A Sparse Representation Model Using the Complete Marginal Fisher Analysis Framework And Its Applications to Visual Recognition

Ajit Puthenputhussery, *Student Member, IEEE,* Qingfeng Liu, and Chengjun Liu

***Abstract*—This paper presents an innovative sparse representation model using the complete marginal Fisher analysis (CMFA-SR) framework for different challenging visual recognition tasks. First, a complete marginal Fisher analysis method is presented by extracting the discriminatory features in both the column space of the local samples based within class scatter matrix and the null space of its transformed matrix. The rationale of extracting features in both spaces is to enhance the discriminatory power by further utilizing the null space, which is not accounted for in the marginal Fisher analysis method. Second, a discriminative sparse representation model is proposed by integrating a representation criterion such as the sparse representation and a discriminative criterion for improving the classification capability. In this model, the largest step size for learning the sparse representation is derived to address the convergence issues in optimization, and a dictionary screening rule is presented to purge the dictionary items with null coefficients for improving the computational efficiency. Experiments on some challenging visual recognition tasks using representative data sets, such as the Painting-91 data set, the fifteen scene categories data set, the MIT-67 indoor scenes data set, the Caltech 101 data set, the Caltech 256 object categories data set, the AR face data set, and the extended Yale B data set, show the feasibility of the proposed method.**

***Index Terms*—Discriminative sparse representation, complete marginal Fisher analysis, dictionary screening rule, discriminatory features, column space, null space, scatter matrix, visual recognition.**

I. Introduction

Visual recognition, which aims to categorize different visual objects into several predefined classes, is a challenging topic in both computer vision and multimedia research areas. Recently, sparse coding algorithms have been broadly applied in multimedia research, for example, in face recognition [1]–[5], in disease recognition [6], in scene and object recognition [1], [2], [7]–[11], in hand written digit recognition [12], and in human action recognition [13]. Pioneer research in cognitive psychology [14], [15] reveals that the biological visual cortex adopts a sparse representation for visual perception in the early stages as it provides an efficient representation for later phases of processing. Besides, manifold learning methods, such as discriminant analysis [16], [17], marginal Fisher analysis [18], have been successfully applied to preserve data locality in the embeded space and learn discriminative feature representations [18]–[20].

The marginal Fisher analysis (MFA) method improves upon the traditional linear discriminant analysis or LDA by means of the graph embedding framework that defines an intrinsic graph and a penalty graph [18]. The intrinsic graph connects each data sample with its neighboring samples of the same class to define the intraclass compactness, while the penalty graph connects the marginal points of different classes to define the interclass separability. The MFA method, however, does not account for the null space of the local samples based within class scatter matrix, which contains important discriminatory imformation. We present a complete marginal Fisher analysis (CMFA) method that extracts the discriminatory features in both the column space of the local samples based within class scatter matrix and the null space of its transformed matrix. The rationale of extracting features in both spaces is to enhance the discriminatory power by further utilizing the null space, which is not accounted for in the marginal Fisher analysis method.

To further improve the classification capability and to ensure an efficient representation, we propose a discriminative sparse representation model using the CMFA framework by integrating a representation criterion such as the sparse coding and a discriminant criterion. Sparse coding facilitates efficient retrieval of data in multimedia as it generates a sparse representation such that every data sample can be represented as a linear combination of a small set of basis vectors due to the fact that most of the coefficients are zero. Another advantage is that the sparse representation may be overcomplete, allowing more flexibility in matching data and yielding a better approximation of the statistical distribution of the data. Sparse coding, however, is not directly related to classification as it does not address discriminant analysis of the multimedia data. We present a discriminative sparse representation model by integrating a representation criterion, such as the sparse representation, and a discriminative criterion, which applies the new within-class and between-class scatter matrices based on the marginal information, for improving the classification capability. Furthermore, we propose the largest step size for learning the sparse representation to address the convergence issues of our proposed optimization procedure. Finally, we present a dictionary screening rule that discards the dictionary items with null coefficients to improve the computational efficiency of the optimization process without affecting the accuracy.

Our proposed CMFA-SR method is assessed on different visual recognition tasks using representative data sets, such as the Painting-91 data set [21], the fifteen scene categories data

Ajit Puthenputhussery, Qingfeng Liu, and Chengjun Liu are with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA email: {avp38@njit.edu, ql69@njit.edu, cliu@njit.edu}

set [22], the MIT-67 indoor scenes data set [23], the Caltech 101 data set [24], the Caltech 256 object categories data set [25], the AR face data set [16], and the extended Yale B data set [26]. The experimental results show the feasibility of our proposed method.

The contributions of the paper are four-fold. First, we propose a novel complete marginal Fisher analysis method that extracts the discriminatory features in both the column space of the local samples based within class scatter matrix and the null space of its transformed matrix. Second, we develop a new discriminative sparse representation model using the CMFA framework (CMFA-SR) that integrates both the discriminant criterion and the sparse criterion to enhance the discriminative ability. Third, we theoretically derive the largest step size for learning the sparse representation to address the convergence issues. And finally, we present an innovative dictionary screening rule for eliminating dictionary items with null coefficients to improve the computational efficiency of the dictionary encoding process.

This manuscript is an extended version of our ECCV conference paper [27] with the following extensions. (i) We improve the efficiency of the CMFA framework by utilizing the column space of the mixture scatter matrix to transform the local samples based within class scatter matrix and between class scatter matrix. We then apply the null space of the transformed within class scatter matrix to salvage any possible missing discriminatory information (see Section III-B). (ii) We further address the convergence issues of the proposed CMFA-SR method by theoretically deriving the largest step size of the sparse representation and discuss the optimization procedure (see Section IV). (iii) In order to improve the computational efficiency of the proposed method for large dictionary sizes, we present a dictionary screening rule that removes the dictionary items with null coefficients (see Section IV-C). An experimental evaluation of the computational time required for the proposed CMFA-SR method on different dictionary sizes is also presented (see Section V-K). (iv) We include a comprehensive comparative assessment of the proposed CMFA-SR method, the sparse representation methods, the deep learning methods, and some popular image descriptors on seven visual recognition data sets (see Section V). (v) Finally, we discuss the performance of the proposed method on different training sizes and dictionary sizes (see Section V-I and V-H). We also visualize the effect of our CMFA-SR method on the initial input features and show how it encourages better clustering and separation between data-points of different classes (see Section V-J).

The rest of this paper is organized as follows. We review some related work on sparse coding and manifold learning methods in Section 2. In Section 3, we introduce the proposed CMFA-SR model. Section 4 describes the derivation of the largest step size, the optimization procedure, as well as the screening rule. In Section 5, we present extensive experimental results and analysis, and Section 6 concludes the paper.

II. Related Work

In visual recognition applications, several manifold learning methods, such as the locality sensitive discriminant analysis (LSDA) [28], the locality preserving projections [29], the marginal Fisher analysis (MFA) [18], have been widely used to preserve data locality in the embedding space. Cai et al. [28] proposed the LSDA method that maximizes the margins between data points of different classes by discovering the local manifold structure. The MFA method based on the graph embedding framework was presented by Yan et al. [18] by designing two graphs that characterize the intraclass compactness and the interclass separability. A geometric l_p norm feature pooling (GLP) method was proposed by Feng et al. [11] to improve the discriminative power of pooled features by preserving their class-specific geometric information. Different deep learning methods, such as the convolutional neural networks (CNN), the deep autoencoders, and the recurrent neural networks, have received increasing attention in the multimedia community for challenging visual recognition tasks. Krizhevsky et al. [30] developed the AlexNet, which was the most notable deep CNN that contains 5 convolution layers followed by max-pooling layers, and 3 fully connected layers. The ZFNet proposed by Zeiler et al. [31] improved upon the AlexNet architecture by using smaller filter sizes, and developed a method to visualize the filters and weights correctly. He et al. [32] developed residual networks with a depth of upto 152 layers that contain skip connections and inter-block activation for better signal propagation between the layers.

Recently, several sparse representation methods based on supervised learning have been developed for learning efficient sparse representations or incorporating discriminatory information by combining multiple class specific dictionary for different visual recognition applications. The sparse representation methods can be roughly classified into three categories. The first category of the sparse representation methods aims to learn a space efficient dictionary by fusing multiple atoms from the initial large dictionary. Fulkerson et al. [33] proposed an object localization framework that efficiently reduces the size of a large dictionary by constructing small dictionaries based on the agglomerative information bottleneck. Lazebnik et al. [34] presented a technique for learning dictionaries by using the information-theoretic properties of sufficient statistics.

The second category combines multiple class specific subdictionaries to improve the discriminatory power of the sparse representation method. Mairal et al. [35] proposed a sparse representation based framework by jointly optimizing both the sparse reconstruction and class discrimination for learning multiple dictionaries. Zhou et al. [36] presented a joint dictionary learning algorithm that jointly learns multiple class-specific dictionaries and a common shared dictionary by exploiting the visual correlation within a group of visually similar objects. A dictionary learning approach for positive definite matrices was proposed by Sivalingam et al. [37] where the dictionary is learned by alternating minimization of sparse coding and dictionary update stages. Yang et al. [9] proposed a Fisher discrimination learning framework to learn a structured dictionary where each sub-dictionary has specific class labels.

The final category of sparse representation methods cotrains the sparse representation and discriminative dictionary by adding a discriminant term to the objective function. Yang et al. [12] proposed supervised hierarchical sparse coding models where the dictionary is learned via back-projection where implicit differentiation is used to relate the sparse codes with the dictionary. Zhang et al. [5] developed a discriminative K-SVD algorithm to learn an over-complete dictionary by directly incorporating labels in the dictionary learning stage. Jiang et al. [8] presented a label consistent K-SVD algorithm where a label consistency constraint and a classification performance criterion are integrated to the objective function to learn a reconstructive and discriminative dictionary.

