Marginalized Denoising Dictionary Learning With Locality Constraint

Shuyang Wang, Zhengming Ding, *Student Member, IEEE*, and Yun Fu, *Senior Member, IEEE*

***Abstract*— Learning good representation for images is always a hot topic in machine learning and pattern recognition fields. Among the numerous algorithms, dictionary learning is a well-known strategy for effective feature extraction. Recently, more discriminative sub-dictionaries have been built by Fisher discriminative dictionary learning with specific class labels. Different types of constraints, such as sparsity, low rankness, and locality, are also exploited to make use of global and local information. On the other hand, as the basic building block of deep structure, the auto-encoder has demonstrated its promising performance in extracting new feature representation. To this end, we develop a unified feature learning framework by incorporating the marginalized denoising auto-encoder into a locality-constrained dictionary learning scheme, named marginalized denoising dictionary learning. Overall, we deploy low-rank constraint on each sub-dictionary and locality constraint instead of sparsity on coefficients, in order to learn a more concise and pure feature spaces meanwhile inheriting the discrimination from sub-dictionary learning. Finally, we evaluate our algorithm on several face and object data sets. Experimental results have demonstrated the effectiveness and efficiency of our proposed algorithm by comparing with several state-of-the-art methods.**

***Index Terms*— Marginalized denoising auto-encoder, locality constraint, dictionary learning.**

I. Introduction

LEARNING discriminative representations is always a challenging problem in image classification, especially when images are in the wild, e.g., various illuminations, partial occlusion and low resolutions. Therefore, it is essential to seek better representations to handle those challenges. Dictionary learning and deep learning are two appealing techniques to learn new effective representations [2], [42]. Specifically, dictionary learning aims at building a better basis on which a discriminative coefficient can be achieved with different constraints. Whilst deep learning is designed with deep structure to capture more information from the images [10]. The key for image classification is to learn more discriminative features by uncovering the global within-class structure and filtering out noises.

Sparse dictionary learning has experienced a rapid growth in both theory and application from recent researches, which has led to interesting results in image classification [11], [29], [40], [41], speech denoising [13], and bio-informatics [27] etc. For each input signal, the key idea is to find a linear combination using atoms from a given over-complete dictionary as a new representation. Therefore, sparse representation is capable to reveal the underlying structure of high dimensional data. However, sparse representation ignores the relationships of samples during feature learning.

Recently, low-rank dictionary learning [22] aims to uncover the global structure by grouping similar samples into one cluster, which has been successfully applied to many applications, e.g. object detection [32], multi-view learning [8], unsupervised subspace segmentation [21], and 3D visual recovery [44]. Moreover, the low-rank dictionary well addresses the noisy data by adding an error term with different norms, e.g., l_1-norm, $l_{2,1}$-norm. Furthermore, supervised information has been well utilized to seek a more discriminative dictionary [42], [45]. In this paper, we also adopt the supervised approach to learn multiple sub-dictionaries so that samples from the same class are drawn from one low-dimensional subspace.

Most recently, deep learning has attracted lots of interest in better feature extraction. Among them, auto-encoder [2], [37] is one of the most popular building blocks to form a deep learning framework. The auto-encoder has drawn increasing attention in feature learning area and has been considered as a simulation of the way that human visual system processes imagery. The auto-encoder architecture explicitly involves an encoder module and a decoder one. The encoder outputs a group of hidden representation units, which is realized by a linear deterministic mapping with a weight matrix and a nonlinear transformation employs logistic sigmoid. The decoder reconstructs the input data based on the responded sparse hidden representation. The aforementioned dictionary learning model can be formalized as a decoder module.

Moreover, there is a well-known trick of the trade to deal with noisy data, that is manually injecting noise into the training samples thereby learning with artificially corrupted data. Denoising auto-encoders (DAEs) [35], learned with artificial corrupted data as input, have been successfully applied to a wide range of machine learning tasks by learning a new

Manuscript received April 11, 2017; revised August 20, 2017 and October 1, 2017; accepted October 4, 2017. Date of publication October 20, 2017; date of current version November 9, 2017. This work was supported in part by NSF IIS Award under Grant 1651902, in part by ONR Young Investigator Award under Grant N00014-14-1-0484, and in part by U.S. Army Research Office Award under Grant W911NF-17-1-0367. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xudong Jiang. *(Corresponding author: Shuyang Wang.)*

S. Wang and Z. Ding are with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: shuyangwang@ece.neu.edu; allanding@ece.neu.edu).

Y. Fu is with the Department of Electrical and Computer Engineering and the College of Computer and Information Science, Northeastern University, Boston, MA 02115 USA (e-mail: yunfu@ece.neu.edu).

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

Digital Object Identifier 10.1109/TIP.2017.2764622

1057-7149 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

denoising representation. During the training process, DAEs reconstruct the input data from partial corruption with a pre-specified corrupting distribution to its original clean version. This process learns robust representation which ensures the toleration to certain distortions in input data. The marginalized denoising auto-encoder [6] is a linear version of DAEs, which works efficiently and achieves comparable results with DAEs.

In this work, we develop a feature learning model by unifying the marginalized denoising auto-encoder and locality-constrained dictionary learning (MDDL) together to benefit from both merits. Specifically, dictionary learning manages to tackle with the corrupted data from the sample space, while marginalized auto-encoder attempts to address the noisy data from feature space. Thus, we aim to fight off the corrupted data both from sample space and feature space by integrating dictionary learning and auto-encoder into a unified framework. The main contributions of our paper are listed as follows:

- We desire to seek a transformation matrix to filter out the noise inside the data with a marginalized denoising auto-encoder, which avoids forward and backward propagation, thus works both efficiently and effectively.
- Secondly, with the transformed data, we aim to build a set of supervised sub-dictionaries with locality constraint. In this way, the sub-dictionaries are discriminative for each class, which makes the new representation preserve the manifold structure while uncovering the global structure.
- The marginalized denoising transformation and locality-constrained dictionary are jointly learned in a unified framework. In this way, our model can integrate auto-encoders and dictionary learning to produce features with denoising ability and discriminative information.

II. Related Works

In this section, we mainly discuss two lines of related works, one is dictionary learning and the other is auto-encoder.

A. Dictionary Learning

Recent researches on dictionary learning have demonstrated that a well-learned dictionary will greatly boost the performance by yielding better representation in human action recognition [7], scene categorization [31], image coloration [20] and transfer learning [9]. In order to learn a compact dictionary with more representation power, several algorithms and regularizations have been introduced into the dictionary learning framework. In FDDL, a set of class-specified sub-dictionary whose atoms correspond to the class labels is updated iteratively based on the Fisher discrimination criterion to include discriminative information. Jiang *et al.* [15] presented a Label Consistent K-SVD (LC-KSVD) algorithm to make a learned dictionary more discriminative for sparse coding. These methods shown that a structured dictionary could dramatically improve the classification performance. However, the performance of these methods will drop a lot if the training data is largely corrupted.

However, sparse representation based methods consider each sample as independent sparse linear combination, this assumption fails to exploit the spatial consistency between neighbor samples. Recent research efforts have yielded more promising results on the task of classification by using the idea of locality [36]. The method named Local Coordinate Coding (LCC), which specifically encourages the coding to rely on local structure, has been presented as a modification to sparse coding. In [36] the author also theoretically proved that locality is more essential than sparsity under certain assumptions.

Most recently, low-rank constraint has been applied in many areas to deal with noisy data. Liu *et al.* [22] proposed that we can convert the task of face image clustering into a subspace segmentation problem with the assumption that face images from different individuals lie in different near independent subspaces. By applying low-rank regularization into dictionary updating, the DLRD [23] algorithm achieved impressive results especially when corruption exists. Jiang and Lai proposed a sparse- and dense- hybrid representation based on a supervised low-rank dictionary decomposition to learn a class-specific dictionary and null out non-class-specific information [14].

Inspired by the above learning techniques, our conference paper proposed a Locality-Constrained Low-Rank Dictionary Learning (LC-LRD) to enhance the identification capability by using the geometric structure information [38], which will be detailed introduced in Section III since it has close connection with this paper.

