Hidden-concept Driven Multi-label Image Annotation and Label Ranking
Bing-Kun Bao∗, Teng Li, and Shuicheng Yan, Senior Member

Abstract—Conventional semi-supervised image annotation algorithms usually propagate labels predominantly via holistic similarities over image representations, and do not fully consider the label locality, inter-label similarity, and intra-label diversity among multi-label images. Taking these problems into consideration, we present the hidden-concept driven image annotation and label ranking algorithm (HDIALR), which conducts label propagation based on the similarity over a visually-semantically consistent hidden-concepts space. The proposed method has such following characteristics: 1) each holistic image representation is implicitly decomposed into label representations to reveal label locality. The decomposition is guided by the so-called hidden concepts, characterizing image regions and reconstructing both visual and non-visual labels of the entire image; 2) each label is represented by a linear combination of hidden concepts, while the similar linear coefficients reveal the inter-label similarity; 3) each hidden concept is expressed as a respective subspace, and different expressions of the same label over the subspace then induce the intra-label diversity; and 4) the sparse coding based graph is proposed to enforce the collective consistency between image labels and image representations, such that it naturally avoids the dilemma of possible inconsistency between the pairwise label similarity and image representation similarity in multi-label scenario. These properties are finally embedded in a regularized nonnegative data factorization formulation, which decomposes images representations into label representations over both labeled and unlabeled data for label propagation and ranking. The objective function is iteratively optimized by a convergence provable updating procedure. Extensive experiments on three benchmark image datasets well validate the effectiveness of our proposed solution to semi-supervised multi-label image annotation and label ranking problem.

I. INTRODUCTION

Rapid advance in technology for image capturing, processing, distribution, storage, and sharing have resulted in the proliferation of image data. How to search for these images effectively and efficiently is an increasingly urgent research topic. Most users prefer searching images by textual queries such as “birds in grass” and expect to obtain images with “bird” as the most relevant object and “grass” as the second most relevant object. Therefore, content-based image annotating and label ranking image labels play an important role in image search, however, it is challenging in three aspects caused by the general multiple labels nature of a single image.

1) Label Locality: As shown in Figure 1a, most labels are only related to their corresponding semantic regions.

2) Inter-Label Similarity: There are several relationships between different labels, such as hierarchical relationship, co-occurrence relationship and so on [3] [4], as shown in Figure 1b. These relationships, named inter-label similarity, are quite necessary to be considered to improve the accuracy in label propagation.

3) Intra-Label Diversity: For each label, its corresponding regions at different images can be different. For example, the label “sky” could infer various expressions, such as cloudy, dark, clear sky and so on, as shown in Figure 1c. The diversity among these expressions for each label is called as intra-label diversity. We need to keep in mind on intra-label diversity so that to eliminate the gap among different expressions in label propagation.

Chen et al. proposed to construct graph on label level to reveal the label correlations [3]. Liu et al. introduced a label similarity matrix to provide a semi-supervised learning algorithm [4]. These methods propagate labels based on holistic image similarity, which does not separate the similarity among different labels. On the other hand, using holistic image similarity will involve other associated labels when propagating a certain label. Thus, the local label similarity should be used instead of holistic image similarity for achieving better annotation performance. Methods based on image decomposition have also been explored. [5] [6] [7] considered each image as a bag of multiple segmented regions and predicts the label of each region by a multi-class bag classifier. This method, however, is limited with the process of segmentation that the manual solution is very time-consuming while the automatic

∗ Correspondence author, email: bingkunbao@gmail.com.
Bing-Kun Bao and Shuicheng Yan are with National University of Singapore, Singapore (email: {elebb, eleyans}@nus.edu.sg).
Teng Li is with Chinese Academy of Sciences, China. (email:tenglwy@gmail.com)

1According to WordNet [2], beach is an area of sand sloping down to the water of a sea or lake.
solution is still far from satisfaction. In this paper, we consider to implicitly decompose the label representation on feature level, which avoids the explicit image segmentation process.

We introduce the so-called hidden concepts, which are expected to represent the visually-semantically consistent subspaces, to resolve the above three challenges. The hidden concepts target at three characteristics corresponding to the above three challenges respectively. Firstly, each label can be represented as the linear combination of hidden concepts, as shown in Figure 2a, so that the inter-label similarity can be revealed by these linear combination coefficients. It is obvious that similar linear coefficients indicate high inter-label similarity, or more specifically, for a given linear combination coefficient matrix as shown in Figure 2b, the inter-label similarity matrix can be expressed as the product of the coefficient matrix and its transpose. Secondly, each hidden concept is expressed as its respective subspace, in which different elements correspond to different expressions of the hidden concept. The intra-label diversity will be covered by these elements. Lastly, the label locality is implicitly revealed by the decomposition from image representation into label representations. The decomposition is guided by factorizing image representation matrix into a basis matrix, which is the combination of the subspace bases from all the hidden concepts, and a coefficient matrix. Then, the label representation can be obtained by summarizing the corresponding hidden concepts with the weighting vector as their linear combination coefficients, as shown in Figure 2c.

Based on the hidden concept, we propose a novel propagation-by-decomposition solution to image annotation and label ranking problems. The basic idea is to arrange the holistic image feature of the labeled and unlabeled images as the columns of a data matrix, and then decompose the data matrix as the product of two matrices that possess only nonnegative elements, i.e., hidden concept basis matrix and hidden concept coefficient matrix. Herein, the hidden concept basis matrix is combined by all the hidden concepts’ subspaces bases, and the element of each column in coefficient matrix is the corresponding coefficient for each basis. Then, a relation matrix is introduced to measure the relationship between hidden concepts and labels, in which each row shows the linear coefficient of hidden concepts to represent each label. With the knowledge of the relation matrix, the hidden concept coefficient matrix will reduce to a label coefficient matrix, which indicates the labels associated to each images and ranking results of these labels. The expectations of the formulation are three-folds:

1) **Representation: image to hidden concepts:** It makes sure the consistency in representation between image and hidden concepts, that is, image representation can be
Fig. 2: (a) The relationship between labels and hidden concepts. The vertical-axis shows 20 labels from the 90 selected labels in Corel5k dataset, and horizontal-axis indicates 60 hidden concepts. The number of hidden concepts is 20 less than that of labels. It shows that each label can be linear represented by hidden concepts. The red color indicates large linear coefficient, while blue color indicates small linear coefficient. The label “sky” and “sun” share 3 hidden concepts, which reveal large the inter-label similarity between them. (b) The inter-label similarity, which is obtained as the product of the coefficient matrix shown in (a) and its transpose matrix. (c) The local label representation is obtained by summarizing the corresponding hidden concepts with the weighting vector as their linear combination coefficients.