III. Sparse Representation Using the Complete Marginal Fisher Analysis

The motivation of this work is to derive a novel learning method by integrating the state-of-the-art feature extraction methods, such as the sparse representation [3] and the marginal Fisher analysis [18], as well as leveraging our research on enhancing discrimination analysis [38], [39]. Specificaly, the pioneer work on the marginal Fisher analysis [18] improves upon the traditional discriminant analysis by introducing K Nearest Neighbors, or KNN samples in the graph embedding framework. Our new complete MFA method further enhances the disriminatory power by introducing two processes that analyze both the column space and the null space of the local (KNN) samples based within-class scatter matrix. In addition, our novel discriminative sparse representation approach fuses both the sparse represetation criterion and the discrimination criterion to improve upon the conventional sparse representation that does not consider classification.

A. Complete Marginal Fisher Analysis

The marginal Fisher analysis or MFA method improves upon the traditional discriminant analysis method by introducing the K Nearest Neighbors or KNN for defining both the intraclass compactness and the interclass separability, respectively [18]. The motivation behind the MFA approach rests on the graph embedding framework that utilizes both the intrinsic graph and the penalty graph [18]. Our recent research also reveals the importance of local smaples, such as the KNN samples, for designing effective learning systems [40], [41]. The application of local samples has its theoretical roots in the statistical learning theory and the stuctrual risk minimization principle in general, and in the design of support vector machines in particular, such as the support vectors, which are local samples. We therefore leverage the ideas of the MFA method and local samples, coupled with the analysis of the column space and the null space of the local (KNN) samples based within-class scatter matrix, and propose our novel complete marginal Fisher analysis method.

Specifically, let the sample data matrix be $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m] \in \mathbb{R}^{h \times m}$, where m is the number of samples of dimension h. Let $\mathbf{W} \in \mathbb{R}^{h \times h}$ be a projection matrix, which will be derived through the following optimization process. The k_1 nearest neighbors based within-class scatter matrix is defined as follows:

$$\mathbf{S}_w = \mathbf{W}^T \mathbf{X}(\mathbf{D} - \mathbf{A})\mathbf{X}^T \mathbf{W} \tag{1}$$

where $\mathbf{A}$ is a binary matrix with nonzero elements $\mathbf{A}_{ij}$ corresponding to the k_1 nearest neighbors of the sample $\mathbf{x}_i$ or the sample $\mathbf{x}_j$ from the same class [27]. $\mathbf{D}$ is a diagonal matrix, whose diagonal elements are defined by the summation of the off-diagonal elements of $\mathbf{A}$ row-wise.

The k_2 nearest neighbors based between-class scatter matrix is defined as follows:

$$\mathbf{S}_b = \mathbf{W}^T \mathbf{X}(\mathbf{D}' - \mathbf{A}')\mathbf{X}^T \mathbf{W} \tag{2}$$

where $\mathbf{A}'$ is a binary matrix with nonzero elements $\mathbf{A}'_{ij}$ corresponding to the k_2 nearest neighbors of the sample $\mathbf{x}_i$ or the sample $\mathbf{x}_j$ from two different classes [27]. $\mathbf{D}'$ is a diagonal matrix, whose diagonal elements are defined by the summation of the off-diagonal elements of $\mathbf{A}'$ row-wise.

Applying the k_1 nearest neighbors based within-class scatter matrix $\mathbf{S}_w$ and the k_2 nearest neighbors based between-class scatter matrix $\mathbf{S}_b$, we are able to derive the optimal projection matrix $\mathbf{W}$ by maximizing the following critirion J_1 [20]:

$$\begin{aligned} J_1 &= \mathbf{tr}(\mathbf{S}_w^{-1}\mathbf{S}_b) \\ &= \mathbf{tr}((\mathbf{W}^T \mathbf{X}(\mathbf{D} - \mathbf{A})\mathbf{X}^T \mathbf{W})^{-1}(\mathbf{W}^T \mathbf{X}(\mathbf{D}' - \mathbf{A}')\mathbf{X}^T \mathbf{W})) \end{aligned} \tag{3}$$

The MFA method first applies pricipal component analysis or PCA for dimensionality reduction [18]. A potential problem with this PCA step is that it may discard the null space of the k_1 nearest neighbors based within-class scatter matrix, which contains important discriminative information. Previous research on linear discriminant analysis shows that the null space of the within-class scatter matrix contains important discriminative information whereas the null space of the between-class scatter matrix contains no useful discriminatory information [42], [43].

We therefore propose a new method, a complete marginal Fisher analysis method, which extracts features from two subspaces, namely, the column space of the k_1 nearest neighbors based within-class scatter matrix $\mathbf{S}_w$ and the null space of the transformed $\mathbf{S}_w$ by removing the null space of the mixture scatter matrix, i.e., $\mathbf{S}_m = \mathbf{S}_w + \mathbf{S}_b$. We then extract two types of discriminantory features in these two subspaces: the discriminantory features in the column space of $\mathbf{S}_w$, and the discriminantory features in the null space of the transformed $\mathbf{S}_w$.

B. Extraction of the Discriminantory Features in Two Subspaces

Let $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_h$ be the eigenvectors of $\mathbf{S}_w$, whose rank is p. The space $\mathbb{R}^h$ is thus divided into the column space, $span\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_p\}$, and its orthogonal complement, i.e., the null space of $\mathbf{S}_w$, $span\{\boldsymbol{\beta}_{p+1}, \boldsymbol{\beta}_{p+2}, ..., \boldsymbol{\beta}_h\}$. Let the transformation matrix $\mathbf{T}_p$ be defined as follows: $\mathbf{T}_p = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_p]$. The k_1 nearest neighbors based within-class scatter matrix $\mathbf{S}_w$ and the k_2 nearest neighbors based between-class scatter matrix $\mathbf{S}_b$ may be transformed into the column space as follows: $\mathbf{S}_w^{'} = \mathbf{T}_p^T \mathbf{S}_w \mathbf{T}_p$, $\mathbf{S}_b^{'} = \mathbf{T}_p^T \mathbf{S}_b \mathbf{T}_p$.

The optimal projection matrix $\boldsymbol{\xi} = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, ..., \boldsymbol{\xi}_p]$ is derived by means of maximizing the following critirion $J_1^{'}$ [20]:

$$\begin{aligned} J_1^{'} &= \mathbf{tr}((\mathbf{S}_w^{'})^{-1}\mathbf{S}_b^{'}) \\ &= \mathbf{tr}((\mathbf{T}_p^T \mathbf{S}_w \mathbf{T}_p)^{-1}\mathbf{T}_p^T \mathbf{S}_b \mathbf{T}_p) \end{aligned} \tag{4}$$

The discriminantory features in the column space of $\mathbf{S}_w$ are derived as follows:

$$\mathbf{U}^c = \boldsymbol{\xi}^T \mathbf{T}_p^T \mathbf{X} \tag{5}$$

The computation of the discriminantory features in the null space of the transformed $\mathbf{S}_w$ consists of the following steps. First, we will discard the null space of the mixture scatter matrix, $\mathbf{S}_m = \mathbf{S}_w + \mathbf{S}_b$, by transforming both $\mathbf{S}_w$ and $\mathbf{S}_b$ into the column space of $\mathbf{S}_m$, respectively: $\mathbf{S}_w^{''}$ and $\mathbf{S}_b^{''}$. The rationale for discarding the null space of the mixture scatter matrix is due to the fact that both the within class scatter matrix and the between class scatter matrix are nullified in this null space. As a result, the null space of the mixture scatter matrix does not carry discriminatory information. Second, we compute the null space of $\mathbf{S}_w^{''}$, and then transform $\mathbf{S}_b^{''}$ into this null space in order to derive the discriminantory features $\mathbf{U}^n$.

Specifically, let $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_k]$ be the transformation matrix that is defined by the eigenvectors of $\mathbf{S}_m$ corresponding to the nonzero eigenvalues, where $k \leq h$. The scatter matrices $\mathbf{S}_w$ and $\mathbf{S}_b$ may be transformed into the column space of $\mathbf{S}_m$ as follows: $\mathbf{S}_w^{''} = \boldsymbol{\alpha}^T \mathbf{S}_w \boldsymbol{\alpha}$, $\mathbf{S}_b^{''} = \boldsymbol{\alpha}^T \mathbf{S}_b \boldsymbol{\alpha}$. Next, we compute the eigenvectors of $\mathbf{S}_w^{''}$, whose null space is spanned by the eigenvectors corresponding to the zero eigenvalues of $\mathbf{S}_w^{''}$. Let $\mathbf{N}$ be the transformation matrix defined by the eigenvectors that span the null space of $\mathbf{S}_w^{''}$. Then, we transform $\mathbf{S}_b^{''}$ into the null space of $\mathbf{S}_w^{''}$ as follows: $\mathbf{S}_b^{'''} = \mathbf{N}^T \mathbf{S}_b^{''} \mathbf{N}$. Finally, we diagonalize the real symmetric matrix $\mathbf{S}_b^{'''}$ and derive its eigenvectors. Let $\boldsymbol{\zeta}$ be the transformation matrix defined by the eigenvectors of $\mathbf{S}_b^{'''}$ corresponding to the non-zero eigenvalues. The discriminantory features in the null space of the transformed $\mathbf{S}_w$ are derived as follows:

$$\mathbf{U}^n = \boldsymbol{\zeta}^T \mathbf{N}^T \boldsymbol{\alpha}^T \mathbf{X} \tag{6}$$

In order to obtain the final set of features, the discriminatory features extracted in the column space and the null space are fused and normalized to zero mean and unit standard deviation.

