Nonnegative Tensor Cofactorization and Its Unified Solution

Xiaobai Liu, Qian Xu, Shuicheng Yan, Gang Wang, Hai Jin, and Seong-Whan Lee

***Abstract*— In this paper, we present a new joint factorization algorithm, called nonnegative tensor cofactorization (NTCoF). The key idea is to simultaneously factorize multiple visual features of the same data into nonnegative dimensionality-reduced representations, and meanwhile, to maximize the correlations of the low-dimensional representations. The data are generally encoded as tensors of arbitrary order, rather than vectors, to preserve the original data structures. NTCoF provides a simple and efficient way to fuse multiple complementary features for enhancing the discriminative power of the desired rank-reduced representations under the nonnegative constraints. We formulate the related objectives with a block-wise quadratic nonnegative function. To optimize, a unified convergence provable solution is developed. This solution is applicable for any nonnegative optimization problems with block-wise quadratic objective functions, and thus offer an unified platform based on which specific solution can be directly derived by skipping over tedious proof about algorithmic convergence. We apply the proposed algorithm and solution on three image tasks, face recognition, multiclass image categorization, and multilabel image annotation. Results with comparisons on public challenging data sets show that the proposed algorithm can outperform both the traditional nonnegative methods and the popular feature combination methods.**

***Index Terms*— Nonnegative matrix/tensor factorization, feature combination, multi-task learning, multi-class image classification.**

I. Introduction

NONNEGATIVE matrix factorization (NMF) [24] is proposed to decompose a matrix into a product of lower-rank nonnegative matrices. It is further extended by tensor representation, known as nonnegative tensor factorization (NTF) [37]. Nonnegative constraints are enforced to allow only additive combinations. An NMF/NTF approach usually results in a rank-reduced representation of the original data. A recent survey on NMF/NTF is referred to [4].

In this work, we further investigate the NMF/NTF and particularly consider the feature combination problem in computer vision. The major motivation beyond feature combination is to adaptively fuse a set of diverse yet complementary feature spaces, e.g., color, shape or texture. Generally, different feature descriptors are discriminative for different classes. For example, color based features usually perform well for differentiating road from sky, but not so effective for classifying the buildings from cars. To enhance the robustness of the desired representations, many efforts have been devoted to exploring multiple complementary features [16], [48]. Our method can also be applied to utilize multiple modalities of multimedia content (e.g. visual, audio, emotions, touch etc) in multimedia processing tasks.

Following the above methodology, we present a novel feature fusion algorithm based on data factorization, called nonnegative tensor co-factorization (NTCoF). The goal is to factorize multiple different yet complementary feature representations of the same data by synchronizing the inter-feature correlation, such that the factorization under one feature representation can well harness the information of other representations. Therefore, there are two general purposes in our approach: i) minimizing the factorization residues to obtain optimal nonnegative low-dimensional representation of the data under each feature; and ii) maximizing the correlation between the desired low-dimensional representations of the same datum. We integrate these two purposes into a block-wise convex quadratic function with nonnegative constraints, which expresses the data with tensors of arbitrary order (e.g. > 2) to preserve the original data structures.

The optimization problem of NTCoF, though intractable, can be divided into several subproblems and solved by the alternate optimization techniques [24]. Based on the same strategy, this work further contributes a unified framework that can provide theoretically provable convergent solutions to general data factorization problems. The solution is under a wild condition, since it only requires the objective function is block-wise quadratic which most of existing nonnegative problems have. This solution can favor the specific formulations by skipping over the tedious mathematical proof, and thus serves as a 'one-stop' toolkit for solving newly proposed data factorization problems.

Manuscript received July 28, 2012; revised September 9, 2013 and January 5, 2014; accepted January 6, 2014. Date of publication June 2, 2014; date of current version July 28, 2014. This work was supported in part by the National Natural Science Foundation of China under Grant 61133008, in part by the National Research Foundation of Korea through the Ministry of Science, ICT and Future Planning under Grant 2012-005741, and in part by the Sub Project of Singapore National Retail Federation/NUS-Tsinghua Extreme Search Centre Center through Live Search in Camera Networks. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Erhardt Barth. *(Corresponding author: Seong-Whan Lee.)*

X. Liu was with the Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Korea. He is now with the Department of Statistics, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: xbliu.lhi@gmail.com).

Q. Xu is with the Department of Statistics, San Diego State University, San Diego, CA 92115 USA (e-mail: qxu@rohan.sdsu.edu).

S. Yan is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: eleyans@nus.edu.sg).

G. Wang is with the Department of School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wanggang@ntu.edu.sg).

H. Jin is with the Department of Computer Science, Huazhong University of Science and Technology, Wuhan 430030, China (e-mail: hjin@hust.edu.cn).

S.-W. Lee is with the Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Korea (e-mail: swlee@image.korea.ac.kr).

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.2014.2327806

The proposed NTCoF method formulates the discriminative data factorization problem in the setting of multi-view learning or multi-task learning [33], [51]. Here one task is referred to inferring the optimal factorization of the data under one visual feature. We aim to jointly estimate the factorizations under multiple feature representations, and impose a cross-task constraint to maximize the correlation between the desired low-dimensionality representations. This cross-task constraint is valuable since factorizations under one feature may favor different coefficients, yet enforcing the cross-task consistency usually leads to robust coefficient estimation. In this context, we can borrow the strength of multi-task learning to improve the discriminative power of the dimensionality-reduced representations. To our best knowledge, the work of NTCoF is the first time to utilize cross-feature strategy to enhance the discriminative power of nonnegative data factorizations. We apply NTCoF for three image tasks, face recognition, multi-class image categorization and multi-label image annotation, and compare it to the traditional nonnegative methods and feature combination methods on public datasets.

A. Related Works

This work is closely related to the advances in machine learning and computer vision, which are reviewed from three aspects, nonnegative matrix factorization (NMF) and related optimization, multi-tasks learning and correlation analysis.

There exist many efforts on **NMF** related problems, and the work of Lee and Seung [24] brings much attention to NMF in both machine learning and computer vision communities. Many alternate **optimization** strategies have been proposed to solve the NMF/NTF related problems, and can be roughly divided into three categories [4]. i) *Alternate least squares methods*. It sequentially minimizes one factor under the nonnegative constraint with other factors fixed. Thus the original problem is divided into multiple sub-problems and each can be solved by traditional numerical optimization techniques. For example, Chu *et al.* [11] propose to use the projected Newton's method for optimization. The convergence of this alternate strategy is proved by Paatero *et al.* in [35]. ii) *Gradient descent methods*. This category generally first reformulates the constrained NMF/NTF related problem into an unconstrained one, and then applies the standard gradient descent approach to obtain the locally optimal factorization, in an alternate way. A key component of this strategy is how to choose an appropriate step size. Lin *et al.* [27] utilize a projected gradient approach to heuristically select the optimal step size, which shows better convergence performance than the ones with fixed step sizes. However, the usage of certain auxiliary constraint for NMF/NTF may break down the bound-constrained optimization assumption, thus limit its applicability. iii) *multiplicative update methods*. For each iteration, each element of the factors is multiplied by a nonnegative factor, and thus all the elements are strictly nonnegative if the initial factor are nonnegative. The pioneering work is presented by Lee *et al.* [24], with many extensions and followups [26], [36], [42]. In particular, Gillis *et al.* [17] proposed to accelerate the multiplicative update procedure by aggressively updating one factor while keeping the other factor fixed. However, one open issue of this strategy is the high complexity and difficulty in proving the algorithmic convergence. The latter becomes even worse when NMF/NTF related formulation contains auxiliary regularization terms. In this work, we will develop a unified solution to above nonnegative problems with block-wise quadratic objective function.

Multi-task learning has been extensively studied in both theory and practice in the past literature. The basic methodology is to learn multiple different yet correlated tasks together, and meanwhile, to maximize the inter-task correlation. In particular, similar idea has been widely used to combine multiple types of features in class-level object recognition and image classification. One popular method in computer vision literature is Multiple Kernel Learning (MKL), that linearly combines similarity functions between images [16], [41], [32]. Yuan *et al.* present a multi-task joint sparsity algorithm for feature fusion and achieve impressive results on multiple datasets [48]. Recently, Han et al. [18] assume the input data in multiple tasks are generated from a latent common domain and proposed a latent probit model to jointly learn the domain transforms. Yang et al. [46] consider the feature correlations as well and presented a feature selection method. In comparisons, our method takes advantages of the non-negativity analysis for feature fusions. Non-negativity is a natural choice while applying factorization techniques for image related problems as shown in [4].

