This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2879626, IEEE Transactions on Circuits and Systems for Video Technology

Robust Discriminative Metric Learning for Image Representation

Zhengming Ding, *Member, IEEE*, Ming Shao, *Member IEEE*, Wonjun Hwang, *Member IEEE*, Sungjoo Suh, *Member IEEE*, Jae-Joon Han, *Member IEEE*, Changkyu Choi, *Member IEEE*, Yun Fu, *Senior Member IEEE*

***Abstract*—Metric learning has attracted significant attentions in the past decades, for the appealing advances in various real-world applications such as person re-identification and face recognition. Traditional supervised metric learning attempts to seek a discriminative metric, which could minimize the pairwise distance of within-class data samples, while maximizing the pairwise distance of data samples from various classes. However, it is still a challenge to build a robust and discriminative metric, especially for corrupted data in the real-world application. In this paper, we propose a Robust Discriminative Metric Learning algorithm (RDML) via fast low-rank representation and denoising strategy. To be specific, the metric learning problem is guided by a discriminative regularization by incorporating the pair-wise or class-wise information. Moreover, low-rank basis learning is jointly optimized with the metric to better uncover the global data structure and remove noise. Furthermore, fast low-rank representation is implemented to mitigate the computational burden and make sure the scalability on large-scale datasets. Finally, we evaluate our learned metric on several challenging tasks, e.g., face recognition/verification, object recognition, and image clustering. The experimental results verify the effectiveness of the proposed algorithm by comparing to many metric learning algorithms, even deep learning ones.**

***Index Terms*—Metric Learning, Fast Low-rank Representation, Denoising Strategy.**

I. Introduction

METRIC learning [1], [2] has been extensively discussed and well developed in computer vision and machine learning fields in the past decades. Existing metric learning models could be generally split into two main categories: unsupervised and supervised metric learning. Specifically, unsupervised metric aims to build a low-dimensional space to keep the geometrical structure within the data, whilst supervised one is developed to learn a distance metric, which maximizes the separability of data from various categories.

Z. M. Ding is with the Department of Computer, Information and Technology, Indiana University-Purdue University Indianapolis, 420 University Blvd Indianapolis, IN 46202, USA. E-mail: zd2@iu.edu

Ming Shao is with the Computer and Information Science, University of Massachusetts Dartmouth, MA, 02747 USA. E-mail: mshao@umassd.edu

Wonjun Hwang is with Dept. of Software and Computer Engineering, College of Information Technology, Ajou University, Korea. E-mail: wjhwang@ajou.ac.kr

Sungjoo Suh, Jae-Joon Han, Changkyu Choi are with Software Solution Lab., Samsung Advanced Institute of Technology, 130, Samsung-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, Korea. E-mail: {sungjoo.suh, jaejoon.han, changkyu_choi}@samsung.com.

Yun 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.

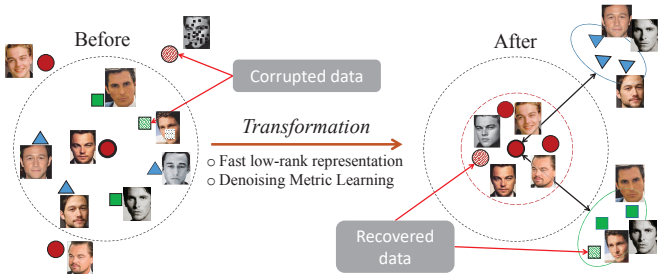


Fig. 1. Framework of our designed robust metric, where the data with the same shape denote the same identity. Originally, data points are mixed together. Then it can be observed that the scatter points from the within class are pulled compacter and points from between classes are pushed far away after metric learning. Our metric learning is robust to noisy samples due to the denoising strategy.

When the training data have labels, supervised metric learning algorithms are more powerful and suitable to build recognition models. Generally, metric learning could be converted to seek linear/non-linear mappings [3], [4], [5], [6], [7], [8], [9]. However, conventional metric learning usually fails to well handle the noisy data and meanwhile preserves the global data structure in real world. This is especially useful for recognition tasks, where classifier can be easily fooled by the corrupted feature. For example, lighting, shadows or occlusions on the face images could prevent face recognition, or messy background in object samples would hurt object recognition performance.

To build discriminative features from corrupted data, many robust feature extractors have been proposed, e.g., sparse representation [10] and low-rank representation [11]. Among them, low-rank representation (LRR) [11] is capable of recovering the global structure within the data by removing the noise samples. However, low-rank based algorithms suffer a heavy computational burden due to the trace-norm optimization, which requires a full SVD operation to solve the proximal operator of the trace norm in each iteration. Hence, it is not scalable to large-scale data analysis challenge. Most recently, many fast implementations of LRR have been proposed to make it scalable to larger benchmarks in reality. Divide-and-conquer strategy is widely explored in low-rank optimization to deal with large-scale issue [12], [13]. Furthermore, Xiao et al. reformulated the conventional LRR model to factorize data with a novel optimization problem [14].

Deep learning can build hierarchical structure to distill knowledge from the data and deal with the noise. Most recently, the idea of metric learning has been proposed to

This is the author's manuscript of the article published in final edited form as:
Ding, Z., Shao, M., Hwang, W., Suh, S., Han, J., Choi, C., & Fu, Y. (2018). Robust Discriminative Metric Learning for Image Representation. IEEE Transactions on Circuits and Systems for Video Technology, 1–1. https://doi.org/10.1109/TCSVT.2018.2879626

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2879626, IEEE Transactions on Circuits and Systems for Video Technology

build the deep structure with a metric loss at the top layer [15], [7], [6], [9], [16], [17], [18], [19], e.g., contrastive loss, and triplet loss. However, the downside is that deep metric learning usually requires large-scale data for training. For some tasks, e.g., person re-identification, large-scale datasets are not common, and therefore, shallow models are still of great use. Notably, shallow models over pre-trained deep features from other large vision datasets work fairly well in this case. On the other hand, some deep metric learning works only focus on fully-connected layer. This motivates us to explore the shallow metric learning together with other discriminant features including the deep ones.

A. Our Contribution

Low-rank based models have provided promising performance in different applications (e.g., subspace clustering, image classification and transfer learning) through uncovering the global structure within the data [11], [20], [21]. However, unlike traditional metric learning, they cannot take full advantages of global structure within the data to seek a discriminative metric. Secondly, conventional metric learning approaches are very sensitive when dealing with the corrupted data. Therefore, the obtained metric has weak generalization ability. To that end, we propose a Robust Discriminative Metric Learning (RDML) framework, which is insensitive to various sources of noises for discriminative metric learning, as Fig. 1 shows. The key idea behind our method is to jointly seek a robust and discriminative denoising metric in a fast fashion, meanwhile preserving more discriminative knowledge. Finally, we summarize our contributions in three folds:

- We design a robust discriminative metric learning framework by simultaneously uncovering the global structure within the data and constructing a compact clean basis. Specifically, low-rank model could help detect and rule out noises within the data under the learned distance metric, where all the features are reconstructed by a clean compact basis. In this way, we fulfill our denoising discriminative metric.
- A fast low-rank model is designed to make our algorithm scalable to large-scale data in real world. In this way, the time-consuming SVD operator in solving the trace norm would be relaxed to a fixed-rank matrix factorization problem. This is particularly useful in supervised learning, as we could approximately estimate the rank of feature matrix ahead of model training. The solutions would finally factorize the original matrix into the product of two low-rank matrices.
- Extensive experiments on various applications, e.g., face recognition, object recognition, image clustering and person re-identification, have been conducted to systemically evaluate our algorithm. Experimental results have validated the superiority of our metric.

The left sections of this paper are organized as follows. In section II, we briefly discuss related works. Then, we provide the proposed robust discriminative metric learning, optimized solutions and complexity/convergence analysis in Section III. Experimental evaluations are reported in Section IV, following by our conclusion in Section V.

II. Related Work

In this part, we briefly introduce two lines of works related to our algorithm: metric learning and fast low-rank representation.

A. Metric Learning

Metric learning [2], [1] becomes appealing in the fields of computer vision and machine learning for decades. It is designed to build a discriminative metric to boost the performance of learning algorithms. Lots of metric learning models attempt to build a Mahalanobis-like distance metric $\mathcal{M}$ ($\mathcal{M}$is positive semi-definite), which could be further decomposed into two smaller matrices, i.e., $\mathcal{M} = PP^\top$.

