Research Article

Nonnegative Signal Decomposition with Supervision

Teng Li,¹ Huan Chang,¹ and Jun Wu²

¹ The Anhui University, Hefei 230601, China

² The Institute of Software Application Technology, Guangzhou & Chinese Academy of Sciences, Guangzhou 511458, China

Correspondence should be addressed to Teng Li; tenglwy@gmail.com

Received 31 July 2013; Accepted 21 August 2013

Academic Editor: Shuping He

Copyright © 2013 Teng Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper presents a novel algorithm to numerically decompose mixed signals in a collaborative way, given supervision of the labels that each signal contains. The decomposition is formulated as an optimization problem incorporating nonnegative constraint. A nonnegative data factorization solution is presented to yield the decomposed results. It is shown that the optimization is efficient and decreases the objective function monotonically. Such a decomposition algorithm can be applied on multilabel training samples for pattern classification. The real-data experimental results show that the proposed algorithm can significantly facilitate the multilabel image classification performance with weak supervision.

1. Introduction

Signal decomposition, which separates a mixed source signal into its constitutive pure components, is an important step in many practical engineering problems. For example, in image classification or speech recognition, a single training or test sample we collected usually contains multiple additive signals from several classes. The decomposition of the mixed signals can yield pure feature representations for better classes model training and recognition. Figure 1 illustrates the decomposition for an image containing three object classes. From the given mixed signal on the top, we aim to derive a specific representation for each of its associated labels, which reflects a corresponding pure region in the image. The decomposed representation can be more accurate and informative for describing object categories.

Unsupervised signal decomposition, for example, blind source separation in signal processing, is a fundamental problem which has been well explored in the past decades [1, 2]. Various methods have been proposed for different tasks, including Independent Component Analysis (ICA) for speech signal [3] and Null Space Pursuit (NSP) for electricity consumption and global surface temperature [4]. In image processing, a straightforward way to decompose the image representation into label-pure representations is manual labeling. It is tedious and impractical in most cases. Most

conventional approaches to automatic image decomposition focused on clustering based on low-level visual cues, such as the bottom-up image segmentation in which pixels are locally grouped on the basis of their appearances [5] and the top-down image parsing in which primitives (e.g., rectangles, sketches, and edges) are correlated based on a few grammar rules [6]. However, all these methods have not taken the supervision information into consideration. In a practical case, online images from internet are usually associated with the textual labels on the webpage, which can be utilized to guide the decomposition.

In this paper, we focus on the signal decomposition problem with label supervision; that is, given a set of mixed signals each with class labels it contains, we decompose each signal to the specific labels. To this aim; the fundamental principle in pattern classification is utilized that signals of the same label should be close to each other while those of different labels should be relatively separated in feature space. Based on this assumption, a novel algorithm to numerically decompose mixed signals with labels in a collaborative way is proposed. The decomposition is formulated as an optimization problem incorporating nonnegative constraint for practical consideration, due to the nonnegative property of most real-life signals. Furthermore, we present a nonnegative data factorization solution, which is shown to be efficient and can decrease the objective function monotonically. Such a decomposition

FIGURE 1: Illustration of the proposed decomposition with an example of an image: given the top image with multiple labels, from its holistic feature representation shown right to it, for each of its associated labels, we aim to derive a label-specific representation, which reflects its actual region in the image. It is more informative to model individual labels compared to the holistic representation.

can be applied on multi-label training samples for pattern classification. The real-data experimental results show that the proposed algorithm can significantly facilitate the multi-label image classification with weak supervision.

2. Notation

In the following, we will use these rules to facilitate presentation: for any matrix $\mathbf{A}$, $\mathbf{A}_i$ means the i th row vector of $\mathbf{A}$, its corresponding lowercase version $\mathbf{a}_i$ means the i th column vector of $\mathbf{A}$, and $\mathbf{A}_{ij}$ denotes the element of $\mathbf{A}$ at the i th row and j th column. Throughout this paper we will also use the following notation:

N : number of total training samples,

$\bar{N}$: number of valid tags in the training set,

$\mathbf{l}_i$: label vector of the i th training sample,

$\mathbf{v}_i$: mixed signal representation vector of the i th training sample,

$\mathbf{V}$: mixed signal representations of all training samples ($\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N]$),

$\mathbf{r}_i^c$: decomposed signal representation of the c th label in i th training sample,

$\mathbf{r}_i$: decomposed representation of the i th valid tag in the full training set,

$\mathbf{R}$: decomposed signal representations from all training samples ($\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_{\bar{N}}]$),

α_{ij} : coefficient of normalized $\mathbf{r}_j$ vector in reconstructing $\mathbf{v}_i$,

$\mathbf{H}$: reconstruction coefficients matrix ($\mathbf{H}_{ij} = \alpha_{ji}$),

$\mathbf{Y}$: membership matrix denoting relation between tags and samples, defined in (5).

