Joint Image Clustering and Labeling by Matrix Factorization

Seunghoon Hong, Jonghyun Choi, Jan Feyereisl, Bohyung Han, Larry S. Davis

Abstract—We propose a novel algorithm to cluster and annotate a set of input images jointly, where the images are clustered into several discriminative groups and each group is identified with representative labels automatically. For these purposes, each input image is first represented by a distribution of candidate labels based on its similarity to images in a labeled reference image database. A set of these label-based representations are then refined collectively through a non-negative matrix factorization with sparsity and orthogonality constraints; the refined representations are employed to cluster and annotate the input images jointly. The proposed approach demonstrates performance improvements in image clustering over existing techniques, and illustrates competitive image labeling accuracy in both quantitative and qualitative evaluation. In addition, we extend our joint clustering and labeling framework to solving the weakly-supervised image classification problem and obtain promising results.

Index Terms—Image clustering, image labeling, label feature, non-negative matrix factorization with sparsity and orthogonality constraints (SO-NMF)

1 Introduction

Image classification and object recognition have been studied intensively over the last decade. Initially, researchers employed supervised learning, where a training dataset is given and each test image is categorized into the predefined classes. However, this approach suffers from several critical challenges including difficulty in handling a large number of categories, requirement of a predefined set of target classes of interest and extensive labeling efforts for constructing training data. To overcome such limitations, semi-supervised or unsupervised learning methods have drawn attention recently. Although semi-supervised or unsupervised algorithms have been successful in categorizing images, they still suffer from critical inherent limitations; semi-supervised learning techniques require a set of classes for categorization in advance just as ordinary supervised learning algorithms do, and unsupervised methods such as clustering cannot provide human interpretable label information for each image or category.

We introduce a data-driven technique to discover categories of query images, where candidate labels are identified automatically based on context within the images with little human supervision. We propose a novel algorithm to cluster and annotate a set of query images simultaneously; *i.e.*, the images are clustered into discriminative groups and each group is assigned representative labels automatically. The overview of our joint image clustering and annotation framework is illustrated in Figure 1.

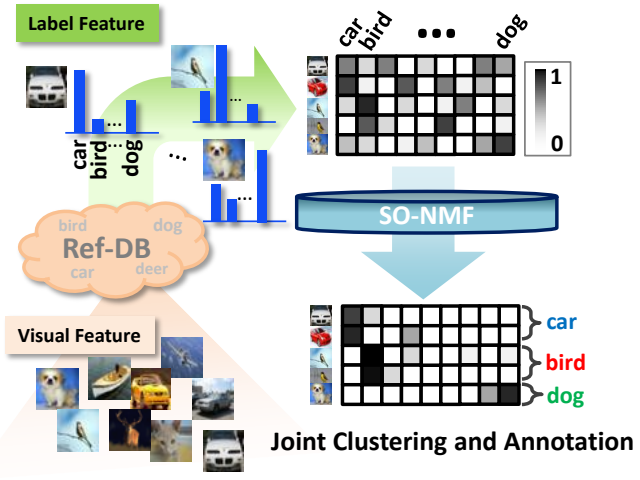


Fig. 1. Overview of our approach. First, query images are represented by noun-typed label features with the help of a reference database (Ref-DB). Based on the label feature representation, we perform the non-negative matrix factorization with sparsity and orthogonality constraints (SO-NMF), which enables us to produce discriminative clusters and label the query images jointly.

We assume that there exists a reference image database (Ref-DB), which contains labeled images and is much larger than the query image set. Ref-DB can be obtained from a public large-scale image data repository or constructed by utilizing external image retrieval systems such as image search engines. Thus, the database can be obtained with little human efforts but it will potentially have inconsistent or missing information; thus, it would be inadvisable to use it with ordinary supervised learning methods.

Given the Ref-DB, we first represent each query image

- *S. Hong, J. Feyereisl and B. Han are with the Department of Computer Science and Engineering, POSTECH, Korea. E-mail: {maga33, feyereisl, bhhan}@postech.ac.kr*
- *J. Choi and L. S. Davis are with University of Maryland Institute of Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD, USA. E-mail: {jhchoi, lsd}@umiacs.umd.edu*

with a label probability vector constructed by weighted label frequencies of its k-nearest neighbors in the Ref-DB. Then, a collection of label features—features based on the label probability vectors—is obtained by a clustering technique using non-negative matrix factorization with sparsity and orthogonality constraints (SO-NMF). Through these steps, our method captures discriminative and representative concepts of the set of query images to label them. In other words, tentative labels for each query image are extracted independently by k-nearest neighbor search in the Ref-DB, and the labels are refined jointly based on the context of the entire set of query images. As a result, each image is eventually represented with only a small subset of label candidates, which are chosen adaptively from the large label set in Ref-DB, and our algorithm annotates query images by the direct use of the refined label candidates. The main characteristics of our algorithm are as follows:

- Our image representation based on label candidates supports the discovery of high-level image content information.
- We propose a novel clustering algorithm, SO-NMF, which is well-suited for label features and captures the discriminative and representative labels of an image.
- We extend our algorithm to a weakly-supervised image classification task, where a set of candidate labels relevant for query images are identified automatically.
- We validate the effectiveness of our approach on two challenging datasets through image clustering performance and image annotation quality.

The rest of this paper is organized as follows. We first discuss existing techniques related to image clustering and labeling in Section 2. Section 3 describes how to construct the label feature using Ref-DB, and Section 4 presents our image clustering and labeling framework based on SO-NMF. Performance of our algorithm is discussed in Section 5.

2 Related Work

This section discusses prior research closely related to our work, which covers image representation, annotation, and clustering. A more comprehensive survey of collective image category discovery can be found in [1].

2.1 Image Representation

Image representation is one of the most important issues in image clustering and labeling. Various visual features, *i.e.*, SIFT [2], GIST [3], HOG [4] and LBP [5], have been used to encode visual information. Recently, learned features using convolutional deep network have shown remarkable performance improvement over hand crafted features [6]. Although such low-level visual descriptors capture exemplar specific characteristics well, they often suffer from significant appearance variations of objects or scenes.

To overcome such limitations, high-level image representation techniques have been proposed. Visual attributes, or *attributes* in short, are widely used in the recent literature to describe high-level information based on visual properties of images [7], [8]. Category membership information is often employed for the construction of image descriptors. For example, images are often represented with a set of relevance scores with respect to known categories [9]–[12], where the scores are obtained from classifier responses. Another example is the Object Bank (OB) representation [13], which is based on a vector of detector responses for a set of objects. The object-graph feature [14], [15] describes an object with a vector of posterior probabilities of surrounding objects.

These approaches, however, require additional annotations to describe visual properties or assume well-trained classifiers for many categories that involves solving a large-scale learning problem. On the other hand, [16] adopted a nearest neighbor based method, where images are represented by their relevance to categories obtained from weighted nearest neighbors in a large-scale image repository. However, their representations may not be accurate enough without proper refinement due to potentially noisy annotations of the images in the repository and errors in nearest neighbor search.

2.2 Automatic Image Annotation

The objective of image annotation is to provide images with relevant labels based on their visual contents. We group the image annotation techniques into three categories: generative methods [9], [17]–[22], discriminative methods [23]–[25] and nearest neighbor methods [16], [26]–[29].

Generative methods often employ topic models to represent image-label relationships. Variants of Latent Dirichlet Allocation (LDA) are integrated to relate labels and images in [9], [18]. Probabilistic Latent Semantic Analysis (pLSA) [19] and constrained Nonnegative Matrix Factorization (NMF) [20] have been used to provide images with multiple labels. Barnard *et al.* [17] model the joint distribution of words and blobs in an image, and perform annotation by finding correspondences between them. Missing tags of images are filled in by a probabilistic model via collaborative filtering [21], which exploits visual features as well as incomplete tags manually annotated. Similarly, [22] proposes an algorithm to complete missing or incorrect tags of images by applying an image-tag matrix completion algorithm. In addition, Li *et al.* [30], [31] use a constrained matrix factorization model to address multi-label image classification. Zhu *et al.* [32] use matrix factorization with a low-rank constraint to address the image tag refinement problem.

On the other hand, discriminative models pose annotation as a classification problem; they learn an individual classifier for each label based on low-level visual features. Various learning techniques have been employed including SVMs [23], boosting classifiers [25], and multiple instance learning [33]. Both generative and discriminative methods assume the availability of a clean and large scale training dataset, which requires extensive human effort for

construction. Also, they are not scalable in general since they are typically designed to annotate images only with a predefined set of labels.

Nearest neighbor based methods identify visually similar images in a database to transfer corresponding labels to unknown images or regions of images, as in [26], [34]. In practice, k-nearest neighbor (k-NN) methods with suitable features and distance functions achieve competitive performance on many visual recognition tasks [35]–[37]. However, k-NN algorithms tend to overfit to local distributions of samples on a given metric. To address this issue, various techniques such as metric learning [27], [28], weighted k-NN [16] and classifier construction [29] have been studied. In addition, Berg *et al.* [29] propose SVM-KNN, which learns locally discriminative SVM classifiers based on k-nearest neighbors. However, the selection of optimal k is not straightforward, and the selected training examples may be biased and lead to poor decision boundaries. This can cause their performance to suffer compared to ordinary SVMs in some settings, as reported in [38], [39].

2.3 Unsupervised Image Clustering

Unsupervised image clustering partitions unlabeled images into disjoint clusters based on a similarity measure [1], [14], [15], [40]–[42]. There are several approaches to this problem: topic modeling [41]–[43], graph-based methods [40], and ensemble methods [44].

Clustering of images or image regions based on pLSA or LDA is proposed in [42], [43]. Liu and Chen [41] exploit correspondences between images to model object configurations and relationship among images based on a pLSA-like formulation. In [40], images are clustered by analyzing the relationship between visual features extracted from images using link analysis, similar to Google's PageRank algorithm. The identity of unknown objects are discovered with the help of surrounding known class objects in an unsupervised manner [14], and this idea is further extended in [15] to curriculum learning, which attempts to learn easy things first.

