Xiao YH, Zhu ZF, Zhao Y *et al.* Class-driven non-negative matrix factorization for image representation. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 28(5): 751–761 Sept. 2013. DOI 10.1007/s11390-013-1374-9

Class-Driven Non-Negative Matrix Factorization for Image Representation

Yan-Hui Xiao (肖延辉), Zhen-Feng Zhu (朱振峰), Yao Zhao* (赵　耀), and Yun-Chao Wei (魏云超)

Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China

E-mail: xiaoyanhui@gmail.com; {zhfzhu, yzhao, 11112065}@bjtu.edu.cn

Received May 5, 2013; revised August 6, 2013.

Abstract Non-negative matrix factorization (NMF) is a useful technique to learn a parts-based representation by decomposing the original data matrix into a basis set and coefficients with non-negative constraints. However, as an unsupervised method, the original NMF cannot utilize the discriminative class information. In this paper, we propose a semi-supervised class-driven NMF method to associate a class label with each basis vector by introducing an inhomogeneous representation cost constraint. This constraint forces the learned basis vectors to represent better for their own classes but worse for the others. Therefore, data samples in the same class will have similar representations, and consequently the discriminability in new representations could be boosted. Some experiments carried out on several standard databases validate the effectiveness of our method in comparison with the state-of-the-art approaches.

Keywords class-driven, non-negative matrix factorization, data clustering, image representation

1 Introduction

Data representation is a fundamental problem in image processing and pattern recognition tasks. A good representation can typically reveal the latent structure of data, and further facilitate these tasks[1-3]. However, in real applications, the input data matrix is generally of very high dimension, which makes learning from example infeasible. To solve this problem, matrix factorization approaches are used to explore two or more lower dimensional matrices whose product provides a good approximation for the original data matrix. For example, singular value decomposition (SVD) and principal component analysis (PCA) decompose the original matrix as the linear combination of principle components.

In recent years, non-negative matrix factorization (NMF)[4] has become popular for data representation owning to its theoretical interpretation and practical performance. Several studies[5-6] have shown that there is psychological and physiological evidence for parts-based representation in human brain. While NMF with non-negative constraints could obtain a parts-based representation since there are only additive, not subtractive, combinations. Specifically, it models data as a linear combination of a set of basis vectors, and both the combination coefficients and the basis vectors are non-negative. For example, a face image can be represented by an additive combination of several versions of mouth, nose, eyes, and other facial parts. In addition, NMF has shown performance superior to PCA and SVD in face recognition[7] and document clustering[8].

Several NMF variants have been developed by integrating additional constraints into the original NMF. Xu and Gong[9] presented a concept factorization (CF) approach which expands NMF to the data containing negative values and can be implemented in the kernel space. To consider the geometric structure in the data, Cai *et al.*[10] presented a graph regularized NMF (GNMF) method. GNMF leads to a new parts-based data representation which respects the geometrical structure of the data space. However, the above NMF approaches ignore the discriminative label information as an unsupervised learning algorithm.

In many real world applications, such as text categorization[11] and data clustering[12], a small amount of labeled data could be used to aid and bias the learning of unlabeled data. Thus, semi-supervised

Regular Paper

This work was supported in part by the National Basic Research 973 Program of China under Grant No. 2012CB316400, the National Natural Science Foundation of China under Grant Nos. 61025013, 61172129, 61210006, the Fundamental Research Funds for the Central Universities of China under Grant No. 2012JBZ012, and the Program for Changjiang Scholars and Innovative Research Team in University of China under Grant No. IRT201206.

*Corresponding Author

GNMF[13] was suggested by incorporating label information into graph structure. Nevertheless, there was no theoretical guarantee that the same class data points would be projected together into the parts-based representation space, and it was still unknown that how to determine the weights in a principled manner. To overcome this limitation, Liu *et al.* developed a constrained NMF method (CNMF)[13], which imposes the label information to the objective function as hard constraints. Mathematically, given the label constraint matrix $\boldsymbol{A}$, CNMF is to find two non-negative matrix factors $\boldsymbol{W}$ and $\boldsymbol{S}$ where the product of the factors $\boldsymbol{W}$, $\boldsymbol{A}$ and $\boldsymbol{S}$ is an approximation of the original matrix $\boldsymbol{X}$ (i.e., $\boldsymbol{X} = \boldsymbol{W}(\boldsymbol{A}\boldsymbol{S})^{\mathrm{T}}$). However, since CNMF mapps the images with the same label onto the same point, it is infeasible when there is only one labeled training example to rely on. That is to say, if there is only one labeled sample, the label matrix $\boldsymbol{A}$ will become an identity matrix and the label constraint will fail to work. In some real-world applications, there are a lot of unlabeled data (such as web pages on the Internet), but there is only one labeled example (such as the current interesting web page). Additionally, since the new representation based on NMF is an additive combination of a set of basis vectors (i.e., parts), the aforementioned NMF approaches fail to consider the correlation between the basis vectors and class labels.

