Neurocomputing 105 (2013) 38–44

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Nonlinear matrix factorization with unified embedding for social tag relevance learning

Zechao Li, Jing Liu*, Hanqing Lu

National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, No. 95, Zhongguancun East Road, Beijing 100080, PR China

ARTICLE INFO

Available online 24 October 2012

Keywords:
Matrix factorization
Unified embedding
Tag relevance
Social image retrieval

ABSTRACT

With the proliferation of social images, social image tagging is an essential issue for text-based social image retrieval. However, the original tags annotated by web users are always noisy, irrelevant and incomplete to interpret the image visual contents. In this paper, we propose a nonlinear matrix factorization method with the priors of inter- and intra-correlations among images and tags to effectively predict the tag relevance to the visual contents. In the proposed method, we attempt to discover the image latent feature space and the tag latent feature space in a unified space, that is, each image or each tag can be described as a point in the unified space. Intuitively, it is more understandable to estimate the relationships between images and tags directly based on their distances or similarities in the unified space. Thus, the task of image tagging or tag recommendation can be efficiently solved by the nearest tag-neighbors search in the unified space. Similarly, we can obtain the top relevant images corresponding to any tag so as to perform the task of image search by keywords. We investigate the performance of the proposed method on tag recommendation and image search respectively and compare to existing work on the challenging NUS-WIDE dataset. Extensive experiments demonstrate the effectiveness and potentials of the proposed method in real-world applications.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

In the web 2.0 era, with the development of Internet technologies and digital devices, image sharing websites such as Flickr and Facebook are increasingly popular. Users cannot only easily upload, distribute and share their digital images and photos, but also tag and comment on their interested images. As a consequence, text-based social image retrieval has become an emerging popular yet rather challenging research topic.

However, due to the diversity of knowledge and cultural background of users, social tagging is often subjective and inaccurate. Consequently, many images are usually not tagged with proper tags, and even completely untagged. The tags associated with social images could be noisy, irrelevant and incomplete as shown in Fig. 1, which may severely deteriorate the performance of text-based image retrieval [1]. Existing studies reveal that many tags provided by Flickr users are imprecise and there are only around 50% tags actually related to the image [2,1]. Hence, a fundamental problem for text-based social image retrieval is how to rank the tags for any given image by the relevance of tags with respect to the visual content.

In this paper, we investigate the tag relevance learning problem to address the above challenge. Fig. 1 shows an exemplary image from Flickr, from which we can see that there are some irrelevant, noisy and incomplete tags. After tag relevance learning, the tags are adjusted and some relevant tags are added. Many approaches have been proposed to tackle the tag relevance learning problem [3–6,2,7,8,1]. The most related work is the multi-correlation probabilistic matrix factorization (MPMF) model [7], which is based on a latent factor model. The image-tag relation matrix is decomposed to two latent feature matrices and the image similarity and tag correlation matrices are exploited simultaneously and seamlessly by the shared latent matrices. It is a linear Gaussian model and the latent factors can be embedded in the different spaces.

Unlike the existing matrix factorization work, this paper proposes a nonlinear matrix factorization approach with unified embedding (MFUE) to learn the tag relevance for social image retrieval. The image latent features and tag latent features are embedded in a unified space and the distance represents the relevance. Compared to the standard matrix factorization, MFUE can also scale to the number of observations and track the sparse data. On the other hand, the structure of latent features embedding in the same space is more intuitive to understand the relationship between images and tags. The new images with few or no tag can be easily mapped into the unified space and

* Corresponding author.
E-mail address: jliu@nlpr.ia.ac.cn (J. Liu).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.02.046

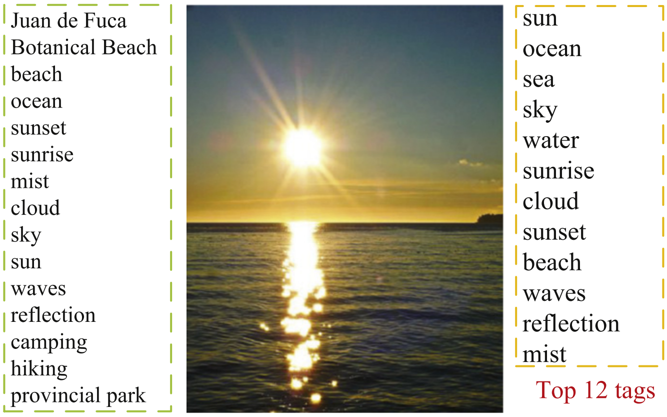


Fig. 1. Illustration of the tag relevance learning. There are many imprecise and meaningless tags in the original tag list. After tag relevance learning, some relevant tags are added and the relevant tags are ranked in the top position.

recommended relevant tags using the nearest neighbor searches. Finally, in the process of matrix factorization, the visual similarity and tag correlation are jointly investigated by the shared latent feature vectors to preserve the visual and semantic local geometry properties. We conduct an extensive set of experiments to evaluate the empirical performance of the proposed MFUE method with the application of social image retrieval and tag recommendation.