Another work related with NTCoF is the **Canonical Correlation Analysis** (CCA) [19], which has been widely used for uncovering the pair-wise correlation between two or multiple sets of variables. For example, Fu *et al.* in [15] propose to fuse multiple features by seeking the optimal subspace and simultaneously maximizing the sum of canonical correlation between different subspace representations of the same sample. In contrast, NTCoF is characterized by following aspects. (i) it utilizes inter-representation correlation as a type of soft constraint for data factorization, and preserve the nonnegativity of the synchronized components in multiple low-dimensional representations. Nonnegativity is intuitively natural, especially for the image tasks to study, which is however not preserved by CCA methods [15]. (ii) NTCoF could seek good decompositions through using multiple cross-modality features (either the visual features or the manual annotations). In contrast, the method in [19] can only handle two feature sets. (iii) NTCoF is formulated based on the tensor representation, which considers each input image as a two-dimensional matrix, instead of vector. It has been recognized [4] that vectorizing images usually leads to the loss of local structure information, which is crucial for classification tasks. The advantages of NTCoF over CCA will be demonstrated by extensive experiments with comparisons.

II. Nonnegative Tensor Co-Factorization

The basic idea of nonnegative tensor co-factorization (NTCoF) is to adaptively combine a set of diverse yet complementary feature spaces, either appearance features based on color, shape and texture, or manually annotated class

information, e.g., labels, keywords or tags. It has been demonstrated in the past literature that combining multiple features will improve the discriminative power as compared to using single feature type. Following the same methodology, in this work, we extend the nonnegative data factorization to multi-task setting, so the desired dimension-reduction representation under one feature can well harness the knowledge from the representations under other features. This new method can naturally take advantages of the nonnegativity analysis, data factorization, and cross-task consistency.

Let $\mathcal{X}^m = [\mathcal{X}_1^m, \mathcal{X}_2^m, \ldots, \mathcal{X}_N^m]$, denote the n^{th} order tensor under the m^{th} feature, and each datum is a $(n-1)^{th}$ order tensor $\mathcal{X}_i^m \in R^{d_1 \times d_2 \times \cdots \times d_{n-1}}$, $m = 1..M$. M is the number of feature representations extracted for each image. d_b denotes the dimension of the b^{th} order where $b = 1..n$. Notice n is usually 3 for image tasks, and d_n is the total number of training images. We assume that $\mathcal{X}^m$ is nonnegative without loss of generalization.

A. Objective-I: Data Reconstruction

The task of nonnegative tensor factorization is to derive a set of nonnegative bases which are linearly mixed by nonnegative encoding coefficients. Let k denote the number of the desired bases. The factorization of $\mathcal{X}^m$ can be represented as the sum of k rank-1 tensors, taking the following form,

$$\mathcal{X}^m \approx \sum_{i=1}^{k} u_i^{1^m} \otimes \ldots u_i^{b^m} \ldots \otimes u_i^{n-1^m} \otimes u_i^{n^m} \tag{1}$$

where $\otimes$ denotes the Kronecker product operator, $u_i^{b^m}$ denotes a rank-1 tensor indexed by $i = 1..k, b = 1..n-1$, and $m = 1..M$. For clarity, we introduce a row vector $H_i{}^m$ so $u_i^{n^m} = H_i{}^{m^T}$ where ^{T} indicates the transpose of a matrix.

Let $\tau_i^m = (u_i^{b^m} \otimes)_{b=1}^{n-1} = u_i^{1^m} \otimes \ldots u_i^{b^m} \ldots \otimes u_i^{n-1^m}$ to facilitate presentation. Each datum $\mathcal{X}^m$ is encoded as a superposition of $\tau_1^m, \ldots, \tau_k^m$, and the reconstruction coefficients are $H_1^m, \ldots, H_k^m$. The corresponding objective function to optimize is,

$$\arg \min_{\{u_i^{b^m}\}, \{h_i{}^m\}} \sum_m \|\mathcal{X}^m - \sum_{i=1}^{k} \tau_i^m \otimes H_i{}^m\|^2 \quad s.t.\ u_i^{b^m}, \quad h_i{}^m \geq 0 \tag{2}$$

where $\|\cdot\|$ indicates the Frobenius norm of a matrix. Let $H^m = [H_1^{m^T}, \ldots, H_k^{m^T}]^T \in R^{k \times d_n}$. Usually, $k < \min(\prod_{b=1}^{n-1} d_b, d_n)$, and H^m could be considered as the dimensionality-reduced representation of $\mathcal{X}^m$ with the objective of best reconstruction under nonnegative constraints.

B. Objective-II: Mutual Correlation

In addition to minimizing the data reconstruction errors in Eq. (2), another goal of nonnegative tensor co-factorization (NTCoF) is to maximize the mutual correlation between the individual factorizations under different features. This can be achieved by maximizing the sum of pair-wise correlation between the desired low-dimensional representations of the same datum. Formally, in order to optimize the above two purposes in the same objective function, we divide the coefficients matrix H^m into two non-overlapping parts. Let

$$H^m = \begin{bmatrix} \tilde{H}^m \\ \hat{H}^m \end{bmatrix}, \tag{3}$$

where $\tilde{H}^m \in R^{q \times d_n}$ and $\hat{H}^m \in R^{(k-q) \times d_n}$. Herein, $\tilde{H}^m$ is desirable for certain discriminative purpose, e.g. classification, while the whole H^m is used for the purpose of data reconstruction. Same strategy has been used in previous works [42], [45]. Denote $U^{b^m} = [u_1^{b^m}, u_2^{b^m}, \ldots, u_k^{b^m}] \in R^{d_b \times k}$ as the basis matrix, and $H = [H^1; H^2; \ldots; H^M]$. Accordingly, we divide U^{b^m} into two parts,

$$U^{b^m} = [\tilde{U}^{b^m}, \hat{U}^{b^m}] \tag{4}$$

where $\tilde{U}^{b^m} \in R^{d_b \times q}$ and $\hat{U}^{b^m} \in R^{d_b \times (k-q)}$. Let $(\tilde{h})_j^m$ denote the j^{th} column of $\tilde{H}^m$. Thus, the goal of maximizing the pairwise correlation of multiple factorizations can be encoded as,

$$\max_H \sum_{j, m \neq n} (\tilde{h})_j^{m^T} (\tilde{h})_j^n \tag{5}$$

Since $\hat{U}^{b^m}$ is the complementary space of $\tilde{U}^{b^m}$, we can revert to optimize

$$\min_H \sum_{j,m,n} e^{(k-q)^T} (\hat{h})_j^m (\hat{h})_j^{n^T} e^{(k-q)} \Rightarrow \min_H Tr(\hat{H} \hat{H}^T E^{M(k-q) \times M(k-q)}), \tag{6}$$

where $Tr(\cdot)$ returns the trace of a matrix, $e^{(k-q)}$ denotes an all-one column vector of $(k-q)$-dimension and $E^{M(k-q) \times M(k-q)} \in R^{M(k-q) \times M(k-q)}$ denotes an all-one matrix.

C. Unified Objective Function

We combine the objectives of Eq. (2) and Eq. (6) into a unified objective function,

$$F(U^{b^m}, H) = \sum_m \|\mathcal{X}^m - \sum_{i=1}^{k} \tau_i^m \otimes H_i{}^m\|^2 + \lambda Tr(\hat{Q} \hat{H} \hat{H}^T \hat{Q}^T E^{M(k-q) \times M(k-q)}), \quad s.t.\ U^{b^m}, H^m \geq 0 \tag{7}$$

where,

$$\hat{Q} = \prod_{b=1}^{n-1} diag\{\hat{Q}^{b^1}, \ldots, \hat{Q}^{b^m}, \ldots, \hat{Q}^{b^M}\}, \tag{8}$$

with,

$$\hat{Q}^{b^m} = diag\{|\hat{u}_1^{b^m}|, \ldots, |\hat{u}_{k-q}^{b^m}|\} \tag{9}$$

$|\cdot|$ denotes the ℓ_1-norm of a vector, and λ is a weighting constant. Herein, $\hat{Q}$ is introduced to scale the coefficients $\hat{H}$ using the norms of basis, and to avoid the trial solution [42].

III. Unified Solution to Nonnegative Data Factorization Problems

Eq. (7) is a quadratic convex function with respect to U^{b^m} and H^m respectively, though intractable, it can be solved by alternate optimization techniques [24], [28], [42]. Here, we aim to develop a unified solution to these problems, which is characterized by: 1) it provides a unified solution to nonnegative data factorization problems with block-wise quadratic objective functions (regularized or not, unsupervised or supervised), and 2) it can be used as a general template to derive update rules for new optimization problems, which are theoretically correct and convergent. The only assumption of this unified solution is that the objective function is block-wise quadratic, and thus it is widely applicable for a large community of nonnegative data factorization problems.

A. Assumption and Update Rules

The objective function of nonnegative problems usually contains one data reconstruction term and one regularization term, as in Eq. (7). Formally, we denote $F((U^b)_{b=1}^{n-1}, H)$ as the objective function to optimize and have the following assumption.

Assumption : The objective function $F((U^b)_{b=1}^{n-1}, H)$ is assumed block-wise quadratic, namely, when $(U^b)_{b=1}^{n-1}$ are fixed, F is quadratic with respect to H, and on the other hand, when H and $(U^p)_{p=1,p\neq b}^{n-1}$ is fixed, F is quadratic with respect to U^b.