To overcome the above problems, we propose a semi-supervised class-driven NMF method for image representation, named cdNMF. Inspired by [14], we associate a class label with each basis vector by introducing an inhomogeneous representation cost constraint. This constraint leads to learn a set of discriminative basis vectors which are enforced to represent better for their own classes but worse for the others. By minimizing the inhomogeneous representation cost, we can learn the basis even with one labeled example. Thus, data samples belonging to the same class will have similar representations, and the obtained new representations can have more discriminative power. In addition, with non-negative constraints, we can evaluate the inhomogeneous representation cost straightforwardly, which is more simple and faster than using L_2-norm in [14]. Furthermore, we utilize both Frobenius norm and KL-divergence to measure the reconstruction cost with the corresponding update rules.

2 Brief Review of NMF

As a matrix factorization algorithm, non-negative matrix factorization (NMF)[4] is utilized to decompose the original data matrix into a basis set and coefficients where the basis and coefficients are assumed to be non-negative. Mathematically, given a data matrix $\boldsymbol{X} = [x_{ij}] = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{m\times n}$, NMF aims to find two non-negative matrices $\boldsymbol{W} = [w_{ik}] \in \mathbb{R}^{m\times t}$ and $\boldsymbol{S} = [s_{jk}] = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times t}$ to approximate the original matrix as follows

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{S}^{\mathrm{T}},$$

where each row $\boldsymbol{s}_j$ of $\boldsymbol{S}$ is a coefficient vector corresponding to the sample vector $\boldsymbol{x}_j$, i.e., the column of $\boldsymbol{X}$. Therefore, the data $\boldsymbol{x}_j$ could be approximated by a linear combination of the columns of $\boldsymbol{W}$ with the coefficient $\boldsymbol{s}_j$. Thus, $\boldsymbol{W}$ and $\boldsymbol{S}$ can be regarded as a basis set and coefficients, respectively. To quantify the quality of the approximation, a cost function can be constructed by some measures of distance. One popular measure is the Euclidean distance (i.e., Frobenius norm).

$$O_F = ||\boldsymbol{X} - \boldsymbol{W}\boldsymbol{S}^{\mathrm{T}}||_F^2. \tag{1}$$

Although the objective function O_F in (1) is not convex with $\boldsymbol{W}$ and $\boldsymbol{S}$ together, the following alternating algorithm[15] converges to a local minimum.

$$w_{ik} \leftarrow w_{ik}\frac{(\boldsymbol{X}\boldsymbol{S})_{ik}}{(\boldsymbol{W}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S})_{ik}},$$

$$s_{jk} \leftarrow s_{jk}\frac{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W})_{jk}}{(\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})_{jk}}. \tag{2}$$

The other measure is not symmetric and referred as "divergence" of $\boldsymbol{X}$ from $\boldsymbol{Y} = [y_{ij}] = \boldsymbol{W}\boldsymbol{S}^{\mathrm{T}}$ instead of distance between $\boldsymbol{X}$ and $\boldsymbol{Y}$. When $\sum_{ij} x_{ij} = \sum_{ij} y_{ij} = 1$, the "divergence" reduces to the Kullback-Leibler divergence (KL-divergence), or relative entropy. Thus, we can obtain the following objective function

$$O_{\mathrm{KL}} = D(\boldsymbol{X}||\boldsymbol{Y}) = \sum_{i,j}\Big(x_{ij}\log\frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij}\Big). \tag{3}$$

In addition, an iterative update algorithm[15] is presented as follows.

$$w_{ik} \leftarrow w_{ik}\frac{\sum_j \Big(x_{ij}s_{jk}\Big/\sum_k w_{ik}s_{jk}\Big)}{\sum_j s_{jk}},$$

$$s_{jk} \leftarrow s_{jk}\frac{\sum_i \Big(x_{ij}w_{ik}\Big/\sum_k w_{ik}s_{jk}\Big)}{\sum_i w_{ik}}. \tag{4}$$

The above update steps would find a local minimum of the objective function O_{KL}[15]. In real applications, we generally have $t \ll m$ and $t \ll n$. Thus, NMF is to explore a compressed approximation of the original data matrix.

3 Semi-Supervised NMF with Inhomogeneous Representation Cost

The original NMF is an unsupervised method and cannot utilize the label information. In this section, we introduce a semi-supervised class-driven NMF method (cdNMF), which employs an inhomogeneous representation cost constraint to associate class label with each basis vector. Additionally, this constraint forces the learned basis vectors to represent better for their own classes but worse for the others.

