IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 2, 2018, accepted July 15, 2018, date of publication August 1, 2018, date of current version August 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2862159

Collaborative Self-Regression Method With Nonlinear Feature Based on Multi-Task Learning for Image Classification

AO LI[1,2], (Member, IEEE), ZHIQIANG WU[2,3], (Senior Member, IEEE), HUAIYIN LU[4], DEYUN CHEN[1], AND GUANGLU SUN[1], (Member, IEEE)

[1]Postdoctoral Research Station of School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China
[2]Department of Electrical Engineering, Wright State University, Dayton, OH 45435, USA
[3]Department of Electronic and Information Engineering, University of Tibet, Lhasa 850000, China
[4]School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510275, China

Corresponding author: Ao Li (dargonboy@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61501147, in part by the Natural Science Foundation of Heilongjiang Province (CN) under Grant F2015040, in part by the China Postdoctoral Science Foundation under Grant 2016M601438, and in part by the Postdoctoral Science Foundation of Heilongjiang Province (CN) under Grant LBH-Z15099.

ABSTRACT Multi-task learning has received great interest recently in the area of machine learning. It shows a considerable capacity to jointly learn multiple latent relationships hidden among tasks, and has been widely used in data mining and computer vision problems. In this paper, we propose a new multi-task based collaborative linear regression framework to address the image classification problem, which allows the class-specific and collaboratively shared latent structure components to be explored simultaneously. The proposed framework takes multi-target regression of each class as a task to transfer shared structures among them. To be more efficient and adaptive, the class-wise nonlinear subspace is also learned in this framework to earn inter-class discrimination and model adaptability. The proposed framework provides a unified and flexible perceptiveness for jointly learning the nonlinear projected features and regression parameters. Furthermore, a numerical scheme via iterative alternating optimization is also developed to solve the novel objective function in the proposed framework and guarantee the convergence. Extensive experimental results tested on several datasets demonstrated that our proposed framework outperforms existing competitive methods and achieves consistently high performance.

INDEX TERMS Linear regression, multi-task learning, image classification, nonlinear feature, numerical optimization.

I. INTRODUCTION

Image classification is a fundamental issue in mid-level vision tasks, and has attracted tremendous attention in the area of computer vision and pattern recognition during the past few decades [1]–[7]. Based on whether the labels for the training set are unknown or known, the image classification task can be divided into two branches, namely unsupervised classification and supervised classification. Without any label information, unsupervised image classification can be seen as a clustering problem, which is realized by some distance metric or knowledge transferring method in feature space [8]–[10]. For the supervised classification problem, the classifiers were first trained with the help of a training set. A well trained classifier can then predict the label for a query sample [11]–[14]. In this paper, we focus on the supervised image classification problem.

Linear regression is a classical and popular method for solving the supervised image classification problem, which is easy to implement and provides competitive performance with low computational cost [15]–[18]. In general, it trains one classifier for each and every class with a linear model by enforcing the training samples close to their label vectors. However, the classical linear regression model working on original training images often cannot obtain a satisfactory classification result. On the one hand, there exist some representation gaps between images themselves and their labels. On the other hand, it is difficult to present a stable and consistent feature in the original image space.

Hence, some user-specific local feature extraction methods, such as Gabor [19], [20], SIFT [21] and HOG [22], are utilized in the linear regression model to obtain more stable performance and improve robustness. Although these user-specific features help to improve classification performance, there remain some drawbacks such as large computational cost, low discrimination ability, and low adaptation for different datasets. With the development of subspace learning, the deficiencies of the user-specific feature extraction methods have been improved progressively. Subspace learning, which aims to project high-dimensional data to a much lower dimensional feature space through dimensionality reduction, such as principal component analysis (PCA) [23] and linear discriminative analysis (LDA) [24], [25], has been considered a significant step in many practical vision applications since the features for visual tasks are always of very high dimensionality. That is, if we train the classification model on them, it not only needs too much computational source but will also encounter the so-called dimensionality curse, which will lead to inferior performance with so high dimensionality. Fortunately, subspace learning-based dimensionality reduction methods can alleviate the aforementioned problem and make the features of the data to be processed in the subsequent operations more compact. Therefore, it has been successfully applied in many visual tasks and studied extensively in the literature [11], [26], [27]. Among these studies, the linear dimensionality reduction method has been widely used due to its ease in implementation and affordable computational cost.

In recent years, multi-task learning has drawn increasing research efforts and shown remarkable performance in many challenging problems of machine learning and computer vision [28], [29]. In the framework of multi-task learning, multiple relevant tasks work synergistically, since the parameters for these tasks can be propagated and learned simultaneously. It is noted that some shared structural information and latent relationships hidden in these relevant tasks are mined sufficiently to be instrumental in improving the performance of the final results. Moreover, for some weakly supervised and semi-supervised problems, multi-task learning is also an effective strategy to tackle the situations where only a small amount of training samples are available for each task. For many recent applications of multi-task learning, they are generally established for high-level visual tasks on relevant but distinct datasets (*e.g. person re-identification from cross-view cameras* [30]), or different features of the same dataset (*e.g. action recognition with multiple modality attributes* [31]).

Though many improvements have been achieved on the conventional linear regression models, there are some shortcomings existed to be further studied. Firstly, due to the discrete structure of label vectors, the gap between instances their labels will degrade the regression parameters learning. Secondly, to better adapt the relationship of input and output, the projection is generally learned to project the data samples to another more suitable subspace. However, the projection learning is implemented from the holistic view in conventional ones. It means that we learn only one projection for all the samples in dataset, and the samples from different classes are projected into the same subspace. By doing so, the significant discrimination may be lost, which will lead to inferior classification performance. Thirdly, for multi-task learning-based classification, the knowledge are transferred across the relevant datasets or different modalities of same dataset. However, the potential knowledge hidden in different classes is ignored and hardly exploited. Furthermore, in some cases, we cannot obtain the relevant datasets or multiple modalities of dataset. Nevertheless, it is worth noting that relevant information still exists among different classes of the same dataset for a common classification issue, which can be sufficiently exploited when it is combined with an elaborated regression model and subspace projection.

Hence, to overcome the above problems, we take each class as a task and explore appropriate relevant tasks for different classes in the same dataset. With the relevant tasks, we want to establish a collaborative classification framework based on multi-task learning to share the knowledge across different classes. Firstly, to bridge the gap between structure of label and instance, we do not force the data to regress to the label vector, but convert the classification problem into a multi-target regression problem with the projected data itself. Hence, it is called a self-regression framework in our method. Secondly, we learn multiple nonlinear projections for different classes, which will not only earn extra discrimination but also facilitate the adaption of self-regression model.Thirdly, to mine the shared and special structures in different classes, we introduce the multi-task learning framework to our self-regression model. It is noted that regression target is self-projected instance in our proposed model. That is, the regression target includes more image structures but not discrete binary values as label, which could be synthesized with class-special structures and basic structures. Nevertheless, the basic structure could be shared and transferred among different classes. To this end, we take the self-regression of each class as a task and expect to transfers the shared knowledge among them. By jointly learning the shared collaborative regression parameters, class-specific regression parameters, and class-wise nonlinear subspace projections, our proposed framework can obtain more promising results.

Based on the above considerations, in this paper we propose a nonlinear projected collaborative self-regression framework based on multi-task learning to address the image classification problem. Unlike conventional regression methods, we do not force the data to regress to the label vector, but convert the classification problem into a multi-target regression problem with the projected data itself. Hence, it is called a self-regression framework in our method. Furthermore, in the proposed framework, shared collaborative regression parameters, class-specific regression parameters, and class-wise nonlinear subspace projections are jointly learned iteratively. Such a joint optimization plays an important role in learning two kinds of latent regression parameters

and transferring useful knowledge among different classes. To further model the intrinsically intertwined relationships between the input and output, we also induce a nonlinear operation to the learned subspace projection, which can help to reveal the potential nonlinear correlation.