Semi-supervised learning—a framework utilizing an incompletely labeled database—has been explored widely in recent years. Shrivastava *et al.* [45] proposed a constrained semi-supervised scene category recognition method, where inter-class relationships using semantic attributes are given by a category ontology. Dai *et al.* [44] construct multiple weak training sets by subsampling query images and learn an ensemble classifier based on the training sets. They claim that an ensemble classifier trained on partially correct data performs comparably to the one with perfect training data. Noisy annotations are exploited to improve labeling performance in a large-scale dataset in [46]. Training data can be augmented by borrowing examples from related categories [47]–[49] or incorporating examples identified with surrounding text [50].

3 Image Representation by Words

One of the critical drawbacks of unsupervised image clustering techniques is that they do not construct human interpretable labels or representations for identified clusters. To overcome this limitation, we represent images with a set of relevant labels. We call the semantic mid-level features *label features*.

3.1 Label Feature Using Reference Database

We refer to large-scale image databases to find the association between labels and images. Although the automatic extraction of accurate label information from an image is extremely difficult, Torralba *et al.* [26] have shown that a simple k-nearest neighbor method with a very large number of images yields reasonable performance in classification. Also, Deng *et al.* [51] have empirically shown that there exist some meaningful correlations between the visual and the semantic spaces. Semantically related labels are expected to be obtained by finding visually similar images in a large image repository, which supports our idea of representing images with the label features extracted from a large-scale reference database based on visual features.

We employ the Tiny Images database [26] and ImageNet [52] as Ref-DBs, but any large image repository containing label information such as image search engines could be integrated naturally in our framework.

Our label features are different from attributes. A label feature represents a distribution of class labels while an attribute denotes a nameable visual property that is shared across classes [8], [53]. Thus, using attribute-based description, it is not straightforward to obtain the label information directly.

3.2 Label Feature Construction

Our algorithm represents each query image by a distribution of candidate labels from Ref-DB. We show that the representation is effective for modeling high-level concepts in images and improving image clustering and annotation performance.

Denote the Ref-DB by $\mathcal{R} = \{(\mathbf{z}_j, \mathbf{l}_j) \in \mathbb{R}^t \times \mathbb{R}^d, j = 1, \ldots, m\}$, which contains m images and their associated labels. Each image $\mathbf{z}_j$ is represented by t-dimensional vector obtained from visual feature descriptors such as GIST, SIFT, and LBP. A binary vector $\mathbf{l}_j$ specifies the label assignments of $\mathbf{z}_j$ to d label candidates. Our goal in this step is to represent each image $\mathbf{x}_i$ ($i = 1, \ldots, n$, $n \ll m$) from a query image dataset using a label probability (frequency) vector, which will be used for clustering and labeling afterwards.

To obtain the label probability vector of $\mathbf{x}_i$, we first determine the k-nearest neighbors (k-NNs) of $\mathbf{x}_i$ from $\mathcal{R}$ in the visual feature space. Then, we retrieve the label vectors of the k-NNs, denoted by $\mathcal{L}_i = \{\mathbf{l}_{i,1}, \ldots, \mathbf{l}_{i,k}\}$, where $\mathbf{l}_{i,l} \in \mathbb{R}^d$ ($l = 1, \ldots, k$) is the label vector of the l^{th} nearest neighbor. A label probability vector of $\mathbf{x}_i$ is then constructed as a weighted sum of the elements in $\mathcal{L}_i$, which creates a label feature for $\mathbf{x}_i$. The weights, denoted by $\omega_{i,l}$, are normalized based on the distances to individual nearest neighbors. This procedure is summarized in Algorithm 1. Since different kinds of visual features capture

Algorithm 1: Label feature construction

Input: Input images $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$
Output: Label frequency matrix $\mathbf{V} \in \mathbb{R}^{n \times d}$

foreach $\mathbf{x}_i$ **do**

Find k-NNs and their corresponding labels:

$$\mathcal{N}(\mathbf{x}_i) = \{(\mathbf{z}_{i,1}, \mathbf{l}_{i,1}) \ldots, (\mathbf{z}_{i,k}, \mathbf{l}_{i,k})\}$$

Compute the weight for each $l \in \{1, \cdots, k\}$:

$$\omega_{i,l} = \frac{\exp\left(-\alpha||\mathbf{x}_i - \mathbf{z}_{i,l}||_2\right)}{\sum_{a=1}^{k} \exp\left(-\alpha||\mathbf{x}_i - \mathbf{z}_{i,a}||_2\right)}.$$

Compute the label feature for $\mathbf{x}_i$:

$$\mathbf{v}_i = \sum_{l=1}^{k} \omega_{i,l} \frac{\mathbf{l}_{i,l}}{||\mathbf{l}_{i,l}||_1}.$$

end

Collect all label feature vectors and construct $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]^{\mathrm{T}}$.

different characteristics, we construct these weighted label probability vectors for each of the multiple features. Then, we sum them up and obtain the final label feature $\mathbf{v}_i$ for $\mathbf{x}_i$ after normalization; the collection of the label features is represented as a label probability (frequency) matrix $\mathbf{V}$.

Label feature representations for several images are shown in Figure 2. An image is typically associated with a few highly weighted labels but the label distribution is often noisy and inconsistent. Our goal is to refine these label probability vectors through the collaboration of multiple images by a machine learning technique—non-negative matrix factorization.

4 Clustering and Labeling by SO-NMF

Our goal is to divide the query images into clusters and assign labels to each cluster. That is, we will provide each cluster with a few representative labels and make the labels discriminative across clusters. This problem is formulated as non-negative matrix factorization with sparsity and orthogonality constraints, which is well-suited for clustering images with our label feature representations. Some recent work [54]–[56] also employs matrix factorization for classification or labeling, but we claim that the benefit of the low rank structure is highlighted when our label feature representation is combined with the two constraints. This section describes our clustering and labeling technique in the matrix factorization framework.

4.1 Non-Negative Matrix Factorization (NMF)

Non-Negative Matrix Factorization (NMF) [57] reconstructs a target matrix based on low-rank and non-negativity properties. It decomposes a non-negative data matrix $\mathbf{V} \in \mathbb{R}_+^{n \times d}$ into a non-negative basis matrix $\mathbf{H} \in \mathbb{R}_+^{d \times c}$ and a coefficient matrix $\mathbf{W} \in \mathbb{R}_+^{n \times c}$, where c denotes the number of bases and $_+$ means non-negativity. Mathematically, NMF solves the following optimization problem:

$$\begin{aligned} \underset{\mathbf{W},\mathbf{H}}{\arg\min} \quad & \|\mathbf{V} - \mathbf{W}\mathbf{H}^{\mathrm{T}}\| \\ \text{s.t.} \quad & \mathbf{W}, \mathbf{H} \geq 0 \end{aligned} \tag{1}$$

where $\|\cdot\|$ denotes a matrix norm, and the constraint on $\mathbf{W}$ and $\mathbf{H}$ enforces the non-negativity of matrix elements. Typically, the ℓ_2-norm is used as a quality measure of the decomposition, but Kullback-Leibler (KL) divergence is an alternative. In our problem, $\mathbf{V} \in \mathbb{R}_+^{n \times d}$ contains n query images in d dimensional label feature space. Each row of $\mathbf{W} \in \mathbb{R}_+^{n \times c}$ describes data membership in each of c clusters and each column of matrix $\mathbf{H} \in \mathbb{R}_+^{d \times c}$ represents a prototype for each cluster.

To solve the optimization problem, we adopt the iterative multiplicative update method proposed in [57] due to its computational efficiency compared to traditional gradient descent algorithms. They show that new values of $\mathbf{W}$ and $\mathbf{H}$ can be found by multiplying the basis or the coefficient matrix by some factor. By rescaling the matrix $\mathbf{W}$ and $\mathbf{H}$ and replacing the step size of the standard gradient descent method with a suitable value, one can derive the following multiplicative update rules:

$$\begin{aligned} \mathbf{H} &\leftarrow \mathbf{H} \odot \left(\frac{\mathbf{V}^{\mathrm{T}}\mathbf{W}}{\mathbf{H}\mathbf{W}^{\mathrm{T}}\mathbf{W}}\right) \quad \text{and} \\ \mathbf{W} &\leftarrow \mathbf{W} \odot \left(\frac{\mathbf{V}\mathbf{H}}{\mathbf{W}\mathbf{H}^{\mathrm{T}}\mathbf{H}}\right). \end{aligned} \tag{2}$$

Using these rules, the objective function in Eq. (1) with ℓ_2-norm can be solved iteratively and the solution converges to a local minimum [57].

We enforce additional constraints on either $\mathbf{W}$, $\mathbf{H}$ or both to achieve our goal. By deriving new multiplicative update rules according to the additional constraints, sparsity and orthogonality, we propose a sparse and orthogonal NMF, called SO-NMF. The details of our algorithm will be presented in the rest of this section.

4.2 Constraints: Sparsity and Orthogonality

In our formulation for matrix factorization, we impose a sparsity requirement on the basis matrix $\mathbf{H}$, and orthogonality constraints on both the basis matrix $\mathbf{H}$ and the coefficient matrix $\mathbf{W}$. Sparsity and orthogonality constraints on the basis matrix ensures that each cluster is represented by only a small number of labels, which are exclusive across clusters. On the other hand, orthogonality constraint on the coefficient matrix encourages each image to be associated with the unique basis vector(s). The combination of the two constraints enforces each cluster to involve only a few exclusive labels, enabling us to describe images based on a few labels while improving the discriminativeness of the clusters themselves. In this subsection, we explain how each constraint is added to the original NMF objective function, Eq. (1), and how new multiplicative update rules are derived given the constraints.

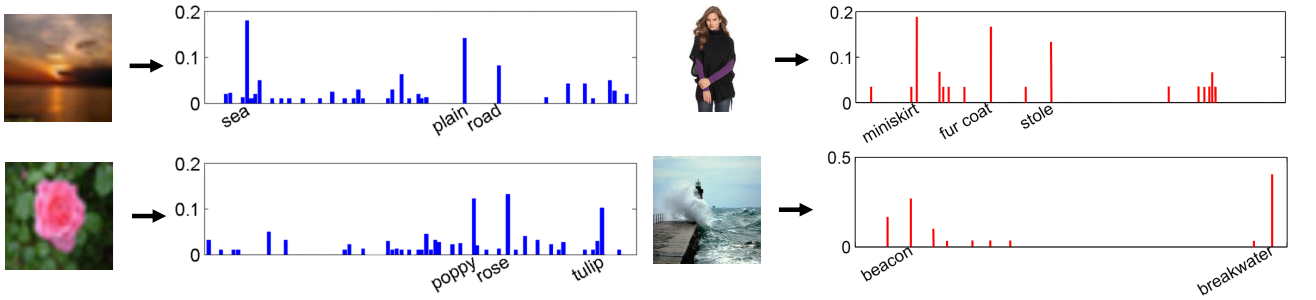


Fig. 2. Examples of label feature representations of images in (left) CIFAR-100 and (right) ImageNet datasets. Labels for highly related classes are annotated in the histograms.

