Transductive Nonnegative Matrix Tri-Factorization

XIAO TENG１,３, LONG LAN２,３, (Member, IEEE), XIANG ZHANG２,３, GUOHUA DONG１,３, AND ZHIGANG LUO１,３, (Member, IEEE)

１Science and Technology on Parallel and Distributed Processing, National University of Defense Technology, Changsha 410073, China
２Institute for Quantum Information and State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha 410073, China
３College of Computer, National University of Defense Technology, Changsha 410073, China
Corresponding author: Long Lan (long.lan@nudt.edu.cn)

ABSTRACT Nonnegative matrix factorization (NMF) decomposes a nonnegative matrix into the product of two lower-rank nonnegative matrices. Since NMF learns parts-based representation, it has been widely used as a feature learning component in many fields. However, standard NMF algorithms ignore the training labels as well as unlabeled data in the test domain. In this paper, we propose a transductive nonnegative matrix tri-factorization method (T-NMTF) to simultaneously exploit the label information of training examples and the statistical structure of features in the test domain. Different from standard NMF, nonnegative matrix tri-factorization (NMTF) decomposes a nonnegative matrix into the product of three lower-rank nonnegative matrices, and thus provides a flexible framework to transduce discriminative information of training examples to test examples. In particular, the proposed T-NMTF projects both training examples and test examples into a unified subspace, and expects the coefficients of training examples close to their label vectors. Since training examples and test examples are assumed to identically distributed, it is reasonable to expect the learned coefficients of test examples approximate their label vectors well. To estimate the T-NMTF parameters, we develop an efficient multiplicative update rule and prove its convergence. In addition, we propose a manifold regularized T-NMTF (MT-NMTF) algorithm that exploits the local geometry structure of the dataset to boost discriminant power. Experimental results on face recognition demonstrate the effectiveness of T-NMTF and MT-NMTF.

INDEX TERMS Nonnegative matrix factorization, nonnegative matrix tri-factorization, transductive learning.

I. INTRODUCTION
Nonnegative matrix factorization (NMF) [1] decomposes a nonnegative matrix into the product of two lower-rank nonnegative matrices. Due to the non-negativity constraints imposed on both factor matrices, NMF learns parts-based representation which has psychological and physiological evidence in human brain [2]. Moreover, NMF, Principal Component Analysis (PCA) [3], Singular Value Decomposition (SVD) [4] are powerful dimension reduction methods because the obtained factor matrices reduce the storage requirement and save the computational cost of subsequent processing. Based on them, Lai et al. [5] proposed a unified sparse learning framework for dimensionality reduction. In [6], a novel discriminative low-rank preserving projection (DLRPP) algorithm was introduced to learn an optimal projection matrix for data dimensionality reduction. In [7], a novel structured optimal graph based sparse feature extraction method was proposed for dimensionality reduction. Among them, NMF can enforce the resulting matrix factors to be nonnegative, which attracted a large amount of attention and has been widely used in many fields such as information retrieval [1], pattern recognition [8], data mining [9], and computational biology [10].

Most of the NMF methods focused on the 2-factor factorization. In this framework, both theoretical and algorithmic aspects of NMF have been extensively studied. On the theoretical side, Donoho and Stodden [11] studied the uniqueness of NMF solution and concluded based on convex duality that the NMF solution is unique unless one of the factor matrices...
is orthogonal or non-overlapping. Vavasis [12] proved that the exact NMF which exactly reconstructs the dataset is equivalent to a NP-hard problem. On the algorithmic side, Lee and Seung proposed an effective multiplicate updating rule (MUR) to solve the optimization problem in NMF in their seminal work [13]. To resolve the convergence issues in MUR, the alternating nonnegative least squares (ANLS) framework were introduced with good convergence properties [14], [15].

Nonnegative matrix tri-factorization (NMTF), which seeks a 3-factor decomposition, has become an emerging tool for co-clustering. NMTF was firstly proposed in [16] to co-cluster the rows and columns of an input data matrix. Under orthogonal constraint, NMTF was shown to have unique decomposition and can be efficiently solved by a new multiplicative updating algorithm [17], [18]. Due to its encouraging empirical results, it has been further investigated and extended for many applications. Wang et al. [19]–[21] decoupled the NMTF and studied its fast implementation by constraining the factor matrices to be cluster indicator. Chakraborty and Sycara [22] proposed the graph regularized NMTF for community detection to incorporate the social relations and user generated content in the decomposition. Li et al. [23] introduce the symmetric NMTF to explore the dual relation and the bilateral information between samples and features.

Since both NMF and NMTF are proposed for unsupervised learning, the decomposition does not take into account of possible label information in the training set and the unlabeled data in the test set. To make use of label information, a series of semi-supervised NMF methods were proposed [24]–[29]. Zafeiriou et al. [24] proposed discriminant NMF method (DNMF) by incorporating Fisher's criterion in NMF and applied it in frontal face verification. Guan et al. [25] proposed manifold regularized discriminative NMF(MD-NMF) to consider more effective margin-based discriminative information for subsequent data representation. To make use of unlabeled test data, several transductive NMF methods which decompose the train and test examples simultaneously were proposed [30]–[32]. Cho and Saul [30] proposed to use a pre-trained SVM classifier to learn the discriminative support vectors and apply them in the decomposition. Liu et al. [31] introduce a label matrix (obtain from the given labels) to constrain the decomposition, which cannot handle the case where only one label is given for each class. However, how to incorporate training labels and unlabelled test data in NMTF is still underexplored.

In this paper, we propose a new transductive nonnegative matrix tri-factorization method (T-NMTF) to simultaneously utilize the labels of training examples and exploit the statistical structure of test examples in the framework of NMTF. Different from previous transductive learning methods [33]–[35] which exploit available test examples to enhance the performance of discriminative classifiers such as SVM, the proposed T-NMTF explores test examples to learn better feature representations and perform classification in a joint framework. In particular, T-NMTF concatenates both training examples, i.e., \( V_t \), and test examples, i.e., \( V_d \), and decomposes the concatenated matrix, i.e., \( V = [V_t, V_d] \), into the product of three lower-rank nonnegative matrices, i.e., \( WDH \), where \( W \) spans the learned subspace and \( DH \) signifies the coefficients. Since the matrix is shared by all examples, it can be considered as a bridge to transduce label information from training examples to test examples. As we all know that the unsupervised NMF could exploit the statistical structure of the whole examples. The basis matrix represents the different centers of all samples while the coefficient matrix denotes the distribution of examples over these centers. As to T-NMTF, the test examples also engage the learning of basis matrix learning and thus help to shape the coefficient matrix from their statistical structure. To this end, T-NMTF constructs a label vector for each training example and expects its coefficient close to the label vector. Assuming that both training examples and test examples follow the same distribution, the learned coefficients of test examples approximate their label vectors well, and thus T-NMTF directly infers the labels of all test examples. To solve T-NMTF, we develop a multiplicative update rule with rigorous convergence analysis. Experimental results on face recognition validate the effectiveness of T-NMTF comparing with NMF, NMTF, and the representative transductive learning methods.

The remainder of this paper is organized as follows: Section 2 briefly reviews NMF and its variants. Section 3 explains T-NMTF in detail. And we compare T-NMTF with other semi-supervised NMF in Section 4. Section 5 validates effectiveness of the proposed methods by face clustering and Section 6 concludes this paper.

II. RELATED WORKS
A. REVIEW OF MORE NMF METHODS

NMF decomposes a nonnegative matrix into the product of two nonnegative matrices. The studies of NMF has two popular branches, the first is the uniqueness of decomposition, and the second is the NMF with label information. For the first aspect, as showed by many works that the composition itself is not unique, for examples, in the decomposition of \( V = WH \), for any nonnegative diagonal matrix \( D \), it can guarantee that the decomposition \( V = WDD^{-1}H \) also satisfies the nonnegativity as both \( WD \geq 0 \) and \( D^{-1}H \geq 0 \). To remove this uncertainty and drive the decomposition to the intend direction, many methods propose to impose task-oriented constraints on the decomposition and achieved better performances.