The contributions of this work can be summarized as the following four aspects:

- By converting conventional label regression into multi-target regression, we develop a multi-task based unified self-regression framework, which is capable of not only transferring the latent knowledge among classes to construct the shared regression parameters, but also learning the class-specific regression parameters and discriminative subspace projection to dimensionality reduction.
- We introduce nonlinear operations to the projected subspace, which can better describe the potential intertwined relationships between the input and output.
- We develop an iterative numerical scheme based on alternating optimization, which enables solving the novel jointly learning objective function in the proposed framework efficiently.
- We also provide some analysis of the architecture of our framework from the perspective of neural networks and demonstrate its novelty.

The rest of the paper is organized as follows. In section II, we briefly review some related works. Section III presents the details of our proposed framework and the developed numerical scheme. Extensive experiments are presented in section VI. Finally, we conclude the paper in section V.

II. RELATED WORKS

A. SUPERVISED IMAGE CLASSIFICATION

To our best knowledge, we follows the following three main branches of supervised image classification in recent literatures: *a)* constructing a sparse representation-based classification model; *b)* learning the deep-level and effective feature descriptor for training samples, and then training the classifiers with them; *c)* training the classifier and feature extractor jointly in a unified framework, which facilitates their mutual benefit. Next, we will briefly describe some previous related works in these branches.

Sparse representation-based image classification is a popular scheme, which is focused on how to establish effective sparsity model and code the query sample with few relevant labeled samples. For instance, John Wright et al proposed a general sparse representation-based classification (SRC) framework to address image-based object recognition problem [32]. In their method, without the dictionary learning, the overcompleted dictionary is comprised of the training samples themselves. With the l_1-norm minimization, the discriminative nature of sparse representation is exploited to perform classification by representing the test sample as a combination of those training samples in the same class. Experiments on face recognition demonstrate that their proposed method shows an advantage on both classification accuracy and robustness to occlusion. Zhang *et al.* [33], presented a collaborative representation based classification method with l_2-norm, which pointed out that the powerful performance of SRC is due to the collaborative representation among training samples but not sparsity based on l_1-norm. Considering the co-sparse model, Shekhar *et al.* [17] investigated a classification method based on analysis coding models. In their work, an analysis operator is learned to extract the sparsity components of training samples as features. Then, an SVM-based classifier is trained on these features. Nevertheless, without embedding the discriminative information into the operator learning, the performance is not optimal. To reveal the intrinsic mechanism of the SRC-based classification, Cai *et al* [34] proposed an approach based on probabilistic collaborative representation to address the pattern recognition problem, in which the probability that a test sample belongs to the subspace of all classes is analyzed and computed. Moreover, a special collaborative classifier is developed which can maximize the likelihood that a test sample belongs to certain class. More extensive SRC-based methods for pattern recognition can be studied in the literature [35]–[38].

Feature description is a key technique in high-level visual tasks. It has been successfully used in many image classification problems and widely studied in recent works. For local features, SIFT and HOG are two popular descriptors in representing images due to their powerful ability to capture the distinguishable local details. Due to the high computational complexity, the local features are rarely used to train the classifier directly but integrated into a global image representation. To this end, Lazebnik *et al.* [39] extended the conventional Bag of Features (BoF) image representation by dividing the image into many sub-regions with multiple scales and integrating their feature descriptors in a so-called Spatial Pyramid Matching (SPM) way. Wang *et al.* [40] proposed a linear locality coding method to take place of the conventional VQ coding in SPM, which could project the descriptor into a more effective local-coordinate system and better preserve its discrimination. To further reduce the unavoidable information loss in the coding process, in [41], a distance coding scheme was studied, where the local features were first transformed into more discriminative distance vectors, and then the distance vectors were further encoded into sparse codes to capture the salient features. Qi *et al.* [42] exploited the Pairwise Transform Invariant (PTI) principle and investigated a novel co-occurrence binary pattern feature. Moreover, the proposed feature can be flexibly extended to incorporate multi-scales and showed some advantage in robustness to geometric and photometric variations.

Though some competitive results have been obtained by these methods, there is still some room for improvement. It is noted that feature extraction and classifier training are two separate phases in the feature description-based classification method, because it only pays attention to feature discrimination and effectiveness but ignores the interaction between feature subspace learning and classifier training.

Compared to the above two categories, the jointly learning feature and classifier can optimize each other in the same framework iteratively. Jiang *et al.* [43] suggested that label information can be used to guide dictionary learning, with which a more discriminative sparsity feature can be extracted. Based on this consideration, they proposed to learn the discriminative dictionary and train the classifiers in a unified framework simultaneously. Inspired by the label propagation based semi-supervised learning, a non-negative sparse graph structure is learned in [44]. Subsequently, the label prediction and projection learning are integrated into a linear regression, where the regression and graph learning are simultaneously performed to guarantee an overall optimum. Yuan and Tang [45] assumed that there was a shared space for the original data space, and that the linear prediction should be contributed from both the original and shared space. As a consequence, they proposed a spectral-spatial shared linear regression and obtained impressive performance on hyperspectral image classification. More recently, Zhen *et al.* [46] explored the inter-target correlations via a robust low-rank learning, and established a general framework to jointly model the nonlinear feature extraction with kernel trick and regression matrices learning. The achieved high performance showed the effectiveness of their proposed method for multivariate regression problems.

B. MULTI-TASK LEARNING

Recently, multi-task learning (MTL) is attracting incremental interest in computer vision [47], [48]. The main advantage of MTL is that some shared information can be propagated across multiple tasks, which can be utilized to improve the performance of many machine learning problems. Meanwhile, a number of existing promising techniques, such as dictionary learning and sparse coding, can also be flexibly integrated with MTL [49]. By exploring task-specific incoherence and low rank structure, Su *et al.* [30] proposed an MTL-based method with low-rank attribute embedding, which has been successfully applied to person re-identification. Hu *et al.* [31] found that features from different channels shared some similar hidden structure and proposed a heterogeneous feature learning method incorporating MTL for activity recognition. Jing et al proposed a novel framework to leverage the different types of visual features and predefined attributes, which can well construct their connections by transferring the shared knowledge among multiple external source [50]. In [51], a novel dictionary learning method with MTL is proposed to address the person re-identification problem, in which the joint semantic and latent attributes are modeled to realize the challenging unsupervised domain adaptation.

Motivated by the above mentioned methods, we establish a MTL-based framework inspired by the fact that some shared structures exist in different classes of the same dataset, which is generally ignored in previous works. Meanwhile, to characterize the potential relationship between the input and output, nonlinear subspace projection learning is also introduced into our proposed framework to obtain further improvement. It is worth noting that aforementioned conventional MTL-based label regression framework for cross-dataset or multi-modal features of the same dataset cannot directly take different classes as related tasks due to only positive instances existing in each task. As a consequence, to integrate with MTL, we convert image classification into a multi-target regression problem, where the regression parameters for each class are matrices rather than vectors, as in conventional methods.

III. DETAILS OF OUR PROPOSED FRAMEWORK

In this section, we present details of our proposed method. A novel objective function is presented to implement MTL-based collaborative self-regression with nonlinear projection. Next, a numerical scheme is developed to obtain a reliable solution with guaranteed convergence. Moreover, a discussion will be presented to show the architecture of our framework from the perspective of neural networks.