The objective function is often of high order although block-wise quadratic, and generally a closed-form solution does not exist. This high-order intractable optimization problem can be transformed into a set of tractable sub-problems, and achieve the convergence to a local optimum in an alternate way. Here, we adopts the multiplicative nonnegative update rules method to optimize $(U^b)_{b=1}^{n-1}$ and H.

For given H and $(U^p)_{p=1,p\neq b}^{n-1}$ at the current step, the objective function F with respect to U^b can be rewritten as,

$$F(U^b) = F(U^b, (U^p)_{p=1,p\neq b}^{n-1}, H), \quad s.t.\ U^b \geq 0. \quad (10)$$

As the objective function $F(U^b)$ is quadratic, its derivative with respect to U^b is then of first order, which can be expressed in the form,

$$\frac{\partial F(U^b)}{\partial U^b} = \sum_{l=1}^{K} A^l U^b B^l + C, \quad (11)$$

where A^l, B^l, and C are real-value constant matrices. Although theoretically K shall be very large to obtain such a general form, many popular objective functions often lead to very small K (even with $K = 1$) as introduced afterward.

Letting ϕ_{ij} denote the Lagrange multiplier for constraint $U_{ij}^b \geq 0$ and $\Phi = [\phi_{ij}]$, we apply the Karush-Kuhn-Tucker (KKT) condition [44] of $\phi_{ij} U_{ij}^b = 0$ to the derivative of Lagrange function and obtain

$$\sum_{l=1}^{K} (A^l U^b B^l)_{ij} U_{ij}^b + C_{ij} U_{ij}^b = 0.$$

We decompose the matrix A^l, B^l and C as the difference of two nonnegative parts, denoted as $A^l = A^{l+} - A^{l-}$, $B^l = B^{l+} - B^{l-}$ and $C = C^+ - C^-$, respectively. Herein, for A_{ij}^l, only one of the A_{ij}^{l+} and A_{ij}^{l-} can be nonzero, which is also applicable for B_{ij}^l and C_{ij}. Thus, we can obtain the relation update rule, which is consequently used as

$$U_{ij}^b \leftarrow U_{ij}^b \times \frac{[\sum_{l=1}^{K} (A^{l+} U^b B^{l-} + A^{l-} U^b B^{l+}) + C^-]_{ij}}{[\sum_{l=1}^{K} (A^{l+} U^b B^{l+} + A^{l-} U^b B^{l-}) + C^+]_{ij}}. \quad (12)$$

Appendix-A gives the mathematical proof about the convergence of the above update rules. The rules for $(U^b)_{b=1}^{n-1}$ and H can be derived in the same fashion. Once initializing the factors, namely U^{b^m} and H^m, the optimization procedure alternately iterates respective multiplicative update rules till convergence.

B. Unified Solution as a General Optimization Template

The above unified solution to nonnegative data factorization problems can be taken as a general template to re-explain a large community of nonnegative data factorization algorithms. Specially, for nonnegative matrix factorization problems, the specific update rules for certain block-wise quadratic objective function can be obtained in two steps: i) calculate the partial derivative with respect to each of two factor matrices to determine the parameters defined in Eqn. (12), including $K, A^l, B^l, C, l = 1..K$, and ii) obtain the update rules by substituting the parameters in Eqn. (12).

Table I summarizes the application of the solution for NMF [24], projective NMF [49], semi-NMF [9], and convex NMF [9]. Herein, X denotes the input matrix, and the basic goal of above algorithms is to factorize X into the product of a base matrix W and a coefficient matrix H. For each algorithm, we show the algorithm name in the 1^{st} column, the corresponding objective function in the 2^{nd} column, and the solution parameters while deriving the update rules of W for given H as well as the derived update rule for W in the 3^{rd} column. Note that: i) we do not report the details on the update rules for H since it has the same form as that of that for W; ii) convex NMF belongs to semi-nonnegative data factorization, and the input data matrix X may have mixed signs; iii) although the update rules for PNMF-I [49] can be derived, it however cannot be proved convergent within the proposed framework, since A_2^+ and A_3^+ are not constant matrices and then the Eqn. (32) cannot be guaranteed, and iv) in the objective function of the manifold NMF [6], the matrix L represents the graph Laplacian matrix, defined as $L = D - G$, where S is the similarity matrix defined on the specific graph and D is a diagonal matrix whose entries are column sums of G, namely $D_{ii} = \sum_j G_{ij}$.

From Table I, we can observe that most popular NMF related algorithms can be unified within the same framework. It is worthy highlighting that this unified solution can relieve the researchers from tedious mathematical deduction on updating rules and proof of the algorithmic convergence.

TABLE I

A UNIFIED RE-DERIVING OF A LARGE COMMUNITY OF NMF/NTF RELATED PROBLEMS. THE FIRST COLUMN INDICATES THE ALGORITHMS NAME. FOR EACH ALGORITHM, THE CORRESPONDING THREE ROWS SHOW, FROM TOP TO BOTTOM, THE OBJECTIVE FUNCTION, PARAMETERS FROM THE PARTIAL DERIVATIVE IN EQN. (12), AND THE DERIVED UPDATE RULE. HERE, WE SHOW THE RULES OF W FOR GIVEN H ONLY, AND THE UPDATE RULE OF H CAN BE OBTAINED BY USING THE SAME PROCEDURE

NMF [24]	Objective function	$F(W,H;X)=\|X-WH\|^2,\ \ X\geq 0$
	Parameters	1) $K=1$; 2) $C^+=0, C^-=2XH^T$; 3) $A^{1+}=2I, A^{1-}=0, B^{1+}=HH^T, B^{1-}=0$.
	Rules	$W_{ij}=W_{ij}\times\frac{(XH^T)_{ij}}{(WHH^T)_{ij}}$.
PNMF-I [50]	Objective function	$F(W;X)=\|X-WW^TX\|^2,\ \ X\geq 0$
	Parameters	1)$K=3$; 2) $C^+=0, C^-=0$; 3)$A^{1+}=0, A^{1-}=2XX^T, B^{1+}=I, B^{1-}=0$; 4) $A^{2+}=WW^TXX^T, A^{2-}=0, B^{2+}=I, B^{2-}=0$; 5) $A^{3+}=XX^TWW^T, A^{3-}=0, B^{3+}=I, B^{3-}=0$.
	Rules	$W_{ij}=W_{ij}\times\frac{2(XX^TW)_{ij}}{(WW^TXX^TW+XX^TWW^TW)_{ij}}$.
Orthogonal NMF [10]	Objective function	$F(W,H;X)=\|X-WH\|^2+tr(W^TW-I),\ \ X\geq 0$
	Parameters	1) $K=2$; 2) $C^+=0, C^-=2(XH^T)$; 3) $A^{1+}=2I, A^{1-}=0, B^{1+}=HH^T, B^{1-}=0$; 4) $A^{2+}=2I, A^{2-}=0, B^{2+}=I, B^{2-}=0$.
	Rules	$W_{ij}=W_{ij}\times\frac{((XH^T))_{ij}}{(WHH^T+W)_{ij}}$.
Convex NMF [9]	Objective function	$F(W,H;X)=\|X-XWH\|^2$,
	Parameters	1)$K=1$; 2)$C^+=2(X^TXH^T)^-, C^-=2(X^TXH^T)^+$; 3)$A^{1+}=2(X^TX)^+, A^{1-}=2(X^TX)^-, B^{1+}=HH^T, B^{1-}=0$.
	Rules	$W_{ij}=W_{ij}\times\frac{[(X^TXH^T)^++(X^TX)^-WHH^T]_{ij}}{[(X^TXH^T)^-+(X^TX)^+WHH^T]_{ij}}$.
Manifold NMF [6]	Objective function	$F(W,H;X)=\|X-WH\|^2+\lambda Tr(HLH^T)$,
	Parameters	1)$K=1$; 2)$C^+=0, C^-=2XH^T$; 3)$A^{1+}=2I, A^{1-}=0, B^{1+}=HH^T, B^{1-}=0$.
	Rules	$W_{ij}=W_{ij}\times\frac{(XH^T)_{ij}}{(WHH^T)_{ij}}$.
Tensor based MNGE [43]	*Refer to the original paper [43], due to space limitation.*	

C. Optimization of Eq. (7)

Based on the solution platform, we briefly introduce the derivation of the update rules for solving Eq. (7).

We first derive an update rule for U^{b^m}, with other factors fixed at the current step. The objective function of Eq. (7) with respect to U^{b^m} can be simplified as,

$$F(U^{b^m})=\sum_m\|X^m_{(b)}-U^{b^m}Z^{u^m}\|^2 +\lambda Tr(\hat{Q}\hat{H}\hat{H}^T\hat{Q}^TE^{M(k-q)\times M(k-q)}), \quad (13)$$

where $X^m_{(b)}\in R^{d_b\times(d_{b+1}\times\ldots\times d_n\times d_1\times\ldots\times d_{b-1})}$ is obtained by flatting the tensor $\mathcal{X}^m$ along the b^{th} direction [42] and Z^{u^m} is a matrix in which the i^{th} row is $[(u_i^{p^m}\otimes)_{p=b+1}^{n-1}\otimes H_i{}^m(\otimes u_i^{p^m})_{p=1}^{b}]^T$.

