Weakly Supervised Deep Matrix Factorization for Social Image Understanding

Zechao Li and Jinhui Tang, *Senior Member, IEEE*

***Abstract*— The number of images associated with weakly supervised user-provided tags has increased dramatically in recent years. User-provided tags are incomplete, subjective and noisy. In this paper, we focus on the problem of social image understanding, i.e., tag refinement, tag assignment, and image retrieval. Different from previous work, we propose a novel weakly supervised deep matrix factorization algorithm, which uncovers the latent image representations and tag representations embedded in the latent subspace by collaboratively exploring the weakly supervised tagging information, the visual structure, and the semantic structure. Due to the well-known semantic gap, the hidden representations of images are learned by a hierarchical model, which are progressively transformed from the visual feature space. It can naturally embed new images into the subspace using the learned deep architecture. The semantic and visual structures are jointly incorporated to learn a semantic subspace without overfitting the noisy, incomplete, or subjective tags. Besides, to remove the noisy or redundant visual features, a sparse model is imposed on the transformation matrix of the first layer in the deep architecture. Finally, a unified optimization problem with a well-defined objective function is developed to formulate the proposed problem and solved by a gradient descent procedure with curvilinear search. Extensive experiments on real-world social image databases are conducted on the tasks of image understanding: image tag refinement, assignment, and retrieval. Encouraging results are achieved with comparison with the state-of-the-art algorithms, which demonstrates the effectiveness of the proposed method.**

***Index Terms*— Deep architecture, matrix factorization, weakly supervised, image tagging, image retrieval.**

I. Introduction

In real-world applications, many photo sharing websites, such as Flickr [1] and Facebook [2], have been becoming popular, which facilitate millions of users to upload, share and tag their images. It leads to the dramatic increase in the number of images associated with user-provided tags available. For example, the Verge reported in March 2013 that Flickr had more than 3.5 million new images uploaded daily [3]. It sheds new light on the problem of image understanding. Unfortunately, these tags are provided by amateur users and are imperfect, i.e., they are often incomplete or inaccurate in describing the visual content of images, which brings challenges to the tasks of image understanding such as tag-based image retrieval [4]. In this work, we focus on refining image tags to complement relevant tags and remove the irrelevant tags, and assigning tags to new images.

Image annotation [5]–[11] is traditionally treated as a machine learning problem, which always depends on a small-scale manually-labeled data. However, they fail to handle large-scale social images due to the weakly-supervised data. Different from the traditional image annotation, tag refinement is to remove irrelevant tags from the initial tags associated with images and add relevant but missing tags [12]–[18]. Li et al. [12] proposed to estimate the tag relevance using a neighbor voting algorithm. In [13], low rank matrix completion with the constraints of the content consistency and textual consistency is used to address the tag refinement problem. By jointly utilizing the labeled and unlabeled data, a kNN-sparse graph-based semi-supervised learning method is proposed to refine tags of social images in [14]. In [16], a latent space is identified based on low rank approximation to link the visual features of images and tags. The image-tag relevance which is consistent with the original image-tag relation and the visual similarity are explored in [17]. Image-specific and tag-specific linear sparse reconstructions are introduced for automatic image tag completion in [18]. However, most of them cannot directly incorporate new images into the learned model, i.e., the out-of-sample problem. Unlike most existing studies, the proposed work simultaneously addresses the problems of image tag refinement and tag assignment assigning tags to new images.

There are also few previous work [19], [20] dealing with the image tag refinement and tag assignment simultaneously. The tags of the learning images are refined while the model is learned. And the learned model can assign tags to new images. Gong et al. [19] proposed a three-view Canonical Correlation Analysis (CCA) model by incorporating the third view, such as topics obtained by clustering tags of social images. With the learned transformation matrix, new images are easily embedded in the latent subspace. In [20], images and tags are embedded in the unified subspace, in which image tag refinement is transformed to the nearest tag-neighbor search for each image. The image latent representation in the subspace is projected from visual feature representations,

Manuscript received October 1, 2015; revised August 16, 2016 and October 3, 2016; accepted October 29, 2016. Date of publication November 1, 2016; date of current version November 18, 2016. This work was supported in part by the 973 Program under Project 2014CB347600 and in part by the National Natural Science Foundation of China under Grant 61522203 and Grant 61402228. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Weisi Lin. *(Corresponding author: Jinhui Tang.)*

The authors are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: echao.li@njust.edu.cn; jinhuitang@njust.edu.cn).

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.2016.2624140

1057-7149 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

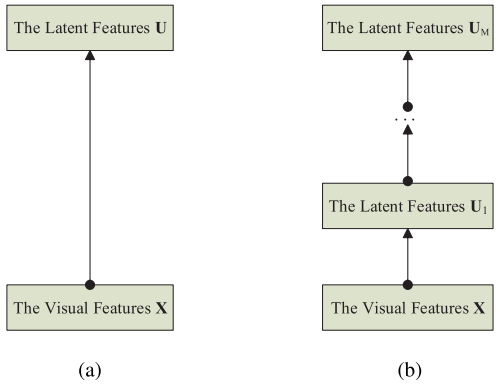


Fig. 1. (a) The shallow model directly maps the visual features into the latent subspace. (b) The deep model learns hierarchical feature representations from the visual features.

which can embed any image into the unified space. They learn the latent relevance between visual features and tags in the latent subspace, which has higher-level semantic, such as topic. However, compared with the tag information, visual feature is a much lower level representation on semantics, and there exists the well-known semantic gap between visual representation and semantic meaning making it challenging [21], [22]. It is unsuitable to directly transform the visual features to the latent representations by a shallow model as shown in Figure 1a and the performance is often unsatisfied.

Towards this end, in this work we propose a novel Weakly-supervised Deep Matrix Factorization (WDMF) framework to refine the initial tags and assign tags to new images. Figure 2 illustrates the flowchart of the proposed framework. It learns the latent image representations and the latent tag representations in the latent subspace by jointly leveraging the weakly-supervised tagging information, visual structure and semantic structure. To better handle the semantic gap, a deep architecture as illustrated in Figure 1b is developed to learn the latent image representations by a progressive way, which can automatically learn the intermediate hidden representations. To obtain a better higher-level feature representation, the semantic structure in the textual space is exploited by expecting that it is consistent with the one in the latent subspace. However, the tagging information is imperfect, and the problem of overfitting may be caused by learning the imperfect data. Consequently, the visual structure is introduced to the proposed model. Besides, a sparse model with the $\ell_{2,1}$ mixed norm is imposed on the transformation matrix of the first layer to filter out the noisy or redundant visual features, since the original visual features are always correlated or redundant to each other, and sometimes noisy. The above principles are jointly formulated into a unified optimization problem, which is optimized using a gradient descent procedure with curvilinear search. To empirically validate the effectiveness of the proposed method, extensive experiments are conducted on two widely utilized real-world datasets on the tasks of image tag refinement, image tag assignment and image retrieval. The achieved outperforming results compared with several representative methods demonstrate the superiority of the proposed method.

The main contributions of this work are summarized as follows.

- To our best knowledge, it is the first work to propose a weakly-supervised deep matrix factorization framework for social image tag refinement, tag assignment and image retrieval.
- We propose a Weakly-supervised Deep Matrix Factorization (WDMF) framework by collaboratively exploring the heterogeneous data of social images, i.e., the weakly-supervised tagging information, the visual information and the textual information.
- In the proposed framework, the hidden representations of images are learned by a hierarchical model in a progressive way, which can alleviate the semantic gap. And the problems of the imperfect tagging information and the redundant or noisy visual features are jointly addressed.

The remainder of this paper is organized as follows. In Section II, we discuss the previous work on image tag refinement and matrix factorization based latent factor learning. Section III elaborates the proposed WSMF framework with the optimization algorithm. The experimental evaluations and discussions are presented in Section IV. Section V concludes this paper with future research directions.

II. Related Work

A. Social Image Analysis

In the multimedia and data mining communities, many researchers focus on the problem of social image analysis [12]–[20], [23]–[30]. Different traditional image annotation methods [5], [6], [8], [9], [31] that usually learn models from small-scale manually-labeled images, these methods exploit massive images associated with weakly-supervised user-provided tags. In this subsection, we present the related work about social image tag refinement and social image tag assignment.

Social image tag refinement is to remove the noisy or irrelevant tags and add the relevant tags. In [23], the group information of images from Flickr is exploited with the assumption that the images within a batch are likely to have a common style. Zhu et al. [13] proposed to decompose the image-tag matrix into a low rank matrix and a sparse matrix and considered the content consistency and tag correlation as regularization terms. The low rank matrix recovery is combined with maximum likelihood estimation to recover the missing tags and de-emphasize the noisy tags in [32]. Zhuang and Hoi [15] discovered the relationships between images and tags by exploiting the textual and visual contents of images to refine tags. Tag co-occurrence is used to find the related tags with the original tags in [33]. In [34], tag refinement is performed using a topic model, i.e., regularized latent dirichlet allocation. In [16], a latent space is identified based on low rank approximation to link the visual features of images and tags. In [17], the relevance between images and tags consistent with the observed tags and the visual similarity is learned. However, most of them cannot directly handle new images out of the learning image set.