The reminder of this paper is organized as follows. We review related work in Section 2. Section 3 elaborates the proposed nonlinear matrix factorization with unified embedding algorithm. In Section 4, extensive experiments are conducted to evaluate the performance of the proposed method and compare it to other related methods. The conclusion of this paper with future work discussion is presented in Section 5.

2. Related work

It is an essential issue to estimate the relevance of tags with respect to images in text-based image retrieval. The related techniques are categorized into two main scenarios, namely tag annotation for untagged images and tag refinement for tagged images.

Methods in the first category predicts relevant tags for images with no tag. A variety of methods have been proposed to annotate images automatically [9–18,7], which can be categorized into two main types, *generative models* and *classification models*. The generative models try to estimate the probabilistic relationship between tags and images. By assigning relevant scores of tags to images, the annotated results can be utilized to help the task of image retrieval.

In the second scenario, given an image labeled with some tags, tag relevance learning can be used to remove noisy tags, recommend new relevant tags or reduce tag ambiguity. Many approaches have been proposed to tackle the tag relevance learning problem [3–6,2,8,19,1]. The random walk with restarts (RWR) algorithm [3] is proposed to leverage the co-occurrence-based tag similarity and the information of the original annotated order of tags. The tag refinement problem is formulated as a Markov process and the candidate tags are treated as the states in [4]. In [5], a neighbor voting algorithm is proposed to estimate a tag's relevance by exploiting tagging redundancies among multiple users. The tag relevance is determined based on the number of such votes from the nearest neighbors. Tag ranking [2] further exploits pairwise similarity between tags by random walk to refine the ranking score. In [7,8,1], both the image similarity and tag correlation are exploited simultaneously to discover the tag relevance. A multi-correlation probabilistic matrix factorization (MPMF) model [7] is proposed to combine the inter- and intra- correlation matrices by the shared latent matrices. In [8], the image labels are refined by decomposing the observed label matrix into a low-rank refined matrix and a sparse error matrix. A two-view learning approach is proposed to address the tag ranking problem in [1].

Different from the previous work, this paper presents a novel non-linear matrix factorization approach to estimate the relevance of tags to social images. The distance is strongly correlated with the relevance between tags and images. The local visual geometry in image space and local textual geometry in tag space are exploited simultaneously. This method can be employed to tag and refine images.

3. Nonlinear matrix factorization with unified embedding

To estimate the relationship between tags and images, we jointly exploit three aspects: matrix factorization, local visual geometry preserving and local textual geometry preserving. In this section, we first present the formulation of the proposed methods with some preliminaries. We then elaborate each part of the objective function and discuss the optimization.

3.1. Formulation

Consider a set of social images $\mathcal{I}=\{x_1,x_2,\ldots,x_n\}$. All initial tags appearing in the collection form a tag set $\mathcal{T}=\{t_1,t_2,\ldots,t_m\}$. For any matrix $\mathbf{A}$, let $\mathbf{a}_i$, $\mathbf{A}_{ij}$, $\mathbf{A}^T$, $\|\mathbf{A}\|$ and Tr[$\mathbf{A}$] denote the ith column vector, the (i,j)th entry, the transpose, the Frobenius and the trace of $\mathbf{A}$ if $\mathbf{A}$ is square, respectively. The tagging records can be represented as an tag-image association matrix $\mathbf{R}\in\mathcal{R}^{m\times n}$, where $R_{ij}=1$ indicates that image x_j is tagged with the tag t_i and $R_{ij}=0$ means that the association is unknown.

For a given tag matrix $\mathbf{R}$, we try to discover the latent image matrix $\mathbf{U}\in\mathcal{R}^{d\times n}$ and the latent tag matrix $\mathbf{V}\in\mathcal{R}^{d\times m}$ embedded in a unified space with the dimension d, which are utilized to fit the observation values by the distance in the unified space. On the other hand, the local visual geometry in image space and the local textual geometry in tag space are incorporated by considering the image similarity matrix $\mathbf{S}$ and tag correlation matrix $\mathbf{C}$ simultaneously. Therefore, our approach is formulated to minimize the following objective function:

$$\mathcal{L}(\mathbf{U},\mathbf{V})=l_1(\mathbf{R},\mathbf{U},\mathbf{V})+\alpha l_2(\mathbf{U},\mathbf{S})+\beta l_3(\mathbf{V},\mathbf{C}), \tag{1}$$

where $l_1(\mathbf{R},\mathbf{U},\mathbf{V})$ measure the estimative loss by the nonlinear matrix factorization with unified embedding. We use $l_2(\mathbf{U},\mathbf{S})$ and $l_3(\mathbf{V},\mathbf{C})$ to measure the local visual geometry and the local textual geometry preservations respectively. The intra-correlations have been considered by many work [7,8,1,20]. α and β are two non-negative trade-off parameters. In the following subsection, we will elaborate on how to define these three items.