4.2.1 Sparsity

We are interested in obtaining sparse representations of the columns in the basis matrix $\mathbf{H} = [\mathbf{h}_1 \dots \mathbf{h}_c]$ as we desire only a few labels to be associated with each cluster. The level of sparsity in columns of $\mathbf{H}$ is controlled by the sparsity measure proposed in [58],

$$s(\mathbf{h}_i) = \frac{1}{\sqrt{d}-1}\left(\sqrt{d} - \frac{\|\mathbf{h}_i\|_1}{\|\mathbf{h}_i\|_2}\right), \tag{3}$$

where d is the dimensionality of $\mathbf{h}_i$, $i = 1, \dots, c$ and $\|\cdot\|_p$ denotes the ℓ_p-norm. This measure exploits the relationship between the ℓ_1 and ℓ_2 norms, and is bounded by $0 \leq s(\mathbf{h}_i) \leq 1$, where $s(\mathbf{h}_i) = 0$ denotes minimal sparsity while $s(\mathbf{h}_i) = 1$ means a maximally sparse vector with a single non-zero element. The detailed optimization technique incorporating this sparsity constraint is described in [58]. The intuitive interpretation of sparsity is related to the composition of basis vectors; maximizing the sparsity of a column of $\mathbf{H}$ is interpreted as minimizing the number of labels that represent a cluster basis.

4.2.2 Orthogonality

Although each cluster is represented with a few labels by imposing a sparsity constraint on $\mathbf{H}$, different clusters may still share some labels. Therefore, in addition to the sparsity constraint, we also impose an orthogonality constraint on the columns of the basis matrix $\mathbf{H}$ and coefficient matrix $\mathbf{W}$. Orthogonality of the basis matrix $\mathbf{H}$ enforces clusters to have mutually exclusive labels, which is desirable to obtain discriminative labels for clusters. On the other hand, orthogonality of the coefficient matrix $\mathbf{W}$ encourages each image to be associated with unique basis vector(s), which facilities straightforward interpretation of clustering results and clustering performance improvement [59].

The objective function incorporating these orthogonality constraints is given by

$$\begin{aligned} &\underset{\mathbf{W},\mathbf{H}\geq 0}{\arg\min} \ \|\mathbf{V} - \mathbf{W}\mathbf{H}^{\mathrm{T}}\| \\ &\text{s.t.} \ \ \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I} \ \ \text{and} \ \ \mathbf{H}^{\mathrm{T}}\mathbf{H} = \mathbf{I} \end{aligned} \tag{4}$$

where $\mathbf{I}$ is the identity matrix. Note that, however, this formulation is problematic because the *orthonormal* constraints on both $\mathbf{W}$ and $\mathbf{H}$ restrict the scale of both factors and increase reconstruction error consequently. The matrix factorization problem with the bi-orthogonality constraint as in Eq. (4) has been discussed in [59], and the detailed optimization procedure with sparsity and orthogonality constraints will be presented next.

4.3 Optimization

Previous work using NMF only involves either sparsity or orthogonality, but not both constraints together. In our formulation, while columns of the basis matrix are sparse due to the constraint in Eq. (3), the orthogonality constraints result in orthonormal coefficient and basis matrices. Use of naïve multiplicative update rules derived from ordinary matrix factorization incorporating all the constraints may not induce a reliable solution due to scaling issue in $\mathbf{H}$ and $\mathbf{W}$ as discussed above. To avoid this problem, we introduce a diagonal scaling matrix $\mathbf{S} \in \mathbb{R}^{c\times c}$, similarly to [59], and preserve the property of representing each image with a sparse and exclusive set of labels. Our problem is thus formulated as a non-negative matrix tri-factorization framework. The new objective function becomes

$$\begin{aligned} &\underset{\mathbf{W},\mathbf{S},\mathbf{H}\geq 0}{\arg\min} \ \|\mathbf{V} - \mathbf{W}\mathbf{S}\mathbf{H}^{\mathrm{T}}\|_2^2 \\ &\text{s.t.} \ \ \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}, \ \ \mathbf{H}^{\mathrm{T}}\mathbf{H} = \mathbf{I} \ \ \text{and} \ \ s(\mathbf{h}_i) = \xi, \end{aligned} \tag{5}$$

where $\|\cdot\|_2$ denotes ℓ_2-norm. The objective function in Eq. (5) is to be minimized under the two orthogonality constraints, $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$ and $\mathbf{H}^{\mathrm{T}}\mathbf{H} = \mathbf{I}$, and the sparsity constraint $s(\mathbf{h}_i) = \xi$, where ξ is the desired sparsity level.

To optimize the new objective function, we follow the standard constrained optimization procedure in [59]. Ignoring sparsity constraints on $\mathbf{H}$, the main objective function in Eq. (5) is rewritten as

$$\begin{aligned} &f(\mathbf{W},\mathbf{S},\mathbf{H}) \\ &= \|\mathbf{V} - \mathbf{W}\mathbf{S}\mathbf{H}^{\mathrm{T}}\|_2^2 + \\ &\quad \mathrm{Tr}[\lambda_{\mathbf{W}}(\mathbf{W}^T\mathbf{W} - \mathbf{I})] + \mathrm{Tr}[\lambda_{\mathbf{H}}(\mathbf{H}^T\mathbf{H} - \mathbf{I})] \\ &= \mathrm{Tr}[\mathbf{V}^{\mathrm{T}}\mathbf{V} - 2\mathbf{H}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\mathbf{W}\mathbf{S} + \mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{S}\mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{S}^{\mathrm{T}} + \\ &\quad \lambda_{\mathbf{W}}(\mathbf{W}^T\mathbf{W} - \mathbf{I}) + \lambda_{\mathbf{H}}(\mathbf{H}^T\mathbf{H} - \mathbf{I})], \end{aligned} \tag{6}$$

where $\lambda_{\mathbf{W}}$ and $\lambda_{\mathbf{H}}$ are $c \times c$ symmetric matrices, which denote the Lagrangian multipliers for orthogonality constraints on $\mathbf{W}$ and $\mathbf{H}$, respectively. Given the objective function in Eq. (6), our goal is to solve Eq. (5) with respect

to factors $\mathbf{W}, \mathbf{S}$ and $\mathbf{H}$ to satisfy the following optimality condition:

$$\bigtriangledown_{\mathbf{W},\mathbf{S},\mathbf{H},\lambda_{\mathbf{W}},\lambda_{\mathbf{H}}} f(\mathbf{W},\mathbf{S},\mathbf{H}) = 0, \tag{7}$$

where the partial derivatives of Eq. (7) with respect to individual factors are given respectively by

$$\bigtriangledown_{\mathbf{W}} f = 2(\mathbf{WSH}^{\mathrm{T}}\mathbf{HS}^{\mathrm{T}} - \mathbf{VHS}^{\mathrm{T}} + \mathbf{W}\lambda_{\mathbf{W}}), \tag{8}$$

$$\bigtriangledown_{\mathbf{S}} f = 2(\mathbf{W}^{\mathrm{T}}\mathbf{WSH}^{\mathrm{T}}\mathbf{H} - \mathbf{W}^{\mathrm{T}}\mathbf{VH}), \tag{9}$$

$$\bigtriangledown_{\mathbf{H}} f = 2(\mathbf{HS}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{WS} - \mathbf{V}^{\mathrm{T}}\mathbf{WS} + \mathbf{H}\lambda_{\mathbf{H}}), \tag{10}$$

$$\bigtriangledown_{\lambda_{\mathbf{W}}} f = \mathbf{W}^{\mathrm{T}}\mathbf{W} - \mathbf{I}, \tag{11}$$

$$\bigtriangledown_{\lambda_{\mathbf{H}}} f = \mathbf{H}^{\mathrm{T}}\mathbf{H} - \mathbf{I}. \tag{12}$$

The Lagrangian multipliers $\lambda_{\mathbf{W}}$ and $\lambda_{\mathbf{H}}$ in Eq. (6) are determined automatically to satisfy the optimality condition in Eq. (7). Given the optimality conditions on $\mathbf{W}$, $\bigtriangledown_{\mathbf{W}} f = 0$ and $\bigtriangledown_{\lambda_{\mathbf{W}}} f = 0$, $\lambda_{\mathbf{W}}$ is obtained by

$$\lambda_{\mathbf{W}} = \mathbf{W}^{\mathrm{T}}\mathbf{VHS}^{\mathrm{T}} - \mathbf{SH}^{\mathrm{T}}\mathbf{HS}^{\mathrm{T}}. \tag{13}$$

Similarly, $\lambda_{\mathbf{H}}$ is obtained by optimality conditions on $\mathbf{H}$, $\bigtriangledown_{\mathbf{H}} f = 0$ and $\bigtriangledown_{\lambda_{\mathbf{H}}} f = 0$, as

$$\lambda_{\mathbf{H}} = \mathbf{H}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\mathbf{WS} - \mathbf{S}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{WS}. \tag{14}$$

To solve Eq. (5) for $\mathbf{W}$, we define an auxiliary function, $g(\mathbf{W}, \mathbf{W}')$, satisfying the following conditions:

$$g(\mathbf{W}, \mathbf{W}') \geq f(\mathbf{W}) \text{ and } g(\mathbf{W}, \mathbf{W}) = f(\mathbf{W}),$$

where $f(\mathbf{W})$ denotes the objective function with respect to $\mathbf{W}$. We can optimize $f(\mathbf{W})$ by iteratively finding