Following this, there are many metric learning algorithms proposed recently in different applications, e.g., kinship verification [22], [6], [23], co-saliency detection [24], image-set classification [25], face verification [9], [16], person re-identification [26], and multi-view learning [27]. Specifically, Xing et al. designed to reduce the distances of similar data pairs whilst maximizing those of different data pairs [28]. Later on, Ding et al. proposed a transfer metric to improve the recognition of unlabeled target data with the help of labeled source data lying different distributions [29]. Notably, a few works recently incorporated a regularizer (e.g., group sparsity or low-rankness) to guide the metric learning, and therefore, most of non-informative features could be removed [30], [31]. For example, Liu et al. presented a rank-constrained metric framework by using a bilinear matrix factorization, which is applicable to high-dimensional data domains [32].

Most recently, deep metric learning has been paid great attentions. The primary idea of deep metric learning is designing different loss functions by exploring the positive and negative information at certain layers of the deep architecture [25], [33], [9], [16], [34], [6], [9], [33]. Along this line, Bromley et al. explored the idea of deep learning and proposed a Siamese metric model for signature verification [35]. Schroff et al. developed FaceNet by using triplet embedding to learn low-dimensional representations for face recognition [15]. Cheng et al. proposed a novel duplex metric learning with two progressive metrics, which was not only explored to seek effective features but also well explored to build a generic classifier [25]. Song et al. designed deep metric learning framework based on structured loss functions, and therefore, such methods could capture the global structure within the data [36], [17]. Later, Duan et al. presented to seek multiple fine-grained localized metrics based on K local subspaces [9], which targets at capturing the global structure within the data and build various metric models per local patch.

B. Fast Low-rank Representation

Low-rank representation (LRR) [11] attempts to capture the global structure within the data. Specifically, LRR aims to learn a new representation amongst all samples constrained by

1051-8215 (c) 2018 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2879626, IEEE Transactions on Circuits and Systems for Video Technology

a very low rank. As a result, LRR generates a block structural representation coefficient matrix, which discovers the multiple subspace structures within the samples. However, low-rank based algorithms suffer a heavy computational burden due to the trace-norm optimization, which requires a full SVD operation to address the proximal operator of the trace norm per iteration. That is, for a matrix $D \in \mathbb{R}^{d\times n}$, the time complexity of SVD would be $O(\min(n^2d, d^2n))$. Hence, the repeated full SVD operations in optimization are computationally expensive, which makes low-rank methods scale poorly in real-world large-scale applications.

There are two strategies to handle the heavy computational cost of trace-norm for large-scale data: one is devide-and-conquer strategy; the other is replacing trace-norm with other constraint to speed up the solutions. Along the first line, Xiao et al. fasten the low-rank learning by proposing a novel optimization objective with factorized data [14]. Divide-and-Conquer strategy is also proposed to decompose large-scale data into small ones [12], [13]. Along the second line, Kim et al. developed a low-rank matrix factorization approach with an elastic-net regularization [37]. For our proposed work, we explore the idea of bi-linear matrix factorization to speed up the low-rank metric learning.

Differently, we design a robust metric learning algorithm, which attempts to jointly capture the global structure within data via low-rank recovery and discriminative local knowledge through pair-wise positive/negative side information. This work is the journal extension of our prior conference version [38], whose key differences lie in two folds. First, the low-rank model could help detect and rule out noises within the data under the learned distance metric, where all the features are reconstructed by a clean compact basis. In this way, we fulfill our denoising discriminative metric. Second, a fast low-rank model is designed to make our algorithm scalable to large-scale data in real world. In this way, the time-consuming SVD operator in solving the trace norm would be relaxed to a fixed-rank matrix factorization problem. This is particularly useful in supervised learning, as we could approximately estimate the rank of feature matrix ahead of model training. The solutions would finally factorize the original matrix into the product of two low-rank matrices.

III. The Proposed Algorithm

In this part, we develop a robust discriminative metric through fast low-rank representation and denoising strategy.

A. Discriminative Metric Learning

Given a training dataset with n labeled samples $\{X, Y\} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, in which $x_i \in \mathbb{R}^d$ is the i-th data point with its label y_i. Conventional supervised metric models [28], [4], [39] were designed to seek a distance metric $\mathcal{M}$ in order to keep the intra-class data compact while making inter-class data discriminative enough. Generally, the objective loss function can be formulated in the following:

$$\begin{aligned} &\min_{\mathcal{M}} \sum_{(x_i,x_j)\in\mathcal{S}} \|x_i - x_j\|^2_{\mathcal{M}} \\ &\text{s.t.} \quad \mathcal{M} \in \mathbb{S}^d_+, \sum_{(x_i,x_j)\in\mathcal{D}} \|x_i - x_j\|^2_{\mathcal{M}} \geq 1, \end{aligned} \tag{1}$$

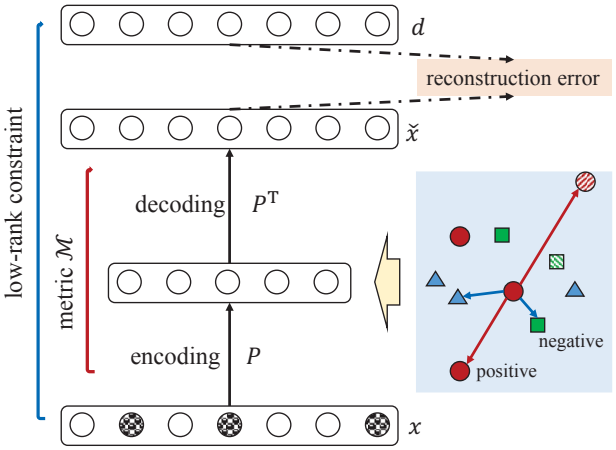


Fig. 2. Illustration of robust discriminative metric learning, in which each sample x is reconstructed using the metric $\mathcal{M}$ to $\check{x} = \mathcal{M}x$. We notice that the metric serves as encoding (P) and decoding ($P^\top$) in one step. The loss function $d - \check{x}$ attempts to reduce the reconstruct error and make our metric denoise the data to a clean low-rank basis $D = \{d_i\}_1^n$. Simultaneously, discriminative term enforces pair-wise constraint into the hidden layer.

in which $\|x_i - x_j\|_{\mathcal{M}}$ denotes $\sqrt{(x_i - x_j)^\top \mathcal{M}(x_i - x_j)}$. $\mathcal{S}$ and $\mathcal{D}$ mean the within-class pair sets and between-class pair sets, respectively.

Therefore, Eq. (1) could be transformed to:

$$\begin{aligned} \mathcal{M} = &\arg\min_{\mathcal{M}\in\mathbb{S}^d_+} \frac{\mathrm{tr}(X\mathcal{L}_{\mathcal{S}}X^\top\mathcal{M})}{\mathrm{tr}(X\mathcal{L}_{\mathcal{D}}X^\top\mathcal{M})} \\ = &\arg\min_{\mathcal{M}\in\mathbb{S}^d_+} \mathrm{tr}(X\mathcal{L}_{\mathcal{S}}X^\top(X\mathcal{L}_{\mathcal{D}}X^\top)^\dagger\mathcal{M}) \\ = &\arg\min_{\mathcal{M}\in\mathbb{S}^d_+} \mathrm{tr}(\mathcal{A}\mathcal{M}) \end{aligned} \tag{2}$$

in which $\mathrm{tr}(\cdot)$ represents the trace of a matrix. $\mathcal{L}_{\mathcal{S}}$ and $\mathcal{L}_{\mathcal{D}}$ denote two Laplacian matrices of $\mathcal{S}$ and $\mathcal{D}$. $\dagger$ is pseudo-inverse operator of a matrix. We transform the trace ratio problem into a ratio-trace problem and $\mathcal{A} = X\mathcal{L}_{\mathcal{S}}X^\top(X\mathcal{L}_{\mathcal{D}}X^\top)^\dagger$.