3.2. Nonlinear matrix factorization

The social image data consists of two media: the visual media and the textual media. It is revealed that multimedia objects tend to obey nonlinear distribution [21] and the nonlinear learning strategy has been studied in multimedia and computer vision [22]. Therefore, we explore the nonlinear model for matrix factorization. As mentioned above, the proposed nonlinear matrix factorization model assumes that all images and tags are embedded in a unified space. The distance is strongly correlated with the relevance of a tag to an image. Therefore, if a tag is close to an image in the unified space, it can describe the semantic information of the image. In this paper, we adopt the Gaussian kernel to transform the distance to the relevance.

Thus, in our nonlinear matrix factorization framework, the observation value R_{ij} can be estimated by

$$R_{ij} \approx \hat{R}_{ij} = e^{-\|\mathbf{u}_j - \mathbf{v}_i\|^2/2\sigma^2}. \tag{2}$$

Here, σ is a free parameter to control the decay rate, and $\mathbf{u}_j$ and $\mathbf{v}_i$ are vectors of the j-th image and the i-th tag in the d-dimensional unified space. $\|\mathbf{u}_j - \mathbf{v}_i\|^2 = (\mathbf{u}_j - \mathbf{v}_i)^T(\mathbf{u}_j - \mathbf{v}_i)$. If they are close in the low-dimensional space, that is, $\hat{R}_{ij}$ is close to 1, the tag is relevant to the image. By adopting Gaussian kernel, the estimated values are within [0,1].

For all images and tags, the objective function to approximate the tag matrix $\mathbf{R}$ is as follows:

$$\min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{m} \sum_{j=1}^{n} (R_{ij} - e^{-\|\mathbf{u}_j - \mathbf{v}_i\|^2/2\sigma^2})^2. \tag{3}$$

To avoid overfitting, due to the distance depends on the relative position of $\mathbf{u}_j$ to $\mathbf{v}_i$ in the embedding space rather than the absolute position of them, we restrict the magnitude of $\mathbf{u}_j - \mathbf{v}_i$ instead of individual points. Therefore, the first term corresponding to matrix factorization in Eq. (1) is the following equation:

$$l_1(\mathbf{R},\mathbf{U},\mathbf{V}) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} \left[\frac{1}{2} (R_{ij} - e^{-\|\mathbf{u}_j - \mathbf{v}_i\|^2/2\sigma^2})^2 + \gamma_1 \|\mathbf{u}_j - \mathbf{v}_i\|^2 \right], \tag{4}$$

$\gamma_1 > 0$ is an algorithmic parameter.

3.3. Local geometry in visual space

To preserve the local visual geometry in image space, the latent image points in the low-dimensional space are supposed to be close to each other if their corresponding images are similar to each other. The distance between the j-th image and the k-th image in the embedded space is measured by $\|\mathbf{u}_j - \mathbf{u}_k\|_2$. The local visual geometry is preserved by minimizing the following distortion:

$$\sum_{j=1}^{n} \sum_{k=1}^{n} (S_{jk} - e^{-\|\mathbf{u}_j - \mathbf{u}_k\|^2/2\sigma^2})^2. \tag{5}$$

Here we adopt the distance between two images in the latent space to approximate their visual similarity. A regularization term is also added to control the magnitude of the Euclidean distance between images in the latent space. As a consequence, the second term in Eq. (1) is defined as

$$l_2(\mathbf{U},\mathbf{S}) = \frac{1}{4} \sum_{j=1}^{n} \sum_{k=1}^{n} \left[\frac{1}{2} (S_{jk} - e^{-\|\mathbf{u}_j - \mathbf{u}_k\|^2/2\sigma^2})^2 + \gamma_2 \|\mathbf{u}_j - \mathbf{u}_k\|^2 \right]. \tag{6}$$

γ_2 is a non-negative regularization parameter.

The predefined similarity matrix $\mathbf{S}$ is the prior knowledge about the data distribution and encodes the local visual geometric information in visual space. In this paper, $\mathbf{S}$ is calculated preliminarily as follows.

As revealed in [23,24], minimizing ℓ_1 norm over the weights enables to suppress the noise contained in data. The constructed graph is more robust than other graph construction strategies and is non-parametric. Additionally, considerable tag-unrelated links between those semantically unrelated images can be removed by the sparse reconstruction to reduce the incorrect information. Therefore, in our implementation, we adopt the linear reconstruction based on sparse coding to define $\mathbf{S}$, similar to [23].

Under the linear reconstruction assumption, a sample is reconstructed by other samples using the following linear equation:

$$\mathbf{x}_i = \mathbf{X}\mathbf{w}_i, \tag{7}$$

where $\mathbf{x}_i$ is the visual feature vector of the i-th image to be reconstructed, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n]$ is used as the over-complete dictionary and $\mathbf{w}_i$ is the vector of the unknown reconstruction coefficient. In practice, probably noises exist in the features and a natural way to recover these elements and provide a robust estimation of $\mathbf{w}_i$ is to formulate $\mathbf{x}_i = \mathbf{X}\mathbf{w}_i + \boldsymbol{\eta} = \mathbf{B}\xi$, where $\boldsymbol{\eta}$ is the sparse noise vector, $\mathbf{B} = [\mathbf{X}\ \mathbf{I}]$ and $\boldsymbol{\xi} = [\mathbf{w}_i; \boldsymbol{\eta}]$. We can then solve the following ℓ_1-norm minimization problem with respect to both reconstruction coefficients and data noises:

$$\min_{\xi} \|\xi\|_1 \quad \text{s.t. } \mathbf{x}_i = \mathbf{B}\xi. \tag{8}$$