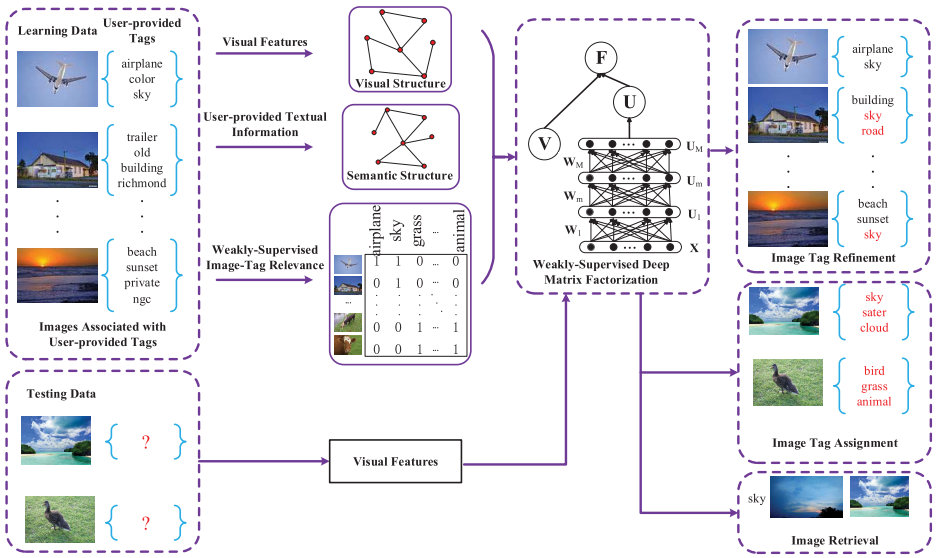


Fig. 2. The flowchart of the proposed framework WDMF for tag refinement, tag assignment and image retrieval. The added tags are marked with red color.

Many neighbor voting based approaches have been proposed for image tag refinement [12], [35], [36]. A tag is selected to annotate a target image if it is used to tag the visually similar images of the target image. In [35], the importance of visual content is discussed in the neighbor voting schemes for image tag refinement. Tags are completed by using the visual neighbor voting and belief theory in [36]. However, they treat each image selected as the neighbor one either equally or simply based on its visual similarity.

Some researchers focus on label propagation to refine tags of social images [14], [37]–[39]. A semi-supervised learning method [14] is developed to refine tags by label propagation over noisily-tagged web images based on the kNN-sparse graph. In [39], a voting graph is constructed to model the relationships among images and an adaptive teleportation random walk is proposed to estimate the tag relevance.

Social images are also mined in terms of other aspects for image tagging, such as metric learning [29], [40], model fusion [41], binary codes embedding [42], semantic diversity [43]. A distance metric is learned by exploiting both visual and textual contents of social images and applied to automated image tagging in [40]. In [29], the user-provided tags and visual information are explored to learn a distance metric, and the learned similarity can well reflect the semantic information. To improve the performance of tag-based image retrieval, ranking scores from both tag-based and content-based models are combined in [41]. Binary codes are constructed by preserving the similarities between images and tags, and then images are tagged based on the Hamming distance in [42]. Qian et al. [43] exploited tag diversity and retagged images with relevant tags with diverse semantics.

Few previous approaches [19], [20], [44] are designed to address the image tag refinement and tag assignment simultaneously. Gong et al. [19] proposed a three-view CCA model to model the statistics of images and associated textual data, which introduces the third view in terms of topics obtained by clustering tags for social images. In [20], images and tags are embedded in the unified subspace uncovered by the matrix factorization framework. The latent image representations are projected from visual feature representations, which can embed any image into the unified space. However, they directly transform the visual features to the latent representations using a linear transformation matrix, which is difficult to reduce the semantic gap. In [44], a subspace learning framework is proposed to uncover a subspace that can preserve the local structure and well predict the label information. The learned subspace is linearly transformed from the visual space, and the proposed framework can be applied to social image tag refinement and tag assignment. Different from the above methods, we propose a deep matrix factorization framework by exploiting the weakly-supervised information, the visual and textual structures simultaneously. The latent image representations are learned under the hierarchical structure, which can well deal with the semantic gap. In addition, the proposed method can deal with both image tag refinement, image tag assignment and image retrieval.

B. Matrix Factorization Based Latent Subspace Learning

Matrix Factorization (MF) is an effective latent factor learning model. Given a data matrix $\mathbf{Y} \in \mathbb{R}^{l \times n}$, matrix factorization tries to two low rank factors whose multiplication can well approximate it.

$$\mathbf{Y} \approx \mathbf{V}\mathbf{U} \tag{1}$$

Here $\mathbf{V} \in \mathbb{R}^{l \times r}$ and $\mathbf{U} \in \mathbb{R}^{r \times n}$ are the latent factor matrices with $r < \min(l, n)$. To avoid overfitting, two regularization terms are introduced.

$$\min_{\mathbf{U},\mathbf{V}} \frac{1}{2} \|\mathbf{Y} - \mathbf{V}\mathbf{U}\|_F^2 + \frac{\lambda_1}{2} \|\mathbf{U}\|_F^2 + \frac{\lambda_2}{2} \|\mathbf{V}\|_F^2 . \tag{2}$$

$\|\cdot\|_F$ denotes the Frobenius norm. λ_1 and λ_2 are two positive parameters. It has a nice probabilistic interpretation

with Gaussian observation noise as detailed in Probabilistic MF (PMF) [45].

Different variants of MF have been proposed, such as Multi-correlation PMF (MPMF) [7], robust matrix factorization [46] and Bayesian PMF [47]. If each element of the latent matrices is required to be nonnegative, it leads to Nonnegative Matrix Factorization (NMF) [48], [49]. Gupta *et al.* [50] proposed a shared NMF method to learn a shared subspace between two sources. Graph regularized Non-negative Matrix Factorization (GNMF) [51] explicitly exploits the local geometrical information as a regularization term. Supervised NMF-based method is proposed to address the problems of classification and annotation in [52]. Trigeorgis et al. [53] proposed a deep semi-NMF method to learn hidden representations. There are some other factor models to find the latent factors, such as tensor factorization [54]. Different from the previous work, a new weakly-supervised deep MF method is proposed to learn a series of transformation matrices for the hidden image features and tag features in the latent subspace. Different from our preliminary work [55], we directly explore the latent structures of data by introducing the corresponding constraints and optimize the proposed optimization problem using a gradient descent procedure with curvilinear search.

III. Weakly-Supervised Deep Matrix Factorization

A. Preliminary

Throughout this work, lowercase italic letters (i.e., i, j, n, etc.) and uppercase italic letters (i.e., A, B, M, etc.) denote scalars while bold uppercase characters (i.e, $\mathbf{W}$, $\mathbf{X}$, etc.) and bold lowercase characters (i.e, $\mathbf{a}$, $\mathbf{x}$, etc.) are utilized to denote matrices and vectors, respectively. For any matrix $\mathbf{A}$, $\mathbf{a}^i$ means the i-th column vector of $\mathbf{A}$, $\mathbf{a}_i$ means the i-th row vector of $\mathbf{A}$, A_{ij} denotes the (i, j)-element of $\mathbf{A}$ and Tr[$\mathbf{A}$] is the trace of $\mathbf{A}$ if $\mathbf{A}$ is square. $\mathbf{A}^T$ denotes the transposed matrix of $\mathbf{A}$. The Frobenius norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as $\|\mathbf{A}\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2 = \text{Tr}[\mathbf{A}^T\mathbf{A}]$. The $\ell_{2,1}$-norm for $\mathbf{A}$ is defined as

$$\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{n} \|\mathbf{a}^i\|_2. \tag{3}$$

Consider a social image set consisting of n images $\{\mathbf{x}^i\}_{i=1}^n$ assigned with l user-provided tags $\mathcal{C} = \{t_1, t_2, \cdots, t_l\}$. For each image $\mathbf{x}^i$, the observed relationships between this image and tags can be represented as a l-dimensional binary-valued vector $\{\mathbf{f}^i\}$. The visual feature matrix is denoted as $\mathbf{X} = [\mathbf{x}^1, \cdots, \mathbf{x}^n]$, in which $\mathbf{x}^i \in \mathbb{R}^d$ is the feature vector of the i-th image while $\mathbf{F} = [\mathbf{f}^1, \cdots, \mathbf{f}^n] \in \mathbb{R}^{l\times n}$ is the tagging matrix, in which $F_{ji} = 1$ indicates that $\mathbf{x}^i$ is associated with the j-th tag, and $F_{ji} = 0$ otherwise. Besides, symmetric matrices $\mathbf{S}$ and $\mathbf{C}$ are utilized to denote the image similarity matrix and the tag correlation matrix, respectively. $\mathbf{I}_m \in \mathbb{R}^{m\times m}$ denotes the identity matrix.

B. Motivation

The key functionality of image tag refinement, image tag assignment and tag-based image retrieval is to uncover the latent relevance of tags with respect to the visual content of a given image by exploiting available resources. In this work we focus on addressing this issue in the MF framework by uncovering the latent subspace.

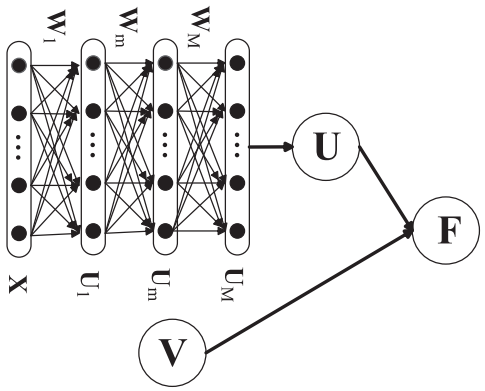


Fig. 3. The illustration of the proposed deep matrix factorization model.

In the problem of image tag refinement and image tag assignment, there are two media types: image and tag, and the content of images and tags are correlated although there exist noisy or irrelevant tags. However, the text space and visual space have inherently different structures. To address this problem, it is crucial and necessary to discover a shared structure to link them. This naturally motivates us to discover a latent space, in which images and tags are embedded. This can be identified by the factor analysis model. What is more important for massive social images, a good model should have good scalability, that is, it can assign tags to new images for our task. Intuitively, it can be achieved by one transformation from the visual space to the latent space. Unfortunately, the visual descriptor is a much lower level representation on semantics compared with the textual information, and there exists the well-known semantic gap making it challenging. Besides, the latent subspace can be treated as a topic space, which is a higher level representation on semantics. Consequently, it is unsuitable to directly transform the visual space to the latent space since the semantic gap may be too large. To well address this problem, we propose a deep architecture to uncover the hidden subspace from the visual space in a progressive way (as shown in Figure 3). The semantic gap is reduced progressively. A better latent space is learned through the proposed progressive learning strategy, which is validated by our experiments in Section IV. In addition, the visual and semantic structures should be preserved. That is to say, visually similar images should have similar representations in the latent subspace, and the latent features of semantically relevant tags should be similar, which are leveraged in the proposed framework.