B. Denoising Metric Learning

In general, $\mathcal{M} \in \mathbb{R}^{d\times d}$ could be factorized into $\mathcal{M} = PP^\top$, in which $P \in \mathbb{R}^{d\times p}$ and $p \leq d$ is the rank of metric matrix. Therefore, we could further reformulate $\|x_i - x_j\|_{\mathcal{M}}$ as $\|x_i - x_j\|_{\mathcal{M}} = \|P^\top(x_i - x_j)\|_2$. Following the idea of principle component analysis (PCA), we could also formulate the metric reconstruction into a PCA-like fashion as follows:

$$\Omega_d = \|X - PP^\top X\|^2_{\mathrm{F}} = \|X - \mathcal{M}X\|^2_{\mathrm{F}}, \tag{3}$$

where $\|\cdot\|_{\mathrm{F}}$ means Frobenius norm of the matrix.

Inspired by denoising fashion, e.g., Denoising Auto-Encoder (DAE) [40], [41] or its marginalized variants (mDAE) [42], [43], [44], we attempt to generate a noise-free metric $\mathcal{M}$. Hence, we proposed a denoising metric learning framework by jointly seeking a low-rank basis as the target to constrain the reconstructed data as follows:

$$\Omega_d = \mathrm{rank}(D) + \lambda\|\mathcal{M}X - D\|^2_{\mathrm{F}}, \tag{4}$$

where we could notice that our metric would reconstruct the real-world data to be as similar as possible to a low-rank basis.

1051-8215 (c) 2018 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2879626, IEEE Transactions on Circuits and Systems for Video Technology

$\mathrm{rank}(\cdot)$ is the rank operator of a matrix and λ is the balanced parameter.

Remark: The proposed denoising metric is much different from mDAE [42] and DAE [40], since our metric could be converted to seek a denoising transformation P ($\mathcal{M} = PP^\top$). On one hand, we can illustrate our denoising metric learning as an auto-encoder format (Fig. 2). $P^\top X$ could be viewed as the hidden-layer representation (*encoding*), and $\mathcal{M}X = PP^\top X$ could be viewed as the reconstruction of original input X (*decoding*). On the other hand, mDAE only seeks a linear rotation to transform the intentionally corrupted data to its original one. Hence, such rotation cannot capture the structure information in the feature space. Furthermore, we assume the real-world data already contain different kinds of noise, e.g., illumination and corruption, which always happens in image representation scenario.

C. Overall Objective Function

To sum up, we formulate our robust discriminative metric objective function by integrating denoising metric and discriminative metric together as follows:

$$\min_{\mathcal{M}\in\mathbb{S}_+^d, D} \mathrm{rank}(D) + \lambda\|\mathcal{M}X - D\|_{\mathrm{F}}^2 + \alpha\mathrm{tr}(\mathcal{A}\mathcal{M}), \tag{5}$$

in which α is the trade-off parameter. With (5), we are able to build a robust and discriminative metric for better image representation, which not only captures the intrinsic structure knowledge in sample space, but also builds more robust knowledge in feature space.

However, rank minimization in Eq. (5) is an NP-hard issue. Along the literature, there exist many solutions to fight off the rank minimization challenge [11]. To this end, we convert Eq. (5) into the equivalent problem:

$$\min_{\mathcal{M}\in\mathbb{S}_+^d, D} \|D\|_* + \lambda\|\mathcal{M}X - D\|_{\mathrm{F}}^2 + \alpha\mathrm{tr}(\mathcal{A}\mathcal{M}) \tag{6}$$

in which $\|\cdot\|_*$ denotes the nuclear norm of a matrix. In general, Eq. (6) suffers a heavy computational burden when dealing with large-scale data [11], since SVD is employed to address the nuclear-norm based objective function at each iteration. Then, low-rank minimization problem could be transformed to a fixed rank problem and we could transform Eq. (6) into an equivalent formulation:

$$\begin{aligned}\min_{\mathcal{M}\in\mathbb{S}_+^d, D, U, V} &\frac{1}{2}(\|U\|_{\mathrm{F}}^2 + \|V\|_{\mathrm{F}}^2) + \lambda\|\mathcal{M}X - D\|_F^2 \\ &+\alpha\mathrm{tr}(\mathcal{A}\mathcal{M}), \quad \text{s.t.} \quad D = UV.\end{aligned} \tag{7}$$

To sum up, the newly proposed model above fulfills our purpose of fast low-rank representation in denoising data reconstruction scheme for robust metric learning. Specifically the time consuming nuclear norm $\|D\|_*$ has been replaced by the sum of two Frobenius norms: $\|U\|_{\mathrm{F}}^2 + \|V\|_{\mathrm{F}}^2$, with an additional constraint $D = UV$. The internal dimension r gives rise to a fixed-rank decomposition scheme, and therefore avoids the time-consuming trace-norm.

Remark: Our goal is to seek a denoising metric by borrowing the idea of denosing auto-encoder. In real-world applications, data are usually contaminated and finding discriminative features is challenging. This is true even for deep features. In our framework, to simulate the denoising process, the learned metric should be able to transform the original feature to a low-rank basis. In supervised learning cases, we know the class number for the data, which could be set as the rank of D, while for unsupervised learning, we would have a preferred rank number.

D. Solution Optimization

To address the minimization problem of Eq. (7), we first transform it to the equivalent optimization issue by applying the augmented Lagrangian function:

$$\begin{aligned}\mathcal{L} = \quad &\frac{1}{2}(\|U\|_{\mathrm{F}}^2 + \|V\|_{\mathrm{F}}^2) + \lambda\|\mathcal{M}X - D\|_F^2 + \alpha\mathrm{tr}(\mathcal{A}\mathcal{M}) \\ &+\langle\Upsilon, Z - UV\rangle + \frac{\mu}{2}\|D - UV\|_{\mathrm{F}}^2,\end{aligned} \tag{8}$$

in which Υ is the Lagrange multiplier while μ is a small positive penalty parameter. Moreover, we obtain the optimization result using an iterative strategy, since we cannot jointly optimize all the variable together. Before that, we convert the optimization issue of Eq. (8) to two sub-problems: the first one is to optimize $\mathcal{M}$ by treating D, U, V as constant; the second one is updating the D, U, V by fixing the metric as constant.

Learning Robust Representation: First of all, we fix metric $\mathcal{M}$ to optimize the low-rank basis variables D, U, V in a leave-one-out fashion. Let's denote the variables at t-th iteration as D_t, U_t, V_t. Hence, the updating to each variable at (t+1)-th iteration could be obtained as:

Updating D:

$$D_{t+1} = \arg\min_D \lambda\|\mathcal{M}_t X - D\|_{\mathrm{F}}^2 + \frac{\mu}{2}\|D - U_t V_t + \frac{\Upsilon_t}{\mu}\|_{\mathrm{F}}^2, \tag{9}$$

which has a closed-form solution as:

$$D_{t+1} = (2\lambda\mathcal{M}_t X + \mu U_t V_t - \Upsilon_t)/(2\lambda + \mu). \tag{10}$$

Updating U:

$$\begin{aligned}U_{t+1} = \arg\min_U &\frac{1}{2}\|U\|_{\mathrm{F}}^2 + \langle\Upsilon_t, D_{t+1} - UV_t\rangle \\ &+\frac{\mu}{2}\|D_{t+1} - UV_t\|_{\mathrm{F}}^2,\end{aligned} \tag{11}$$

which has a closed-form solution as:

$$U_{t+1} = (\mu D_{t+1}V_t^\top + \Upsilon_t V_t^\top)(\mathrm{I}_r + \mu V_t^\top V_t)^{-1}. \tag{12}$$

Updating V:

$$\begin{aligned}V_{t+1} = \arg\min_V &\frac{1}{2}\|V\|_{\mathrm{F}}^2 + \langle\Upsilon_t, D_{t+1} - U_{t+1}V\rangle \\ &+\frac{\mu}{2}\|D_{t+1} - U_{t+1}V\|_{\mathrm{F}}^2,\end{aligned} \tag{13}$$

which has a closed-form solution as:

$$V_{t+1} = (\mathrm{I}_r + \mu U_{t+1}^\top U_{t+1})^{-1}(\mu U_{t+1}^\top D_{t+1} - U_{t+1}^\top\Upsilon_t). \tag{14}$$

Robust Discriminative Metric Learning: With positive semi-definite constraint, we are not easy to directly update the metric $\mathcal{M}$ with D fixed. Thus, we define $h(\mathcal{M}) = \lambda\|\mathcal{M}X -$