The derivative of $F(U^{b^m})$ with respect to U^{b^m} is

$$\frac{\partial F}{\partial U^{b^m}}=-2X^m_{(b)}Z^{u^m\,T}+2U^{b^m}Z^{u^m}Z^{u^m\,T}+2\lambda[O^{d_b\times q},S^{\hat{H}}], \quad (14)$$

with

$$S^{\hat{H}}=E^{d_b\times(k-q)}\sum_m\hat{Q}^m\hat{H}^m\hat{H}^{m^T}(\prod_{p\neq b}\hat{Q}^{p^m})^T, \quad (15)$$

where $O^{d_b\times q}\in R^{d_b\times q}$ is an all-zero matrix and $E^{d_b\times(k-q)}\in R^{d_b\times(k-q)}$ is an all-one matrix.

Following [24], we can obtain the update rule for U^{b^m},

$$(U^{b^m})_{ij}\leftarrow U^{b^m}{}_{ij}\frac{(X^m_{(b)}Z^{u^m\,T})_{ij}}{(U^{b^m}Z^{u^m}Z^{u^m\,T}+\lambda[O^{d_b\times q},S^{\hat{H}}])_{ij}}. \quad (16)$$

After updating the matrix U^{b^m}, we normalize the vectors $u_i^{b^m}$ as conventionally [42], and consequently convey the norms to the coefficient matrix H^m to keep the objective value at the current step. Let H_i^m denote the i^{th} row of H^m, we have,

$$H_i^m\Leftarrow H_i^m\times\prod_b^{n-1}|u_i^{b^m}|,\forall\ i, \quad (17)$$

$$u_i^{b^m}\Leftarrow u_i^{b^m}/|u_i^{b^m}|,\forall\ i. \quad (18)$$

Then, we simplify the Eq. (7) with respect to H^m as

$$F(H^m)=\sum_m\|X^m_{(n)}-H^{m^T}Z^{h^m}\|^2 +\lambda Tr(\hat{H}\hat{H}^TE^{M(k-q)\times M(k-q)}) \quad (19)$$

where $X^m_{(n)}\in R^{d_n\times(\prod_{b=1}^{n-1}d_b)}$ and $H^m\in R^{k\times d_n}$. $Z^{h^m}\in R^{k\times(\prod_{b=1}^{n-1}d_b)}$ is a matrix, where the i^{th} row is $[u_i^1{}^m(\otimes u_i^{b^m})_{b=2}^{n-1}]^T$.

Algorithm 1 Nonnegative Tensor Co-Factorization

1: **Input:** M n-order tensors$\mathcal{X}^m = [\mathcal{X}_1^m, \ldots, \mathcal{X}_N^m], m = 1, .., M$;
2: **Initialization**, $U^{b^m} = rand(d_b, k), b = 1, .., n-1$, $H^m = rand(k, N)$;
3: For t=1: T_{max} (iteration body),
   1) Update $U^{b^m}, \forall b, m$ by Eq. (18);
   2) ${H_i}^m \Leftarrow {H_i}^m \times \prod_{b=1}^{n-1} |{u_i^b}^m|, \forall\ i$;
   3) ${u_i^b}^m \Leftarrow {u_i^b}^m / |{u_i^b}^m|, \forall\ i$;
   4) Update $H^m, \forall m$ by Eq. (23);
4: **Output:** Base matrices U^{b^m}; coefficient matrices H^m.

The partial derivative of $F(H^m)$ with respect to H^m is,

$$\frac{\partial F}{\partial H^m} = -2{Z_h}^m {X_{(n)}^m}^T + 2Z^{h^m} {Z^{h^m}}^T H^m + 2\lambda \begin{bmatrix} O^{q\times d_n} \\ E^{(k-q)\times(k-q)} \sum_m \hat{H}^m \end{bmatrix}, \tag{20}$$

where $E^{(k-q)\times(k-q)} \in R^{(k-q)\times(k-q)}$ is an all-one matrix.

Following [24], we can obtain the update rule for H^m,

$$H_{ij}^m \leftarrow H_{ij}^m \times \frac{({Z_h}^m {X_{(n)}^m}^T)_{ij}}{(Z^{h^m} {Z^{h^m}}^T H^m + \lambda \begin{bmatrix} O^{q\times d_n} \\ E^{(k-q)\times(k-q)} S^{\hat{H}} \end{bmatrix})_{ij}}. \tag{21}$$

Algorithm 1 summarizes the entire procedure of NTCoF. The update rules are performed iteratively to optimize the objective function in Eq. (7). NTCoF can be used for a number of data analysis tasks, under supervised or unsupervised settings. It is worthy highlighting that while the input data are provided in the form of matrix, NTCoF degenerates to two-dimensional matrix co-factorization (NTCoF-2D), which can also be solved by Algorithm 1.

IV. Classification via Co-Factorization

We discuss in this section how to infer discriminative representations, based on the proposed co-factorization algorithm. We consider three multi-class classification tasks, including face recognition, general image categorization, and multi-label image annotation.

Suppose multiple factorizations of the test data and the training data are obtained by NTCoF or other algorithms, one common solution to learning discriminative models [42] is simply based on the Nearest Neighbor (NN) classifier. First, NN is used to compute for the test sample the confidences of belonging to specific categories under each feature. Then, the confidences over all features are accumulated and the class that achieves the highest accumulated confidence is assigned to the test sample. However, this simply voting strategy does not consider the inter-representation correlation that may provide additional discriminative power.

In this work, we utilize the Multi-Task Joint Sparse Representation (MTJSR) [48] to fully take advantage of the factorization results of NTCoF. Methodologically, MTJSR belongs to the sparse learning methods called Multi-Task Joint Covariate Selection (MTJCS) [33], which can be regarded as a combined model of group Lasso and multi-task Lasso [51]. By penalizing the sum of ℓ_2-norms of the blocks of coefficients associated with each variable group across different tasks, similar sparsity patterns in all models are encouraged. Particularly, in this work, the sparse reconstruction of one test datum under one feature is referred to as one task. The goal of joint sparsity can be achieved by imposing $\ell_{1,2}$-norm constraint on the reconstruction coefficients.

We extract multiple different visual features, e.g. colors, gradients, for each image pixel, and collect multiple feature matrices as descriptors. These two-dimension matrixes are further concatenated along the 3^{th} dimension to form the 3^{th} tensors, served as the inputs of NTCoF. Once obtained the low-dimensional representations of these images and the desired basis by NTCoF, we can project the test image into the same low-dimensional subspaces using the method in [42], assuming the learnt basis are fixed.

Let $H^y = [{h^y}^1, \ldots, {h^y}^M]$ denote the low-dimensional representation of the test image, ${h^j}^m$ denotes the dimensionality-reduced representation of the j^{th} training image under the m^{th} feature, $\alpha_{[j]}^m$ denotes the reconstruction coefficient associated with the j^{th} sample, and $\alpha_j = [\alpha_{[j]}^1, \ldots, \alpha_{[j]}^M]$ denotes the coefficient vector associated with the j^{th} training sample under different features. The joint reconstructions of ${h^y}^m, m = 1..M$, over all the training images can be obtained by solving following program,

$$\arg\min_{\alpha^m} \frac{1}{2} \sum_{m=1}^{M} \|{h^y}^m - \sum_{j=1}^{N} {h^j}^m \alpha_{[j]}^m\|^2 + \beta \sum_{j=1}^{M} \|\alpha_j\| \tag{22}$$

where β is a tunable constant. Eq (22) is a convex but non-smooth quadratic function, and we choose to use the Accelerated Proximal Gradient (APG) method [48] to efficiently solve it.

A. App-I: Multi-Class Image Classification

We classify the test image based on how well it can be recovered from the reconstruction coefficients associated with all the training images of each category. Letting L denote the total number of categories and α^{m*} denote the optimal coefficients solved from Eq. (22), image classification is performed in favor of the category with the lowest total reconstruction error accumulated over all the M tasks,

$$\arg\min_{c\in\{1,2,\ldots,L\}} \sum_{m=1}^{M} \|{h^y}^m - \sum_{j=1}^{N} \delta(j,c) h_j^m \alpha^{m*}_{[j]}\|^2 \tag{23}$$

where $\delta(j, c)$ takes 1 when the j^{th} training sample contains the label c, or 0 otherwise.