C. The Proposed Formulation

In the deep architecture, the latent image representations in the uncovered subspace are learned in layers. Let us assume that the proposed hierarchical structure has M layers. The proposed DMF model factorizes the observed image tagging matrix $\mathbf{F}$ into $M + 1$ factor matrices, i.e., $\mathbf{V}$, $\mathbf{U}_M$, $\cdots$, $\mathbf{U}_1$.

To better exploit the visual features of images and deal with new images, the output of the first layer is transformed from the visual space, i.e., $\mathbf{U}_1 = \mathbf{W}_1\mathbf{X}$. Besides, in this work, since we focus on explaining our basic idea rather than designing a complex objective function, a deep neural network is constructed to discover the hidden representations using multiple layers of linear transformations rather complex nonlinear transformations. As a consequence, the proposed factorization model is obtained as follows.

$$\begin{aligned} \mathbf{F} &\leftarrow \mathbf{V}\mathbf{U}_M \\ \mathbf{U}_M &= \mathbf{W}_M\mathbf{U}_{M-1} \\ &\vdots \\ \mathbf{U}_2 &= \mathbf{W}_2\mathbf{U}_1 \\ \mathbf{U}_1 &= \mathbf{W}_1\mathbf{X} \end{aligned} \tag{4}$$

Here $\mathbf{W}_m (m = 1, \cdots, M)$ is the transformation matrix of the m-th layer. $\mathbf{V}$ is the latent tag feature matrix in the subspace and $\mathbf{U}_m$ is the implicit representation matrix of images in the m-th layer. In this work, the output of the most top layer $\mathbf{U}$ as shown in Figure 3 is computed as $\mathbf{U} = \mathbf{U}_M$. From above equations, it can be observed that the problem of learning $M + 1$ factor matrices becomes the problem of learning one factor $\mathbf{V}$ and M transformation matrices $\mathbf{W}_M, \cdots, \mathbf{W}_1$.

As one can see from the formulation (4), the latent image representation at the top layer learned by the proposed deep MF model have a similar interpretation as one in the traditional MF modal. However, the transformation from the original visual space to the latent space is now further analyzed as a product of multiple factors. The learned subspace can be deemed as a topic space, which has higher-level semantic meanings. Due to the well known semantic gap, it is difficult to directly map the visual features into the latent subspace using a transformation matrix, which often leads to poor performance. The constructed deep model is able to (1) uncover a better high-level, top-layer representation for images; (2) well alleviate the semantic gap using the progressive way; and (3) find suitable hidden representations at each layer. Besides, to better reduce the semantic gap and obtain a better underlying subspace, $\mathbf{U}_{M-1}$ is required to be analogous to the tag space, and then transformed to the latent subspace. To this end, a cost function is defined to measure the semantic difference between the structures of the hidden space $\mathbf{U}_{M-1}$ and the text space. In this work, we utilize the Laplacian regularization to constrain this different, i.e., the semantically similar images should be similar to each other in the hidden space $\mathbf{U}_{M-1}$.

$$\begin{aligned} &\min_{\mathbf{U}_{M-1}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=n}^{l} A_{ij}\|\mathbf{u}_{M-1}^i - \mathbf{u}_{M-1}^j\|_2^2 \\ \Leftrightarrow\ &\min_{\mathbf{U}_{M-1}} \mathrm{Tr}[\mathbf{U}_{M-1}^T\mathbf{T}\mathbf{U}_{M-1}] \end{aligned} \tag{5}$$

Here A_{ij} measures the semantic relevance between the i-th image and the j-th image, which is defined using the cosine similarity based on the tagging vectors. $\mathbf{T} = \mathbf{D}^A - \mathbf{A}$ is the positive semi-definite Laplacian matrix, in which $\mathbf{D}^A$ is a diagonal matrix with $D_{ii}^A = \sum_{j=1}^{l} A_{ij}$.

Semantic tags associated with images naturally correlate with each other at the semantic level and always appear correlatively. Therefore, this prior should be exploited, which is benefit to the discovery of the latent subspace. Specifically, the semantic consistency is exploited to learn the latent subspace. That is, the tags with stronger correlations are assumed to be closer to each other in the learned subspace, which is analogous to the Laplace-Beltrami operator on manifolds [56]. Towards this end, a smooth regularization is imposed on the underlying geometric structure between tags in the latent subspace, which can also avoid the overfitting problem induced by the noisy and sparse tag-image correlations. The semantic consistency is preserved in the underlying latent subspace to exploit the semantic structure by the following constraint.

$$\min_{\mathbf{V}} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} C_{ij}\|\mathbf{v}_i - \mathbf{v}_j\|_2^2 = \mathrm{Tr}[\mathbf{V}^T\mathbf{L}\mathbf{V}] \tag{6}$$

Here C_{ij} measures the relationship between the i-th tag and the j-th tag. $\mathbf{L} = \mathbf{D}^C - \mathbf{C}$ is the positive semi-definite Laplacian matrix, in which $\mathbf{D}^C$ is a diagonal matrix with $D_{ii}^C = \sum_{j=1}^{l} C_{ij}$. Following [17], the semantic correlation is quantified as follows.

$$C_{ij} = \frac{P(t_i, t_j)}{P(t_i, t_i) + P(t_j, t_j) - P(t_i, t_j)} \tag{7}$$

The above equation can well estimate the co-occurrence probability to measure the co-occurrence frequency. $P(t_i, t_j)$ is defined as $P(t_i, t_j) = \sum_{t=1}^{n} F_{it}F_{jt}$, which essentially estimate the number of images associated with these two tags.

On the other hand, the visual structure ought to be leveraged to discover a better subspace. The discriminative structure is exploited to well preserve the visual structure. Since a global predictor might not be good for the entire space [57], we focus on exploiting the local discriminative structure. For each image $\mathbf{x}^i$, its k nearest neighborhood $\mathcal{N}_k(\mathbf{x}^i)$ is constructed. Following [13] and [32], k is empirically set $k = 0.001n$. Each dimension of the latent subspace can be considered as a topic. For $\mathcal{N}_k(\mathbf{x}^i)$, we construct predicting function $h_i^q(\mathbf{x})$ $(1 \le q \le r)$ to predict the topic of images within $\mathcal{N}_k(\mathbf{x}^i)$. $h_i^q(\mathbf{x})$ is learned by minimizing the following loss function.

$$J_i^q = \frac{1}{k}\sum_{\mathbf{x}^j \in \mathcal{N}_k(\mathbf{x}^i)} (h_i^q(\mathbf{x}^j) - U_{qi})^2 + \gamma_i\Omega(h_i^q) \tag{8}$$

The second term measures the smoothness of the predictor and γ_i is a positive regularization parameter. In this work, we set $\gamma_i = \gamma$. The predictor $h_i^q(\mathbf{x})$ can be written as [58]:

$$\mathbf{h}_i^q = \mathbf{K}_i\rho_i^q, \tag{9}$$

where $\mathbf{K}_i \in \mathbb{R}^{k\times k}$ is the kernel matrix defined on $\mathcal{N}_k(\mathbf{x}^i)$. ρ_i^q is the corresponding coefficient vector. Thus, we have

$$J_i^q = \frac{1}{k}\|\mathbf{K}_i\rho_i^q - \mathbf{u}_q^i\|_2^2 + \gamma\,(\rho_i^q)^T\mathbf{K}_i\rho_i^q, \tag{10}$$

where $\mathbf{u}_q^i$ is the vector corresponding to images in $\mathcal{N}_k(\mathbf{x}^i)$. Through simple deduction, ρ_i^q is obtained and then substituted into $h_i^q(\mathbf{x}^i)$.

$$h_i^q = \mathbf{k}_i^T(\mathbf{K}_i + \gamma\, k\mathbf{I}_k)^{-1}\mathbf{u}_q^i = \mathbf{g}_i^T\mathbf{u}_q^i \tag{11}$$

$\mathbf{k}_i$ is the kernel vector defined on $\mathcal{N}_k(\mathbf{x}^i)$. Combining all the local predictors together, we have

$$J = \sum_{q=1}^{r}\sum_{i=1}^{n} \|h_i^q(\mathbf{x}_i) - U_{qi}\| = \sum_{q=1}^{r} \|\mathbf{S}\mathbf{u}_q - \mathbf{u}_q\|_2^2 = \mathrm{Tr}[\mathbf{U}\mathbf{M}\mathbf{U}^T] \quad (12)$$

Here $\mathbf{M} = (\mathbf{S} - \mathbf{I}_n)(\mathbf{S} - \mathbf{I}_n)$ and $\mathbf{S}$ is defined: $S_{ij} = g_{ij}$ if $\mathbf{x}^j \in \mathcal{N}_k(\mathbf{x}^i)$ and $S_{ij} = 0$ otherwise. The local discriminative structure can be well maintained, which can also address the overfitting introduced by the noisy tagging information.