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}^c \\ \mathbf{U}^n \end{bmatrix} \tag{7}$$

C. Discriminative Sparse Representation Model

In this section, we present a sparse representation model CMFA-SR that uses a discriminative sparse representation criterion with the rationale to integrate a representation criterion such as sparse coding and a discriminative criterion so as to improve the classification performance.

Given m training samples, our complete marginal Fisher analysis method derives the feature matrix: $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m] \in \mathbb{R}^{l \times m}$. Let $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_r] \in \mathbb{R}^{l \times r}$ be the dictionary defined by the r basis vectors and $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m] \in \mathbb{R}^{r \times m}$ be the sparse representation matrix denoting the sparse representation of the m samples. Note that the coefficients $\mathbf{a}_i$ correspond to the items in the dictionary $\mathbf{D}$.

In our proposed CMFA-SR model, we optimize a sparse representation criterion and a discriminative analysis criterion to derive the dictionary $\mathbf{D}$ and the sparse representation $\mathbf{S}$ from the training samples. We use the representation criterion of the sparse representation to define new discriminative within-class matrix $\hat{\mathbf{H}_w}$ and discriminative between-class matrix $\hat{\mathbf{H}}_b$ by considering only the k nearest neighbors. Specifically, using the sparse representation criterion the descriminative within class matrix is defined as $\hat{\mathbf{H}_w} = \sum_{i=1}^m \sum_{(i,j) \in N_k^w(i,j)} (\mathbf{s}_i - \mathbf{s}_j)(\mathbf{s}_i - \mathbf{s}_j)^T$, where $(i,j) \in N_k^w(i,j)$ represents the (i,j) pairs where sample $\mathbf{u}_i$ is among the k nearest neighbors of sample $\mathbf{u}_j$ of the same class or vice versa. The discriminative between class matrix is defined as $\hat{\mathbf{H}}_b = \sum_{i=1}^m \sum_{(i,j) \in N_k^b(i,j)} (\mathbf{s}_i - \mathbf{s}_j)(\mathbf{s}_i - \mathbf{s}_j)^T$, where $(i,j) \in N_k^b(i,j)$ represents k nearest (i,j) pairs among all the (i,j) pairs between samples $\mathbf{u}_i$ and $\mathbf{u}_j$ of different classes. As a result, the new optimization criterion is as follows:

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{S}} \sum_{i=1}^{m} \{||\mathbf{u}_i - \mathbf{D}\mathbf{s}_i||^2 + \lambda ||\mathbf{s}_i||_1\} + \alpha \mathbf{tr}(\beta \hat{\mathbf{H}_w} - (1-\beta)\hat{\mathbf{H}}_b) \\ &s.t. ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., r) \end{aligned} \tag{8}$$

where the parameter λ controls the sparseness term, the parameter α controls the discriminatory term, the parameter β balances the contributions of the discriminative within class matrix $\hat{\mathbf{H}_w}$ and between class matrix $\hat{\mathbf{H}}_b$, and **tr(.)** denotes the trace of a matrix. In order to derive the discriminative sparse representation for the test data, as the dictionary $\mathbf{D}$ is already learned, we only need to optimize the following criterion: $\min_B \sum_{i=1}^t \{||\mathbf{y}_i - \mathbf{D}\mathbf{b}_i||^2\} + \lambda ||\mathbf{b}_i||_1$ where $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_t$ are the test samples and t is the number of test samples. The discriminative sparse representation for the test data is defined as $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_t] \in \mathbb{R}^{r \times t}$. Since the dictionary $\mathbf{D}$ is learned from the training optimization process, it contains both the sparseness and the discriminative information, therefore the derived representation $\mathbf{B}$ is the discriminative sparse representation for the test set.

IV. The Optimization Procedure

In this section, we provide a detailed analysis of the largest step size for learning the sparse representation to address the convergence issues of the algorithm. We also introduce a screening rule to safely remove the dictionary items with null coefficients without affecting the performance to improve the computational efficiency of the proposed model.

A. Largest Step Size for Learning the Sparse Representation

In this section, we present and prove the largest step size for learning the sparse representation using the FISTA algorithm [44]. In particular, after applying some linear algebra transformations, the scatter matrices $\hat{\mathbf{H}_w}$ and $\hat{\mathbf{H}}_b$ in equation 8 can be defined as :

$$\begin{aligned} \hat{\mathbf{H}_w} &= 2\mathbf{S}(\mathbf{D}_{\hat{\mathbf{H}_w}} - \mathbf{W}_{\hat{\mathbf{H}_w}})\mathbf{S}^T \\ \hat{\mathbf{H}}_b &= 2\mathbf{S}(\mathbf{D}_{\hat{\mathbf{H}}_b} - \mathbf{W}_{\hat{\mathbf{H}}_b})\mathbf{S}^T \end{aligned} \tag{9}$$

where $\mathbf{W}_{\hat{\mathbf{H}_w}}$ and $\mathbf{W}_{\hat{\mathbf{H}}_b}$ are matrices whose values $\mathbf{W}_{\hat{\mathbf{H}_w}}(i,j) = 1$ if the pair (i,j) is among the k nearest pairs in the same class otherwise 0, $\mathbf{W}_{\hat{\mathbf{H}}_b}(i,j) = 1$ if the pair (i,j) is among the set $\{(i,j), i \in \pi_c, j \notin \pi_c\}$ otherwise 0, $\mathbf{D}_{\hat{\mathbf{H}_w}}$ and $\mathbf{D}_{\hat{\mathbf{H}}_b}$ are diagonal matrices whose values are $\mathbf{D}_{\hat{\mathbf{H}_w}}(i,i) = \sum_j \mathbf{W}_{\hat{\mathbf{H}_w}}(i,j)$ and $\mathbf{D}_{\hat{\mathbf{H}}_b}(i,i) = \sum_j \mathbf{W}_{\hat{\mathbf{H}}_b}(i,j)$.

Therefore, the objective function of the sparse representation in equation 8 can be converted to the following form:

$$\min_{\mathbf{D},\mathbf{S}} \sum_{i=1}^{m} \{||\mathbf{u}_i - \mathbf{D}\mathbf{s}_i||^2 + \lambda||\mathbf{s}_i||_1\} + \alpha \mathbf{tr}(\mathbf{S}\mathbf{M}\mathbf{S}^T) \\ s.t. ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., r) \tag{10}$$

where $\mathbf{M} = 2(\beta(\mathbf{D}_{\hat{\mathbf{H}_w}} - \mathbf{W}_{\hat{\mathbf{H}_w}}) - (1-\beta)(\mathbf{D}_{\hat{\mathbf{H}}_b} - \mathbf{W}_{\hat{\mathbf{H}}_b}))$ for the proposed CMFA-SR method. We further optimize the objective function in equation 10 by alternatively updating the sparse representation and the discriminative dictionary by decomposing into two separate objective functions for each training sample $\mathbf{u}_i$ given as follows:

$$\min_{\mathbf{s}_i} ||\mathbf{u}_i - \mathbf{D}\mathbf{s}_i||^2 + \alpha M_{ii}\mathbf{s}_i^t\mathbf{s}_i + \alpha\mathbf{s}_i^t\mathbf{g}_i + \lambda||\mathbf{s}_i||_1 \tag{11}$$

where $\mathbf{g}_i = \sum_{j \neq i} M_{ij}\mathbf{s}_j = [g_{i1}, g_{i2}, ..., g_{ik}]^t$ and $M_{ij}(i, j = 1, 2, .., m)$ is the value of the element in the i-th row and j-th column of the matrix $\mathbf{M}$. We optimize the above objective function by alternatively applying the FISTA algorithm [44] to learn the sparse representation and the Lagrange dual method [45] for updating the dictionary. In order to derive the largest step size for learning the sparse representation, we rewrite the objective function in equation 11 in the form of $a(\mathbf{s}_i) + b(\mathbf{s}_i)$, where $a(\mathbf{s}_i) = ||\mathbf{u}_i - \mathbf{D}\mathbf{s}_i||^2 + \alpha M_{ii}\mathbf{s}_i^t\mathbf{s}_i + \alpha\mathbf{s}_i^t\mathbf{g}_i$ and $b(\mathbf{s}_i) = \lambda||\mathbf{s}_i||_1$.

To guarantee the convergence of the FISTA algorithm, an important quantity to be determined is the step size. Given the objective function $F(x) = f(x) + g(x)$, where $f(x)$ is a smooth convex function and $g(x)$ is a non-smooth convex function, the theoretical analysis [46] shows that

$$F(x_k) - F(x^*) \leq \frac{2||x_0 - x^*||^2}{s * (k+1)^2} \tag{12}$$

where x_k is the solution generated by the FISTA algorithm at the k-th iteration, x^* is the optimal solution, and s is the largest step size for convergence. This theoretical result means that the number of iterations of the FISTA algorithm required to obtain an ϵ-optimal solution (x_t), such that $F(x_t) - F(x^*) \leq \epsilon$, is at most $\lceil C/\sqrt{\epsilon} - 1 \rceil$, where $C = \sqrt{2||x_0 - x^*||^2/s}$ Therefore, the step size plays an important role for the convergence of the algorithm and the largest step size can lead to less required iterations for the convergence of the FISTA algorithm.

We now theoretically derive the largest step size required for learning the sparse representation for each training sample.