B. App-II: Multi-Label Image Annotation

The task of multi-label image annotation is to predict the class labels for the test image, given a set of training images that are provided with label annotations. Let $z^j \in R^{L\times 1}$ indicate the annotated label vector of the j^{th} training image,

where the binary component $z^j(c)$ takes 1 when the j^{th} image contains the label c, or 0 otherwise. Given the coefficients α^{m*} solved from Eq. (22), we can propagate the label annotations of the training images to the test image as follows,

$$z^y = \sum_{m=1}^{M} \sum_{j} z^j \alpha^{m*}_{[j]}. \quad (24)$$

The desired label vector z^y associates each potential label with one confidence. To obtain the final annotation for the test image, we can simple select a fixed number of top-ranked labels, or select the labels scored larger than a threshold.

V. Experiments

In this section, we evaluate the proposed nonnegative tensor co-factorization (NTCoF) algorithm for three image problems, including face recognition, multi-class image categorization and multi-label image annotation, and compare it to respective popular algorithms on publicly available image datasets.

A. Exp-I: Face Recognition

We evaluate the proposed NTCoF for face recognition on two databases, Yale,[1] and CMU PIE [39]. **Yale** contains 165 grayscale images of 15 subjects. There are 11 images per subject, and we use 5 images for training and the rest 6 images for testing. **CMU PIE** database contains 41,368 images of 68 people, we use the subset selected in [7]. It contains 170 images for each of the 68 subjects. Among the total of 11,560 images, 20 images per subject are randomly selected as the training set and the rest images are used for testing. We crop the face regions and resize the cropped images to be 32×32 pixels.

We extract three visual features for each image pixel. The first one is the pixel intensity, which is normalized to be within $[0, 1]$. The second one is the sum of square of the gradient magnitudes on vertical and horizontal directions. The third one uses Local Binary Pattern (LBP) feature [1] and describes each pixel as one of 59 8-bit patterns. The pattern is extracted for each pixel from the surrounding window of 8×8 pixels. In order to make the binary patterns of the LBP feature comparable with each other, we transform the patterns to one-dimensional decimals while preserving the pair-wise similarities between patterns measured by the Hamming metric. Formally, let p_i, p_j denote two local binary patterns, $\mathcal{H}(p_i, p_j)$ denote the Hamming distance between patterns p_i and p_j. Letting s_i, s_j denote the desired values, the related optimization has following form,

$$\min_{\{s_i\}} \sum_{i \neq j} \|s_i - s_j\|^2 \exp\{-\mathcal{H}(p_i, p_j)\}, \quad s.t. \ \sum_i s_i^2 = 1, \quad (25)$$

which is a typical embedding problem and thus can be efficiently solved by the Laplacian Eigenmap method [2].

[1] http://cvc.yale.edu/projects/yalefaces/yalefaces.html

We evaluate following algorithms for comparisons: 1) principal component analysis (*PCA*) [40]; 2) nonnegative matrix factorization (*NMF*) [24]; 3) nonnegative tensor factorization (*NTF*) [37]; and 4) multiple feature fusion via canonical correlation analysis (*mCCA*) [15]. For PCA or NMF, we first apply it under every single feature to obtain the dimensionality-reduction representations of training samples independently. Then, NN classifier (augmented with the voting schema as introduced in Section IV) or MTJSR classifier is utilized to estimate the category label for the testing sample. For these two algorithms, each image is described as a histogram of quantized intensities, a histogram of quantized gradient magnitudes, or a histogram of binary patterns, and the feature dimensions are 256, 64 or 59 respectively. Differently, NTF and mCCA can directly handle multiple feature representations. For NTF, we describe each image under one feature as a feature matrix, and concatenate all feature matrices under one feature along the 3^{th}-dimension to form one 3-order tensor. These tensors are further concatenated to form a 4-order tensor as the algorithm input. For mCCA, each image is described as above three feature histograms. In addition, we implement two variants of the proposed NTCoF method. 5) *NTCoF-3D*, that describes each image under a feature as a two-dimension feature matrix (as introduced in Section IV). 6) *NTCoF-2D*, that describes each image under a feature as one histogram.

For all algorithms, the dimension of the subspace (k) is tuned within $\{7^2, \ldots, 11^2, 12^2\}$. for NTCoF-2D and NTCoF-3D, the parameter q is fixed as $q = 0.6 \times k$. We optimize the parameter λ for NTCoF, the parameter β for MTJSR, and the subspace dimension for all algorithms using the 10-fold cross validation procedure on the training set.

We report the mean accuracy of recognition, i.e. the percentage of agreements between the groundtruth classes and the predicted classes. We also calculated the standard derivations of those recognition results calculated from ten random splits of the datasets.

Fig. 1 shows the convergence curves of NTCoF-3D and NTCoF-2D algorithms on the CMU PIE dataset. The dimension of desired subspace is set to be $k = 100$. From the curves, one can observe that both algorithms will converge after about 2000 iterations. We implement the algorithms using MATLAB 2008a and conduct the experiments on a computer with Intel(R) Core(TM)2 Duo 2.66GHz CPU and 8GB RAM.

Tables II and III report the quantitative comparisons of various face recognition algorithms on CMU PIE and YALE datasets. From these results, we could obtain following observations. 1) The proposed NTCoF-2D and NTCoF-3D algorithms usually achieve higher accuracies as compared to various baseline algorithms over two datasets, while using either NN or MTJSR classifiers. Particularly, our method is much better than mCCA [15] although both utilize multi-modal strategy. The advantages of NTCoF over mCCA come from the nonnegative constraints and the proposed data factorization formulation in the setting of muti-task learning. 2) The adopted MTJSR classifier outperforms the traditional NN classifier consistently in the previous work [42] and [44]. 3) The algorithms NTCoF-2D and NTCoF-3D, that utilizes three types of features, clearly outperform the algorithms that

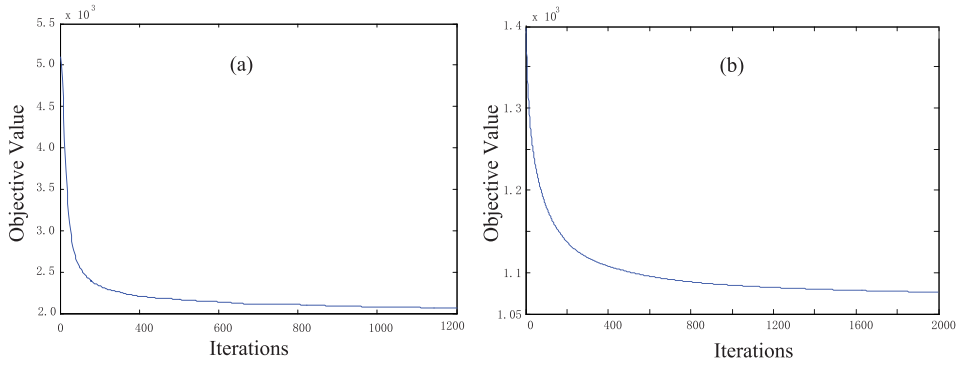


Fig. 1. Convergence curves of (a) NTCoF-2D and (b) NTCoF-3D. X-direction: iterations; Y-direction: objective function values of two methods (see Eq. 7). The data used here is from the CMU PIE dataset.

TABLE II

FACE RECOGNITION ACCURACIES ON CMU PIE DATASET. THE FIRST COLUMN INDICATES THE ALGORITHM TO EVALUATE AND THE REMAINING COLUMNS INDICATE THE AVERAGE ACCURACIES (%) AND STANDARD DEVIATIONS (%, IN THE PARENTHESES) OF THE RESULTS OBTAINED FROM TEN RANDOM SPLITS OF THE DATASETS

Features	Intensity+Gradient		Gradient+LBP		Intensity+LBP		Intensity+Gradient+LBP	
Classifiers	NN	MTJSR	NN	MTJSR	NN	MTJSR	NN	MTJSR
PCA [41]	80.49(±1.38)	81.39(±1.47)	81.23(±1.03)	82.15(±1.51)	80.18(±1.21)	82.43(±1.69)	82.31(±0.84)	84.57(±1.22)
NMF [24]	82.18(±2.49)	84.68(±2.26)	83.47(±2.13)	84.35(±2.57)	83.31(±2.16)	85.26(±2.64)	85.48(±0.97)	86.72(±1.04)
mCCA [15]	83.85(±2.53)	84.26(±2.11)	84.17(±2.14)	85.19(±2.39)	83.72(±2.58)	84.68(±2.71)	85.68(±1.47)	86.33(±2.48)
NTF [38]	85.93(±1.54)	86.29(±2.11)	86.48(±2.09)	87.41(±2.01)	86.48(±2.09)	87.41(±2.01)	87.53(±0.79)	88.39(±0.96)
NTCoF-2D	87.54(±2.33)	**89.54**(±2.30)	85.19(±2.09)	88.31(±2.38)	87.68(±2.73)	88.27(±2.59)	88.13(±1.32)	**89.75**(±1.42)
NTCoF-3D	86.22(±2.43)	86.49(±1.52)	86.51(±2.23)	**89.56**(±1.63)	87.35(±2.01)	**90.83**(±1.60)	88.77(±0.53)	89.32(±0.24)