This optimization problem is convex and can be transformed into a general linear programming problem. There exists a globally optimal solution, and the optimization can be solved efficiently using many available ℓ_1-norm optimization toolboxes like Least Angle Regression (LAR) algorithm [25]. Based on the reconstruction coefficient matrix $\mathbf{W}$, $\mathbf{S}$ is defined as $\mathbf{S} = \max(\mathbf{W}, \mathbf{W}^T)$.

3.4. Local geometry in concept space

Similarly, we calculate the tag correlation $\mathbf{C}$ in tag space, and have the local textual geometry preserving objective function

$$l_3(\mathbf{V},\mathbf{C}) = \frac{1}{4} \sum_{i=1}^{m} \sum_{l=1}^{m} \left[\frac{1}{2} (C_{il} - e^{-\|\mathbf{v}_i - \mathbf{v}_l\|^2/2\sigma^2})^2 + \gamma_3 \|\mathbf{v}_i - \mathbf{v}_l\|^2 \right]. \tag{9}$$

Here $\gamma_3 \geq 0$ is to avoid overfitting.

Similar to $\mathbf{S}$, $\mathbf{C}$ contains the local geometric information in tag space. To utilize this prior information about the tag distribution, the tag correlation matrix $\mathbf{C}$ should be defined. To estimate the tag correlation between tags t_i and t_j, we first calculate $corr(t_i, t_j)$, the number of images where the two tags co-occur. Flickr distance [26] is to globally measure tag correlation based on the web source. In our work, we adopt the local correlations. As tags that appear very often in the dataset tend to co-occur more frequently that most of the other words in the vocabulary, the tag correlation is defined by normalizing $corr(t_i, t_j)$ by the tag frequency

$$C_{ij} = \frac{corr(t_i, t_j)}{corr(t_i, t_i) + corr(t_j, t_j) - corr(t_i, t_j)}. \tag{10}$$

3.5. Implementation

Based on the definitions of the terms regarding matrix factorization, local visual geometry and local semantic geometry, the objective function in Eq. (1) can be rewritten as

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \mathcal{L}(\mathbf{U},\mathbf{V}) = {} & \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} \left[\frac{1}{2} (R_{ij} - e^{-\|\mathbf{u}_j - \mathbf{v}_i\|^2/2\sigma^2})^2 + \gamma_1 \|\mathbf{u}_j - \mathbf{v}_i\|^2 \right] \\ & + \frac{\alpha}{4} \sum_{j=1}^{n} \sum_{k=1}^{n} \left\{ \frac{1}{2} [S_{jk} - e^{-\|\mathbf{u}_j - \mathbf{u}_k\|^2/2\sigma^2}]^2 + \gamma_2 \|\mathbf{u}_j - \mathbf{u}_k\|^2 \right\} \\ & + \frac{\beta}{4} \sum_{i=1}^{m} \sum_{l=1}^{m} \left\{ \frac{1}{2} [C_{il} - e^{-\|\mathbf{v}_i - \mathbf{v}_l\|^2/2\sigma^2}]^2 + \gamma_3 \|\mathbf{v}_i - \mathbf{v}_l\|^2 \right\}. \end{aligned} \tag{11}$$

Since the latent features are in the same space, the parameters to control the magnitude of the distance are set to be the same. That is, we set $\gamma_1 = \alpha\gamma_2 = \beta\gamma_2 = \gamma$. $\mathcal{L}(\mathbf{U},\mathbf{V})$ can be rewritten as

$$\min_{\mathbf{U},\mathbf{V}} \mathcal{L}(\mathbf{U},\mathbf{V}) = \frac{1}{4} \sum_{i=1}^{m} \sum_{j=1}^{n} (R_{ij} - e^{-\|\mathbf{u}_j - \mathbf{v}_i\|^2/2\sigma^2})^2$$

$$+\frac{\alpha}{8}\sum_{j=1}^{n}\sum_{k=1}^{n}(S_{jk}-e^{-\|\mathbf{u}_j-\mathbf{u}_k\|^2/2\sigma^2})^2$$
$$+\frac{\beta}{8}\sum_{i=1}^{m}\sum_{l=1}^{m}(C_{il}-e^{-\|\mathbf{v}_i-\mathbf{v}_l\|^2/2\sigma^2})^2+\frac{\gamma}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\|\mathbf{u}_j-\mathbf{v}_i\|^2$$
$$+\frac{\gamma}{4}\sum_{j=1}^{n}\sum_{k=1}^{n}\|\mathbf{u}_j-\mathbf{u}_k\|^2+\frac{\gamma}{4}\sum_{i=1}^{m}\sum_{l=1}^{m}\|\mathbf{v}_i-\mathbf{v}_l\|^2. \tag{12}$$