1051-8215 (c) 2018 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2879626, IEEE Transactions on Circuits and Systems for Video Technology

Algorithm 1 Solving Eq. (8)

Input: X, λ, α

Initialize: $D_0 = U_0 = V_0 = \Upsilon_0 = 0, \mu_0 = 10^{-6}$
$\rho = 1.1, \max_\mu = 10^6, \epsilon = 10^{-6}, \eta = 10^{-2}$.

while not converged **do**

1. Update D_{t+1} with others fixed through Eq. (10).
2. Update U_{t+1} with others fixed through Eq. (12).
3. Update V_{t+1} with others fixed through Eq. (14).
4. Update $\mathcal{M}_{t+1}$ with others fixed through Eq. (15).
5. Update Υ_{t+1} through $\Upsilon_{t+1} = \Upsilon_t + \mu_t(\mathcal{M}_t X - D_{t+1})$;
6. Update the μ_{t+1} through $\mu_{t+1} = \min(\rho\mu_t, \max_\mu)$
7. Check the convergence $\|D_{t+1} - U_{t+1}V_{t+1}\|_\infty < \epsilon$.
8. $t = t + 1$.

end while

output: $\mathcal{M}, D, U, V$.

$D\|_F^2 + \alpha \mathrm{tr}(\mathcal{A}\mathcal{M})$. Next, we explore a linear approximation to $h(\mathcal{M})$ to address the optimization by following [32]. Define $\mathcal{M}_t$ is the result at t-th iteration, then $\mathcal{M}_{t+1}$ is achieved at (t+1)-th iteration as:

Updating $\mathcal{M}$:

$$\begin{aligned}\mathcal{M}_{t+1} &= \arg\min_{\mathcal{M}\in\mathbb{S}_+^d} h(\mathcal{M}) \\ &= \arg\min_{\mathcal{M}\in\mathbb{S}_+^d} \tfrac{1}{2\eta}\|\mathcal{M} - \mathcal{M}_t\|_F^2 + h(\mathcal{M}_t) \\ &\quad + \langle \nabla_{\mathcal{M}} h(\mathcal{M})|_{\mathcal{M}=\mathcal{M}_t}, \mathcal{M} - \mathcal{M}_t\rangle \\ &= \arg\min_{\mathcal{M}\in\mathbb{S}_+^d} \tfrac{1}{2\eta}\|\mathcal{M} - (\mathcal{M}_t - \eta\mathcal{H}_t)\|_{\mathrm{F}}^2 \\ &= \mathcal{P}_{\mathbb{S}_+^d}(\mathcal{M}_t - \eta\mathcal{H}_t),\end{aligned} \tag{15}$$

in which $\mathcal{H}_t = \nabla_{\mathcal{M}} h(\mathcal{M})|_{\mathcal{M}=\mathcal{M}_t} = \lambda(2\mathcal{M}_t X X^\top - D_t X^\top - D_t^\top X) + \alpha\mathcal{A}$ and $\eta > 0$ denotes the step size, which is set as 0.01 in our experiments. Moreover, $\mathcal{P}_{\mathbb{S}_+^d}(\cdot)$ means the projection operator to $\mathbb{S}_+^d$. That is to say, $\mathcal{P}_{\mathbb{S}_+^d}(\mathcal{K})$ can be formulated as $\sum_{i=1}^d [\gamma_i]_+ k_i k_i^\top$ for a symmetric matrix $\mathcal{K} \in \mathbb{R}^{d\times d}$, where $\{k_i, \gamma_i\}|_{i=1}^d$ mean the eigenvector-eigenvalue pairs.

Algorithm 1 lists the detailed steps of the optimization to our model. To be specific, the parameters μ_0, ρ, ϵ, η and $\max_\mu$ are set empirically, while other parameters λ, and α need to be tuned using cross-validation during the experiments. To achieve a fast convergence in optimization, we initialize $\mathcal{M}$ using Eq. (2). For other variables, e.g., D, U, V, Υ, we initialize with zero matrices for simplicity. In experiments, we observed that their initial values do not affect the convergence much.

E. Complexity Analysis

In this part, we would present time complexity analysis of our proposed method.

The major time-consuming components are matrix multiplication and inverse in Step 1, 2, 3, and SVD-projection in Step 4. Specifically, step 1 would take about $\mathcal{O}(d^2 n)$ for $D \in \mathbb{R}^{d\times n}$ (Generally, the feature dimensionality d is smaller than the sample size n). Step 2 takes about $\mathcal{O}(d^2 r)$ while Step 3 would take about $\mathcal{O}(n^2 r)$. Step 4 takes $\mathcal{O}(d^3)$ when mapping $\mathcal{M}$ onto $\mathbb{S}_+^d$ via SVD-based projection. When d is large, it is very time consuming to update $\mathcal{M}$. Fortunately, we could explore recent advances in efficient metric learning, e.g., incremental SVD [45] to save the optimization time. Let $\mathcal{M}$ be a rank-p matrix. $\mathcal{H}_t$ can be decomposed to $A_t A_t^\top$, where $A_t \in \mathbb{R}^{d\times q}$. Then, the Eigen-decomposition of $\mathcal{M}_t - \eta\mathcal{H}_t$ can be calculated in $\mathcal{O}(d(p+q)^2 + (p+q)^3)$, which is almost linear to d.

Fig. 3. Samples of COIL-100 dataset, where the first row is the original images whilst the second row is the 10% corrupted images.

TABLE I
RECOGNITION RESULTS (%) OF 7 ALGORITHMS ON COIL-100 IN DIFFERENT EVALUATION SIZES, FROM 20 TO 100 OBJECTS.

Methods	20	40	60	80	100	Average
DML-eig [46]	86.62	82.98	79.25	78.93	76.37	80.83
ITML [47]	87.12	83.32	80.18	79.65	78.45	81.74
SILD [4]	85.54	82.88	78.74	76.34	72.32	79.16
Sub-SML [39]	90.38	89.49	84.68	84.12	82.92	86.32
SRRS [20]	92.03	92.51	90.82	88.75	85.12	89.85
DLML [38]	91.83	91.91	90.95	88.98	85.83	89.90
Ours	**93.14**	**93.16**	**92.16**	**90.49**	**87.48**	**91.29**

IV. EXPERIMENTAL RESULTS

In this part, we evaluate our presented algorithm from different image representation tasks by comparing with other state-of-the-art metrics. In this end, we analyze some properties including parameter influence, optimization stability, and model convergence.

A. Object Classification

The COIL-100 dataset[1] is composed of 100 different objects with different illuminations under 72 different views, which are captured 5 degree apart. We first convert the images to gray-scale and resize them to 32×32. We also evaluate the robustness of different methods to noisy data, where we add 10% random corruption to the original images in pixel level by replacing original values with 0 (Fig. 3). The pixel value is directly adopted as the feature input. We randomly choose 10 samples per object for training, while the rest for testing. We do 20 trials to calculate the average recognition rates. Additionally, we conduct scalability evaluations, by tuning the size of objects.

In the experiments, we compare our proposed model with DML-eig [46], ITML [47], SILD [4], Sub-SML [39], SRRS [20], DLML [38]. The comparison results are provided in TABLE I for original data and Table II for 10% corrupted data. We observe that our algorithm outperforms the competitive methods in the original data. Furthermore, in the corrupted data situations, our algorithm could outperform other algorithms with a large margin, which further demonstrates the superiority of our proposed algorithm.

[1] http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php

1051-8215 (c) 2018 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2879626, IEEE Transactions on Circuits and Systems for Video Technology

TABLE II
RECOGNITION RESULTS (%) OF 7 ALGORITHMS ON 10% CORRUPTED COIL-100 IN DIFFERENT EVALUATION SIZES, FROM 20 TO 100 OBJECTS.