TABLE III

FACE RECOGNITION ACCURACIES ON YALE DATASET

Features	Intensity+Gradient		Gradient+LBP		Intensity+LBP		Intensity+Gradient+LBP	
Classifiers	NN	MTJSR	NN	MTJSR	NN	MTJSR	NN	MTJSR
PCA [41]	53.25(±2.45)	55.47(±2.16)	54.19(±2.37)	55.98(±2.46)	54.37(±2.97)	56.10(±2.58)	57.15(±1.68)	58.36(±1.74)
NMF [24]	57.47(±3.37)	59.32(±3.27)	59.23(±3.45)	60.15(±3.18)	58.15(±3.62)	61.24(±3.47)	60.38(±2.19)	60.97(±1.69)
mCCA [15]	57.35(±2.49)	58.92(±2.03)	58.47(±2.34)	59.90(±2.13)	58.85(±2.61)	60.12(±2.98)	61.77(±1.31)	62.49(±2.17)
NTF [38]	61.52(±2.60)	62.87(±2.05)	62.47(±2.75)	62.58(±2.53)	61.28(±2.38)	62.57(±2.46)	63.02(±1.12)	63.28(±1.31)
NTCoF-2D	64.25(±2.11)	64.98(±1.13)	64.08(±2.27)	63.18(±1.92)	63.65(±2.41)	65.52(±2.00)	65.13(±1.11)	66.32(±1.34)
NTCoF-3D	65.77(±1.70)	**67.59**(±1.80)	65.24(±1.85)	**67.32**(±1.59)	65.19(±1.72)	**67.67**(±1.62)	65.81(±2.34)	**68.23**(±2.10)

utilizes only two features. This is due to the fact that utilizing more complementary features could improve the robustness of face recognition. We obtain consistent observations on the benefits of feature combination, which is however not consistent in [15], partially because we use the evaluation strategy of multiple splits of the datasets that will reduce the effects of uncertainties in each evaluation routine. These comparisons well demonstrate the effectiveness of our proposed co-factorization algorithm.

B. Exp-II: Multi-Class Image Categorization

In this subsection, we apply the proposed NTCoF-2D based classifiers for multi-class image categorization, and compare them with various Multiple Kernel Learning (MKL) methods [16], [32], [41] on Oxford Flower Datasets [32] and Caltech 101 datasets [25]. As reported in the past literature, MKL methods can achieve the state-of-the-art algorithms on these datasets.

The Oxford Flower dataset [32] consists of 8,189 images divided into 102 flower categories. Each category consists of 40-250 images. The categorization is carried out based on four features, HSV, HOG, SIFTint and SIFTbdy. The dataset is divided into a training set, a validation set and a testing set. The training set and validation set each consist of 10 images per category. The test set consists of the remaining 6,149 images (minimum 20 per class). A predefined training/validation/test split and the above features are publicly available on the database website.[2] We use the predefined splits as aforementioned for training and parameter selection. The ten-fold cross-validation procedure is conducted on the validation set. Accuracy is first measured per class and then averaged over all categories. For comparisons, we implement the NTF + NN and NTF + MTJSR algorithms introduced in Exp-I. The accuracies by the proposed NTCoF-2D + MTJSR algorithm and the baselines methods are reported in Table IV. We also list the results of the MKL method [32] for comparisons. We can observe that our algorithms are slightly better than the MKL method and much better than two baselines.

The Caltech101 data set [25] contains images of 101 categories of objects plus a background class. Following the standard experimental protocol [25], 15 images per category

[2]http://www.robots.ox.ac.uk/∼vgg/data/flowers/102/index.html

TABLE IV

ACCURACIES COMPARISON ON THE OXFORD FLOWERS-102 DATASET

Algorithms	Accuracy(%)
NTF+NN	60.8
NTF+MTJSR	67.3
MKL [33]	72.8
NTCoF+MTJSR	73.3

TABLE V

ACCURACIES COMPARISON ON THE CALTECH 101 DATASET

Algorithms	Accuracy(%)
NTF+NN	52.6(±0.9)
NTF+MTJSR	54.8(±1.1)
MKL [42]	70.0(±1.0)
LPBoost [16]	70.7(±0.4)
NTCoF+MTJSR	71.5(±0.6)

are selected for training and 15 images are selected for testing. Evaluations includes all 102 classes averaged over three random training/test splits, and the performance is measured as the mean accuracy per class. We extract from images four features, including geometric Blur [3], Phow-gray [5], color [23] and SSIM [38], among which the later three are represented in spatial pyramid with two levels. Table V reports the accuracies of various methods and the results from past literature [16] and [41]. Again, we can observe that our algorithms perform comparably to the state-of-the-art results achieved by MKL.

C. Exp-III: Multi-Label Image Annotation

The public dataset COREL5K [13] is used for this experiment. There are 5,000 images from 374 class labels. We use the standard partition strategy as in [13], 4,500 images are used for training and the rest images are used for testing.

We apply the proposed NTCoF-2D method for multi-label image annotation and compare it with various popular algorithms, including the co-occurrence model (co-occ) [30], the machine translation model (MT) [12], the cross-media relevance model (CMRM) [20], the continuous relevance model (CRM) [22], CRM with rectangular regions as input (CRM-Rect) [14], the multiple bernoulli relevance model (MBRM) [14] and the supervised multiclass labeling model (SML) [8]. For SML, we use the results corresponding to the best parameters in [8]. All the results of above baselines are directly from the original papers.

In implementation, we extract both global and local features commonly used for image categorization. The global image descriptor used here is the GIST feature [34], and the local descriptors include the SIFT feature [29] as well as the robust hue descriptor [43], extracted densely on a multi-scale grid or for Harris-Laplacian interest points. Each local feature descriptor is first extracted and then quantized using K-Means on training samples. Images are then represented as a Bag-of-Word histogram. These results in 5 distinct descriptors, namely one Gist descriptor and four bag-of-features descriptors ("Dense"+"SIFT", "Harris"+"SIFT", "Dense"+"Hue", "Harris"+"Hue"). Each descriptor is normalized to be a ℓ_2-norm unit and preprocessed by Principal Component Analysis (PCA) so the feature dimension is reduced to

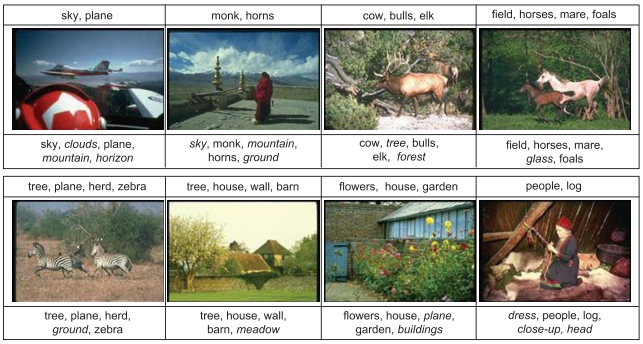


Fig. 2. Image annotation results on COREL5K dataset [13]. For each image, the labels predicted by NTCoF-2D algorithm are shown at the bottom line, the ground-truth labels are shown at the top line.

be 256. In order to evaluate the benefits brought by different feature combinations, we implement two versions of NTCoF-2D for this task: 1) NTCoF-2D (M=3), that uses GIST feature, "Dense"+"SIFT" feature and "Harris"+"SIFT" feature, and 2) NTCoF-2D (M = 5) that uses all five features. In the low-dimensional feature space obtained by NTCoF-2D, we utilize the MTJSR classifier or the multi-label K-Nearest Neighborhood (MLKNN) [50] classifier to propagate the image labels annotated for training images to the test images. We choose to use NTCoF-2D, rather than NTCoF-3D, since we try to utilize the same image descriptors (all in the form of vectors) as in other multi-label image annotations algorithms.

There are several free parameters, including the desired dimension of subspace k and the parameters λ for NTCoF, the tuning parameter β for MTJSR, and the number of neighbors K for MLKNN classifier. We choose the parameter values using a 10-fold cross validation procedure on the training set. For this dataset, they are fixed to be $k = 64$, $\lambda = 0.001$, $\beta = 0.03$, $K = 20$.

We use two standard metrics, *precision* and *recall* rates. As in previous works [8], [14], [22], we first compute the precision or recall for each label and then average over all labels. For each algorithm, we list the number of labels with nonzero recalls, which provides an indication of how many labels the system has effectively learned. All comparisons are conducted for the 260 labels appeared in test set. In addition, we also evaluate the top 49 annotations to make a direct comparison with the works in [12], [14], [20], and [22].