To solve this optimization problem, the gradient descent algorithm is adopted to update **U** and **V**

$$\frac{\partial \mathcal{L}}{\partial \mathbf{u}_j}=\sum_{i=1}^{m}H_{ij}(\mathbf{u}_j-\mathbf{v}_i)+\alpha\sum_{k=1}^{n}P_{jk}(\mathbf{u}_j-\mathbf{u}_k), \tag{13}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}_i}=\sum_{j=1}^{n}H_{ij}(\mathbf{v}_i-\mathbf{u}_j)+\beta\sum_{l=1}^{m}Q_{il}(\mathbf{v}_i-\mathbf{v}_l). \tag{14}$$

Here $H_{ij}=(1/2\sigma^2)(R_{ij}-\hat{R}_{ij})\hat{R}_{ij}+\gamma$ and $\hat{R}_{ij}=e^{-\|\mathbf{u}_j-\mathbf{v}_i\|^2/2\sigma^2}$. $P_{ij}=(1/2\sigma^2)(S_{ij}-\hat{S}_{ij})\hat{S}_{ij}+\gamma$ and $\hat{S}_{jk}=e^{-\|\mathbf{u}_j-\mathbf{u}_k\|^2/2\sigma^2}$. $Q_{ij}=(1/2\sigma^2)(C_{ij}-\hat{C}_{ij})\hat{C}_{ij}+\gamma$ and $\hat{C}_{il}=e^{-\|\mathbf{v}_i-\mathbf{v}_l\|^2/2\sigma^2}$.

In our approach, a new image x_o can be incorporated into the model. We find its nearest neighbors $\{x_{o_1},x_{o_2},\ldots,x_{o_K}\}$ in image space, and compute the similar weights $\{w_{o_1},w_{o_2},\ldots,w_{o_K}\}$ by ℓ_1-norm optimization. To preserve the local visual geometry, the new image can be reconstructed by its nearest neighbors in the embedding space. Therefore, we can estimate the latent image vector by

$$\mathbf{u}_o=\sum_{i=1}^{K}\frac{w_{o_i}\mathbf{u}_{o_i}}{\sum_{j=1}^{K}w_{o_j}}. \tag{15}$$

Therefore, for any tag t, we can estimate its relevance to the image as

$$R_{to}=e^{-\|\mathbf{u}_o-\mathbf{v}_t\|^2/2\sigma^2}. \tag{16}$$

Based on the estimated relevancies, we can recommend top T tags with highest relevancies to the image.

4. Experimental analysis

To validate the effectiveness of our proposed approach on tag relevance learning, we conduct extensive experiments, and apply our method to text-based social image retrieval and automatic image recommendation. All of the experiments are implemented via MATLAB on a 2.39 GHz PC with 16 GB RAM.

4.1. Experimental setting

We conduct experiments on the real-world image dataset NUS-WIDE-Lite [27], which contains 55,615 images with 5018 unique tags. For feature representation, we extract four types of global features: 64-D color histogram (LAB), 144-D color autocorrelation (HSV), 73-D edge direction histogram and 128-D wavelet texture. For local feature, we use grid-based features: 225-D block-wise color moments (LAB). Thus, we sequentially combine these five groups into 634-D features. To evaluate the performance, we evaluate the performance on 81 concepts in NUS-WIDE-Lite where the ground-truth annotations of these tags have been provided. The criteria to compare the performance include *Average Precision* (AP) for each concept and *Mean Average Precision* (MAP) for all concepts. MAP is obtained by averaging the APs on 81 concepts. For tag recommendation, Precision, Recall and MAP are adopted.

For experimental setting, there are several parameters used in our algorithm. For most of our results, we set the dimensionality d of the embedding space to 300 and σ is set by a "grid-search" strategy [28] in the set $\{2^{-8},2^{-7.5},\ldots,2^{2}\}$. The trade-off parameters α and β in Eq. (12) are set to 0.001 and 0.01 empirically. $\gamma=0.005$ is set to avoid overfitting in Eq. (12). For tag recommendation, we set $T=10$.

To evaluate the performance of the proposed MFUE method, we compared it extensively with the following methods:

- **OT**: i.e., the original tags associated with images.
- **CIAR**: the tag refinement algorithm of content-based image annotation refinement proposed in [4].
- **PRW**: the tag ranking method proposed in [2], which can be viewed as a combination of the Probabilistic tag ranking and random walk-based tag ranking.
- **TRNV**: the tag relevance by neighbor voting learning algorithm [5].
- **TWTV**: the two-view tag weighting method combines the local information in the tag space and visual space [1].
- **MPMF**: the proposed multi-correlation probabilistic matrix factorization method in [7];
- **LRES**: tag refinement based on low-rank approximation and error sparsity with content-tag prior while considering the content consistency and tag consistency simultaneously [8].