Proposition 1. *The largest step size that guarantees convergence of the FISTA algorithm is $\frac{1}{Lip(a)}$, where $Lip(a)$ is the smallest Lipschitz constant of the gradient ∇a and $Lip(a) = 2E_{\max}(\boldsymbol{D}^t\boldsymbol{D} + \alpha M_{ii}\boldsymbol{I})$ which is twice the largest eigenvalue of the matrix $(\boldsymbol{D}^t\boldsymbol{D} + \alpha M_{ii}\boldsymbol{I})$.*

Proof. Function $a(\mathbf{s}_i)$ can be generalized as follows:

$$a(\mathbf{x}) = ||\mathbf{D}\mathbf{x} + \mathbf{b}||^2 + \alpha M_{ii}\mathbf{x}^t\mathbf{x} + \alpha\mathbf{x}^t\mathbf{c} \tag{13}$$

Taking the first derivative and finding the difference, we get

$$\nabla a(\mathbf{x}) - \nabla a(\mathbf{y}) = 2(\mathbf{D}^t\mathbf{D} + \alpha M_{ii}\mathbf{I})(\mathbf{x} - \mathbf{y}) \tag{14}$$

The Lipschitz constant of the gradient ∇a satisfies the following inequality

$$||\nabla a(\mathbf{x}) - \nabla a(\mathbf{y})|| \leq Lip(a)||\mathbf{x} - \mathbf{y}|| \tag{15}$$

Therefore, the smallest Lipschitz constant of the gradient ∇a is

$$Lip(a) = 2E_{\max}(\mathbf{D}^t\mathbf{D} + \alpha M_{ii}\mathbf{I}) \tag{16}$$

which is twice the largest eigenvalue of the matrix $(\mathbf{D}^t\mathbf{D} + \alpha M_{ii}\mathbf{I})$.

Hence, as shown in the FISTA algorithm [44], the largest step size that assures the convergence of the FISTA algorithm is the reciprocal of the smallest Lipschitz constant of the gradient ∇a.

□

B. Updating the Dictionary

After the sparse representation $\mathbf{S}$ is learned using the FISTA algorithm, we have to learn the optimal dictionary $\mathbf{D}$. The objective function in equation 10 is a constrained optimization problem with inequality constraints, which may be solved using the Lagrange optimization method and the Kuhn-Tucker condition [45]. In order to solve the primal optimization, we take the first derivative with respect to $\mathbf{D}$ and set it to zero. The dual optimization problem can be formulated as follows:

$$\Lambda^* = \min_{\Lambda} \mathbf{tr}(\mathbf{U}\mathbf{S}^t(\mathbf{S}\mathbf{S}^t + \Lambda)^{-1}\mathbf{S}\mathbf{U}^t + \Lambda - \mathbf{U}^t\mathbf{U}) \tag{17}$$

where Λ is a diagonal matrix whose diagonal values are the dual parameters of the primal optimization problem. We solve the dual problem defined in equation 17 using the gradient descent method and the dictionary $\mathbf{D}$ is updated using the following equation:

$$\mathbf{D} = \mathbf{U}\mathbf{S}^t(\mathbf{S}\mathbf{S}^t + \Lambda^*)^{-1} \tag{18}$$

C. The Dictionary Screening Rule

In this section, we present a dictionary screening rule to improve the computational efficiency during the optimization of the objective function defined in equation 11. During the optimization procedure, the computational complexity is generally introduced due to an oversized dictionary. In our proposed dictionary screening rule, we first identify dictionary items with corresponding coefficient score set as zero by checking the sparse coefficient vectors. We then derive a trimmed dictionary by deleting the zero coefficient dictionary items to improve the computational efficiency. The trimmed dictionary is utilized by the FISTA algorithm [44] to obtain a compact sparse representation. We finally reintroduce the deleted zero coefficients back to compute the final sparse representation. Therefore, the dictionary screening rule improves the computational efficiency of the proposed sparse representation framework by computing a trimmed dictionary utilized by the FISTA algorithm.

The following proposition rule identifies the zero coefficients, so that the corresponding dictionary items may be deleted in order to compute the trimmed dictionary.

Proposition 2. *Given a training sample $\boldsymbol{u}_i(i = 1, 2, .., m)$ and a dictionary item $\boldsymbol{d}_j(j = 1, 2, .., k)$, the sparse coefficient s_{ij} is zero if $|\boldsymbol{u}_i\boldsymbol{d}_j - \frac{\alpha}{2}\boldsymbol{g}_i^t\boldsymbol{I}_j| < (\lambda_{\max} - \sqrt{(||\boldsymbol{d}_j||^2 + \alpha M_{ii})(||\boldsymbol{u}_i||^2 + \frac{\alpha}{4M_{ii}}||\boldsymbol{g}_i||^2)}(\frac{\lambda_{\max}}{\lambda} - 1)$ where s_{ij} is the j-th element of the sparse representation $\boldsymbol{s}_i$, $\lambda_{\max} = \max_{1 \leq j \leq k} |\boldsymbol{u}_i^t\boldsymbol{d}_j - \frac{\alpha}{2}\boldsymbol{g}_i^t\boldsymbol{I}_j|$ and $\boldsymbol{I}_j \in \mathbb{R}^{k \times 1}$ is a vector with zero values for all elements except the j-th element which has a value 1.*

V. Experiments

Our proposed CMFA-SR method has been evaluated on some challenging visual recognition tasks: (i) fine art painting categorization using the Painting-91 data set [21], (ii) scene recognition using the fifteen scene categories [22] and the MIT-67 indoor scenes data set [23], (iii) object recognition using the Caltech 101 data set [24] and the Caltec 256 object categories [], and (iv) face recognition using the AR face database [16] and the extended Yale B data set [26]. Specifically, the data sets used in our experiments are detailed in table I and some sample images are shown in figure 1.

A. Painting-91 Data Set

The Painting-91 data set [21] is a challenging data set of fine art painting images collected from the Internet and contains two tasks: artist classification and style classification. We follow the experimental protocol in [21] which uses a fixed train and test split for both the tasks. The initial features used are fused Fisher vector (FFV) features [49] which are extracted using a hybrid feature extraction step as described in [50]. We further compute the FFV features in different color spaces namely RGB, XYZ, YUV, YCbCr, YIQ, LAB, HSV and oRGB to incorporate color information as the color cue provides powerful discriminatory information.

1) Artist Classification: The artist classification task classifies a painting image to its respective artist and is a challenging task as there are large variations in the appearance, styles and subject matter of the paintings of the same artist. The dictionary size is set as 512, and the parameters λ = 0.05, α = 0.2 and β = 0.4 are selected for the CMFA-SR method. The experimental results are summarized in column 3 of table II. MSCNN is the abbreviation for multi-scale convolutional neural networks. The classification is performed using RBF-SVM with parameters C = 20 and γ = 0.00007. Our proposed method consistently outperforms other popular image descriptors and state-of-the-art deep learning methods for the artist classification task.

2) Style Classification: The style classification task deals with the problem of categorizing a painting to the 13 style classes defined in the data set. For the CMFA-SR method, the dictionary size is set as 256 and the same parameters are used as the artist classification task. The fourth column in table II shows the recognition results. Experimental results demonstrate that our proposed CMFA-SR method achieves better performance compared to other popular image descriptors and deep learning methods for style classification.

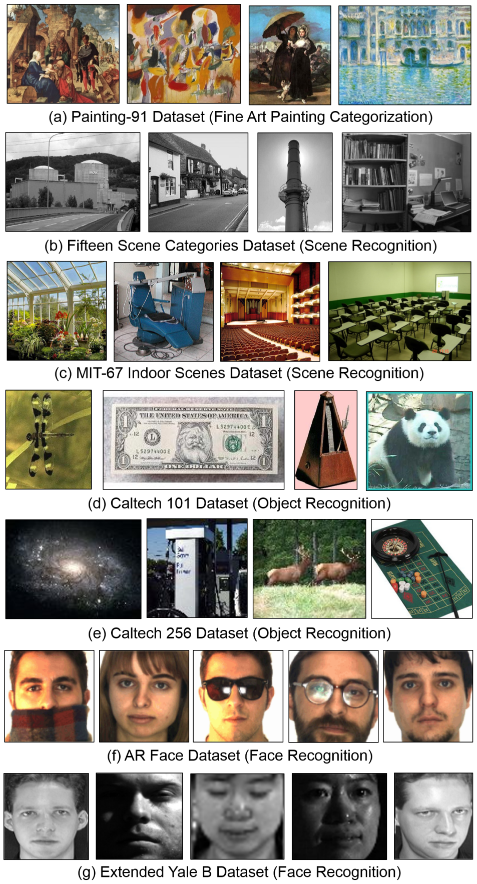
(a) Painting-91 Dataset (Fine Art Painting Categorization)
(b) Fifteen Scene Categories Dataset (Scene Recognition)
(c) MIT-67 Indoor Scenes Dataset (Scene Recognition)
(d) Caltech 101 Dataset (Object Recognition)
(e) Caltech 256 Dataset (Object Recognition)
(f) AR Face Dataset (Face Recognition)
(g) Extended Yale B Dataset (Face Recognition)

Fig. 1. Some example images of the different data sets used for evaluation.

Figure 2 (a) shows the confusion matrix for the 13 style categories of the Painting-91 data set. It can be seen that the style categories with the best performance are 1 (abstract expressionism) and 13(symbolism) with classification rates of 93% and 89% respectively. The most difficult style category to classify is category 6 (neoclassical) as there are large confusions between the style categories baroque and neoclassical. The other style category pairs that create confusion are the styles neoclassical : renaissance and the styles renaissance : baroque.