Ding et al. [16] studied the uniqueness of NMF and proposed the nonnegative matrix tri-factorization (NMTF), in their studies, to drive to a unique and intend decomposion, several deliberately designed constraints are imposed. Based on the NMTF, Chen et al. [36] proposed orthogonal NMTF for collaborative filtering and solved the sparsity and scalability of user-item matrix factorization very well. Ding et al. [16] introduce the bi-orthogonal 3-factor NMF to clustering and show its ability on simultaneously clustering
rows and columns of input data matrix. Chakraborty and Sycara [22] proposed the graph regularized NMTF for community detection to incorporate the social relations and user generated content in the decomposition. Wang et al. [20] decoupled the NMTF and studied its fast implementation by constraining the factor matrices to be cluster indicator. However, all these studied NMTF methods ignore the label information in the decomposition. Li et al. [37] proposed a NMTF to sentiment classification with lexical prior knowledge. Ma et al. [38] introduce the orthogonal NMTF to semi-supervised document clustering. As proved, with the prior knowledge or given label information, the NMTF can achieve better performances. To better utilize the label information in the NMF, many semi-supervised NMF are well studied, Zaferiou et al. [24] proposed discriminant NMF method (DNMF) by incorporating Fisher’s criterion in NMF and applied it in frontal face verification. Guan et al. [25] proposed manifold regularized discriminative NMF (MD-NMF) to consider more effective margin-based discriminative information for subsequent data representation. The discriminative NMF methods perform better than standard NMF; however, they did not exploit the statistical structure of test examples if they are available during training. Several transductive NMF are studied recently, Cho and Saul [30] proposed to use a pre-trained SVM classifier to learn the discriminative support vectors and apply them in the decomposition. Liu et al. [31] introduce a label matrix (obtain from the given labels) to constrain the decomposition. The works of [30] and [31] decompose the train and test examples simultaneously to improve the performance on testing. However, both Cho and Saul [30] and Liu et al. [31] fail to handle the case where only one label example is given of each class.

Ding et al. [16] studied the relationship between 2-factor and 3-factor matrix factorization in deep, and provide a systematic analysis of 3-factor NMF that unconstrained 3-factor NMF is equivalent to unconstrained 2-factor NMF, and only constrained 3-factor NMF make a difference. Chen et al. [36] and Ding et al. [16] introduced the orthogonal 3-factor NMF and show that the orthogonality in the decomposition bring better performances in both collaborative filtering document clustering. Chakraborty and Sycara [22] studied the graph regularized 3-factor NMF and concluded that the informative graph with additional social relations greatly help the community detection in social network. To acquire a more robust representations of NMF for more accurate measure and representation, a novel unsupervised Nonnegative Adaptive Feature Extraction (NAFE) algorithm was proposed in [39] by integrating the sparsity constrained nonnegative matrix factorization (NMF), representation learning, and adaptive reconstruction weight learning into a unified model. For the nonnegative matrix factorization with label information, Zaferiou et al. [24] proposed discriminant NMF (DNMF) to incorporate Fisher’s criterion in NMF. In specific, based on the given label, the within-class scatter and between-class scatter are constructed and force the final decomposition approximate the two scatters. Guan et al. [25] proposed manifold regularized discriminative NMF (MD-NMF) to preserve more effective margin-based discriminative information in NMF subspace. Choi et al. pre-trained a classifier with the labeled examples to learn the support vectors and proposed to use the support vectors to guide the decomposition. Liu et al. [31] constructed a label matrix with the given label information to constrain the NMF, in the decomposition, the label matrix enforced the coefficients of two same label examples to approximate to each other. Chen et al. [40] proposed to apply NMF to data clustering. Then Shao et al. [41] applied it to high dimensional data clustering, and Hu et al. [42] modified NMF to make it more stable. Kannan et al. [43] proposed to apply NMF to recommender system, Then Yu et al. [44] paralleled it. Nie et al. [45] proposed to apply NMF to missing data recovery. Recently, Pecli et al. [46] have provided a comparative study.

Concept Factorization (CF) is another variation of NMF by representing each cluster with a linear combination of data points and using a linear combination of the cluster centers to represent each data [47]. Zhang et al. [48] proposed a novel Robust Flexible Auto-weighted Local-coordinate Concept Factorization (RFA-LCF) for unsupervised clustering by integrating the robust flexible CF, robust sparse local-coordinate coding and adaptive weighting learning into a unified model. To improve the representation and clustering abilities, a Deep Self-representative Concept Factorization Network (DSCF-Net) framework was proposed in [49] by integrating the robust deep concept factorization, deep self-expressive representation and adaptive locality preserving feature learning into a unified framework. To improve the data representations by enhancing the robustness to outliers and noise in data, Ren et al. [50] proposed a joint robust factorization and projective dictionary learning (J-RFDL) model by performing the robust representation in a factorized compressed space. Similar to our work, Zhang et al. [51] proposed a joint label prediction based Robust Semi-Supervised Adaptive Concept Factorization (RS2ACF) framework recently, which utilized class information of labeled data and propagated it to unlabeled data by jointly learning an explicit label indicator for unlabeled data. However, the RS2ACF framework use a fixed label constrained matrix to guide the decomposition, that means, the label constrained matrix is given beforehand and not changed during the optimization. While our proposed T-NMTF relax the constraints and use the label indicator matrix instead.

B. TRANSDUCTIVE METHODS

Transductive learning methods [33]–[35] exploit available test examples to enhance the discriminative power of the learned model. Along this direction, Joachims [33] firstly proposed transductive SVM (TSVM) for binary classification. TSVM randomly assigns labels for test examples and iterates the following two steps to correct these labels: 1) training a classifier based on both training and test examples, and 2) finding a pair of positive and negative test examples that are most possibly wrongly labeled and
exchanging their labels. TSVM significantly outperforms SVM especially when the training examples are limited. Since TSVM corrects labels of test examples pair by pair, it fails when the numbers of positive and negative examples are imbalanced. Joachims [34] proposed spectral graph transducer (SGT) method that can be viewed as a transductive version of the k nearest neighbor (KNN) classifier. The SGT problem is relaxed to a convex formulation and solved globally optimally via the spectral method. However, TSVM and SGT are designed for binary classification, and thus they are unsuitable for multi-class cases. Liu et al. [35] proposed transductive component analysis (TCA) to learn a subspace where multi-class classification tasks can be conducted. TCA keeps both the smoothness property of test examples and discriminative information of training examples, and thus learns a powerful discriminant subspace based on only a few training examples. However, the complex parameter tuning procedure prohibits TCA from practical applications.

Furthermore, some novel transductive frameworks were proposed and achieved effectiveness in the classification task. Zhang et al. [52] proposed an enhanced semi-supervised classification approach termed Nonnegative Sparse Neighborhood Propagation (SparseNP) to ensure the outputted soft labels of points to be sufficiently sparse, discriminative, robust to noise and be probabilistic values. Jia et al. [53] proposed a new transductive label propagation method, termed Adaptive Neighborhood Propagation (Adaptive-NP) by joint L2,1-norm regularized sparse coding, for semi-supervised classification. In [54], a novel adaptive transductive label propagation approach was proposed by joint discriminative clustering on manifolds for representing and classifying high-dimensional data. In [55], to acquire a more accurate prediction in classification, the triple matrix recovery-based robust auto-weighted label propagation framework (ALP-TMR) was proposed by introducing a TMR mechanism to remove noise or mixed signs from the estimated soft labels and improve the robustness to noise and outliers in the steps of assigning weights and predicting the labels simultaneously.

III. BACKGROUND

A. NMF
The standard NMF minimizes the squared Euclidean distance between the examples \( V \in \mathbb{R}_{++}^{m \times n} \) and the product of two lower-rank matrices \( W \in \mathbb{R}_{++}^{r \times m} \) and \( H \in \mathbb{R}_{++}^{r \times n} \), i.e.,

\[
\min_{W \geq 0, H \geq 0} \| V - WH \|_F^2,
\]

where \( r \ll \min(m, n) \) denotes the reduced dimensionality and \( \| \cdot \|_F \) signifies the Frobenius norm. NMF performs poorly in some pattern recognition tasks because it completely ignores the labels of training examples.