Fig. 2: (a) The relationship between labels and hidden concepts. The vertical-axis shows 20 labels from the 90 selected labels in Corel5k dataset, and horizontal-axis indicates 60 hidden concepts. The number of hidden concepts is 20 less than that of labels. It shows that each label can be linear represented by hidden concepts. The red color indicates large linear coefficient, while blue color indicates small linear coefficient. The label “sky” and “sun” share 3 hidden concepts, which reveal large the inter-label similarity between them. (b) The inter-label similarity, which is obtained as the product of the coefficient matrix shown in (a) and its transpose matrix. (c) The local label representation is obtained by summarizing the corresponding hidden concepts with the weighting vector as their linear combination coefficients.

described by the product of hidden concept basis matrix and hidden concept coefficient matrix.

2) Known-Knowledge Consistency: This consistency infers that the relationship between ground-truth and the predicted labels on labeled data. That is, given the knowledge of hidden concept coefficient matrix, the derived label coefficients should be consistent with the ground-truth label vectors of the labeled data.

3) Collective Label-Feature Consistency: This consistency infers that if an image can be sparsely represented by other images, its labels also can sparsely represented by these images’ labels. This fold aims to smoothly propagate the labels from labeled data to unlabeled data. To guarantee it, \( \ell_1 \)-graph \[8\] is employed to robustly represent one image as the sparse linear combination of the other images, which introduces \( \ell_1 \)-graph weight matrix as a parameter matrix in the formulation.

All these folds involve four parameter matrices: hidden concept basis matrix, hidden concept coefficient matrix, relation matrix, and \( \ell_1 \)-graph weight matrix. The optimization of this formulation can be transformed into a set of tractable sub-problems, and achieves the convergence to a local optimum by a multiplicative nonnegative iterative procedure. Based on the optimal hidden concept coefficient matrix, the predicted labels and the corresponding ranking results can be at last obtained by summarizing the relative coefficients.

II. RELATED WORKS

In this section, we introduce some existing methods in multi-label image annotation and label ranking.

A. Multi-label Image Annotation

Traditional multi-label learning algorithms transfer the original problem into multiple independent binary classification problems \[9\] \[10\] \[11\]. These algorithms, however, neglect the underlying correlations among different labels or are of the higher computational cost. \[12\] \[13\] \[14\] incorporated hierarchical mixture models and concepts hierarchy into semantic image classification. \[15\] employed Bayesian network for automatic image classification. Since these relationships are based on prior knowledge of object parts and their locations,
which are predefined and fixed. If there are some corruptions in the context or predefined contextual relationships, the final annotation is also corrupted. Methods from natural language processing area such as probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA), which also utilize the idea of hidden or latent semantics, have been successfully employed in multimedia circles in the last many years [16][17], where the hidden semantics are represented by words occurrence likelihoods, and the visual words are considered independently in the generative model. In this paper the hidden concepts are modeled with the vectors of visual words aggregation expected to fully describe the local conceptual areas, in lighting of the local representation in [18]. Comparing to pLSA or LDA, the proposed method not only incorporates semi-supervised learning idea, but also can fully utilizes the information of local concept combinations from the whole images. Therefore, the label similarities based on decomposed bag of words vectors can be more accurate than visual words level models.

The label co-occurrence is also useful to increase the accuracy in label propagation. [19] utilized co-occurrence of objects to improve the performance of object detection. [20] explored the co-occurrence of the labels in its large-scale image annotation algorithm. Several methods are also proposed to utilize the relations among concepts in the learning process. In [21], Qi et al. proposed a novel Correlative Multi-Label (CML) framework to simultaneously classify concepts and model correlations among them. Chen et al. presented a graph-based semi-supervised multi-label learning algorithm by solving a sylvestre equation [3], which is denoted as SMSE in this paper. This algorithm firstly constructs two graphs on sample level and category level respectively, then defines a quadratic energy function on each graph, and finally obtains the labels of the unlabaled images by minimizing the combination of the two energy functions. Liu et al. provided another semi-supervised learning algorithm to optimize the consistency between image similarity and label similarity by solving a constrained non-negative matrix factorization (CNMF) [4] problem.

Most previous approaches considering the label similarity are based on a common assumption that the global image similarity and the label similarity are consistent. This assumption ignores that each label often only characterizes a local region within an image while the image similarity is calculated based on the whole image. Some recent works explored the relationships between labels and regions by considering each image as a bag of multiple regions and predict the label of each region by a multi-class bag classifier [5][6][7]. These works, however, heavily rely on the performance of image segmentation, and are thus very sensitive to image noises.

B. Label Ranking

Label ranking is to learn a mapping from samples to rankings over the labels associated. Some early works are based on extensions of algorithms for binary classification. [22] proposed to learn a binary preference models for each pair of labels, then combine these models to rank all the labels. [23] constrained classification and log-linear models for label ranking. [24] proposed to firstly estimate the tag relevance scores using kernel density estimation, and further use random walk procedure to boost this primary estimation based on a similarity graph built on the testing image. This method can also be utilized in the label ranking problem. All these existing algorithms, however, are based on the information of predicted labels associated to each sample. Different from these algorithms, our approach simultaneously annotates image and ranks labels, which decreases the running time cost and makes those online applications possible.

III. Problem Formulation

Let $X = [x_1, x_2, \ldots, x_{n_1}, x_{n_1+1}, \ldots, x_{n-1}, x_n]$ be the set of holistic image representation vectors for $n$ images in the dataset, where $x_i \in \mathbb{R}^{r_i}$ denotes the $i$-th image with $r$-dimensional feature, $n_1$ is the number of labeled images, and $n_2 = n - n_1$ is the number of unlabeled images. Note that the symbol $\mathbb{R}^+$ means the nonnegative real value set in this work. The label ground truth for labeled data is set as $H_l = [h_{1,1}, h_{1,2}, \ldots, h_{1,n_1}]$, where $h_{i,j}$ is a $c$-dimensional label vector with the value of 0 or 1, $c$ is the number of labels, and the value “1” indicates that the image is fully relevant to a single label, while “0” indicates irrelevant. Denote the label coefficient matrix for unlabeled images as $H_u = [\hat{h}_{n_1+1,1}, \hat{h}_{n_1+2,1}, \ldots, \hat{h}_{n_2,1}]$, where $\hat{h}_{i,j} \in [0,1]^c$ indicates the weights of the labels assigned to $i$-th image. The task for this work is to propagate the labels of those labeled images to unlabeled images, meanwhile rank the associated labels, that is, to seek the values for $H_u^2$.

Denote the $i$-th hidden concept subspace basis as $W^i \in \mathbb{R}^{n \times k'}$, where $k'$ is the dimension for each hidden concept subspace. Let the number of hidden concept be $c'$, then all the hidden concept subspace bases are combined to hidden concept basis matrix $W$, that is $W = [W^1, W^2, \ldots, W^{c'}] \in \mathbb{R}^{n \times k'}$, where $k = k' \times c'$. The hidden concept coefficient matrix is denoted as $H \in \mathbb{R}^{+}_{c \times n}$, and relation matrix, which expresses the relationship between hidden concepts and labels, is denoted as $P \in \mathbb{R}^{+}_{c \times c'}$.