Table VI reports the comparison results, where the winner of each comparison term (split by double vertical lines) is indicated with bold font. For CO-occ [30] and SML [8], there are no statistics for the top 49 keywords in the corresponding papers. From the results, we can draw following observations. First, NTCoF-2D achieves the best performances among all the evaluated algorithms. Particularly, we obtain a gain of 3 percents in terms of recall rate and 5 percents in terms of precision rate, as compared to SML [8], which is one of the most popular and effective image annotation algorithms. Second, the algorithm NTCoF-2D (M = 5) that uses five types of feature descriptors clearly outperform the algorithm NTCoF-2D (M = 3) that uses three types of features in terms of both recall and precision rates while utilizing NN or MTJSR

TABLE VI
ACCURACIES COMPARISON OF VARIOUS IMAGE ANNOTATION ALGORITHMS ON THE COREL5K DATASET

Algorithms	#Words (Recall> 0)		Results on all 260 words				Results on 49 best words			
			Recall		Precision		Recall		Precision	
Co-occ [31]	19		0.02		0.03					
MT [12]	49		0.04		0.06		0.34		0.20	
CMRM [20]	66		0.09		0.10		0.48		0.40	
CRM [22]	107		0.19		0.16		0.70		0.59	
CRM-Rect [14]	119		0.23		0.22		0.75		0.72	
MBRM [14]	122		0.25		0.24		0.78		0.74	
SML [8]	137		0.29		0.23					
–	MLKNN	MTJSR	MLKNN	MTJSR	MLKNN	MTJSR	MLKNN	MTJSR	MLKNN	MTJSR
NTCoF-2D (M=3)	129	137	0.22	0.27	0.20	0.23	0.72	0.76	0.70	0.76
NTCoF-2D (M=5)	135	**142**	0.25	**0.31**	0.24	**0.28**	0.76	**0.85**	0.75	**0.80**

classifier. Third, integrating MTJSR classifier with NTCoF-2D achieve better annotation results than those by NN classifier, which well demonstrates the effectiveness of the multi-task joint sparse representation formulation.

Fig. 2 shows several exemplar annotations (at the bottom of each image) produced by NTCoF-2D+MTJSR. Each image contains at least one mismatched label compared with the ground-truth labels (at the top of the image). The images with completely matched annotations are not listed. We can observe that the labels estimated by our algorithm but not contained in the ground-truth annotations are still frequently plausible.

VI. CONCLUSIONS

In this paper, we proposed a novel nonnegative data factorization algorithm, nonnegative tensor co-factorization (NTCoF), for multi-modal learning problems. A quadratic block-wise convex function was defined and an efficient multiplicative update method was developed. NTCoF provides a general framework for utilizing multiple representations to obtain discriminative dimensionality-reduced representations. We applied NTCoF for face recognition, multi-class image categorization and multi-label image annotation, and obtained superiorities over other subspace learning methods or the state-of-the-art methods on several public benchmark databases.

APPENDIX

A. Preliminaries

We first introduce the concept of auxiliary function and the lemmas which shall be used for the deduction of the unified solution.

Definition : Function $G(A, A')$ is an auxiliary function for function $F(A)$ if the following conditions are satisfied:

$$G(A, A') \geq F(A), \quad G(A, A) = F(A). \forall A, A' \tag{26}$$

From the above definition, we have the following lemma with proof omitted [24].

Lemma 3.1: If G is an auxiliary function, then F is non-increasing under the update

$$A^{t+1} = \arg\min_{A} G(A, A^t), \tag{27}$$

where t means the t^{th} iteration.

B. Convergence of the Update Rule Eq. (14)

Letting $U = U^b$, we denote F_{ij} as the part of $F(U)$ relevant to U_{ij}, and we have,

$$F'_{ij}(U) = \sum_{l=1}^{K} (A^l U B^l + C)_{ij}, \tag{28}$$

$$F''_{ij}(U) = \sum_{l=1}^{K} (A^l)_{ii} (B^l)_{jj}. \tag{29}$$

The auxiliary function of F_{ij} is then designed as:

$$G(U_{ij}, U^t_{ij}) = F_{ij}(U^t_{ij}) + F'_{ij}(U^t_{ij})(U_{ij} - U^t_{ij}) + \frac{(\sum_{l=1}^{K} (A^{l+} U^t B^{l+} + A^{l-} U^t B^{l-}) + C^+)_{ij}}{2U^t_{ij}} (U_{ij} - U^t_{ij})^2. \tag{30}$$

Lemma A.1: $G(U_{ij}, U^t_{ij})$ is the auxiliary function of F_{ij}.

Proof: Since $G(U_{ij}, U_{ij}) = F_{ij}(U_{ij})$, we need only to show that $G(U_{ij}, U^t_{ij}) \geq F(U_{ij})$.

First, we can obtain the Taylor series expansion of F_{ij}, which is quadratic with respect to U_{ij}, as follows,

$$F_{ij}(U_{ij}) = F_{ij}(U^t_{ij}) + F'_{ij}(U^t_{ij})(U_{ij} - U^t_{ij}) + \frac{1}{2} F''_{ij}(U^t_{ij})(U_{ij} - U^t_{ij})^2. \tag{31}$$

Then, since,

$$(A^{l+} U^t B^{l+} + A^{l-} U^t B^{l-})_{ij} \geq U^t_{ij} (A^l)_{aa} (B^l)_{bb}, \tag{32}$$

we have the following inequality

$$\frac{(\sum_{l=1}^{K} (A^{l+} U^t B^{l+} + A^{l-} U^t B^{l-}) + C^+)_{ij}}{U^t_{ij}} \geq \sum_{l=1}^{K} (A^l)_{ii} (B^l)_{jj}$$

Thus, $G(U_{ij}, U^t_{ij}) \geq F_{ij}(U_{ij})$ holds. ■

Lemma A.2: Eqn. (12) could be obtained by minimizing the auxiliary function $G(U_{ij}, U^t_{ij})$ with respect to U_{ij}.

Proof: Let $\frac{\partial G(U_{ij}, U^t_{ij})}{\partial U_{ij}} = 0$, we have,

$$0 = \frac{(\sum_{l=1}^{K} (A^{l+} U^t B^{l+} + A^{l-} U^t B^{l-}) + C^+)_{ij}}{U^t_{ij}} (U_{ij} - U^t_{ij}) + F'_{ij}(U^t_{ij}).$$

Then we can obtain the iterative update rule for U as,

$$U_{ij}^{t+1} \leftarrow U_{ij}^{t} \times \frac{[\sum_{l=1}^{K}(A^{l+}U^{t}B^{l-} + A^{l-}U^{t}B^{l+}) + C^{-}]_{ij}}{[\sum_{l=1}^{K}(A^{l+}U^{t}B_{l+} + A^{l-}U^{t}B^{l-}) + C^{+}]_{ij}},$$

and the lemma is proved. ■

ACKNOWLEDGMENT

This work was originally done when X. Liu worked in National University of Singapore and was further studied when he worked in Korea University.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikȧinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[2] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. ECCV*, 2002, pp. 97–112.

[3] A. Berg, T. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences," in *Proc. IEEE Int. Conf. CVPR*, 2005, pp. 26–33.

[4] M. Berry, M. Browne, A. Langville, P. Pauca, and R. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Comput. Statist. Data Anal.*, vol. 52, no. 1, pp. 155–173, Jun. 2007.

[5] A. Bosch, A. Zisserman, and X. Munoz, "Image classifcation using random forests and ferns," in *Proc. IEEE ICCV*, Oct. 2007, pp. 1–8.

[6] D. Cai, X. He, X. Hu, and J. Han, "Non-negative matrix factorization on manifold," in *Proc. 8th IEEE ICDM*, Dec. 2008, pp. 63–72.

[7] D. Cai, X. He, Y. Hu, J. Han, and T. Huang, "Learning a spatially smooth subspace for face recognition," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2007, pp. 1–7.

[8] G. Carneiro, A. Chan, and P. Moreno, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, Mar. 2007.

[9] D. Chris, T. Li, and M. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.

[10] D. Chris, T. Li, P. Wei, and P. Haesun, "Orthogonal nonnegative matrix tri-factorization for clustering," in *Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2006, pp. 126–135.

[11] M. Chu, F. Diele, R. Plemmons, and S. Ragni. (2004). Optimality, computation and interpretation of nonnegative matrix factorizations. *SIAM J. Matrix Anal.* [Online]., pp. 4–8030. Available: http://citeseerx.ist.psu.edu/ viewdoc/summary?doi=10.1.1.61.5758
M. Chu, F. Diele, R. Plemmons, and S. Ragni, "Optimality, computation and interpretation of nonnegative matrix factorizations," in *Proc. SIAM J. Matrix Anal.*, 2004, pp. 4––8030. [Online]. Available: citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.5758

[12] P. Duygulu and K. Barnard, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. 7th ECCV*, 2002, pp. 97–112.

[13] P. Duygulu, K. Barnard, J. Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. ECCV*, 2002, pp. 97–112.

[14] S. Feng, R. Manamatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jul. 2004, pp. II-1002–II-1009.

[15] Y. Fu, L. Cao, G. Guo, and T. Huang, "Multiple feature fusion by subspace learning," in *Proc. CIVR*, 2008, pp. 127–134.

[16] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Oct. 2009, pp. 221–228.