By jointly incorporating the deep MF model, local semantic and visual structures, the proposed framework is formulated as the following unified objective function.

$$\begin{aligned}\min_{\mathbf{V},\mathbf{W}_M,\cdots,\mathbf{W}_1} & \frac{1}{2}\|\mathbf{F} - \mathbf{V}\mathbf{U}\|_F^2 + \frac{\alpha}{2}\mathrm{Tr}[\mathbf{U}_{M-1}\mathbf{T}\mathbf{U}_{M-1}^T] \\ & + \frac{\beta}{2}\mathrm{Tr}[\mathbf{V}^T\mathbf{L}\mathbf{V}] + \frac{\mu}{2}\mathrm{Tr}[\mathbf{U}\mathbf{M}\mathbf{U}^T] \\ & + \frac{\lambda_1}{2}(\|\mathbf{V}\|_F^2 + \sum_{m=1}^{M}\|\mathbf{W}_m\|_F^2) + \frac{\lambda_2}{2}\|\mathbf{W}_1\|_{2,1} \\ \text{s.t.} \quad & \mathbf{W}_m^T\mathbf{W}_m = \mathbf{I}_{r_m} \end{aligned} \quad (13)$$

Here $\mathbf{U} = \mathbf{W}_M \cdots \mathbf{W}_1\mathbf{X}$ and r_m is the number of nodes in the m-th layer, where $r_1 = d$. α, β and μ are three positive trade-off parameters. The orthogonal constraints restrict $\mathbf{W}_m$ to converge to reasonable solutions in practice. λ_1 is a regularization parameter to avoid fitting, and λ_2 is a regularization parameter to control the sparsity in columns of $\mathbf{W}_1$. Since the visual features are often correlated or redundant to each other, and sometimes noisy, the sparse model in columns with $\ell_{2,1}$ mixed norm is introduced, which ensures $\mathbf{W}_1$ sparse in columns. It can compress the redundant or noisy features.

By exploiting the weakly-supervised tagging information, the prior of semantic consistency and the local discriminative structure in the visual space simultaneously under the deep framework, the proposed WDMF method can learn a better unified subspace, in which images and tags are embedded. Image tagging refinement is implemented by identifying the nearest tags in the latent subspace. And the coming images can be embedded in the subspace and associated with tags.

D. Optimization

The joint optimization problem in Eq. 13 is not convex over all the variables $\mathbf{V}$ and $\mathbf{W}_m$ $(1 \le m \le M)$ simultaneously. Thus, we propose an iterative optimization algorithm using the sub-gradient descent scheme for local optimal solutions. For ease of representation, we use notation $\mathcal{O}$ to denote the objective function in Eq. 13. The variables $\mathbf{V}$ and $\mathbf{W}_m$ $(1 \le m \le M)$ are alternately updated by fixing other variables.

First, $\mathbf{V}$ is solved with $\mathbf{W}_m$ $(1 \le m \le M)$ fixed. The derivative of $\mathcal{O}$ *w.r.t.* $\mathbf{V}$ is calculated as follows.

$$\frac{\partial \mathcal{O}}{\partial \mathbf{V}} = \mathbf{E}\mathbf{U}^T + \beta\mathbf{L}\mathbf{V} + \lambda_1\mathbf{V} \quad (14)$$

Here $\mathbf{E} = \mathbf{V}\mathbf{U} - \mathbf{F}$. And then $\mathbf{V}$ is updated by using the following rule with the learning rate η.

$$\mathbf{V} = \mathbf{V} - \eta\frac{\partial \mathcal{O}}{\partial \mathbf{V}} \quad (15)$$

Then, we solve $\mathbf{W}_m$ by fixing $\mathbf{V}$. Due to the orthogonal constraints, $\mathbf{W}_m$ is updated using a gradient descent procedure with curvilinear search [59] in this work. The derivatives of $\mathcal{O}$ with respect to $\mathbf{W}_m$ are obtained.

$$\begin{aligned}\frac{\partial \mathcal{O}}{\partial \mathbf{W}_1} = & \prod_{i=2}^{M}\mathbf{W}_i^T\mathbf{V}^T\mathbf{E}\mathbf{X}^T + \alpha\prod_{i=2}^{M-1}\mathbf{W}_i^T\mathbf{U}_{M-1}\mathbf{T}\mathbf{X}^T \\ & + \mu\prod_{i=2}^{M}\mathbf{W}_i^T\mathbf{U}\mathbf{M}\mathbf{X}^T + \lambda_1\mathbf{W}_1 + \lambda_2\mathbf{W}_1\mathbf{D} \end{aligned} \quad (16)$$

$$\begin{aligned}\frac{\partial \mathcal{O}}{\partial \mathbf{W}_m} = & \prod_{i=m+1}^{M}\mathbf{W}_i^T\mathbf{V}^T\mathbf{E}\mathbf{U}_{m-1}^T \\ & + \alpha\prod_{i=m+1}^{M-1}\mathbf{W}_i^T\mathbf{U}_{M-1}\mathbf{T}\mathbf{U}_{m-1}^T \\ & + \mu\prod_{i=m+1}^{M}\mathbf{W}_i^T\mathbf{U}\mathbf{M}\mathbf{U}_{m-1}^T + \lambda_1\mathbf{W}_m \end{aligned} \quad (17)$$

$$\begin{aligned}\frac{\partial \mathcal{O}}{\partial \mathbf{W}_{M-1}} = & \mathbf{W}_M^T\mathbf{V}^T\mathbf{E}\mathbf{U}_{M-2}^T + \alpha\mathbf{U}_{M-1}\mathbf{T}\mathbf{U}_{M-2}^T \\ & + \mu\mathbf{W}_M^T\mathbf{U}\mathbf{M}\mathbf{U}_{M-2}^T + \lambda_1\mathbf{W}_{M-1} \end{aligned} \quad (18)$$

$$\frac{\partial \mathcal{O}}{\partial \mathbf{W}_M} = \mathbf{V}^T\mathbf{E}\mathbf{U}_{M-1}^T + \mu\mathbf{U}\mathbf{M}\mathbf{U}_{M-1}^T + \lambda_1\mathbf{W}_M \quad (19)$$

Here $\mathbf{D} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with $D_{ii} = \frac{1}{2\|\mathbf{w}^i\|_2}$. It is worth noting that in practice, $\|\mathbf{w}_i\|_2$ could be close to zero but not zero. Theoretically, it could be zeros. For this case, we can regularize $D_{ii} = \frac{1}{2(\|\mathbf{w}^i\|_2+\epsilon)}$, where ϵ is very small constant. In each iteration of the gradient descent procedure, letting $\mathbf{G}_m = \frac{\partial \mathcal{O}}{\partial \mathbf{W}_m}$, we can obtain the following skew-symmetric matrix.

$$\mathbf{Z}_m = \mathbf{G}_m\mathbf{W}_m^T - \mathbf{W}_m\mathbf{G}_m^T \quad (20)$$

Then the new solution can be searched as a curvilinear function of a step size τ:

$$\mathbf{Y}_m(\tau) = (\mathbf{I} + \frac{\tau}{2}\mathbf{Z}_m)^{-1}(\mathbf{I} - \frac{\tau}{2}\mathbf{Z}_m)\mathbf{W}_m \quad (21)$$

From the theoretical analysis in [59], $(\mathbf{Y}_m(\tau))^T\mathbf{Y}_m(\tau) = \mathbf{I}$ for any $\tau \in \mathbb{R}$ and the derivative with respect to τ is

$$\mathbf{Y}_m^{'}(\tau) = -(\mathbf{I} + \frac{\tau}{2}\mathbf{Z}_m)^{-1}\mathbf{Z}_m\frac{\mathbf{W}_m + \mathbf{Y}_m(\tau)}{2} \quad (22)$$

When $\tau = 0$, we have $\mathbf{Y}_m^{'}(0) = -\mathbf{Z}_m\mathbf{W}$. Hence, $\mathbf{Y}_m(\tau)$ is a descent curve when $\tau > 0$. To identify a suitable step τ, the Armijo-Wolfe based monotone curvilinear search algorithm [59] is introduced to find one satisfying the following conditions.

$$\mathcal{O}(\mathbf{Y}_m(\tau)) \le \mathcal{O}(\mathbf{Y}_m(0)) + \rho_1\tau\mathcal{O}_\tau^{'}(\mathbf{Y}_m(0)) \quad (23)$$

$$\mathcal{O}_\tau^{'}(\mathbf{Y}_m(\tau)) \ge \rho_2\mathcal{O}_\tau^{'}(\mathbf{Y}_m(0)) \quad (24)$$

Here $0 < \rho_1 < \rho_2 < 1$ are two parameters. $\mathcal{O}_\tau^{'}(\mathbf{Y}_m(\tau))$ is the derivative of $\mathcal{O}$ with respect to τ.

$$\mathcal{O}_\tau^{'}(\mathbf{Y}_m(\tau)) = -\mathrm{Tr}[(\partial\mathcal{O}(\mathbf{Y}_m(\tau))/\partial\mathbf{W}_m)^T\mathbf{Y}_m^{'}(\tau)] \quad (25)$$

$$\mathcal{O}_\tau^{'}(\mathbf{Y}_m(0)) = -\frac{1}{2}\|\mathbf{Z}_m\|_F^2 \quad (26)$$

Algorithm 1 The Proposed WDMF Algorithm

Input:

Visual feature matrix $\mathbf{X}$, the tagging matrix $\mathbf{F}$, the number of network layers M, learning rate η, $0 < \varepsilon < 1$ and $0 < \rho_1 < \rho_2 < 1$.

Calculate $\mathbf{T}$, $\mathbf{L}$ and $\mathbf{M}$ according to $\mathbf{X}$ and $\mathbf{F}$;
Initialize $\mathbf{V}$ and $\mathbf{W}_m$ $(1 \leq m \leq M)$; Set $\mathbf{D}$ as an identity matrix;
repeat
  //Forward propagation
  for $m = 1, 2, \cdots, M$
    Do forward propagation to get $\mathbf{U}_m$;
  end
  //Computing gradient
  Compute gradient according to Eq. 14;
  for $m = M, M-1, \cdots, 1$
    Compute $\mathbf{Z}_m$ according to Eq. 20;
    $\tau = 1$;
    repeat
      $\tau = \varepsilon\tau$;
      Compute $\mathbf{Y}_m(\tau)$ via Eq. 21;
    until Armijo-Wolfe conditions satisfied
  end
  //Back propagation
  Update $\mathbf{V}$ according to Eq. 15;
  for $m = 1, 2, \cdots, M$
      Update $\mathbf{W}_m$ according to Eq. 27;
  end
  Update the diagonal matrix $\mathbf{D}$;
until Convergence criterion satisfied

Output:

The latent matrix $\mathbf{V}$, and transformation matrices $\mathbf{W}_m$.