A. Representation: Image to Hidden Concepts

As aforementioned, information-mixed holistic image representation may not be the best way to characterize label information due to label locality. To ensure the consistency between feature and label, we decompose each image representation into a set of hidden concepts representations, which can be linearly combined into label representation without explicit image segmentation process.

Mathematically, the image representation matrix $X$ is factorized as $X = WH$, where $W$ is hidden concept basis matrix, and $H = [h_{1,1}, h_{1,2}, \ldots, h_{1,n}]$ is hidden concept coefficient matrix,

\[H = [H_l, H_u]\] as the hidden concept coefficient matrix, $H = [H_l, H_u]$ as the label ground truths matrix over the whole dataset, and denote $H = [H_l, H_u]$ as the predicted label coefficient matrix from our formulation.

The intrinsic dimension for each subspace can be different, but a sufficiently large $k'$ is enough, since some components in $W^i$ can be zeros or linearly dependent.
where \( h_i \) is the hidden concept coefficient vector for the \( i \)-th image. In practice, we express this factorization as to optimize the objective function,

\[
\min_{W,H} \| X - WH \|_F^2, \quad \text{s.t. } W,H \succeq 0.
\]

(1)

For a given image representation \( x_i \), the minimization of \( \| x_i - Wh_i \|_F^2 \) actually enforces \( x_i \) to be reconstructed by the linear combination of the column vectors of \( W \). By doing so, we decompose holistic image representation as the linear representation over the hidden concept subspaces.

**B. Known-Knowledge Consistency**

The formulation of Eqn. (1) is exactly the same as that for Nonnegative Matrix Factorization [25], and thus extra regularization terms are required to achieve the characteristics for the hidden concept basis matrix. One of such terms is to enforce the consistence between the derived hidden concept coefficient vectors and the ground truth label vectors for each labeled data.

Denote hidden concept coefficient matrix as \( H = [H_L, H_U] \), where \( H_L \) is the hidden concept coefficient matrix for labeled data, and \( H_U \) is that for unlabeled matrix. From \( H_L \), we can predict the label coefficient matrix \( \tilde{H}_L \) for the labeled data, that is

\[
\tilde{H}_L = PEQH_L.
\]

(2)

Herein, \( P \) is relation matrix, every row of which constructs the mapping from all the hidden concepts to alabel. The constant matrix \( E \in \{0,1\}^{c \times k} \) is used to combine every \( k' \) coefficient elements in \( H_L \) to obtain the total coefficient for each hidden concept, that is,

\[
E_{i,j} = \begin{cases} 
1 & \text{if } i = 1,2,\cdots,c' \text{ and } j = k'(i-1) + 1,\cdots,k'; \\
0 & \text{otherwise}. 
\end{cases}
\]

\( Q = \text{diag}(e^T W) \) is used to convey the \( \ell_1 \) norms of the column vectors of \( W \) to \( H \) such that \( W \) is virtually columnly normalized, and \( e \) is an all-one vector. Mathematically, to enforce consistence between the predicted labels and original labels on the labeled data, the following objective function is optimized,

\[
\min_{P,W,H_L} \| \tilde{H}_L - PEQH_L \|_F^2, \quad \text{s.t. } P, W, H_L \succeq 0.
\]

(3)

The minimization of the above formulation reveals that the predicted labels should be approximate to the ground truths of the labeled data.

**C. Collective Label-Feature Consistency**

For multi-label image data, the pairwise global representation similarity and label similarity may be inconsistent due to the inter-label similarity and intra-label diversity. However, in semi-supervised learning, the consistency between image representations and labels plays an important role for propagating labels from labeled data to unlabeled ones. In this work, beyond pairwise image representation similarity, we utilize collective representation similarity to propagate labels among data, that is, to construct an \( \ell_1 \)-graph [8] to enhance the consistency between labels and features. The graph vertex corresponds to the combination of image representation and image label, and they can be represented by as less other nodes as possible, referred to sparse coding [26]. The weight between two vertices is set as the sparse coefficient between the corresponding two images.

More specifically, denote \( G = \{V, S\} \) as an \( \ell_1 \)-graph with \( V = (X \tilde{H})^T \) as vertices and \( S \) as weight matrix. Each vertex \( v_i = (x_i \tilde{h}_i)^T \) is assumed to be reconstructed with the remaining data, denoted as \( V^i = [v_1, \cdots, v_{i-1}, v_{i+1}, \cdots, v_n] \), with sparse representation \( s^i \) [27]. Generally, such a solution shall be sparse, and thus \( s^i \) characterizes the sparse and collective relation between the data \( v_i \) and all the other data, namely,

\[
\min_{s^i} \| v_i - V^i s^i \|_F^2 + \| s^i \|_1, \quad \text{s.t. } s \succeq 0.
\]

(4)

For the whole image set, let \( S \in \mathbb{R}_{+}^{n \times n} \), referred to \( \ell_1 \)-graph weight matrix, characterize such sparse reconstruction relations, and \( s^i \) constitutes the \( i \)-th row vector of the matrix \( S \) (with \( S_{ii} \) ignored and set as zero). So, the collective label-feature consistency can be optimized by

\[
\min_{S} \left\| \left( \begin{array}{c} X \\ \tilde{H} \end{array} \right) - \left( \begin{array}{c} X \\ \tilde{H} \end{array} \right) S \right\|_F^2 + \gamma \| S \|_1, \quad \text{s.t. } S \succeq 0.
\]

(5)

Recall that, the label coefficient matrix \( \tilde{H} = PEQH \), similar to Eqn. (2), then in our objective function, Eqn. 5 can be rewritten as

\[
\min_{P,W,H,S} \left\| \left( \begin{array}{c} X \\ PEQH \end{array} \right) - \left( \begin{array}{c} X \\ PEQH \end{array} \right) S \right\|_F^2 + \gamma \| S \|_1, \quad \text{s.t. } S \succeq 0.
\]

(6)

As the extension of our previous work [28], the vertex of \( \ell_1 \)-graph is the combination of image representation and image label instead of image representation only. Recall that, in \( \ell_1 \)-graph, each vertex should be sparsely represented by clean data to remedy the noise. In practice, the clean images are not easy to guarantee. so the \( \ell_1 \)-graph on only image representation is possible to be unreliable. Thus, we add the label information, which is robust to image noise, into \( \ell_1 \)-graph construction in this paper. Experiments as introduced later demonstrate that the extended version achieves better performance than the previous version in [28].