B. Auto-Encoder

As a typical single hidden layer neural network with identical input and target, auto-encoder [4] aims to discover data's intrinsic structure by encouraging the output to be as similar to the target as possible. Essentially, the neurons in the hidden layer can be seen as a good representation since they are able to reconstruct the input data. To encourage structured feature learning, further constraints has been imposed on parameters during training. Denoising auto-encoders (DAEs) is proposed to enforce the hidden layer be capable to discover more robust features meanwhile prevent it from simply learning the identity [2], [34]. The DAEs is trained to have the ability to reconstruct the input signal from its corrupted version which artificially added with noise.

On this basis, stacked denoising autoencoders (SDAs) [35] have been successfully used to learn new representations and attained record accuracy on standard benchmark for domain adaptation. However, there are two crucial limitations of SDAs, 1) high computational cost, and 2) lack of scalability to high-dimensional features. To address these two problems, [6] proposed marginalized SDAs (mSDA). Different with SDAs, mSDA marginalizes noise and thus the parameters can be computed in closed-form rather than using stochastic gradient descent or other optimization algorithms. Consequently, mSDA significantly speeds up SDAs by two orders of magnitude.

In our conference paper [38], different from previous locality linear coding works [28], [31], we also imposed the locality on auto-encoder method to extract feature with

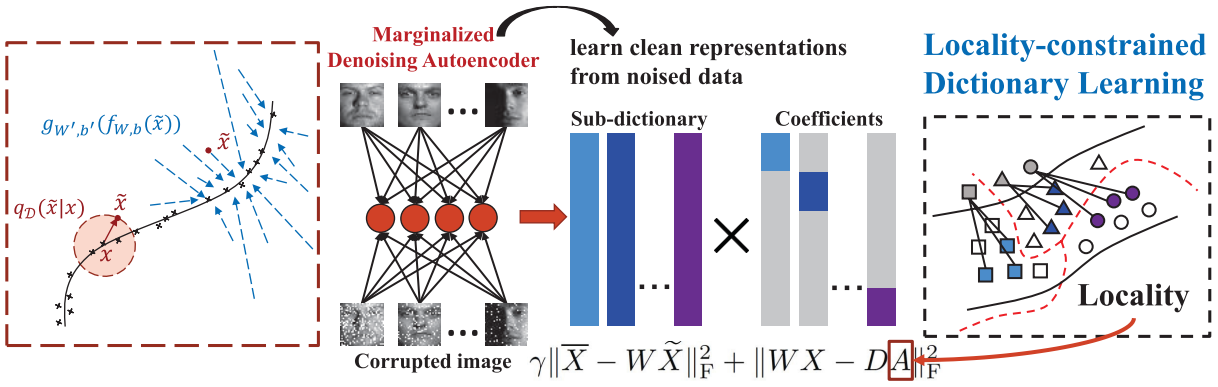

Fig. 1. Illustration of our methods. The marginalized denoising auto-encoder is adopted in dictionary learning (DL) schemes. The weights in auto-encoder and sub-dictionaries in DL are trained jointly. Each sub-dictionary is learned of low-rank, which can narrow the negative effect of noise contained in training samples. For marginalized denoising auto-encoder, the input is manually added with noise.

local information for enhancing the classification ability. The proposed locality-constrained on auto-encoder (LCAE) can be trained with the backpropagation algorithm.

Our previous paper [38] explored the effectiveness of locality constraint on two different types of feature learning techniques, i.e., dictionary learning and auto-encoder, respectively. In our journal extension, we jointly learn auto-encoder and dictionary to benefit from both techniques. Fig. 1 illustrates our framework. To make our model fast, we adopt a lite version of auto-encoder, i.e., marginalized denoising auto-encoder [6], which has shown appealing performance and efficiency. Furthermore, we use several benchmarks to evaluate our proposed algorithm, and the experimental results show its better performance comparing state-of-the-arts.

III. Marginalized Denoising Dictionary Learning With Locality Constraint

In this section, we first revisit our locality-constrained dictionary learning and marginalized denoising auto-encoder. Then we propose our novel marginalized denoising dictionary learning with locality constraint. Finally, we develop an efficient solution to optimize our proposed algorithm.

A. LC-LRD Revisit

Given a set of training data $X = [X_1, X_2, \ldots, X_c] \in \mathbb{R}^{d\times n}$, where d is the feature dimensionality, n is the number of total training samples, c is the number of classes, and $X_i \in \mathbb{R}^{d\times n_i}$ is the samples from class i which has n_i samples. The goal of dictionary learning is to learn an m atoms dictionary $D \in \mathbb{R}^{d\times m}$ which yields sparse representation matrix $A \in \mathbb{R}^{m\times n}$ from X for future classification tasks. Then we can write $X = DA + E$, where E is the sparse noise matrix. Rather than learning the dictionary as a whole from all the training samples, we learn sub-dictionary D_i for the i-th class separately. Then A and D could be written as $A = [A_1, A_2, \ldots, A_c]$ and $D = [D_1, D_2, \ldots, D_c]$, where A_i is the sub-matrix that denotes the coefficients for X_i over D.

In our conference paper [38], we have proposed the following LC-LRD model for each sub-dictionary:

$$\min_{D_i,A_i,E_i} R(D_i, A_i) + \alpha\|D_i\|_* + \beta\|E_i\|_1 + \lambda\sum_{k=1}^{n_i}\|l_{i,k}\odot a_{i,k}\|^2, \quad \text{s.t. } X_i = DA_i + E_i, \quad (1)$$

where $R(D_i, A_i)$ is the Fisher discriminant regularization on each sub-dictionary, $\|D_i\|_*$ is nuclear norm to enforce low-rank properties, and $\|l_{i,k}\odot a_{i,k}\|^2$ is locality constraint to replace sparsity on coding coefficient matrix. $a_{i,k}$ designates the k-th column in A_i, which means the coefficient for k-th sample in class i. We will break down this model into the following modules: discriminative sub-dictionaries, low-rank regularization term, the locality constraint on the coding coefficients.

Sub-dictionary D_i should be endowed with the discrimination power to well represent samples from i-th class. Mathematically, the coding coefficients of X_i over D can be written as $A_i = [A_i^1; A_i^2; \ldots; A_i^c]$, where A_i^j is the coefficient matrix of X_i over D_j. The discerning power of D_i is produced by following two aspects: first, it is expected that X_i should be well represented by D_i but not by D_j, $j \neq i$. Therefore, we will have to minimize $\|X_i - D_iA_i^i - \mathcal{E}_i\|_F^2$, where $\mathcal{E}_i$ is the residual. Meanwhile, D_i should not be good at representing samples from other classes, that is each A_j^i, where $j \neq i$ should have nearly zero coefficients so that $\|D_iA_j^i\|_F^2$ is as small as possible. Thus we denote the discriminative fidelity term for sub-dictionary D_i as follows:

$$R(D_i, A_i) = \|X_i - D_iA_i^i - \mathcal{E}_i\|_F^2 + \sum\nolimits_{j=1,j\neq i}^{c}\|D_iA_j^i\|_F^2. \quad (2)$$

In the task of image classification, the within-class samples are linearly correlated and lie in a low dimensional manifold. Therefore, we want to find the dictionary with the most concise atoms by minimizing the rank of D_i, which suggest to be replaced by $\|D_i\|_*$ [5], where $\|.\|_*$ denotes nuclear norm of a matrix (i.e., the sum of singular values of the matrix).

In addition, locality constraint is deployed on the coefficient matrix instead of the sparsity constraint. As suggested by

LCC [43], locality is more essential than sparsity under certain assumptions, as locality must lead to sparsity but not necessary vice versa. Specifically, the locality constraint uses the following criteria:

$$\min_{A} \sum_{i=1}^{n} \|l_i \odot a_i\|^2, \quad \text{s.t. } \mathbf{1}^T a_i = 1, \ \forall i, \tag{3}$$

where $\odot$ denotes the element-wise multiplication, and $l_i \in \mathbb{R}^m$ is the locality adaptor that gives different freedom for each basis vector proportional to its similarity to the input sample x_i. Specifically, $l_i = \exp(\frac{\text{dist}(x_i, D)}{\delta})$, where $\text{dist}(x_i, D) = [\text{dist}(x_i, d_1), \text{dist}(x_i, d_2), \ldots, \text{dist}(x_i, d_m)]^T$, and $\text{dist}(x_i, d_j)$ is the Euclidean distance between sample x_i and j-th dictionary atom d_j. δ controls the bandwidth of the distribution.