B. NMTF
In contrast to NMF, nonnegative matrix tri-factorization (NMTF) decomposes a nonnegative matrix into the product of three lower-rank nonnegative matrices, i.e.,

\[
\min_{W \geq 0, D \geq 0, H \geq 0} \frac{1}{2} \| V - WDH \|_F^2,
\]

where \( W \in \mathbb{R}_{++}^{m \times r}, D \in \mathbb{R}_{++}^{r \times r} \) and \( H \in \mathbb{R}_{++}^{r \times n} \) are three lower-rank nonnegative matrices decomposed from the nonnegative matrix \( V \). \( W \) is the basis matrix, \( D \) denotes the middle matrix and label indicator matrix. In our method the product of \( D \) and \( H \) denote the coefficient matrix. Since the matrix \( D \) can absorb different scales of \( V \), \( W \) and \( H \), NMTF provides a more flexible way for data representation than NMF. However, NMTF shrinks to the standard NMF without any additional constraints on the factor matrices.

Ding et al. [16] proposed bi-orthogonal NMTF that enforces the columns of \( W \) and the rows of \( H \) to be orthogonal, i.e.,

\[
\min_{W \geq 0, D \geq 0, H \geq 0} \frac{1}{2} \| V - WDH \|_F^2, \text{ s.t.,} \quad W^TW = I, \quad HH^T = I,
\]

where \( I \) signifies the identity matrix. The orthogonality constraints over \( W \) and \( H \) prevent NMTF from shrinking to NMF and make bi-orthogonal NMTF a powerful tool for simultaneously clustering both rows and columns of \( V \).

IV. TRANSDUCTIVE NONNEGATIVE MATRIX TRI-FACTORIZATION

According to [33]–[35], transductive learning enhances the discriminant power of the learned model by exploiting test examples if they are available during training. In this paper, we propose transductive NMTF (T-NMTF) to jointly learn from both training examples and test examples.

Given \( l \) training examples arranged in columns of \( V_t \in \mathbb{R}_{++}^{m \times l} \) and \( u \) test examples arranged in columns of \( V_u \in \mathbb{R}_{++}^{m \times u} \), T-NMTF concatenates both training examples and test examples in a nonnegative matrix \( V = [V_t, V_u] = [\tilde{v}_1, \ldots, \tilde{v}_n] \in \mathbb{R}_{++}^{m \times n} \), where \( n = l + u \). Assuming both training and test examples belong to known \( c \) classes and there is at least one training example in each class, T-NMTF constructs a label vector \( \tilde{y} \in \mathbb{R}_+^c \) for each training example \( \tilde{v} \) as follows:

\[
\tilde{y}_i = \begin{cases} 
1, & \text{if } \tilde{v} \text{ is labeled by } i \\
0, & \text{otherwise}
\end{cases}
\]

Based on definition (4), it is easy to construct a label matrix \( Y_t \in \mathbb{R}_{++}^{c \times l} \) for training examples. T-NMTF aims to predict the label matrix \( Y_u \in \mathbb{R}_{++}^{c \times u} \) for test examples.

As mentioned above, NMTF provides a flexible framework for developing new data representation methods. T-NMTF follows the idea of NMTF and decomposes the concatenated examples into the product of three matrices, i.e.,

\[
V \approx WDH,
\]

where \( W \in \mathbb{R}_{++}^{m \times c} \) signifies the basis matrix, \( D \in \mathbb{R}_{++}^{c \times c} \) and \( H \in \mathbb{R}_{++}^{c \times n} \), and \( DH \) represents the coefficients of all examples. We set the reduced dimensionality to \( c \) to match
the dimensionality of the learned subspace and label vector for the sake of utilizing the labels of the training examples. Since both training and test examples share an identical \( W \), T-NMTF exploits the test examples when learning the subspace. According to Section 3.2, the NMTF model 2 will shrink to the standard NMF without any other constraints. Unlike bi-orthogonal NMTF [16], T-NMTF constrains the coefficients of training examples by using their labels. Based on the definition of the label vector, it is natural to believe that T-NMTF learns a discriminative data representation if coefficients of examples are exactly their label vectors. If labels of training examples are available, we have the following constraint

\[
DH_l = Y_l, \tag{6}
\]

where \( H_l \) is the first \( l \) columns of \( H \) and \( DH_l \) contains the coefficients of training examples. Assuming the remaining columns of \( H \) is \( H_u \), i.e., \( H = [H_l, H_u] \), then the coefficients of the test examples are \( DH_u \). Since both \( DH_l \) and \( DH_u \) share an identical matrix \( D \) which transduces discriminative information from training examples to test examples, T-NMTF significantly boosts the discriminant power of the coefficients of the test examples.

By incorporating (5) and (6), we have the objective of T-NMTF

\[
\min_{w \geq 0, d \geq 0, h \geq 0} \frac{1}{2} \|V - WDH\|^2_F, \text{s.t., } DH_l = Y_l \tag{7}
\]

Since the inequality and equality mix-constrained problem is not easy to solve, we rewrite eq. (7), based on the Lagrangian multiplier method [56], to the following inequality constrained problem

\[
\min_{w \geq 0, d \geq 0, h \geq 0} \frac{1}{2} \|V - WDH\|^2_F + \lambda \|DH_l - Y_l\|^2_F, \tag{8}
\]

where \( \lambda > 0 \) is the Lagrangian multiplier of the constraint (6). Since both training and test examples are assumed to be sampled from identical probabilistic distribution. The coefficients of training examples are equal to their label vectors, it is reasonable to expect that the learned coefficients of test examples equal their label vectors. Let \( C \) denote the coefficient of a test example, since the equality constraint (6) is approximately satisfied, \( C \) actually approximates its label vector and its entries reflect the possibility that it belongs to all classes. The larger an entry is, the greater the possibility that it belongs to the corresponding class. Therefore, T-NMTF infers the label vectors of test examples from their coefficients as follows:

\[
[Y_u]_{ij} = \begin{cases} 
1, & \text{if } i = \arg \max C_{u,ik} \\
0, & \text{otherwise},
\end{cases}
\tag{9}
\]

where \( Y_u \) is defined as (4), \( C_u = DH_u \) and \( X_{ij} \) signifies the \((i, j)\)-th element of \( X \). From (9), it is easy to predict labels for test examples.

### A. Multiplicative Update Rule

Although (8) is jointly non-convex with respect to \( W, D, \) and \( H \), it is separately convex with respect to each of them. Due to the separability of the squared Frobenius norm and the concatenation of \( V \), based on the Lagrangian multiplier method [56], we have the Lagrangian function of \( f(W, D, H) \) as

\[
\mathcal{L} = \frac{1}{2} \|V_l - WDH_l\|^2_F + \frac{1}{2} \|V_u - WDH_u\|^2_F + \lambda \|DH_l - Y_l\|^2_F \\
- \langle W, \varphi \rangle - \langle D, \phi \rangle - \langle H_l, \theta \rangle - \langle H_u, \psi \rangle \tag{10}
\]

where \( \varphi \geq 0, \phi \geq 0, \theta \geq 0 \) and \( \psi \geq 0 \) are the Lagrangian multipliers of constraints \( W \geq 0, D \geq 0, H_l \geq 0 \) and \( H_u \geq 0 \), respectively, and \( \langle \cdot, \cdot \rangle \) signifies the inner product.

