Machine Vision and Applications (2014) 25:1647–1659
DOI 10.1007/s00138-013-0548-3

SPECIAL ISSUE PAPER

Inductive hierarchical nonnegative graph embedding for "verb–object" image classification

Chao Sun · Bing-Kun Bao · Changsheng Xu

Received: 20 February 2013 / Revised: 7 August 2013 / Accepted: 19 September 2013 / Published online: 11 October 2013

Abstract Most existing image classification algorithms mainly focus on dealing with images with only "object" concepts. However, in real-world cases, a great variety of images contain "verb–object" concepts, rather than only "object" ones. The hierarchical structure embedded in these "verb–object" concepts can help to enhance classification. However, traditional feature representation methods cannot utilize it. To tackle this problem, we present in this paper a novel approach, called inductive hierarchical nonnegative graph embedding. By assuming that those "verb–object" concept images which share the same "object" part but different "verb" part have a specific hierarchical structure, we integrate this hierarchical structure into the nonnegative graph embedding technique, together with the definition of inductive matrix, to (1) conduct effective feature extraction from hierarchical structure, (2) easily transfer each new testing sample into its low-dimensional nonnegative representation, and (3) perform image classification of "verb–object" concept images. Extensive experiments compared with the state-of-the-art algorithms on nonnegative data factorization demonstrate the classification power of proposed approach on "verb–object" concept images classification.

C. Sun · B.-K. Bao · C. Xu (✉)
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
e-mail: csxu@nlpr.ia.ac.cn

C. Sun
e-mail: csun@nlpr.ia.ac.cn

B.-K. Bao
e-mail: bkbao@nlpr.ia.ac.cn

C. Sun · B.-K. Bao · C. Xu
China-Singapore Institute of Digital Media, Singapore, Singapore

Keywords Inductive · Hierarchical · Graph embedding · Verb–object

1 Introduction

Image understanding and classification applications have been wildly researched for decades. Most existing image classification methods focus on handling images with only "object" concepts [2,4,13,14], such as "horse", "bike" and "tree" etc. However, in real-world, a huge number of images contain "verb–object" concepts, such as "ride horse", "repair boat", and "cut tree", rather than only "object" concepts. These kinds of "verb–object" concept images could represent more abundant semantic meaning while having much smaller size comparing with videos [6,7,22]. Hence, research on "verb–object" concept images, like image classification, is significant and essential. However, by using traditional image representation techniques, each "verb–object" concept could be treated as a whole "object". The embedded information in it will be ignored. In such a way, the classification performance would be discounted.

It is observed that, some concepts, like "ride horse" and "feed horse", "repair boat" and "row boat", "cut tree" and "plant tree", are sharing the same "object" part. Figure 1 illustrates several images under a set of "verb–object" concepts sharing same "object" but different "verb" parts. Intuitively, this kind of set of concepts very likely share a common latent information or pattern in images, which can be helpful for image classification. Motivated by these observations, we try to more effectively interpret "verb–object" concepts images. We assume that those "verb–object" concept images which share the same "object" but different "verb" part have a specific hierarchical structure, which can be utilized for image classification. By applying this hierarchical structure,

Fig. 1 An illustration of a set of "verb–object" concept images

one "verb–object" concept can be used not only to separate a set of concepts from other concepts which have different "object" parts, but also to discriminated itself from concepts in this set by the different "verb" parts. For example, images of "carry bike" can help to discriminate images of "repair bike" from images of "feed horse", while itself should be classified from images of "repair bike". We regard this structure as hierarchical structure and utilize it in classification.

Our previous work, hierarchical nonnegative graph embedding (HNGE) [19], has utilized hierarchical structure together with nonnegative graph embedding algorithm to preform "verb–object" concept image classification, and has achieved a remarkable classification accuracy. In that work, we reconstructed the testing sample with the training samples and then used the derived reconstruction coefficients to combine the encoding coefficient vectors of training samples. However, in its subsequent work, we found that it suffered from an out-of-sample extension problem, that is, how to easily and accurately transform each new testing sample into its low-dimensional nonnegative representation. This problem could also be described as how to obtain the encoding coefficient vector for the feature vectors of testing samples, without increasing computational cost or violating the basic nonnegative assumption.

To tackle this out-of-sample extension problem, in this paper, we reformulate the problem, extend the HNGE, and propose a new algorithm, named inductive hierarchical nonnegative graph embedding (IHNGE). This approach could not only combine hierarchical structure with nonnegative graph embedding to preform effective "verb–object" concept image classification, but also propose a conventional and effective way for out-of-sample extension in classification. We conduct experiments on a web image corpus composed of 9000 images on 45 "verb–object" concepts. The experimental results demonstrate the effectiveness of our approach.

The contributions of our work can be summarized into threefold:

- We utilize the hierarchical information to extend the nonnegative graph embedding, which is proved to be suitable for image classification within hierarchical structure.
- Based on the HNGE, we propose the IHNGE, which brings in the inductive matrix to deal with the problems of high computational cost and nonnegative assumption satisfaction within the testing procedure in image classification.
- We use the IHNGE to tackle the classification of "verb–object" concept images, and have obtained a remarkable classification accuracy.

The rest of the paper is organized as follows: in Sect. 2, we introduce the related work. We elaborate our IHNGE and its formulation in Sect. 3. Experimental results are reported in Sect. 4. Finally, we give conclusions in Sect. 5.

2 Related work

Nonnegative and sparse representation techniques have been well researched in recent decades to find nonnegative basis of data features with few nonzero elements [9]. A pioneer work for such a purpose is nonnegative matrix factorization (NMF) [12]. It decomposes the data features matrix as the arithmetical product of two matrices which possess only nonnegative elements. Generally, NMF belongs to the techniques of feature extraction and dimensionality reduction, as it results in a dimension-reduced representation of the primal data features [16]. In recent years, NMF [8] and its variants, localized NMF (LNMF) [15], convex NMF (CNMF) [3], and Fisher NMF [23], have been proved effective in many applications. Some work extended and applied NMF in many different fields, such as face and object recognition [18,27], biomedical applications [5,10], color science [17], and so on.

Recently, beyond the original nonnegative data factorization, Yan et al. [24] proposed a graph embedding framework which provided a unified formulation for dimensionality reduction, and possessed the algorithmic properties in convergency, sparsity, and classification power. Yang et al. [25] extended this work and proposed an approach, named nonnegative graph embedding (NGE), to obtain customized nonnegative data factorization by simultaneously realizing the specific purpose characterized by the intrinsic and penalty graphs. This work was further refined by Wang et al. [20] with the efficient multiplicative updating rule, namely multiplicative nonnegative graph embedding (MNGE). MNGE achieved a satisfactory performance, however, it lacked the direct way to obtain the encoding coefficient vector for the

new testing sample. Our previous work [19] HNGE also suffered from this out-of-sample extension problem, which is solved by IHNGE in this paper.

A similar work was proposed by Zhang et al. [28], which tried to simultaneously learn a set of classifiers for "verb–object" concepts in the same group. However, in this paper, we do not focus on designing classifiers, but focus on feature extraction and representation, which refer to nonnegative and sparse representation techniques.

Besides, our work seems to be similar to human action recognition, like [26], as "verb–object" concepts look like actions. However, the main difference between our work and human action recognition is that human action recognition treats every action as unrelated to each others while our IHNGE considers "verb–object" concepts having hierarchical structure to explore a new layer of linkage among actions. Moreover, in human action recognition, an action does not always have an object, while IHNGE focuses on handling images containing both "verb" and "object".

3 Inductive hierarchical nonnegative graph embedding

In this section, we elaborate IHNGE and formulate the problem within the framework of nonnegative data decomposition.