Generally speaking, LC-LRD is based on the following three observations: 1) Locality is more essential than sparsity to ensure obtain the similar representations for similar samples; 2) Each sub-dictionary should have discerning ability by introducing the discriminative term; 3) Low-rank is introduced to each sub-dictionary to separate noise from samples and discover the latent structure.

B. Marginalized Denoising Auto-Encoder (mDA)

Given the vector input $x \in \mathbb{R}^d$, with d as the dimensionality of the visual descriptor. There are two important transformation which can be considered as encoder and decoder processes involved in the auto-encoder: "input→hidden units", and "hidden units→output" as:

$$h = \sigma(Wx + b_h), \quad \hat{x} = \sigma(W^T h + b_o), \tag{4}$$

where $h \in \mathbb{R}^z$ is the hidden representation unit, and $\hat{x} \in \mathbb{R}^d$ is interpreted as a reconstruction of input x. The parameter set includes a weight matrix $W \in \mathbb{R}^{z \times d}$, and two offset vectors $b_h \in \mathbb{R}^z$ and $b_o \in \mathbb{R}^d$ for hidden and output, respectively. σ is a non-linear mapping such as the sigmoid function as the form $\sigma(x) = (1 + e^{-x})^{-1}$. In general, auto-encoder is a single layer hidden neural network, with identical input and target, meaning the auto-encoder encourages the output to be as similar to the target as possible, namely,

$$\min_{W, b_h, b_o} L(x) = \min_{W, b_h, b_o} \frac{1}{2n} \sum_{i=1}^{n} \|x_i - \hat{x}_i\|_2^2, \tag{5}$$

where n is the number of images, x_i is the target and $\hat{x}_i$ is the reconstructed input. In this way, the neurons in the hidden layer can be seen as a good representation for the input, since they are able to reconstruct the data.

Since auto-encoder deploys non-linear functions, it takes more time to train the model especially when the dimension of the data is very high. Recently, marginalized denoising auto-encoder (mDA) [6] was developed to address the data reconstruction in a linear fashion and achieved comparable performance with the original auto-encoder. The general idea of mDA is to learn a linear transformation matrix W to reconstruct the data with the transformation matrix by minimizing the squared reconstruction loss

$$\frac{1}{2n} \sum_{i=1}^{n} \|x_i - W\widetilde{x}_i\|^2, \tag{6}$$

where $\widetilde{x}_i$ is the corrupted version of x_i. The above objective solution is correlated to the randomly corrupted features of each input. To make the variance lower, marginal denoising auto-encoder was proposed to minimize the overall squared loss of t kinds of different corrupted versions

$$\frac{1}{2tn} \sum_{j=1}^{t} \sum_{i=1}^{n} \|x_i - W\widetilde{x}_{i,(j)}\|^2, \tag{7}$$

where $\widetilde{x}_{i,(j)}$ denotes the j^{th} corrupted version of the original input x_i. Define $X = [x_1, \cdots, x_n]$ and its t-times repeated version as $\overline{X} = [X, \cdots, X]$ with its t different corrupted version $\widetilde{X} = [\widetilde{X}_{(1)}, \cdots, \widetilde{X}_{(t)}]$, $\widetilde{X}_{(i)}$ denotes i^{th} corrupted version of X. In this way, Eq. (7) can be reformulated as

$$\frac{1}{2tn} \|\overline{X} - W\widetilde{X}\|_{\text{F}}^2, \tag{8}$$

which has the well-known closed-form solution for ordinary least squares. When $t \to \infty$, it can be solved with expectation with the weak law of large numbers [6].

C. Our Proposed Model

Previous discussion on mDA gives a brief idea, that with a linear transformation matrix, mDA can be implemented in several lines of Matlab code and works very efficiently. The learned transformation matrix can well reconstruct the data and dig out the noisy data.

Inspired by this, we aim at jointly learning a dictionary and a marginalized denoising transformation matrix in a unified framework. We formulate our objective function as

$$\begin{aligned} \min_{D_i, A_i, E_i, W} \ & \mathcal{F}(D_i, A_i, E_i) + \|\overline{X} - W\widetilde{X}\|_{\text{F}}^2, \\ \text{s.t. } & WX_i = DA_i + E_i \end{aligned} \tag{9}$$

where $\mathcal{F}(D_i, A_i, E_i) = R(D_i, A_i) + \alpha\|D_i\|_* + \beta_1\|E_i\|_1 + \lambda \sum_{k=1}^{n_i} \|l_{i,k} \odot a_{i,k}\|^2$ is our locality-constrained dictionary learning part in Eq. (1). And $R(D_i, A_i) = \|WX_i - D_i A_i^i - \mathcal{E}_i\|_F^2 + \sum_{j=1, j \neq i}^{c} \|D_i A_j^i\|_F^2$ is discriminative term in Eq. (2). α, β_1, and λ are trade-off parameters.

Discussion: The proposed marginalized denoising regularized dictionary learning (MDDL) aims to learn a more discriminative dictionary on transformed data. Since the marginalized denoising regularizer could generate a better transformation matrix to address the corrupted data, therefore, the dictionary could be learned on denoised clean data. In our framework, we unify the marginal denoising auto-encoder and locality-constrained dictionary learning together. Generally, dictionary learning seeks a well-represented basis in order to achieve more discriminative coefficients for original data. Therefore, dictionary learning can handle noisy data to some extent. While denoising auto-encoder has been demonstrated its denoising power in many applications. To this end, our joint learning scheme can benefit from both marginal denoising auto-encoder and locality-constrained dictionary learning.

D. Optimization

We consider solving the proposed objective function in Eq. (9) by dividing it into two sub-problems: First updating each coefficient $A_i (i = 1, 2, \ldots, c)$ one by one and W by

fixing dictionary D, all other $A_j (j \neq i)$ and putting together to get coding coefficient matrix A; Second, updating D_i by fixing others. This two steps are iteratively operated to get the discriminative low-rank sub-dictionaries D, the marginal denoising transformation W, and the locality-constrained coefficients A. One problem arises in the second sub-problem, remember in Eq. (2), the coefficients A_i^i corresponding to X_i over D_i should be updated to meet the condition $\|X_i - D_i A_i^i - \mathcal{E}_i\|_F^2$. Therefore, when we update D_i in the second sub-problem the related variance A_i^i is also updated.

1) Sub-Problem I: In the first sub-problem, assume that the structured dictionary D is given, the coefficients matrix $A_i (i = 1, 2, \ldots, c)$ is updated one by one, then the original objective function Eq. (9) reduces to the following locality-constrained coding problem for each class's coefficient and W:

$$\begin{aligned} \min_{A_i, E_i, W} \ & \lambda \sum_{k=1}^{n_i} \|l_{i,k} \odot a_{i,k}\|^2 + \beta_1 \|E_i\|_1 + \|\overline{X} - W\widetilde{X}\|_{\mathrm{F}}^2 \\ \text{s.t. } & WX_i = DA_i + E_i \end{aligned} \tag{10}$$

which can be solved by the following Augmented Lagrange Multiplier method [3]. We transform Eq. (10) into its Lagrange function as follows:

$$\begin{aligned} \min_{A_i, E_i, W, T_1} \ & \sum_{i=1}^{c} \Big(\lambda \sum_{k=1}^{n_i} \|l_{i,k} \odot a_{i,k}\|^2 + \beta_1 \|E_i\|_1 \\ & + \langle T_1, WX_i - DA_i - E_i \rangle + \frac{\mu}{2} \|WX_i \\ & - DA_i - E_i\|_F^2 \Big) + \|\overline{X} - W\widetilde{X}\|_{\mathrm{F}}^2, \end{aligned} \tag{11}$$

where T_1 is the Lagrange multiplier, and μ is a positive penalty parameter. Different from traditional locality-constrained linear coding (LLC) [36], we add an error term which could handle large noise in samples. In the following, we provide the iterative optimization on A_i, E_i, and W.

Updating A_i:

$$\begin{aligned} A_i \ &= \arg\min_{A_i} \frac{\mu}{2} \|Z_i - DA_i\|_F^2 + \lambda \sum_{k=1}^{n_i} \|l_{i,k} \odot a_{i,k}\|^2, \\ &\Rightarrow A_i = \mathrm{LLC}(Z_i, D, \lambda, \delta), \end{aligned} \tag{12}$$

where $Z_i = WX_i - E_i + \frac{T_1}{\mu}$, and $l_{i,k} = \exp(\mathrm{dist}(z_{i,k}, D)/\delta)$. Function LLC$(\cdot)$ is locality-constrained linear coding function[1] [36].