Assuming \( W, D, H_l, H_u \) is a stationary point of the constrained optimization problem (8). According to the K.K.T. condition [56], the stationary point satisfies the primal conditions

\[
\begin{align*}
\frac{\partial \mathcal{L}}{\partial W} &= WDHH^T D^T - VH^T D^T - \varphi = 0 \\
\frac{\partial \mathcal{L}}{\partial D} &= W^T WDHH^T + \lambda DH_l H_l^T - W^T VH \\
&\quad - \lambda Y_l H_l^T - \phi = 0 \\
\frac{\partial \mathcal{L}}{\partial H_l} &= D^T W^T WDH_l + \lambda D^T DH_l^T - D^T W^T V_l \\
&\quad - \lambda D^T Y_l - \theta = 0 \\
\frac{\partial \mathcal{L}}{\partial H_u} &= D^T W^T WDH_u \\
&\quad - D^T W^T V_u - \psi = 0,
\end{align*}
\tag{11}
\]

and the slackness conditions

\[
\begin{align*}
W \odot \varphi &= 0 \\
D \odot \phi &= 0 \\
H_l \odot \theta &= 0 \\
H_u \odot \psi &= 0.
\end{align*}
\tag{12}
\]

By substituting (11) into (12) and using simple algebra, we have

\[
\begin{align*}
W &\odot (WDHH^T D^T) = W \odot (VH^T D^T) \\
D &\odot (W^T WDHH^T + \lambda DH_l H_l^T) \\
&= D \odot (W^T VH^T + \lambda Y_l H_l^T) \\
H_l &\odot (D^T W^T WDH_l + \lambda D^T DH_l^T) \\
&= H_l \odot (D^T W^T V_l + \lambda D^T Y_l) \\
H_u &\odot (D^T W^T WDH_u) = H_u \odot (D^T W^T V_u).
\end{align*}
\tag{13}
\]

From (13), we get the following multiplicative update rules (MUR):

\[
W \leftarrow W \odot \frac{VH^T D^T}{WDHH^T D^T}, \tag{14}
\]

\[
D \leftarrow D \odot \frac{W^T VH^T + \lambda Y_l H_l^T}{W^T WDHH^T + \lambda DH_l H_l^T}, \tag{15}
\]

\[
H_l \leftarrow H_l \odot \frac{D^T W^T V_l + \lambda D^T Y_l}{D^T W^T WDH_l + \lambda D^T DH_l D^T W^T V_l}. \tag{16}
\]
TNMTF alternatively updates \( W, D, \) and \( H \) until the objective value does not change.

The main time complexity of MUR is spent on (14), (15), and (16). Since (14) includes \( DH \) three times, the matrix multiplication \( DH \) can be computed in \( O(nc^2) \) before (14) and the product can be repeatedly utilized. The time complexity of (14) is \( O(nc^2 + mnc) \). The main time cost of (15) is caused by \( W^T VH^T, Y_lH_l^T, W^T WDHH^T \) and \( DH_lH_l^T \) whose time complexities are \( O(nc^2 + mnc), O(lc^2), O(mmc^2 + mc^2 + c^3) \) and \( O(lc^2 + c^3) \) respectively. The time complexity of (15) is \( O(mmc + nc^2 + mc^2 + c^3) \) as \( l \leq n \). In the same way, we know that the time complexity of (16) is \( O(mmc + nc^2 + mc^2 + c^3) \). In summary, the total time complexity of MUR is \( O(mmc + nc^2 + mc^2 + c^3) \times K \), where \( K \) is the number of iterations.

### B. Convergence Analysis

The convergence of MUR can be easily established. Since the update rule (14) is same as [13], which decreases the objective function \( f(W, D, H) \) with both \( D \) and \( H \) fixed. The following Proposition 1 and Proposition 2 prove that both (15) and (16) decrease the objective function \( f(W, D, H) \), respectively.

**Proposition 1.** Given both \( W \) and \( H \), the update rule (15) does not increase the objective function \( f(W, D, H) \).

**Proof:** Fixing both \( W \) and \( H \), we rewrite the objective function \( f(W, D, H) \) with respect to \( D \) as

\[
J(D) = \text{Tr}\left(-DHV^TW - \lambda DH_lY_l^T + \frac{1}{2} D^T WDHH^T + \frac{\lambda}{2} D^T DH_lH_l^T \right).
\]

Assuming (15) updates from \( D' \) to \( D'' \), i.e.,

\[
D'' = D' \circ \frac{W^T VH^T + \lambda Y_lH_l^T}{W^T WDHH^T + \lambda D'Y_lH_l^T}.
\]

Our goal is to prove \( J(D'') \leq J(D') \). To this end, we construct an auxiliary function for \( J(D) \) as

\[
Z(D, D') = - \sum_{ik} (W^T VH^T + \lambda Y_lH_l^T)_{ik} D_{ik} + \frac{1}{2} \sum_{ik} \frac{(W^T WDHH^T)_{ik} D_{ik}^2}{D_{ik}'} + \frac{\lambda}{2} \sum_{ik} \frac{(D'Y_lH_l^T)_{ik} D_{ik}^2}{D_{ik}'}.
\]

In the following, we will prove that \( Z(D, D') \) is an upper bound of \( J(D) \), i.e., \( J(D) \leq Z(D, D') \), and the equality is satisfied when \( D = D' \). It is easy to verify that \( J(D') = Z(D', D') \) and remains to prove \( J(D) \leq Z(D, D') \) for any \( D \).

According to [57], we have

\[
\sum_{i=1}^{n} \sum_{p=1}^{k} (AD'B)_{ip} D_{ip}^2 \geq \text{Tr}(D^T ADB),
\]

where both \( A \) and \( B \) are symmetric matrices. By substituting \( A = W^T W \) and \( B = HH^T \) into (20), we have

\[
\sum_{ik} \frac{(W^T WDHH^T)_{ik} D_{ik}^2}{D_{ik}'} \geq \text{Tr}(D^T W^T WDHH^T).
\]

By substituting \( A = I \) and \( B = H_lH_l^T \) into (19), we have

\[
\sum_{ik} \frac{(D'Y_lH_l^T)_{ik} D_{ik}^2}{D_{ik}'} \geq \text{Tr}(D^T D_lH_lH_l^T).
\]

Based on (21) and (22), it is easy to check that \( J(D) \leq Z(D, D') \). It is obvious that \( Z(D, D') \) is a quadratic function with respect to \( D \) and \( D'' \) is its minima. Since \( J(D) \leq Z(D, D') \) and \( J(D') = Z(D', D') \), we have

\[
J(D') \leq Z(D', D') \leq Z(D', D') \leq J(D').
\]

This completes the proof.

**Proposition 2.** Given \( W \) and \( D \), the update rule (16) does not increase the objective function \( f(W, D, H) \).

**Proof:** For the convenience of derivative, we divide \( H \) into two parts, i.e., \( H_l \) and \( H_u \). Given \( W \) and \( D \), the objective functions with respect to \( H_l \) and \( H_u \) can be written as

\[ J(H_l) = \frac{1}{2} \|V_l - WD_lH_l\|^2 + \frac{\lambda}{2} \|DH_l - Y_l\|^2. \]

\[ J(H_u) = \frac{1}{2} \|V_u - WD_uH_u\|^2. \]

We can easily prove the update rule (16) for \( H_u \) decreases \( J(H_u) \) according to [6]. It suffices to prove the update rule (16) for \( H_l \) decreases \( J(H_l) \). To this end, we rewrote \( J(H_l) \) as

\[
J(H_l) = \text{Tr}\left(-V_l^T WD_lH_l - \lambda Y_l^T D_lH_l + \frac{1}{2} H_l^T D^T WD_lH_l + \frac{1}{2} H_l^T D^T D_lH_l \right).
\]

Assuming (16) updates \( H_l \) from \( H_l' \) to \( H_l'' \), i.e.,

\[
H_l'' = H_l' \circ \frac{D^T W^T V_l + \lambda D^T Y_l}{D^T WDHH^T + \lambda D^T DH_lH_l^T}.
\]

We need to prove \( J(H_l'') \leq J(H_l') \). Following the proof of Proposition 1, we construct the auxiliary function for \( J(H_l) \) as:

\[
Z(H_l, H_l') = - \sum_{ik} (D^T W^T V_l + \lambda D^T Y_l)_{ik} (H_l')_{ik} + \frac{1}{2} \sum_{ik} \frac{(D^T WDHH^T)_{ik} (H_l')_{ik}^2}{(H_l')_{ik}'} + \frac{\lambda}{2} \sum_{ik} \frac{(D^T D_lH_lH_l^T)_{ik} (H_l')_{ik}^2}{(H_l')_{ik}'}
\]
By substituting $A = D^T D$ and $B = I$ into (20), we have
\[
\sum_{ik} \left( \frac{(D^T D)_{ik} (H_i)^2}{(H_i')^2} \right) \geq \text{Tr}(H_i^T D^T D H_i). \tag{30}
\]

By combining (27), (28), and (29), we know that $Z(H_i, H_i') \geq J(H_i)$. It is easy to verify that $H_i''$ is the minima of $Z(H_i, H_i')$ and it satisfies
\[
J(H_i') \leq Z(H_i'', H_i') \leq Z(H_i', H_i') \leq J(H_i'). \tag{31}
\]
This completes the proof. \hfill \Box