**D. Unified Problem Formulation as Regularized Nonnegative Data Factorization**

By integrating all the above three folds, we can formulate the semi-supervised multi-label learning problem as a regularized nonnegative data factorization problem in the following form,

\[
\begin{aligned}
\min_{W,H,P,S} & \| X - WH \|_F^2 + \alpha \| PEQH_L - \tilde{H}_L \|_F^2 \\
& + \beta \left\| \left( \begin{array}{c} X \\ PEQH \end{array} \right) (I - S) \right\|_F^2 + \beta \gamma \| S \|_1 \\
\text{s.t. } & W, H, P, S \succeq 0.
\end{aligned}
\]

The parameter \( \alpha \) and \( \beta \) in Eqn. (7) are the weights to balance the three terms. The \( W, H, P, S \) are all set to be nonnegative.
W represents the hidden concept basis matrix which captures the specific characteristics of each label. \( H \) implicitly indicates the confidences of each data associated to each hidden concept subspace basis. \( P \) indicates the relevances between the hidden concept and original labels. \( S \) implicates the \( \ell_1 \)-graph weights for whole image set.

In our previous work [28], we only focused on propagating the labels from labeled data to unlabeled ones, yet did not consider to rank these labels. However, label ranking is in urgent need in many areas, such as image search [29], label recommendation [30] etc. Recent research studied how to rank the labels on labeled data, that is, we need to know the labels first, and then rank them [24]. In this paper, we can obtain the label rank based on the resulting coefficient matrix as follows.

As aforementioned, each sample can be represented as the linear combination of the associated label representations. Obviously, the combination coefficients provide us weights to rank the associated labels. If the coefficient value is high, the rank of the corresponding label will be high. Specifically, in our approach, a sample \( x \) is decomposed into the coefficient on each hidden concept subspace, that is, \( x = Wh \). Given this hidden concept coefficient \( h \), the label coefficient can be easily derived by combining the relative hidden concepts with the corresponding relative coefficients, that is,

\[
\hat{H} = PEQH.
\]

Thus, the rank of labels is obtained simultaneously with the image label propagation. The proposed label ranking can be taken not only for test images, but also for the training ones.

**IV. ITERATIVE OPTIMIZING RULES AND ALGORITHM**

The problem defined in Eqn. (7) falls into the framework of regularized nonnegative matrix factorization [25] [31]. The objective function is of high order, and here we optimize the objective function in an iterative way with a multiplicative updating rule, which guarantees the nonnegativity of the 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 for whole image set.

Based on the update rules, we conclude the whole process in Algorithm 1, which is referred to as Hidden-concept Drive Image Annotation and Label Ranking Algorithm (HDIALR).

**V. EXPERIMENTS**

In this section, the proposed solution is systematically evaluated on three datasets for multi-label image annotation. Furthermore, label ranking results on MSRC [32] and Corel5k [1] is also given.

**A. Datasets and Representations**

The three benchmark datasets used in experiments are MSRC [32], Corel5k [1] and NUS-WIDE [33]. In order to show that the proposed algorithm is robust to the different sub-dataset selections, we experiment the algorithms on three different sub-dataset selection methods over these three datasets respectively, that is, no sub-dataset selection on MSRC, sub-dataset selection by labels with sufficient associated images on Corel5k, and sub-dataset selection by images with sufficient associated labels on NUS-WIDE. Two parameters are selected to describe the datasets used, that is, label cardinality, which shows how many labels have been assigned to a dataset example in average, and label density, which is defined as the fraction of the average number of the labels used in each dataset to total number of available labels [34].

The MSRC dataset contains 591 images associated with 374 labels. Since some labels are relatively low-level, where each pixel is labeled as one of 23 labels or “void”. As suggested by MSRC, “horse” and “mountain” are treated as “void” because they only have few labeled images. Therefore, there are 21 labels in our experiment. The label cardinality and label density are 2.4602 and 0.1172 respectively. The Corel5k dataset\(^4\) in our experiments contains 5000 images associated with 374 labels. Since some labels are relatively low-level, the specific characteristics of each label. \( H \) implicitly indicates the confidences of each data associated to each hidden concept subspace basis. \( P \) indicates the relevances between the hidden concept and original labels. \( S \) implicates the \( \ell_1 \)-graph weights for whole image set.

In our previous work [28], we only focused on propagating the labels from labeled data to unlabeled ones, yet did not consider to rank these labels. However, label ranking is in urgent need in many areas, such as image search [29], label recommendation [30] etc. Recent research studied how to rank the labels on labeled data, that is, we need to know the labels first, and then rank them [24]. In this paper, we can obtain the label rank based on the resulting coefficient matrix as follows.

As aforementioned, each sample can be represented as the linear combination of the associated label representations. Obviously, the combination coefficients provide us weights to rank the associated labels. If the coefficient value is high, the rank of the corresponding label will be high. Specifically, in our approach, a sample \( x \) is decomposed into the coefficient on each hidden concept subspace, that is, \( x = Wh \). Given this hidden concept coefficient \( h \), the label coefficient can be easily derived by combining the relative hidden concepts with the corresponding relative coefficients, that is,

\[
\hat{H} = PEQH.
\]

Thus, the rank of labels is obtained simultaneously with the image label propagation. The proposed label ranking can be taken not only for test images, but also for the training ones.

The problem defined in Eqn. (7) falls into the framework of regularized nonnegative matrix factorization [25] [31]. The objective function is of high order, and here we optimize the objective function in an iterative way with a multiplicative updating rule, which guarantees the nonnegativity of the 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. The proposed iterative procedure also follows this philosophy and optimizes \( W \), \( H \), \( P \), and \( S \) iteratively. From the deduction, which is detailed in Appendix, we can obtain the convergence provable update rules for all variable matrix as follows,

\[
W_{ij} = W_{ij} \times \frac{(XH^T + A^1_{W})_{ij}}{(WHH^T + A^1_{W} + A^3_{W})_{ij}},
\]

\[
P_{ij} = P_{ij} \times \frac{(A^2_{P})_{ij}}{(A^1_{P} + A^3_{P})_{ij}},
\]

\[
H_{ij} = H_{ij} \times \frac{(A^2_{H} + WTX)_{ij}}{(WTHW + A^1_{H} + A^3_{H})_{ij}},
\]

\[
S = \min_{S} \left\{ \| XPEQH S - \left( \begin{array}{c} X \\ PEQH \end{array} \right) \|_F^2 + \gamma \| S \|_1, \ s.t. \ S \geq 0 \right\}
\]

\[
\text{subject to } S \geq 0.
\]

\[
A^1_{W} = \alphaee^TW[H_LH_L^T] \cdot ((PE)^T(PE)),
\]

\[
A^2_{W} = \alpha ee_{1 \times 1}[\text{diag}(H_LH_L^T)PE]^T,
\]

\[
A^3_{W} = \beta ee^TW[(H(I - S)(I - S)^TH^T) \cdot ((PE)^TPE)],
\]

\[
A^1_{P} = \alpha PEQH_LH_L^TQTE^T,
\]

\[
A^2_{P} = \alpha H_LH_L^TQTE^T,
\]

\[
A^3_{P} = \beta PEQH(I - S)(I - S)^TH^TQTE^T,
\]

\[
A^1_{H} = \alpha (PEQ)^TPEQH \left[ \begin{array}{cc} I_{N_1 \times N_1} & 0_{N_1 \times N_2} \\ 0_{N_2 \times N_1} & 0_{N_2 \times N_2} \end{array} \right],
\]

\[
A^2_{H} = \alpha \left[ (PEQ)^T \hat{H}_L \begin{array}{c} \begin{array}{c} 0_{K \times N_1} \end{array} \end{array} \right],
\]

\[
A^3_{H} = \beta (PEQ)^TPEQH(I - S)(I - S)^T.
\]

The initial value of \( S \) can be set as

\[
S^0 = \min_{S} \| X - XS \|_F^2 + \gamma_0 \| S \|_1, \ s.t. \ S \geq 0,
\]

and the value of \( \gamma \) in Eqn. (7) can be simply set equal to \( \gamma^0 \).