Here a label represents a class, and a tag is a specific label instance. For a specific sample $i \in \{1, \dots, N\}$, its signal representation can be calculated as a K -dimensional normalized vector $\mathbf{v}_i$ with $\|\mathbf{v}_i\|_1 = 1$, where $\|\cdot\|_1$ represents the ℓ^1 -norm operator. The labels of this sample are denoted in a vector $\mathbf{l}_i \in \mathcal{R}^C$ with only 0 or 1 elements:

$$l_i^c = \begin{cases} 1 & \text{if label } c \text{ is contained in sample } i, \\ 0 & \text{if there is no label } c \text{ in sample } i, \end{cases} \quad (1)$$

where C denotes the number of labels. Let a K -dimensional vector $\mathbf{r}_i^c$ denote the representation of the corresponding part for the c th label in the i th sample. For a set of training samples, we collect the representations for all nonzero label instances, that is, “tag”s, and reindex as vectors $\{\mathbf{r}_i\}$, $i \in \{1, \dots, \tilde{N}\}$, where $\tilde{N}$ is the number of valid tags in the set obtained by $\tilde{N} = \sum_{i=1}^N \sum_{c=1}^C l_i^c$. Note that $\{\mathbf{r}_i\}$ are unknown as the objective to be computed.

3. Problem Formulation

In this section, the above decomposition problem is solved by nonnegative data factorization. The presented solution appears to be efficient. Furthermore, the monotonic decrease of the objective function can be proved regarding the convergence properties.

With N training samples, to formulate the decomposition problem, we use such representations as follows: a matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N]$ to enclose the input holistic signal representations in columns; a $K \times \tilde{N}$ matrix $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_{\tilde{N}}]$ to enclose the decomposed representations for all the tags, which are also arranged in columns, where $\tilde{N}$ is the total number of tags in the training set. The decomposition problem formulation consists of the following two parts.

(1) The first part is minimizing the error of reconstruction from the decomposed representations to the additive signal representations; that is, for all $i \in \{1, \dots, N\}$,

$$\min \left\| \mathbf{v}_i - \sum_j \mathbf{r}_j \right\|^2, \quad (2)$$

if the j th tag is associated with the i th sample.

Considering that, in the above formulation, $\{\mathbf{r}_j\}$ are not normalized, we introduce the coefficients and yield:

$$\min \left\| \mathbf{v}_i - \sum_j \alpha_{ij} \mathbf{r}_j \right\|^2, \quad (3)$$

if the j th tag is associated with the i th sample,

where α_{ij} represents the coefficient of $\mathbf{r}_j$ in the reconstruction of $\mathbf{v}_i$. Using matrix form and an $\tilde{N} \times N$ matrix $\mathbf{H}$ with elements $\mathbf{H}_{ij} = \alpha_{ji}$ to enclose the information of coefficients, the reconstruction error minimization can be formulated as $\arg \min_{\mathbf{R}, \mathbf{H}} \|\mathbf{V} - \mathbf{R}\mathbf{H}\|^2$.

Note that a mixed holistic representation should be reconstructed only from representations of the associated labels. Therefore, some elements of $\mathbf{H}$ matrix must be 0; that is,

$$\mathbf{H}_{ij} = 0, \quad \text{if the } i\text{th tag is not associated with the } j\text{th sample.} \quad (4)$$

This information can be derived from the training labels. Here we release the 0-elements constraint for $\mathbf{H}$ matrix and introduce a membership matrix $\mathbf{Y}$:

$$\mathbf{Y}_{ij} = \begin{cases} 1 & \text{if } \mathbf{r}_i \text{ is associated to } \mathbf{v}_j, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Then the minimization problem becomes

$$\arg \min_{\mathbf{R}, \mathbf{H}} \|\mathbf{V} - \mathbf{R}(\mathbf{H} \bullet \mathbf{Y})\|^2. \quad (6)$$

Note that $\mathbf{Y}$ matrix will be abandoned later in the proposed solution when the 0-elements constraint for $\mathbf{H}$ matrix is already considered.

(2) The second part is minimizing intralabel distance between the decomposed signal representations of the same label. Here we use $\mathbf{S}$ matrix to denote the intralabel information with

$$\mathbf{S}_{ij} = \begin{cases} 1 & \text{if } i \neq j, \mathbf{r}_i \text{ and } \mathbf{r}_j \text{ correspond to the same label;} \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

To force the decomposed representations belonging to the same label to be similar to each other, we will minimize $\sum_{ij} \|\mathbf{r}_i - \mathbf{r}_j\|^2 \mathbf{S}_{ij}$ (here for convenience $\{\mathbf{r}_i\}$ are assumed as normalized, which will be described later), and we can derive in the matrix form

$$\begin{aligned} & \sum_{ij} \|\mathbf{r}_i - \mathbf{r}_j\|^2 \mathbf{S}_{ij} \sum_{ij} (\mathbf{r}_i^T \mathbf{r}_i + \mathbf{r}_j^T \mathbf{r}_j - 2\mathbf{r}_i^T \mathbf{r}_j) \mathbf{S}_{ij} \\ &= 2 \sum_{ij} \mathbf{r}_i^T \mathbf{r}_i - 2 \sum_{ij} \mathbf{r}_i^T \mathbf{S}_{ij} \mathbf{r}_j \\ &= 2\text{Tr}(\mathbf{R}\mathbf{D}\mathbf{R}^T) - 2\text{Tr}(\mathbf{R}\mathbf{S}\mathbf{R}^T) = 2\text{Tr}(\mathbf{R}\mathbf{L}\mathbf{R}^T), \end{aligned} \quad (8)$$

where

$$\begin{aligned} \mathbf{L} &= \mathbf{D} - \mathbf{S}, \\ \mathbf{D}_{ij} &= \begin{cases} \sum_k \mathbf{S}_{ik} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (9)$$

Representing the context among the decomposed local representations in a graph as shown in Figure 2, $\mathbf{L}$ matrix is therefore the Laplacian matrix for the graph, based on which graph preserving energy [7] (similar to (8)) can be derived.