Before introducing the inductive nonnegative graph embedding, we first list the notations used in this paper here. Let $\mathbf{X} = [x_1, x_2, \ldots, x_N]$ denote the data sample set, in which $x_i \in \mathbb{R}^k$ denotes the feature descriptor of the ith sample and N is number of total samples. Here we assume that the matrix X is nonnegative. Let m be the dimension of the desired dimension-reduced feature space, the task of our data factorization is to derive a nonnegative basis matrix $W \in \mathbb{R}^{k\times m}$ and a nonnegative encoding coefficient matrix $H \in \mathbb{R}^{m\times N}$, while the data matrix X can be approximated as the product of matrices W and H. In this paper, we utilize the following rule to facilitate presentation: for any matrix A, A_i denotes the ith row vector of A, its corresponding lowercase version a_i denotes the ith column vector of A. A_{ij} denotes the element of A at the ith row and jth column, and $A_{p\times q}$ means that A has p rows and q columns.

3.1 Motivation

Practically, we believe that the "verb–object" concept images contain hierarchical structure. Figure 3 illustrates the hierarchical structure in "verb–object" concept images. As shown, we divide whole "verb–object" concepts into two levels. On the first level, those "verb–object" concepts containing the same "object" are treated as a class, here "class" is as the same meaning as "group" or "set" mentioned above. On the second level, those "verb–object" concepts in the same class on the first level are divided into sub-classes, according to the different "verb" parts they have. Although the final aim of our classification is to discriminate all the sub-classes on the second level, we do not directly perform classification on only second level. Instead, as shown in Fig. 2, on one hand, we enlarge interclass distance for classes on the first level. On the other hand, we reduce intraclass distance and enlarge interclass distance for sub-classes on the second level. These two steps are performed simultaneously while the one on the first level will compensate the one on the seconde level and improve the performance on the final "verb–object" concept classification.

Figure 2 shows an illustration of this procedure. Suppose "repair car", "drive car", "wash face" and "make up face" are four illustrative sub-classes on the second level in our dataset, while "repair car" and "drive car" belong to the same class "car" on the first level and "wash face" and "make up face" belong to another same class "face" on the first level. Our goal is to classify four sub-classes "repair car", "drive car", "wash face" and "make up face". However, classification in our IHNGE is not directly conducted by only treating these four as different classes. As Fig. 2 shown, for an illustrative sample in sub-class "repair car", on one hand, we treat it as a sample in class "car" on the first level and tend to separate it from all samples in class "face" by enlarging the inter-

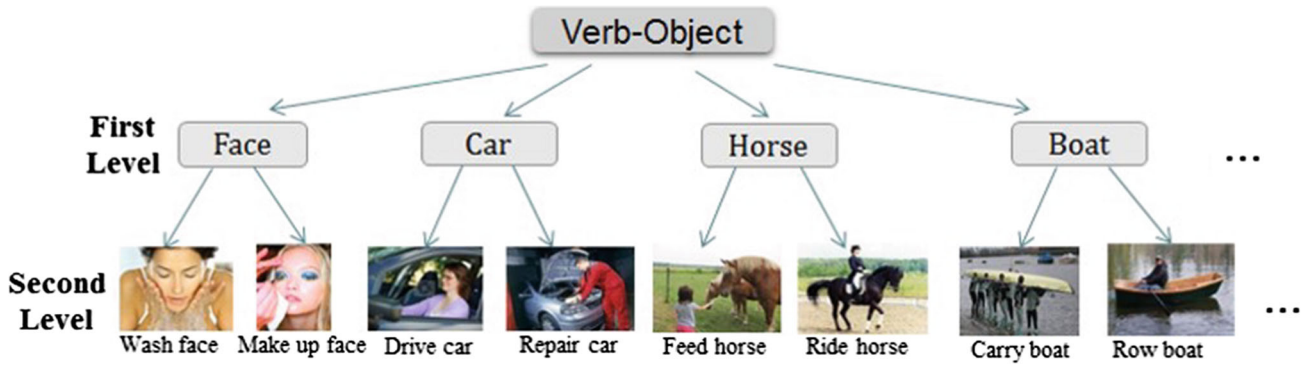

Fig. 2 Intraclass distance and interclass distance on two levels

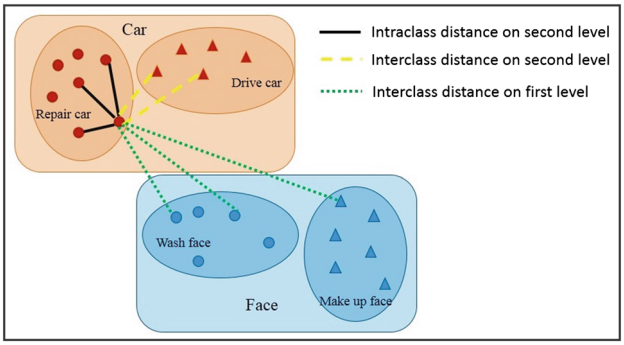

Fig. 3 An illustration of hierarchical structure

class distance on first level. In this step, all samples in anther sub-class "drive car" but same class "car" will contribute to classify this illustrative sample. On the other hand, we tend to separate it from those samples in same class "car" but different sub-class "drive car", by reducing the intraclass distance on second level and enlarging interclass distance on second level. Specifically, these two steps are integrated into the our IHNGE model and performed simultaneously in hierarchical graph embedding process.

To achieve our goal, we make use of inductive nonnegative data decomposition together with HNGE. The target of inductive nonnegative data decomposition and the purpose of HNGE coexist harmoniously and do not mutually compromise, while satisfying those distance requirements in two levels. We elaborate each part as well as the unified one in the following.

3.2 Objective for inductive nonnegative data decomposition

Nonnegative matrix factorization (NMF) factorizes the data matrix X into one lower-rank nonnegative basis matrix W and one nonnegative coefficient matrix H. Its usual objective function is shown as follows, which was also be utilized in [19]:

$$\min_{W,H} \|X - WH\|_F^2, \quad \text{s.t.} \quad W, H \geq 0, \tag{1}$$

The most important improvement of IHNGE different from HNGE [19] is that, IHNGE imposes the extra constraint that the coefficient of each data point lies within the subspace spanned by the column vectors of an inductive matrix. Theoretically, based on the coefficient matrix H, we assume that the coefficient vector h_i can be derived by linear transformation from the feature descriptor of sample x_i, shown as follows:

$$h_i = Cx_i \tag{2}$$

or:

$$H = CX \tag{3}$$

where $C \in \mathbb{R}^{m\times k}$ indicates the inductive matrix which transforms the k-dimensional feature vector into the m-dimensional feature vector. Then the objective function of NMF in (1) can be refined as:

$$\min_{W,C} \|X - WCX\|_F^2, \quad \text{s.t.} \quad W, C \geq 0, \tag{4}$$

Based on this assumption, we can easily obtain the encoding coefficient vector h_y for a new testing sample y. After getting the inductive matrix C, we have:

$$h_y = Cy \tag{5}$$

The inductive matrix C can ensure the nonnegative ability of vector h_y required by NMF, while avoiding increasing computational cost when testing new samples.

3.3 Objective for hierarchical nonnegative graph embedding

As proposed by Yan et al. [24], most dimensionality reduction algorithms can be explained within a unified framework, called graph embedding [24]. The purpose of graph embedding-based algorithm is characterized by the so-called intrinsic and penalty graphs. In this paper, we formulate our IHNGE within the general graph embedding framework [24].

Specifically, to serves for graph embedding, we first divide the coefficient matrix H into two parts, namely,[1]

$$H = \begin{bmatrix} H^1 \\ H^2 \end{bmatrix} \tag{6}$$

where $H^1 = [h_1^1, h_2^1, \ldots, h_N^1] \in \mathbb{R}^{d\times N}$, $d < N$, denotes the desired low-dimensional representations for the training data on parent classes, and $H^2 = [h_1^2, h_2^2, \ldots, h_N^2] \in \mathbb{R}^{(m-d)\times N}$.

As existing hierarchical structure, we then divide the matrix H^1 into two parts, namely,

$$H^1 = \begin{bmatrix} H^{11} \\ H^{12} \end{bmatrix} \tag{7}$$

where $H^{11} = [h_1^{11}, h_2^{11}, \ldots, h_N^{11}] \in \mathbb{R}^{r\times N}$, $r < d$, denotes the desired low-dimensional representations for the training data on child classes, and $H^{12} = [h_1^{12}, h_2^{12}, \ldots, h_N^{12}] \in \mathbb{R}^{(d-r)\times N}$.