When the suitable value τ is found, $\mathbf{W}_m$ is updated by using the following rule.

$$\mathbf{W}_m = \mathbf{Y}_m(\tau) \tag{27}$$

Algorithm 1 summarizes the detailed procedure of the proposed WDMF approach. The utilized convergence criterion is that the number of iterations is more than N_t or $|\mathcal{O}_{t-1} - \mathcal{O}_t|/\mathcal{O}_{t-1} < \xi$, where $\mathcal{O}_t$ is the value of the objective function in the t-th iteration.

E. Implementation

To expedite the proposed method and obtain better approximation of the latent factors, the designed deep architecture is first pre-trained, which can obtain initial matrices of $\mathbf{V}$ and $\mathbf{W}_m$. It has been widely implemented on deep networks, which can shorten the learning time. For the pre-training, we first utilize the regress model to learn $\mathbf{V}_1$: $\min_{\mathbf{V}_1} \|\mathbf{F} - \mathbf{V}_1\mathbf{X}\|_F^2 + \lambda_1\|\mathbf{V}_1\|_F^2$. With the learned $\mathbf{V}_1$, we then decompose it: $\mathbf{V}_1 \leftarrow \mathbf{V}_2\mathbf{W}_1$. Now, we obtain $\mathbf{W}_1$ and $\mathbf{V}_2$, and then further decompose $\mathbf{V}_2$. Following this, we have $\mathbf{V}_M \leftarrow \mathbf{V}\mathbf{W}_M$. By now, all the layers have been pre-trained. Afterwards, each layer is fine-tuned by the proposed optimization algorithm in **Algorithm 1**.

TABLE I

STATISTICS OF THE USED DATASETS WITH IMAGE AND TAG COUNTS IN THE FORMAT MEAN / MAXIMUM

	MIRFlickr	**NUS-WIDE**
Tag size	457	3137
Concept size	18	81
Image size	25,000	269,648
Tags per image	2.7 / 45	7.9 / 201
Concepts per image	4.7 / 17	1.9 / 13
Images per tag	145.4 / 1,483	677.1 / 20,140
Images per concept	3,102.8 / 10,373	6,220.3 / 74,190

IV. EXPERIMENTS

In this section, extensive experiments on two real-world datasets are carried out for image tag refinement, image tag assignment and tag-based image retrieval. The experimental results empirically validate the effectiveness of the proposed WDMF framework, and real the necessity of the deep architecture and the structure preservation.

A. Datasets

Social image sharing sites allow users to freely upload, tag and comment images, which generates massive images associated with user-provided tags. Therefore, social image datasets can be easily built for experimental purpose. In this work, we conduct extensive experiments on two widely utilized social image datasets: MIRFlickr [60] and NUS-WIDE [61]. Table I summarizes some statistics of these data sets.

MIRFlickr [60]. The MIRFlickr dataset has 25,000 images associated with 1,386 user-provided tags collected from Flickr. It also provides the ground-truth annotation of 38 concepts to evaluate the performance. Since there are some obviously noisy tags, we select tags that appear at least 50 times, and obtain a vocabulary consisting of 457 tags. Accordingly, only 18 concepts are preserved and utilized to validate the performance in our experiments. To describe the visual content, two types of global descriptors (Gist features and color histograms with 16 bins in each color channel for LAB and HSV representations) and one type of local feature (SIFT feature) are utilized. The features are available at http://lear.inrialpes.fr/data/.

NUS-WIDE [61]. The NUS-WIDE dataset is a large-scale community-contributed image dataset as a benchmark to evaluate multimedia tasks. There are $269,648$ images associated with $5,018$ tags annotated by amateur users. It provides the ground-truth annotations of 81 concepts, which are used to evaluate the performance. Note that these 81 concepts are different from the user-provided tags since they are manually labeled. To reduce too noisy tags, we removed those tags whose occurrence numbers are below 125. Consequently, 3,137 unique tags including 81 concepts were obtained. To represent the visual content of images, 1,134-D features provided by the dataset are used.

Data are randomly partitioned into two groups in our experiments: the learning data for image tag refinement and the testing data for image tag assignment. n images are randomly chosen as the learning data while the rest ones are used as the testing data. The learning data is utilized to learn the proposed

model and evaluate the performance of image tag refinement. The testing images are utilized to validate the effectiveness of image tag assignment. In our experiments, we set $n = 20,000$ and $n = 50,000$ for the MIRFlickr and NUS-WIDE datasets, respectively. To alleviate the instability introduced by the randomly partition, experiments are independently repeated 5 times to generate different learning and testing data, and report the average values of all the results. The experimental results on the learning data and the testing data are both reported.

B. Evaluation Metrics

To compare the effectiveness of methods, in our experiments we adopt the area under the receiver operating characteristic (ROC) curve, known as the AUC, as evaluation metric. AUC is currently considered to be the standard method for model comparison and a more faithful criterion used in many applications. Both the microaveraging and macroaveraging measures are utilized to evaluate both the global performance across multiple concepts and the average performance of all the concepts. The concept indicator vectors of all concepts are first concatenated and then the average AUC is computed to calculate the microaveraging result. The macroaveraging result is obtained by averaging the mean AUC values of all the concepts. F1 is also introduced to measure the performance for image tagging. Please refer to [44] for more details. Besides, the ranking information provided by tag refinement and assignment is important for tag-based image retrieval. To evaluate the ranking order, we analyze experimental results with single-tag queries. Average Precision (AP) is the standard measure used for retrieval benchmark. It corresponds to the average of the precision at each position where a relevant image appears. Mean Average Precision (MAP) over concepts is obtained by averaging average precision over all concepts.

C. Compared Algorithms

In order to validate the performance of our method, we compare the proposed WDMF method with one baseline and a number of related state-of-the-art algorithms. The compared methods are enumerated as follows.

- **OT**: The original user-provided tags from Flickr as the baseline.
- **MF**: The traditional matrix factorization model in Eq. 2.
- **LR**: Tags of images are refined by the low-rank matrix decomposition without considering the content consistency and tag correlation in [13].
- **MPMF** [7]: Multiple correlations are jointly exploited by multi-correlation probabilistic matrix factorization algorithm for image annotation. In our experiments, we adopt it for image tag refitment.
- **LRES** [13]: Tags of images are refined by the low-rank matrix decomposition with the constraints of content consistency and tag correlation.
- **C2MR** [16]: The latent semantic space is modeled under the low-rank matrix approximation framework by considering both context and content correlation.

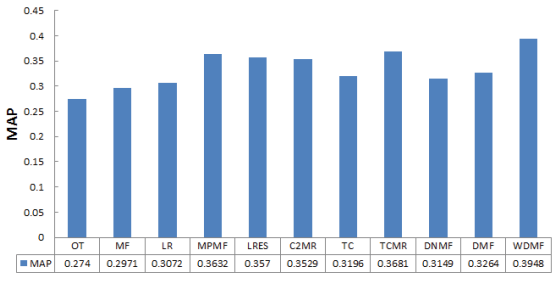


Fig. 4. Image retrieval results on the MIRFlickr dataset in terms of MAP.

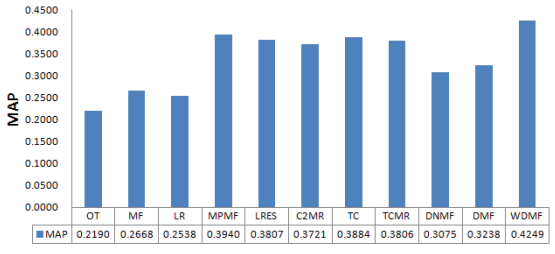


Fig. 5. Image retrieval results on the NUS-WIDE dataset in terms of MAP.

- **TC** [17]: The tags of images are refined to be consistent with the observed tags and the visual similarity.
- **TCMR** [32]: Image tag completion is performed by uncovering a low-rank tagging matrix from the noisy matrix.
- **DNMF** [53]: It learns hidden representations by factorizing the tagging matrix in a multi-layer structure.
- **DMF** [55]: The preliminary version of the proposed deep matrix factorization method without considering the data structures, i.e., $\alpha = \beta = \mu = 0$.
- **WDMF**: The proposed deep matrix factorization with the structure preservation for image tag refinement and assignment.

For all the compared methods except the baseline method, there are some parameters to be set in advance. All the parameters corresponding to the overfitting regularization terms in the compared algorithms are set to 0.005. For the other hyper-parameters, we adopt the same parameter configuration as described in their original reports. For the proposed WDMF method, we set $\lambda_1 = 0.005$, $\lambda_2 = 0.01$ and empirically set the learning rate to 0.001. For all the matrix factorization methods, the dimension of the latent subspace r is empirically set to 50. A deep network with three layers ($M = 3$) are learned and the numbers of nodes on the first, second and last layers are set to $\frac{2d}{3}$, $\frac{d}{3} + \frac{r}{2}$ and r, respectively. For the curvilinear search, we set $\rho_1 = 0.25$, $\rho_2 = 0.85$ and $\varepsilon = 0.3$. The parameters α, β and μ are tuned using the learning data. We will discuss the parameter setting in more details in Subsection IV-G.