Note that $\mathbf{R}$ may not be normalized. When measuring distance between different $\{\mathbf{r}_i\}$ in the graph based representation, we introduce $\mathbf{Q}$ for normalizing $\mathbf{R}$. Since $\mathbf{V}$ is known to be normalized in column, by minimizing the reconstruction error with (6), $\mathbf{R}(\mathbf{H} \bullet \mathbf{Y})$ can be considered to be approximately normalized. Thus $\mathbf{Q}$ is designed as a diagonal matrix:

$$\mathbf{Q} = \text{diag}(\mathbf{Y}\mathbf{H}^T \mathbf{e}), \quad (10)$$

FIGURE 2: An illustration showing the graph based representation which encodes the context among decomposed signals of multiple labels: each node represents the decomposed representation to be obtained for a specific tag; tags belonging to the same label are linked with edge weight 1, while edge weights between those of different labels are set to 0.

where $\mathbf{e} = [1, \dots, 1, \dots, 1]$ is a $\tilde{N}$ dimensional vector; that is,

$$\mathbf{Q}_{ij} = \begin{cases} \sum_k \mathbf{Y}_{ik} \mathbf{H}_{ik} & \text{if } i = j; \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

Multiplying $\{\mathbf{r}_i\}$ with $\{\mathbf{Q}_{ij}\}$, respectively, we can see that $\mathbf{RQ}$ can be approximately columnwise normalized. If we take into account the normalization factor $\mathbf{Q}$, the second factor of the objective function minimizing intralabel distance is then yielded as

$$\min \text{Tr}(\mathbf{RQ}\mathbf{LQ}^T\mathbf{R}^T). \quad (12)$$

Based on the above two parts, the presented decomposition problem can be formulated as a regularized nonnegative data factorization problem as

$$\arg \min_{\mathbf{R}, \mathbf{H}} \|\mathbf{V} - \mathbf{R}(\mathbf{H} \cdot \mathbf{Y})\|^2 + \lambda \text{Tr}(\mathbf{RQ}\mathbf{LQ}^T\mathbf{R}^T) \quad (13)$$

s.t. $\mathbf{R}, \mathbf{H} \geq 0$,

where λ is a weight to be empirically set in the experiments. The constraints show that $\mathbf{R}$ and $\mathbf{H}$ should be nonnegative to satisfy the reconstruction assumption, since in many real applications only when the signal representations and the reconstruction coefficients are nonnegative, can the proposed decomposition be physically meaningful. The decomposed label representations $\mathbf{R}$ can be obtained from the fission of $\mathbf{V}$ by solving (13). The solution will be given in the following. In the formulation, the given $\mathbf{V}$ is normalized, but there is no constraint that $\mathbf{R}$ or $\mathbf{H}$ matrix should be normalized. In the following optimization process we will normalize $\mathbf{H}$.

4. Optimizing Solution

The problem of (13) falls into the framework of nonnegative matrix factorization (NMF) [8] for its nonnegative constraints. NMF has shown its effectiveness in practical signal

processing for image classification [9]. Here similar to the optimization strategy adopted in the previous work by Lee and Seung [8], we optimize the objective function of (13) in an iterative way with a multiplicative updating rule, which guarantees the non-negativity. For initialization, to get an original guess of the target decomposed representations, we initialize $\mathbf{r}_i^c = \mathbf{v}_i \cdot (\bar{\mathbf{v}}^c / \sum_{c'=1, c' \neq i}^C \bar{\mathbf{v}}^{c'})$. Note that “ $\cdot$ ” stands for dot product in this paper, and $\bar{\mathbf{v}}^c$ is the average representation of c th class from training samples associated with label c .

Most iterative procedures for solving high-order optimization problems transform the original intractable problem into a set of tractable subproblems and finally obtain the convergence to a local optimum. The proposed iterative procedure also follows this philosophy and optimizes $\mathbf{H}$ and $\mathbf{R}$ alternately. The proposed iteration rules can be proved to monotonically decrease the objective function value, and the theoretical proof is given in Section 5. In fact, although stationary is the necessary condition for achieving local minima, it is difficult to justify that any limit points are the stationary point. Lin [10] has slightly modified the original iterative algorithm of NMF to achieve the convergence safely. Their modification, however, will increase the computational complexity while achieving his similar performance, as reported in [10]. Therefore, in this work, we do not adopt that the modification and experiment results show that the proposed solution could usually converge to local minima. We give the iterative procedure as follows.