Therefore, to clearly reveal the structure, we rewrite matrix H as:

$$H = \begin{bmatrix} \begin{bmatrix} H^{11} \\ H^{12} \end{bmatrix} \\ H^2 \end{bmatrix} \tag{8}$$

[1] Superscript numbers of matrices, 1, 2, 11, 12, etc., are symbols, not the power in math.

Meanwhile, to facilitate presentation, we define matrix $\bar{H}^2$ as the combination of H^{12} and H^2:

$$\bar{H}^2 = \begin{bmatrix} H^{12} \\ H^2 \end{bmatrix} \tag{9}$$

where $\bar{H}^2 = [\bar{H}_1^2, \bar{H}_2^2, \ldots, \bar{H}_N^2] \in \mathbb{R}^{(m-r)\times N}$.

Correspondingly, the basis matrix W is also divided as:

$$W = \begin{bmatrix} W^1 & W^2 \end{bmatrix} \tag{10}$$

and:

$$W^1 = \begin{bmatrix} W^{11} & W^{12} \end{bmatrix} \tag{11}$$

where $W^1 \in \mathbb{R}^{k\times d}$, $W^2 \in \mathbb{R}^{k\times(m-d)}$, $W^{11} \in \mathbb{R}^{k\times r}$ and $W^{12} \in \mathbb{R}^{k\times(d-r)}$.

The combination of W^{12} and W^2 is:

$$\bar{W}^2 = \begin{bmatrix} W^{12} & W^2 \end{bmatrix} \tag{12}$$

where $\bar{W}^2 \in \mathbb{R}^{k\times(m-r)}$.

As $H = CX$, inductive matrix C is also divided as:

$$C = \begin{bmatrix} C^1 \\ C^2 \end{bmatrix} \tag{13}$$

and

$$C^1 = \begin{bmatrix} C^{11} \\ C^{12} \end{bmatrix} \tag{14}$$

where $C^1 \in \mathbb{R}^{d\times k}$, $C^2 \in \mathbb{R}^{(m-d)\times k}$, $C^{11} \in \mathbb{R}^{r\times k}$ and $C^{12} \in \mathbb{R}^{(d-r)\times k}$.

The combination of C^{12} and C^2 is denoted as:

$$\bar{C}^2 = \begin{bmatrix} C^{12} \\ C^2 \end{bmatrix} \tag{15}$$

where $\bar{C}^2 \in \mathbb{R}^{(m-r)\times k}$.

Let $G = \{X, S\}$ be an undirected weighted graph with vertex set X and similarity matrix $S \in \mathbb{R}^{N\times N}$. Each element of the real symmetric matrix S measures similarity between a pair of vertices. The Laplacian matrix L and diagonal matrix D of a graph G are defined as:

$$L = D - S, \quad D_{ii} = \sum_{j\neq i} S_{ij}, \quad \forall i \tag{16}$$

Graph embedding generally involves an intrinsic graph G, which characterizes the favorite relationship among the data samples, and a penalty graph $G^u = \{X, S^u\}$, which characterizes the unfavorable relationship among the data samples. Correspondingly, penalty graph G^u also has its Laplacian matrix L^u and diagonal matrix D^u. In our work, we assume that the sample data set has two levels. Therefore, there are an intrinsic graph and a penalty graph on the first level, and an intrinsic graph and a penalty graph on the second level, respectively.

Let $G = \{X, S\}$ be the intrinsic graph on the first level, $G^u = \{X, S^u\}$ be the penalty graph on the first level, $\tilde{G} = \{X, \tilde{S}\}$ be the intrinsic graph on the second level, and $\tilde{G}^u = \{X, \tilde{S}^u\}$ be the penalty graph on the second level. Their Laplacian matrices and diagonal matrices are L, D, L^u, D^u, $\tilde{L}$, $\tilde{D}$, $\tilde{L}^u$, and $\tilde{D}^u$, respectively.

According to graph embedding, the target of graph preserving on the first level is:

$$\max_{H^1} \sum_{i\neq j} \|h_i^1 - h_j^1\|^2 S_{ij}^u \tag{17}$$

As (W^2, H^2) are considered as the complementary space of (W^1, H^1). From the complementary property between H^1 and H^2, the objective is transformed into:

$$\min_{H^2} \sum_{i\neq j} \|h_i^2 - h_j^2\|^2 S_{ij}^u \tag{18}$$

On the second level, our two targets of graph preserving are given as:

$$\begin{cases} \min_{H^{11}} \sum_{i\neq j} \|h_i^{11} - h_j^{11}\|^2 \tilde{S}_{ij} \\ \max_{H^{11}} \sum_{i\neq j} \|h_i^{11} - h_j^{11}\|^2 \tilde{S}_{ij}^u \end{cases} \tag{19}$$

As $(\bar{W}^2, \bar{H}^2)$ are considered as the complementary space of (W^{11}, H^{11}). From the complementary property between H^{11} and $\bar{H}^2$, the second objective above is transformed into:

$$\min_{\bar{H}^2} \sum_{i\neq j} \|\bar{h}_i^2 - \bar{h}_j^2\|^2 \tilde{S}_{ij}^u \tag{20}$$

3.4 Unified formulation

To achieve the objectives in Eqs. (4), (18), (19), and (20) which are required for HNGE, we have the unified objective function as:

$$\begin{aligned} \min_{W,C} \; & \sum_{i\neq j} \|h_i^2 - h_j^2\|^2 S_{ij}^u + \sum_{i\neq j} \|h_i^{11} - h_j^{11}\|^2 \tilde{S}_{ij} \\ & + \sum_{i\neq j} \|\bar{h}_i^2 - \bar{h}_j^2\|^2 \tilde{S}_{ij}^u + \lambda \|X - WCX\|_F^2, \\ \text{s.t.} \quad & W, C \geq 0 \end{aligned} \tag{21}$$

where λ is a positive parameter to balance the two parts for graph embedding and data reconstruction, and is always set as 1 empirically.

From the definitions of Laplacian matrix and diagonal matrix, together with definition in Eq. (3), we have:

$$\begin{aligned} \sum_{i\neq j} \|h_i^2 - h_j^2\|^2 S_{ij}^u &= Tr(H^2 L^u H^{2^T}) \\ &= Tr(C^2 X L^u X^T C^{2^T}) \end{aligned} \tag{22}$$

$$\begin{aligned} \sum_{i\neq j} \|h_i^{11} - h_j^{11}\|^2 \tilde{S}_{ij} &= Tr(H^{11} \tilde{L} H^{11^T}) \\ &= Tr(C^{11} X \tilde{L} X^T C^{11^T}) \end{aligned} \tag{23}$$

$$\sum_{i\neq j}\|\bar{h}_i^2-\bar{h}_j^2\|^2\tilde{S}_{ij}^u = Tr(\bar{H}^2\tilde{L}^u\bar{H}^{2^T}) = Tr(\bar{C}^2X\tilde{L}^uX^T\bar{C}^{2^T}) \quad (24)$$

Furthermore, as W is the basis matrix, it is natural to require that each column vector of W is normalized, that is $\|W_i\|=1,\forall i$. But this constraint makes the optimization problem much more complicated. Hence, we compensate the norms of the bases into the coefficient matrix and rewrite (22), (23), and (24) as:

$$Tr(C^2XL^uX^TC^{2^T}) \Rightarrow Tr(E_2C^2XL^uX^TC^{2^T}E_2^T) \quad (25)$$

$$Tr(C^{11}X\tilde{L}X^TC^{11^T}) \Rightarrow Tr(E_{11}C^{11}X\tilde{L}X^TC^{11^T}E_{11}^T) \quad (26)$$

$$Tr(\bar{C}^2X\tilde{L}^uX^T\bar{C}^{2^T}) \Rightarrow Tr(\bar{E}_2\bar{C}^2X\tilde{L}^uX^T\bar{C}^{2^T}\bar{E}_2^T) \quad (27)$$

where the matrix $E_2=\text{diag}\{\|w_1^2\|,\|w_2^2\|,\ldots,\|w_{m-d}^2\|\}$, $E_{11}=\text{diag}\{\|w_1^{11}\|,\|w_2^{11}\|,\ldots,\|w_r^{11}\|\}$, and $\bar{E}_2=\text{diag}\{\|\bar{w}_1^2\|,\|\bar{w}_2^2\|,\ldots,\|\bar{w}_{m-d}^2\|\}$, where w_i^k denotes the ith column vector of matrix W^k, $k=1,2,11,22$.