Based on the update rules, we conclude the whole process in Algorithm 1, which is referred to as Hidden-concept Drive Image Annotation and Label Ranking Algorithm (HDIALR).

\[\text{Download at http://research.microsoft.com/vision/cambridge/recognition/}\]

\[\text{Download at http://kobus.ca/research/data/eccv_2002/index.html}\]
Algorithm 1 Hidden-concept Driven Image Annotation and Label Ranking Algorithm (HDIALR)

1: **Input:** Image representation matrix $X$, annotated label matrix $H_L$;
2: **Initialization:** Randomly choose $W^0$, $H^0$, and $P^0$ as nonnegative matrices;
3: for $t = 0, 1, 2, \ldots, T_{\text{max}}$, do
   1) For given $H = H^t$, $P = P^t$, $S = S^t$, update the hidden concept basis matrix $W$ as:
      $$W_{ij}^{t+1} = W_{ij}^t \times \frac{(XH^T + A^W_{ij})}{(W^T H H^T + A^W_{ij})_{ij}};$$
   2) For given $W = W^t$, $H = H^t$, $S = S^t$, update the relation matrix $P$ as:
      $$P_{ij}^{t+1} = P_{ij}^t \times \frac{(A^P_{ij})}{(A^P + A^S_{ij})};$$
   3) For given $W = W^t$, $P = P^t$, $S = S^t$, update the hidden concept coefficient matrix $H$ as:
      $$H_{ij}^{t+1} = H_{ij}^t \times \frac{(A^H_{ij} + W^T X)}{(W^T W H^T + A^H_{ij} + A^S_{ij})_{ij}};$$
   4) For given $W = W^t$, $H = H^t$, $P = P^t$, update $\ell_1$-graph weight matrix $S$ as:
      $$S^{t+1} = \min_S \left( \left\| X \right\|_F^2 + \gamma \|S\|_1 \right), \quad \text{s.t. } S \geq 0;$$
5: if $\|W^{t+1} - W^t\| < \epsilon$, $\|P^{t+1} - P^t\| < \epsilon$, $\|H^{t+1} - H^t\| < \epsilon$, and $\|S^{t+1} - S^t\| < \epsilon$ ($\epsilon$ is set as $10^{-3}$ in this work), then break.
4: **Output:** The label ranking matrix: $\hat{H} = PEQH.$
   The predict label matrix: $\hat{H}_U = \hat{H}(n_1 + 1 : n, :).$

Scale-Invariant-Feature-Transform (SIFT) [38] features over the local areas defined by the detected salient points; and iii) we perform the vector quantization on SIFT region descriptors to construct the visual vocabulary by K-Means clustering approach. In this work we generate 500 clusters, and thus the dimension of the BOW image representation vector is $r = 500$.

B. Evaluation Criteria and Baselines

Many measurements can be used to evaluate multi-label image annotation performance for labels propagated to the unlabeled images, e.g., ROC curve, precision recall curve, average precision, and so on. In this work, we use label-based and example-based criteria provided in [34]. The label-based one is selected as AUC (area under ROC curve) [39], and the example-based is accuracy.

To measure the performance of image label ranking, we utilized normalized discount cumulative gain (NDCG) [40], which is widely used in evaluating information retrieval systems. Given the ground truths of the relevant labels, and the predicted ranked labels of length $G$, the cumulated gain vector CG is defined recursively as

$$CG[i] = \begin{cases} g(i) & i = 1; \\ CG[i-1] + g(i), & \text{otherwise.} \end{cases}$$

where $g(i) = 0$, if the $i$-th predicted ranked label is not relevant, and $g(i) = G - j + 1$, if the $i$-th predicted ranked label is ranked as $j$-th relevant in the ground truths. Based on CG vector, a discounting function is introduced to reduces the relevant score as the rank increases but not steeply. Let $b$ is the discounting parameter, which is a constant. Then DCG is defined as

$$DCG[i] = \begin{cases} CG[i], & i < b; \\ DCG[i-1] + \frac{g(i)}{(\log(b)/(\log(i))}, & i \geq b \end{cases}$$

Let DCG value of the best resultant list be $DCG_{\text{best}}$, NDCG is obtained as

$$NDCG = DCG / DCG_{\text{best}}$$

For all the three datasets, our proposed solution is compared with 3 supervised algorithms and 5 semi-supervised algorithms. The supervised ones are: 1) Binary support vector machines (SVM) classifier with linear kernel (SVM_L); 2) Binary SVM classifier with polynomial kernel (SVM_P); 3) Binary SVM classifier with radial basis kernel (SVM_R). The semi-supervised ones are 4) Constrained non-negative matrix factorization (CNMF) [4]; 5) Semi-supervised learning by Sylvester equation (SMSE) [3]; 6) Multi-label Green’s function (MLGF) [41]; 7) Multi-label correlated Green’s function (MCGF) [41]; and (8) Semi-supervised multi-label (SSML) [28]. The first three algorithms are the traditional methods, which convert the multi-label image annotation problem into a set of binary classification problems for different labels. The latter five are the state-of-the-art algorithms for semi-supervised multi-label learning without image segmentation requirement. CNMF algorithm infers the labels for the unlabeled data by optimizing the consistency between image similarity and label similarity. SMSE algorithm first constructs

\(^4\)Download at [http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm](http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm)
two graphs on sample level and label level respectively, then defines a quadratic energy function on each graph, and finally obtains the labels of the unlabeled data by minimizing the combination of the two energy functions. MLGF and MCGF are the algorithms to annotate images over a graph. SSML is our previous algorithm without S updating.