(1) *Optimize $\mathbf{H}$ for Given $\mathbf{R}$.* For a fixed $\mathbf{R}$, the objective function in (13) with respect to $\mathbf{H}$ can be written as

$$F(\mathbf{H}) = \|\mathbf{V} - \mathbf{RH}\|^2 + \lambda \text{Tr}(\mathbf{R} \text{diag}(\mathbf{YH}^T \mathbf{e}) \mathbf{L} \text{diag}(\mathbf{YH}^T \mathbf{e}) \mathbf{R}^T) \quad (14)$$

s.t. $\mathbf{H} \geq 0$.

Let Φ_{ij} be the Lagrange multiplier for constraint $\mathbf{H}_{ij} \geq 0$; the Lagrange factor is $\sum_{ij} \Phi_{ij} \mathbf{H}_{ij} = \text{Tr}(\Phi \mathbf{H})$ and the Lagrange function $\mathcal{L}$ is

$$\begin{aligned} \mathcal{L} &= \|\mathbf{V} - \mathbf{RH}\|^2 + \lambda \text{Tr}(\mathbf{R} \text{diag}(\mathbf{YH}^T \mathbf{e}) \mathbf{L} \text{diag}(\mathbf{YH}^T \mathbf{e}) \mathbf{R}^T) \\ &\quad + \text{Tr}(\Phi \mathbf{H}^T) \\ &= \text{Tr}(\mathbf{V}\mathbf{V}^T) - 2\text{Tr}(\mathbf{VH}^T \mathbf{R}^T) + \text{Tr}(\mathbf{RHH}^T \mathbf{R}^T) \\ &\quad + \lambda \text{Tr}(\mathbf{R} \text{diag}(\mathbf{YH}^T \mathbf{e}) \mathbf{L} \text{diag}(\mathbf{YH}^T \mathbf{e}) \mathbf{R}^T) \\ &\quad + \text{Tr}(\Phi \mathbf{H}^T). \end{aligned} \quad (15)$$

The partial derivation of $\mathcal{L}$ with respect to $\mathbf{H}$ is

$$\frac{\partial \mathcal{L}}{\partial \mathbf{H}} = -2\mathbf{R}^T \mathbf{V} + 2\mathbf{R}^T \mathbf{R} \mathbf{H} + 2\lambda (\mathbf{L}(\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T \mathbf{H} \mathbf{Y}^T \mathbf{Y} + \Phi, \quad (16)$$

where we have used the following deductions:

$$\begin{aligned} O &= \mathbf{R} \text{diag}(\mathbf{YH}^T \mathbf{e}) \mathbf{L} \text{diag}(\mathbf{YH}^T \mathbf{e}) \mathbf{R}^T \\ &= \mathbf{R} (\mathbf{L} (\mathbf{YH}^T \mathbf{e} \mathbf{e}^T \mathbf{HY}^T)) \mathbf{R}^T \end{aligned} \quad (17)$$

$$\begin{aligned} &= \sum_i \mathbf{R}_i (\mathbf{L} (\mathbf{YH}^T \mathbf{e} \mathbf{e}^T \mathbf{HY}^T)) \mathbf{R}_i^T, \\ \frac{\partial O}{\partial \mathbf{H}} &= \sum_i (\mathbf{L}^T (\mathbf{R}_i^T \mathbf{R}_i)) 2\mathbf{e} \mathbf{e}^T \mathbf{HY}^T \mathbf{Y} \\ &= 2 (\mathbf{L} (\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T \mathbf{HY}^T \mathbf{Y}. \end{aligned} \quad (18)$$

Using the KarushKuhnTucker (KKT) condition [11] $\Phi_{ij} \mathbf{H}_{ij} = 0$, from (16), the following equations can be obtained for $\mathbf{H}_{ij}$:

$$\begin{aligned} &-(\mathbf{R}^T \mathbf{V})_{ij} \mathbf{H}_{ij} + (\mathbf{R}^T \mathbf{RH})_{ij} \mathbf{H}_{ij} \\ &+ \lambda ((\mathbf{L} \cdot (\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T \mathbf{HY}^T \mathbf{Y})_{ij} \mathbf{H}_{ij} \\ &= -(\mathbf{R}^T \mathbf{V})_{ij} \mathbf{H}_{ij} + (\mathbf{R}^T \mathbf{RH})_{ij} \mathbf{H}_{ij} \\ &+ \lambda ((\mathbf{D} (\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T \mathbf{HY}^T \mathbf{Y})_{ij} \mathbf{H}_{ij} \\ &- \lambda ((\mathbf{S} (\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T \mathbf{HY}^T \mathbf{Y})_{ij} \mathbf{H}_{ij} = 0, \end{aligned} \quad (19)$$

which leads to the update rule

$$\mathbf{H}_{ij} \leftarrow \mathbf{H}_{ij} \frac{[\mathbf{R}^T \mathbf{V} + \lambda ((\mathbf{S} (\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T \mathbf{HY}^T \mathbf{Y})]_{ij}}{[\mathbf{R}^T \mathbf{RH} + \lambda ((\mathbf{D} (\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T \mathbf{HY}^T \mathbf{Y})]_{ij}}. \quad (20)$$