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} g(\mathbf{W}, \mathbf{W}') \tag{15}$$

because $f(\mathbf{W}) = g(\mathbf{W}, \mathbf{W}) \geq g(\mathbf{W}^*, \mathbf{W}) \geq f(\mathbf{W}^*)$. The following auxiliary function presented in [59] is also used in our problem:

$$g(\mathbf{W}, \mathbf{W}') = -\sum_{ik} 2(\mathbf{VHS}^{\mathrm{T}})_{ik}\mathbf{W}'_{ik}(1 + \log\frac{\mathbf{W}_{ik}}{\mathbf{W}'_{ik}}) + \sum_{ik} \frac{\mathbf{W}'(\mathbf{SH}^{\mathrm{T}}\mathbf{HS}^{\mathrm{T}} + \lambda_{\mathbf{W}})\mathbf{W}^2_{ik}}{\mathbf{W}'_{ik}}. \tag{16}$$

By setting element-wise derivatives of $g(\mathbf{W}, \mathbf{W}')$ to zeros, we minimize the coefficient matrix $\mathbf{W}$ as

$$\mathbf{W}_{ik} = \mathbf{W}'_{ik}\left(\frac{(\mathbf{VHS}^{\mathrm{T}})_{ik}}{(\mathbf{WSH}^{\mathrm{T}}\mathbf{HS}^{\mathrm{T}} + \mathbf{W}\lambda_{\mathbf{W}})_{ik}}\right)^{\frac{1}{2}}. \tag{17}$$

Plugging Eq. (13) into Eq. (17), the multiplicative update rule of $\mathbf{W}$ is obtained as

$$\mathbf{W} \leftarrow \mathbf{W} \odot \sqrt{\frac{\mathbf{VHS}^{\mathrm{T}}}{\mathbf{WW}^{\mathrm{T}}\mathbf{VHS}^{\mathrm{T}}}}. \tag{18}$$

We refer to [59] for more detailed derivation.

The update rule of the scaling matrix $\mathbf{S}$ is derived in a similar way using another auxiliary function $g(\mathbf{S}, \mathbf{S}')$, which is given by

$$g(\mathbf{S}, \mathbf{S}') = -\sum_{ik} 2(\mathbf{W}'\mathbf{VH})_{ik}\mathbf{S}'_{ik}(1 + \log\frac{\mathbf{S}_{ik}}{\mathbf{S}'_{ik}}) + \sum_{ik} \frac{(\mathbf{W}^{\mathrm{T}}\mathbf{WS}'\mathbf{H}^{\mathrm{T}}\mathbf{H})_{ik}\mathbf{S}^2_{ik}}{\mathbf{S}'_{ik}}. \tag{19}$$

Algorithm 2: Sparse and Orthogonal NMF (**SO-NMF**)

Input: Matrix $\mathbf{V} \in \mathbb{R}^{n \times d}$, Sparsity ξ, Iterations r
Output: $\mathbf{W}$, $\mathbf{S}$ and $\mathbf{H}$

1 Initialize $\mathbf{W}$, $\mathbf{S}$ and $\mathbf{H}$ with random positive values.
2 Project each column $\mathbf{h}_i$ of $\mathbf{H}$, such that
  $\mathbf{h}_i \geq 0$, $\|\mathbf{h}_i\|_2 = 1$,
  $\|\mathbf{h}_i\|_1$ is set to satisfy Eq. (3) s.t. $s(\mathbf{h}_i) = \xi$.
3 **while** *iter* $< r$ **do**
4 Update $\mathbf{W}$:
  $$\mathbf{W} \leftarrow \mathbf{W} \odot \sqrt{\frac{\mathbf{VHS}^{\mathrm{T}}}{\mathbf{WW}^{\mathrm{T}}\mathbf{VHS}^{\mathrm{T}}}}$$
5 Update $\mathbf{S}$:
  $$\mathbf{S} \leftarrow \mathbf{S} \odot \left(\sqrt{\frac{\mathbf{W}^{\mathrm{T}}\mathbf{VH}}{\mathbf{W}^{\mathrm{T}}\mathbf{WSH}^{\mathrm{T}}\mathbf{H}}}\right)_{ii}$$
6 Update $\mathbf{H}$:
  $$\mathbf{H} \leftarrow \mathbf{H} - \mu\left(\mathbf{HH}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\mathbf{WS} - \mathbf{V}^{\mathrm{T}}\mathbf{WS}\right)$$
7 Repeat projection step 2
end

Finding $\mathbf{S}$ to minimize Eq. (19) results in the multiplicative update rule of $\mathbf{S}$ as

$$\mathbf{S} \leftarrow \mathbf{S} \odot \left(\sqrt{\frac{\mathbf{W}^{\mathrm{T}}\mathbf{VH}}{\mathbf{W}^{\mathrm{T}}\mathbf{WSH}^{\mathrm{T}}\mathbf{H}}}\right)_{ii}, \tag{20}$$

where only diagonal elements of $\mathbf{S}$ are updated to maintain the diagonal matrix property of $\mathbf{S}$ while non-diagonal elements remain zero.

Finally, we impose a sparsity constraint on $\mathbf{H}$ based on Hoyer's work [58], and derive the following update rule for $\mathbf{H}$ by combining Eq. (10) and (14)

$$\mathbf{H} \leftarrow \mathbf{H} - \mu\left(\mathbf{HH}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\mathbf{WS} - \mathbf{V}^{\mathrm{T}}\mathbf{WS}\right), \tag{21}$$

where μ denotes the gradient step size determined automatically. Hoyer's projected gradient descent algorithm first takes a step in the direction of the negative gradient and makes a projection onto the constraint space satisfying $s(\mathbf{h}_i) = \xi$, $\forall i$. Since the gradient of $\mathbf{H}$ is obtained from Eq. (6), $\mathbf{H}$ is updated to satisfy both the sparsity and orthogonality constraints in each iteration.

This is followed by iteratively updating $\mathbf{W}$, $\mathbf{S}$ and $\mathbf{H}$ using Eq. (18), (20), and (21), respectively, and a further projection step. The entire process of the optimization is summarized in Algorithm 2.

4.4 Simultaneous Clustering and Labeling

Once the algorithm converges, each image should be assigned to a specific cluster based on the coefficient matrix $\mathbf{W}$. For this purpose, $\mathbf{W}$ needs to be normalized with respect to the basis matrix $\mathbf{H}$, such that values in each column add to unity [60]. The procedure for obtaining the cluster membership probability is as follows:

$$\mathbf{V} \approx \mathbf{WSH}^{\mathrm{T}} = \mathbf{WSD}(\mathbf{HD}^{-1})^{\mathrm{T}} \equiv \widetilde{\mathbf{W}}\widetilde{\mathbf{H}} \tag{22}$$

where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{H}_{ji}$, which is used to normalize $\mathbf{H}$ and subsequently adjust the scale of $\mathbf{W}$ according to the scale of $\mathbf{S}$. The adjusted coefficient matrix $\widetilde{\mathbf{W}}$ is then used to determine the final cluster membership of each of the n images. For each row $\tilde{\mathbf{w}}_i$ of $\widetilde{\mathbf{W}}$ representing the i^{th} image $\mathbf{v}_i$, we calculate the cluster assignment by taking the maximum responses in $\tilde{\mathbf{w}}_i$ as

$$C(\mathbf{v}_i) = \arg\max_j \tilde{\mathbf{w}}_{ij}. \tag{23}$$

To assign labels to a cluster, we exploit the normalized basis matrix $\widetilde{\mathbf{H}}$, where each column vector $\mathbf{h}_j \in \mathbb{R}^d$ represents a cluster prototype in label feature space. Therefore, we can annotate clusters by extracting labels that have high response in $\mathbf{h}_j$. By the sparse nature of $\mathbf{h}_j$, the resulting annotations consist of only a small set of labels. For our experiments, we have chosen labels that have associated values in $\mathbf{h}_i$ corresponding to the cumulative top 60% of the column's normalized sum.

When we want to rank labels for an individual image, we reconstruct the label feature matrix $\mathbf{V}' = \mathbf{WSH}^{\mathrm{T}}$, where each row $\mathbf{v}'_i \in \mathbb{R}^d$ can then be interpreted as an association of labels with the i-th image.

5 Experiments

We evaluate our algorithm on two large scale standard benchmark datasets, CIFAR-100 [61] and ImageNet [52]. We compare our algorithm to existing methods in terms of both clustering and labeling performance. In addition, we present an extension of our algorithm to image classification and evaluate its performance as well.

5.1 Datasets and Experimental Setup

According to our scenario, categories in query images are not as diverse as Ref-DB and image labeling is performed by the relative label information of a query image within the context of a small set of the query images as well as the absolute label information. To simulate such a scenario, we randomly select a subset of categories for the query image set out of all the classes available in the Ref-DB.

5.1.1 CIFAR-100

CIFAR-100 is a subset of the Tiny Images database [26] and comprises 100 classes with 600 images each. Having been collected from the internet, the images have highly diverse appearances but the sizes of all images are normalized to 32×32.

In our experiments with CIFAR-100, 50K images are used as a Ref-DB and the remaining 10K images are used as a test dataset. We randomly select 10 classes (10% of entire classes) with 100 images per class for query image set construction, and measure the performance based on 10 different sets. For each image, the label feature matrix $\mathbf{V}$ is constructed by k-nearest neighbor method ($k = 15$) as described in Section 3. We use GIST and a simple color histogram as visual features.

5.1.2 ImageNet

ImageNet is a large-scale image database with 21K+ categories organized according to the WordNet hierarchy [62]. Like the Tiny Images dataset, all images are collected from the internet and vary significantly in appearance.

We use 1.2M training images and 50K validation images from ILSVRC2012 1K dataset [63] as Ref-DB and testing dataset, respectively. As in our experiment for CIFAR-100 dataset, we randomly select 20 classes out of the 1K classes and generate a query image set by sampling 50 images per class from the testing dataset. The performance of our algorithm is evaluated based on 20 different sets. Although the size of Ref-DB in the ImageNet database is significantly larger then CIFAR-100, the size of our query image set remains small since one of our goals is to identify a small number of data-driven labels depending on the query image set. To construct label features, we represent images based on the deep convolutional network using the Caffe library [64] with the pre-trained network from [6] for k nearest neighbor search ($k = 30$). Specifically, we apply input images to the forward propagation until the next to last layer and generate 4,096 dimensional vectors.