Updating E_i:

$$E_i = \arg\min_{E_i} \frac{\beta_1}{\mu} \|E_i\|_1 + \frac{1}{2} \|E_i - (WX_i - DA_i + \frac{T_1}{\mu})\|_F^2, \tag{13}$$

which can be solved by the shrinkage operator [39].

Updating W:

$$\begin{aligned} W \ &= \arg\min_{W} \sum_{i=1}^{c} \big(\frac{\mu}{2} \|WX_i - DA_i - E_i + \frac{T_1}{\mu}\|_F^2 \big) + \|\overline{X} - W\widetilde{X}\|_{\mathrm{F}}^2 \\ &= \arg\min_{W} \frac{\mu}{2} \|WX - D_A\|_{\mathrm{F}}^2 + \|\overline{X} - W\widetilde{X}\|_{\mathrm{F}}^2, \end{aligned} \tag{14}$$

[1]We set Z_i, D, λ and σ as the input of function LLC [36] and the code can be downloaded from http://www.ifp.illinois.edu/ jyang29/LLC.htm.

where $X = [X_1, \cdots, X_c]$ and $D_A = [DA_1 + E_1 - \frac{T_{1,1}}{\mu}, \cdots, DA_c + E_c - \frac{T_{1,c}}{\mu}]$. Eq. (14) has a well-known closed-form solution as follows:

$$W = (\mu D_A X^{\mathrm{T}} + 2\overline{X}\widetilde{X}^{\mathrm{T}})(\mu XX^{\mathrm{T}} + 2\widetilde{X}\widetilde{X}^{\mathrm{T}})^{-1} \tag{15}$$

where $\overline{X}$ is t-times repeated version of X and $\widetilde{X}$ consists of its t kinds of corrupted version. We define $P = \mu D_A X^{\mathrm{T}} + 2\overline{X}\widetilde{X}^{\mathrm{T}}$ and $Q = \mu XX^{\mathrm{T}} + 2\widetilde{X}\widetilde{X}^{\mathrm{T}}$. And we would like the repeated number t to be ∞, therefore, the denoising transformation W could be effectively learned from infinitely many copies of noisy data. Practically we cannot generate $\widetilde{X}$ with infinitely versions of corruption, however fortunately, the matrices P and Q converge to their expectations when t becomes very large. In this way, we can derive the expected values of P and Q, and calculate the corresponding mapping W as:

$$\begin{aligned} W \ &= \mathbb{E}[P]\mathbb{E}[Q]^{-1} \\ &= \mathbb{E}[\mu D_A X^{\mathrm{T}} + 2\overline{X}\widetilde{X}^{\mathrm{T}}]\mathbb{E}[\mu XX^{\mathrm{T}} + 2\widetilde{X}\widetilde{X}^{\mathrm{T}}]^{-1} \\ &= \Big(\mu D_A X^{\mathrm{T}} + 2\mathbb{E}[\overline{X}\widetilde{X}^{\mathrm{T}}]\Big)\Big(\mu XX^{\mathrm{T}} + 2\mathbb{E}[\widetilde{X}\widetilde{X}^{\mathrm{T}}]\Big)^{-1} \end{aligned} \tag{16}$$

where D_A and μ are treated as constant values when optimizing W. The expectations $\mathbb{E}[\overline{X}\widetilde{X}^{\mathrm{T}}]$ and $\mathbb{E}[\widetilde{X}\widetilde{X}^{\mathrm{T}}]$ are easy to be computed through mDA [6].

2) Sub-Problem II: For the procedure of updating sub-dictionary, we have the same method with [23]. Considering the second sub-problem, when A_i is fixed, sub-dictionary $D_i (i = 1, 2, \ldots, c)$ is updated one by one. The objective function Eq. (9) is converted to the following problem:

$$\begin{aligned} \min_{D_i, \mathcal{E}_i, A_i^i} \ & \sum_{j=1, j\neq i}^{c} \|D_i A_j^i\|_F^2 + \alpha \|D_i\|_* + \beta_2 \|\mathcal{E}_i\|_1 \\ & + \lambda \sum_{k=1}^{n_i} \|l_{i,k}^i \odot a_{i,k}^i\|^2, \\ \text{s.t. } & WX_i = D_i A_i^i + \mathcal{E}_i \end{aligned} \tag{17}$$

where $a_{i,k}^i$ is the k-th column in A_i^i, which means the coefficient for k-th sample in class i over D_i. Problem Eq. (17) can be solved using the Augmented Lagrange Multiplier method [3] by introducing a relaxing variable J:

$$\begin{aligned} \min_{D_i, \mathcal{E}_i, A_i^i} \ & \lambda \sum_{k=1}^{n_i} \|l_{i,k}^i \odot a_{i,k}^i\|^2 + \sum_{j=1, j\neq i}^{c} \|D_i A_j^i\|_F^2 + \alpha \|J\|_* \\ & + \beta_2 \|\mathcal{E}_i\|_1 + \langle T_2, WX_i - D_i A_i^i - \mathcal{E}_i \rangle + \langle T_3, D_i - J \rangle \\ & + \frac{\mu}{2} (\|WX_i - D_i A_i^i - \mathcal{E}_i\|_F^2 + \|D_i - J\|_F^2), \end{aligned} \tag{18}$$

where T_2 and T_3 are Lagrange multipliers, and μ is a positive penalty parameter. In the following, we provide the iterative optimization on D_i and A_i^i.

Updating A_i^i: Similar as Eq. (12), we have the solution for A_i^i as follow:

$$A_i^i = \mathrm{LLC}((WX_i - \mathcal{E}_i + \frac{T_2}{\mu}), D_i, \lambda, \delta), \tag{19}$$

where function LLC$(\cdot)$ is locality-constrained linear coding function [36].

Algorithm 1 Optimization for MDDL

Input: Training data $X = [X_1, \ldots, X_c]$, Parameters α, λ, δ, β_1, β_2
Output: W, A=[$A_1, A_2, \ldots, A_c$], D=[$D_1, D_2 \ldots, D_c$]
Initialize: $W = I$, PCA initialized D, $E_i = \mathcal{E}_i = 0$, $T_1 = 0$, $T_2 = 0$, $T_3 = 0$, $\mu_{max} = 10^{30}$, $\rho = 1.1$, $\epsilon = 10^{-8}$, $maxiter = 10^4$
repeat
  $iter = 0$, $\mu = 10^{-6}$
  %Solving Eq. (11) via ALM
  while *not converge and iter* $\leq maxiter$ **do**
    Fix others and update A_i with Eq. (12)
    Fix others and update E_i with Eq. (13)
    Fix others and update W with Eq. (16)
    Update multipliers T_1 by:
      $T_1 = T_1 + \mu(WX_i - DA_i - E_i)$
    Update parameter μ by:
      $\mu = \min(\rho\mu, \mu_{max})$
    Check the convergence conditions:
      $\|WX_i - DA_i - E_i\|_\infty < \epsilon$
  end
  $iter = 0$, $\mu = 10^{-6}$
  %Solving Eq. (17) via ALM
  while *not converge and iter* $\leq maxiter$ **do**
    Fix others and update A_i^i with Eq. (19)
    Fix others and update J and D_i with Eq. (20) and Eq. (21)
    Fix others and update $\mathcal{E}_i$ with Eq. (22)
    Update multipliers T_2 and T_3 by:
      $T_2 = T_2 + \mu(WX_i - D_iA_i^i - \mathcal{E}_i)$
      $T_3 = T_3 + \mu(D_i - J)$
    Update parameter μ by: $\mu = \min(\rho\mu, \mu_{max})$
    Check convergence: $\|WX_i - DA_i^i - E_i\|_\infty < \epsilon$
  end
until The sub-dictionary converges or the maximal iteration is reached;

Updating J and D_i: Here we convert the Eq.(18) to related with J and D_i as:

$$\begin{aligned}\min_{J,D_i} \sum_{j=1,j\neq i}^{c} \|D_iA_j^i\|_F^2 + \alpha\|J\|_* \\ + \langle T_2, WX_i - D_iA_i^i - \mathcal{E}_i\rangle + \langle T_3, D_i - J\rangle \\ + \frac{\mu}{2}(\|D_i - J\|_F^2 + \|WX_i - D_iA_i^i - \mathcal{E}_i\|_F^2), \end{aligned} \quad (20)$$

where $J = \arg\min_J \alpha\|J\|_* + \langle T_3, D_i - J\rangle + \frac{\mu}{2}(\|D_i - J\|_F^2)$, and the solution for D_i is:

$$\begin{aligned} D_i &= (J + WX_iA_i^{iT} - \mathcal{E}_iA_i^{iT} + (T_2A_i^{iT} - T_3)/\mu) \\ &\quad \times (I + A_i^iA_i^{iT} + V)^{-1}, \\ \text{where}\quad V &= \frac{2\lambda}{\mu}\sum_{j=1,j\neq i}^{c} A_j^iA_j^{iT} \end{aligned} \quad (21)$$

Updating $\mathcal{E}_i$:

$$\begin{aligned}\mathcal{E}_i = \arg\min_{\mathcal{E}_i} \beta_2\|E_i\|_1 + \langle T_2, WX_i - D_iA_i^i - \mathcal{E}_i\rangle \\ + \frac{\mu}{2}\|WX_i - D_iA_i^i - \mathcal{E}_i\|_F^2, \end{aligned} \quad (22)$$

which can be solved by the shrinkage operator [39]. The detail of the Optimization Solution for proposed model can be referred to Algorithm 1.