By combining equations above, the final objective function is then reformulated as:

$$\begin{aligned}\min_{W,C}\ & Tr(E_2C^2XL^uX^TC^{2^T}E_2^T)\\ &+Tr(E_{11}C^{11}X\tilde{L}X^TC^{11^T}E_{11}^T)\\ &+Tr(\bar{E}_2\bar{C}^2X\tilde{L}^uX^T\bar{C}^{2^T}\bar{E}_2^T)+\lambda\|X-WCX\|_F^2,\\ \text{s.t.}\ & W,C\geq 0\end{aligned} \quad (28)$$

3.5 Convergent iterative procedure

As the final objective function is biquadratic and generally there does not exist closed-form solution, we use an iterative procedure to get the nonnegative solution.

Most iterative procedures for solving high-order optimization problems transform the original intractable problem into a set of tractable sub-problems, and finally obtain the convergence to a local optimum. Our proposed iterative procedure also follows this philosophy and optimizes W and C alternately.

3.5.1 Optimize W for given C

For a fixed matrix C, the objective function in (28) with respect to basis matrix W can be rewritten as:

$$F(W) = Tr(WQ^1W^T)+Tr(WQ^2W^T)+Tr(WQ^3W^T)+\lambda\|X-WCX\|_F^2 \quad (29)$$

where:

$$Q^1=\begin{bmatrix}0 & 0\\ 0 & C^2XL^uX^TC^{2^T}\end{bmatrix}\circ I,$$

$$Q^2=\begin{bmatrix}C^{11}X\tilde{L}X^TC^{11^T} & 0\\ 0 & 0\end{bmatrix}\circ I,$$

$$Q^3=\begin{bmatrix}0 & 0\\ 0 & \bar{C}^2X\tilde{L}^uX^T\bar{C}^{2^T}\end{bmatrix}\circ I$$

and $Q^1,Q^2,Q^3\in\mathbb{R}^{m\times m}$, operator $\circ$ indicates the element-wise matrix multiplication, I indicates the identity matrix.

We integrate the nonnegative constraints into the objective function with respect to W, and set $\psi_{i,j}$ as the Lagrange multiplier for constraint $W_{i,j}\geq 0$. Set Matrix Ψ, where $\Psi_{i,j}=\psi_{i,j}$. Then the Lagrange $L(W)$ is defined as:

$$L(W)=Tr(WQ^1W^T)+Tr(WQ^2W^T)+Tr(WQ^3W^T)+\lambda\|X-WCX\|_F^2+Tr(\Psi W^T) \quad (30)$$

By setting the partial derivation of $L(W)$ with respect to W as zero,

$$\frac{\partial L(W)}{\partial W} = W(2Q^1+2Q^2+2Q^3)+2\lambda WCXX^TC^T-2\lambda XX^TC^T+\Psi \quad (31)$$

Along with the Karush–Kuhn–Tucker (KKT) condition [11] of $\psi_{i,j}W_{i,j}=0$, we obtain the following equation:

$$(WQ^1+WQ^2+WQ^3)_{i,j}W_{i,j}+(\lambda WCXX^TC^T)_{i,j}W_{i,j}-(\lambda XX^TC^T)_{i,j}W_{i,j}=0 \quad (32)$$

Then we set:

$$Q^1=Q_+^1-Q_-^1,\quad Q^2=Q_+^2-Q_-^2,\quad Q^3=Q_+^3-Q_-^3$$

where:

$$Q_+^1=\begin{bmatrix}0 & 0\\ 0 & C^2XD^uX^TC^{2^T}\end{bmatrix}\circ I,$$

$$Q_+^2=\begin{bmatrix}C^{11}X\tilde{D}X^TC^{11^T} & 0\\ 0 & 0\end{bmatrix}\circ I,$$

$$Q_+^3=\begin{bmatrix}0 & 0\\ 0 & \bar{C}^2X\tilde{D}^uX^T\bar{C}^{2^T}\end{bmatrix}\circ I$$

and:

$$Q_-^1=\begin{bmatrix}0 & 0\\ 0 & C^2XS^uX^TC^{2^T}\end{bmatrix}\circ I,$$

$$Q_-^2=\begin{bmatrix}C^{11}X\tilde{S}X^TC^{11^T} & 0\\ 0 & 0\end{bmatrix}\circ I,$$

$$Q_-^3=\begin{bmatrix}0 & 0\\ 0 & \bar{C}^2X\tilde{S}^uX^T\bar{C}^{2^T}\end{bmatrix}\circ I$$

Equation (32) can be rewritten as:

$$\begin{aligned}&(WQ_+^1 + WQ_+^2 + WQ_+^3)_{i,j}W_{i,j}\\&+(\lambda WCXX^TC^T)_{i,j}W_{i,j}\\&-(WQ_-^1 + WQ_-^2 + WQ_-^3)_{i,j}W_{i,j}\\&-(\lambda XX^TC^T)_{i,j}W_{i,j} = 0\end{aligned} \tag{33}$$

which leads to the following update rule for W:

$$W_{ij} \leftarrow W_{ij}\frac{[\lambda XX^TC^T + WQ_-^1 + WQ_-^2 + WQ_-^3]_{ij}}{[\lambda WCXX^TC^T + WQ_+^1 + WQ_+^2 + WQ_+^3]_{ij}} \tag{34}$$

3.5.2 Optimize C for given W

After updating the matrix W, we normalize the column vectors of it and consequently convey the norm to the coefficient matrix C, namely,

$$C_{ij} \leftarrow C_{ij} \times \|w_i\| \quad \forall i, j \tag{35}$$

$$w_i \leftarrow w_i/\|w_i\| \quad \forall i \tag{36}$$

Note that the updating of C and W here will not change the value of the objective function in (28).

Then based on the normalized W, the objective function in (28) with respect to C is then rewritten as:

$$\begin{aligned}F(C) = &Tr(C^2XL^uX^TC^{2^T}) + Tr(C^{11}X\tilde{L}X^TC^{11^T})\\&+Tr(\bar{C}^2X\bar{L}^uX^T\bar{C}^{2^T}) + \lambda\|X - WCX\|^2\end{aligned} \tag{37}$$

Let $R^1 = [0_{(m-d)\times d}, I_{(m-d)\times(m-d)}]$, $R^2 = [I_{r\times r}, 0_r \times (m-r)]$, $R^3 = [0_{(m-r)\times r}, I_{(m-r)\times(m-r)}]$

$F(C)$ can be rewritten as:

$$\begin{aligned}F(C) = &Tr(R^1CXL^uX^TC^TR^{1^T})\\&+Tr(R^2CX\tilde{L}X^TC^TR^{2^T})\\&+Tr(R^3CX\tilde{L}^uX^TC^TR^{3^T}) + \lambda\|X - WCX\|^2\end{aligned} \tag{38}$$

We integrate the nonnegative constraints into the objective function with respect to C, and set $\phi_{i,j}$ as the Lagrange multiplier for constraint $C_{i,j} \geq 0$. Set Matrix Φ, where $\Phi_{i,j} = \phi_{i,j}$. Then the Lagrange $L(C)$ is defined as:

$$\begin{aligned}L(C) = &Tr(R^1CXL^uX^TC^TR^{1^T})\\&+Tr(R^2CX\tilde{L}X^TC^TR^{2^T})\\&+Tr(R^3CX\tilde{L}^uX^TC^TR^{3^T}) + \lambda\|X - WCX\|^2\\&+Tr(\Phi C^T)\end{aligned} \tag{39}$$