4.2. Results of social image retrieval

In this experiment, we compare the proposed algorithm with seven baseline algorithms. For each method, we report its best results by tuning its parameters. MAP of these eight methods is illustrated in Table 1. We also present the detailed performances for the 81 concepts in Fig. 3.

First of all, it is observed that our proposed algorithm MFUE works well. It produces very competitive results with MPMF and LRES. Actually it achieves the best performance. MPMF, LRES and MFUE are superior to other methods, because they exploit inter- and intra- correlations among images and tags simultaneously. Second, all the tag relevance learning methods outperform OT significantly, which verifies the necessity. This also demonstrates that the relevant tags may not be placed at the top position. Finally, using factor analysis, MFUE, MPMF and LRES performs well on the sparse data, which coincides with the motivation.

4.3. Sensitiveness of parameters

There are several free parameters in our method. α and β control the trade-off between the information of local visual geometric information and local semantic geometric information. d controls the dimension of the unified embedding space. They may be the most important parameters and should be tuned to the sensitiveness of them.

Fig. 2(a) shows the influence of the values α and β. From the results, several interesting observations can be gained. First, the retrieval performance varies with different values of α and β, which indicates the information of local visual geometry and tag geometry is useful. Second, when $\alpha=0.001$ and $\beta=0.01$, the best MAP is achieved. Finally, when $\alpha=0$ or $\beta=0$, the performance is

Table 1
The results of the comparison of MFUE and the seven baselines.

Methods	OT	CAIR	PRW	TRNV	TWTV	MPMF	LRES	MFUE
MAP	0.3876	0.4064	0.4367	0.4117	0.4469	0.4504	0.4519	**0.4565**

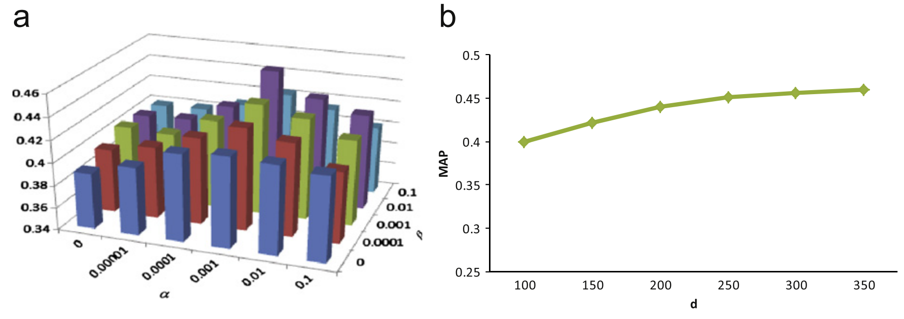


Fig. 2. MAPs of MFUE with the varying parameters of (a) α and β, and (b) the dimensionality of the unified space d.