E. Classification Based on Our Model

We use a linear classifier for classification. After the dictionary is learned, the locality-constrained coefficients A of training data X and A_{test} of test data X_{test} are calculated.

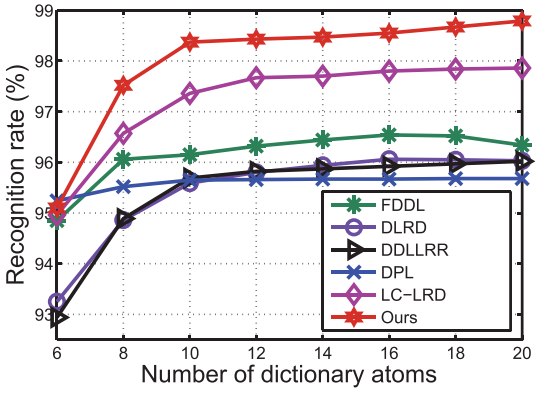

Fig. 2. The recognition rates of six DL based methods versus the number of dictionary atoms with 20 training samples per class on Extend YaleB dataset.

The representation a_i for test sample i is the i-th column vector in A_{test}. We use the multivariate ridge regression model [46] to obtain a linear classifier $\hat{P}$:

$$\hat{P} = \arg\min_P \|L - PA\|_F^2 + \gamma\|P\|_F^2 \quad (23)$$

where L is the class label matrix. This yields $\hat{P} = LA^T(AA^T + \gamma I)^{-1}$. When testing points A_{test} comes in, we first compute $\hat{P}A_{\text{test}}$. Then label for sample i is assigned by the position corresponding to the largest value in the label vector, that is: $label = \arg\max_{label}(\hat{P}a_i)$.

IV. Experiments

To verify the effectiveness and generality of the proposed MDDL, we conduct experiments on various visual classification applications. The method is tested on six datasets including four face datasets: ORL [30], Extend YaleB [18], AR [24], CMU PIE [33], one object categorization dataset COIL-100 [26], and digits recognition dataset MNIST [17].

We conduct the experiments in comparison with LDA [1], linear regression classification (LRC) [25] and several latest dictionary learning based classification methods, i.e., FDDL [42], DLRD [23], D^{2}L^{2}R^2 [19], DPL [12] and also our conference paper LC-LRD [38]. What's more, for verifying the advantage of joint learning, we proposed a simple combination framework as a baseline, named as AE+DL, which first uses a traditional SAE to learn a new representation, then feeds in our conference paper's dictionary learning framework.

A. Parameter Selection

The number of atoms in every sub-dictionary, which denoted as m_i, is one of the most important parameters in most of dictionary learning algorithms. We conduct the experiment on Extended YaleB with different number of dictionary atomsm_i and analyze its effect on the performance of our proposed new model and other competitors. Fig. 2 shows that all comparisons obtain an increasing performance with larger dictionary size. In the experiments, we fix the dictionary columns of each class as training size for ORL, Extend YaleB, AR and COIL-100 datasets, while fix as 30 for CMU PIE and

TABLE I

AVERAGE RECOGNITION RATE(%) OF DIFFERENT ALGORITHMS ON EXTENDED YALEB DATASET WITH DIFFERENT NUMBER OF TRAINING SAMPLES PER CLASS. RED DENOTES THE BEST RESULTS WHILE THE BLUE MEANS THE SECOND BEST RESULTS

Training	LDA [1]	LRC [25]	FDDL [42]	DLRD [23]	$D^2L^2R^2$ [19]	DPL [12]	LC-LRD [38]	AE+DL	MDDL[ours]
5	74.12±1.52	60.24±2.02	77.75±1.34	76.17±1.16	75.96±1.20	75.17±1.86	78.62±1.20	78.64±1.12	**79.11±1.16**
10	86.67±0.90	82.98±0.82	91.16±0.85	89.94±0.89	89.60±0.89	89.31±0.62	92.07±0.89	92.10±0.88	**92.19±0.80**
20	90.64±1.07	91.80±0.97	96.15±0.66	96.03±0.85	96.02±0.91	95.69±0.90	97.86±0.91	96.56±0.89	**98.77±0.67**
30	86.84±0.92	94.60±0.60	97.86±0.35	97.90±0.47	97.87±0.42	97.80±0.36	99.23±0.47	98.64±0.47	**99.35±0.20**
40	95.27±0.79	96.10±0.58	98.84±0.46	98.80±0.37	98.09±0.39	98.67±0.43	99.54±0.44	99.23±0.39	**99.78±0.16**

TABLE II

AVERAGE RECOGNITION RATE(%) OF DIFFERENT ALGORITHMS ON EXTENDED YALEB DATASET WITH VARIOUS CORRUPTION PERCENTAGE(%). RED DENOTES THE BEST RESULTS WHILE THE BLUE MEANS THE SECOND BEST RESULTS

Corruption	LDA [1]	LRC [25]	FDDL [42]	DLRD [23]	$D^2L^2R^2$ [19]	DPL [12]	LC-LRD [38]	AE+DL	MDDL[ours]
0	86.84±0.92	94.60±0.60	97.86±0.35	97.90±0.47	97.87±0.42	97.80±0.36	99.23±0.47	98.64±0.47	**99.35±0.20**
5	29.03±0.82	80.49±1.10	63.55±0.87	91.84±1.07	91.90±1.14	78.27±1.22	93.31±0.69	93.15±0.66	**93.64±0.44**
10	18.53±1.15	67.61±1.33	44.65±1.22	85.82±1.54	85.71±1.51	64.58±1.09	86.97±0.86	87.05±0.89	**87.54±0.76**
15	13.63±0.53	56.81±1.24	32.76±1.03	80.89±1.37	80.46±1.64	53.77±0.86	81.71±0.81	81.55±0.86	**82.10±0.64**
20	11.30±0.46	47.23±1.59	25.26±0.42	73.56±1.63	73.59±1.54	44.95±1.38	74.14±1.01	74.18±1.60	**76.33±1.45**

MNIST datasets. All the dictionaries are initialized with PCA on input data.

There are five parameters in Algorithm 1: α, λ, δ along with β_1, β_2 as two error term parameters respectively for updating dictionary and coefficients. These five are associated with dictionary learning part in our new model and are chosen by 5-fold cross validation. Experiments show that β_1 and β_2 play more important roles than the other parameters, therefore we set $\alpha = 1$, $\lambda = 1$ and $\delta = 1$ in this paper. For Extended YaleB, $\beta_1 = 15$, $\beta_2 = 100$; for ORL, $\beta_1 = 5$, $\beta_2 = 50$; for AR, $\beta_1 = 5$, $\beta_2 = 100$; for CMU PIE, $\beta_1 = 5$, $\beta_2 = 1.5$; for COIL-100, $\beta_1 = 3$, $\beta_2 = 150$; for MNIST, $\beta_1 = 2.5$, $\beta_2 = 2.5$.