V. MANIFOLD REGULARIZED TNMTF

TNMTF provides a flexible framework to transduce discriminative information from training examples to unlabeled test examples. Based on this framework, other statistical information such as local geometric structure can be easily incorporated to further boost the discriminant power. To validate the flexibility of TNMTF, we introduced a manifold regularized TNMTF algorithm (MT-NMTF for short). MT-NMTF learns to represent examples $V = \{v_1, \ldots, v_l, \ldots, v_n\}$ as a linear combination of low-dimensional $c$ bases represented by a nonnegative matrix $W \in \mathbb{R}^{m \times c}$ and transduces the discriminative information from training examples to test examples based on TNMTF; i.e., $V \approx WDH$, where $H \in \mathbb{R}^{c \times n}$ denotes the encodings of $V$ in the low-dimensional space. Thus, it can be deemed as a function $h = f(v)$ subject to $W D h = v$. Our aim is to preserve the local geometry of the distribution of examples $V$. Suppose $V$ is sampled from a probability distribution $P_v$ on a manifold $M$ embed in a high-dimensional ambient space, the function $f$ can be achieved by penalizing the gradient $\nabla_M f$ along the manifold $M$, i.e.,
\[
\int_{V \in M} \left\| \nabla_M f \right\|^2 dP_v. \tag{32}
\]
where the integral is taken over the probability distribution $P_v$. However, since both the manifold $M$ and the marginal distribution $P_v$ are unknown in practice, we use its empirical estimate instead.

According to the manifold regularization theory [25], regularization (32) can be approximated by using the graph Laplacian of examples $V$. MT-NMTF constructs an adjacent graph $G$ whose vertexes are examples and the edge weights $S_{ij}$ reflect the extent to which two examples are close. By using '0-1 weighting' schema, the entries of $S$ are defined as
\[
S_{ij} = \begin{cases} 1, & \text{if } \tilde{v}_j \in N(\tilde{v}_i) \text{ or } \tilde{v}_i \in N(\tilde{v}_j) \\ 0, & \text{otherwise.} \end{cases} \tag{33}
\]
where $N(\tilde{v})$ denotes the sets of $k$ nearest neighbors of $\tilde{v}$. In our work, $k$ is set to 5, as we find the method works well in most cases under this setting. By using the strategy of defining the edge weight matrix, regularization (32) can be approximated by $fL^D f$, where $L = T - S$ is the graph Laplacian of $G$, and the $i$-th element of the diagonal matrix $T$ is defined as $T_{ij} = \sum_j S_{ij}$.

Since the basis $W$ contains $c$ independent components, the penalty (32), by taking summation of $c$ regularizations and using simple algebra, is equivalent to
\[
\sum_{i=1}^c f_i L_i^D = \text{Tr}\left( H L H^T \right), \tag{34}
\]
where $f_i$ is the function corresponding to the $i$-th components, i.e., $f_i = h_i$ denotes the $i$-th row of $H$. Like TNMTF, MT-NMTF constructs a label vector for each training example, and expects the coefficient of each training example to approximate its label vector. By exploiting (35) into (5), we have the objective of MT-NMTF
\[
\min_{w \geq 0, D \geq 0, h \geq 0} g(W, D, H) = \frac{1}{2} \|V - WDH\|^2_F + \frac{\lambda}{2} \|DH_l - Y_l\|^2_F + \frac{\beta}{2} \text{Tr}(HLH^T) - \langle H, \psi \rangle, \tag{35}
\]
where $\beta$ trades off the manifold regularization.

Similar to TNMTF, the objective function $g(W, D, H)$ is jointly non-convex with respect to $W, D$ and $H$, but it is convex with respect to each of them. Therefore, we can develop a MUR-based algorithm for solving MT-NMTF by using the Lagrangian multiplier method. Since the manifold regularization is independent of both $W$ and $D$, we keep their update rules consistent with (14) and (15) and focus on the update rule of $H$. To this end, similar to (10), we construct a Lagrangian function of $g(W, D, H)$ as
\[
\mathcal{G} = \frac{1}{2} \|V - WDH\|^2_F + \frac{\lambda}{2} \|DH_l - Y_l\|^2_F + \frac{\beta}{2} \text{Tr}(HLH^T) - \langle H, \psi \rangle, \tag{36}
\]
where $\psi$ denotes the Lagrangian multiplier for constraint $H \geq 0$.

Assume $H$ is a stationary point of the constrained optimization problem (35) with both $W$ and $D$ fixed. According to the K.K.T. condition [56], the stationary point satisfies the primal conditions
\[
\frac{\partial \mathcal{G}}{\partial H} = D^T W^T WD - D^T W^T V + \lambda \nabla + \beta H L - \psi = 0, \tag{37}
\]
where $\nabla = [D^T DH_l^T - D^T Y_l, 0]$, and the slackness conditions
\[
H \circ \psi = 0. \tag{38}
\]
By substituting (38) into (37) and using simple algebra, we have
\[
H \circ \left( D^T W^T WD + \lambda \nabla^+ + \beta HT \right) = H \circ \left( D^T W^T V + \lambda \nabla^- + \beta HS \right). \tag{39}
\]
where $\nabla^+ = [D^T DH_l^T, 0]$ and $\nabla^- = [D^T Y_l, 0]$ represent the positive and negative parts of $\nabla$, respectively. From (39), we obtain the following multiplicative update rule for $H$:
\[
H \leftarrow H \circ \frac{D^T W^T V + \lambda \nabla^- + \beta HS}{D^T W^T WD + \lambda \nabla^+ + \beta HT}. \tag{40}
\]
It can be easily proved that (40) does not increase the objective function \( g(W, D, H) \). We omit its proof due to the space limitations.

In the proposed MT-NMTF, the local geometry is expected to help to transduce the discriminative information from training examples to test examples. This is because neighbor examples are expected to have close encodings and their coefficients are also expected to be close to one another. Since the coefficient of each example approximates their label vector under the T-NMTF framework, it is likely that the neighbor examples will produce the same label vectors. In this sense, MT-NMTF enhances the clustering performance of unlabeled test examples by exploiting the local geometry structure of the whole dataset. Our experiments validate the effectiveness of both T-NMTF and MT-NMTF.

VI. EXPERIMENT

In this section, we conducted several experiments to validate the effectiveness of both T-NMTF and MT-NMTF on four popular face image datasets including ORL [58], FERET [59], YALE [60], UMIST [61], JAFFE [62] and MIT CBCL [63] by comparing them with NMF [1], NMTF [16], CNMF [31], and NMF-\( \alpha \) [30]. The Cambridge ORL dataset [64] consists of 400 images collected from 40 subjects. Ten images were collected from each subject with varying lighting, facial expressions and facial details (with-glasses or without-glasses). The FERET dataset [59] contains 13,539 photos in total taken from 1,565 subjects. We randomly selected 100 individuals, and each individual has 7 photos. The YALE dataset [60] contains 165 frontal view face photos of 15 subjects. Eleven photos were taken from each subject with varying facial expressions (smiling or sad) and configurations. The UMIST dataset [61] contains 575 face photos collected from subjects. At least 19 and at most 48 photos were taken from each subject in varying poses with both profile and frontal views. The JAFFE dataset [62] contains 213 face photos while the MIT CBCL dataset [63] contains 3240 face photos. All photos were cropped to a pixel array and reshaped into a long vector. Table 1 summarizes the datasets used.

| Name              | Size of Pixel Array | Number of Subjects | Number of Images |
|-------------------|---------------------|--------------------|-----------------|
| ORL [60]          | 32×32               | 40                 | 400             |
| FERET [59]        | 40×40               | 100                | 700             |
| YALE [60]         | 32×32               | 15                 | 165             |
| UMIST [61]        | 40×40               | 20                 | 213             |
| JAFFE [62]        | 256×256             | 10                 | 213             |
| MIT CBCL [63]     | 200×200             | 10                 | 3240            |

The recognition performance is evaluated by both the accuracy (AC) and the normalized mutual information (NMI) of test examples. Refer to [8], [9], [64] for more details of AC and NMI. In this experiment, both T-NMTF and MT-NMTF output labels using (6) and NMF, NMTF, CNMF, and NMF-\( \alpha \) output labels using K-means for the purpose of obtaining the best performance.