D. Experimental Results for Tag Refinement

In the first set of experiments, different algorithms are evaluated for the task of image tag refinement. The corresponding experimental results in terms of the mean MicroAUC and mean MacroAUC on the MIRFlickr and NUS-WIDE datasets

TABLE II

EXPERIMENTAL RESULTS (MEAN MICROAUC ± STANDARD DEVIATION, MEAN MACROAUC ± STANDARD DEVIATION AND MEAN F1 ± STANDARD DEVIATION) ON THE MIRFLICKR AND NUS-WIDE DATASETS FOR IMAGE TAG REFINEMENT. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Method	MIRFlickr			NUS-WIDE		
	MicroAUC	MacroAUC	F1	MicroAUC	MacroAUC	F1
OT	0.587 ± 0.001	0.558 ± 0.001	0.274 ± 0.002	0.703 ± 0.003	0.619 ± 0.004	0.264 ± 0.007
MF	0.617 ± 0.005	0.588 ± 0.006	0.286 ± 0.001	0.724 ± 0.006	0.680 ± 0.002	0.273 ± 0.002
LR	0.634 ± 0.005	0.603 ± 0.004	0.288 ± 0.004	0.704 ± 0.004	0.693 ± 0.003	0.284 ± 0.005
MPMF	0.651 ± 0.004	0.642 ± 0.004	0.438 ± 0.002	0.751 ± 0.003	0.726 ± 0.002	0.327 ± 0.005
LRES	0.653 ± 0.005	0.646 ± 0.008	0.444 ± 0.003	0.770 ± 0.005	0.744 ± 0.007	0.340 ± 0.006
C2MR	0.655 ± 0.003	0.640 ± 0.004	0.374 ± 0.006	0.774 ± 0.002	0.745 ± 0.002	0.348 ± 0.003
TC	0.640 ± 0.003	0.637 ± 0.006	0.426 ± 0.010	0.759 ± 0.004	0.653 ± 0.003	0.354 ± 0.007
TCMR	0.651 ± 0.002	0.633 ± 0.004	0.402 ± 0.006	0.779 ± 0.008	0.728 ± 0.006	0.364 ± 0.005
DNMF	0.639 ± 0.007	0.632 ± 0.005	0.294 ± 0.003	0.746 ± 0.006	0.736 ± 0.004	0.304 ± 0.011
DMF	0.647 ± 0.006	0.641 ± 0.004	0.378 ± 0.003	0.754 ± 0.002	0.741 ± 0.003	0.329 ± 0.003
WDMF	$\mathbf{0.696 \pm 0.006}$	$\mathbf{0.676 \pm 0.004}$	$\mathbf{0.467 \pm 0.003}$	$\mathbf{0.804 \pm 0.002}$	$\mathbf{0.781 \pm 0.009}$	$\mathbf{0.389 \pm 0.004}$

TABLE III

EXPERIMENTAL RESULTS (MEAN MICROAUC ± STANDARD DEVIATION, MEAN MACROAUC ± STANDARD DEVIATION AND MEAN F1 ± STANDARD DEVIATION) ON THE MIRFLICKR AND NUS-WIDE DATASETS FOR IMAGE TAG ASSIGNMENT. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Method	MIRFlickr			NUS-WIDE		
	MicroAUC	MacroAUC	F1	MicroAUC	MacroAUC	F1
OT	0.576 ± 0.002	0.561 ± 0.003	0.224 ± 0.009	0.693 ± 0.003	0.564 ± 0.002	0.126 ± 0.003
MF	0.603 ± 0.001	0.582 ± 0.006	0.238 ± 0.004	0.694 ± 0.002	0.567 ± 0.008	0.138 ± 0.001
LR	0.614 ± 0.003	0.598 ± 0.007	0.248 ± 0.003	0.705 ± 0.002	0.574 ± 0.007	0.134 ± 0.004
MPMF	0.630 ± 0.001	0.606 ± 0.004	0.279 ± 0.003	0.739 ± 0.003	0.654 ± 0.004	0.165 ± 0.002
LRES	0.629 ± 0.001	0.613 ± 0.003	0.271 ± 0.001	0.720 ± 0.008	0.642 ± 0.006	0.174 ± 0.003
C2MR	0.637 ± 0.003	0.624 ± 0.003	0.283 ± 0.002	0.723 ± 0.007	0.655 ± 0.003	0.180 ± 0.004
TC	0.625 ± 0.004	0.607 ± 0.007	0.284 ± 0.010	0.714 ± 0.005	0.638 ± 0.007	0.159 ± 0.006
TCMR	0.628 ± 0.001	0.587 ± 0.005	0.268 ± 0.004	0.740 ± 0.003	0.641 ± 0.004	0.163 ± 0.005
DNMF	0.627 ± 0.003	0.612 ± 0.005	0.258 ± 0.003	0.719 ± 0.005	0.592 ± 0.004	0.153 ± 0.003
DMF	0.636 ± 0.004	0.621 ± 0.005	0.281 ± 0.003	0.741 ± 0.004	0.634 ± 0.005	0.178 ± 0.003
WDMF	$\mathbf{0.652 \pm 0.005}$	$\mathbf{0.636 \pm 0.006}$	$\mathbf{0.337 \pm 0.004}$	$\mathbf{0.760 \pm 0.002}$	$\mathbf{0.692 \pm 0.009}$	$\mathbf{0.213 \pm 0.004}$

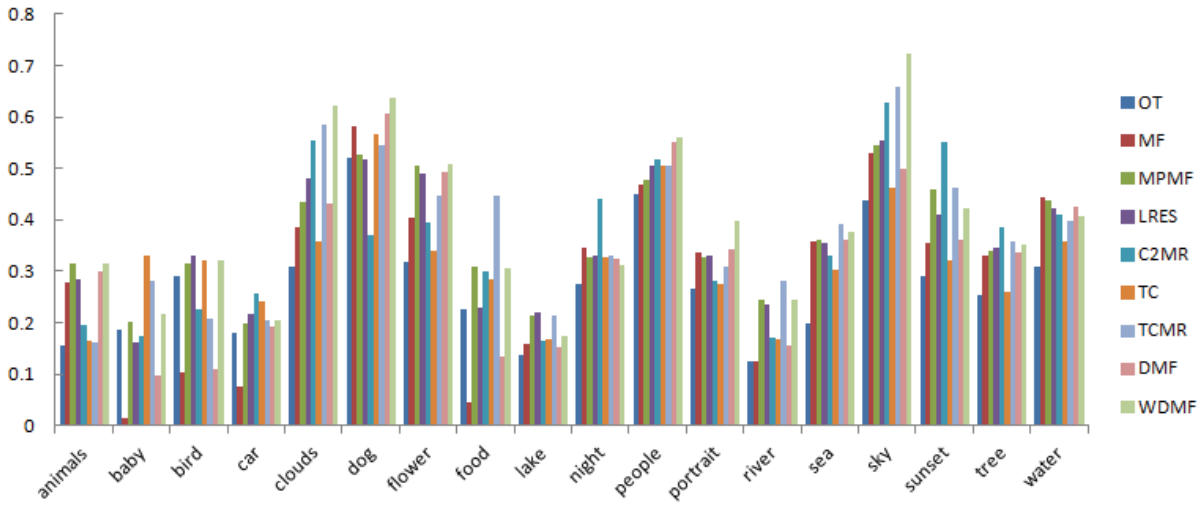
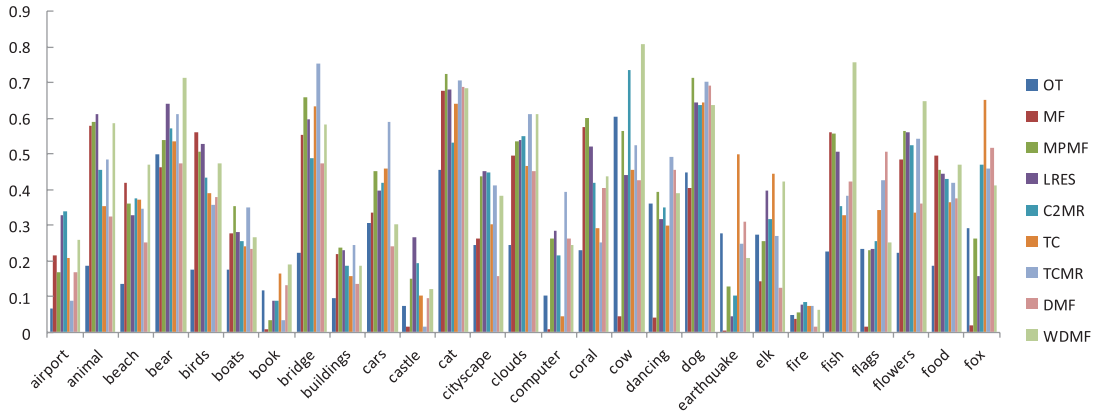


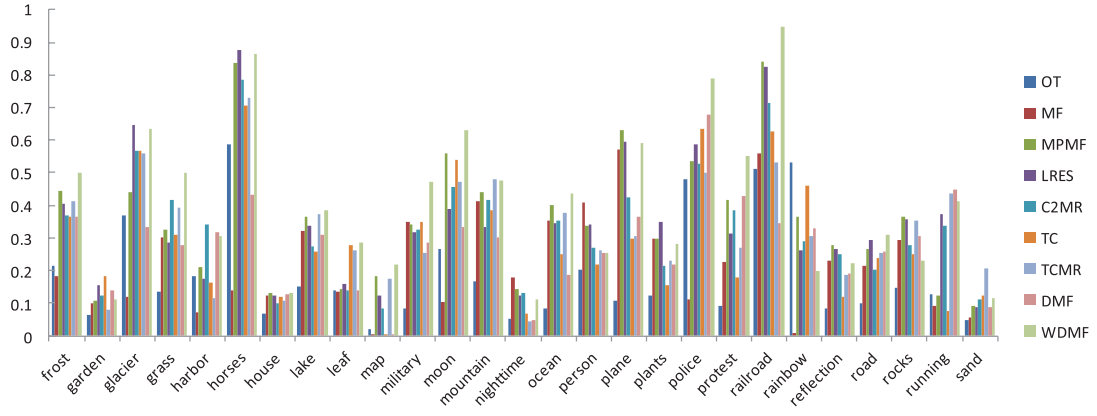
Fig. 6. The detailed performance comparison in terms of AP over 18 concepts on the MIRFlickr dataset.

are presented in Table II. From the above experimental results, the following interesting observations are revealed.