Then the partial derivation of $L(C)$ with respect to C is:

$$\begin{aligned}\frac{\partial L(C)}{\partial C} = &2R^{1^T}R^1CXL^uX^T + 2R^{2^T}R^2CX\tilde{L}X^T\\&+2R^{3^T}R^3CX\tilde{L}^uX^T - 2\lambda W^TXX^T\\&+2\lambda W^TWCXX^T + \Phi\end{aligned} \tag{40}$$

Algorithm 1

1: Input: Image representation matrix X, graphs G, G^u, $\tilde{G}$, $\tilde{G}^u$
2: Initialization: Randomly choose W^0,C^0 as nonnegative matrices.
3: For $t = 0, 1, 2, \ldots, T_{max}$, Do
1) For given $C = C^t$, update matrix W as:
$W_{ij} \leftarrow W_{ij}\frac{[\lambda XX^TC^T + WQ_-^1 + WQ_-^2 + WQ_-^3]_{ij}}{[\lambda WCXX^TC^T + WQ_+^1 + WQ_+^2 + WQ_+^3]_{ij}}$ $\forall i, j$
2) Normalize the column vectors of W^{t+1}
$C_{ij} \leftarrow C_{ij} \times \|w_i^{t+1}\|$ $\forall i, j$
$w_i^{t+1} \leftarrow w_i^{t+1}/\|w_i^{t+1}\|$ $\forall i$
3) For given $W = W^{t+1}$, update the matrix C as:
$C_{ij} \leftarrow C_{ij} \cdot [\lambda W^TXX^T + R^{1^T}R^1CXS^uX^T + R^{2^T}R^2CX\tilde{S}X^T + R^{3^T}R^3CX\tilde{S}^uX^T]_{ij} / [\lambda W^TWCXX^T + R^{1^T}R^1CXD^uX^T + R^{2^T}R^2CX\tilde{D}X^T + R^{3^T}R^3CX\tilde{D}^uX^T]_{ij}$ $\forall i, j$
4) If $\|W^{t+1} - W^t\| < \epsilon$ and $\|C^{t+1} - C^t\| < \epsilon$ (ϵ is a small positive number), then break.
4: Output $W = W^t$ and $C = C^t$

Along with the KKT condition [11] of $\phi_{i,j}C_{i,j} = 0$, we obtain the following equation:

$$\begin{aligned}&(R^{1^T}R^1CXL^uX^T)_{i,j}C_{i,j} + (R^{2^T}R^2CX\tilde{L}X^T)_{i,j}C_{i,j}\\&+(R^{3^T}R^3CX\tilde{L}^uX^T)_{i,j}C_{i,j} - (\lambda W^TXX^T)_{i,j}C_{i,j}\\&+(\lambda W^TWCXX^T)_{i,j}C_{i,j} = 0\end{aligned} \tag{41}$$

As $L^u = D^u - S^u$, $\tilde{L} = \tilde{D} - \tilde{S}$, and $\tilde{L}^u = \tilde{D}^u - \tilde{S}^u$, we the rewrote Eq. (41) as:

$$\begin{aligned}&(R^{1^T}R^1CXD^uX^T)_{i,j}C_{i,j} + (R^{2^T}R^2CX\tilde{D}X^T)_{i,j}C_{i,j}\\&+(R^{3^T}R^3CX\tilde{D}^uX^T)_{i,j}C_{i,j} + (\lambda W^TWCXX^T)_{i,j}C_{i,j}\\&-(R^{1^T}R^1CXS^uX^T)_{i,j}C_{i,j} - (R^{2^T}R^2CX\tilde{S}X^T)_{i,j}C_{i,j}\\&-(R^{3^T}R^3CX\tilde{S}^uX^T)_{i,j}C_{i,j} - (\lambda W^TXX^T)_{i,j}C_{i,j}\\&= 0\end{aligned} \tag{42}$$

which leads to the following update rule for C:

$$\begin{aligned}C_{ij} \leftarrow &C_{ij} \cdot [\lambda W^TXX^T + R^{1^T}R^1CXS^uX^T\\&+R^{2^T}R^2CX\tilde{S}X^T + R^{3^T}R^3CX\tilde{S}^uX^T]_{ij}\\&/ [\lambda W^TWCXX^T + R^{1^T}R^1CXD^uX^T\\&+R^{2^T}R^2CX\tilde{D}X^T + R^{3^T}R^3CX\tilde{D}^uX^T]_{ij}\ \forall i, j\end{aligned} \tag{43}$$

The aim of training part for the IHNGE is to obtain the matrices W and C from training samples. Its entire procedure is described in Algorithm 1.

Table 1 "Verb–object" concepts

Concept names				
answer phone	play phone	feed horse	ride horse	build boat
repair boat	row boat	buy car	drive car	repair car
buy vegetable	cook vegetable	cut vegetable	carry bike	ride bike
repair bike	carry water	drink water	pour water	clap hands
wave hand	comb hair	cut hair	wash hair	cut tree
plant tree	prune tree	iron clothes	wash clothes	use computer
fix computer	put on shoes	fix shoe	feed dog	walk dog
sit on chair	repair chair	water flowers	arrange flower	make up face
wash face	lie on bed	sit on bed	read book	write on book

3.6 Classification

After obtaining the inductive matrix C from training samples according to algorithm 1, we can easily conduct the testing procedure of classification.

Suppose y is the raw feature vector of a new testing sample extracted from an image, where $y \in \mathbb{R}^k$. Then its corresponding encoding coefficient vector h_y is calculated by $h_y = Cy$, where $h_y \in \mathbb{R}^m$.

After that, we get the first r rows of h_y, and form them into a new vector h_y^*, where $h_y^* \in \mathbb{R}^r$. The h_y^* is the desired dimension-reduced feature vector of the sample.

To prove the dimension reduction power of our IHNGE, we then use some common classifiers, like 1NN and SVMs, to perform classification on those dimension-reduced feature vectors of samples.

4 Experiments

In this section, we evaluate the effectiveness of our proposed IHNGE compared with other subspace algorithms including linear discriminant analysis (LDA) [1], marginal fisher analysis (MFA) [24], and other nonnegative basis pursuit algorithm nonnegative matrix factorization (NMF) [12], multiplicative nonnegative graph embedding (MNGE) [20], and hierarchical nonnegative graph embedding (HNGE) in our previous work [19].

4.1 Database

The most existing image databases do not satisfy the requirement of containing "verb–object" concepts with hierarchical structure. Hence we setup a "verb–object" concept image database and conduct experiments on it. We predefined 45 "verb–object" concepts including 19 different "object", while each "object" containing 2–3 "verb–object" concepts. Table 1 lists all the "verb–object" concepts in our database. Then a total number of 9000 images were collected from Google Image[2] and Flickr[3], with 200 images on each "verb–object" concept, while every image was labeled manually. Each image should have two labels, one indicates class category on the first level, the other indicates class category on the second level. Figure 4 illustrates some "verb–object" concepts and their corresponding image samples.

4.2 Parameters

To describe the image content, each image was resized to 256×256 pixels. Followed [21], we divided each image by three-level spatial pyramid. The subregions are 1×1, 2×2 and 4×4 and the bases of codebook are 1024. A set of 21504-dimensional ($(1^2 + 2^2 + 4^2) \times 1024 = 21504$) Locality-constrained Linear Coding (LLC) features [21] were extracted from each image. We randomly choose 60 % images from each "verb–object" concept and combine them as training data. The rest 40 % images are then used as testing data.

As LDA, MFA, NMF, and MNGE algorithms can not handle hierarchical structure, to serve these algorithms, we just treat each "verb–object" concept in the second level as a class by ignoring the first level when performing experiments using these algorithms. For MFA, NMF and MNGE algorithms, the intrinsic graph and penalty graph are set as the same, and each dimension of the desired dimension-reduced feature space, m, is set as 1000. For HNGE, parameter m is also set as 1000, parameter d is set as $0.8 \times m$, parameter r is set as $0.8 \times d$. In our IHNGE, those parameters are set as same as in HNGE.