5.2 Image Clustering Performance

The proposed algorithm retrieves a label set for each image through clustering. The overall accuracy of our label extraction algorithm thus depends on the clustering performance, which is evaluated by three popular measures in the literature: Adjusted Rand Index (ARI) [65], Normalized Mutual Information (NMI) [66] and Accuracy (AC) [60]. We compare the performance of our algorithm (SO-NMF) with three existing clustering methods: standard k-means clustering (KM), self-tuning spectral clustering (SC) [67], and probabilistic Latent Semantic Indexing (pLSI) [68]. To assess the contribution of the imposed constraints, comparison to two baseline algorithms is conducted additionally: NMF with orthogonality constraint [59] (O-NMF) and with sparsity constraint [58] (S-NMF). Note that pLSI and NMF with KL-divergence have the same objective function [69] and pLSI can work as baseline NMF with no constraint.

Table 1 illustrates comparative clustering results on the CIFAR-100 and the ImageNet datasets. The proposed algorithm (SO-NMF) operating in the label feature space, substantially outperforms all other methods in all three clustering quality measures. Imposing both sparsity and orthogonality constraints is beneficial for query image clustering in the label feature space. It is interesting that NMF methods with partial or full constraints (O-NMF, S-NMF and SO-NMF) perform poorly in visual feature space; this is because our assumptions on properties of good clusters, enforced by the sparsity and orthogonality constraints, are more appropriate for the label-based representation. When the orthogonality constraint is applied to either $\mathbf{W}$ or $\mathbf{H}$, we observe that clusters are often not sufficiently discriminative and the overall clustering accuracy drops substantially.

Our clustering performance in the label feature space depends on the quality of the Ref-DB. In order to see the effect of the quality of Ref-DB, we conduct two experiments,

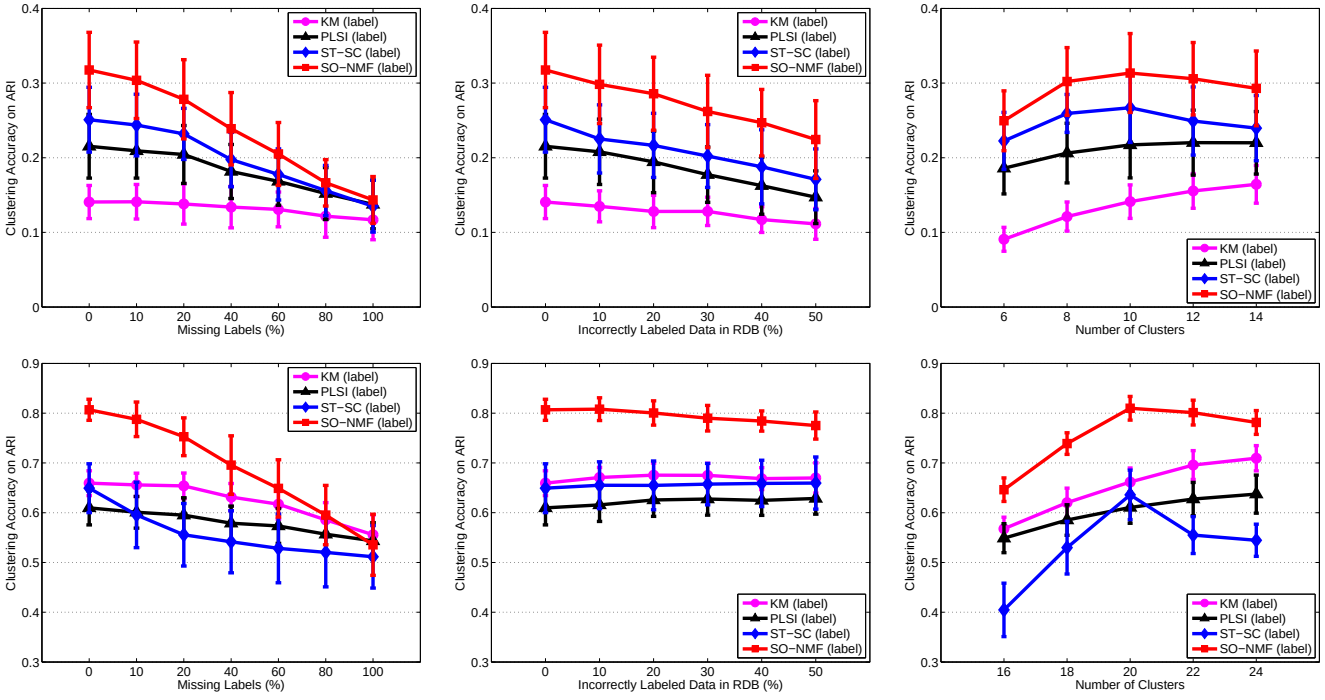
TABLE 1
Image clustering performance by SO-NMF compared to other methods with visual and label features.

CIFAR-100 dataset

	Visual feature						Label feature					
	KM	SC	PLSI	O-NMF	S-NMF	SO-NMF	KM	SC	PLSI	O-NMF	S-NMF	SO-NMF
ARI	0.07±0.02	0.08±0.03	0.09±0.03	0.06±0.02	0.06±0.02	0.04±0.02	0.14±0.02	0.25±0.04	0.22±0.04	0.17±0.03	0.30±0.05	**0.32±0.05**
NMI	0.14±0.03	0.13±0.03	0.17±0.04	0.12±0.03	0.12±0.03	0.09±0.03	0.34±0.03	0.37±0.05	0.32±0.05	0.34±0.03	0.36±0.04	**0.39±0.05**
AC	0.26±0.02	0.26±0.03	0.28±0.03	0.24±0.02	0.24±0.02	0.21±0.03	0.41±0.03	0.51±0.06	0.44±0.04	0.44±0.03	0.54±0.04	**0.57±0.05**

ImageNet dataset

	Visual feature						Label feature					
	KM	SC	PLSI	O-NMF	S-NMF	SO-NMF	KM	SC	PLSI	O-NMF	S-NMF	SO-NMF
ARI	0.55±0.03	0.74±0.06	0.74±0.05	0.68±0.03	0.68±0.04	0.67±0.03	0.66±0.02	0.65±0.05	0.61±0.03	0.71±0.03	0.76±0.02	**0.81±0.02**
NMI	0.73±0.02	0.84±0.04	0.84±0.03	0.82±0.02	0.80±0.03	0.79±0.02	0.80±0.02	0.82±0.02	0.75±0.02	0.82±0.02	0.84±0.02	**0.87±0.02**
AC	0.67±0.02	0.84±0.05	0.82±0.04	0.76±0.03	0.77±0.03	0.75±0.03	0.73±0.02	0.83±0.02	0.71±0.02	0.80±0.02	0.86±0.01	**0.89±0.01**

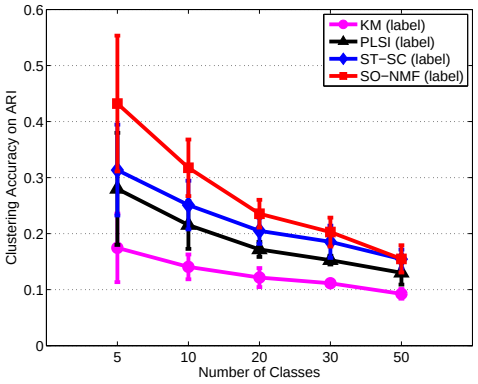


(a) Missing classes in Ref-DB (b) Incorrect labels in Ref-DB (c) Different number of Clusters

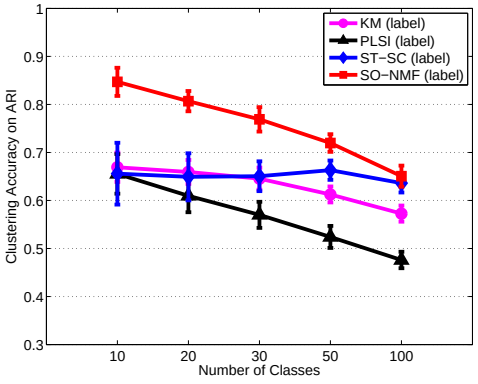
Fig. 3. Clustering performance comparison under various situations in (top) CIFAR-100 and (bottom) ImageNet dataset. (a) Clustering accuracy in the presence of missing labels in the Ref-DB. (b) Clustering accuracy in the presence of incorrect labels in Ref-DB. (c) Clustering accuracy by varying the number of clusters. Note that SO-NMF based on label features outperforms all other methods. The performance with visual features is not illustrated in these graphs since the ARI for all algorithms is very low (almost below 0.1) as seen in Table 1.

which investigate the effect of missing classes and incorrectly labeled classes in Ref-DB, and present the results in Figure 3. Figure 3(a) shows that our method outperforms other approaches even when a significant portion of the true labels of the query images are missing in the Ref-DB. This means that other visually similar images in the Ref-DB can still describe the query images effectively and facilitate image clustering. Although the performance of SO-NMF degraded quickly as the number of missing labels increases, our algorithm nevertheless outperforms the other methods over a large reasonable range of missing rate. Figure 3(b) illustrates the effect of incorrectly labeled images in the Ref-DB. For this experiment, we randomly exchange image labels of some fraction of pairs of images in the database, from 10% to 50%. In addition, we investigate the robustness of our method by varying the number of clusters. Note that there are 10 and 20 true underlying clusters in our experiments for the CIFAR-100 and ImageNet datasets, respectively. According to Figure 3(c), our method is robust with respect to the number of clusters and increasing the number of desired clusters results in less performance degradation.

Our algorithm assumes that the number of classes in query images, c, is small compared to the entire set of

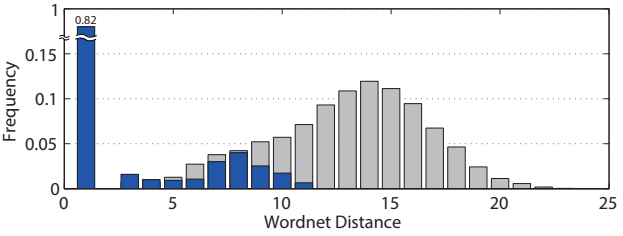


(a) CIFAR-100 dataset

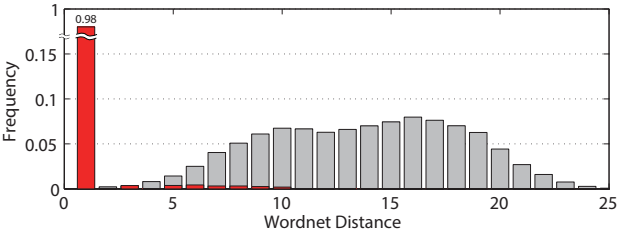


(b) ImageNet dataset

Fig. 4. Comparison of clustering performance with increasing number of classes in query images in (top) CIFAR-100 and (bottom) ImageNet dataset.