C. Annotation Results

We use 3-fold cross validation to select the parameters in each algorithm on the labeled set, then the obtained parameters are utilized for performance comparison over the whole dataset. \( \alpha \) is tuned from \( 10^{-7} \) to \( 10^{-3} \), \( \beta \) is tuned from \( 10^{-7} \) to \( 10^{-3} \), \( \gamma \) is tuned from 0.05 to 0.15. We uniformly select 10 values for each of parameter range, then choose the highest one to fine tune. All the tuning parameters are shown in the tables below. For MSRC, \( \mu \) and \( \nu \) are set as 1 in SMSE approach, the number of hidden-concepts is set equally to the number of the original labels in SSML and HDIALR approaches, and the balance parameter in lasso \( \gamma \) is set as 0.07 in HDIALR approaches. For Corel5k, \( \mu \) and \( \nu \) are also set as 1 in SMSE approach. The number of hidden-concepts is set as 60, which is smaller than that of the original labels, namely 90, in SSML and HDIALR approaches, and the balance parameter in lasso \( \gamma \) is fixed to 0.12 in HDIALR approaches. For NUS-WIDE, \( \mu \) and \( \nu \)'s settings are the same as before in SMSE. The number of hidden-concepts is set as 58 in SSML and HDIALR approaches, and the balance parameter in lasso \( \gamma \) is set as 0.1 in HDIALR approaches.

Table I, III and V show the label-based performance comparison (AUC) on test data for different algorithms on MSRC, Corel5k and NUS-WIDE respectively. The second row indicates the average AUC over all the labels, and the other rows indicate the label-specific AUC's for the 5 labels randomly selected in label set. It is shown that the proposed algorithm achieves best performance on the label-based criterion.

Table II, IV and VI show the example-based performance comparison (accuracy) under different lengths of predicted labels \( G \) for all the semi-supervised algorithms over MSRC, Corel5k and NUS-WIDE respectively. On the example-based criterion, the proposed algorithm is also better than all others. For MSRC, the AUC value of our algorithm is 0.0182 higher than our previous algorithm, and 0.047 higher than SMSE, which is the best one among baselines. While the accuracy value of our algorithm is the highest when \( G \), the length of predicted label, is equal to 2, 3 and 5, but it is 0.0241 lower than linear kernel SVM, which is the best result when \( G \) is set as 7. Since the label cardinality is 2.4602 for MSRC, we need not set \( G \) larger than 5.

For Corel5k, the AUC of our algorithm is 0.0201 higher than SSML, and 0.0493 higher than SMSE, which is the best result of other baselines. The accuracy value of our algorithm is the highest when \( G = 2, 3, 5 \), but it is 0.0131 lower than polynomial kernel SVM when \( G = 7 \). Since the label cardinality is 2.8164, \( G \) is suggested to set no larger than 5.

For NUS-WIDE, the AUC value of our algorithm is 0.0205 higher than our previous version, and 0.0441 higher than SMSE, which is the best result of other baselines. The accuracy value of our algorithm is the highest when \( G = 2, 3, 5, 7 \).

D. Label Ranking Results

In MSRC, we rank the labels for the whole dataset as the ground truths according to the region size and conceptual importance to evaluate the performance on label ranking. In Corel5k, the labels of each image are ranked manually, such that the more salient labels have higher ranks. NUS-WIDE dataset is too large, so it is tedious to manually rank all the labels. Therefore, we do not experiment label ranking on it.

Table VII and VIII show the NDCGs of label ranking using different methods on the MSRC and Corel5k dataset under different parameter settings. A good ranking of labels requires not only the accurate multi-label classification, but also the reasonable estimation of relevance of the labels. For MSRC, our proposed algorithm is better than other baselines under different parameter settings. When \( b = 2, G = 3 \), the NDCGs of our algorithm is 0.0146 higher than SSML. When \( b = 2, G = 5 \), our algorithm is 0.0204 higher than SSML. When \( b = 5, G = 7 \), the improvement increases to 0.0494. For Corel5k, our proposed algorithm is also better than other baselines under different parameter settings. And the improvement over SSML ranges from 0.0021 to 0.0506.

Figure 3 demonstrates some exemplar results on label ranking of different algorithms from MSRC dataset. Our proposed solution is much more accurate than others.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed the hidden-concept driven image annotation and label ranking algorithm, which possesses three characteristics, namely, 1) label propagation is built on virtual local label representations instead of holistic image representations; 2) the inter-label similarity is well modeled with hidden concepts; and 3) the intra-label diversity is taken into consideration in the label propagation and label ranking process. The entire problem is formulated within the nonnegative data factorization framework, and an efficient multiplicative iterative procedure is proposed for optimizing the objective function with theoretically provable convergency. In our future work, we plan to further exploit new formulation and solution for images with only partially labeled images.

APPENDIX

A. Preliminaries

Before formally describing the iterative procedure for multi-label annotation, we first introduce the concept of auxiliary function, and lemma which shall be used for the algorithm deduction.

**Definition A.1:** Function \( G(Z, Z') \) is an auxiliary function for function \( F(Z) \) if the following conditions are satisfied:

\[
G(Z, Z') \geq F(Z), \quad G(Z, Z) = F(Z)
\]  
(24)

From the above the definition, we have the following lemma with proof omitted [42].

**Lemma A.1:** If \( G(Z, Z') \) is an auxiliary function , then \( F(Z) \) is non-increasing under the update

\[
Z^{t+1} = \arg \min \limits_{Z} G(Z, Z^t),
\]  
(25)

where \( t \) is the \( t \)-th iteration.
TABLE I: Performance comparison (AUC) of different algorithms for label propagation to unlabeled data on the MSRC dataset.

<table>
<thead>
<tr>
<th>SVM_L</th>
<th>SVM_R</th>
<th>CNMF</th>
<th>SMSE</th>
<th>MLGF</th>
<th>MCGF</th>
<th>SSML[28]</th>
<th>HDIALR</th>
</tr>
</thead>
<tbody>
<tr>
<td>G = 2</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
<tr>
<td>G = 3</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
<tr>
<td>G = 4</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
</tbody>
</table>

TABLE II: Performance comparison (Accuracy) of different algorithms for label propagation to unlabeled data on the MSRC dataset.

<table>
<thead>
<tr>
<th>SVM_L</th>
<th>SVM_R</th>
<th>CNMF</th>
<th>SMSE</th>
<th>MLGF</th>
<th>MCGF</th>
<th>SSML[28]</th>
<th>HDIALR</th>
</tr>
</thead>
<tbody>
<tr>
<td>G = 2</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
<tr>
<td>G = 3</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
<tr>
<td>G = 4</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
</tbody>
</table>

TABLE III: Performance comparison (AUC) of different algorithms for label propagation to unlabeled data on the Corel5k dataset.

<table>
<thead>
<tr>
<th>SVM_L</th>
<th>SVM_R</th>
<th>CNMF</th>
<th>SMSE</th>
<th>MLGF</th>
<th>MCGF</th>
<th>SSML[28]</th>
<th>HDIALR</th>
</tr>
</thead>
<tbody>
<tr>
<td>G = 2</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
<tr>
<td>G = 3</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
<tr>
<td>G = 4</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
</tbody>
</table>

TABLE IV: Performance comparison (Accuracy) of different algorithms for label propagation to unlabeled data on the Corel5k dataset.