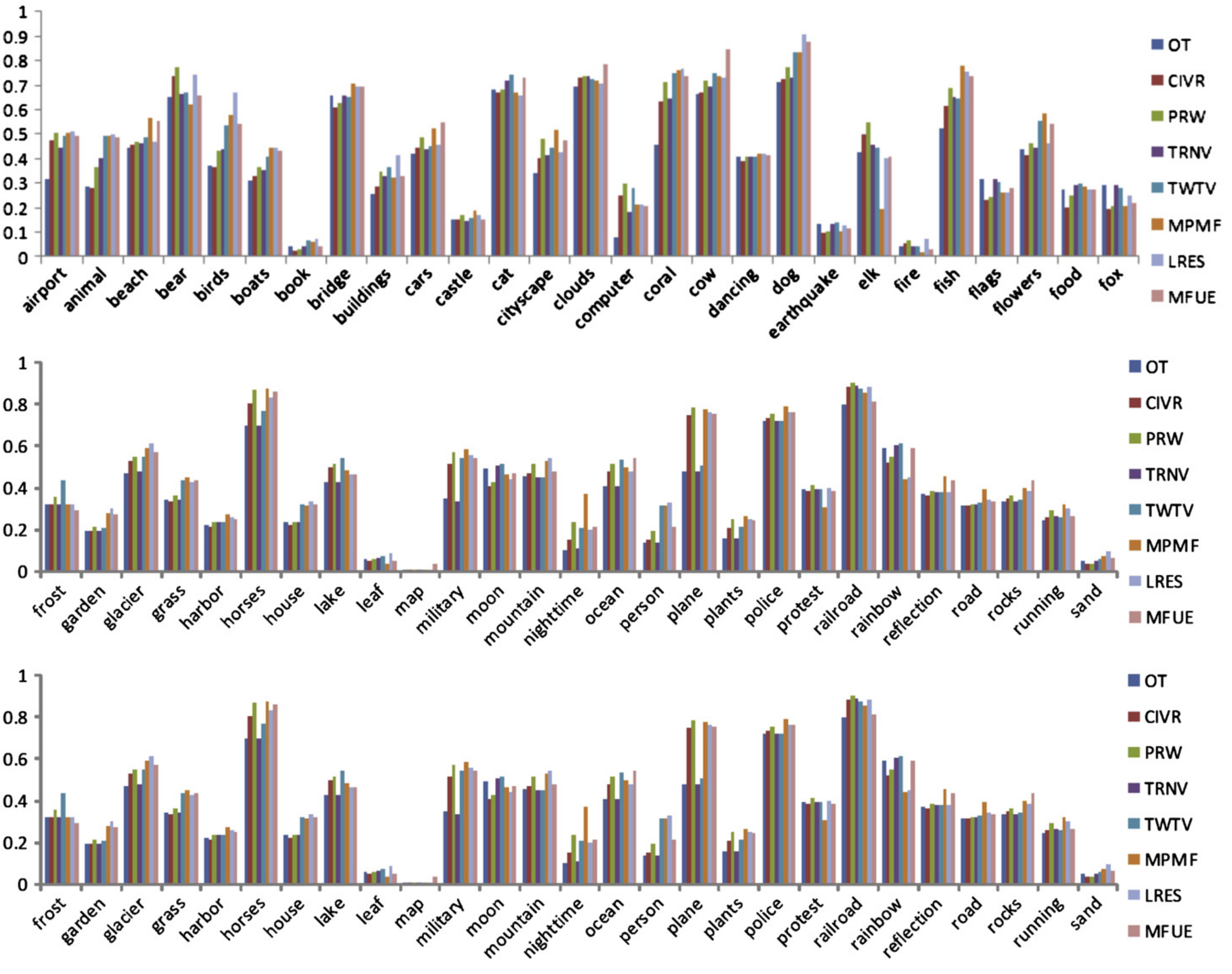


Fig. 3. The comparisons of 81 concepts using eight methods.

poor. The results demonstrate the effectiveness of the complementary property of visual and tag space.

Fig. 2(b) illustrates the results of implementing MFUE while varying the dimensionality of the unified space. From the results, we observe that the MAP is improved by increasing the dimensionality of the unified space to some extent. Considering high dimensional corresponds to expensive computing cost, we set $d=300$ to leverage the performance and the cost.

Table 2
Precision, recall and MAP for tag recommendation by MFUE-n and MFUE-a. The MFUE-n refers to mapping new samples into the space learned by the rest samples as in Section 3.5. MFUE-a denotes that the space is learned by all samples.

Methods	Precision	Recall	MAP
MFUE-n	0.301	0.346	0.417
MFUE-a	0.312	0.357	0.451

Recommended Tags	water sea girl people female beach young outdoor single pink	mast wind sky solar windmills energy sun windmill panel arc	sky smokestack chimney smoke sunrise pollution sun building factory red
Top 10 Original Tags	water people sea beach travel girl summer pink woman love	wind energy mast arc hull solar	sky snow sunrise smokestack
Recommended Tags	jet plane sky airplane blue aircraft military yellow clouds	ship water tugboat ocean boat vehicle cloud sky bay blue	zoo lion animal grass green mountain sky wild zoos wildlife
Top 10 Original Tags	sky blue yellow plane california view jet airplane aircraft merge	blue water clouds ship tugboat bay california color pacific harbor	travel zoo vocation lion wildcat family holiday africa trip zoos

Fig. 4. An illustration of tag recommendation with the proposed MFUE method. Best viewed in color.

4.4. Tag recommendation

In this subsection, we adopt tag recommendation to indicate the effectiveness for tag recommendation and incorporating new images. For each image, the top 10 tags are recommended, that is $T=10$.

As discussed earlier, we could map the images in the existing space jointly exploiting visual and textual information. To evaluate its effectiveness, we randomly select 5615 images as new images. The space is learned via the rest images. The results are shown in Table 2. For sake of comparison, we also present the results of the regular setting when the space is learned by all images. We can see that our method performs very well for new images. Some examples by applying the proposed method for tag recommendation are shown in Fig. 4. The obvious incorrect tags are marked with the blue color. It is clear to see that the recommended tags are relevant to the images.

5. Conclusion

In this paper, we propose a nonlinear matrix factorization approach to estimate the tag relevance. The latent factors are mapped into a unified space and their relevance can be measured by distance between them. The image visual similarity and tag correlation are incorporated simultaneously to preserve the local visual geometry and local textual geometry. The latent factors in the same space makes their relationship more understandable, and allows to fast return top tags (images) to a query image (tag) via nearest neighbor search. News images can be better incorporated into the model by considering their visual and semantic information. Empirical results on a real dataset have demonstrated the effectiveness of our method.

In future, we will first extent our method to further incorporate users of social images and then apply our method to group recommendation. Besides, by transforming images and tag in the same space, we can incorporate the content-based image retrieval, keyword-based image retrieval, image tagging and tag recommendation in a unified framework.

Acknowledgments

This work was supported by 973 Program (Project no. 2010CB327905) and the National Natural Science Foundation of China (Grant nos. 60833006, 60903146, 61272329).

References

[1] J. Zhuang, S.C.H. Hoi, A two-view learning approach for image tag ranking, in: WSDM, 2011.

[2] D. Liu, X.-S. Hua, L. Yang, M. Wang, H.-J. Zhang, Tag ranking, in: WWW, 2009.