B. Fifteen Scene Categories Data Set

For the fifteen scene categories data set [22], we follow the experimental protocol as in [22] where for 10 iterations, 100 images per class are randomly selected for each iteration from the data set for training and the remaining images are used for testing. The initial input features used are the spatial pyramid features provided by [8] obtained by using a four-level spatial pyramid with a codebook of size 200. For the CMFA-SR method, the dictionary size is set as 1024 and the parameters λ = 0.05, α = 0.2, and β = 0.4 are selected. The RBF-SVM is used for classification with parameters set as C = 7 and γ = 0.0001. The experimental results in table III show that the proposed method improves upon other popular sparse representation and deep learning methods by more than 5%. Figure 2 (b) shows the confusion matrix for the fifteen scene categories data set.

data set	Task	# Classes	Total # Images	# Train Images	# Test images	Train/Test split	# Folds	Reference of above setting
Painting-91 [21]	artist classification	91	4266	2275	1991	specified [21]	1	[21]
Painting-91 [21]	style classification	13	2338	1250	1088	specified [21]	1	[21]
15 Scenes [22]	scene recognition	15	4485	1500	2985	random	10	[22]
MIT-67 Scenes [23]	scene recognition	67	subset of 15620	5360	1340	specified [23]	1	[23]
Caltech 101 [24]	object recognition	101	9144	random split of 10, 15, 20, 25 and 30 images per class	at most 50 images per class	random	5	[47]
Caltech 256 [25]	object recognition	256	30607	random split of 15, 30, 45 and 60 images per class	at most 25 images per class	random	3	[47]
AR Face [16]	face recognition (setting 1)	126	4000	2520	1480	random	10	[8]
AR Face [16]	face recognition (setting 2)	126	subset of 4000	13 images per class	13 images per class	random	10	[48]
Extended Yale B [26]	face recognition	38	2414	760	1654	random	10	[5]

TABLE I
DIFFERENT TASKS AND THEIR ASSOCIATED DATA SETS USED FOR EVALUATION OF THE PROPOSED CMFA-SR METHOD.

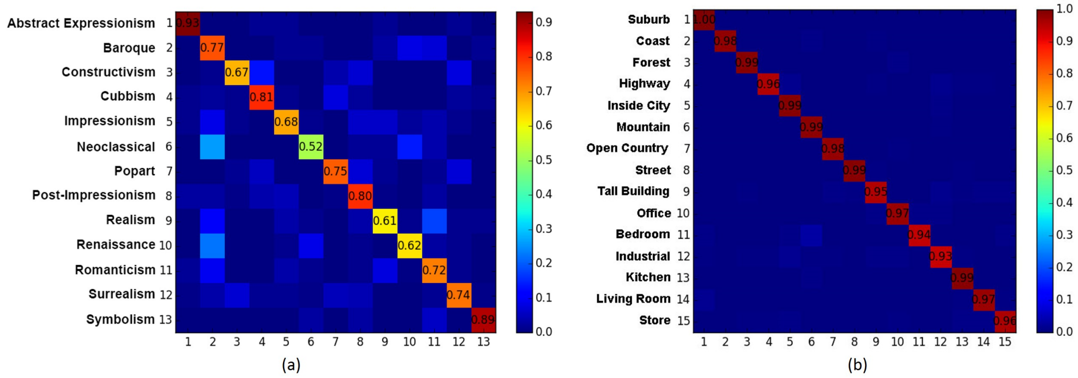

Fig. 2. The confusion matrix for (a)13 style categories of the Painting-91 data set (b) 15 scene categories data set.

C. MIT-67 Indoor Scenes Data Set

The MIT-67 indoor scenes data set [23] is a challenging indoor scenes recognition data set with a variable number of images per category where each category has atleast 100 images. We use experimental settings as in [23] where 80*67 images are used for training and 20*67 images are used for testing. The performance measure provided is the average classification accuracy over all the categories. We extract features for images of the MIT-67 indoor scenes data set using a pre-trained Places-CNN [62]. For the proposed CMFA-SR method, the dictionary size is set as 512 and the parameters $\lambda = 0.05$, $\alpha = 0.1$, and $\beta = 0.5$ are selected, whereas for the RBF-SVM, parameters are set as $C = 2$ and $\gamma = 0.0001$. It can be seen from table IV that our method improves over the performance of Places-CNN by 13%. Our proposed CMFA-SR method helps to significantly improve the initial CNN features by encouraging better separation between the samples of different class and assist in the formation of compact clusters for the samples of same class (see subsection V-J). Experimental results in table IV show that the proposed method is able to achieve significantly better results and outperform other popular sparse representation and deep learning methods.

D. Caltech 101 Data Set

For the Caltech 101 data set [24], we use the experimental settings as in [47], where we randomly split the data set into 10, 15, 20, 25 and 30 training images per category and at the most 50 test images per category in order to have a fair comparison with other methods. The performance measure provided is the average accuracy over all the classes. We evaluate our methods with features that are extracted using a pre-trained convolutional neural network CNN-M [69]. The dictionary size is selected as 512 and the parameters are set as $\lambda = 0.05$, $\alpha = 0.1$, and $\beta = 0.5$ for the CMFA-SR method. The parameters of the RBF-SVM are $C = 4$ and $\gamma = 0.00001$. The experimental results shown in table V show that even without using different fine tuning techniques as in [69], our proposed method is able to achieve comparable results to other state-of-the-art deep learning methods.

E. Caltech 256 Data Set

The Caltech 256 data set [25] is an extended version of the Caltech 101 data set and a more challenging object recognition data set. We follow the experimental settings as specified in [47], where the data set is randomly divided to 15, 30, 45 and

No.	Method	Artist Cls.	Style Cls.
1	LBP [21], [51]	28.50	42.20
2	Color-LBP [21]	35.00	47.00
3	PHOG [21], [52]	18.60	29.50
4	Color-PHOG [21]	22.80	33.20
5	GIST [21], [53]	23.90	31.30
6	Color-GIST [21]	27.80	36.50
7	SIFT [21], [54]	42.60	53.20
8	CLBP [21], [55]	34.70	46.40
9	CN [21], [56]	18.10	33.30
10	SSIM [21], [57]	23.70	37.50
11	OPPSIFT [21], [58]	39.50	52.20
12	RGBSIFT [21], [58]	40.30	47.40
13	CSIFT [21], [58]	36.40	48.60
14	CN-SIFT [21]	44.10	56.70
15	Combine(1 - 14) [21]	53.10	62.20
16	MSCNN-1 [59]	58.11	69.67
17	MSCNN-2 [59]	57.91	70.96
18	CNN F_3 [60]	56.40	68.57
19	CNN F_4 [60]	56.35	69.21
20	**CMFA-SR**	**65.78**	**73.16**

TABLE II
Comparison between the proposed method and other popular methods for artist and style classification task of the Painting-91 data set

Method	Accuracy (%)
LLC [47]	80.57
KSPM [22]	81.40
DHFVC [61]	86.40
D-KSVD [5]	89.10
LaplacianSC [10]	89.70
LC-KSVD [8]	90.40
Places-CNN [62]	90.19
Hybrid-CNN [62]	91.59
DAG-CNN [63]	92.90
CMFA-SR	**98.45**

TABLE III
Comparison between the proposed method and other popular methods on the fifteen scene categories data set

60 training images per category and at the most 25 test images for 3 iterations. The methods are evaluated using features extracted from a pre-trained ZFNet [31]. For the CMFA-SR method, we set the dictionary size to 1024, and the parameters as λ = 0.05, α = 0.1, and β = 0.5. The RBF-SVM is used for classification with C = 2 and γ = 0.0001. The experimental results in table VI show that our proposed method is able to achieve better results compared to other learning methods.

F. AR Face Data Set

For the AR face data set, a subset of the data [16] is selected containing 50 male and 50 female subjects and the images are cropped to 165*120 in order to follow the standard evaluation procedure. We evaluate our proposed method using two common experimental settings to have a fair comparison with other methods. We follow the first experimental setting as in [8] and [5] where we randomly select 20 training images and the remaining are selected for testing, for each person for 10 iterations. The model parameters are set as λ = 0.1, α = 0.2, and β = 0.6 and the dictionary size is selected as 512 for the CMFA-SR method. RBF-SVM is used for classification with parameters set as C= 4, γ = 0.0001.

Method	Accuracy (%)
ROI + GIST [23]	26.10
Object Bank [64]	37.60
Discriminative parts [65]	51.40
VC + VQ [66]	52.30
DP + IFV [67]	60.80
Places-CNN [62]	68.24
Hybrid-CNN [62]	70.80
DAG-CNN [63]	77.50
CMFA-SR	**81.12**

TABLE IV
Comparison between the proposed method and other popular methods on the MIT-67 indoor scenes data set

Method	10	15	20	25	30
SVM-KNN [68]	55.80	59.10	62.00	–	66.20
SPM [22]	–	56.40	–	–	64.60
LLC [47]	59.77	65.43	67.74	70.16	73.44
D-KSVD [5]	59.50	65.10	68.60	71.10	73.00
SRC [3]	60.10	64.90	67.70	69.20	70.70
LC-KSVD [8]	63.10	67.70	70.50	72.30	73.60
CNN-M + Aug [69]	–	–	–	–	87.15
CMFA-SR	**83.11**	**85.88**	**86.95**	**87.61**	**88.28**

TABLE V
Comparison between the proposed method and other popular methods on the Caltech 101 data set

The second experimental setting is defined in [48] where we randomly consider 26 images per person of which 13 images are used for training and the remaining 13 for testing for total of 10 iterations. The dictionary size is set to 512, and the parameters are set as λ = 0.1, α = 0.2, β = 0.5, and C= 1, γ = 0.0007 for the RBF-SVM classifier. The experimental results in table VII using our proposed CMFA-SR method for both the experimental settings show that our method is able to improve upon other popular methods.