A. FACE RECOGNITION

To study the effectiveness of both T-NMTF and MT-NMTF, we first randomly selected one example from each individual for training and used the remaining examples for testing. To eliminate the effect of randomness, the trial was repeated ten times and the average AC and NMI on the test examples was used to evaluate the recognition performance. We varied the number of individuals from 2 to 10 in this experiment to study its effectiveness for multi-class clustering tasks.

Figures 1 and 2 give the average ACs and NMIs versus the number of classes of T-NMTF, MT-NMTF, CNMF, and NMF on both ORL and FERET datasets, respectively. They show that CNMF performs closely to NMF because it degenerates to NMF in this case (see Section 2.1). However, both T-NMTF and MT-NMTF overcome this deficiency and significantly boost recognition performance by transducing
the discriminative information from quite limited training examples to test examples.

To sufficiently mine the discriminant power of CNMF, we randomly selected two images from each individual for training in the following experiments and varied the number of classes from 2 to 10. We repeated the trial ten times and evaluated both T-NMTF and MT-NMTF by comparing them with NMF, NMTF, CNMF, and NMF-α in terms of both...
average AC and average NMI. Figures 3 to 6 show both average ACs and average NMIs versus the number of classes on the ORL, FERET, YALE and UMIST datasets, respectively. Figures 3 and 4 show that T-NMTF significantly outperforms NMF, NMTF, CNMF and NMF-α on both ORL and FERET datasets. Figures 5(a) and 6(a) show that T-NMTF
is superior to NMF, NMTF, CNMF, and NMF-\(\alpha\) in terms of AC on both YALE and UMIST datasets. Figure 5(b) shows that the performance of T-NMTF is comparable to both CNMF and NMF-\(\alpha\) in terms of NMI because the YALE dataset suffers seriously from noise by occlusion and varying profiles. However, the following section shows that T-NMTF
performs progressively better with the increase in the size of the training examples. From Figures 3 to 6, we see that MT-NMTF outperforms T-NMTF in most cases, and this observation confirms that local geometry structure further boosts the performance of T-NMTF.

To investigate the discriminant power of the data representation learned by T-NMTF, Figures 11 to 14 illustrate the bases (cluster centroid) learned by T-NMTF, MT-NMTF, CNMF, NMTF and NMF on ORL, FERET, YALE, and UMIST datasets, respectively. The fifth and sixth rows of Figure 11 show that the test examples do help NMF and NMTF to learn good representation. The fourth row of Figure 11 shows that the bases learned by CNMF favor the training examples due to the utilized constraint. However, the bases learned by both T-NMTF and MT-NMTF best represent the training examples because their coefficients are enforced close to the corresponding label vectors. Therefore, both T-NMTF and MT-NMTF succeeds in transducing the discriminative information of the training examples to the test examples and boosting the subsequent processing. Figure 8 makes a similar observation to Figure 11 on the FERET dataset. Figures 13 and 14 confirm that both T-NMTF and MT-NMTF represent the training examples well on both YALE and UMIST datasets, although the photos of both datasets were taken with serious occlusions or significantly varying profiles. This confirms our observations in Figures 5 and 6. Table 2 shows the result of Average ACs and NMIs of Kmeans, NMF, NMTF, CNMF, NMF-α and T-NMTF versus the number of classes on the MIT CBCL face database. As the MIT CBCL face database contains 3240 images, we randomly choose 30% samples from each class as labeled set. It can be seen from the table that the performance of these methods tend to perform worse with the increasing of the number of classes, and the reason is probably that clustering data of more categories is relatively more challenging. Furthermore, The semi-supervised CNMF, NMF-α and T-NMTF can deliver better results than the unsupervised NMF, NMTF and K-means methods, and the proposed T-NMTF method is superior to other methods in the experiment. These results show the superiority of our proposed T-NMTF, and the reason is probably that it can
TABLE 2. Average ACs and NMIs of Kmeans, NMF, NMTF, CNMF, NMF-α and T-NMTF versus the number of classes on the MIT CBCL face database. (30% samples from each class are regarded as labeled set).

| number of class | AC | NMI |
|----------------|----|-----|
|                | Kmeans | NMF | NMTF | CNMF | NMF-α | T-NMTF | Kmeans | NMF | NMTF | CNMF | NMF-α | T-NMTF |
| 2              | 0.611 | 0.601 | 0.581 | 0.741 | 0.658 | 0.929 | 0.202 | 0.149 | 0.111 | 0.235 | 0.301 | 0.641 |
| 3              | 0.685 | 0.649 | 0.694 | 0.742 | 0.662 | 0.947 | 0.591 | 0.313 | 0.603 | 0.584 | 0.579 | 0.833 |
| 4              | 0.511 | 0.574 | 0.578 | 0.616 | 0.632 | 0.723 | 0.480 | 0.457 | 0.544 | 0.508 | 0.542 | 0.636 |
| 5              | 0.487 | 0.470 | 0.508 | 0.559 | 0.550 | 0.704 | 0.447 | 0.369 | 0.406 | 0.544 | 0.499 | 0.614 |
| 6              | 0.447 | 0.405 | 0.399 | 0.536 | 0.499 | 0.648 | 0.425 | 0.318 | 0.388 | 0.457 | 0.470 | 0.551 |
| 7              | 0.469 | 0.463 | 0.489 | 0.556 | 0.542 | 0.673 | 0.500 | 0.470 | 0.510 | 0.547 | 0.545 | 0.608 |
| 8              | 0.458 | 0.430 | 0.401 | 0.605 | 0.550 | 0.739 | 0.544 | 0.523 | 0.466 | 0.636 | 0.606 | 0.715 |
| 9              | 0.445 | 0.454 | 0.451 | 0.563 | 0.445 | 0.680 | 0.558 | 0.359 | 0.546 | 0.634 | 0.545 | 0.609 |
| 10             | 0.475 | 0.507 | 0.471 | 0.565 | 0.504 | 0.736 | 0.586 | 0.607 | 0.482 | 0.672 | 0.626 | 0.717 |

simultaneously exploit the label information of training examples and the statistical structure of features in the test domain, thus it can get a better representation of the original data.

B. EFFECT OF SIZE OF TRAINING SET

In this section, we study the effectiveness of the size of training set. We randomly selected a different number of training images from each individual and fixed the number of classes at 10. To eliminate the effect of randomness, we repeated this trial ten times and compared the average ACs and NMIs of different methods. Since NMTF performs comparably with NMF, we have not included it in this experiment. Figures 7 to 10 give average ACs and NMIs versus the number of training images per individual on the ORL, FERET, YALE, and UMIST datasets, respectively.

From Figures 7 to 10, we can see that both T-NMTF and MT-NMTF significantly outperform NMF, CNMF and NMF-α with different sizes of training set on the four face image datasets. The more images per individual are utilized for training, the more superior the results of both T-NMTF and MT-NMTF are. This observation shows that the T-NMTF framework takes full advantage of the discriminative information of the training examples in the reduced dimensional space. It also confirms that the bases learned by both T-NMTF and MT-NMTF become more representative as the size of training set increases, and thus both T-NMTF and MT-NMTF are able to infer the labels of the test examples more accurately. Table 3 shows the Average ACs and NMIs of Kmeans, NMF, NMTF, CNMF, NMF-α and T-NMTF versus the number of training images per class on the JAFFE database. From the table we can find that semi-supervised CNMF, NMF-α and T-NMTF tend to perform better with the increasing number of training samples in general and they can deliver better results than the unsupervised NMF, NMTF and K-means methods. Furthermore, the proposed T-NMTF is superior to other semi-supervised methods even with very limited training samples as it can transduce the discriminative information from quite limited training examples to test examples.

Figure 15 gives both the ACs and NMIs of T-NMTF versus the parameter $\lambda$ on the ORL, FERET, YALE, and UMIST datasets. It shows that our T-NMTF model is stable over a wide range of the trade-off parameter $\lambda$ on all the tested datasets. Throughout this experiment, we set the trade-off parameter $\lambda = 1$.