Remember that $\mathbf{H}_{ij}^{(t)} = 0$ if $\mathbf{Y}_{ij} = 0$, where t means the t th iteration. Note that it could be satisfied by initializing $\mathbf{H}_{ij}^{(0)} = 0$ if $\mathbf{Y}_{ij} = 0$; therefore $\mathbf{Y}$ actually can be neglected without any influence. It is obvious that the updated $\mathbf{H}^{(t+1)}$ is still nonnegative if the matrices $\mathbf{R}$ and $\mathbf{H}^{(t)}$ are nonnegative. After getting $\mathbf{H}^{(t+1)}$, we normalize the column vectors of $\mathbf{H}^{(t+1)}$ and consequently convey the norms to the basis matrix; namely,

$$\mathbf{R} \leftarrow \mathbf{R} \text{diag}(\mathbf{YH}^T \mathbf{e}), \quad (21)$$

$$\mathbf{H} \leftarrow \mathbf{H} \text{diag}\left(\frac{1}{\mathbf{H}^T \mathbf{e}}\right). \quad (22)$$

The above updating of $\mathbf{R}$ and $\mathbf{H}$ does not change the value of the objective function in (13).

(2) *Optimize $\mathbf{R}$ for Given Normalized $\mathbf{H}$.* Then based on the normalized $\mathbf{H}$ in (22), the objective function in (13) with respect to $\mathbf{R}$ for given $\mathbf{H}$ can be written as

$$\begin{aligned} F(\mathbf{R}) &= \|\mathbf{V} - \mathbf{R}(\mathbf{H} \cdot \mathbf{Y})\|^2 + \lambda \text{Tr}(\mathbf{RLR}^T) \\ &= \|\mathbf{V} - \mathbf{RH}\|^2 + \lambda \text{Tr}(\mathbf{RLR}^T). \end{aligned} \quad (23)$$

Let Ψ_{ij} be the Lagrange multiplier for constraint $\mathbf{R}_{ij} \geq 0$, the Lagrange function $\mathcal{L}$ is

$$\begin{aligned} \mathcal{L} &= \|\mathbf{V} - \mathbf{RH}\|^2 + \lambda \text{Tr}(\mathbf{RLR}^T) + \text{Tr}(\Psi \mathbf{R}^T) \\ &= \text{Tr}(\mathbf{V}\mathbf{V}^T) - 2\text{Tr}(\mathbf{VH}^T \mathbf{R}^T) + \text{Tr}(\mathbf{RHH}^T \mathbf{R}^T) \\ &\quad + \lambda \text{Tr}(\mathbf{RLR}^T) + \text{Tr}(\Psi \mathbf{R}^T). \end{aligned} \quad (24)$$

The partial derivation of $\mathcal{L}$ with respect to $\mathbf{R}$ is

$$\frac{\partial \mathcal{L}}{\partial \mathbf{R}} = -2\mathbf{VH}^T + 2\mathbf{RHH}^T + 2\lambda \mathbf{RL} + \Psi. \quad (25)$$

Using the KKT condition $\Psi_{ij} \mathbf{R}_{ij} = 0$, from (25), we obtain the following equations for $\mathbf{R}_{ij}$:

$$-(\mathbf{VH}^T)_{ij} \mathbf{R}_{ij} + (\mathbf{RHH}^T)_{ij} \mathbf{R}_{ij} + \lambda (\mathbf{RL})_{ij} \mathbf{R}_{ij} = 0, \quad (26)$$

which leads to the following update rule:

$$\mathbf{R}_{ij} \leftarrow \mathbf{R}_{ij} \frac{(\mathbf{VH}^T + \lambda \mathbf{RS})_{ij}}{(\mathbf{RHH}^T + \lambda \mathbf{RD})_{ij}}. \quad (27)$$

5. Convergence Proof of the Proposed Solution

Definition 1. Function $G(\mathbf{A}, \mathbf{A}')$ is an auxiliary function for function $F(\mathbf{A})$ if the conditions

$$G(\mathbf{A}, \mathbf{A}') \geq F(\mathbf{A}), \quad G(\mathbf{A}, \mathbf{A}) = F(\mathbf{A}) \quad (28)$$

are satisfied.

Theorem 2. If G is an auxiliary function, then F is nonincreasing under the update

$$\mathbf{A}^{(t+1)} = \arg \min_{\mathbf{A}} G(\mathbf{A}, \mathbf{A}^{(t)}), \quad (29)$$

where t means the t th iteration.