<table>
<thead>
<tr>
<th>SVM_L</th>
<th>SVM_R</th>
<th>CNMF</th>
<th>SMSE</th>
<th>MLGF</th>
<th>MCGF</th>
<th>SSML[28]</th>
<th>HDIALR</th>
</tr>
</thead>
<tbody>
<tr>
<td>G = 2</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
<tr>
<td>G = 3</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
<tr>
<td>G = 4</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
</tbody>
</table>

TABLE V: Performance comparison (AUC) of different algorithms for label propagation to unlabeled data on the NUS-WIDE dataset.

<table>
<thead>
<tr>
<th>SVM_L</th>
<th>SVM_R</th>
<th>CNMF</th>
<th>SMSE</th>
<th>MLGF</th>
<th>MCGF</th>
<th>SSML[28]</th>
<th>HDIALR</th>
</tr>
</thead>
<tbody>
<tr>
<td>G = 2</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
<tr>
<td>G = 3</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
<tr>
<td>G = 4</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
</tbody>
</table>

TABLE VI: Performance comparison (Accuracy) of different algorithms for label propagation to unlabeled data on the NUS-WIDE dataset.

<table>
<thead>
<tr>
<th>SVM_L</th>
<th>SVM_R</th>
<th>CNMF</th>
<th>SMSE</th>
<th>MLGF</th>
<th>MCGF</th>
<th>SSML[28]</th>
<th>HDIALR</th>
</tr>
</thead>
<tbody>
<tr>
<td>G = 2</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
<tr>
<td>G = 3</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
<tr>
<td>G = 4</td>
<td>0.7027</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
<td>0.7135</td>
</tr>
</tbody>
</table>

B. Optimize W for given H, P, and S

1) Update Rule for \( W \): For a fixed \( H, P, \) and \( S \), the objective function in (7) with respect to \( W \) can be written as,

\[
F(W) = \|X - WH\|^2_F + \alpha \|PEQH_L - \tilde{H}_L\|^2_F + \beta \left( \|X\|_F^2(I - S)\right)^2_F \text{ s.t. } W \geq 0.
\]
Fig. 3: This figure illustrates the label ranking results of different algorithms on MSRC dataset. The second row shows the groundtruth, and the blue words show the incorrect predicted labels in each algorithm.

<table>
<thead>
<tr>
<th>Alg.</th>
<th>Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground Truth</td>
<td>water, boat, mountain, sky</td>
</tr>
<tr>
<td>HDIALR</td>
<td>mountain, water, sky, boat</td>
</tr>
<tr>
<td>SSML</td>
<td>mountain, water, sky, boat</td>
</tr>
<tr>
<td>CNMF</td>
<td>mountain, airplane, boat, car</td>
</tr>
<tr>
<td>SMSE</td>
<td>grass, sky, building, tree</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alg.</th>
<th>Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDIALR</td>
<td>sky, airplane, grass, building, tree</td>
</tr>
<tr>
<td>SSML</td>
<td>sky, airplane, grass, building, tree</td>
</tr>
<tr>
<td>CNMF</td>
<td>airplane, grass, boat, car</td>
</tr>
<tr>
<td>SMSE</td>
<td>grass, sky, building, tree, road</td>
</tr>
</tbody>
</table>

Table VII: NDCGs of different algorithms on the MSRC dataset.

<table>
<thead>
<tr>
<th>Alg.</th>
<th>b = 2, G = 3</th>
<th>b = 2, G = 5</th>
<th>b = 2, G = 7</th>
<th>b = 3, G = 3</th>
<th>b = 5, G = 5</th>
<th>b = 5, G = 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMSE</td>
<td>0.2956</td>
<td>0.3778</td>
<td>0.4164</td>
<td>0.4575</td>
<td>0.5402</td>
<td></td>
</tr>
<tr>
<td>CNMF</td>
<td>0.0913</td>
<td>0.1049</td>
<td>0.1368</td>
<td>0.1918</td>
<td>0.1907</td>
<td></td>
</tr>
<tr>
<td>MLGF</td>
<td>0.2999</td>
<td>0.3693</td>
<td>0.4015</td>
<td>0.4374</td>
<td>0.5086</td>
<td></td>
</tr>
<tr>
<td>MCGF</td>
<td>0.2950</td>
<td>0.3750</td>
<td>0.4120</td>
<td>0.4436</td>
<td>0.3189</td>
<td></td>
</tr>
<tr>
<td>SSML [28]</td>
<td>0.4252</td>
<td>0.4878</td>
<td>0.5288</td>
<td>0.5618</td>
<td>0.6517</td>
<td></td>
</tr>
<tr>
<td>HDIALR</td>
<td>0.4398</td>
<td>0.5082</td>
<td>0.5557</td>
<td>0.5989</td>
<td>0.7011</td>
<td></td>
</tr>
</tbody>
</table>

Table VIII: NDCGs of different algorithms on the Corel5k dataset.

<table>
<thead>
<tr>
<th>Alg.</th>
<th>b = 2, G = 3</th>
<th>b = 2, G = 5</th>
<th>b = 2, G = 7</th>
<th>b = 3, G = 3</th>
<th>b = 5, G = 5</th>
<th>b = 5, G = 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMSE</td>
<td>0.2667</td>
<td>0.3012</td>
<td>0.2846</td>
<td>0.3163</td>
<td>0.3657</td>
<td></td>
</tr>
<tr>
<td>CNMF</td>
<td>0.2070</td>
<td>0.2615</td>
<td>0.2849</td>
<td>0.3165</td>
<td>0.3660</td>
<td></td>
</tr>
<tr>
<td>MLGF</td>
<td>0.1271</td>
<td>0.1509</td>
<td>0.1685</td>
<td>0.1695</td>
<td>0.2010</td>
<td></td>
</tr>
<tr>
<td>MCGF</td>
<td>0.0508</td>
<td>0.0567</td>
<td>0.0709</td>
<td>0.0644</td>
<td>0.1113</td>
<td></td>
</tr>
<tr>
<td>SSML [28]</td>
<td>0.2208</td>
<td>0.2855</td>
<td>0.3289</td>
<td>0.3584</td>
<td>0.4538</td>
<td></td>
</tr>
<tr>
<td>HDIALR</td>
<td>0.2311</td>
<td>0.3043</td>
<td>0.3310</td>
<td>0.3867</td>
<td>0.5044</td>
<td></td>
</tr>
</tbody>
</table>

Let $\Phi_{ij}$ be the Lagrange multiplier for the constraint $W_{ij} \geq 0$, and $\Phi = [\Phi_{ij}]$, the Lagrange $L$ is:

$$
L = \|X - W\|_F^2 + \alpha \sum_i P_{EQH} L_i - \hat{H}_L \|_F^2
+ \beta \left( \frac{X}{P_{EQH}} \right) (I - S) \|_F^2 + Tr(\Phi W^T).
$$

The partial derivation of $L$ with respect to $W$ is:

$$
\frac{\partial L}{\partial W} = -2XH^T + 2WHH^T + 2A_1^W - 2A_2^W + 2A_3^W + \Phi,
$$

where $A_1^W$, $A_2^W$, $A_3^W$ is defined in (11) (12) and (13).