A. NOVEL OBJECTIVE FUNCTION

For clarity, we first list some notations used in our paper. The training data set is denoted as $\{X_i\}_i^C$, where $X_i \in R^{d\times N_i}$ represents a training subset of ith class with N_i instances, and d denotes the instance dimensionality. For each training instance $x_k \in R^d$, let $y_k \in R^C$ be its label vector, where C denotes the total number of classes in the training set. Furthermore, if x_k is chosen from the cth class($c = 1, 2, 3 \ldots C$), only the cth entry of y_k is one and all others are zero. For example if x_k is chosen from the second class, its label y_k is denoted as $y_k = [0, 1, 0, \ldots]^T$. Let $W \in R^{C\times d}$ be the regression parameters as $W = \left[w_1^T, w_2^T, \ldots w_c^T\right]^T$, where each w_c denotes the regression vector for the cth class and can be seen as the classifier for the cth class in the regression-based classification problem. With these notations, the general training model for W with linear regression can be formulated as follows:

$$\min_W \|Y - WX\|_F^2 + \lambda\phi(W) \tag{1}$$

where $Y = [y_1, y_2, \ldots y_k, \ldots]$ denotes the label matrix for all instances ($X = [x_1, x_2, \ldots x_k, \ldots]$) in training data set, $\phi(W)$ presents certain constraint on the parameters, $\|\cdot\|_F$ denotes the Frobinus norm and λ is the positive scalar to balance the two terms.

By enforcing the training data close to their labels, Eq.(1) aims to learn the regression parameters from a set of instances statistically. Then, the label vector y_t of a test instance x_t can be predicted with the parameters as $\hat{y}_t = \hat{W}x_t$, and the label $\hat{c}_t$ can be derived with $\hat{c}_t = \underset{c}{argmin}\left(\hat{y}(c)\right)$.

However, it is very difficult to learn the effective regression parameters in the original data space and there is enormous computational cost due to the high dimensionality of instances. Hence, subspace learning is usually incorporated into the regression model in the following form:

$$\min_{W,P} \|Y - WPX\|_F^2 + \gamma\varphi(P) + \lambda\phi(W) \tag{2}$$

where $P \in R^{p} \times d$ denotes the subspace projection matrix, and satisfies $p \ll d$ to implement the dimensionality reduction by projecting the original data to a low dimensional space. $\varphi(P)$ denotes the constraint on P to control its projection structure, γ and λ are two positive scalars to balance the three terms. By optimizing W and P alternatively in the same objective function, they benefit each other to obtain an overall optimal solution, which will help to gain better classification accuracy with suitable learned projection subspace. For more clarity, Eq.(2) is equivalent to the following form:

$$\min_{\{w_c\}_{c=1}^{C}, P} \sum_{i=1}^{C} \sum_{k=1}^{N_i} \left(y_k(i) - w_i^T P x_k \right) + \gamma \varphi(P) + \lambda \phi(W) \quad (3)$$

where $y_k(i)$ is a scalar and denotes the i th entry of the k th instance. From the equivalent Eq.(3), we can see that, in fact, with the training instances from all the classes, Eq.(2) trained each regression vector by constraining its relevant positive instances to 1 and negative ones to 0.

In light of the superiority of model in Eq.(2), our goal is to introduce nonlinear projection and extend it with a multi-task learning framework. Nevertheless, the form in Eq.(2) cannot be extended and incorporated into our proposed MTL-based framework directly. This is because, in our proposed framework, we aim to explore the shared structure among the different classes in the same dataset. It follows that the data from each class (X_i) are taken as a task. In this way, there exist no negative instances in each task and the labels for the instances of each task are all 1. That is, the mentioned label prediction formulation $\hat{c}_t = \underset{i}{argmin} \left(\hat{y}_t(i) \right)$ is no longer effective due to the fact that the classifier w_i can not regress a smaller value for instances from other classes, because it is trained without negative instances in each task.

Hence, to realize our idea, we convert the conventional label-based regression into multi-target self-regression (MTSR), which can be extended and incorporated into our proposed MTL-based framework. To the best of our knowledge, this is the first attempt to explore the shared structure of different classes in the same dataset with MTL. The MTSR with MTL is formulated as follows:

$$\min_{\{W_i, P_i\}_{i=1}^{C}, W_S} \sum_{i=1}^{C} \Big(\| \mathcal{A}(X_i) - (W_S + W_i) P_i X_i \|_F^2 + \varphi(P_i) + \lambda \| W_i \|_F^2 \Big) + \lambda_S \| W_S \|_F^2 \quad (4)$$

where $\mathcal{A}(X_i) \in R^{d_A \times N_i}$ ($d_A < d$) is a designed self-projected operator, which is used to transform the data matrix X_i to a low-dimensional self-regression multi-target matrix with respect to its columns. $W_S \in R^{d_A \times p}$ and $W_i \in R^{d_A \times p}$ denote the regression parameters to explore the shared regression component among the tasks and the class-specific regression component, respectively. P_i represents the subspace projection for the ith class. With the objective function in Eq.(4), regression parameters and class-specific subspace projections are learned simultaneously in a unified framework. Also, for a classification problem including C classes, the first term in the proposed objective function is a similar task for each class. By summarizing them, a MTL framework based on shared parameters is established by transferring the learned knowledge with W_S among the different classes. Unlike conventional label-based regression, in our proposed framework, the shared and class-specific regression parameters for each class are no longer vectors but matrices, which are used to enforce the instances close to their self-regression multi-target components.

Furthermore, inspired by the success of nonlinear operations in the neural network, we also introduce a nonlinear operation into our proposed framework to describe the potential nonlinear relationship between the self-regression multi-target component and the original data. We reformulate the objective function in Eq.(4) to the following nonlinear projection form:

$$\min_{\{W_i, P_i\}_{i=1}^{C}, W_S} \sum_{i=1}^{C} \Big(\| \mathcal{A}(X_i) - (W_S + W_i) \Theta(P_i X_i) \|_F^2 + \varphi(P_i) + \lambda \| W_i \|_F^2 \Big) + \lambda_S \| W_S \|_F^2 \quad (5)$$

where $\Theta(\cdot)$ denotes the nonlinear function. In our method, motivated by [52], two kinds of element-wise nonlinear function are employed as follows:

$$\Theta_1(t) = \frac{t}{1 + e^{-t}}, \quad or \ \Theta_2(t) = sign(t) \, Shrink_\tau(t) \quad (6)$$

where t is the scalar variable, $sign(t)$ denotes the sign function and $Shrink_\tau(t) = \max(|t| - \tau, 0)$.With the achievement of nonlinear projection in neural network, we hope to the introduced nonlinear operation can earn extra improvement by considering the potential nonlinear relation between input and output. Moreover, in our proposed method, samples are enforced to regress to its self-projection counterpart but not the conventional label vector. Hence, the nonlinear feature should be capable of preserving more structure of original sample. The above nonlinear functions can be seen as the extended version of sigmoid function and ReLU function, which has been applied in the neural network successfully. In fact, the functions in Eq.(6) implement the nonlinear shrinkage operation on original feature, which is believed to preserve more significant structures existed in the original ones.

To enhance the discrimination of our proposed framework, the subspace projection is also learned individually. Hence, the regularization term $\varphi(P_i)$ plays a key role in learning the suitable discriminative subspace projection. As a direct consequence, in our framework, the form of $\varphi(P_i)$ is considered as follows:

$$\varphi(P_i) = \gamma \| P_i \|_F^2 + \theta \| P_i \bar{X}_i \|_F^2 \quad (7)$$

where $\bar{X}_i$ denotes an instance matrix, which includes all the instances of the training set except the instances in the ith class. γ and θ are two positive scalars to balance the two terms. The first term in Eq.(7) is used to control the power of P_i, while the second one is used to alleviate the correlation

between P_i and instances of other classes, which can make P_i more discriminative.