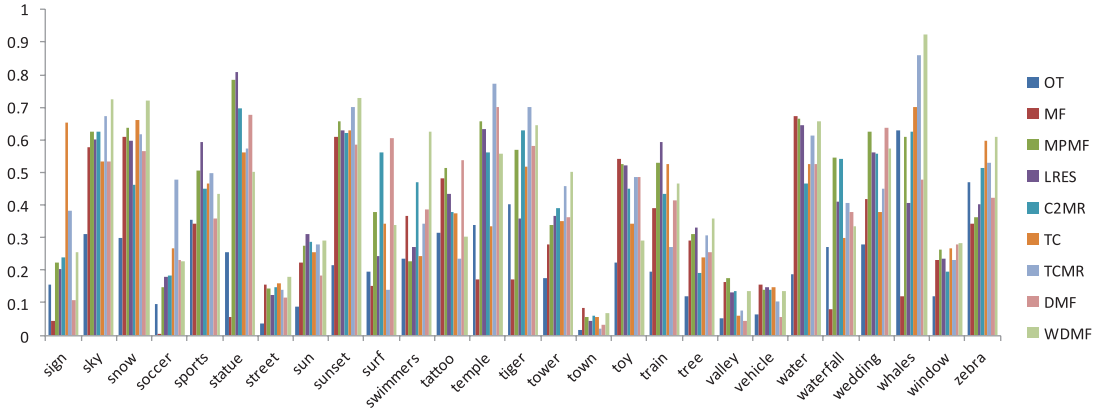
- The proposed method achieves the best performance for image tag refinement, which well demonstrates the effectiveness of the proposed method. Because it can well exploit the user-provided tagging information, local textual and semantic structures simultaneously.
- Compared with the original tags, the latent factor models achieve better results. It demonstrates that the latent factor models enable to complete the image tagging matrix.
- By jointly considering the tag correlation and image visual similarity, MPMF and LRES are superior to MF and LR, respectively, which indicates the importance of the prior of the semantic and visual consistencies.
- By introducing the deep decomposition framework, the performance is significantly improved, which is indicated by comparing MF and DMF.
- It is beneficial for social tag refinement to exploit the latent visual and semantic structures. It is also verified by the improvement of WDMF over DMF.



(a)



(b)



(c)

Fig. 7. The detailed performance comparison in terms of AP over 81 concepts on the NUS-WIDE dataset. (a) Airport - fox. (b) Frost - sand. (c) Sign - zebra.

- Experimental results demonstrates that the proposed WDMF outperforms MPMF, LRES, C2MR, TC and TCMR. The reason is that the proposed method under the deep framework can well handle the gap between the low-level visual features and the high-level semantic and the introduced $\ell_{2,1}$ mixed norm

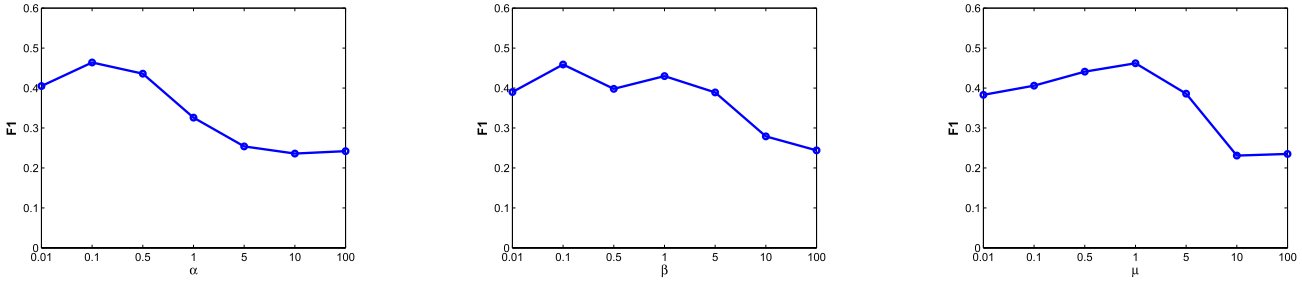


Fig. 8. The parameter sensitiveness of α, β and μ in terms of F1 on the MIRFlickr dataset. (a) α. (b) β. (c) μ.

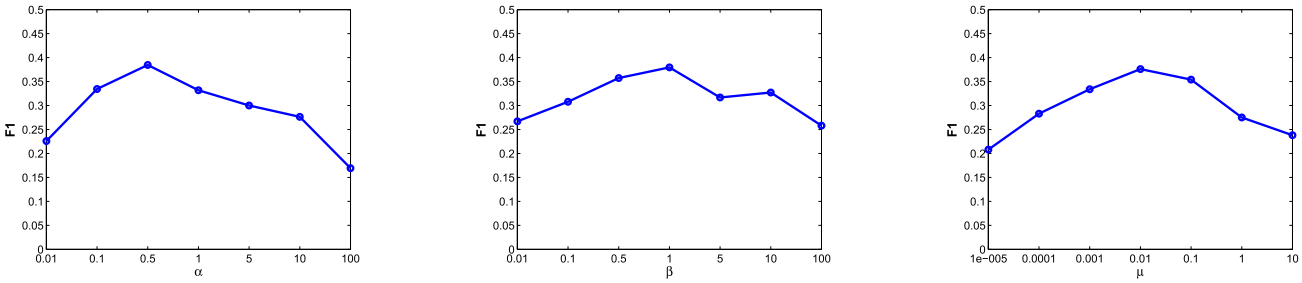


Fig. 9. The parameter sensitiveness of α, β and μ in terms of F1 on the NUS-WIDE dataset.(a) α. (b) β. (c) μ.

can also deal with the noisy or redundant visual features.

E. Experimental Results for Tag Assignment

With the learned latent factors $\mathbf{V}$ and $\mathbf{W}_m (1 \leq m \leq M)$, we can estimate the relationships between any image and tags, that is, we can assign the relevant semantic tags to any image. Now in this section, experiments are conducted on the MIRFlickr and NUS-WIDE datasets to verify the effectiveness of the proposed method for tag assignment. The corresponding quantitative results are shown in Table III.

From the compared results, it can be seen that the proposed method achieves the best performance on both the MIRFlickr and NUS-WIDE datasets for the task of image tag assignment. We can also observe the conclusions in the above section. In addition, by directly learning the transformation matrices which transform the visual features to the latent subspace, the proposed method is significantly effective to assign relevant tags to new images without any tag. That is, the motivation of this work is empirically verified.

F. Experimental Results for Image Retrieval

To better empirically evaluate the effectiveness of the proposed method, we conduct experiments on these two datasets for the task of tag-based image retrieval. The performance is measured in terms of MAP. The quantitative results on the MIRFlickr and NUS-WIDE datasets are presented in Figure 4 and Figure 5, respectively. From the experimental results, we can draw the observations as follows. First, the proposed method also achieves the best results on both datasets for image retrieval. Second, DMF outperforms DNMF, which shows that DMF is more suitable for social image analysis. Third, by leveraging the data structures, the performance of image retrieval increases, which is indicated by comparing WDMF vs. DMF and LRES vs. LR. Finally, we also present the detailed performances in terms of APs on the MIRFlickr and NUS-WIDE datasets over the 18 and 81 concepts in Figure 6 and Figure 7, respectively. Since LRES and WMF are superior to LR and WNMF, the results of LR and WNMF are not presented in Figure 6 and Figure 7. The advantages of the proposed method are demonstrated again.

G. Parameter Sensitivity Analysis

The proposed formulation contains several regularization terms and the corresponding hyper-parameters. It is a common problem to set the parameters to tradeoff each component in a joint optimization objective. In the previous experiments, the hyper-parameters are optimally set in order to evaluate the effectiveness of the proposed method for social image tag refinement, assignment and retrieval. Now, experiments are conducted to study the sensitivity analysis of the hyper-parameters towards image tag refinement on the learning data.

The corresponding results by varying α, β and μ on the MIRFlickr and NUS-WIDE datasets are illustrated in Figure 8 and Figure 9, respectively. The performance is measured using F1. From the experimental results, we observe that the it is necessary to leverage the underlying structures for uncovering the latent subspace of images and tags. Besides, it can be seen that best results are achieved with $\alpha = 0.1$, $\beta = 0.1$, $\mu = 1$ and $\alpha = 0.5$, $\beta = 1$, $\mu = 0.01$ on the MIRFlickr and NUS-WIDE datasets, respectively. If they are too large or too small, the performance dramatically

decreases. That is, each regularization term has its contribution in improving the performance of social image analysis, which also verifies our motivations to introduce these constraints.

V. Conclusion

In this work, we proposed a novel Weakly-supervised Deep Matrix Factorization (WDMF) algorithm for social image tag refinement, assignment and retrieval, which uncovers the latent image representations and tag representations embedded in the latent subspace by collaboratively exploiting the weakly-supervised tagging information, the visual structure and the semantic structure. To well handle the out-of-sample problem, the underlying image representations are assumed to be progressively transformed from the visual feature space. Besides, the proposed approach can deal with the noisy, incomplete or subjective tags and the noisy or redundant visual features. The proposed problem is formulated as a joint optimization problem with a well-defined objective function, which is solved by a gradient descent procedure with curvilinear search. Extensive experiments on two real-world social image databases are conducted to demonstrate the effectiveness of the problem. There are several potential research directions. First, we will explore the effective of the depth in the deep MF framework and how to adaptively identify better parameters for different vision tasks. Second, we will explore how to integrate the proposed deep MF model and CNN into a unified framework. Besides, we will extend the proposed deep MF framework to make it applicable and investigate its new applications such as image suggestion.