Methods	20	40	60	80	100	Average
DML-eig [46]	58.76	48.89	43.42	39.13	37.79	45.60
ITML [47]	66.35	65.78	62.92	59.98	54.24	68.85
SILD [4]	48.34	46.24	44.68	40.68	38.12	43.61
Sub-SML [39]	73.43	72.23	70.56	68.20	65.18	69.92
SRRS [20]	86.45	82.03	82.05	79.83	74.95	81.06
DLML [38]	82.52	79.54	71.15	54.68	39.69	65.58
Ours	**87.42**	**87.12**	**86.12**	**84.84**	**83.26**	**85.75**

TABLE III
RECOGNITION RESULTS (%) OF 7 ALGORITHMS ON CMU-PIE IN DIFFERENT TRAINING SIZES, FROM 10 TO 60 SAMPLES PER SUBJECT.

Methods	10	20	30	40	50	60
DML-eig [46]	58.24	68.65	74.28	80.64	84.72	88.24
ITML [47]	65.60	70.46	84.84	87.72	90.74	91.92
SILD [48]	69.62	79.54	88.12	91.24	92.92	93.76
Sub-SML [39]	69.70	79.62	89.72	92.13	93.04	93.82
SRRS [20]	70.38	80.17	89.24	92.38	93.86	94.78
DLML [38]	71.15	82.52	90.26	92.85	94.12	94.93
Ours	**75.72**	**85.24**	**92.16**	**95.43**	**96.26**	**96.84**

Besides, we could observe that our conference version DLML [38] could achieve competitive performance in the original data; however, its performance degrades significantly in the corrupted data. That demonstrates our statement that the pre-learned low-dimensional features would introduce noises into the low-rank constraint and in turn contaminate the metric. Our proposed model could well handle this problem, since we optimize a clean and compact basis in our algorithm instead of low-dimensional pre-learned features. In this way, our current version can still achieve very good results in the corrupted data and beat competitive methods.

B. Face Recognition

1) CMU-PIE Face Dataset: is composed of 68 subjects and each individual in CMU-PIE has 21 kinds of illumination variations with environmental illuminations changing. We pick up five pose images (C05, C07, C09, C27, C29) with large variance for each subject for evaluations. We cropped the images and resized into 32×32, and then we used the raw feature as the input. In this dataset, we compare with DML-eig [46], ITML [47], SILD [4], Sub-SML [39], SRRS [20], DLML [38]. Faces under five poses are combined together first, and then we randomly select $l(l = 10, 20, 30, 40, 50, 60)$ samples per subject for training while the rest samples are adopted as the testing data. We do 50 random trials. The nearest neighbor classifier (NNC) is adopted to evaluate all the algorithms. Table III reports the recognition performance of 7 different algorithms.

From Table III, we observe that our model can consistently perform better than other baselines. For face recognition, pose variations can be treated one kind of real-world noises. Our experimental results show that our model could handle such pose variation well.

Fig. 4. Samples of face images from LFW, where a column denotes one pair.

TABLE IV
COMPARISON RESULTS (%) ON LFW DATASET IN THE IMAGE RESTRICTED SETTING WITH LBP-LD AND LBP-HD FEATURES.

Methods	LBP-LD	LBP-HD
Xing [28]	74.64±0.45	80.82±0.38
DML-eig [46]	82.28±0.41	87.94±0.55
SILD [4]	80.07±1.35	86.04±1.45
ITML [47]	79.98±0.39	85.94±0.25
LDML [49]	80.65±0.47	86.64±0.75
KISSME [50]	83.37±0.52	88.92±0.61
Sub-SML [39]	85.47±0.55	91.02±0.62
DDML [6]	-	92.62±0.35
SvDML [27]	85.70±0.41	91.22±0.45
DLML [38]	85.35±0.51	91.15±0.59
Ours	86.56±0.48	92.38±0.54

C. Face Verification

In this part, we aim to testify our robust discriminative metric on large-scale data. As we know, Labeled Faces in the Wild (LFW) is one of the most challenging real-world facial datasets including over 13000 face images from 5749 individuals (Fig. 4). LFW is collected online and the face samples have large variations in expression, illumination, age and other factors [51]. We adopt the standard protocol using "view 2" that contains 3000 positive pairs and 3000 negative pairs. The samples will be further averagely divided into 10 folds, and 9 folds would be used to train the model while the left fold is used to evaluate. We adopt the restricted scenario, where only similar/dissimilar pairs are accessible while the identities of samples are unavailable.

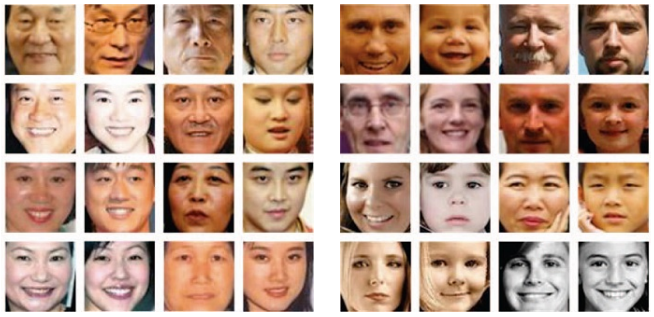
In these experiments, we compare our algorithm with several shallow metric learning approaches, e.g., Xing [28], SILD [4], DML-eig [46], ITML [47], LDML [49], KISSME [50], DLML [38] and two deep metric methods, i.e., DDML [6] and SvDML [27]. Specifically, we adopt the image restricted setting of LFW and use two different types of LBP features [52]: one is low-dimensional LBP (LBP-LD) with 5900 dimensions and the other is high-dimensional LBP (LBP-HD) with 127440 dimensions[2].

The results are reported in Table IV, where we could notice that our method outperforms other competitive methods in both LBP-LD and LBP-HD features. Compared with DDML, we can achieve comparable performance over LBP-HD features. Since most of the individuals have a small number of samples, it is hard for low-rank representation to uncover the global structures. Therefore, our algorithm only slightly improves the verification performance compared to other competitors.

[2]http://home.ustc.edu.cn/∼chendong/

1051-8215 (c) 2018 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2879626, IEEE Transactions on Circuits and Systems for Video Technology



(a) KinFaceW I (b) KinFaceW II

Fig. 5. Samples from KinFaceW-I (a) and KinFaceW-II (b) datasets. Four relationships, i.e., F-S, F-D, M-S and M-D are provided from top to bottom.

TABLE V
CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS ON DIFFERENT SUBSETS OF THE KINFACEW-I DATA SET

Features	Methods	F-S	F-D	M-S	M-D	Avg.
LBP	CSML [53]	63.7	61.2	55.4	62.4	60.7
	LMNN [54]	62.7	63.2	57.4	63.4	61.7
	NRML [22]	64.7	65.2	59.4	65.4	63.7
	DSML [6]	70.8	67.2	72.5	74.0	71.1
	DDML [6]	78.4	71.9	75.8	75.8	75.5
	Ours	76.2	70.4	75.8	76.2	74.7
DSIFT	CSML [53]	66.5	60.0	60.0	56.4	59.8
	LMNN [54]	69.5	63.0	63.0	59.4	62.8
	NRML [22]	70.5	64.0	64.0	60.4	63.8
	DSML [6]	70.0	70.9	73.9	78.1	73.2
	DDML [6]	78.0	75.9	76.5	83.3	78.4
	Ours	76.2	74.2	76.9	82.2	77.3

D. Kinship Verification

KinFaceW-I/II[3][22] are two widely-used kinship datasets collected publicly (Fig. 5). For each image per dataset, a corresponding parent or child image is also provided. In total, there exist four different kin relationships, i.e., mother-son (M-S), father-son (F-S), mother-daughter (M-D) and father-daughter (F-D). KinFaceW-I contains 156, 134, 116, and 127 pairs of kinship samples for four relationships. While KinFaceW-II provides 250 pairs of kinship samples. For two benchmarks, any aligned 64×64 sample is used for feature extraction and two kinds of features are extracted, i.e., LBP feature vector with 3776 dims, and Dense SIFT (DSIFT) feature vector with 6272 dims. Similar to [22], we explore 5-fold cross validation using the image restricted setting. The average verification results are reported in Tables V and VI.

From the results, we observe that our model achieves better performance over several shallow metric learning approaches, and obtain very close verification results to deep metric learning algorithm, i.e., DDML.