B. DEVELOPED NUMERICAL SCHEME

In this section, we will develop an iterative numerical scheme for solving the novel objective function in the proposed framework, which can monotonically decrease the objective function with a guaranteed convergence. It is noted that the minimization problem in Eq.(5) is difficult to solve for all variables simultaneously due to the non-convexity of the objective function. Hence, we utilize the alternative optimization for efficiently updating them by solving the objective function with respect to one variable while keeping others fixed. The details are shown as follows. For separating the nonlinear operation, we reformulate the objective function as the following constraint equation:

$$\begin{aligned}\min_{\{W_i,P_i,\Phi_i\},W_S} \quad & \sum_{i=1}^{C}\Big(\|\mathcal{A}(X_i)-(W_S+W_i)\,\Phi_i\|_F^2 \\ & +\varphi(P_i)+\lambda\|W_i\|_F^2\Big)+\lambda_S\|W_S\|_F^2 \\ s.t.\ \ & \Phi_i=\Theta(P_iX_i), \quad i=1,2,\ldots,C \end{aligned} \tag{8}$$

Then, by fixing $\{W_i, P_i, \Phi_i\}_{i=1}^{C}$, we minimize the following function J_{W_S} with respect to W_S.

$$\min_{W_S}\sum_{i=1}^{C}\|\mathcal{A}(X_i)-(W_S+W_i)\,\Phi_i\|_F^2+\lambda_S\|W_S\|_F^2 \tag{9}$$

From Eq.(9), it can be seen that J_{W_S} is a convex and differentiable function. By forcing its derivative to be zero, the closed-form solution can be obtained as

$$\frac{\partial J_{WS}}{\partial W_S}=0\Rightarrow$$
$$W_S^*=\left(\sum_{i=1}^{C}(\mathcal{A}(X_i)-W_i\Phi_i)\,\Phi_i^T\right)\left(\sum_{i=1}^{C}\Phi_i\Phi_i^T+\lambda_S I\right)^{-1} \tag{10}$$

where $(\cdot)^T$ and $(\cdot)^{-1}$ denote the transpose and inverse matrix operations. I represents the identity matrix.

Fixing $\{P_i, \Phi_i\}_{i=1}^{C}$ and W_S, the class-specific regression parameters W_i can be optimized by the minimization function J_{W_i} in each task as follows:

$$\min_{W_i}\|\mathcal{A}(X_i)-(W_S+W_i)\,\Phi_i\|_F^2+\lambda_i\|W_i\|_F^2 \tag{11}$$

Similarly, forcing its derivative to be zero, we can obtain the closed-form solution for each W_i as

$$\frac{\partial J_{Wi}}{\partial W_i}=0\Rightarrow$$
$$W_i^*=\left(\sum_{i=1}^{C}(\mathcal{A}(X_i)-W_S\Phi_i)\,\Phi_i^T\right)\left(\Phi_i\Phi_i^T+\lambda_i I\right)^{-1} \tag{12}$$

When the regression matrices W_S and $\{W_i\}_{i=1}^{C}$ are fixed, Eq.(8) can be rewritten as the following individual optimization for each pair of P_i and Φ_i.

$$\begin{aligned}\min_{P_i,\Phi_i}\ & \|\mathcal{A}(X_i)-(W_S+W_i)\,\Phi_i\|_F^2+\varphi(P_i) \\ s.t.\ & \Phi_i=\Theta(P_iX_i)\end{aligned} \tag{13}$$

To further decouple P_i and Φ_i, we introduce an extra auxiliary variable Q_i, and convert it into the following equivalent problem:

$$\begin{aligned}\min_{P_i,\Phi_i}\ & \|\mathcal{A}(X_i)-(W_S+W_i)\,\Phi_i\|_F^2+\varphi(P_i) \\ s.t.\ & \Phi_i=\Theta(Q_i), \quad Q_i=P_iX_i\end{aligned} \tag{14}$$

Next, the unconstrained augmented Lagrangian function of Eq.(14) can be given as

$$\begin{aligned}\min_{P_i,\Phi_i,Q_i,U_1,U_2}\ & \|\mathcal{A}(X_i)-(W_S+W_i)\,\Phi_i\|_F^2+\varphi(P_i) \\ & +\frac{\rho}{2}\|\Phi_i-\Theta(Q_i)\|_F^2+\frac{\rho}{2}\|Q_i-P_iX_i\|_F^2 \\ & +\langle U_1,\Phi_i-\Theta(Q_i)\rangle+\langle U_2,Q_i-P_iX_i\rangle\end{aligned} \tag{15}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product matrix operation and ρ is a positive scalar to penalize the approximation between the two variables. U_1 and U_2 are two Lagrangian multiplier matrices. With some algebra, the minimization problem in Eq.(15) can be solved with the iterative steps as follows.

First, by keeping others fixed, we update Φ_i within the tth iteration as follows:

$$\begin{aligned}\Phi_i^{(t+1)}=\arg\min_{\Phi_i}\ & \|\mathcal{A}(X_i)-(W_S+W_i)\,\Phi_i\|_F^2 \\ & +\frac{\rho}{2}\|\Phi_i-\Theta\left(Q_i^{(t)}\right)+\frac{1}{\rho}U_1^{(t)}\|_F^2\end{aligned} \tag{16}$$

With the above quadratic form, the closed-form solution of the updated Φ_i can be computed by forcing its derivative to be zero.

Second, by substituting Eq.(7), P_i can be updated as

$$\begin{aligned}P_i^{(t+1)}=\arg\min_{P_i}\ & \frac{\rho}{2}\|Q_i^{(t)}-P_iX_i+\frac{1}{\rho}U_2^{(t)}\|_F^2 \\ & +\gamma\|P_i\|_F^2+\theta\|P_i\bar{X}_i\|_F^2\end{aligned} \tag{17}$$

Similarly, with the differentiable function in Eq.(17), the solution can be easily obtained.

Finally, the term Q_i is solved when P_i and Φ_i are fixed. The minimization defined in Eq.(15) can be rewritten as

$$\begin{aligned}Q_i^{(t+1)}=\arg\min_{Q_i}\ & \frac{\rho}{2}\|\Phi_i-\Theta(Q_i)+\frac{1}{\rho}U_1^{(t)}\|_F^2 \\ & +\frac{\rho}{2}\|Q_i-P_iX_i+\frac{1}{\rho}U_2^{(t)}\|_F^2\end{aligned} \tag{18}$$

Due to the separable form of nonlinear function defined in Eq.(6), the updated Q_i can be computed in an element-wise manner.

The overall scheme for our proposed framework is presented in **Alogrithm 1**.

Algorithm 1 Iterative Scheme for Solving the Proposed Framework

Input: Training data set $\{X_i\}_{i=1}^{C}$,self-projected operator $\mathcal{A}$ regularization parameters $\gamma, \theta, \lambda, \lambda_S$ ρ initial random $W_S^{(0)}$,$\{W_i^{(0)}, P_i^{(0)}, \Phi_i^{(0)}\}_{i=1}^{C}$ $U_1^{(0)} = U_2^{(0)} = 0$, iteration number T, initial $k = 0$.

Repeat

1. Calculate matrix $W_S^{(k+1)}$ with Eq.(10);
2. Calculate matrix $W_i^{(k+1)}$ for each task with Eq.(12);
3. **for** $t = 0$ to $T - 1$ **do**, $(i = 1, 2, \ldots, C)$
   Update $\Phi_i^{(t+1)}$ in subproblem in Eq.(16);
   Update $P_i^{(t+1)}$ in subproblem in Eq.(17);
   Update $Q_i^{(t+1)}$ in subproblem in Eq.(18);
   Update Lagrangian multipliers:
   $$U_1^{(t+1)} = U_1^{(t)} + \rho\left(\Phi_i^{(t+1)} - \Theta\left(Q_i^{(t+1)}\right)\right)$$
   $$U_2^{(t+1)} = U_2^{(t)} + \rho\left(Q_i^{(t+1)} - P_i^{(t+1)} X_i\right)$$
   end for
4. For each i, $\Phi_i^{(k+1)} \leftarrow \Phi^T$, $P_i^{(k+1)} \leftarrow P^T$, $Q_i^{(k+1)} \leftarrow Q^T$
5. $k \leftarrow k + 1$;

Until convergence

Output: W_S^*,$\{W_i^*, P_i^*\}_{i=1}^{C}$

C. CLASSIFICATION PRINCIPLE AND ANALYSIS OF THE PROPOSED FRAMEWORK

The details of our framework have been described in previous subsections. In this subsection, we will show the classification principle with respect to our method, and analyze the proposed framework from the perspective of network architecture. In our method, the combined regression parameters with W_S and W_i are trained to enforce an instance within the ith class close to its multi-target counterpart by self-projected operator. It follows that, with the learned parameters W_S^* and $\{W_i^*, P_i^*\}_{i=1}^{C}$, a query sample x_t can be classified into a category by the self-regression loss term in our framework as follows:

$$\hat{l}_t = \arg\min_i \|\mathcal{A}(x_t) - (W_S + W_i)\,\Theta\,(P_i x_t)\|_F^2 \qquad (19)$$

where $\hat{l}_t$ denotes the estimated label indicator.