G. Extended Yale B Data Set

As for the extended Yale B data set, a common evaluation procedure is to use a cropped version of the data set [26] where the images are manually aligned, cropped and resized to 192 x 168 pixels. The experimental setting as in [74] is followed wherein 20 images per subject are randomly selected for training and the remaining images are used for testing, for a total of 10 iterations. Note that this experimental setting is more difficult than that in [5]. We first scale the image to 42 X 48 and and we obtain the pattern vector using random faces [3]. The dictionary size is selected as 512. We set the parameters λ = 0.06, α = 0.2, and β = 0.5 for the CMFA-SR method. The classification is done using RBF-SVM with parameters C = 4 and γ = 0.001. Experimental results in table VIII show that the proposed method achieves better results compared to other popular methods.

H. Evaluation of the Size of the Dictionary

In this section, we analyze the impact of different dictionary sizes on the performance of the CMFA-SR method. In particular, dictionary sizes of 1024, 512, and 256 are used for a comparative assessment of the performance. The results are

Method	15	30	45	60
ScSPM [70]	27.73	34.02	37.46	40.14
IFK [71]	34.70	40.80	45.00	47.90
LLC [47]	34.36	41.19	45.31	47.68
M-HMP [72]	40.50	48.00	51.90	55.20
ZFNet CNN [31]	65.70	70.60	72.70	74.20
CMFA-SR	**67.85**	**71.44**	**74.27**	**76.31**

TABLE VI
COMPARISON BETWEEN THE PROPOSED METHOD AND OTHER POPULAR METHODS ON THE CALTECH 256 DATA SET

Experimental Setting 1 Method	Accuracy (%)
D-KSVD [5]	95.00
LC-KSVD [8]	97.80
CMFA-SR	**98.95**
Experimental Setting 2 Method	Accuracy (%)
SRC [3]	93.75 ± 1.01
ESRC [73]	97.36 ± 0.59
SSRC [48]	98.58 ± 0.40
CMFA-SR	**98.65** ± 0.42

TABLE VII
COMPARISON BETWEEN THE PROPOSED METHOD AND OTHER POPULAR METHODS ON THE AR FACE DATA SET

presented in figure 3 and we can deduce that the performance of the CMFA-SR method increases upto a certain dictionary size and then reaches a stable performance. We can also observe that for small data sets, a fairly good performance is achieved with a small dictionary size, whereas in case of large data sets such as the Caltech 101, a larger dictionary size is required. This indicates that a large data set requires a larger dictionary as the dictionary captures the variability of the data set.

I. *Evaluation of the Size of the Training Data*

We now evaluate the performance of our proposed CMFA-SR method when different sizes of training images per category are used. Figure 4 shows the performance of the CMFA-SR method for different training data sizes per category on the Caltech 101 data set and 15 scenes data set. The model parameters for both the data sets are set to values used in the corresponding experimental section. It can be observed from figure 4 that the performance of the CMFA-SR method improves with the increase in the size of the training data upto a certain value. After a certain training size, the performance only has minor variations indicating the robustness of the proposed method.

J. *Evaluation of the Effect of the Proposed CMFA-SR Method*

In order to understand the effectiveness of the proposed method, we first examine the effect of the CMFA-SR method using the deep learning features on the MIT-67 data set. We extract the input CNN features extracted using the Places-CNN [62] on the MIT-67 data set. The proposed method then processes these input CNN features to obtain the CMFA-SR features. Finally, the SVM classifier is used for classification.

Method	Accuracy (%)
D-KSVD [5]	75.30
SRC [3]	90.00
FDDL [74]	91.90
CMFA-SR	**94.94**

TABLE VIII
COMPARISON BETWEEN THE PROPOSED METHOD AND OTHER POPULAR METHODS ON THE EXTENDED YALE B DATA SET

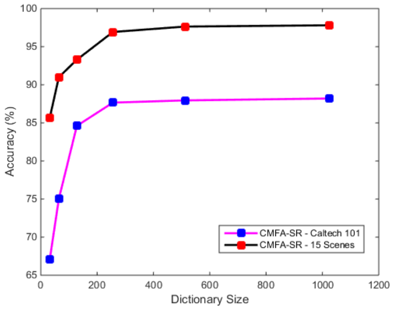

Fig. 3. The performance of the proposed CMFA-SR method for different dictionary sizes on the Caltech 101 data set and the 15 scenes data set.

Table IX shows the comparative evaluation of the proposed method and the deep learning method [62]. Specifically, our proposed method improves upon the performance of the deep learning method by a large margin.

To demonstrate the general importance of our proposed method, we conduct additional experiments on the Painting 91 data set (artist classification task). The input features used are Fisher vector features computed as described in [50]. We then apply the proposed method to extract the CMFA-SR features and the final classification is performed by using the SVM classifier with the RBF kernel. Table X shows that our proposed method achieves the classification accuracy of 65.78%, compared to only 59.04% by the Fisher vector features method.

We further discuss the effects of our proposed method on the initial features and how it encourages better clustering and discrimination among different classes of a data set. To visualize the effect of our proposed method, we use the popular t-SNE visualization technique [75] that produces visualization of high dimensional data in scatter plots. Figure 5 shows the t-SNE visualizations of the initial features used as input and the features extracted after applying the CMFA-SR method for different data sets. It can be seen from figure 5 that the proposed CMFA-SR method helps to reduce the distance among the data points of the same class, which leads to the formation of higher density clusters for these data points. Meanwhile the CMFA-SR method also helps increase the distance among the clusters of different classes resulting in better discrimination among them. Applying two types of discriminatory information, coupled with a discriminative sparse representation model, our proposed CMFA-SR method, which leads to better separation among the data samples from different classes, thus improves recognition performance.

To evaluate the contribution of the individual steps to the overall recognition rate, we conduct experiments on the MIT-

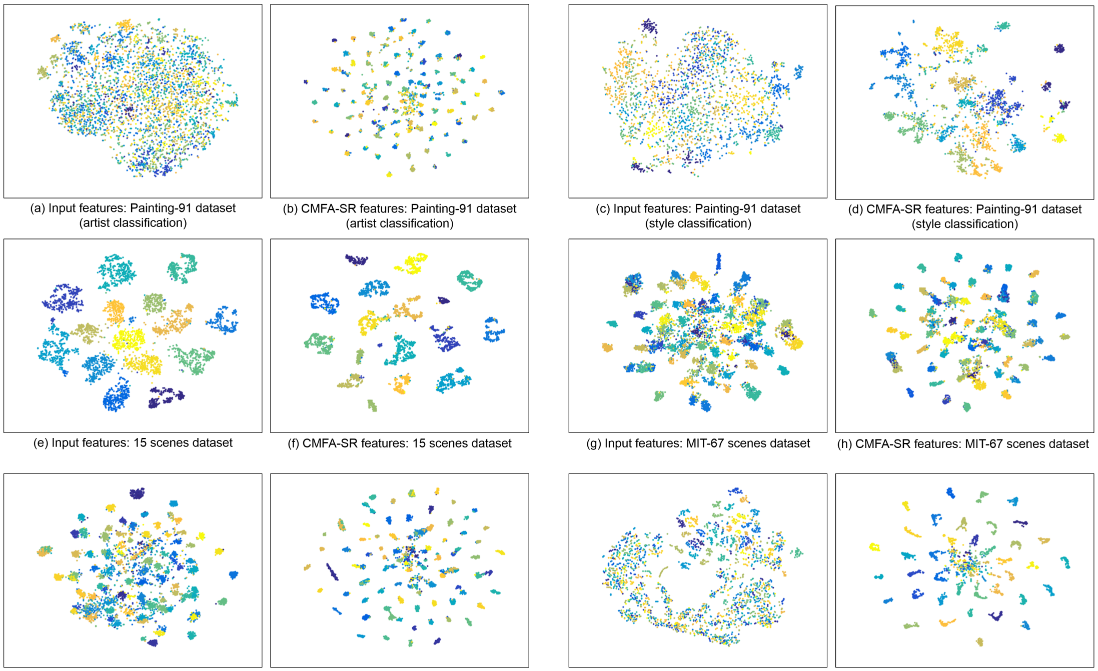
(a) Input features: Painting-91 dataset (artist classification)
(b) CMFA-SR features: Painting-91 dataset (artist classification)
(c) Input features: Painting-91 dataset (style classification)
(d) CMFA-SR features: Painting-91 dataset (style classification)
(e) Input features: 15 scenes dataset
(f) CMFA-SR features: 15 scenes dataset
(g) Input features: MIT-67 scenes dataset
(h) CMFA-SR features: MIT-67 scenes dataset
(i) Input features: Caltech 101 dataset
(j) CMFA-SR features: Caltech 101 dataset
(k) Input features: Extended Yale B dataset
(l) CMFA-SR features: Extended Yale B dataset

Fig. 5. The t-SNE visualization of the initial input features and the features extracted after applying the proposed CMFA-SR method for different data sets.

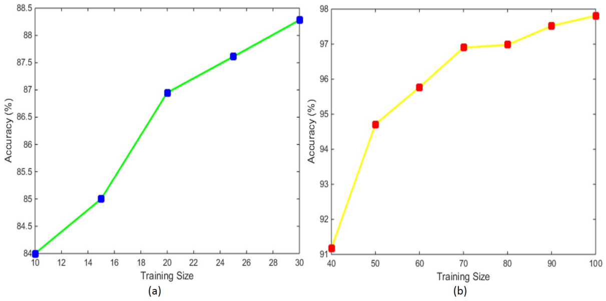

Fig. 4. The performance of the proposed CMFA-SR method when the size of the training data varies on (a) Caltech 101 data set (b) 15 Scenes data set.