Using the KKT condition [43] $\Phi_{ij} W_{ij} = 0$, from (27), we can obtain the following update rule:

$$
W_{ij} = W_{ij} \times \frac{(XH^T + A_1^W)_{ij}}{(WHH^T + A_1^W + A_2^W + A_3^W)_{ij}}.
$$

2) Convergence of Update Rule for $W$: We denote $F_{ij}$ as the part of $F(W)$ relevant to $W_{ij}$, and we have,

$$
F_{ij}'(W) = (-2XH^T + 2WHH^T + 2A_1^W - 2A_2^W + 2A_3^W)_{ij}
$$

$$
F_{ij}''(W) = (2HH^T)_{ij}
+ 2(\alpha e e^T)_{ij} (H_L H_L^T) \cdot [(PE)^T(PE)]_{ij}
+ 2(\beta e e^T)_{ij} (H(I - S)(I - S)^T H^T) \cdot [(PE)^T(PE)]_{ij}.
$$

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

$$
G(W_{ij}, W_{ij}') = F_{ij}(W_{ij}) + F_{ij}'(W_{ij} - W_{ij}') + \frac{3}{2}(W_{ij}^T H H^T + 2A_1^W + 4A_2^W + 2A_3^W)_{ij} (W_{ij} - W_{ij}')^2
$$

where

$$
A_1^W = \alpha e e^T W^T (H_L H_L^T) \cdot [(PE)^T(PE)];
$$

$$
A_2^W = \beta e e^T W^T (H(I - S)(I - S)^T H^T) \cdot [(PE)^T(PE)].
$$

Lemma A.2: Eqn. (29) is an auxiliary function for $F_{ij}$, namely the part of $F(W)$ relevant to $W_{ij}$.

Proof: Since $G(W_{ij}, W_{ij}') = F_{ij}(W_{ij})$, we need only to show that $G(W_{ij}, W_{ij}') > F_{ij}(W_{ij})$.

First, we get the Taylor series expansion of $F_{ij}$ as,

$$
F_{ij}(W_{ij}) = F_{ij}(W_{ij}') + F_{ij}'(W_{ij} - W_{ij}') + F_{ij}''(W_{ij} - W_{ij}')^2.
$$
Then, since
\[(W^T H H^T)_{ij} = \sum_k W^T_k (H H^T)_{kj} \geq W^T_k (H H^T)_{jj},\]
\[(\mathbf{w} e^T W^T (H H^T)_{ij} \cdot ((P E^T (P E))_{ij}) \geq (\mathbf{w} e^T W^T (H H^T)_{ij} \cdot ((P E^T (P E))_{jj}),\]
and
\[(\mathbf{w} e^T W^T (H (I-S) (I-S)^T T H^T) \cdot ((P E^T P E))_{ij}) \geq (\mathbf{w} e^T W^T (H (I-S) (I-S)^T H^T) \cdot ((P E^T P E))_{jj},\]
Thus, \(G(W_{ij}, W^T_{ij}) \geq F_{ij}(W_{ij}) \text{ holds.}\)

\[\partial G(W_{ij}, W^T_{ij}) = 0, \text{ then we can obtain the iterative update rule for } W.\]

C. Optimize P for given W, H, and S

1) Update Rule for P: The objective function in Eqn. (7) with respect to \(P\) for given \(W, H,\) and \(S\) can be written as,
\[F(P) = \alpha \|PEQH_L - H_L\|_F^2 + \beta \|PEQH(I-S)\|_F^2,\]
\[s.t. \quad P \succeq 0.\]

Let \(\Psi_{ij}\) be the Lagrange multiplier for constraint \(P_{ij} \succeq 0,\) and \(\Psi = [\Psi_{ij}]\), the Lagrange \(L\) is,
\[L = \alpha \|PEQH_L - H_L\|_F^2 + \beta \|PEQH(I-S)\|_F^2 + \text{Tr}(\Psi P^T).\]

The partial derivation of \(L\) with respect to \(P\) is,
\[\frac{\partial L}{\partial P} = 2 A^P F - 2 A^F P + 2 A^S + \Psi,\]
where \(A^P, A^F, A^S\) is defined in (14) (15) and (16).

Using the KKT condition \(\Psi_{ij} P_{ij} = 0,\) from (32), we can get the following update rule:
\[P_{ij} = P_{ij} \times \frac{(A^P_{ij})}{(A^F_{ij} + A^S_{ij})}.\]

Convergence analysis of update rule for \(P\) is similar to B2.

D. Optimize H for given W, P, and S

1) Update Rule for H: For the fixed \(W, P,\) and \(S,\) the objective function in (7) with respect to \(H\) can be written as,
\[F(H) = \|X - WH\|_F^2 + \alpha \|PEQH_L - H_L\|_F^2 + \beta \|PEQH(I-S)\|_F^2,\]
\[s.t. \quad H \succeq 0.\]

Let \(\Theta_{ij}\) be the Lagrange multiplier for constraint \(H_{ij} \succeq 0,\) and \(\Theta = [\Theta_{ij}]\), the Lagrange \(L\) is,
\[L = \|X - WH\|_F^2 + \alpha \|PEQH_L - H_L\|_F^2 + \beta \|PEQH(I-S)\|_F^2 + \text{Tr}(\Phi H^T).\]

The partial derivation of \(L\) with respect to \(H\) is,
\[\frac{\partial L}{\partial H} = -2 W^T X + 2 W^T W H + 2 A^H - 2 A^F + 2 A^S + \Theta\]
where \(A^H, A^F, A^S\) is defined in (17) (18) and (19).

Using the KKT condition [43] \(\Theta_{ij} H_{ij} = 0,\) from (36), we can get the following update rule:
\[H_{ij} = H_{ij} \times \frac{(A^H_{ij} + W^T X)_{ij}}{(W^T W H + A^F_{ij} + A^S_{ij})_{ij}}.\]

Convergence analysis of update rule for \(H\) is similar to B2.

E. Optimize S for given W, H, P

Given \(W, H, P,\) (7) can be deduced to
\[S = \min_S \left\| \left( \frac{X}{PEQH} \right) S - \left( \frac{X}{PEQH} \right) \right\|_F^2 + \gamma \|S\|_1, \quad \text{s.t. } S \succeq 0.\]

This is a standard Lasso problem [44], and can be easily solved by off-the-shelf algorithm, e.g., SLEP [45].

ACKNOWLEDGEMENT

This work is partially support by project grant NRF2007IDM-IDM002-069 on “Life Spaces” from the IDM Project Office, Media Development Authority of Singapore.

REFERENCES


D. L. Donoho, “For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparest solution,” *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–826, 2006.


“http://www.flickr.com/”