Taking the loss term in Eq.(19) as the classification principle, the classification results are greatly influenced by the residual error between multi-target counterpart $\mathcal{A}(x_t)$ and the regression component. It is worth noting that the residual error may increase when the regression model is only with the class-specific regression parameters W_i^* due to the limited training instances and structure information embedded in them. That is, the shared component generated by W_S^* will help the class-specific component to force a smaller error, because some latent shared structures exist in the different classes. By W_S^*, these shared structures can be transferred among the classes to facilitate the classification performance.

Considering the success of neural networks in many visual applications, next we will present some analysis of the architecture of our framework from the network perspective. The basic unit of each task to construct the network (framework) in our method is presented in Figure 1. The global architecture of the proposed MTL-based framework is presented in Figure 2.

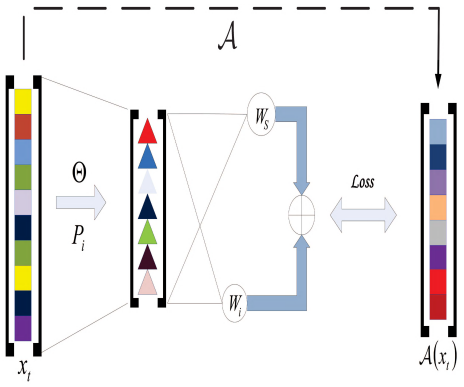


FIGURE 1. The unit structure of our proposed framework for each instance in *i*th task.

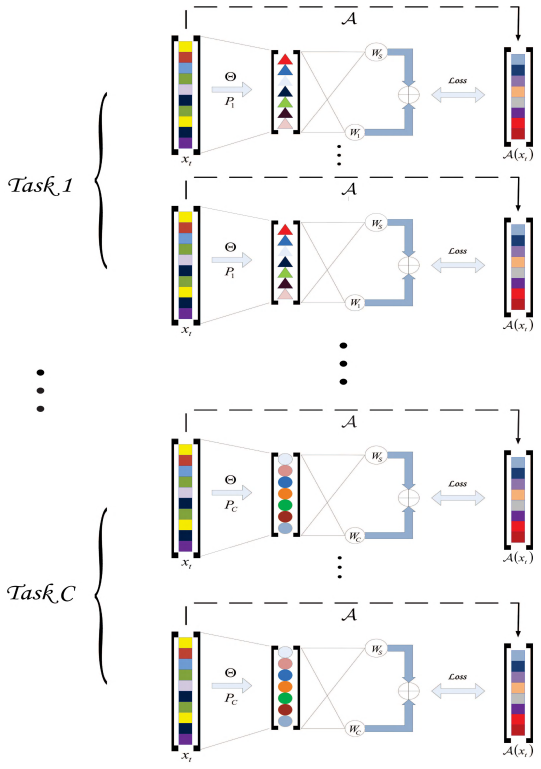


FIGURE 2. Global architecture of our proposed framework from the perspective of network.

From Figure 2, we can observe that, in fact, our proposed framework can be seen as a network architecture with MTL, which includes input layer, hidden layer and output layer. The introduced nonlinear projection is used as a neural cell of the network to model the potential intrinsic relationships between inputs and outputs. Furthermore, on the one hand,

the parameter W_S is learned across all the tasks to transfer the shared knowledge among them. On the other hand, W_i is learned in each task to construct the class-specific components. Next, the combination of the shared and class-specific components will allow better adaption in the loss function, which is believed to help to improve the performance and obtain more promising classification results.

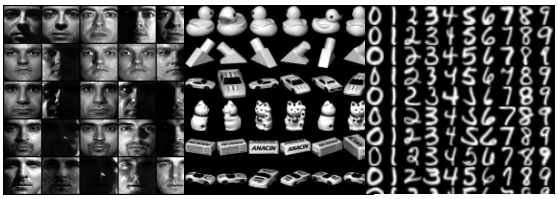
IV. EXPERIMENTAL RESULTS

In this section, we will conduct extensive experiments on four public datasets and one handcrafted dataset constructed ourselves. The proposed framework will be compared to a number of existing representative algorithms to evaluate the performance. Furthermore, the convergence of the developed numerical scheme will be verified on two test datasets and, finally, the advantages of our framework are presented and analyzed.

A. DATASETS DESCRIPTION AND SETTING

1) FACE DATASET

Extended YaleB face dataset includes 2414 frontal images of 38 persons, each person having approximately 64 images taken under different lighting conditions. Some sample images are shown in Figure 3(a). The images used in our experiment are cropped to 32×32, and the dimensionalities of nonlinear projection and self-projected operator are reduced to 300 and 500, respectively. 32 images of each person are randomly selected as the training instance while the rest are used as test instances.



(a) (b) (c)

FIGURE 3. Some instances from Extended YaleB, COIL20 and USPS. (a) 20 selected face images from Extended YaleB and 5 images for each person in a line. (b) 36 selected images from COIL20 and 6 images for each object in a line. (c) 90 selected images from USPS and 10 images for each digital.

2) OBJECT DATASET

The COIL20 contains 1440 images from 20 objects, and each object has 72 images obtained from continuous angles at an interval of five degrees. Some sample images are shown in Figure 3(b). In our experiment, all of the images in dataset are cropped and resized to 32×32. The dimensionalities of nonlinear projection and the self-projected operator are also reduced to be 300 and 500, respectively. 36 images of each object are used for training and the rest are used for testing.

3) HANDWRITTEN DATASET

The USPS dataset has a total of 9298 handwritten digit images, and the size of each is 16×16. Some sample images are shown in Figure 3(c). Due to the simple structure of data USPS, 30 images of each class are used as the training instances and the remaining ones are used for testing. Due to the not very high dimensionality of the original images, we take the images themselves as the self-regression counterpart, which means that an identity self-projected operator is used for USPS. The dimensionality of subspace projection is reduced to 100.

4) SCENE DATASET

The fifteen scene dataset was built gradually by [11], [39], and [53], and has 15 scene categories, including bedroom, coast, forest, highway, industrial, inside city, kitchen, living room, mountain, office, open country, store, street, suburb, and tall building. Some sample images are shown in Figure4. For each category, there are 210 to 410 images.Following the protocol in [39], the three levels of SPM with SIFT feature for each image is adopted as the input in our experiment. The dimensionalities of nonlinear projection and the self-projected operator are also reduced to be 300 and 500 respectively. 100 images per category are used for constructing the training set and the rest are used for constructing the test set. Compactly, we take scene-15 as fifteen scene categories dataset in experiments.

FIGURE 4. Some instances from Scene-15. we presets only 6 scene categories and 8 selected sample images for each category.

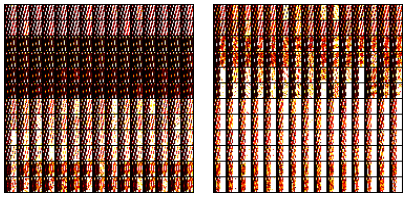
5) WSU HIGH-ORDER SPECTRA (WSU-HOS) DATASET

We collect a new hand-crafted dataset to evaluate the compared algorithms.We name it the ''WSU High-order Spectra'' dataset and it can be used for evaluating performance for modulation recognition problems. Some sample images are shown in Figure 5. This dataset contains 600 high-order spectra images of communication signals with 6 different kinds of modulation in Gaussian and Rayleigh channels respectively. Within each channel, there are 100 images per class, which are generated by the method in [54] and [55]. For each class, the communication signals are modulated in the same way but with differing channel noise power, widely known as additive white Gaussian noise (AWGN). More specifically, in each class, there are 20 images of communication signals for signal

TABLE 1. The comparison classification results(%) on different datasets.