5.1. Convergence of the Update Rule for $\mathbf{H}$. Firstly let us go back to (14) and consider one element $\mathbf{H}_{ab}$ of matrix $\mathbf{H}$. Using F_{ab} to denote the part of $F(\mathbf{H})$ relevant to $\mathbf{H}_{ab}$, we can get

$$\begin{aligned} F'_{ab} &= [-2\mathbf{R}^T \mathbf{V} + 2\mathbf{R}^T \mathbf{RH} + 2\lambda (\mathbf{L} (\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T \mathbf{HY}^T \mathbf{Y}]_{ab}; \\ F''_{ab} &= 2(\mathbf{R}^T \mathbf{R})_{aa} + 2\lambda [(\mathbf{L} (\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T]_{aa} (\mathbf{Y}^T \mathbf{Y})_{bb}. \end{aligned} \quad (30)$$

The auxiliary function of F_{ab} is designed as

$$\begin{aligned} G(\mathbf{H}_{ab}, \mathbf{H}_{ab}^{(t)}) &= F_{ab}(\mathbf{H}_{ab}^{(t)}) + F'_{ab}(\mathbf{H}_{ab}^{(t)}) (\mathbf{H}_{ab} - \mathbf{H}_{ab}^{(t)}) \\ &\quad + \frac{(\mathbf{R}^T \mathbf{RH})_{ab} + \lambda [(\mathbf{D} (\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T \mathbf{HY}^T \mathbf{Y}]_{ab}}{\mathbf{H}_{ab}^{(t)}} \\ &\quad \times (\mathbf{H}_{ab} - \mathbf{H}_{ab}^{(t)})^2. \end{aligned} \quad (31)$$

Lemma 3. Equation (32) is an auxiliary function for F_{ab} , the part of $F(\mathbf{H})$ only relevant to $\mathbf{H}_{ab}$.

Proof. Since $G(\mathbf{H}_{ab}, \mathbf{H}_{ab}) = F_{ab}(\mathbf{H}_{ab})$ is obvious, we need only show that $G(\mathbf{H}_{ab}, \mathbf{H}_{ab}^{(t)}) \geq F_{ab}(\mathbf{H}_{ab})$.

To this aim, we compare the Taylor series expansion of $F_{ab}(\mathbf{H}_{ab})$

$$F_{ab}(\mathbf{H}_{ab}) = F_{ab}(\mathbf{H}_{ab}^{(t)}) + F'_{ab}(\mathbf{H}_{ab}^{(t)})(\mathbf{H}_{ab} - \mathbf{H}_{ab}^{(t)}) + \frac{1}{2}F''_{ab}(\mathbf{H}_{ab} - \mathbf{H}_{ab}^{(t)})^2 \quad (32)$$

with (32) to find that $G(\mathbf{H}_{ab}, \mathbf{H}_{ab}^{(t)}) \geq F_{ab}(\mathbf{H}_{ab})$ is equivalent to

$$\frac{(\mathbf{R}^T \mathbf{R} \mathbf{H})_{ab} + \lambda[(\mathbf{D}(\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T \mathbf{H} \mathbf{Y}^T \mathbf{Y}]_{ab}}{\mathbf{H}_{ab}^T} \geq (\mathbf{R}^T \mathbf{R})_{aa} + \lambda[(\mathbf{L}(\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T]_{aa} (\mathbf{Y}^T \mathbf{Y})_{bb}, \quad (33)$$

where

$$\begin{aligned} (\mathbf{R}^T \mathbf{R} \mathbf{H})_{ab} &= \sum_k (\mathbf{R}^T \mathbf{R})_{ak} \mathbf{H}_{kb} \geq (\mathbf{R}^T \mathbf{R} \mathbf{H})_{aa} \mathbf{H}_{ab}^{(t)}, \\ \lambda[(\mathbf{D}(\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T \mathbf{H} \mathbf{Y}^T \mathbf{Y}]_{ab} &= \lambda \sum_k [(\mathbf{D}(\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T]_{ak} \mathbf{H}_{kb} (\mathbf{Y}^T \mathbf{Y})_{bb} \\ &\geq \lambda[(\mathbf{D}(\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T]_{aa} \mathbf{H}_{ab} (\mathbf{Y}^T \mathbf{Y})_{bb} \\ &\geq \lambda[(\mathbf{D} - \mathbf{S})(\mathbf{R}^T \mathbf{R}) \mathbf{e} \mathbf{e}^T]_{aa} \mathbf{H}_{ab} (\mathbf{Y}^T \mathbf{Y})_{bb} \\ &= \lambda[(\mathbf{L}(\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T]_{aa} \mathbf{H}_{ab} (\mathbf{Y}^T \mathbf{Y})_{bb}. \end{aligned} \quad (34)$$

Thus, (33) holds and $G(\mathbf{H}_{ab}, \mathbf{H}_{ab}^{(t)}) \geq F_{ab}(\mathbf{H}_{ab})$. $\square$

Lemma 4. Equation (20) could be obtained by minimizing the auxiliary function $G(\mathbf{H}_{ab}, \mathbf{H}_{ab}^{(t)})$.