[17] N. Gillis and F. Glineur, "Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization," *Neural Comput.*, vol. 24, no. 4, pp. 1085–1105, 2012.

[18] S. Han, X. Liao, and L. Carin, "Cross-domain multitask learning with latent probit models," in *Proc. IEEE ICML*, Jun. 2012, pp. 1463–1470.

[19] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 312–377, 1936.

[20] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retr.*, 2003, pp. 119–126.

[21] C. Lampert, "Detecting objects in large image collections and videos by efficient subimage retrieval," in *Proc. IEEE 12th Conf. Comput. Vis.*, Oct. 2009, pp. 987–994.

[22] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2003.

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

[24] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Aug. 1999.

[25] F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. CVPR Workshop Generat.-Model Based Vis.*, Jun. 2004, p. 178.

[26] S. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2001, pp. 207–212.

[27] C. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 10, pp. 2756–2779, 2007.

[28] X. Liu, S. Yan, J. Yan, and H. Jin, "Unified solution to nonnegative data factorization problems," in *Proc. 9th IEEE ICDM*, Dec. 2009, pp. 307–316.

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

[30] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *Proc. First Int. Workshop Multimedia Intell. Storage Retrieval Manag. (MISRM)*, Oct. 1999.

[31] Y. Mu, J. Dong, X. Yuan, and S. Yan, "Accelerated low-rank visual recovery by random projection," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 2609–2616.

[32] M. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. ICVGIP*, Dec. 2008, pp. 722–729.

[33] G. Obozinski, B. Taskar, and M. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *J. Statist. Comput.*, vol. 20, no. 2, pp. 231–252, 2009.

[34] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[35] P. Paatero, "The multilinear engine—A table-driven, least squares program for solving multilinear problems," *J. Comput. Graph. Statist.*, vol. 8, no. 4, pp. 854–888, 1999.

[36] F. Sha, Y. Lin, L. Saul, and D. Lee,"Multiplicative updates for nonnegative quadratic programming," *Neural Comput.*, vol. 19, no. 8, pp. 2004–2031, 2007.

[37] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Proc. 22nd ICML*, 2005, pp. 792–799.

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

[39] S. B. Terence Sim 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.

[40] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[41] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. IEEE 11th ICCV*, Oct. 2007, pp. 1–8.

[42] C. Wang, Z. Song, S. Yan, L. Zhang, and H. Zhang, "Multiplicative nonnegative greph embedding," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2009, pp. 389–396.

[43] J. Weijer and C. Schmid, "Coloring local feature extraction," in *Proc. 9th ECCV*, 2006, pp. 334–348.

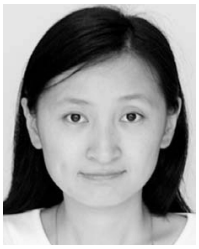
[44] S. Yan, D. Xu, B. Zhang, Q. Yang, H. Zhang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[45] J. Yang, S. Yan, Y. Fu, X. Li, and T. Huang, "Non-negative graph embedding," in *Proc. CVPR*, Jun. 2008, pp. 1–8.

[46] Y. Yang, Z. Ma, A. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 661–669, Apr. 2013.

[47] Z. Yang, T. Hao, O. Dikmen, X. Chen, and E. Oja, "Clustering by nonnegative matrix factorizationusing graph random walk," In *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran & Associates Inc., 2012.
[48] X. Yuan and S. Yan, "Visual classification with multi-task joint sparse representation," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2010, pp. 3493–3500.
[49] Z. Yuan and E. Oja, "Projective nonnegative matrix factorization for image compression and feature extraction," in *Proc. Scandin. Conf. Image Anal.*, 2005, pp. 333–342.
[50] M. Zhang and Z. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007.
[51] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2567, Nov. 2006.

Xiaobai Liu is currently a Post-Doctoral Research Scholar with the Department of Statistics, University of California at Los Angeles, Los Angeles, CA, USA. He received the Ph.D. degree from the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China, in 2011. Before that, he was with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, as the Research Associate for Prof. S. Yan. He was also a Research Associate with the Lotus Hill Institute, Pearl River Delta, China, from 2007 to 2008. He has authored more than 30 research papers over a series of research topics, and now his research interests fall on scene grammar for 3D parsing, commonsense knowledge modeling and reasoning, and typical computer vision problems.

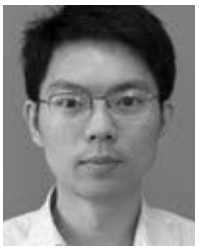


Qian Xu received the B.S. degree from the School of Science, Beihang University, Beijing, China, in 2006, and the M.S. degree from the Department of Mathematics and Statistics, San Diego State University, San Diego, CA, USA, in 2011, where she is currently pursuing the Ph.D. degree with the Department of Mathematics and Statistics. Her research interest falls in the typical nonparametric models and their applications in image data.

Shuicheng Yan is currently an Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and the Founding Lead of the Learning and Vision Research Group.

His research areas include computer vision, multimedia, and machine learning, and he has authored and co-authored over 370 technical papers over a wide range of research topics, with the Google Scholar citation of more than 12 000 and an H-index of 47. He is listed as the ISI Highly-Cited Researcher of 2014.

Dr. Yan is an Associate Editor of the *Journal of Computer Vision and Image Understanding*, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *ACM Transactions on Intelligent Systems and Technology*, and has been serving as the Guest Editor of the special issues for the IEEE TRANSACTIONS ON MULTIMEDIA and the Computer Vision and Image Understanding. He was a recipient of the Best Paper Awards from ACM MM'13 (Best Paper and Best Student Paper), ACM MM'12 (demo), PCM'11, ACM MM'10, ICME'10, and ICIMCS'09, the winner prizes of the classification task in PASCAL VOC from 2010 to 2012, the winner prize of the segmentation task in PASCAL VOC in 2012, the honorable mention prize of the detection task in PASCAL VOC'10, the TCSVT Best Associate Editor Award in 2010, the Young Faculty Research Award in 2010, the Singapore Young Scientist Award in 2011, the NUS Young Researcher Award in 2012, and the co-author of the Best Student Paper Awards of PREMIA'09, PREMIA'11, and PREMIA'12. He has been the General/Program Co-Chair of MMM'13, PCM'13, and MM'15.



Gang Wang (M'08) is an Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, and a Research Scientist with the Advanced Digital Science Center, Singapore. He received the B.S. degree in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2005, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Champaign, IL, USA, in 2010. He was a recipient of the prestigious Harriett & Robert Perry Fellowship from 2009 to 2010, and the CS/AI Award at UIUC in 2009. His research interests include computer vision and machine learning. In particular, he is focusing on object recognition, scene analysis, and large-scale machine learning.

Hai Jin (SM'06) is a Professor of Computer Science and Engineering with the Huazhong University of Science and Technology (HUST), Wuhan, China, where he is currently the Dean of the School of Computer Science and Technology. He received the Ph.D. degree in computer engineering from HUST in 1994. He was a recipient of the German Academic Exchange Service Fellowship in 1996 to visit the Technical University of Chemnitz, Chemnitz, Germany. He was with the University of Hong Kong, Hong Kong, from 1998 and 2000, and was a Visiting Scholar with the University of Southern California, Los Angeles, CA, USA, from 1999 and 2000. He was also a recipient of the Excellent Youth Award from the National Science Foundation of China in 2001. He is the Chief Scientist of ChinaGrid, the largest grid computing project in China.

He is a member of the Association for Computing Machinery. He is a member of the Grid Forum Steering Group. He has co-authored 15 books and published over 400 research papers. His research interests include computer architecture, virtualization technology, cluster computing and grid computing, peer-to-peer computing, network storage, and network security.

Dr. Jin is the Steering Committee Chair of the International Conference on Grid and Pervasive Computing and the Asia-Pacific Services Computing Conference. He is a Steering Committee Member of the IEEE/ACM International Symposium on Cluster Computing and the Grid, the IFIP International Conference on Network and Parallel Computing, the International Conference on Grid and Cooperative Computing, the International Conference on Autonomic and Trusted Computing, and the International Conference on Ubiquitous Intelligence and Computing.

Seong-Whan Lee (S'84–M'89–SM'96–F'10) is the Hyundai-Kia Motor Chair Professor with Korea University, Seoul, Korea, where he is the Head of the Department of Brain and Cognitive Engineering. He received the B.S. degree in computer science and statistics from Seoul National University, Seoul, in 1984, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 1986 and 1989, respectively. From 1989 to 1995, he was an Assistant Professor with the Department of Computer Science, Chungbuk National University, Cheongju, Korea. In 1995, he joined the faculty of the Department of Computer Science and Engineering at Korea University, where he is currently a Full Professor. In 2001, he was with the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA, as a Visiting Professor. He is a fellow of the International Association for Pattern Recognition and the Korean Academy of Science and Technology, and has served several professional societies as the Chairman/Governing Board Member. His research interests include pattern recognition, computer vision, and brain engineering. He has more than 300 publications and authored 10 books.