B. Extended YaleB Dataset

The Extended Yale Face Database B contains 2414 frontal-face images from 38 human subjects captured under various laboratory-controlled illumination conditions. The size of image is cropped to 32×32. Two experiments are deployed on this dataset. First, we choose random subsets with $p(= 5, 10, \ldots, 40)$ images per individual taken with labels to form the training set, and the rest of the dataset was considered to be the testing set. For each given p, there are 10 randomly splits; Second, we replace a certain percentage of randomly selected pixels from the images with pixel value of 255 (show in Fig. 3 (c)). Then randomly take 30 images as training samples, with the rest as testing and the experiment is also repeated ten times. These two experimental results are given in Table I and Table II, respectively.

We can observe from Table I that with different training sizes setting our three methods (including conference model) archive the top accuracy, and the proposed MDDL performs all the best. Our method's robustness to noise is demonstrated in Table II, along with the percentage of corruption increases our algorithms still produce best recognition results. The performance of LDA as well as LRC, FDDL and DPL drops rapidly under larger corruption, while our methods (LC-LRD, MDDL), $D^2L^2R^2$ and DLRD can still get much better

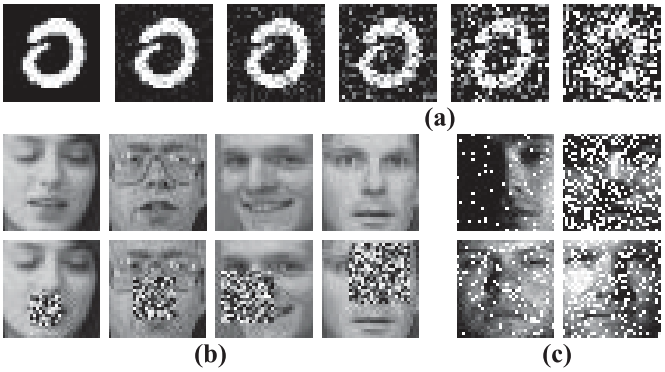

Fig. 3. Example images from three datasets. (a) images with 30db, 20db, 10db 5db, 1db SNR addition white gaussian noise from MNIST digit dataset; (b) ORL with 10%, 20%, 30% block occlution; (c) Extended YaleB with 10%, 15%, 20%, 25% random pixel corruption.

recognition accuracies. This demonstrates the effectiveness of low-rank regularization and the error term when noise exists. Our conference model and simply AE+DL equally matched in different situations, while MDDL performs best constantly.

C. ORL Face Database

The ORL dataset consists 400 images of 40 individuals, such that there are 10 images for each subject with varying pose and illumination. The subjects of the images are in frontal and upright posture while the background is dark and uniform. The images are resized to 32×32, converted to gray scale, normalized and the pixels are concatenated to form a vector. Each image is manually corrupted by an random located and unrelated block image. Fig. 3 (b) shows four examples of images with increasing block corruptions. For each subject, we select 5 samples for training and the rest as testing and repeat the experiment on 10 random splits for evaluation. Furthermore, SIFT and Gabor filter features are extracted to evaluate our methods generality.

We illustrate the recognition rates under different percentages of occlusions in Table III. From the table, we can observe

TABLE III

AVERAGE RECOGNITION RATE(%) OF DIFFERENT ALGORITHMS ON ORL DATASET WITH VARIOUS OCCLUSION PERCENTAGE(%). RED DENOTES THE BEST RESULTS WHILE THE BLUE MEANS THE SECOND BEST RESULTS

Occlusion	LDA [1]	LRC [25]	FDDL [42]	DLRD [23]	$D^2L^2R^2$ [19]	DPL [12]	LC-LRD [38]	AE+DL	MDDL[ours]
0	92.50±1.81	91.75±1.60	96.00±1.24	93.50±1.52	93.90±1.81	94.10±1.77	96.70±1.42	96.25±1.22	**96.75±1.25**
0 (SIFT)	95.75±1.27	92.85±2.06	95.20±1.34	93.65±1.33	93.85±1.49	95.25±1.43	93.50±2.60	96.00±1.25	**96.25±1.40**
0 (Gabor)	88.95±3.27	93.40±1.74	96.00±1.31	96.30±1.31	96.60±1.25	97.00±1.36	94.60±1.80	96.70±1.06	**97.40±1.15**
10	71.65±3.23	82.20±2.15	86.60±1.90	91.25±1.87	91.00±1.86	84.50±2.73	**92.25±1.25**	91.45±1.23	92.00±1.38
20	54.25±2.01	71.30±2.80	75.25±3.39	82.80±2.97	82.80±3.32	71.15±1.70	83.95±2.31	83.55±1.76	**84.25±2.19**
30	40.45±3.69	63.65±3.06	63.75±2.67	78.90±3.11	78.80±3.28	59.80±3.85	**80.15±2.94**	78.90±2.86	78.00±2.44
40	25.65±2.46	48.00±3.04	48.05±2.40	67.30±3.24	67.40±3.42	43.00±2.94	**67.95±3.01**	67.30±2.60	67.50±3.21
50	20.65±3.02	40.85±3.71	36.70±1.24	58.65±3.10	58.70±3.27	32.20±3.52	**58.85±3.47**	58.65±3.24	58.50±3.01

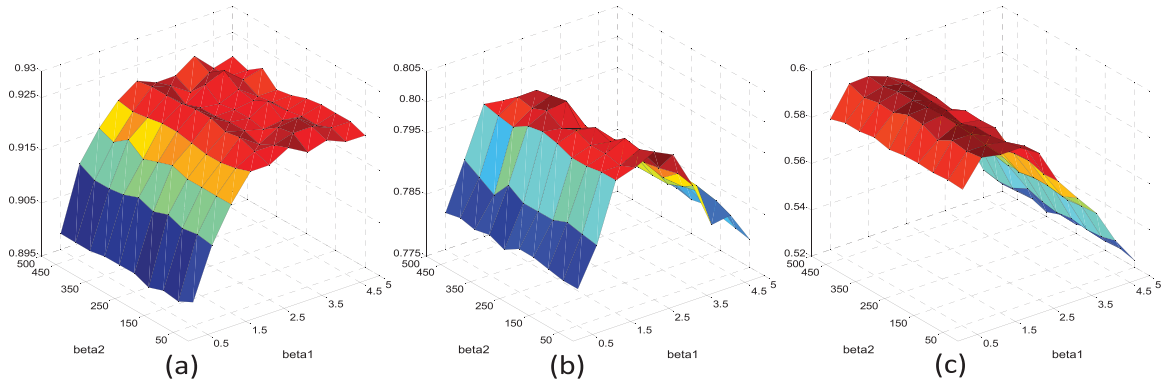

Fig. 4. Our method's performance (a)-(c) under increasing percentage of corrupted pixels versus different parameters. As more occlusion applied, the best result appears when the parameter β_1 is smaller, which means the error term plays more important role when noise exists.

two phenomenons: first, our methods achieve the top results all the settings; second, the new model performs best when the data is clean, however, along with the percentage of occlusions increases MDDL drop behind with our conference paper. That makes sense because in this experiment we add occlusion on to the images, while the denoising auto-encoder module in our methods are introduced to tackle with gaussian noise. In conclusion: first, our method can achieve top results in no occlusion situation, because of the locality term; second, in larger occlusion situation, low-rank term outweighs DAEs.

The effects from two parameters of the error term β_1 and β_2 are demonstrated in Fig. 4. From the three sub-figures under increasing percentage of corrupted pixels, the parameter β_1 in the coefficients updating procedure makes larger different. As more occlusion applied, the best result appears when the parameter β_1 is smaller, which means the error term plays more important role when noise exists.

The results show that our methods have significant improvement on some datasets, and for some other datasets, the significance increases along with the noise level in the input.

D. AR Dataset

The AR dataset consists of over 4,000 frontal-face images of 126 individuals, that is, there are 26 pictures for each subject taken in two separated sessions. We follow the experimental setting in [42], for fair comparison, to choose a subset consisting of 50 male subjects and 50 female subjects. For each subject, the 13 images with with different facial expressions, illumination conditions, and occlusions from session 1 were used for training, and the other 13 images with the same

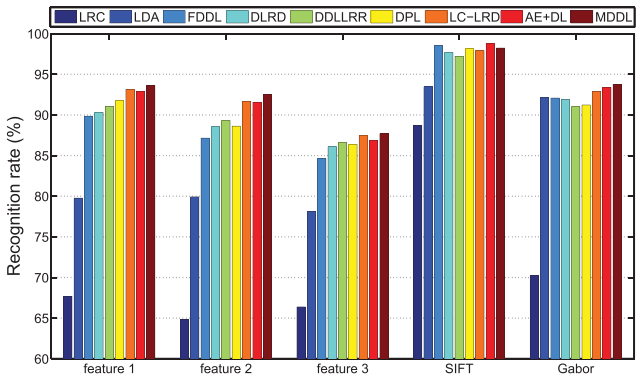

Fig. 5. Average recognition rate(%) of different algorithms on AR dataset with five different features. Feature 1: row pixel 60×43; feature 2: row pixel 27×20; feature 3: feature provided by [16].