For those NMF-based algorithms (MFA, MNGE, HNGE), after obtaining matrices W and H by training, the desired dimension-reduced feature of a testing sample y is calculated as $h_y = W^{\perp} y$, where $W^{\perp}$ denotes the pseudo-inverse of

[2] http://images.google.com.

[3] http://www.flickr.com/.

Fig. 4 Sampling images in our database

matrix W. For our IHNGE, we directly get $h_y = Cy$ after obtaining inductive matrix C by training.

4.3 Results

First of all, we validate the selection of parameters m, d and r.

For m, we first set $r = d = m$. The hierarchical graph embedding is then degenerated into the general NMF. m is tuned among $\{500, 600, 700, \ldots, 1000\}$. Considering the computing complexity, m will not set above 1000 empirically. The average accuracy of classification on all image classes with the change of m is shown in Fig. 5. We can see that the average accuracy is highest when m is set as 1000.

For r, we first set $m = 1000$ and $d = r$. The hierarchical graph embedding is then degenerated into the general graph embedding and loses the hierarchical structure. We tuned r among $\{0.1 \times m, 0.2 \times m, \ldots, 0.9 \times m\}$. The average accuracy of classification on all image classes with the change of r is shown in Fig. 6. We notice that the average accuracy could achieve highest when r is set as $0.8 \times m$. Moreover, considering the hierarchical structure of our IHNGE model, we then empirically set d as $0.8 \times m$ and r as $0.8 \times d$.

We implement our algorithm using MATLAB 2009a and conduct the experiments on a computer with Inter(R) Xeon(R) E7-4860 2.27Ghz CPU and 32GB RAM. As the high-dimensional feature (21504-dimensional), each iteration of IHNGE costs about 1000 s in average.

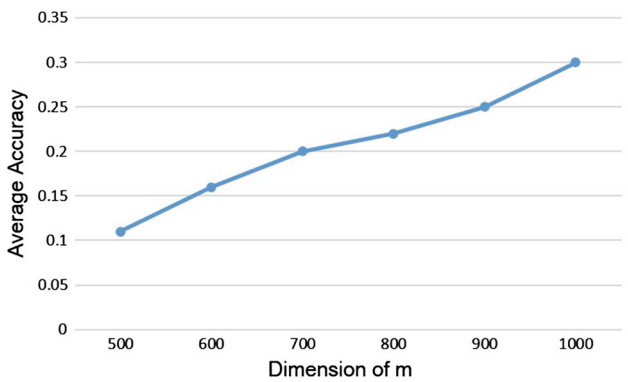

Fig. 5 Average accuracy over different parameter m for IHNGE

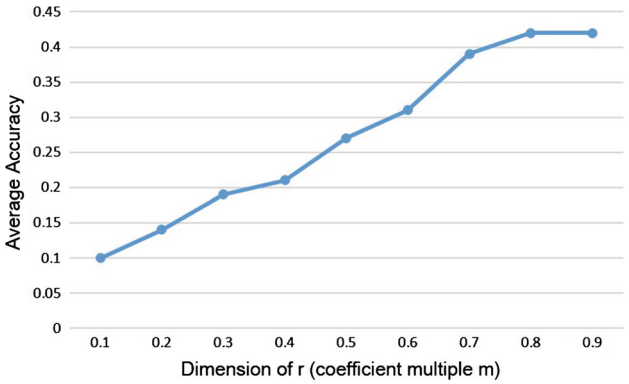

Fig. 6 Average accuracy over different parameter r for IHNGE

We conduct all experiments on our self-built database. Comparison experimental results of different algorithms are shown in Tables 2 and 3. Table 2 indicates the precision of classification and Table 3 indicates the recall of classification. For Table 2, the fist line illustrates the average of precisions over all 45 "verb–object" concepts using our proposed IHNGE as well other baselines, and the next ten lines are the precisions of classification on randomly selected 10 "verb–object" concepts from 4 different "objects". For Table 3, the fist line illustrates the average of recalls over all 45 "verb–object" concepts using our proposed IHNGE as well other baselines, and the next ten lines are the recalls of classification on the same 10 "verb–object" concepts as in Table 2.

From the results, we can infer that:

1. The average of precisions and recalls of LDA are lower than other NMF-based algorithms, which demonstrates the discriminative power of nonnegative data factorization.
2. The average of precisions and recalls of IHNGE and HNGE are higher than other NMF-based algorithms. This demonstrates that the hierarchical structure in "verb–object" could benefits the classification. By taking

Table 2 Classification precision (%) of different algorithms

	LDA	NMF	MFA	MNGE	HNGE	**IHNGE**
Precision	26.51	31.96	37.15	41.58	45.28	**47.32**
Play phone	11.58	10.35	8.37	12.45	17.20	**18.11**
Answer phone	14.75	21.81	34.95	**35.84**	35.12	35.24
Row boat	67.81	69.98	89.25	75.38	88.34	**89.45**
Repair boat	27.19	26.64	27.53	29.44	35.48	**38.10**
Build boat	23.74	26.89	32.57	30.83	37.75	**38.92**
Fix computer	32.57	37.78	39.19	44.72	48.56	**48.79**
Use computer	15.11	29.42	21.39	25.92	35.67	**36.52**
Prune tree	24.28	31.78	45.67	41.47	47.82	**48.00**
Plant tree	21.37	39.11	**57.49**	50.02	55.28	56.91
Cut tree	34.28	41.38	67.41	60.83	72.45	**74.71**

The fist line is the average precision over all concepts. The next ten lines are the precisions on randomly selected 10 concepts
Bold values in each line indicate highest one among all methods

Table 3 Classification recall (%) of different algorithms

	LDA	NMF	MFA	MNGE	HNGE	**IHNGE**
Recall	25.82	30.12	37.01	40.49	44.10	**45.09**
Play phone	10.00	8.75	6.25	10.00	13.75	**16.25**
Answer phone	13.75	17.50	32.50	33.75	30.00	**37.50**
Row boat	61.25	57.50	77.50	67.50	75.00	**85.00**
Repair boat	28.75	21.25	27.50	25.00	28.75	**33.75**
Build boat	18.75	25.00	28.75	28.75	32.50	**33.75**
Fix computer	28.75	35.00	37.50	36.25	**41.25**	40.00
Use computer	13.75	23.75	16.25	21.25	30.00	**36.25**
Prune tree	21.25	31.25	42.50	32.50	38.75	**51.25**
Plant tree	22.50	33.75	**67.50**	46.25	48.75	51.25
Cut tree	31.25	38.75	53.75	52.50	**78.75**	72.50

The fist line is the average recall over all concepts. The next ten lines are the recalls on randomly selected 10 concepts (same as in Table 2)
Bold values in each line indicate highest one among all methods

hierarchical structure into consideration and jointly optimizing on both two levels, IHNGE and HNGE outperform other algorithms.

3. The average of precisions and recalls of IHNGE are higher than all other baselines, including HNGE. This result demonstrates the contribution of inductive matrix. Practically, all baselines produce desired dimension-reduced feature of a testing sample y by $h_y = W^{\perp}y$, where $W^{\perp}$ is the pseudo-inverse of matrix W. Obviously, process of calculating the pseudo-inverse of matrix W could not guarantee the non-negativity of h_y, which sometimes causes the violation of the non-negativity requirement of non-negative data factorization, and hence harms the classification accuracy. At the same time, our IHNGE directly get $h_y = Cy$ by avoiding to calculate the pseudo-inverse of matrix W and hence guarantees the non-negativity of h_y.
4. On some concepts, like *Answer phone*, the precisions of IHNGE are a little bit lower than baselines. We believe that this is caused by image diversity. Specifically, a base assumption of our hierarchical graph embedding is that, all "verb–object" concepts should have hierarchical structure. Under this assumption, all samples in one sub-class should have a smaller interclass distance on the second level than the one on the first level. However, as all images in our dataset are collected from Internet, the diversity of images in one sub-class may sometimes be huge. This makes the interclass distances on the second level of some samples be higher than the interclass distances on the second level. This huge diversity could harm the classification in IHNGE and hence leads to the lower classification accuracy on some concepts. However, the average of precisions and recalls over all 45 concepts of IHNGE still outperform other baselines.