E. Image Clustering

The CUB-200-2011 dataset[4] consists of 200 bird categories with 11,788 image samples in total. We adopt the first 100 birds for training (5,864 samples) while the remaining birds to do evaluation (5,924 samples). As we know, birds are notoriously challenging to recognize, as the intra-class variation is quite siginificant when compared to the inter-class variation. Specifically, we compare with several deep metric learning methods, e.g., FaceNet [15], Lifted Struct [18], Npairs [36] and Proxy NCA [19]. For fair comparisons, we adopt pool5 activation features with GoogLeNet [55] pre-trained on ImageNet as the input features for metric learning.

Table VII lists the performance of the quantitative comparison between our approach and other algorithms. We utilize NMI score to measure the clustering result, also Recall@K metric. From the results, we notice that our RDML can achieve better results over the state of the art in most cases. That is to say, based on deep features, shallow structure metric learning could further improve the performance. More importantly, we could adopt our proposed metric learning as the loss function at the top layer of deep structure, thus, we could integrate our metric learning into deep architecture to formulate a unified framework.

TABLE VI
CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS ON DIFFERENT SUBSETS OF THE KINFACEW-II DATA SET

Features	Methods	F-S	F-D	M-S	M-D	Avg.
LBP	CSML [53]	66.0	65.5	64.8	65.0	65.3
	LMNN [54]	68.0	68.5	68.8	67.0	68.2
	NRML [22]	69.0	69.5	69.8	69.0	69.5
	DSML [6]	72.4	64.3	67.6	71.2	68.9
	DDML [6]	81.4	73.8	78.1	77.2	77.6
	Ours	77.6	69.2	75.4	76.3	74.6
DSIFT	CSML [53]	62.0	58.9	56.8	57.4	58.8
	LMNN [54]	65.0	57.9	58.8	59.4	60.4
	NRML [22]	68.9	60.9	60.8	61.4	62.8
	DSML [6]	75.6	63.8	70.0	74.7	71.0
	DDML [6]	82.5	75.7	79.1	79.2	79.1
	Ours	79.3	72.3	77.4	78.3	76.8

TABLE VII
RETRIEVAL AND CLUSTERING PERFORMANCE ON THE CUB200 DATASET.

Methods	R@1	R@2	R@4	R@8	NMI
FaceNet [15]	42.59	55.03	66.44	77.23	55.38
Lifted Struct [18]	43.57	56.55	68.59	79.63	56.50
Npairs [36]	45.37	58.41	69.51	79.49	57.24
Facility Location [17]	48.18	61.44	71.83	81.92	59.23
Proxy NCA [19]	49.21	61.90	67.90	72.40	59.53
GoogLeNet+Ours	49.68	62.46	73.25	80.32	60.78

F. Person Re-identification

Viewpoint Invariant Pedestrian Recognition (VIPer) [56] includes 632 sample pairs of pedestrian collected from two camera views in the wild. The benchmark shows large viewpoint variations and relatively low resolution, and therefore it is much challenging for person re-identification. Following the standard single-shot protocol, i.e., one sample per person per view, the dataset could be randomly split into training and test sets, each with 316 image pairs. The performances for all evaluations were achieved by averaging over 10 splits. We adopted the local maximal occurrence (LOMO) representation to illustrate each sample.

We mainly compare with metric learning based methods, ITML [47], LFDA [57], PCCA [58] and XQDA [59]. Tables

[3]http://www.kinfacew.com/datasets.html
[4]http://www.vision.caltech.edu/visipedia/CUB-200-2011.html

1051-8215 (c) 2018 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2879626, IEEE Transactions on Circuits and Systems for Video Technology

Fig. 6. Some examples from the viewpoint invariant pedestrian recognition (VIPeR) dataset. Each column is one of 632 same-person example pairs.

TABLE VIII
TOP RANKED MATCHING RATE (%) ON THE VIPER DATASET IN SAME RESOLUTION.

Methods	$k=1$	$k=5$	$k=10$	$k=20$
LFDA [57]	32.30	65.80	79.70	90.90
ITML [47]	24.64	35.93	48.76	60.08
PCCA [58]	19.27	48.89	64.91	80.28
XQDA [59]	40.00	68.13	80.51	91.08
Ours	**42.25**	**70.58**	**87.32**	**94.20**

TABLE IX
TOP RANKED MATCHING RATE (%) ON THE VIPER DATASET IN DIFFERENT RESOLUTIONS.

Methods	$k=1$	$k=5$	$k=10$	$k=20$
LFDA [57]	9.57	28.80	43.33	60.94
ITML [47]	8.92	26.32	34.26	46.26
PCCA [58]	8.55	27.39	41.17	58.68
XQDA [59]	23.26	53.86	70.03	84.68
Ours	**27.12**	**59.34**	**74.26**	**87.42**

VIII and IX show the top 1, 5, 10, and 20 matching rates of our proposed approach, and other baselines on the VIPeR benchmark. Here we do two experiments by using the original images (Table VIII) and down-sampling images from one view to the rate 1/8 (Table IX). The second experiment further differentiates two views of images. We can see that our proposed model outperforms all the other algorithms for all the ranks, most cases with margins.

G. Property Analysis

In this part, we mainly evaluate some properties of our proposed model, e.g., convergence, parameter, robustness to noise and computational cost.

1) Convergence Analysis: First, we empirically analyze the convergence of our proposed approach during optimization. To be specific, we evaluate on CMU-PIE face dataset with 40 training images per individual. Fig. 7 reports he convergence curve along with the recognition results in different iterations.

From Figure 7, we notice our proposed model converges generally after about 100 iterations. Also, the recognition performance reaches the peak after about 200 iterations while remains there afterwards. Note that different scales of evaluation datasets need various iterations to converge. Usually, large-scale datasets need more iterations compared to the small ones.

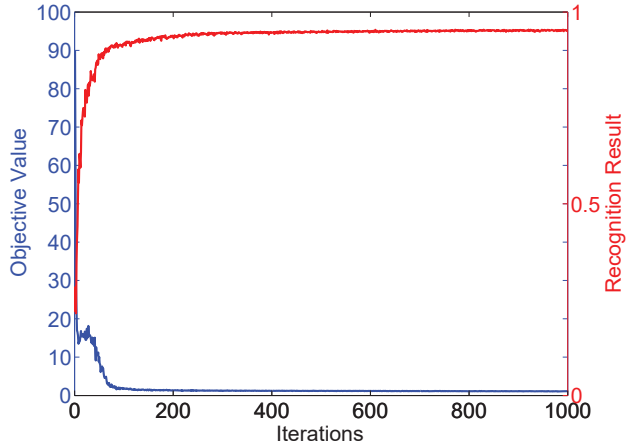


Fig. 7. Convergence curve (blue) and recognition curve (red) of our approach on CMU-PIE face dataset with 40 training samples per subject. We report the results in 1000 iterations with $\lambda = 10^{-2}, \alpha = 10^{2}$.

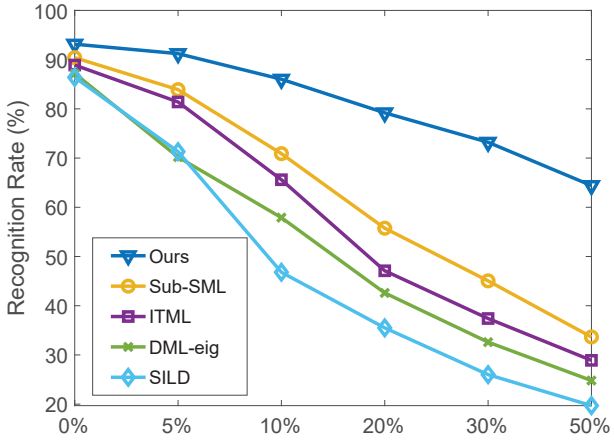


Fig. 8. Recognition rates on COIL database with different levels of noise.

2) Robustness Evaluation: Secondly, we testify the influence of different noise levels by comparing with other algorithms. To be specific, we adopt 0%, 5%, 10%, 20%, 30%, and 50% corruptions under the 20-objects COIL task. Fig. 8 lists the robustness results, in which we notice that our approach consistently works better than others. It indicates that our proposed model is very robust to noise, especially heavy noise, which makes our approach applicable to real-world scenarios.