Proof. By setting $\partial G(\mathbf{H}_{ab}, \mathbf{H}_{ab}^{(t)}) / \partial \mathbf{H}_{ab} = 0$, we have

$$\begin{aligned} \frac{\partial G(\mathbf{H}_{ab}, \mathbf{H}_{ab}^{(t)})}{\partial \mathbf{H}_{ab}} &= F'_{ab}(\mathbf{H}_{ab}^{(t)}) \\ &+ \frac{2[\mathbf{R}^T \mathbf{R} \mathbf{H} + \lambda((\mathbf{D}(\mathbf{R}^T \mathbf{R})) \mathbf{e} \mathbf{e}^T \mathbf{Y}^T \mathbf{Y})]_{ab}}{\mathbf{H}_{ab}^{(t)}} \\ &\times (\mathbf{H}_{ab} - \mathbf{H}_{ab}^{(t)}) = 0. \end{aligned} \quad (35)$$

Then the update rule for $\mathbf{H}$ in (20) can be obtained. $\square$

5.2. Convergence of the Update Rule for $\mathbf{R}$. Let us go back to (23) and consider any element $\mathbf{R}_{ab}$ of matrix $\mathbf{R}$. Now F_{ab} is

used to denote the part of $F(\mathbf{R})$ relevant to $\mathbf{R}_{ab}$. It is easy to check that

$$\begin{aligned} F'_{ab} &= (-2\mathbf{V} \mathbf{H}^T + 2\mathbf{R} \mathbf{H} \mathbf{H}^T + 2\lambda \mathbf{R} \mathbf{L})_{ab}, \\ F''_{ab} &= 2(\mathbf{H} \mathbf{H}^T)_{bb} + 2\lambda \mathbf{L}_{bb}. \end{aligned} \quad (36)$$

The auxiliary function of F_{ab} is designed as

$$\begin{aligned} G(\mathbf{R}_{ab}, \mathbf{R}_{ab}^{(t)}) &= F_{ab}(\mathbf{R}_{ab}^{(t)}) + F'_{ab}(\mathbf{R}_{ab}^{(t)})(\mathbf{R}_{ab} - \mathbf{R}_{ab}^{(t)}) \\ &+ \frac{(\mathbf{R} \mathbf{H} \mathbf{H}^T)_{ab} + \lambda(\mathbf{R} \mathbf{D})_{ab}}{\mathbf{R}_{ab}^T} (\mathbf{R}_{ab} - \mathbf{R}_{ab}^{(t)})^2. \end{aligned} \quad (37)$$

Lemma 5. Equation (37) is an auxiliary function for F_{ab} , the part of $F(\mathbf{R})$ only relevant to $\mathbf{R}_{ab}$.

Proof. Since $G(\mathbf{R}_{ab}, \mathbf{R}_{ab}) = F_{ab}(\mathbf{R}_{ab})$ is obvious, we need only show that $G(\mathbf{R}_{ab}, \mathbf{R}_{ab}^{(t)}) \geq F_{ab}(\mathbf{R}_{ab})$. To do this, we compare the Taylor series expansion of $F_{ab}(\mathbf{R}_{ab})$

$$\begin{aligned} F_{ab}(\mathbf{R}_{ab}) &= F_{ab}(\mathbf{R}_{ab}^{(t)}) + F'_{ab}(\mathbf{R}_{ab}^{(t)})(\mathbf{R}_{ab} - \mathbf{R}_{ab}^{(t)}) \\ &+ \frac{1}{2}F''_{ab}(\mathbf{R}_{ab} - \mathbf{R}_{ab}^{(t)})^2 \end{aligned} \quad (38)$$

with (37), and $G(\mathbf{R}_{ab}, \mathbf{R}_{ab}^{(t)}) \geq F_{ab}(\mathbf{R}_{ab})$ is equivalent to

$$\frac{(\mathbf{R} \mathbf{H} \mathbf{H}^T)_{ab} + \lambda(\mathbf{R} \mathbf{D})_{ab}}{\mathbf{R}_{ab}^T} \geq (\mathbf{H} \mathbf{H}^T)_{bb} + \lambda \mathbf{L}_{bb}, \quad (39)$$

where

$$(\mathbf{R} \mathbf{H} \mathbf{H}^T)_{ab} = \sum_k \mathbf{R}_{ak}^{(t)} (\mathbf{H} \mathbf{H}^T)_{kb} \geq \mathbf{R}_{ab}^{(t)} (\mathbf{H} \mathbf{H}^T)_{bb},$$

$$\lambda(\mathbf{R} \mathbf{D})_{ab} = \lambda \sum_k \mathbf{R}_{ak}^{(t)} \mathbf{D}_{kb} \geq \lambda \mathbf{R}_{ab}^{(t)} \mathbf{D}_{bb} \geq \lambda \mathbf{R}_{ab}^{(t)} (\mathbf{D} - \mathbf{S})_{bb} \quad (40)$$

$$= \lambda \mathbf{R}_{ab}^{(t)} \mathbf{L}_{bb}.$$

Thus, (39) holds and $G(\mathbf{R}_{ab}, \mathbf{R}_{ab}^{(t)}) \geq F_{ab}(\mathbf{R}_{ab})$. $\square$

Lemma 6. Equation (27) could be obtained by minimizing the auxiliary function $G(\mathbf{R}_{ab}, \mathbf{R}_{ab}^{(t)})$.