(a) WordNet distance distribution comparison in CIFAR-100 dataset



(b) WordNet distance distribution comparison in ImageNet dataset

Fig. 5. Quality of per-cluster labels obtained by our approach. The semantic distance is defined as the sum of distances to the lowest common ancestor in the WordNet hierarchy of ground truth and identified labels. Blue and red bar graphs indicate the distributions of WordNet distances between ground truth and identified labels by SO-NMF. Gray bar graphs illustrate the distribution of WordNet distances between the labels of randomly chosen images in the Ref-DB. Note that our method tends to find more relevant labels, even if they are not correct.

classes in Ref-DB, d, ($c \ll d$), which is reasonable for many practical applications. However, to observe the effect of c close to d, we conduct additional experiments, where the number of classes in the query images varies, and present the results in Figure 4.

As the true number of classes increases in query image sets, the clustering performance generally degrades in all methods. The advantage of the proposed algorithm is particularly significant when the number of classes in the query image set is small, which corresponds to our target scenario. This is partly because our assumptions about sparsity and orthogonality can be met better with a small number of classes. However, the benefit of our algorithm is not limited to improving image clustering performance; another crucial contribution is the capability to label individual clusters and images, which will be evaluated next.

Our matrix factorization takes 3 and 16 seconds for the matrices with size of 1000×100 in CIFAR-100 and 1000×1000 in ImageNet dataset.

5.3 Image Labeling Performance

Our method can identify the labels for each cluster and each image as described in Section 4.4. We first evaluate the quality of label per cluster based on semantic similarity using the WordNet tree structure [62]. For this purpose, we select the most relevant label for each cluster based on Eq. (23), and compute the WordNet distance between the identified labels and ground truth labels. Note that the ground truth label for each cluster is determined by the majority labels of the images in the cluster. As illustrated in Figure 5, SO-NMF identifies the ground truth labels successfully (WordNet distance = 1) for 82% and 98% of clusters on average on the CIFAR-100 and ImageNet datasets, respectively. The quality of unmatched labels is also measured by computing their semantic distances from true labels. These distances are compared with distances between all unmatched pairs of labels in Ref-DB. Figure 5 shows that the distances from extracted labels to true labels are close compared to random distances in both datasets. This suggests that labels identified by SO-NMF tend to be semantically related even though they might be incorrect.

We also evaluate per-image labeling quality by comparing three different algorithms—k-nearest neighbors (KNN), SVM-KNN [29], and our algorithm (SO-NMF). Note that a simple k-nearest neighbor method can successfully categorize images with a large auxiliary database [26]. To label images by KNN, we simply select the label with maximum frequency from the label feature representation of an image. SVM-KNN is the combination of SVM and KNN approaches, and attempts to overcome the limitation of the standard KNN; it trains a classifier to identify the label

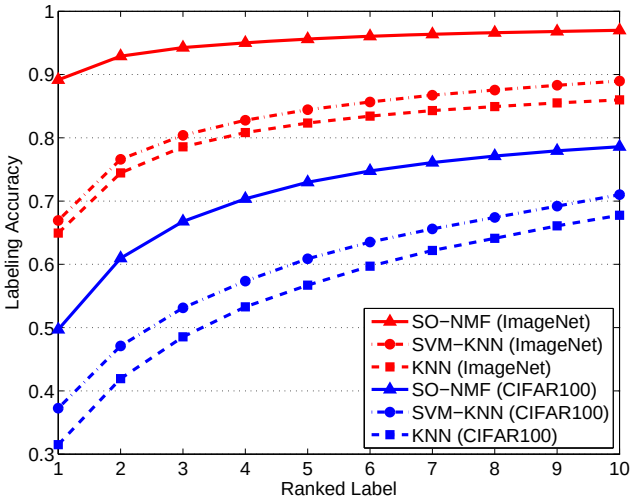


Fig. 6. Cumulative Match Characteristic (CMC) curves for image labeling accuracy. SO-NMF is substantially better than other methods.

of each image, where the training images and their labels are obtained adaptively based on k-nearest neighbor search. For k-nearest neighbor search from Ref-DB, k is set to 500 and 100 in CIFAR-100 and ImageNet dataset, respectively. SVM's with a polynomial kernel ($d = 6, r = 1, \gamma = 0.01$) and a RBF kernel ($\gamma = 1$) are used to train classifiers for CIFAR-100 and ImageNet dataset, respectively. These parameters achieve best performance of SVM-KNN in our experiment.

Cumulative Match Characteristic (CMC) curves are employed to illustrate per-image labeling performance and the results are presented in Figure 6. The horizontal axis denotes the number of ranked labels extracted by our algorithm and the vertical one indicates the percentage of times that the correct label is among the top n label candidates. In both datasets, our algorithm is substantially better than the other two, which suggests that image clustering based on SO-NMF can reduce the noise in the initial representation of images. However, if images in Ref-DB involve many labels annotated, label features become dense and are difficult to enforce sparsity constraint; the performance of our algorithm would be degraded in this case.

We present examples of identified clusters with associated images and per-image annotations in Figure 7 and 8, respectively, for both datasets. SO-NMF provides visually and semantically relevant groupings with proper labels and identifies scenes or objects within images accurately.

5.4 Extension to Weakly-Supervised Learning

Since our framework can provide labels for each identified cluster, it is possible to collect training images from the Ref-DB based on the extracted labels. Specifically, we assign the most relevant label to each cluster and train a classifier using the images in Ref-DB with the obtained

Labels	(conf.)	Labels	(conf.)	Labels	(conf.)
television	**(0.30)**	**ray**	**(0.25)**	**train**	**(0.22)**
wardrobe	(0.10)	shark	(0.16)	streetcar	(0.19)
couch	(0.05)	turtle	(0.12)	bus	(0.07)
plate	**(0.44)**	**fox**	**(0.33)**	**spider**	**(0.25)**
clock	(0.10)	tiger	(0.13)	beetle	(0.10)
bowl	(0.10)	squirrel	(0.08)	cockroach	(0.06)
sunflower	**(0.55)**	**couch**	**(0.23)**	sunflower	(0.16)
bee	(0.10)	television	(0.07)	**spider**	**(0.06)**
poppy	(0.10)	bed	(0.07)	poppy	(0.03)
orange	**(0.25)**	**porcupine**	**(0.72)**	orange	(0.26)
plate	(0.20)	leopard	(0.06)	**sunflower**	**(0.22)**
apple	(0.06)	shrew	(0.04)	poppy	(0.07)

(a) Image labeling performance in CIFAR-100 dataset.

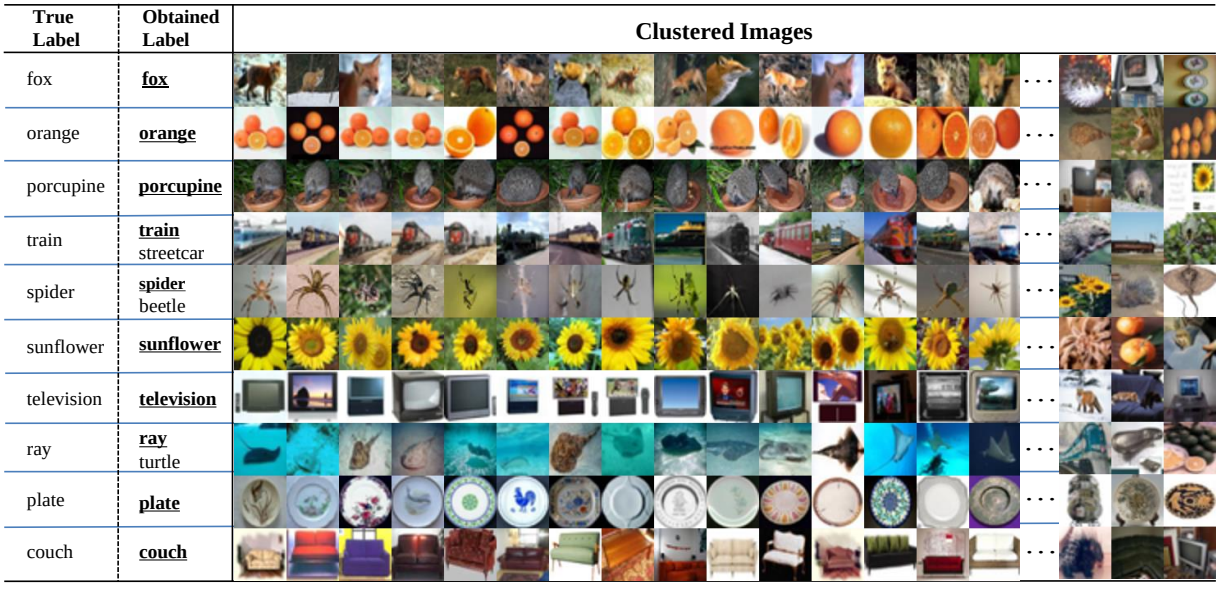
Labels	(conf.)	Labels	(conf.)	Labels	(conf.)
poncho	**(0.75)**	**toilet seat**	**(0.91)**	**black bear**	**(0.80)**
stole	(0.18)	tissue	(0.07)	sloth bear	(0.19)
harvester	**(0.83)**	**redbone**	**(0.56)**	**crock pot**	**(0.94)**
thresher	(0.16)	ridgeback	(0.32)	hot pot	(0.03)
turtle	**(0.94)**	**chain saw**	**(0.32)**	**breakwater**	**(0.63)**
loggerhead	(0.05)	harvester	(0.16)	beacon	(0.16)
toucan	**(0.93)**	**pug**	**(0.39)**	**scotch ter.**	**(0.88)**
hornbill	(0.06)	border ter.	(0.22)	schnauzer	(0.11)
border ter.	**(0.94)**	**boa**	**(0.85)**	**carpenterkit**	**(0.88)**
Irish ter.	(0.05)	python	(0.12)	oscilloscope	(0.05)
whale	**(0.87)**	**hay**	**(0.95)**	**scoreboard**	**(0.64)**
greywhale	(0.12)	megalith	(0.03)	oscilloscope	(0.21)

(b) Image labeling performance in ImageNet dataset.