References

[1] *Flickr*. [Online]. Available: http://www.flickr.com

[2] *Facebook*. [Online]. Available: http://www.facebook.com

[3] A. Jeffries. The Verge, *The Man Behind Flickr on Making the Service 'Awesome Again'*. (Mar. 2013). [Online]. Available: http://www.theverge.com/2013/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer

[4] L. Xu *et al.*, "Multi-task rank learning for image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, 2016, doi: 10.1109/TCSVT.2016.2543099.

[5] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 97–112.

[6] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, Mar. 2003.

[7] Z. Li, J. Liu, X. Zhu, T. Liu, and H. Lu, "Image annotation using multi-correlation probabilistic matrix factorization," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1187–1190.

[8] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "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] R. Wong and C. Leung, "Automatic semantic annotation of real-world Web image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1933–1944, Nov. 2008.

[10] J. Tang, Z.-J. Zha, D. Tao, and T.-S. Chua, "Semantic-gap-oriented active learning for multilabel image annotation," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2354–2360, Apr. 2012.

[11] Z. Li, J. Liu, C. Xu, and H. Lu, "Mlrank: Multi-correlation learning to rank for image annotation," *Pattern Recognit.*, vol. 10, no. 46, pp. 2700–2710, 2013.

[12] X. Li, C. G. M. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1310–1322, Nov. 2009.

[13] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 461–470.

[14] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain, "Image annotation by kNN-sparse graph-based label propagation over noisily tagged Web images," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, p. 14, Feb. 2011.

[15] J. Zhuang and S. C. Hoi, "A two-view learning approaches for image tag ranking," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2011, pp. 625–634.

[16] G.-J. Qi, C. Aggarwal, Q. Tian, H. Ji, and T. S. Huang, "Exploring context and content links in social media: A latent space method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 850–862, May 2011.

[17] L. Wu, R. Jin, and A. K. Jain, "Tag completion for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 716–727, Mar. 2013.

[18] Z. Lin, G. Ding, M. Hu, J. Wang, and X. Ye, "Image tag completion via image-specific and tag-specific linear sparse reconstructions," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1618–1625.

[19] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.

[20] Z. Li, J. Liu, J. Tang, and H. Lu, "Projective matrix factorization with unified embedding for social image tagging," *Comput. Vis. Image Understand.*, vol. 124, pp. 71–78, Jul. 2014.

[21] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[22] Z. Li and J. Tang, "Unsupervised feature selection via nonnegative spectral analysis and redundancy control," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5343–5355, Dec. 2015.

[23] A. Ulges, M. Worring, and T. Breuel, "Learning visual contexts for image annotation from flickr groups," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 330–341, Feb. 2009.

[24] D. Liu, S. Yan, X.-S. Hua, and H.-J. Zhang, "Image retagging using collaborative tag propagation," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 702–712, Apr. 2011.

[25] Z. Li, J. Liu, Y. Jiang, J. Tang, and H. Lu, "Low rank metric learning for social image retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 853–856.

[26] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 64–78, Jan. 2015.

[27] S. Qian, T. Zhang, R. Hong, and C. Xu, "Cross-domain collaborative learning in social multimedia," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 99–108.

[28] H. Zhang, X. Shang, H.-B. Luan, Y. Yang, and T.-S. Chua, "Learning features from large-scale, noisy and social image-tag collection," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 1079–1082.

[29] Z. Li and J. Tang, "Weakly supervised deep metric learning for community-contributed image retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1989–1999, Nov. 2015.

[30] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. M. Snoek, and A. D. Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval," *ACM Comput. Surv.*, vol. 49, no. 1, p. 14, 2016.

[31] Y.-H. Kuo, W.-H. Cheng, H.-T. Lin, and W. H. Hsu, "Unsupervised semantic feature discovery for image object retrieval and tag refinement," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1079–1090, Apr. 2012.

[32] Z. Feng, S. Feng, R. Jin, and A. K. Jain, "Image tag completion by noisy matrix recovery," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 424–438.

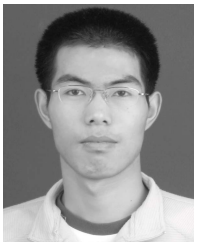
[33] A. Sun, S. S. Bhowmick, and J.-A. Chong, "Social image tag recommendation by concept matching," in *Proc. ACM Int. Conf. Multimedia*, 2011, pp. 1181–1184.

[34] J. Wang, J. Zhou, H. Xu, T. Mei, X.-S. Hua, and S. Li, "Image tag refinement by regularized latent Dirichlet allocation," *Comput. Vis. Image Understand.*, vol. 124, pp. 61–70, Jul. 2014.

[35] B. Q. Truong, A. Sun, and S. S. Bhowmick, "Content is still king: The effect of neighbor voting schemes on tag relevance for social image retrieval," in *Proc. ACM Int. Conf. Multimedia Retr.*, vol. 9. 2012, pp. 1–8.

[36] A. Znaidia, H. L. Borgne, and C. Hudelot, "Tag completion based on belief theory and neighbor voting," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2013, pp. 49–56.

[37] J. Yu, X. Jin, J. Han, and J. Luo, "Collection-based sparse label propagation and its application on social group suggestion from photos," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, pp. 1–21, 2011.
[38] D. Rafailidis, A. Axenopoulos, J. Etzold, S. Manolopoulou, and P. Daras, "Content-based tag propagation and tensor factorization for personalized item recommendation based on social tagging," *ACM Trans. Interact. Intell. Syst.*, vol. 3, no. 4, pp. 1–27, 2014.
[39] X. Zhu, W. Nejdl, and M. Georgescu, "An adaptive teleportation random walk model for learning social tag relevance," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2014, pp. 223–232.
[40] P. Wu, S. C.-H. Hoi, P. Zhao, and Y. He, "Mining social images with distance metric learning for automated image tagging," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2011, pp. 197–206.
[41] C. Chen, Q. Zhu, L. Lin, and M.-L. Shyu, "Web media semantic concept retrieval via tag removal and model fusion," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 4, p. 61, 2013.
[42] Q. Wang, B. Shen, S. Wang, L. Li, and L. Si, "Binary codes embedding for fast image tagging with incomplete labels," in *Proc. Eur. Conf. Comput. Vis. II*, 2014, pp. 425–439.
[43] X. Qian, X.-S. Hua, Y. Y. Tang, and T. Mei, "Social image tagging with diverse semantics," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2493–2508, Dec. 2014.
[44] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2085–2098, Oct. 2015.
[45] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. Neural Inf. Process. Syst.*, 2007, pp. 1257–1264.
[46] N. Wang, T. Yao, J. Wang, and D.-Y. Yeung, "A probabilistic approach to robust matrix factorization," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 126–139.
[47] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov–Chain Monte–Carlo," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 880–887.
[48] 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.
[49] E. Esser, M. Möller, S. Osher, G. Sapiro, and J. Xin, "A convex model for nonnegative matrix factorization and dimensionality reduction on physical space," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3239–3252, Jul. 2012.
[50] S. K. Gupta, D. Phung, B. Adams, T. Tran, and S. Venkatesh, "Non-negative shared subspace learning and its application to social media retrieval," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 1169–1178.
[51] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized non-negative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.
[52] L. Jing, C. Zhang, and M. K. Ng, "SNMFCA: Supervised NMF-based image classification and annotation," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4508–4521, Nov. 2012.
[53] G. Trigeorgis, K. Bousmalis, S. Zafeurios, and B. W. Schuller, "A deep semi-nmf model for learning hidden representations," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1692–1700.
[54] X. He, D. Cai, and P. Niyogi, "Tensor subspace analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 499–506.
[55] Z. Li and J. Tang, "Deep matrix factorization for social image tag refinement and assignment," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Oct. 2015, pp. 1–6.
[56] F. R. K. Chung, *Spectral Graph Theory* (MS Regional Conference Series in Mathematics). Providence, RI, USA: American Mathematical Society, 1996, no. 92.
[57] L. Bottou and V. Vapnik, "Local learning algorithms," *Neural Comput.*, vol. 4, no. 6, pp. 888–900, 1992.
[58] B. Schölkopf and A. Smola, *Learning With Kernels* (MS Regional Conference Series in Mathematics), vol. 92. Cambridge, MA, USA: MIT Press, 2002.
[59] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Program.*, vol. 142, no. 1, pp. 397–434, 2013.
[60] M. Huiskes and M. Lew, "The MIRFLICKR retrieval evaluation," in *Proc. ACM Int. Conf. Multimedia Inf. Retr.*, 2008, pp. 39–43.
[61] J. Tang, X. Shu, Z. Li, G.-J. Qi, and J. Wang, "Generalized deep transfer networks for knowledge propagation in heterogeneous domains," *ACM Trans. Multimedia Comput. Commun. Appl.*, 2016.



Zechao Li (M'10) the B.E. degree from the University of Science and Technology of China in 2008 and received the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2013. He is currently an Associate Professor with the Nanjing University of Science and Technology, China. His research interests include large-scale multimedia understanding, social media mining, and subspace learning. He received the 2015 Excellent Doctoral Dissertation of Chinese Academy of Sciences, the Excellent Doctoral Theses of China Computer Federation, the Top 10% Paper Award of the IEEE MMSP 2015, and the 2013 President Scholarship of Chinese Academy of Science.

Jinhui Tang (SM'14) received the B.E. and Ph.D. degrees from the University of Science and Technology of China in 2003 and 2008, respectively. He is currently a Professor with the Nanjing University of Science and Technology. His current research interests include large-scale multimedia search, social media mining, and computer vision. He has authored over 100 papers in these areas. He serves as an Editorial Board Member of Pattern Analysis and Applications, Multimedia Tools and Applications, Information Sciences, Neurocomputing, and a Technical Committee Member for about 30 international conferences. He is a co-recipient of the Best Paper Award in the ACM Multimedia 2007, the PCM 2011 and the ICIMCS 2011. He is a member of the ACM.