M. Martinez, "The ar face database," *CVC Technical Report24*, 1998.
[18] S. Li and Y. Fu, "Learning robust and discriminative subspace with low-rank constraints," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 11, pp. 2160–2173, 2016.