Fig. 8. Per-image annotation results in both datasets. Only relevant labels are shown, where true labels are underlined and marked in bold.

label[1]. To evaluate the performance of our algorithm with SVM learning for identified classes (SO-NMF+SVM), several baseline algorithms are employed. We investigate the performance of supervised techniques such as KNN and SVM-KNN, which are comparable to our method, and an upper bound of our algorithm obtained by training an SVM on ground truth labels, denoted by SVM-GT. We also present how much the categorization performance is improved through integrating supervised learning by comparing the performance between SO-NMF and SO-NMF+SVM. We use an SVM with a polynomial kernel ($d = 6$, $r = 1$, $\gamma = 0.01$, $C = 1$) for the CIFAR-100 dataset, and a simple linear SVM ($C = 1$) for the ImageNet dataset, respectively. Figure 9 illustrates clustering performance of all compared algorithms, where classification tasks are also evaluated by clustering quality measures. SO-NMF+SVM is competitive with SVM-GT, and outperforms all others. To investigate classification accuracy, sample confusion tables of SO-NMF+SVM, SVM-KNN, and KNN are presented in Figures 10 and 11. The overall accuracy of our algorithm is better than the other two techniques as described in the captions of the figures.

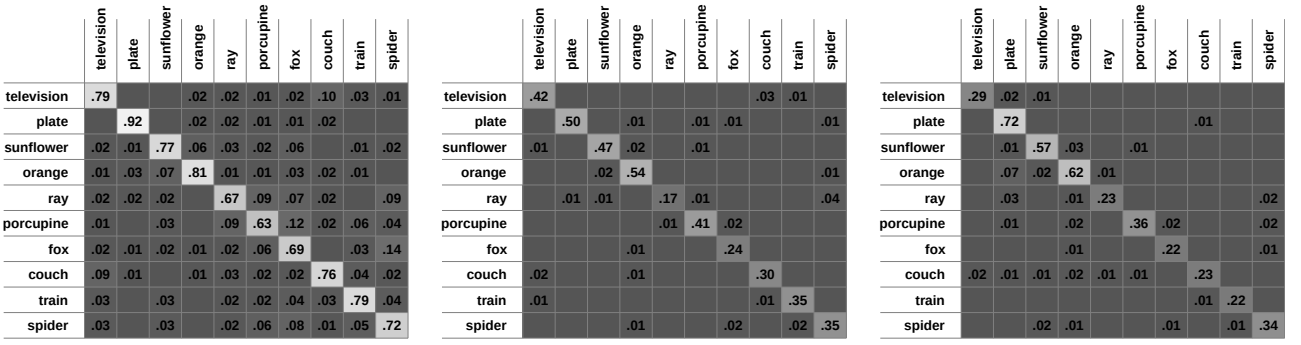
1. We only need to train models for the identified classes from clustering procedure, which reduces the number of candidate classes significantly.



(a) Joint clustering results in CIFAR-100 dataset.

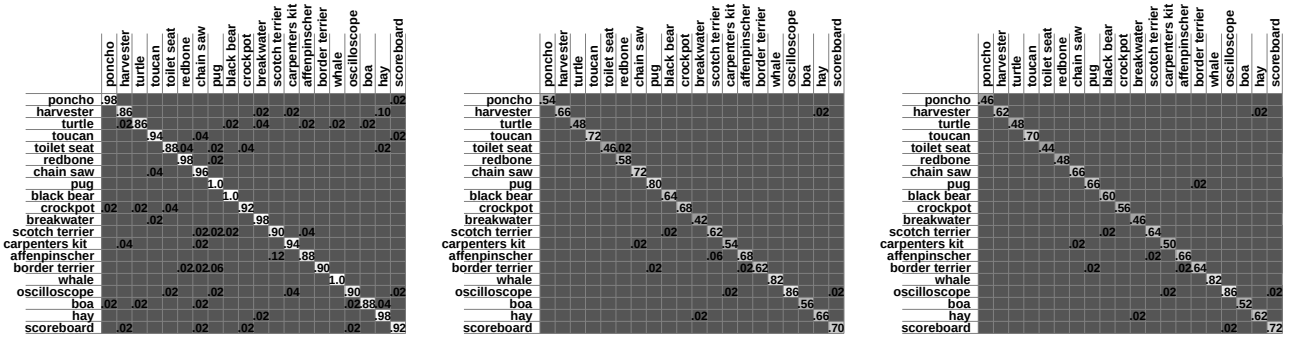
(b) Joint clustering results in ImageNet dataset.

Fig. 7. Joint clustering and labeling performance in (top) CIFAR-100 and (bottom) ImageNet datasets. In (a) and (b), 15 most confident and 3 least confident images are presented for each class (sorted by confidence).



(a) SO-NMF with SVM (b) SVM-KNN (c) KNN

Fig. 10. Performance evaluation by confusion tables in CIFAR-100 dataset. (a) SO-NMF+SVM (b) SVM-KNN (c) KNN. Our algorithm performs notably better in identifying correct class labels. Average precisions of entire 10 subsets of the three algorithms are 0.66, 0.36, and 0.32, respectively.



(a) SO-NMF with SVM (b) SVM-KNN (c) KNN

Fig. 11. Performance evaluation by confusion tables in ImageNet dataset. (a) SO-NMF+SVM (b) SVM-KNN (c) KNN. Average precisions of entire 20 subsets of the three algorithms are 0.94, 0.67, and 0.65, respectively.

6 Conclusion

We proposed a novel joint image clustering and annotation algorithm using label features employing a non-negative matrix factorization framework. The label feature is extracted from a large-scale Ref-DB, and SO-NMF is employed to cluster images and retrieve a sparse set of labels from noisy label frequency information. The semantic gap between visual and semantic spaces is reduced by representing images using label features, and SO-NMF captures discriminative and representative labels of an image successfully. Our algorithm was tested with two challenging datasets and obtained superior performance to other image clustering and labeling methods. Also, we extended the proposed algorithm to image classification, and obtained outstanding performance to a few comparable algorithms.

Acknowledgments

This work was supported by ICT R&D program of MSIP/ ITP [14-824-09-006; 14-824-09-014]. The work of J. Choi and L. S. Davis was supported by NSF Grant IIS1262121: Video analytics in large heterogeneous repositories.

References

[1] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, and W. Buntine, "Unsupervised Object Discovery: A comparison," *Int'l. J. of Computer Vision*, vol. 88, no. 2, pp. 284–302, 2010.

[2] D. Lowe, "Object Recognition from Local Scale-Invariant Features," in *Proc. 7th Int'l. Conf. on Computer Vision,* Kerkyra, Greece, vol. 2, 1999, pp. 1150–1157.

[3] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int'l. J. of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* San Diego, CA, vol. 1, 2005, pp. 886–893.

[5] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51 – 59, 1996.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proc. of the Neural Information Processing Systems*, 2012, pp. 1097–1105.

[7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing Objects by their Attributes," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Miami, FL, 2009.

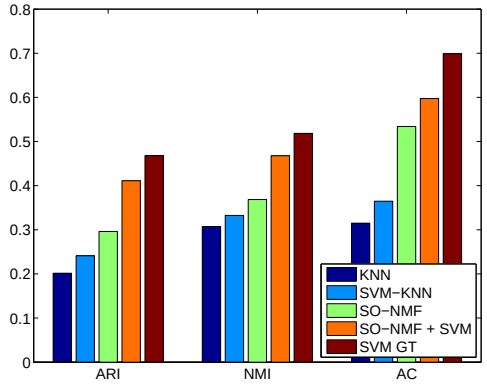
[8] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Miami, FL, 2009.

[9] C. Wang, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Miami, FL, 2009, pp. 1903–1910.

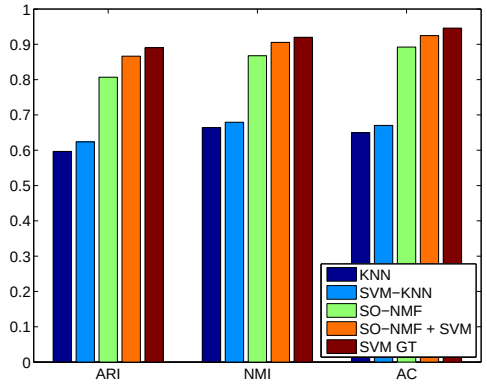
[10] J. Deng, A. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Providence, RI, 2011, pp. 785–792.

[11] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient Object Category Recognition Using Classemes," in *Proc. European Conf. on Computer Vision,* Crete, Greece, 2010, pp. 776–789.

[12] A. Bergamo and L. Torresani, "Meta-Class Features for Large-Scale Object Categorization on a Budget," in *Proc. IEEE Conf. on*



(a) CIFAR-100 dataset



(b) ImageNet dataset

Fig. 9. Performance of the extension to supervised image categorization. Our algorithm with SVM (SO-NMF+SVM) outperforms all other methods except SVM-GT, which represents an upper bound on performance of our algorithm.

Computer Vision and Pattern Recognition, Providence, RI, 2012, pp. 3085–3092.

[13] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei, "Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification," in *Proc. of the Neural Information Processing Systems*, 2010, pp. 1378–1386.

[14] Y. J. Lee and K. Grauman, "Object-Graphs for Context-Aware Category Discovery," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* San Francisco, CA, 2010, pp. 1–8.

[15] ——, "Learning the easy things first: Self-paced visual category discovery," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Colorado Springs, CO, 2011, pp. 1721–1728.

[16] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Int'l. J. of Computer Vision*, 2009, pp. 309–316.

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

[18] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proc. 26th Int'l. ACM SIGIR Conf. on Research and development in informaion retrieval*, 2003, pp. 127–134.

[19] F. Monay and D. Gatica-Perez, "Plsa-based image auto-annotation: constraining the latent space," in *Proc. 12th Annu. ACM Int'l. Conf. on Multimedia*, 2004, pp. 348–351.

[20] Y. Liu, R. Jin, and L. Yang, "Semi-supervised multi-label learning by constrained non-negative matrix factorization," in *Proc. 21st Nat. Conf. on Artificial intelligence*, 2006, pp. 421–426.