5 Conclusions

In this paper, we proposed an IHNGE algorithm for "verb–object" concept images classification. Our IHNGE takes hierarchical structure involved in "verb–object" concepts into consideration, and develops the method of feature extraction

and dimensionality reduction based on nonnegative graph embedding. Moreover, we introduce the inductive matrix into our formulations, which could tackle the out-of-sample extension problem against our previous work. The entire "verb–object" concept image classification problem is formulated within the nonnegative data factorization framework, and an efficient iterative procedure is proposed for optimizing the objective function with theoretically and practically convergency. Experiments on the self-collected "verb–object" concept image database demonstrate the effectiveness of our algorithm in "verb–object" concept images classification.

Currently, our IHNGE is a linear projection technique. Those non-linear techniques may improve the classification performance. In future work, we will investigate it.

Acknowledgments This work is supported in part by National Basic Research Program of China (No. 2012CB316304), National Natural Science Foundation of China (No. 61225009, No. 61201374) and Beijing Natural Science Foundation (No. 4131004). This work is also supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office.

Appendix

Here we present the convergence proof of update rule for both matrix W and matrix C.

Preliminaries

First of all, we introduce the concept of auxiliary function and the lemma which will be used for algorithmic derivation.

Definition 1 Function $G(A, A')$ is an auxiliary function for function $F(A)$ if the following conditions are satisfied

$$G(A, A') \geq F(A), \qquad G(A, A) = F(A) \tag{44}$$

From this definition, we have the following lemma with proof omitted [12].

Lemma 1 *If G is an auxiliary function, then F is non-increasing under the update*

$$A^{t+1} = \arg\min_A G(A, A') \tag{45}$$

where t denotes the tth iteration.

Convergence proof of update rule for W

Let F_{ij} as the part of $F(W)$ relevant to W_{ij}, we have

$$F'_{ij}(W) = [W(2Q^1 + 2Q^2 + 2Q^3) + 2\lambda WCXX^TC^T - 2\lambda XX^TC^T]_{ij} \tag{46}$$

$$F''_{ij}(W) = [2(Q^1 + Q^2 + Q^3) + 2\lambda CXX^TC^T]_{jj} \tag{47}$$

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

$$G(W_{ij}, W^t_{ij}) = F_{ij}(W^t_{ij}) + F'_{ij}(W_{ij})(W_{ij} - W^t_{ij}) + \frac{[W^t(Q^1_+ + Q^2_+ + Q^3_+) + \lambda W^tCXX^TC^T]_{ij}}{W^t_{ij}} \times (W_{ij} - W^t_{ij})^2 \tag{48}$$

Lemma 2 *Equation (48) is an auxiliary function for F_{ij}, which is the part of $F(W)$ relevant to W_{ij}.*

Proof Obviously, $G(W_{ij}, W_{ij}) = F_{ij}(W_{ij})$. We only need to prove that $G(W_{ij}, W^t_{ij}) \geq F_{ij}(W_{ij})$.

First, we have the Taylor series expansion of F_{ij}

$$F_{ij}(W_{ij}) = F_{ij}(W^t_{ij}) + F'_{ij}(W^t_{ij})(W_{ij} - W^t_{ij}) + \frac{1}{2}F''_{ij}(W^t_{ij})(W_{ij} - W^t_{ij})^2 \tag{49}$$

Then, it is easy to verify that

$$[\lambda W^tCXX^TC^T]_{ij} \geq W^t_{ij}[\lambda CXX^TC^T]_{jj} \tag{50}$$

$$[W^t(Q^1_+ + Q^2_+ + Q^3_+)]_{ij} \geq W^t_{ij}[(Q^1_+ + Q^2_+ + Q^3_+)]_{jj} \tag{51}$$

Thus we have

$$\frac{[W^t(Q^1_+ + Q^2_+ + Q^3_+) + \lambda W^tCXX^TC^T]_{ij}}{W^t_{ij}} \geq [(Q^1 + Q^2 + Q^3) + \lambda CXX^TC^T]_{jj} \tag{52}$$

Then, $G(W_{ij}, W^t_{ij}) \geq F_{ij}(W_{ij})$ holds.

Lemma 3 *Equation (34) could be obtained by minimizing the auxiliary function $G(W_{ij}, W^t_{ij})$.*

Proof Let $\partial G(W_{ij}, W^t_{ij})/\partial W_{ij} = 0$, we have

$$F'_{ij}(W^t_{ij}) + 2\frac{[W^t(Q^1_+ + Q^2_+ + Q^3_+) + \lambda W^tCXX^TC^T]_{ij}}{W^t_{ij}} \times (W_{ij} - W^t_{ij}) = 0 \tag{53}$$

Finally we can obtain the update rule for W

$$W^{t+1}_{ij} \leftarrow W^t_{ij}\frac{[\lambda XX^TC^T + W^t(Q^1_- + Q^2_- + Q^3_-)]_{ij}}{[\lambda W^tCXX^TC^T + W^t(Q^1_+ + Q^2_+ + Q^3_+)]_{ij}} \tag{54}$$

and the lemma is proved.

Convergence proof of update rule for C

Let F_{ij} as the part of $F(C)$ relevant to C_{ij}, we have

$$F'_{ij}(C) = [2R^{1T}R^1CXL^uX^T + 2R^{2T}R^2CX\tilde{L}X^T + 2R^{3T}R^3CX\tilde{L}^uX^T - 2\lambda W^TXX^T + 2\lambda W^TWCXX^T]_{ij} \tag{55}$$

$$F''_{ij}(C) = 2[R^{1T}R^1]_{ii}[XL^uX^T]_{jj} + 2[R^{2T}R^2]_{ii}[X\tilde{L}X^T]_{jj} + 2[R^{3T}R^3]_{ii}[X\tilde{L}^uX^T]_{jj} + 2\lambda[W^TW]_{ii}[XX^T]_{jj} \quad (56)$$

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

$$\begin{aligned} &G(C_{ij}, C^t_{ij}) \\ &= F_{ij}(C^t_{ij}) + F'_{ij}(C_{ij})(C_{ij} - C^t_{ij}) \\ &\quad + [R^{1T}R^1C^tXD^uX^T + R^{2T}R^2C^tX\tilde{D}X^T \\ &\quad + R^{3T}R^3C^tX\tilde{D}^uX^T + \lambda W^TWC^tXX^T]_{ij}/C^t_{ij} \\ &\quad \times (C_{ij} - C^t_{ij})^2 \end{aligned} \quad (57)$$

Lemma 4 *Equation (57) is an auxiliary function for F_{ij}, which is the part of $F(C)$ relevant to C_{ij}.*

Proof Obviously, $G(C_{ij}, C_{ij}) = F_{ij}(C_{ij})$. We only need to prove that $G(C_{ij}, C^t_{ij}) \geq F_{ij}(C_{ij})$.

First, we have the Taylor series expansion of F_{ij}

$$F_{ij}(C_{ij}) = F_{ij}(C^t_{ij}) + F'_{ij}(C^t_{ij})(C_{ij} - C^t_{ij}) + \frac{1}{2}F''_{ij}(C^t_{ij})(C_{ij} - C^t_{ij})^2 \quad (58)$$

Then, it is easy to verify that

$$[W^TWCXX^T]_{ij} \geq C^t_{ij}[W^TW]_{ii}[XX^T]_{jj} \quad (59)$$

$$[R^{1T}R^1C^tXD^uX^T]_{ij} \geq [R^{1T}R^1]_{ii}C^t_{ij}[XL^uX^T]_{jj} \quad (60)$$

$$[R^{2T}R^2C^tX\tilde{D}X^T]_{ij} \geq [R^{2T}R^2]_{ii}C^t_{ij}[X\tilde{L}X^T]_{jj} \quad (61)$$

$$[R^{3T}R^3CX\tilde{D}^uX^T]_{ij} \geq [R^{3T}R^3]_{ii}C^t_{ij}[X\tilde{L}^uX^T]_{jj} \quad (62)$$

Thus we have $G(C_{ij}, C^t_{ij}) \geq F_{ij}(C_{ij})$.