C. EFFECT OF TRADE-OFF PARAMETERS

In T-NMTF (5), the critical trade-off parameter $\lambda$ balances the reconstruction error of all examples and the distance between the label vectors of the training examples and their corresponding coefficients. Here, we evaluate its effect on our T-NMTF model by cross-validating $\lambda$ over a set $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. According to Section 6.1, we randomly selected ten individuals from each of ORL, FERET, YALE, and UMIST datasets and randomly selected two images per individual for training according to Section 6.2. To eliminate the effect of randomness, we repeated this trial ten times and evaluated the performance of T-NMTF in terms of both average AC and average NMI. In MT-NMTF (36), another critical parameter $\beta$ controls the weight of the manifold regularization. According to the above analysis, the T-NMTF framework is insensitive to the trade-off parameter $\lambda$, and we therefore fixed $\lambda = 1$ and studied the effect of $\beta$ on the performance of MT-NMTF in terms of both AC and NMI. In this experiment, we cross-validated $\beta$ over a set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$.
We randomly selected two images from each dataset for training and reported the average AC and NMI as the final result. Figure 16 gives the cross-validation results on four datasets. It shows that MT-NMTF is stable over a wide range of $\beta$ from $10^{-5}$ to 1. Therefore, we fixed $\beta = 0.1$ in our experiments, and the experimental results are encouraging.

**D. STUDY OF GENERALIZATION**

T-NMTF generalizes well on unseen examples because the learned basis can be viewed as a good representation of both the training and test examples according to Figures 7-10. This experiment validates the generalization ability of T-NMTF by clustering unseen examples in the learned subspace. To this end, we partitioned each dataset into two subsets including a training set and a test set, and further partitioned the training set into a labeled set whose labels are known and an unlabeled set whose labels are unknown. T-NMTF was conducted on both the labeled set and the unlabeled set to learn the subspace. The evaluation was conducted by recognizing the test examples. For each test example $x$, its label vector is
calculated by
\[ y_i = \begin{cases} 1, & \text{if } i = \arg \max_k \{h_k\} \\ 0, & \text{otherwise} \end{cases} \] (41)

where \( h = \min_{h \geq 0} \|x - Wh\|_2^2 \) and \( W \) stands for the learned basis.

In this experiment, we respectively randomly selected \( p = 5, 6, 10 \) images for each individual from ORL, YALE, and UMIST datasets to form the training set and the remaining images composed the test set. Next, we randomly assigned labels for 1 to \( p \) training images of each individual.
To eliminate the effect of randomness, we repeated the trial ten times and output the average AC and NMI as the final result. Since the FERET dataset only contains seven images for each individual, it was not included in this experiment. Figures 17 to 19 give the average AC and NMI of T-NMTF, MT-NMTF, NMF, CNMF, and NMF-α on the test examples of ORL, YALE, and UMIST datasets, respectively. These results show that the performance of both T-NMTF and MT-NMTF is significantly superior to other methods, and this superiority increases with the increase in the percentage of labeled examples in the training set.

VII. CONCLUSION

This paper proposes an efficient Transductive Nonnegative Matrix Tri-Factorization method (T-NMTF) to simultaneously exploit the discriminative information of training examples and the features of all examples. Since T-NMTF forces the coefficient of the training examples to equal its label vector in the reduced dimensional space, T-NMTF successfully transduces the discriminative information from the training examples to the test examples and directly infers the labels for the test examples. We develop an efficient Multiplicative Update Rule (MUR) for optimizing T-NMTF and theoretically analyze MUR’s convergence. T-NMTF can be viewed as a flexible transductive learning framework under which we develop a manifold regularized T-NMTF (MT-NMTF) method. The experimental results on four face image datasets demonstrate the effectiveness of both T-NMTF and its variant.

ACKNOWLEDGMENT

The authors would like to thank J. Lin for her assistance with the experiment.

REFERENCES

[1] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, no. 6755, pp. 788–791, Oct. 1999.
[2] E. Wachsmuth, M. W. Oram, and D. I. Perrett, “Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque,” Cerebral Cortex, vol. 4, no. 5, pp. 509–522, 1994.
[3] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” Chemometrics Intell. Lab. Syst., vol. 2, nos. 1–3, pp. 37–52, 1987.
[4] G. H. Golub and C. Reinsch, “Singular value decomposition and least squares solutions,” in Linear Algebra. Berlin, Germany: Springer, 1971, pp. 134–151.
[5] Z. Lai, W. K. Wong, Y. Xu, J. Yang, and D. Zhang, “Approximate orthogonal sparse embedding for dimensionality reduction,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 4, pp. 723–735, Apr. 2016.
[6] Z. Liu, J. Wang, G. Liu, and L. Zhang, “Discriminative low-rank preserving projection for dimensionality reduction,” Appl. Soft Comput., vol. 85, Dec. 2019, Art. no. 105768.
[7] Z. Liu, Z. Lai, W. Ou, K. Zhang, and R. Zheng, “Structured optimal graph based sparse feature extraction for semi-supervised learning,” Signal Process., vol. 170, May 2020, Art. no. 107456.
[8] H. Sun, X. He, X. Wu, and J. Han, “Non-negative matrix factorization on manifold,” in Proc. 8th IEEE Int. Conf. Data Mining, Dec. 2008, pp. 63–72.
[9] W. Xu, X. Liu, and Y. Gong, “Document clustering based on non-negative matrix factorization,” in Proc. 26th Annul. Int. ACM SIGIR Conf. Res. Develop. Information Retr. SIGIR, 2003, pp. 267–273.
[10] S. Li, X. Wen Hou, H. Jiang Zhang, and Q. Sheng Cheng, “Learning spatially localized, parts-based representation,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR, Dec. 2001, pp. 207–212.
[11] D. L. Donoho and V. Stodden, “When does non-negative matrix factorization give a correct decomposition into parts,” in Proc. Adv. Neural Inf. Process. Syst., 2003, pp. 1141–1148.
[12] S. A. Vavasis, “On the complexity of nonnegative matrix factorization,” SIAM J. Optim., vol. 20, no. 3, pp. 1364–1377, Jan. 2010.
[13] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in Proc. Adv. Neural Inf. Process. Syst., 2001, pp. 556–562.
[14] H. Kim and H. Park, “Nonnegative matrix factorization based on alternating non-negativity constrained least squares and active set method,” SIAM J. Matrix Anal. Appl., vol. 30, no. 2, pp. 713–730, Jan. 2008.
[15] J. Kim and H. Park, “Toward faster nonnegative matrix factorization: A new algorithm and comparisons,” in Proc. 8th IEEE Int. Conf. Data Mining, Dec. 2008, pp. 353–362.
[16] C. Ding, T. Li, W. N. Y. H. Park, “Orthogonal nonnegative matrix t-factorizations for clustering,” in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining KDD, 2006, pp. 126–135.
[17] J. Yoo and S. Choi, “Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds,” Inf. Process. Manage., vol. 46, no. 5, pp. 559–570, Sep. 2010.
[18] D. Wang, F. Nie, and H. Huang, “Fast robust non-negative matrix factorization for large-scale data clustering,” in Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI), 2016, pp. 2104–2110.
[19] H. Wang, F. Nie, H. Huang, and F. Makedon, “Fast nonnegative matrix tri-factorization for large-scale data co-clustering,” in Proc. 22nd Int. Joint Conf. Artif. Intell., vol. 22, no. 1, Jun. 2011, p. 1553.
[20] H. Wang, F. Nie, H. Huang, and C. Ding, “Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation,” in Proc. 11th Int. Conf. Data Mining, Dec. 2011, pp. 774–783.
[21] H. Wang, H. Huang, and C. Ding, “Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization,” in Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. CIKM, 2011, pp. 279–284.
[22] V. N. Chakrabarti and K. Sycara, “Nonnegative matrix TRI-factorization with graph regularization for community detection in social networks,” in Proc. 24th Int. Joint Conf. Artif. Intell., Jun. 2015, pp. 2083–2089.
[23] P. Li, J. Bu, C. Chen, Z. He, and D. Cai, “Relational multimatrix co-clustering,” IEEE Trans. Cybern., vol. 43, no. 6, pp. 1871–1881, Dec. 2013.
[24] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, “Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification,” IEEE Trans. Neural Netw., vol. 17, no. 3, pp. 683–695, May 2006.
[25] N. Guan, D. Tao, Z. Luo, and B. Yuan, “Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent,” IEEE Trans. Image Process., vol. 20, no. 7, pp. 2030–2048, Jul. 2011.
[26] N. Guan, X. Huang, L. Lan, Z. Luo, and X. Zhang, “Graph based semi-supervised non-negative matrix factorization for document clustering,” in Proc. 11th Int. Conf. Mach. Learn. Appl., Dec. 2012, pp. 404–408.
[27] L. Lan, N. Guan, X. Zhang, D. Tao, and Z. Luo, “Soft-constrained non-negative matrix factorization via normalization,” in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2014, pp. 3025–3030.
[28] L. Lan, X. Huang, N. Guan, Z. Luo, and X. Zhang, “Semi-supervised non-negative patch alignment framework,” in Proc. 11th Int. Conf. Mach. Learn. Appl., Dec. 2012, pp. 174–178.
[29] N. Guan, D. Tao, L. Lan, Z. Luo, and X. Yang, “Activity recognition in still images with transductive non-negative matrix factorization,” in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer, 2014, pp. 802–817.
[30] Y. Cho and L. K. Saul, “Nonnegative matrix factorization for document clustering,” in Proc. 11th Int. Conf. Mach. Learn. Appl., Dec. 2012, pp. 1141–1148.
[31] X. Teng et al.; T-NMTF
XIAO TENG received the B.S. degree from the National University of Defense Technology, Changsha, China, in 2018, where he is currently pursuing the M.S. degree with the College of Computer. His current research interests include data mining, computer vision, and machine learning.