	Extended YaleB	COIL20	USPS	WSU-HOS-G	WSU-HOS-R
SRC	95.42±0.47	95.08±0.33	87.80±0.79	92.87±1.56	79.85±2.79
CRC	97.23±0.35	96.47±0.98	89.11±0.60	96.42±0.66	83.62±0.93
ProCRC	93.25±1.28	96.33±1.00	88.65±0.82	95.87±1.93	88.89±0.62
NNGR	92.32±1.43	97.00±2.63	87.18±2.59	95.47±1.40	82.86±1.33
Ours with Θ_1	96.98±0.28	**98.94±0.57**	**92.59±0.60**	98.63±1.29	**90.63±1.23**
Ours with Θ_2	**97.85±0.41**	98.89±0.44	92.06±0.91	**98.79±0.23**	89.80±0.88

to noise ratios of 0dB, 5dB, 10dB, 15dB and 20dB respectively. The size of the images in WSU-HOS is 16×16. The self-regression counterpart are also the instances themselves due to their not very high dimensionalities. The dimensionality of nonlinear projection is reduced to 50. For Gaussian channel, 5 images of each class are used for training and the rest are used for testing. We refer to it as WSU-HOS-G. For Rayleigh channel, 10 images of each class are used for training and the remaining are used for testing. We refer to it as WSU-HOS-R.



(a) (b)

FIGURE 5. Some instances from WSU-HOS. (a) 180 selected images within Gaussian channel and 30 images of each modulation in every two lines. (b) 180 selected images within Rayleigh channel and 30 images of each modulation in every two lines.

B. EVALUATION RESULTS

In our experiments, the shared regression parameters $W_S^{(0)}$ and class-specific parameters $\{W_i^{(0)}\}_{i=1}^{C}$ are initialized by random Gaussian matrices. The regularization parameters are tuned experimentally and set empirically for all of the experiments as $\gamma = 10^{-3}, \theta = 10^{-6}, \rho = 0.1, \lambda = \lambda_S = 0.02$. The total iteration steps for inner and outer loop in **Algorithm 1** are both set to 20. The self-projected operator in **Algorithm 1** is simply implemented by PCA. We compared our proposed approach with the existing representative methods including SRC, collaborative representation-based classification (CRC) in [32], probabilistic collaborative representation classification (ProCRC) in [34], nonnegative graph regression(NGR) in [44], SPM+SVM in [39], linear locality coding (LLC) in [40], LLC with distance coding (LLCDC) and LLC with both of distance coding and SIFT (LLCDCSIFT) in [41]. Our methods with the two kinds of nonlinear functions in Eq.(6) are named as ours with Θ_1 and Θ_2 respectively. All the experiments are run 5 times and the average classification results with standard deviation are reported in Table 1 and Table 2. Due to the scene diversity of Scene-15, for clarity, we also present its confusion matrix of classification results in Figure 6.

TABLE 2. The comparison classification results(%) on Scene-15.

	Scene-15
SRC	79.70±0.71
CRC	79.29±0.30
ProCRC	80.01±0.46
SPM	79.95±0.27
LLC	79.81±0.35
LLCDC	80.30±0.62
LLCDCSIFT	82.40±0.35
Ours with Θ_1	82.23±0.49
Ours with Θ_2	81.51±0.54
Ours-LSIFT with Θ_1	**83.15±0.37**
Ours-LSIFT with Θ_2	82.61±0.32

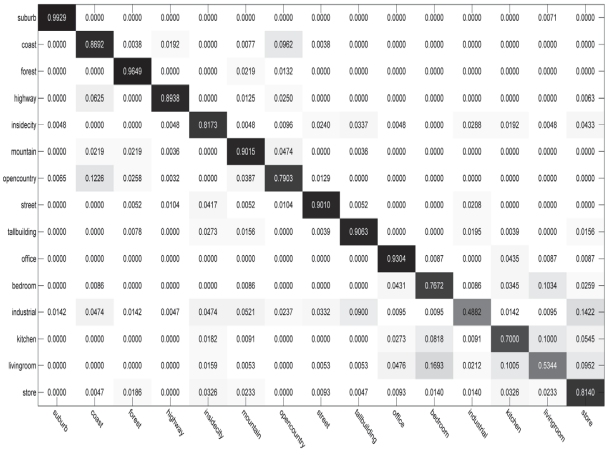


FIGURE 6. The confusion matrix of classification results of Scene-15.

From Table 1, it can be observed that our proposed approach significantly outperforms the compared methods and consistently achieves the best classification accuracy on all of the datasets. In the compared methods, the collaborative correlation of instances is exploited intuitively. Nevertheless, the deep-level potential relationships in the more suitable subspaces are not employed sufficiently. Due to the joint learning of the regression parameters and nonlinear subspace projection, our model better explores the inter-class intrinsic difference statistically and achieves greater discrimination than the others. For scene classification, we take the local SIFT with SPM in [34] as the input features of comparison methods. Comparing the results for scene-15 in Table 2, we can observe that our approach can still achieve competitive performance on the scene classification task. The classification accuracy of our approach is only slightly lower than that of the LLCDCSIFT. Nevertheless, the LLCDCSIFT are

told specially in [36] that it introduce the extra LLC with local SIFT into LLCDC and concatenated them as the final feature. That is, more features are used to facilitate discrimination. So, to be comparable, we also concatenated the extra LCC with local SIFT to our proposed method. Its classification results are presented as ours-LSFT in Table.2. The superior classification results on scene-15 also demonstrate that some latent shared structure information indeed exists among different scene categories. With the MTL framework in our approach, it is effectively learned from multiple tasks and transferred among classes to enable the performance on the scene classification problem.

TABLE 3. The comparison classification results(%) on Extended YaleB with a few training instances.

	Extended YaleB (10)	Extended YaleB (16)
SCLRR	85.44±1.75	91.06±2.21
SCRLRR	85.90±2.92	90.38±2.55
SLNNLRS	89.00±2.96	92.71±1.75
RSCNNLRR	88.00±0.73	93.67±0.79
Ours with Θ_1	**89.46±0.88**	**94.11±0.75**
Ours with Θ_2	88.90±0.57	93.51±0.64

TABLE 4. The comparison classification results(%) on COIL20 with a few training instances.

	COIL20 (6)	COIL20 (8)	COIL20 (10)
SCLRR	82.89±2.97	85.95±1.78	88.87±2.06
SCRLRR	83.93±3.10	87.76±2.24	90.54±2.75
SLNNLRS	84.50±2.81	88.86±2.53	89.43±2.99
RSCNNLRR	86.78±3.22	89.09±1.72	90.80±1.62
Ours with Θ_1	**87.67±1.67**	89.92±1.33	91.10±1.26
Ours with Θ_2	87.26±0.83	**90.70±1.34**	**93.18±1.93**

To further evaluate the performance, we also compare our method to some state-of-the-art semi-supervised classification methods with only a few training instances. The experiments are done on Extended YaleB and COIL20 respectively. The classification results are shown in Table 3 and Table 4. The numbers in the parenthesis denotes the number of training instances. The state-of-the-art methods include, subspace clustering with latent low-rank representation (SCLLR) [11], subspace clustering with robust low-rank representation (SCRLRR) [56], semi-supervised learning based on non-negative low rank and sparse graph (SLNNLRS) [57], and Robust subspace clustering via non-negative low rank representation (RSCNNLRR) [58]. From the results of Table 3 and Table 4, it can be observed that our algorithm consistently achieves the best performance on classification tasks with only a few training instances. This is due to the fact that our MLT-based framework outperforms the comparison methods in exploring the latent relationships among instances and is more suitable to solving scenarios with limited available training data by transferring knowledge among tasks.