[21] N. Zhou, W. Cheung, G. Qiu, and X. Xue, "A hybrid probabilistic model for unified collaborative and content-based image tagging," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 33, no. 7, pp. 1281–1294, 2011.

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

[23] C. Cusano, G. Ciocca, and R. Schettini, "Image annotation using SVM," *Proceedings of SPIE*, vol. 5304, pp. 330–338, 2004.

[24] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 3, pp. 394–410, 2007.

[25] K. Tieu and P. Viola, "Boosting image retrieval," *Int'l. J. of Computer Vision*, vol. 56, no. 1-2, pp. 17–36, 2004.

[26] A. Torralba, R. Fergus, and W. Freeman, "80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, pp. 1958–1970, 2008.

[27] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 6, pp. 607–616, 1996.

[28] C. Domeniconi and D. Gunopulos, "Adaptive Nearest Neighbor Classification using Support Vector Machines," in *Proc. of the Neural Information Processing Systems*, 2001, pp. 665–672.

[29] A. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* New York, NY, vol. 2, 2006, pp. 2126–2136.

[30] T. Li, S. Yan, T. Mei, X.-S. Hua, and I. S. Kweon, "Image decomposition with multilabel context: Algorithms and applications," *IEEE Trans. Image Process.*, 2011.

[31] T. Li, T. Mei, S. Yan, I.-S. Kweon, and C. Lee, "Contextual decomposition of multi-label images," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Miami, FL, 2009.

[32] G. Zhu, S. Yan, and Y. Ma, "Image Tag Refinement Towards Low-Rank, Content-Tag Prior and Error Sparsity," in *Proc. of the Int'l Conf. on Multimedia*, 2010.

[33] C. Yang, M. Dong, and J. Hua, "Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* New York, NY, vol. 2, 2006, pp. 2057–2063.

[34] J. Tighe and S. Lazebnik, "SuperParsing: Scalable Nonparametric Image Parsing with Superpixels," in *Proc. European Conf. on Computer Vision,* Crete, Greece, 2010.

[35] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 4, pp. 509–522, 2002.

[36] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int'l. J. of Computer Vision*, vol. 62, no. 1-2, pp. 61–81, 2005.

[37] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Anchorage, AK, 2008, pp. 1–8.

[38] M. P. Kumar, P. H. S. Torr, and A. Zisserman, "An Invariant Large Margin Nearest Neighbour Classifier," in *Proc. 11th Int'l. Conf. on Computer Vision,* Rio de Janeiro, Brazil, 2007, pp. 1–8.

[39] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang, "Hierarchical Gaussianization for Image Classification," in *Proc. 11th Int'l. Conf. on Computer Vision,* Rio de Janeiro, Brazil, 2007, pp. 1971–1977.

[40] G. Kim, C. Faloutsos, and M. Hebert, "Unsupervised Modeling of Object Categories Using Link Analysis Techniques," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Anchorage, AK, 2008, pp. 1–8.

[41] D. Liu and T. Chen, "Unsupervised Image Categorization and Object Localization using Topic Models and Correspondences between Images," in *Proc. 11th Int'l. Conf. on Computer Vision,* Rio de Janeiro, Brazil, 2007, pp. 1–7.

[42] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman, "Using Multiple Segmentations to Discover Objects and their Extent in Image Collections," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* New York, NY, vol. 2, 2006, pp. 1605–1614.

[43] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images," in *Proc. 10th Int'l. Conf. on Computer Vision,* Beijing, China, vol. 1, 2005, pp. 370–377.

[44] D. Dai, M. Prasad, C. Leistner, and L. J. V. Gool, "Ensemble partitioning for unsupervised image categorization." in *Proc. European Conf. on Computer Vision,* Florence, Italy, 2012, pp. 483–496.

[45] A. Shrivastava, S. Singh, and A. Gupta, "Constrained semi-supervised learning using attributes and comparative attributes," in *Proc. European Conf. on Computer Vision,* Florence, Italy, 2012, pp. 369–383.
[46] R. Fergus, Y. Weiss, and A. Torralba, "Semi-supervised Learning in Gigantic Image Collections," in *Proc. of the Neural Information Processing Systems*, 2009.
[47] R. Salakhutdinov, A. Torralba, and J. Tenenbaum, "Learning to Share Visual Appearance for Multiclass Object Detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Colorado Springs, CO, 2011, pp. 1481–1488.
[48] J. J. Lim, R. Salakhutdinov, and A. Torralba, "Transfer Learning by Borrowing Examples for Multiclass Object Detection," in *Proc. of the Neural Information Processing Systems*, 2011, pp. 118–126.
[49] J. Choi, M. Rastegari, A. Farhadi, and L. S. Davis, "Adding unlabeled samples to a category by learned attributes," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Portland, OR, 2013.
[50] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* San Francisco, CA, 2010, pp. 902–909.
[51] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei, "What Does Classifying More Than 10,000 Image Categories Tell Us?" in *Proc. European Conf. on Computer Vision,* Crete, Greece, 2010, pp. 71–84.
[52] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Miami, FL, 2009, pp. 248–255.
[53] D. Parikh and K. Grauman, "Interactively Building a Discriminative Vocabulary of Nameable Attributes," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Colorado Springs, CO, 2011.
[54] Z. Harchaoui, M. Douze, M. Paulin, M. Dud'ık, and J. Malick, "Large-scale image classification with trace-norm regularization," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Providence, RI, 2012.
[55] N. Loeff and A. Farhadi, "Scene Discovery by Matrix Factorization," in *Proc. European Conf. on Computer Vision,* Marseille, France, 2008.
[56] R. S. Cabral, F. D. la Torre, J. P. Costeira, and A. Bernardino, "Matrix Completion for Multi-label Image Classification," in *Proc. of the Neural Information Processing Systems*, 2011.
[57] D. D. Lee and H. S. Seung, "Learning the Parts of Objects by Non-Negative Matrix Factorization," *Nature*, vol. 401, pp. 788–791, 1999.
[58] P. O. Hoyer, "Non-Negative Matrix Factorization with Sparseness Constraints," *J. of Machine Learning Research*, vol. 5, no. 5, pp. 1457–1469, 2004.
[59] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal Nonnegative Matrix Tri-Factorizations for Clustering," in *ACM Conf. on Knowledge, Discovery and Data Mining*, 2006, p. 126.
[60] T. Li and C. Ding, "The Relationships Among Various Nonnegative Matrix Factorization Methods for Clustering," in *Proc. IEEE Int'l Conf. on Data Mining,* Hong Kong, China, 2006, pp. 362–371.
[61] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," University of Toronto, Tech. Rep., 2009.
[62] G. A. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[63] A. Berg, J. Deng, and L. Fei-Fei, "Large scale visual recognition challenge (ilsvrc) 2012," http://www.image-net.org/challenges/LSVRC/2012/, 2012.
[64] Y. Jia, "Caffe: An open source convolutional architecture for fast feature embedding," http://caffe.berkeleyvision.org/, 2013.
[65] L. Hubert and P. Arabie, "Comparing Partitions," *J. of Classification*, vol. 2, no. 1, pp. 193–218, 1985.
[66] A. Strehl and J. Ghosh, "Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions," *J. of Mach. Learning Research*, vol. 3, pp. 583–617, 2002.
[67] L. Zelnik-Manor and P. Perona, "Self-Tuning Spectral Clustering," in *Proc. of Neural Inform. Process. Syst.*, 2005, pp. 1601–1608.
[68] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22th Int'l. ACM SIGIR Conf. on Research and development in informaion retrieval*, 1999, pp. 50–57.
[69] E. Gaussier and C. Goutte, "Relation between plsa and nmf and implications," in *Proc. of the 28th Annual Int'l. ACM SIGIR Conf. on Research and Development in Informaion Retrieval*, 2005, pp. 601–602.

Seunghoon Hong received the B.S. degree from the Department of Computer Science and Engineering at POSTECH, Pohang, Korea in 2011. He is currently pursuing the Ph.D. degree in the Department of Computer Science and Engineering at POSTECH. His current research interests include computer vision, machine learning, and artificial intelligence. He received Microsoft Research Asia Fellowship in 2014.

Jonghyun Choi received the B.S. and M.S. degrees in electrical engineering and computer science from Seoul National University, Seoul Korea in 2003 and 2008 respectively. He is currently pursuing a PhD degree at the Computer Vision Laboratory in the University of Maryland, College Park. His research interest includes object recognition, regularized classifier learning and multi-stage learning using various visual features. He is a student member of the IEEE.

Jan Feyereisl received his B.Sc. and Ph.D. degrees from the School of Computer Science at The University of Nottingham in 2005 and 2010, respectively. He was then a Research Fellow at the Intelligent Modelling and Analysis Research Group at The University of Nottingham and in the Computer Vision Lab. at POSTECH, Korea. Currently he is a Senior Research Engineer at Samsung DMC R&D Center. His current research interests include machine learning and computer vision, in particular large-scale learning, deep learning and learning using privileged information.

Bohyung Han received the B.S. and M.S. degrees from the Department of Computer Engineering at Seoul National University, Korea, in 1997 and 2000, respectively, and the Ph.D. degree from the Department of Computer Science at the University of Maryland, College Park, MD, USA, in 2005. He was with Samsung Electronics Research and Development Center, Irvine, CA, USA, and Mobileye Vision Technologies, Princeton, NJ, USA. He is currently an Associate Professor with the Department of Computer Science and Engineering at POSTECH, Korea. He served as an Area Chair in ACCV 2012/2014 and WACV 2014, and as a Demo Chair in ACCV 2014. His current research interests include computer vision and machine learning.

Larry S. Davis received the BA degree from Colgate University in 1970 and the MS and PhD degrees in computer science from the University of Maryland in 1974 and 1976, respectively. From 1977 to 1981, he was an assistant professor in the Department of Computer Science at the University of Texas, Austin. He returned to the University of Maryland as an associate professor in 1981. From 1985 to 1994, he was the director of the University of Maryland Institute for Advanced Computer Studies. From 1999 to 2012, he was the chair of the Computer Science Department in the institute. He is currently a professor in the institute and in the Computer Science Department. He is a Fellow of ACM, IEEE, and IAPR.