LONG LAN (Member, IEEE) received the Ph.D. degree in computer science from the National University of Defense Technology, in 2017. He was a Visiting Ph.D. Student with the University of Technology, Sydney, from 2015 to 2017. He is currently a Lecturer with the College of Computer, National University of Defense Technology. His research interests include multiobject tracking, computer vision, and discrete optimization.

ZHIGANG LUO received the M.S. and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 2010 and 2015, respectively, where he is currently a Research Assistant with the State Key Laboratory of High Performance Computing, College of Computer, Institute for Quantum Information. His current research interests include computer vision and machine learning.

GUOHUA DONG is currently pursuing the Ph.D. degree with the College of Computer, National University of Defense Technology. Her current research interests include image retrieval, cross-modal retrieval, hashing methods, and graph models.

X. Teng et al.: T-NMTF

[38] H. Ma, W. Zhao, Q. Tan, and Z. Shi, “Orthogonal nonnegative Matrix TRI-factorization for semi-supervised document co-clustering,” in Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining, 2010, pp. 189–200.

[39] Y. Zhang, Z. Zhang, S. Li, J. Qin, G. Liu, M. Wang, and S. Yan, “Unsupervised nonnegative adaptive feature extraction for data representation,” IEEE Trans. Knowl. Data Eng., vol. 31, no. 12, pp. 2423–2440, Dec. 2019.

[40] Y. Chen, M. Rege, M. Dong, and J. Hua, “Non-negative matrix factorization for semi-supervised data clustering,” Knowl. Inf. Syst., vol. 17, no. 3, pp. 355–379, Dec. 2008.

[41] J. Shao, X. Wang, Q. Yang, C. Plant, and C. Böhm, “Synchronization-based scalable subspace clustering of high-dimensional data,” Knowl. Inf. Syst., vol. 52, no. 1, pp. 33–111, Jul. 2017.

[42] J. Hu, Q. Qian, J. Pei, R. Jin, and S. Zhu, “Finding multiple stable clusterings,” Knowl. Inf. Syst., vol. 51, no. 3, pp. 991–1021, Jun. 2017.

[43] R. Kannan, M. Ishteva, and H. Park, “Bounded matrix factorization for recommender system,” Knowl. Inf. Syst., vol. 39, no. 3, pp. 491–511, Jun. 2014.

[44] H.-F. Yu, C.-J. Hsieh, S. Si, and I. S. Dhillon, “Parallel matrix factorization for recommender systems,” Knowl. Inf. Syst., vol. 41, no. 3, pp. 793–819, Dec. 2014.

[45] F. Nie, H. Wang, H. Huang, and C. Ding, “Joint schatten p-norm and $l_p$-norm robust matrix completion for missing value recovery,” Knowl. Inf. Syst., vol. 42, no. 3, pp. 525–544, Mar. 2015.

[46] A. Peclii, M. C. Cavalcanti, and R. Goldscheidt, “Automatic feature selection for supervised learning in link prediction applications: A comparative study,” Knowl. Inf. Syst., vol. 56, no. 1, pp. 85–121, Jul. 2018.

[47] W. Xu and Y. Gong, “Document clustering by concept factorization,” in Proc. 27th Annu. Int. Conf. Res. Develop. Inf. Retr. SIGIR, 2004, pp. 202–209.

[48] Z. Zhang, Y. Zhang, S. Li, G. Liu, D. Zeng, S. Yan, and M. Wang, “Flexible auto-weighted local-coordinate concept factorization: A robust framework for unsupervised clustering,” IEEE Trans. Knowl. Data Eng., early access, Sep. 13, 2019, doi: 10.1109/TKDE.2019.2940576.

[49] Y. Zhang, Z. Zhang, Z. Zhang, M. Zhao, L. Zhang, Z. Zha, and M. Wang, “Deep self-representative concept factorization network for representation learning,” 2019, arXiv:1912.06444. [Online]. Available: http://arxiv.org/abs/1912.06444

[50] J. Ren, Z. Zhang, S. Li, Y. Wang, G. Liu, S. Yan, and M. Wang, “Learning hybrid representation by robust dictionary learning in factorized compressed space,” IEEE Trans. Image Process., vol. 29, pp. 3941–3956, 2020.

[51] Z. Zhang, Y. Zhang, G. Liu, J. Tang, S. Yan, and M. Wang, “Joint label prediction based semi-supervised adaptive concept factorization for robust data representation,” IEEE Trans. Knowl. Data Eng., vol. 32, no. 5, pp. 952–970, May 2020.

[52] Z. Zhang, L. Zhang, M. Zhao, W. Jiang, Y. Liang, and F. Li, “Semi-supervised image classification by nonnegative sparse neighborhood propagation,” in Proc. 5th ACM Int. Conf. Multimedia Retr. ICMR, 2015, pp. 139–146.

[53] L. Jia, Z. Zhang, L. Wang, W. Jiang, and M. Zhao, “Adaptive neighborhood propagation by joint $l_2,1$-norm regularized sparse coding for representation and classification,” in Proc. IEEE 16th Int. Conf. Data Mining (ICDM), Dec. 2016, pp. 201–210.

[54] Z. Zhang, L. Jia, M. Zhang, B. Li, L. Zhang, and F. Li, “Discriminative clustering on manifold for adaptive transductive classification,” Neural Netw., vol. 94, pp. 260–273, Oct. 2017.

[55] H. Zhang, Z. Zhang, M. Zhao, Q. Ye, M. Zhang, and M. Wang, “Robust Triple-Matrix Recovery-Based auto-weighted label propagation for classification,” IEEE Trans. Neural Netw. Learn. Syst., early access, Jan. 22, 2020, doi: 10.1109/TNNLS.2019.2956015.

[56] D. P. Bertsekas, “Nonlinear programming,” J. Oper. Res. Soc., vol. 48, no. 3, p. 334, 1997.

[57] C. Ding, T. Li, and M. Jordan, “Convex and semi-nonnegative matrix factorizations for clustering and low-dimension representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 1, pp. 45–55, Jan. 2010.

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

[59] P. J. Phillips, H. Moon, P. J. Rauss, and S. A. Rizvi, “The FERET evaluation methodology for face recognition algorithm,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

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

[61] D. B. Graham, “Characterizing virtual eigensignatures for general purpose face recognition,” Face Recognit. Form Theory Appl., vol. 163, no. 2, pp. 446–456, 1998.

[62] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with Gabor wavelets,” in Proc. 3rd IEEE Int. Conf. Autom. Face Recognit., Apr. 1998, pp. 200–205.

[63] H. A. Rowley, S. Baluja, and T. Kanade, “Neural network-based face detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 1, pp. 23–38, Jan. 1998.

[64] W. Liu, N. Zheng, and X. Lu, “Non-negative matrix factorization for visual coding,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Apr. 2003, pp. 293–296.