3) Parameter Analysis: Thirdly, we aim to evaluate two parameters λ and α. To be specific, we jointly analyze two parameters on CMU-PIE benchmark under 40 training images per person. The impacts of parameters on performance are reported in Fig. 9, where we notice that larger α would lead to better results, which indicates the fact that the term $\mathrm{tr}(\mathcal{AM})$ plays an important role in our discriminative metric learning. Moreover, we also observe that λ influences a little to the final performance. That means, in this minimization problem (Eq. 6), compared to the term $\|\mathcal{M}X - D\|_{\mathrm{F}}^2$, we will need to punish more on the trace norm. This can be observed from the Fig. 9, too, and we also tune our model parameters $\{\alpha, \lambda\}$ in this way.

On the other hand, when removing the third term in Eq.

1051-8215 (c) 2018 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2879626, IEEE Transactions on Circuits and Systems for Video Technology

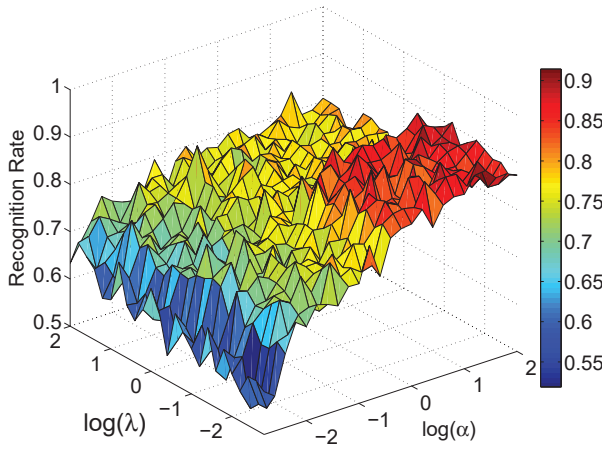


Fig. 9. Impacts of parameters λ and α on CMU-PIE face dataset with 40 training samples per subject.

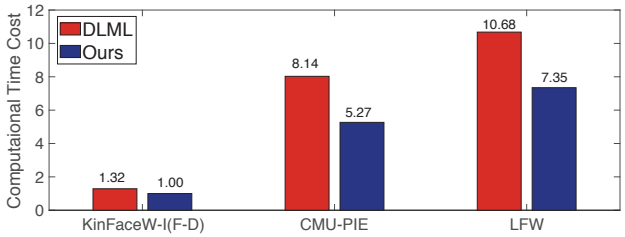


Fig. 10. Computational time cost of our current version and our previous conference version. Here we adopt $\log()$ to scale the time.

6, we find that the overall performance is compromised. That means the third term helps a lot. Therefore, we generally set $\alpha = 10^2, \lambda = 10^{-2}$ as default in our experiments.

4) Computational Cost: We calculate the computational costs of our conference work [38] and this current work. We run experiments on three datasets ranged from small- to large-scale in 20 iterations to calculate the training time. Specifically, we evaluate on F-D from KinFaceW-I (LBP feature), CMU-PIE (60 training samples per subject), and LFW. We evaluate on Matlab 2017b with Intel i7-3770 CPU and 64GB memory. Fig. 10 shows the training time (in second). Note that we use $\log()$ to rescale the training time axis for better illustrations.

From the results, we witness that the proposed model is more efficient than our previous conference work. The most time-consuming part of our previous work is the low-rank constraint on reconstruction coefficient matrix D with SVD in the optimization. To address this, we speed up the SVD by introducing a fixed-rank matrix decomposition, and experimental results demonstrate that the speedup version in this paper works fairly well on large-scale data, in terms of both recognition performance and running time, especially on real-world dataset.

V. Conclusions

In this paper, we proposed a robust discriminative metric learning algorithm via seeking a fast low-rank representation and building a compact basis in a unified framework. Specifically, low-rank representation aimed to capture the global structure within the data to facilitate the discriminative linear projection learning. In addition, our algorithm was accelerated so that it could well handle the large-scale datasets in real world. Furthermore, a compact basis was incorporated for denoising linear projection, especially when data was corrupted. Experimental evaluations on several datasets had witnessed the effectiveness of our approach by comparing with other methods.

References

[1] B. Kulis, "Metric learning: A survey," *Machine Learning*, vol. 5, no. 4, pp. 287–364, 2012.
[2] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *arXiv preprint arXiv:1306.6709*, 2013.
[3] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1875–1882.
[4] M. Kan, S. Shan, D. Xu, and X. Chen, "Side-information based linear discriminant analysis for face recognition." in *British Machine Vision Conference*, 2011, pp. 1–12.
[5] Z. Gu, M. Shao, L. Li, and Y. Fu, "Discriminative metric: Schatten norm vs. vector norm," in *21st International Conference on Pattern Recognition*. IEEE, 2012, pp. 1213–1216.
[6] J. Lu, J. Hu, and Y.-P. Tan, "Discriminative deep metric learning for face and kinship verification," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4269–4282, 2017.
[7] J. Lu, J. Hu, and J. Zhou, "Deep metric learning for visual understanding: An overview of recent advances," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 76–84, 2017.
[8] J. Hu, J. Lu, and Y.-P. Tan, "Deep transfer metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 325–333.
[9] Y. Duan, J. Lu, J. Feng, and J. Zhou, "Deep localized metric learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.
[10] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.
[11] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 171–184, 2013.
[12] Y. Pan, H. Lai, C. Liu, and S. Yan, "A divide-and-conquer method for scalable low-rank latent matrix pursuit," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 524–531.
[13] Y. Pan, R. Xia, J. Yin, and N. Liu, "A divide-and-conquer method for scalable robust multitask learning," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 12, pp. 3163–3175, 2015.
[14] S. Xiao, W. Li, D. Xu, and D. Tao, "Falrr: A fast low rank representation solver," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4612–4620.
[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.
[16] Y. Duan, W. Zheng, X. Lin, J. Lu, and J. Zhou, "Deep adversarial metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2780–2789.
[17] H. O. Song, S. Jegelka, V. Rathod, and K. Murphy, "Deep metric learning via facility location," pp. 5382–5390, 2017.
[18] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4004–4012.
[19] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies," pp. 360–368, 2017.
[20] S. Li and Y. Fu, "Learning robust and discriminative subspace with low-rank constraints," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–1, 2015.
[21] Z. Ding, M. Shao, and Y. Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *The 28th AAAI Conference on Artificial Intelligence*, 2014, pp. 1192–1198.
[22] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou, "Neighborhood repulsed metric learning for kinship verification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 331–345, 2014.

1051-8215 (c) 2018 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2879626, IEEE Transactions on Circuits and Systems for Video Technology

[23] S. Wang, Z. Ding, and Y. Fu, "Cross-generation kinship verification with sparse discriminative metric," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[24] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[25] G. Cheng, P. Zhou, and J. Han, "Duplex metric learning for image set classification," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 281–292, 2018.

[26] K. Li, Z. Ding, K. Li, Y. Zhang, and Y. Fu, "Support neighbor loss for person re-identification," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 1492–1500.

[27] J. Hu, J. Lu, and Y.-P. Tan, "Sharable and individual multi-view metric learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 9, pp. 2281–2288, 2018.

[28] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Neural information processing systems*, 2002, pp. 505–512.

[29] Z. Ding and Y. Fu, "Robust transfer metric learning for image classification," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 660–670, 2017.

[30] G. Zhong, K. Huang, and C.-L. Liu, "Low rank metric learning with manifold regularization," in *IEEE 11th International Conference on Data Mining*, 2011, pp. 1266–1271.

[31] D. Lim, G. Lanckriet, and B. McFee, "Robust structural metric learning," in *The 30th International Conference on Machine Learning*, 2013, pp. 615–623.

[32] W. Liu, C. Mu, R. Ji, S. Ma, J. R. Smith, and S.-F. Chang, "Low-rank similarity metric learning in high dimensions," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 2792–2799.

[33] H. Junlin, L. Jiwen, T. Yap-Peng, and Z. Jie, "Deep transfer metric learning." *IEEE transactions on image processing*, vol. 25, no. 12, pp. 5576–5588, 2016.

[34] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: remote sensing image scene classification via learning discriminative cnns," *IEEE transactions on geoscience and remote sensing*, vol. 56, no. 5, pp. 2811–2821, 2018.

[35] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," in *Advances in Neural Information Processing Systems*, 1994, pp. 737–744.

[36] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems*, 2016, pp. 1857–1865.