Next, to evaluate the speed, we take the experiments on Extended YaleB as examples and show the running time of comparison methods in Table 5. The experiments are performed on Matlab 8.3.0 on a computer with Intel(R) Core(TM) i5-5200 CPU(2×2.20GHz), 8GB memory and Windows 10 operating system. By the results, we can see that the speed of CRC and ProCRC is more quickly than other methods, because the iteration is not needed in these two methods. NNGR obtained the worst result due to the large computation of low rank representation. Though the quadratic minimization is modeled for each parameter, our proposed method need some more time than most of comparison methods. It is because that many individual parameters, such as projection and regression matrix, need learning for each class respectively, which will cost extra computation.

TABLE 5. The running time of comparison methods on Extended YaleB.

	Time(s)
SRC	21.03
CRC	2.70
ProCRC	1.27
NNGR	360
Ours with Θ_1	37.12
Ours with Θ_2	31.26

FIGURE 7. Visualization results on test instances of Extended YaleB. Left shows the original face images. Middle shows the shared components extracted by W_S. Right shows the class-specific components extracted by W_i.

C. DISCUSSION ON ADVANTAGE OF MTL

To illustrate the advantage of introducing MTL in our proposed framework intuitively, we show some visualization results on the Extended YaleB dataset in Figure 7. For more clarity, we take the identity operator as $\mathcal{A}$ in this experiment. That is, the self-regression component is the face image itself. From Figure 7, we can observe that the shared structure components extracted with W_S presents some special information, which can be learned from all of the tasks, such as imperceptible illumination variation, global facial sketch and fine structures existing in different persons. In fact, these structures are tremendously helpful for class-specific components to further decrease regression residual error. Meanwhile, the decreased residual error can significantly facilitate higher classification performance due to the loss function-based classification principle (Eq.19) adopted in our framework. Nevertheless, the shared structures cannot be learned sufficiently from only a certain class itself due to the limited instances in each class. Hence, thanks to the MTL framework, the latent shared structures are collaboratively learned and transferred to mutual benefit among classes.

To evaluate the individual effectiveness of shared and class-specific parameters respectively, we test the classification performance when the proposed model is with only

TABLE 6. The classification accuracy(%) of proposed method with only one kind of regression parameter.

	With only W_S	With only W_i
Extended YaleB	90.65	95.83
COIL20	94.36	96.53
USPS	90.15	91.80
WSU-HOS-G	98.03	98.11
WSU-HOS-R	85.31	89.70

one of them. The accuracies are presented in Table 6. From Table 6, it can be observed that the performance with both of the regression parameters are better than with only anyone of them. It is due to the fact that collaborative shared structures and class-specific structures can be employed simultaneously to synthesis the self-projected component. Furthermore, the results also shows that the accuracy with W_i is a little higher than that with W_S. In our opinion, it is because that the regression object is the self-projected counterpart of sample but not a binary label vector here. So, it will be effectiveness to train respective regression matrix for each class to be adaptive to the complex self-projected objective component.

D. CONVERGENCE AND PARAMETERS

In this section, some analysis will be discussed on the parameters and convergence of our proposed model respectively. For testing the effectiveness of parameters, we still conduct experiments on Extended YaleB with nonlinear function Θ_1 for analyzing. In our proposed model, we introduce the λ and λ_S to constrain the regression parameters matrices, which can control the generality ability of the model. In the experiments, these two parameters are both selected from {0, 0.02, 0.2, 2}. The classification accuracy with pair of varying values are presented in Table7. It can be observed that performance is not obvious sensitive to λ and λ_S, and robustness with most of parameter couples.

TABLE 7. The classification accuracy(%) varying on Extended YaleB with regularization parameters λ and λ_S.

λ \ λ_S	0	0.02	0.2	2
0	94.49	94.74	94.91	94.32
0.02	94.41	96.23	95.17	94.49
0.2	94.07	95.49	94.07	94.24
2	94.46	94.74	93.07	93.79

The parameters γ and θ are used to trade-off between the generality and discrimination of projection operator P. In our test experiments, we found that the classification performance is inferior when both of the two parameters increased. It may be due to that if the parameters for projection are comparable to λ and λ_S, it will degrade the regression parameter learning, which is more significant to the final classification result. The relative stable performance can be obtained with the varying range of parameters in Table 8.

TABLE 8. The classification accuracy(%) varying on Extended YaleB with regularization parameters γ and θ.

γ \ θ	10^{-4}	10^{-5}	10^{-6}	10^{-7}
10^{-2}	94.66	95.83	96.91	96.81
10^{-3}	95.16	95.74	96.95	96.95
10^{-4}	93.74	94.24	95.87	96.21

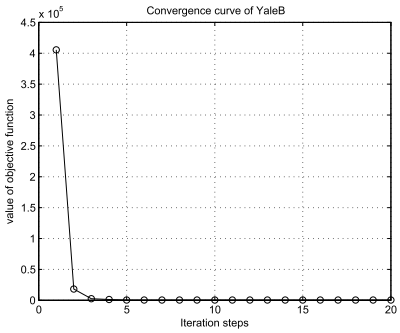


FIGURE 8. Convergence results of the objective function on Extended YaleB.

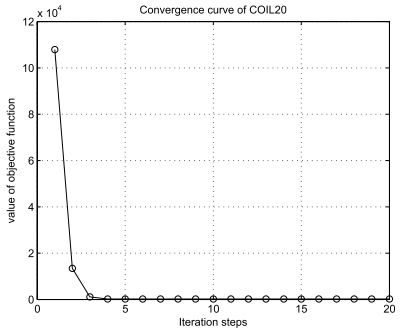


FIGURE 9. Convergence results of the objective function on COIL20.

The effective convergence is significant for our method in practical usage. So, to investigate the convergence, we examine the objective function value during the iteration on Extended YaleB and COIL20 dataset, respectively. The two convergence curves with respect to the outer loop steps are depicted in Figure 8 and Figure 9. It can be observed that the two objective functions both decrease monotonically with respect to the iterative steps. In fact, though we only show the objective function on two representative datasets, all the evaluation datasets in our experiments reach convergence within 20 iterations. The quickly decreasing curves shown in these two figures guarantee the great efficiency in convergence of our developed iterative numerical scheme.

V. CONCLUSION

In this paper, we have proposed an MTL-based nonlinear collaborative regression method for image classification problems. The novel framework in the proposed method enables jointly learning the shared regression parameters, class-specific regression parameters and class-specific subspace projection. To take the instances of different classes

as related tasks, in our method, the image classification is converted into a multi-target regression problem, and the knowledge from different tasks can be transferred to better learn the latent shared structures among classes. Meanwhile, to explore the potential nonlinear relationship between input and output, the nonlinear operation is introduced into our framework to obtain extra performance facilitation. The objective function in the novel framework can be efficiently solved by a developed iterative numerical scheme with guaranteed convergence. The competitive classification results on test datasets validates the outstanding performance of our approach.

REFERENCES

[1] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 524–531.

[2] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1794–1801.

[3] Y. Su, H. Wang, P. Jing, and C. Xu, "A spatial-temporal iterative tensor decomposition technique for action and gesture recognition," *Multimedia Tools Appl.*, vol. 76, no. 8, pp. 10635–10652, 2017.

[4] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2010, pp. 141–154.

[5] J. Zhang, Z. Li, P. Jing, Y. Liu, and Y. Su, "Tensor-driven low-rank discriminant analysis for image set classification," in *Multimedia Tools and Applications*. New York, NY, USA: Springer, 2017, pp. 1–20.

[6] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.

[7] J. Zhang, X. Li, P. Jing, J. Liu, and Y. Su, "Low-rank regularized heterogeneous tensor decomposition for subspace clustering," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 333–337, Mar. 2017.

[8] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.

[9] R. C. De Amorim and B. Mirkin, "Minkowski metric, feature weighting and anomalous cluster initializing in K-means clustering," *Pattern Recognit.*, vol. 45, no. 3, pp. 1061–1075, Mar. 2012.

[10] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1–10.