TABLE IV

CLASSIFICATION ERROR RATES(%) ON CMU PIE DATASET

Methods	CMU (near frontal poses)	CMU (all poses)
LRC [25]	4.12	9.65
FDDL [42]	3.30	11.20
DLRD [23]	3.33	10.64
$D^2L^2R^2$ [19]	3.29	10.14
DPL [12]	3.47	9.30
LC-LRD [38][ours]	3.01	8.98
MDDL [ours]	**2.74**	**7.64**

condition from session 2 were used for testing. We do experiments on different features: original 60×43 images, resized 27×20 images, SIFT, Gabor and the feature provided by [16].

We illustrate the recognition rates under different features in Fig. 5. From the figure, we can observe that our proposed method achieves the best results on most the features.

TABLE V

AVERAGE RECOGNITION RATE(%) WITH STANDARD DEVIATIONS OF DIFFERENT ALGORITHMS ON COIL-100 DATASET WITH DIFFERENT NUMBER OF CLASSES. RED DENOTES THE BEST RESULTS WHILE THE BLUE MEANS THE SECOND BEST RESULTS

Class No.	LDA [1]	LRC [25]	FDDL [42]	DLRD [23]	$D^2L^2R^2$ [19]	DPL [12]	LC-LRD [38]	AE+DL	MDDL[ours]
20	81.94±1.21	90.74±0.71	85.74±0.77	88.61±0.95	90.98±0.38	87.55±1.32	**92.15±0.34**	91.26±0.45	91.57±0.41
40	76.73±0.30	89.00±0.46	82.05±0.40	86.39±0.54	88.27±0.38	85.05±0.21	89.86±0.49	89.09±0.66	**92.23±0.26**
60	66.16±0.97	86.57±0.37	77.22±0.74	83.46±0.15	86.36±0.53	81.22±0.21	87.12±0.66	87.23±0.29	**88.05±0.30**
80	59.19±0.73	85.09±0.34	74.81±0.55	81.50±0.47	84.69±0.45	78.78±0.85	85.40±0.61	85.06±0.47	**86.15±0.31**
100	52.48±0.53	83.16±0.64	73.55±0.63	79.91±0.59	83.06±0.37	76.28±0.94	84.15±0.39	84.12±0.39	**85.31±0.33**

E. CMU PIE Dataset

The CMU PIE dataset contains 41,368 face images from 68 identities, each with 13 different poses, 4 different expressions, and 43 different lighting conditions. We deploy two experiments on two subsets of CMU PIE respectively. First of all, we select five near frontal poses (C05, C07, C09, C27, C29) under different illuminations and expressions as a first subset (11,554 samples in total). Thus, there are about 170 images for each person. We select 60 images per person as training. Secondly, we choose more poses to build a relatively large-scale dataset, which contains totally 24,245 samples. In total, there are around 360 images for each person. Each image is normalized to the size of 32×32 pixels for both experiments. The training set is constructed by randomly selected 200 images per person, while the rest is used for evaluation. Table IV shows that our method outperforms the compared methods.

F. COIL-100 Dataset

In this section, we evaluate our approach on object categorization by using the COIL-100 dataset. The training set is constructed by randomly selected 10 images per object, and the testing set contains the rest of the images. We repeat this random selection ten times, and report the average results of all the compared methods. To evaluate the scalability of different methods, we separately utilize images of 20, 40, 60, 80 and 100 objects from the dataset. Table V shows the average recognition rates with standard deviations of all compared methods. The results show our algorithm could not only work on face recognition but also on object categorization.

G. MNIST Dataset

We test our algorithm on a subset of MNIST handwritten digit dataset, which includes first 2000 training images and first 2000 test images with the size of each digit image is 16×16. This experimental setting follows [19], and we get consistent results. Table VI summarizes the recognition rates and traing/testing time by different algorithms. Our algorithm achieves the highest accuracy than its competitors. Compared with our conference method LC-LRD, our MDDL costs only slightly more computational time thanks to the easy updating of marginalized auto-encoder.

Another experiment setting is conducted on this dataset to evaluate the effect from denoising auto-encoders. All the training and testing images in MNIST are added with additive white Gaussian noise corresponding with signal-to-noise ratio (snr) from 50dB to 1dB (show in Fig. 3 (a)). Fig 6 illustrates the recognition rate curves on 8 noised version of datasets. X-axis verses the noise ratio used in input reconstruction process in DAEs, close to 1 means more noise added, 0 means no DAEs evolved. From the figure, we can observe that, with the increasing noise added in the datasets (50dB to 1dB), the highest recognition rate appears when the noise parameter goes larger (from nearly 0.004 for 50dB to nearly 0.1 for 1dB). In another word, denoising auto-encoders plays more important role when the datasets carries heavier noise.

To verify if our improvement is statistically significant, we further conduct a significance test (t-test) for the results shown in Fig. 7. A significance level of 0.05 was used, that is to say, when p-value is less than 0.05, the performance difference of two methods is statistically significant. The p-values of our method and other competitors are listed in Fig. 7 lists. Since we do $-\log(p)$ processing, the comparison shows that our method outperforms others significantly if the

TABLE VI

AVERAGE RECOGNITION RATE(%) & RUNNING TIME(*second*) ON MNIST DATASET

Methods	Accuracy	Training time	Testing time
LDA [1]	77.45	0.164	0.545
LRC [25]	82.70	227.192	–
FDDL [42]	85.35	240.116	97.841
DLRD [23]	86.05	156.575	48.373
$D^2L^2R^2$ [19]	84.65	203.375	48.154
DPL [12]	84.65	1.773	0.847
LC-LRD [38]	88.25	80.581	48.970
AE+DL	87.95	176.525	49.230
MDDL [ours]	**89.75**	81.042	49.781

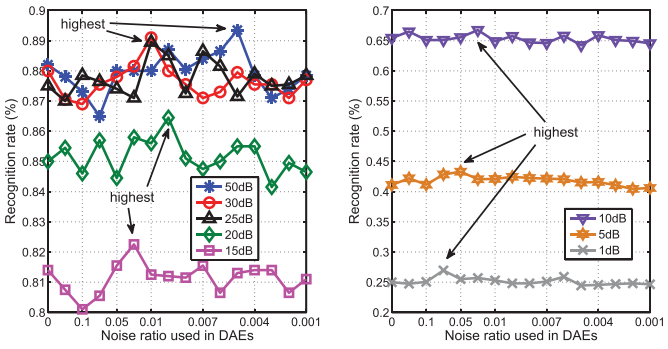

Fig. 6. The performance on MNIST datsets with different snr noise. As snr goes lower, the best result appears when the noise on reconstruction process is larger, which means the DAEs plays more important role when noise goes heavier on MNIST dataset.

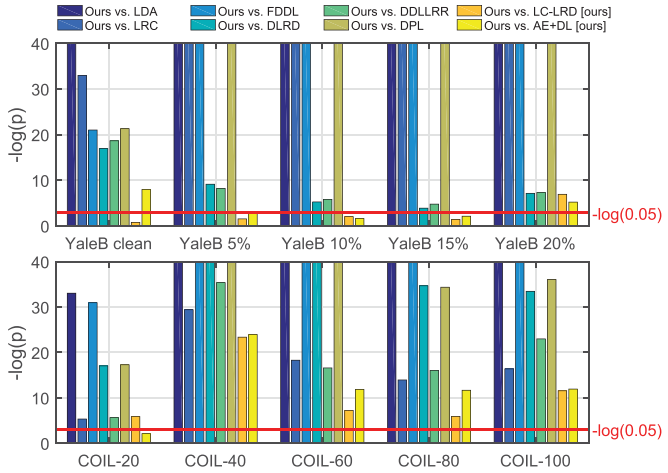

Fig. 7. p-value of t-test between our method and others on Extended YaleB (upper figure, with 0% to 20% corruption) and COIL-100 (lower figure, with 20-100 classes) datasets. We do pre-processing using $-\log(p)$ so that the larger value shown in the figure means the more significance of our algorithm compared with others.

values are greater than $-\log(0.05)$. The results show that our methods have significant improvement on COIL-100 dataset, and for Extended YaleB dataset, the significance increases along with the noise level in the input data.