Proof. By setting $\partial G(\mathbf{R}_{ab}, \mathbf{R}_{ab}^{(t)}) / \partial \mathbf{R}_{ab} = 0$, we can get

$$\begin{aligned} \frac{\partial G(\mathbf{R}_{ab}, \mathbf{R}_{ab}^{(t)})}{\partial \mathbf{R}_{ab}} &= F'_{ab}(\mathbf{R}_{ab}^{(t)}) \\ &+ \frac{2(\mathbf{R} \mathbf{H} \mathbf{H}^T + \lambda \mathbf{R} \mathbf{D})_{ab}}{\mathbf{R}_{ab}^{(t)}} (\mathbf{R}_{ab} - \mathbf{R}_{ab}^{(t)}) = 0. \end{aligned} \quad (41)$$

Then the update rule for $\mathbf{R}$ in (27) can be obtained. $\square$

FIGURE 3: Exemplar results of multilabel image classification on the KAIST campus images using the proposed decomposed training samples: corresponding regions of the correctly detected labels are marked in the images.

6. Experiment

A typical application of the proposed decomposition in real life is as follows: for a set of images related to a scene such as campus and each associated with multiple labels, we can decompose the image feature representations to pure label representations, to provide better label samples for modeling object categories, which therefore can facilitate object recognition, detection, and so on. It leads to a useful tool for automatic labeling of online images or personal photo album. There are two main reasons that the proposed decomposition is necessary here. First, with limited training samples, it is difficult to learn good label models using the original image presentations directly without decomposition. Since each sample is a mixture of several labels and the background for each label is various, it is better to decompose the image representation to labels optimally. Second, instances of the same label related to a scene usually have short intrasubspace distance, which supports the assumption of the proposed decomposition and guarantees the effectiveness.

To demonstrate a practical application of the proposed decomposition in automatic image annotation, we manually crawled images of “KAIST campus” from the Flickr photo sharing website [12], excluded the irrelevant images, and obtained 183 images (we share the data at http://rcv.kaist.ac.kr/~tengli/Teng_resource.html). They are labeled with nine concepts: “sky,” “plant (grass/tree/flower),” “water,” “ground,” “man-made (building/sign/road),” “car/bus,” “bicycle,” “person,” and “animal.” 150 images are used for training samples, which are used for class models training and can be applied to the proposed decomposition, and the rest are used for testing. Labels are highly mixed in the images.

In the implementation, the images and regions are represented using the Bag-of-Words (BoW) feature [13], where the number of visual words is set as 500 in the experiments for the tradeoff of classification accuracy and computational cost, and the feature dimension is therefore 500. With the decomposed label representations from the N training images, that is, normalized $\mathbf{R}$ matrix, a typical classifier, support vector machine (SVM) [14] is applied to learn label models and classify the testing samples. In the experiment

TABLE 1: Comparison of classification performance (AUC) on the KAIST campus images.

Method	BoW + SVM	Proposed + SVM	Proposed + SRC
AUC	0.5	0.77	0.76

of multilabel image classification, we adopted the area under ROC curve (AUC) measurement [15].

Table 1 gives the performance comparison between the image classification adopting the proposed nonnegative signal decomposition and the classic BoW method [13]. To further examine the performance, except for the well-known SVM, we also apply a recent popular Sparse Representation Classification (SRC) [16] as the classifier. To train using the original BoW image representations without applying decomposition yield AUC as low as around 0.5, whereas, to train label models using the decomposed results, the AUCs on the test samples are 0.77 and 0.76, in combination with SVM and SRC, respectively. Figure 3 gives the multilabel classification results for some sample images of KAIST campus Flickr dataset. The results appear to be promising.

7. Conclusion

A new technique for the nonnegative signals decomposition considering the supervised additive label information has been developed in this paper, with the assumption that intralabel signals should be relatively closer to each other than interlabel signals. The nonnegative constraint makes it applicable for many practical cases. We formulated it as an optimization problem and proposed an efficient iterative solution based on nonnegative data factorization. Convergence analysis showed that the solution guarantees monotonic decrease of the objective function.

References

- [1] J. F. Cardoso, “Source separation using higher order moments,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2109–2112, May 1989.

- [2] P. Comon, "Separation of sources using higher-order cumulants," in *Advanced Algorithms and Architectures for Signal Processing*, Proceedings of SPIE, pp. 170–181, 1989.
- [3] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.
- [4] S. Peng and W. L. Hwang, "Null space pursuit: an operator-based approach to adaptive signal separation," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2475–2483, 2010.
- [5] J. Shi and J. Malik, "Normalized cuts and image segmentation," in *Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 731–737, June 1997.
- [6] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu, "Image parsing: unifying segmentation, detection, and recognition," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 113–140, 2005.
- [7] S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.
- [8] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
- [9] C. Lang, S. Feng, B. Cheng, B. Ni, and S. Yan, "A unified supervised codebook learning framework for classification," *Neurocomputing*, vol. 77, no. 1, pp. 281–288, 2012.
- [10] C. J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Transactions on Neural Networks*, vol. 18, no. 6, pp. 1589–1596, 2007.
- [11] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," in *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*, pp. 481–492, University of California Press, Berkeley, Calif, USA, 1951.
- [12] Flickr, <http://www.flickr.com/>.
- [13] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: a comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.
- [14] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
- [15] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
- [16] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.