[11] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1615–1622.

[12] J. Yuan, M. Yang, and Y. Wu, "Mining discriminative co-occurrence patterns for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 2777–2784.

[13] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1465–1472.

[14] P. Jing, Y. Su, C. Xu, and L. Zhang, "HyperSSR: A hypergraph based semi-supervised ranking method for visual search reranking," *Neurocomputing*, vol. 274, pp. 50–57, Jan. 2018.

[15] X. Chai, S. Shan, X. Chen, and W. Gao, "Locally linear regression for pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 7, pp. 1716–1725, Jul. 2007.

[16] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106–2112, Nov. 2010.

[17] S. Shekhar, V. M. Patel, and R. Chellappa, "Analysis sparse coding models for image-based classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 5207–5211.

[18] P. Jing, Y. Su, L. Nie, X. Bai, J. Liu, and M. Wang, "Low-rank multi-view embedding learning for micro-video popularity prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 8, pp. 1519–1532, Aug. 2017.

[19] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer. A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[20] J.-K. Kamarainen, "Gabor features in image analysis," in *Proc. 3rd Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Oct. 2012, pp. 13–14.

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

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[23] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal Component Analysis*. New York, NY, USA: Springer, 2002, pp. 150–166, ch. 7.

[24] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data—With application to face recognition," *Pattern Recognit.*, vol. 34, no. 10, pp. 2067–2070, 2001.

[25] Y. A. Ghassabeh, F. Rudzicz, and H. A. Moghaddam, "Fast incremental LDA feature extraction," *Pattern Recognit.*, vol. 48, no. 6, pp. 1999–2012, 2015.

[26] X. Fang, Y. Xu, X. Li, Z. Fan, H. Liu, and Y. Chen, "Locality and similarity preserving embedding for feature selection," *Neurocomputing*, vol. 128, pp. 304–315, Mar. 2014.

[27] F. Nie, S. Xiang, Y. Song, and C. Zhang, "Orthogonal locality minimizing globality maximizing projections for feature extraction," *Opt. Eng.*, vol. 48, no. 1, p. 017202, 2009.

[28] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 895–903.

[29] J. Chen, J. Liu, and J. Ye, "Learning incoherent sparse and low-rank patterns from multiple tasks," *ACM Trans. Knowl. Discovery Data*, vol. 5, no. 4, p. 22, Feb. 2012.

[30] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3739–3747.

[31] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5344–5352.

[32] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[33] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 471–478.

[34] S. Cai, L. Zhang, W. Zuo, and X. Feng, "A probabilistic collaborative representation based approach for pattern classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2950–2959.

[35] Y. Chi and F. Porikli, "Connecting the dots in multi-class classification: From nearest subspace to collaborative representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3602–3609.

[36] W. Deng, J. Hu, and J. Guo, "Extended SRC: Undersampled face recognition via intraclass variant dictionary," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1864–1870, Sep. 2012.

[37] Z. Wen, B. Hou, and L. Jiao, "Discriminative nonlinear analysis operator learning: When cosparse model meets image classification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3449–3462, Jul. 2017.

[38] X. Jiang and J. Lai, "Sparse and dense hybrid representation via dictionary decomposition for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1067–1079, May 2015.

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

[40] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3360–3367.

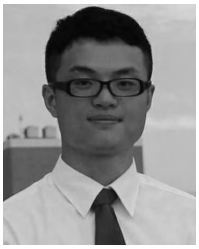
[41] Z. Wang, J. Feng, S. Yan, and H. Xi, "Linear distance coding for image classification," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 537–548, Feb. 2013.

[42] X. Qi, R. Xiao, C.-C. Li, Y. Qiao, J. Guo, and X. Tang, "Pairwise rotation invariant co-occurrence local binary pattern," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2199–2213, Nov. 2014.

[43] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013.
[44] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, "Learning a nonnegative sparse graph for linear regression," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2760–2771, Sep. 2015.
[45] H. Yuan and Y. Y. Tang, "Spectral–spatial shared linear regression for hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 934–945, Apr. 2017.
[46] X. Zhen, M. Yu, X. He, and S. Li, "Multi-target regression via robust low-rank learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 497–504, Feb. 2017.
[47] X.-T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4349–4360, Oct. 2012.
[48] L. Chen, Q. Zhang, and B. Li, "Predicting multiple attributes via relative multi-task learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1027–1034.
[49] P. Ruvolo and E. Eaton, "Online multi-task learning via sparse dictionary optimization," in *Proc. AAAI*, 2014, pp. 2062–2068.
[50] P. Jing, Y. Su, L. Nie, and H. Gu, "Predicting image memorability through adaptive transfer learning from external sources," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1050–1062, May 2017.
[51] P. Peng, Y. Tian, T. Xiang, Y. Wang, M. Pontil, and T. Huang, "Joint semantic and latent attribute modelling for cross-class transfer learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1625–1638, Jul. 2017.
[52] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Netw.*, Jan. 2018, doi: 10.1016/j.neunet.2017.12.012.
[53] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[54] A. Napolitano and I. Perna, "Cyclic spectral analysis of the GPS signal," *Digit. Signal Process.*, vol. 33, pp. 13–33, Oct. 2014.
[55] T. Fusco, L. Izzo, A. Napolitano, and M. Tanda, "On the second-order cyclostationarity properties of long-code DS-SS signals," *IEEE Trans. Commun.*, vol. 54, no. 10, pp. 1741–1746, Oct. 2006.
[56] H. Zhang, Z. Lin, C. Zhang, and J. Gao, "Robust latent low rank representation for subspace clustering," *Neurocomputing*, vol. 145, pp. 369–373, Dec. 2014.
[57] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu, "Non-negative low rank and sparse graph for semi-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2328–2335.
[58] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, "Robust semi-supervised subspace clustering via non-negative low-rank representation," *IEEE Trans. Cybern.*, vol. 46, no. 8, pp. 1828–1838, Aug. 2016.

AO LI (M'17) received the B.S. and Ph.D. degrees from Harbin Engineering University, Harbin, China, in 2009 and 2014, respectively. He was a Research Assistant with Wright State University from 2017 to 2018. He is currently with the Harbin University of Science and Technology, Harbin. His current research interests include sparse representation, pattern recognition, machine learning, and their applications to computer vision.

ZHIQIANG WU (M'02–SM'16) received the B.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 1993, the M.S. degree from Peking University, Beijing, in 1996, and the Ph.D. degree from Colorado State University, Fort Collins, CO, USA, in 2002, all in electrical engineering. From 2003 to 2005, he was an Assistant Professor with the Department of Electrical Engineering, West Virginia University Institute of Technology. In 2005, he joined the Department of Electrical Engineering, Wright State University, Dayton, OH, USA, where he is currently a Full Professor. He has also held visiting positions at Peking University, Harbin Engineering University, Guizhou Normal University, and Tibet University. His research has been supported by NSF, AFRL, AFOSR, NASA, NRL, and OFRN.



HUAIYIN LU received the B.S. degree (Hons.) in electronic information science and technology from Sun Yat-sen University, Guangzhou, China, in 2015, where he is currently pursuing the Ph.D. degree in information and communication engineering. His research interests include cognitive radio network, chaotic communication, visible light communication, and 5G technology.

DEYUN CHEN received the Ph.D. degree from the Harbin University of Science and Technology, Harbin, China, in 2006. He is currently the Dean with the School of Computer Science and Technology and a Full Professor with the Harbin University of Science and Technology. He has authored over 100 scientific papers. His current research interests include electrical capacitance tomography, image processing, and pattern recognition.

GUANGLU SUN (M'07) received the B.S., M.S., and Ph.D. degree from the Harbin Institute of Technology. He was a Visiting Scholar with Northwestern University, USA, from 2014 to 2015. He is currently a Professor and the Director with the Center of Information Security and Intelligent Technology, Harbin University of Science and Technology, Harbin, China. His current research interests include computer networks and security, machine learning, and intelligent information processing.