V. Conclusion

In this paper, we developed an efficient marginalized denoising dictionary learning (MDDL) framework with locality constraint. Our proposed algorithm was designed to take advantage of two feature learning schemes, dictionary learning and auto-encoder. Specifically, we adopted a lite version of auto-encoder to seek a denoising transformation matrix. Then, dictionary learning with locality constraint was built on the transformed data. These two strategies were iteratively optimized so that a marginalized denoising transformation and a locality-constrained dictionary were jointly learned. Experiments on several image datasets, e.g., face, object, digits, demonstrated the superiority of our proposed algorithm by comparing other existing dictionary algorithms.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[2] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[3] D. P. Bertsekas, "Constrained optimization and lagrange multiplier methods," in *Computer Science and Applied Mathematics*. Boston, MA, USA: Academic, 198.

[4] M. Ranzato, Y.-L. Boureau, and Y. L. Cun, "Sparse feature learning for deep belief networks," in *Proc. NIPS*, 2008, pp. 1185–1192.

[5] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, 2011, Art. no. 11.

[6] M. Chen, Z. Xu, K. Weinberger, and F. Sha. (2012). "Marginalized denoising autoencoders for domain adaptation." [Online]. Available: https://arxiv.org/abs/1206.4683

[7] Y. Chen and X. Guo, "Learning non-negative locality-constrained linear coding for human action recognition," in *Proc. IEEE VCIP*, Nov. 2013, pp. 1–6.

[8] Z. Ding and Y. Fu, "Low-rank common subspace for multi-view learning," in *Proc. IEEE ICDM*, Dec. 2014, pp. 110–119.

[9] Z. Ding, M. Shao, and Y. Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1192–1198.

[10] Z. Ding, M. Shao, and Y. Fu, "Deep low-rank coding for transfer learning," in *Proc. IJCAI*, 2015, pp. 3453–3459.

[11] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *Proc. IEEE CVPR*, Jun. 2011, pp. 1873–1879.

[12] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Projective dictionary pair learning for pattern classification," in *Proc. NIPS*, 2014, pp. 793–801.

[13] M. G. Jafari and M. D. Plumbley, "Fast dictionary learning for sparse representations of speech signals," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 1025–1031, Sep. 2011.

[14] 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.

[15] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. IEEE CVPR*, Jun. 2011, pp. 1697–1704.

[16] 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.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[18] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[19] L. Li, S. Li, and Y. Fu, "Learning low-rank and discriminative dictionary for image classification," *Image Vis. Comput.*, vol. 32, no. 10, pp. 814–823, 2014.

[20] Y. Liang, M. Song, J. Bu, and C. Chen, "Colorization for gray scale facial image by locality-constrained linear coding," *J. Signal Process. Syst.*, vol. 74, no. 1, pp. 59–67, 2014.

[21] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[22] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. ICML*, 2010, pp. 663–670.

[23] L. Ma, C. Wang, B. Xiao, and W. Zhou, "Sparse representation for face recognition based on discriminative low-rank dictionary learning," in *Proc. IEEE CVPR*, Jun. 2012, pp. 2586–2593.

[24] A. Martínez and R. Benavente, "The AR face database," CVC, Luxembourg, Tech. Rep. 24, 1998.

[25] 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.

[26] S. Nayar, S. A. Nene, and H. Murase, "Columbia object image library (COIL 100)," Dept. Comp. Sci., Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-006-96, 1996.

[27] R. Pique-Regi, J. Monso-Varona, A. Ortega, R. C. Seeger, T. J. Triche, and S. Asgharzadeh, "Sparse representation and Bayesian detection of genome copy number alterations from microarray data," *Bioinformatics*, vol. 24, no. 3, pp. 309–318, 2008.

[28] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Action classification with locality-constrained linear coding," in *Proc. IEEE ICPR*, Aug. 2014, pp. 3511–3516.

[29] F. Rodriguez and G. Sapiro, "Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries," Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep., 2008.

[30] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. 2nd IEEE Workshop Appl. Comput. Vis.*, Dec. 1994, pp. 138–142.

[31] A. Shabou and H. LeBorgne, "Locality-constrained and spatially regularized coding for scene categorization," in *Proc. IEEE CVPR*, Jun. 2012, pp. 3618–3625.

[32] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. IEEE CVPR*, Jun. 2012, pp. 853–860.

[33] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.

[34] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. ACM ICML*, 2008, pp. 1096–1103.
[35] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, Dec. 2010.
[36] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE CVPR*, Jun. 2010, pp. 3360–3367.
[37] S. Wang, Z. Ding, and Y. Fu, "Feature selection guided auto-encoder," in *Proc. AAAI*, 2017, pp. 2725–2731.
[38] S. Wang and Y. Fu, "Locality-constrained discriminative learning and coding," in *Proc. CVPRW*, Jun. 2015, pp. 17–24.
[39] J. Yang, W. Yin, Y. Zhang, and Y. Wang, "A fast algorithm for edge-preserving variational multichannel image restoration," *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 569–592, 2009.
[40] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE CVPR*, Jun. 2009, pp. 1794–1801.
[41] M. Yang, D. Dai, L. Shen, and L. Van Gool, "Latent dictionary learning for sparse representation based classification," in *Proc. CVPR*, Jun. 2014, pp. 4138–4145.
[42] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE ICCV*, Nov. 2011, pp. 543–550.
[43] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. NIPS*, 2009, pp. 2223–2231.
[44] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma, "Image classification by non-negative sparse coding, low-rank and sparse decomposition," in *Proc. IEEE CVPR*, Jun. 2011, pp. 1673–1680.
[45] D. Zhang, P. Liu, K. Zhang, H. Zhang, Q. Wang, and X. Jinga, "Class relatedness oriented-discriminative dictionary learning for multiclass image classification," *Pattern Recognit.*, vol. 59, pp. 168–175, Nov. 2015.
[46] G. Zhang, Z. Jiang, and L. S. Davis, "Online semi-supervised discriminative dictionary learning for sparse representation," in *Proc. ACCV*, 2013, pp. 259–273.

Shuyang Wang received the B.Eng. degree in technology and apparatus of measuring and control from Beihang University, Beijing, China, in 2013. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA. His current research interests include low-rank matrix recovery, machine learning, and computer vision. He has received the ACM MM Travel Award for the ACM International Conference on Multimedia in 2014 and the AAAI Student Travel Award for the AAAI Conference on Artificial Intelligence in 2016. He has served as the reviewers for IEEE journals, such as the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.

Zhengming Ding (S'14) received the B.Eng. degree in information security and the M.Eng. degree in computer software and theory from the University of Electronic Science and Technology of China, China, in 2010 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Northeastern University, USA. His research interests include machine learning and computer vision. Specifically, he devotes himself to develop scalable algorithms for challenging problems in transfer learning scenario. He is an AAAI Student Member. He was a recipient of the Student Travel Grant of ACM MM 14, ICDM 14, AAAI 16, and IJCAI 16. He received the National Institute of Justice Fellowship. He was a recipient of the Best Paper Award (SPIE). He has served as the reviewers for IEEE journals, such as the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.

Yun Fu (S'07–M'08–SM'11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, China, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign. He has been an Interdisciplinary Faculty Member with the College of Engineering and the College of Computer and Information Science, Northeastern University, since 2012. He has extensive publications in leading journals, books/book chapters, and international conferences/workshops. His research interests are machine learning, computational intelligence, big data mining, computer vision, pattern recognition, and cyber-physical systems. He is a fellow of IAPR, a Lifetime Senior Member of ACM and SPIE, a Lifetime Member of AAAI, OSA, and the Institute of Mathematical Statistics, a member of the Global Young Academy, INNS, and a Beckman Graduate Fellow from 2007 to 2008. He received seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, Grainger Foundation; seven Best Paper Awards from IEEE, IAPR, SPIE, SIAM; and three major Industrial Research Awards from Google, Samsung, and Adobe. He serves as an associate editor, the chair, a PC member, and a reviewer for many top journals and international conferences/workshops. He is currently an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEANING SYSTEMS.