Method	Accuracy (%)
Places-CNN [62]	68.24
CMFA-SR features	**81.12**

TABLE IX
COMPARISON OF THE PROPOSED CMFA-SR FEATURES AND THE DEEP LEARNING FEATURES USING THE MIT-67 INDOOR SCENES DATA SET.

67 data set using the input CNN features extracted from the Places-CNN [62] as specified in [62]. Table XI shows the performance evaluation of the individual steps in the proposed CMFA-SR method. Specifically, the CMFA-SR features (both CMFA and dictionary learning) achieves the best classification accuracy of 81.12% since it incorporates both the discriminatory features extracted using the CMFA method and the discriminative dictionary learning.

Method	Accuracy (%)
Fisher Vector features [50]	59.04
CMFA-SR features	**65.78**

TABLE X
COMPARATIVE EVALUATION OF THE PROPOSED CMFA-SR FEATURES AND THE HAND CRAFTED FEATURES USING THE PAINTING-91 DATA SET (ARTIST CLASSIFICATION TASK).

Method	Accuracy (%)
Places-CNN (input features) [62]	68.24
CMFA features only (only subspace learning)	73.96
Dictionary learning features only	76.19
CMFA-SR features	**81.12**

TABLE XI
EVALUATION OF THE CONTRIBUTION OF INDIVIDUAL STEPS IN THE PROPOSED CMFA-SR METHOD USING THE MIT-67 INDOOR SCENES DATA SET.

K. Evaluation of the Dictionary Screening Rule

We evaluate the performance of the proposed CMFA-SR method with and without the dictionary screening rule to understand the effectiveness of the screening rule. In particular, the performance is evaluated by calculating the average training time (s/per image), which is determined by dividing the total train time with the training sample size. The assessment is performed on the Caltech 101 data set with the same settings as provided in the experiments section. Table XII provides the average training time per image of the CMFA-SR method with and without dictionary screening rule for different dictionary sizes of 256, 512 and 1024 on the Caltech 101 data set. It can be observed that the training time significantly reduces as the dictionary size of the CMFA-SR method increases. The training time efficiency is marginal for small dictionary sizes but for the dictionary size 1024, the screening rule improves the average training time per image by almost 33%. Table XIII shows the performance comparison of the proposed CMFA-SR

Method	256	512	1024
CMFA-SR without screening rule	0.45	2.62	5.78
CMFA-SR with screening rule	0.40	2.05	3.84

TABLE XII
EVALUATION OF THE DICTIONARY SCREENING RULE FOR THE PROPOSED CMFA-SR METHOD ON THE CALTECH 101 DATA SET WITH THE DICTIONARY SIZE OF 256, 512, AND 1024.

Method	Accuracy (%)
CMFA-SR without screening rule	88.49
CMFA-SR with screening rule	88.20

TABLE XIII
COMPARATIVE EVALUATION OF THE PROPOSED CMFA-SR METHOD WITH AND WITHOUT THE DICTIONARY SCREENING RULE FOR THE DICTIONARY SIZE 1024 USING THE CALTECH 101 DATA SET.

method, with and without the screening rule for the dictionary size 1024 using the Caltech 101 data set. It can be seen that there is a marginal loss of performance of less than 0.5% for the proposed method with the screening rule but it provides a significant improvement in the average training time by almost 33%.

L. Comparison with the L2 norm regularizer

We compare the proposed method with the L1 (sparsity regularizer) and L2 norm on the Painting-91 data set and the 15 scenes data set, respectively. The same input features are used for the two data sets as described in Section V-A and V-B. The L2 norm based method is optimized using stochastic gradient decent algorithm and the RBF-SVM classifier is used for the final classification. Experimental results in Table XIV show that the L1 norm performs better than the L2 norm by a margin of between 5% and 8%. The L2 norm based method, even though possesses good analytical properties due to its differentiability, does not encourage model compression and removal of irrelevant features, which can be crucial for high-dimensional data. The L1 norm based method implicitly filters out a lot of noise from the model as well as stabilizes the estimates if there is high collinearity between the features resulting in a better generalized model. Another advantage of the L1 norm based method is that it is less sensitive to outliers, and therefore improves the pattern recognition performance.

VI. CONCLUSION

We have presented in this paper a complete marginal Fisher analysis (CMFA) method that extracts the discriminatory features in both the column space of the local samples based within class scatter matrix and the null space of its transformed matrix. We have also presented a discriminative sparse representation model by integrating a representation criterion, such as the sparse representation, and a discriminative criterion, which applies the new within-class and between-class scatter matrices based on the marginal information, for improving the classification capability. We have finally proposed the largest step size for learning the sparse representation to address the convergence issues in optimization, and a dictionary screening rule to purge the dictionary items with null coefficients for improving the computational efficiency. Our experiments on different visual recognition tasks using representative data sets show the feasibility of our proposed method.

data set	Method	Accuracy (%)
Painting-91 Artist Cls. Task	Proposed method with L2 norm **Proposed method with L1 norm**	59.82 **65.78**
Painting-91 Style Cls. Task	Proposed method with L2 norm **Proposed method with L1 norm**	64.32 **73.16**
15 Scenes	Proposed method with L2 norm **Proposed method with L1 norm**	92.26 **98.45**

TABLE XIV
COMPARISON OF THE PROPOSED METHOD WITH L1 AND L2 NORM USING THE PAINTING-91 AND 15 SCENES DATA SET.

APPENDIX
PROOF OF PROPOSITION 3

Proof. We first establish a relation between our proposed method and the traditional sparse representation lasso method. The objective function in equation 11 is identical to the following equation:

$$\min_{\mathbf{s}_i} ||\mathbf{u}_i - \mathbf{D}\mathbf{s}_i||^2 + ||\sqrt{\alpha M_{ii}}\mathbf{s}_i + \sqrt{\frac{\alpha}{4M_{ii}}}\mathbf{g}_i||^2 + \lambda||\mathbf{s}_i||_1 \tag{19}$$

Therefore, the objective function in equation 11 can be rewritten as follows:

$$\min_{\mathbf{s}_i} ||\mathbf{u}_i^* - \mathbf{D}^*\mathbf{s}_i||^2 + \lambda||\mathbf{s}_i||_1 \tag{20}$$

where $\mathbf{u}_i^* = (\mathbf{u}_i^t - \sqrt{\frac{\alpha}{4M_{ii}}}\mathbf{g}_i^t)^t \in \mathbb{R}^{(n+k)\times 1}$ and $\mathbf{D}^* = (\mathbf{D}^t, \sqrt{\alpha M_{ii}}\mathbf{I})^t \in \mathbb{R}^{(n+k)\times k}$. Note that $||\mathbf{d}_j^*||^2 = ||\mathbf{d}_j||^2 + \alpha M_{ii} \leq 1 + \alpha M_{ii}$ and $||\mathbf{u}_i^*||^2 = ||\mathbf{u}_i||^2 + \frac{\alpha}{4M_{ii}}||\mathbf{g}_i||^2$.

According to the projection theorem in [76], we observe that $||\boldsymbol{\theta}_i(\lambda) - \boldsymbol{\theta}_i(\lambda_{\max})||_2 \leq ||\frac{\mathbf{u}_i^*}{\lambda} - \frac{\mathbf{u}_i^*}{\lambda_{\max}}||_2$, where $\boldsymbol{\theta}_i(\lambda)$ and $\boldsymbol{\theta}_i(\lambda_{\max})$ are the solutions of the dual problem associated with the values of λ. The condition given in proposition 3 for identifying dictionary items with zero coefficients is $|\mathbf{u}_i\mathbf{d}_j - \frac{\alpha}{2}\mathbf{g}_i^t\mathbf{I}_j| < (\lambda_{\max} - \sqrt{(||\mathbf{d}_j||^2 + \alpha M_{ii})(||\mathbf{u}_i||^2 + \frac{\alpha}{4M_{ii}}||\mathbf{g}_i||^2)}(\frac{\lambda_{\max}}{\lambda} - 1)$, which is equal to $|(\mathbf{d}_j^*)^t\boldsymbol{\theta}_i(\lambda_{\max})| < 1 - ||\mathbf{u}_i^*||_2||\mathbf{d}_j^*||_2|\frac{1}{\lambda} - \frac{1}{\lambda_{\max}}|$.

Thus, we have the following relations.

$$\begin{aligned}
|\boldsymbol{\theta}_i^t(\lambda)\mathbf{d}_j^*| &= |(\mathbf{d}_j^*)^t\boldsymbol{\theta}_i(\lambda)| \\
&\leq |(\mathbf{d}_j^*)^t\boldsymbol{\theta}_i(\lambda) - (\mathbf{d}_j^*)^t\boldsymbol{\theta}_i(\lambda_{\max})| + |(\mathbf{d}_j^*)^t\boldsymbol{\theta}_i(\lambda_{\max})| \\
&\leq ||(\mathbf{d}_j^*)||_2||\boldsymbol{\theta}_i(\lambda) - \boldsymbol{\theta}_i(\lambda_{\max})||_2 \\
&+ 1 - ||(\mathbf{d}_j^*)||_2||\frac{\mathbf{u}_i^*}{\lambda} - \frac{\mathbf{u}_i^*}{\lambda_{\max}}||_2 \\
&\leq ||(\mathbf{d}_j^*)||_2||\frac{\mathbf{u}_i^*}{\lambda} - \frac{\mathbf{u}_i^*}{\lambda_{\max}}||_2 \\
&+ 1 - ||(\mathbf{d}_j^*)||_2||\frac{\mathbf{u}_i^*}{\lambda} - \frac{\mathbf{u}_i^*}{\lambda_{\max}}||_2 \\
&= 1
\end{aligned} \tag{21}$$

It is shown in [77] that the dual variable $\boldsymbol{\theta}_i$ in the Lagrange dual function of the lasso problem defined in equation 20 satisfies

$$|\boldsymbol{\theta}_i^t\mathbf{d}_j^*| \leq 1 \implies s_{ij} = 0 \tag{22}$$

Hence, the proposition 3 is proved. □

References

[1] C. Luo, B. Ni, S. Yan, and M. Wang, "Image classification by selective regularized subspace learning," *IEEE Transactions on Multimedia*, vol. 18, no. 1, pp. 40–50, Jan 2016.