[37] E. Kim, M. Lee, and S. Oh, "Elastic-net regularization of singular values for robust subspace learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 915–923.

[38] Z. Ding, S. Suh, J.-J. Han, C. Choi, and Y. Fu, "Discriminative low-rank metric learning for face recognition," in *12th IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[39] Q. Cao, Y. Ying, and P. Li, "Similarity metric learning for face recognition," in *IEEE International Conference on Computer Vision*, 2013, pp. 2408–2415.

[40] 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," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[41] Z. Ding, M. Shao, and Y. Fu, "Deep robust encoder through locality preserving low-rank dictionary," in *European Conference on Computer Vision*. Springer, 2016, pp. 567–582.

[42] M. Shao, Z. Ding, H. Zhao, and Y. Fu, "Spectral bisection tree guided deep adaptive exemplar autoencoder for unsupervised domain adaptation," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 2023–2029.

[43] S. Wang, Z. Ding, and Y. Fu, "Coupled marginalized auto-encoders for cross-domain multi-view learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016, pp. 2125–2131.

[44] ——, "Marginalized denoising dictionary learning with locality constraint," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 500–510, 2018.

[45] J. Zhang and L. Zhang, "Efficient stochastic optimization for low-rank distance metric learning." in *31st AAAI Conference on Artificial Intelligence*, 2017, pp. 933–940.

[46] Y. Ying and P. Li, "Distance metric learning with eigenvalue optimization," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1–26, 2012.

[47] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *The 24th International Conference on Machine Learning*, 2007, pp. 209–216.

[48] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[49] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *IEEE 12th International Conference on Computer Vision*, 2009, pp. 498–505.

[50] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2288–2295.

[51] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[52] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[53] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Asian conference on computer vision*. Springer, 2010, pp. 709–720.

[54] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.

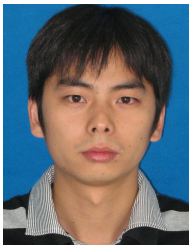
[55] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[56] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3610–3617.

[57] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *European conference on computer vision*. Springer, 2014, pp. 1–16.

[58] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2666–2672.

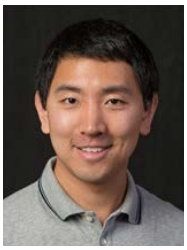
[59] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.



Zhengming Ding (S'14-M'18) received the B.Eng. degree in information security and the M.Eng. degree in computer software and theory from University of Electronic Science and Technology of China (UESTC), China, in 2010 and 2013, respectively. He received the Ph.D. degree from the Department of Electrical and Computer Engineering, Northeastern University, USA in 2018. He is a faculty member affiliated with Department of Computer, Information and Technology, Indiana University-Purdue University Indianapolis since 2018. His research interests include transfer learning, multi-view learning and deep learning. He received the National Institute of Justice Fellowship during 2016-2018. He was the recipients of the best paper award (SPIE 2016) and best paper candidate (ACM MM 2017). He is currently an Associate Editor of the Journal of Electronic Imaging (JEI). He is a member of IEEE.

1051-8215 (c) 2018 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2879626, IEEE Transactions on Circuits and Systems for Video Technology



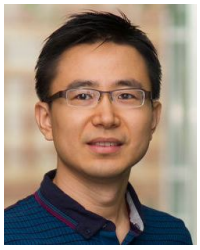
Ming Shao (S'11-M'16) received the B.E. degree in computer science, the B.S. degree in applied mathematics, and the M.E. degree in computer science from Beihang University, Beijing, China, in 2006, 2007, and 2010, respectively. He received the Ph.D. degree in computer engineering from Northeastern University, Boston MA, 2016. He is a tenure-track Assistant Professor affiliated with College of Engineering at the University of Massachusetts Dartmouth since 2016 Fall. His current research interests include sparse modeling, low-rank matrix analysis, deep learning, and applied machine learning on social media analytics. He was the recipient of the Presidential Fellowship of State University of New York at Buffalo from 2010 to 2012, and the best paper award winner/candidate of IEEE ICDM 2011 Workshop on Large Scale Visual Analytics, and ICME 2014. He has served as the reviewers for many IEEE Transactions journals including TPAMI, TKDE, TNNLS, TIP, and TMM. He has also served on the program committee for the conferences including AAAI, IJCAI, and FG. He is the Associate Editor of SPIE Journal of Electronic Imaging, and IEEE Computational Intelligence Magazine. He is a member of IEEE.

Wonjun Hwang (M'15) received both B.S. and M.S. degrees from the Department of Electronics Engineering, Korea University, Korea, in 1999 and 2001, respectively, and Ph.D. degree in the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Korea, in 2016. From 2001 to 2008, he was a research staff member in Samsung Advanced Institute of Technology (SAIT), Korea, where he contributed to the promotion of Advanced Face Descriptor, Samsung and NEC joint proposal, to MPEG-7 international standardization in 2004. He proposed the SAIT face recognition engine which achieved the best results under the uncontrolled illumination situation at Face Recognition Grand Challenge (FRGC) and Face Recognition Vendor Test (FRVT) in 2006. He developed the real-time face recognition engine for the Samsung cellular phone, SGH-V920, in 2006. From 2009 to 2011, he was a senior engineer in Samsung Electronics, Korea, where he worked on developing face and gesture recognition modules for Samsung humanoid robot, a.k.a RoboRay. In 2011, he rejoined the SAIT as a research staff member for face recognition. His research interests are in face recognition, computer vision and pattern recognition.

Sungjoo Suh received the B.S. and M.S. degrees in Electronics Engineering from Korea University, Seoul, Korea, and the Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University. Since 2009, he has been an R&D Staff Member with Samsung Advanced Institute of Technology, Suwon, Korea. His current research interests include image processing, pattern recognition, and biometrics. He has also worked on interactive display architecture, computational photography, and image watermarking. He has served as a Program Committee Member for VECTaR 2014 in conjunction with European Conference on Computer Vision 2014.

Jae-Joon Han (M'07) received the B.S. degree in electronic engineering from Yonsei University, Korea, in 1997, the M.S. degree in electrical and computer engineering from the University of Southern California, Los Angeles, in 2001, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2006. He was with Purdue University, as a Post-Doctoral Fellow in 2007. Since 2007 he has been with the Samsung Advanced Institute of Technology, Gyeonggi-do, Korea, since 2007, as a Principal Researcher. His research interests include statistical machine learning and data mining, computer vision, and real-time recognition technologies. He also participated in the development of standards, such as ISO/IEC 23005 (MPEG-V) and ISO/IEC 23007 (MPEG-U), and served as the Editor of ISO/IEC 23005-1/4/6.R 2014 in conjunction with European Conference on Computer Vision 2014.

Changkyu Choi is a Vice President of Samsung Advanced Institute of Technology, where he is in charge of a biometric authentication project including face, fingerprint, etc. Prior to leading the project, he conducted various researches on speech technologies including blind source separation and sound source localization, human-robot interaction utilizing audio-visual cues, and 3-D interaction techniques using full-body and hand posture recognition. Dr. Choi received his Ph.D. degree in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, Korea in 1999. He was a visiting scholar at Imperial College, London, U.K. from 2013 to 2014.



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, respectively, 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, respectively. He is an interdisciplinary faculty member affiliated with College of Engineering and the College of Computer and Information Science at Northeastern University since 2012. His research interests are Machine Learning, Computational Intelligence, Big Data Mining, Computer Vision, Pattern Recognition, and Cyber-Physical Systems. He has extensive publications in leading journals, books/book chapters and international conferences/workshops. He serves as associate editor, chairs, PC member and reviewer of many top journals and international conferences/workshops. He received seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, Grainger Foundation; nine Best Paper Awards from IEEE, IAPR, SPIE, SIAM; many major Industrial Research Awards from Google, Samsung, and Adobe, etc. He is currently an Associate Editor of the IEEE Transactions on Neural Networks and Leaning Systems (TNNLS). He is fellow of IAPR, OSA and SPIE, Lifetime Distinguished Member of ACM, Lifetime Member of AAAI and Institute of Mathematical Statistics, member of ACM Future of Computing Academy, Global Young Academy, AAAS, INNS and Beckman Graduate Fellow during 2007-2008.

1051-8215 (c) 2018 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.