Lemma 5 *Equation (43) could be obtained by minimizing the auxiliary function $G(C_{ij}, C^t_{ij})$.*

Proof Let $\partial G(C_{ij}, C^t_{ij}) / \partial C_{ij} = 0$, we have

$$\begin{aligned} &F'_{ij}(C_{ij}) + [R^{1T}R^1C^tXD^uX^T + R^{2T}R^2C^tX\tilde{D}X^T \\ &\quad + R^{3T}R^3C^tX\tilde{D}^uX^T + \lambda W^TWC^tXX^T]_{ij}/C^t_{ij} \\ &\quad \cdot (C_{ij} - C^t_{ij}) = 0 \end{aligned} \quad (63)$$

Finally we can obtain the update rule for C

$$\begin{aligned} C^{t+1}_{ij} \leftarrow C^t_{ij} \cdot &[\lambda W^TXX^T + R^{1T}R^1C^tXS^uX^T \\ &+ R^{2T}R^2C^tX\tilde{S}X^T + R^{3T}R^3C^tX\tilde{S}^uX^T]_{ij} \\ &/ [\lambda W^TWC^tXX^T + R^{1T}R^1C^tXD^uX^T \\ &+ R^{2T}R^2C^tX\tilde{D}X^T + R^{3T}R^3C^tX\tilde{D}^uX^T]_{ij} \end{aligned} \quad (64)$$

and the lemma is proved. □

References

1. Belhumeur, P., Hespanha, J.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 711–720 (1997)
2. Carneiro, G., Chan, A., Moreno, P., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **29**(3), 394–410 (2007)
3. Ding, C.H., Li, T., Jordan, M.I.: Convex and semi-nonnegative matrix factorizations. IEEE Trans. Pattern Anal. Mach. Intell. **32**(1), 45–55 (2010)
4. Gao, Y., Fan, J., Xue, X., Jain, R.: Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers. Proceedings of the 14th annual ACM international conference on Multimedia, pp. 901–910. ACM, New York (2006)
5. Heger, A., Holm, L.: Sensitive pattern discovery with fuzzyalignments of distantly related proteins. Bioinformatics **19**(suppl 1), i130–i137 (2003)
6. Hong, R., Tang, J., Tan, H.-K., Ngo, C.-W., Yan, S., Chua, T.-S.: Beyond search: event-driven summarization for web videos. TOMCCAP **7**(4), 35 (2011)
7. Hong, R., Wang, M., Li, G., Nie, L., Zha, Z.-J., Chua, T.-S.: Multimedia question answering. IEEE Multimed. **19**(4), 72–78 (2012)
8. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. **5**, 1457–1469 (2004)
9. Hu, C., Zhang, B., Yan, S., Yang, Q., Yan, J., Chen, Z., Ma, W.: Mining ratio rules via principal sparse non-negative matrix factorization. In Fourth IEEE International Conference on Data Mining, 2004. ICDM'04, pp. 407–410. IEEE (2004)
10. Kim, P., Tidor, B.: Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res. **13**(7), 1706–1718 (2003)
11. Kuhn, H.W., Tucker, A.W.: Nonlinear programming. In: Second Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 481–492 (1951)
12. Lee, D., Seung, H., et al.: Learning the parts of objects by nonnegative matrix factorization. Nature **401**(6755), 788–791 (1999)
13. Li, L., Jiang, S., Huang, Q.: Learning hierarchical semantic description via mixed-norm regularization for image understanding. IEEE Trans. Multimed. **14**(5), 1401–1413 (2012)
14. Li, L.-J., Su, H., Fei-Fei, L., Xing, E.P.: Object bank: a high-level image representation for scene classification & semantic feature sparsification, pp. 1378–1386. In: Advances in Neural Information Processing Systems (2010)
15. Li, S.Z., Hou, X.W., Zhang, H.J., Cheng, Q.S.: Learning spatially localized, parts-based representation. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001, vol. 1, pp. I-207. IEEE (2001)
16. Liu, X., Yan, S., Jin, H.: Projective nonnegative graph embedding. IEEE Trans. Image Process. **19**(5), 1126–1137 (2010)
17. Ramanath, R., Kuehni, R., Snyder, W., Hinks, D.: Spectral spaces and color spaces. Color Res. Appl. **29**(1), 29–37 (2004)
18. Ramanath, R., Snyder, W., Qi, H.: Eigenviews for object recognition in multispectral imaging systems. In: Applied Imagery Pattern Recognition Workshop, 2003. Proceedings. 32nd, pp. 33–38. IEEE (2003)
19. Sun, C., Bao, B.-K., Xu, C.: Verb-object concepts image classification via hierarchical nonnegative graph embedding. In: Proceeding of 19th International Conference on Multimedia Modeling (MMM), pp. 58–69 (2013)
20. Wang, C., Song, Z., Yan, S., Zhang, L., Zhang, H.: Multiplicative nonnegative graph embedding. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, pp. 389–396. IEEE (2009)

21. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 3360–3367. IEEE (2010)
22. Wang, M., Hong, R., Li, G., Zha, Z.-J., Yan, S., Chua, T.-S.: Event driven web video summarization by tag localization and key-shot identification. IEEE Trans. Multimed. **14**(4), 975–985 (2012)
23. Wang, Y., Jia, Y.: Fisher non-negative matrix factorization for learning local features. In: Proc. Asian Conf. on Comp. Vision, Citeseer (2004)
24. Yan, S., Xu, D., Zhang, B., Zhang, H., Yang, Q., Lin, S.: Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. **29**(1), 40–51 (2007)
25. Yang, J., Yang, S., Fu, Y., Li, X., Huang, T.: Non-negative graph embedding. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pp. 1–8. IEEE (2008)
26. Yao, B., Jiang, X., Khosla, A., Lin, A., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: IEEE International Conference on Computer Vision (ICCV), 2011, pp. 1331–1338. IEEE (2011)
27. Yun, X.: Non-negative matrix factorization for face recognition. PhD thesis, Hong Kong Baptist University (2007)
28. Zhang, X., Zha, Z., Xu, C.: Learning verb-object concepts for semantic image annotation. Proceedings of the 19th ACM International Conference on Multimedia, pp. 1077–1080. ACM, New York (2011)

Author Biographies

Chao Sun received his B.E. degree at University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2006. He is currently a Ph.D. student in Multimedia Computing Group, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. In 2012, he was an intern student in China-Singapore Institute of Digital Media, Singapore. His current research interests include multimedia and computer vision.

Bing-Kun Bao received the Ph.D. degree in Control Theory and Control Application, Department of Automation, University of Science and Technology of China (USTC), China, in 2009. From 2009 to 2011, she was a research engineer in Electrical and Computer Engineering at National University of Singapore (NUS). She is currently a assistant researcher in Institute of Automation, Chinese Academy of Science, and a researcher in China-Singapore Institute of Digital Media.

Changsheng Xu is Professor in National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences and Executive Director of China-Singapore Institute of Digital Media. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition and computer vision. He has hold 30 granted/pending patents and published over 200 refereed research papers in these areas. Dr. Xu is an Associate Editor of IEEE Transactions on Multimedia and ACM Transactions on Multimedia Computing, Communications. He served as Program Chair of ACM Multimedia 2009. He has served as associate editor, guest editor, general chair, program chair, area/track chair, special session organizer, session chair and TPC member for over 20 IEEE and ACM prestigious multimedia journals, conferences and workshops. He is ACM Distinguished Scientist.