[2] M. Jian and C. Jung, "Semi-supervised bi-dictionary learning for image classification with smooth representation-based label propagation," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 458–473, March 2016.

[3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[4] J. Yang and C. Liu, "A general discriminant model for color face recognition," in *Proc. ICCV*, 2007, pp. 1–6.

[5] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Proc. IEEE Conf. CVPR*, 2010, pp. 2691–2698.

[6] Q. Feng and Y. Zhou, "Kernel combined sparse representation for disease recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 10, pp. 1956–1968, 2016.

[7] U. L. Altintakan and A. Yazici, "Towards effective image classification using class-specific codebooks and distinctive local features," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 323–332, March 2015.

[8] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2651–2664, 2013.

[9] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *International Journal of Computer Vision*, vol. 109, no. 3, pp. 209–232, 2014.

[10] S. Gao, I. W. H. Tsang, and L. T. Chia, "Laplacian sparse coding, hypergraph laplacian sparse coding, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 92–104, 2013.

[11] J. Feng, B. Ni, Q. Tian, and S. Yan, "Geometric lp-norm feature pooling for image classification," in *CVPR 2011*, June 2011, pp. 2609–2704.

[12] J. Yang, K. Yu, and T. Huang, "Supervised translation-invariant sparse coding," in *Proc. IEEE Conf. CVPR*, 2010, pp. 3517–3524.

[13] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576–1588, 2012.

[14] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[15] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[16] A. M. Martinez and A. C. Kak, "Pca versus lda," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[17] T.-K. Kim and J. Kittler, "Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 318–327, 2005.

[18] S. Yan, D. Xu, B. Zhang, H. j. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, Jan 2007.

[19] C. Liu, "Discriminant analysis and similarity measure," *Pattern Recognition*, vol. 47, no. 1, pp. 359 – 367, 2014.

[20] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, CA, USA: Academic Press Professional, Inc., 1990.

[21] F. Khan, S. Beigpour, J. van de Weijer, and M. Felsberg, "Painting-91: a large scale database for computational painting categorization," *Machine Vision and Applications*, vol. 25, no. 6, pp. 1385–1397, 2014.

[22] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. CVPR*, vol. 2, 2006, pp. 2169–2178.

[23] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *CVPR 2009*, 2009, pp. 413–420.

[24] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *CVPRW*, 2004, pp. 178–178.

[25] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.

[26] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[27] A. Puthenputhussery, Q. Liu, and C. Liu, "Sparse representation based complete kernel marginal fisher analysis framework for computational art painting categorization," in *ECCV*, 2016, pp. 612–627.

[28] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," in *IJCAI*, 2007, pp. 708–713.

[29] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 3, pp. 328–340, 2005.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[33] B. Fulkerson, A. Vedaldi, and S. Soatto, *Localizing Objects with Smart Dictionaries*. Springer, 2008, pp. 179–192.

[34] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1294–1309, 2009.

[35] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Conf. CVPR*, 2008, pp. 1–8.

[36] N. Zhou, Y. Shen, J. Peng, and J. Fan, "Learning inter-related visual dictionary for object recognition," in *Proc. IEEE Conf. CVPR*, 2012, pp. 3490–3497.

[37] R. Sivalingam, D. Boley, V. Morellas, and N. Papanikolopoulos, "Positive definite dictionary learning for region covariances," in *Proc. ICCV*, 2011, pp. 1013–1019.

[38] Q. Liu and C. Liu, "A new locally linear knn method with an improved marginal fisher analysis for image classification," in *Biometrics (IJCB), 2014 IEEE International Joint Conference on*. IEEE, 2014, pp. 1–6.

[39] S. Chen and C. Liu, "Clustering-based discriminant analysis for eye detection," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1629–1638, 2014.

[40] Q. Liu and C. Liu, "A novel locally linear knn method with applications to visual recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. IEEE Early Access Articles, pp. 1–12, 2016.

[41] Q. Liu and C. Liu, "A novel locally linear knn model for visual recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1329–1337.

[42] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new lda-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713 – 1726, 2000.

[43] H. Yu and J. Yang, "A direct lda algorithm for high-dimensional datawith application to face recognition," *Pattern recognition*, vol. 34, no. 10, pp. 2067–2070, 2001.

[44] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring," in *ICASSP*, 2009, pp. 693–696.

[45] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. NIPS*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 801–808.

[46] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[47] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. CVPR*, June 2010, pp. 3360–3367.

[48] W. Deng, J. Hu, and J. Guo, "In defense of sparsity based face recognition," in *Proc. IEEE Conf. CVPR*, 2013, pp. 399–406.

[49] F. Perronnin, J. Snchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010, pp. 143–156.

[50] A. Puthenputhussery, Q. Liu, and C. Liu, "Color multi-fusion fisher vector feature for fine art painting categorization and influence analysis," in *WACV*, March 2016, pp. 1–9.

[51] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, Jul 2002.

[52] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *CIVR*, ser. CIVR '07, 2007, pp. 401–408.

[53] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[54] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[55] Z. Guo, D. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *Image Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1657–1663, June 2010.

[56] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *Image Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1512–1523, July 2009.

[57] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Conf. CVPR*, June 2007, pp. 1–8.

[58] K. van de Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1582–1596, Sept 2010.

[59] K.-C. Peng and T. Chen, "A framework of extracting multi-scale features using multiple convolutional neural networks," in *ICME*, June 2015, pp. 1–6.

[60] K.-C. Peng and T. Chen, "Cross-layer features in convolutional neural networks for generic classification tasks," in *ICIP*, Sept 2015, pp. 3057–3061.

[61] H. Goh, N. Thome, M. Cord, and J. H. Lim, "Learning deep hierarchical visual feature coding," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 12, pp. 2212–2225, 2014.

[62] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. NIPS*, 2014, pp. 487–495.

[63] S. Yang and D. Ramanan, "Multi-scale recognition with dag-cnns," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[64] L. jia Li, H. Su, L. Fei-fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification &amp; semantic feature sparsification," in *Proc. NIPS*, 2010, pp. 1378–1386.

[65] J. Sun and J. Ponce, "Learning discriminative part detectors for image classification and cosegmentation," in *Proc. ICCV*, 2013, pp. 3400–3407.

[66] Q. Li, J. Wu, and Z. Tu, "Harvesting mid-level visual concepts from large-scale internet images," in *Proc. IEEE Conf. CVPR*, 2013, pp. 851–858.

[67] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proc. IEEE Conf. CVPR*, 2013, pp. 923–930.

[68] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "Svm-knn: Discriminative nearest neighbor classification for visual category recognition," in *Proc. IEEE Conf. CVPR*, vol. 2, 2006, pp. 2126–2136.

[69] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.

[70] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[71] F. Perronnin, J. Sanchez, and T. Mensink, *Improving the Fisher Kernel for Large-Scale Image Classification*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 143–156.

[72] L. Bo, X. Ren, and D. Fox, "Multipath sparse coding using hierarchical matching pursuit," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[73] W. Deng, J. Hu, and J. Guo, "Extended src: Undersampled face recognition via intraclass variant dictionary," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1864–1870, 2012.

[74] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. ICCV*, 2011, pp. 543–550.

[75] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[76] J. Wang, J. Zhou, J. Liu, P. Wonka, and J. Ye, "A safe screening rule for sparse logistic regression," in *Proc. NIPS*, 2014, pp. 1053–1061.

[77] Z. J. Xiang, H. Xu, and P. J. Ramadge, "Learning sparse representations of high dimensional data on large scale dictionaries," in *Proc. NIPS*, 2011, pp. 900–908.

Ajit Puthenputhussery is currently pursuing the Ph.D. degree with the Department of Computer Science, New Jersey Institute of Technology, NJ, USA. His current work includes designing robust sparse representation models and developing image descriptors for visual recognition applications. His current research interests include machine learning, pattern recognition, sparse representation, metric learning, content based image classification and kinship verification.

Qingfeng Liu is currently pursuing the Ph.D. degree with the Department of Computer Science, New Jersey Institute of Technology, NJ, USA. His current research interests include computer vision, machine learning, pattern recognition, sparse representation, metric learning, image classification, face recognition, and kinship verification.

Chengjun Liu is currently the Director of the Face Recognition and Video Processing Laboratory with the New Jersey Institute of Technology, NJ, USA. He has developed the class of new methods that includes the evolutionary pursuit method, the enhanced Fisher models, the Gabor Fisher classifier, the Bayesian discriminating features method, the kernel Fisher analysis method, new color models, and similarity measures. His current research interests include pattern recognition, machine learning, computer vision, image and video analysis, and security.