[3] C. Wang, F. Jing, L. Zhang, Image annotation refinement using random walk with restarts, in: MM, 2006.

[4] C. Wang, F. Jing, L. Zhang, H.-J. Zhang, Content-based image annotation refinement, in: CVPR, 2007.

[5] X. Li, C.G.M. Snoek, M. Worring, Learning tag relevance by neighbor voting for social image retrieval, in: MIR, 2008, pp. 180–187.

[6] L. Wu, L. Yang, N. Yu, X.-S. Hua, Learning to tag, in: WWW, 2009, pp. 361–370.

[7] Z. Li, J. Liu, X. Zhu, T. Liu, H. Lu, Image annotation using multi-correlation probabilistic matrix factorization, in: MM, 2010.

[8] G. Zhu, S. Yan, Y. Ma, Image tag refinement towards low-rank, content-tag prior and error sparsity, in: MM, 2010, pp. 461–470.

[9] P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, in: ECCV, 2002, pp. 97–112.

[10] J. Jeon, V. Lavrenko, R. Manmatha, Automatic image annotation and retrieval using cross-media relevance models, in: SIGIR, 2003.

[11] S. Feng, R. Manmatha, V. Lavrenko, Multiple bernoulli relevance models for image and video annotation, in: CVPR, 2004, pp. 1002–1009.

[12] V. Lavrenko, R. Manmatha, J. Jeon, A model for learning the semantics of pictures, in: NIPS, 2004, pp. 553–560.

[13] J. Liu, B. Wang, M. Li, W. Ma, H. Lu, S. Ma, Dual cross-media relevance model for image annotation, in: MM, 2007, pp. 605–614.

[14] X. Liu, R. Ji, H. Yao, P. Xu, X. Sun, T. Liu, Cross-media manifold learning for image retrieval & annotation, in: MIR, 2008, pp. 141–148.

[15] A. Makadia, V. Pavlovic, S. Kumar, A new baseline for image annotation, in: ECCV, 2008, pp. 316–329.

[16] J. Liu, M. Li, Q. Liu, H. Lu, S. Ma, Image annotation via graph learning, Pattern Recogn. 42 (2) (2009) 218–228.

[17] C. Wang, S. Yan, L. Zhang, H.-J. Zhang, Multi-label sparse coding for automatic image annotation, in: CVPR, 2009, pp. 1463–1650.

[18] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation, in: ICCV, 2009, pp. 309–316.

[19] X. Li, C.G.M. Snoek, M. Worring, Unsupervised multi-feature tag relevance learning for social image retrieval, in: CIVR, 2010.

[20] Z. Li, M. Wang, J. Liu, C. Xu, H. Lu, News contextualization with geographic and visual information, in: MM, 2011.

[21] Y. Yang, Y. Zhuang, F. Wu, Y. Pan, Harmonizing hierarchical manifolds for multimedia document semantics understanding cross-media retrieval, IEEE Trans. MM 10 (3) (2008) 437–446.

[22] Y. Yang, F. Nie, S. Xiang, Y. Zhuang, W. Wang, Local and global regressive mapping for manifold learning with out-of-sample extrapolation, in: AAAI, 2010.

[23] X. Chen, Y. Mu, S. Yan, T.-S. Chua, Efficient large-scale image annotation by probabilistic collaborative multi-label propagation, in: MM, 2010, pp. 35–44.

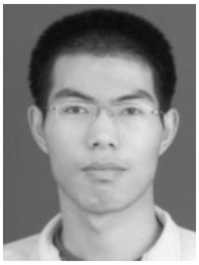
[24] Z. Li, J. Liu, H. Lu, Sparse constraint nearest neighbour selection in cross-media retrieval, in: ICIP, 2010.

[25] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Statist. 32 (2) (2004) 407–499.

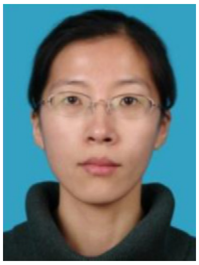
[26] L. Wu, X.-S. Hua, W.-Y. Ma, S. Li, Image retagging, in: MM, 2010.

[27] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y.-T. Zheng, Nus-wide: a real-world web image database from National University of Singapore, in: CIVR, 2009.

[28] C.-W. Hsu, C.-C. Chang, C.-J. Lin, A Practical Guide to Support Vector Classification 〈http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf〉.



Zechao Li received the BE degree from University of Science and Technology of China (USTC), Anhui, China, in 2008. He is currently pursuing the PhD degree at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.



Jing Liu received the BE and ME degrees from Shandong University, Shandong, in 2001 and 2004, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2008. She is an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her current research interests include multimedia analysis, understanding, and retrieval.

Hangqing Lu received his BS and MS from Department of Computer Science and Department of Electric Engineering in Harbin Institute of Technology in 1982 and 1985. He got his PhD from Department of Electronic and Information Science in Huazhong University of Sciences and Technology. He is a professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Current research interests include Image similarity measure, Video Analysis, Multimedia Technology and System.