To begin with some definitions, let $\boldsymbol{X} = [x_{ij}] = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{l+u}] \in \mathbb{R}^{m\times n}(l+u=n)$ denote a given dataset, where the first l samples are labeled data and the remaining u ones are unlabeled, usually $l \ll u$. In addition, the sample $\boldsymbol{x}_j$ is labeled as $b_j \in \{1, \ldots, c\}$ where c is the total number of classes. Our goal is to learn a discriminative basis set $\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W_2}, \ldots, \boldsymbol{W}_c] \in \mathbb{R}^{m\times t}$, where $\boldsymbol{W}_i \in \mathbb{R}^{m\times r}$ is the basis subset that can sparsely represent the i-th class well but not others, r is the number of basis vectors of each subset and $t = r \times c$.

Denote $\boldsymbol{S} = [s_{jk}] = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times t}$ as coefficient matrix and $\boldsymbol{D} = [d_{jk}] = [\boldsymbol{d}_{b_1}, \ldots, \boldsymbol{d}_{b_n}]^{\mathrm{T}} \in \mathbb{R}^{n\times t}$ as indicator matrix for the inhomogeneous representation. Our goal is to utilize class information to learn discriminative basis vectors, which represent better for their own classes but worse for the others. Thus, we hope that approximated parts-based representation $\boldsymbol{s}_j$ for sample $\boldsymbol{x}_j$ with label b_j in (1) and (3) will have the following property:

$$\boldsymbol{d}_{b_j}^{\mathrm{T}} \boldsymbol{s}_j = 0, \tag{5}$$

where $\boldsymbol{d}_{b_j}$ selects the inhomogeneous representation coefficients of $\boldsymbol{s}_j$, i.e., coefficients corresponding to basis vectors other than $\boldsymbol{W}_{b_j}$. For example, assuming $\boldsymbol{W} = [\boldsymbol{W}_1, \ldots, \boldsymbol{W}_3]$, $\boldsymbol{W}_i \in \mathbb{R}^{m\times 2}$ (i.e., $t = 6$), there are n data samples among which $\boldsymbol{x}_1$ belongs to the 1st class, $\boldsymbol{x}_2$ belongs to the 2nd class, $\boldsymbol{x}_3$ belongs to the 3rd class, and the other $n-3$ samples are unlabeled, i.e., $l = 3$ and $u = n-3$. Thus, the indicator matrix $\boldsymbol{D}^{\mathrm{T}} = [\boldsymbol{d}_{b_1}, \boldsymbol{d}_{b_2}, \boldsymbol{d}_{b_3}, \ldots, \boldsymbol{d}_{b_n}]$ can be defined as

$$\begin{bmatrix} \overbrace{0 & 1 & 1}^{l} & \overbrace{0 & \cdots & 0}^{u} \\ 0 & 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 \end{bmatrix},$$

where if $\boldsymbol{x}_j$ (such as $\boldsymbol{x}_4$) is the unlabeled sample, we set all the elements in $\boldsymbol{d}_{b_j}$ (such as $\boldsymbol{d}_{b_4}$) to 0. For convenience, we term the (5) as the inhomogeneous representation cost. (5) shows that the parts-based representation $\boldsymbol{s}_j$ in terms of basis matrix $\boldsymbol{W}$ will only concentrate on the basis subset $\boldsymbol{W}_{b_j}$. The ideal basis matrix $\boldsymbol{W}$ should consist of basis vectors where each subset $\boldsymbol{W}_{b_j}$ can represent data samples from the b_j-th class rather than others class.

4 Class-Driven NMF Based on Frobenius Norm Cost

Based on the Frobenius norm, we incorporate the inhomogeneous representation cost into the function (1):

$$\begin{aligned} O_F &= ||\boldsymbol{X} - \boldsymbol{W}\boldsymbol{S}^{\mathrm{T}}||_F^2 + \lambda \sum_j \boldsymbol{d}_{b_j}^{\mathrm{T}} \boldsymbol{s}_j \\ &= Tr((\boldsymbol{X} - \boldsymbol{W}\boldsymbol{S}^{\mathrm{T}})(\boldsymbol{X} - \boldsymbol{W}\boldsymbol{S}^{\mathrm{T}})^{\mathrm{T}}) + \lambda Tr(\boldsymbol{D}\boldsymbol{S}^{\mathrm{T}}) \\ &= Tr(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}) + Tr(\boldsymbol{W}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}) - \\ &\quad 2Tr(\boldsymbol{X}\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}) + \lambda Tr(\boldsymbol{D}\boldsymbol{S}^{\mathrm{T}}), \end{aligned}$$

where $Tr(\cdot)$ denotes the trace of a matrix, $\lambda \geqslant 0$ is the regularization parameter and the steps of derivation employ the matrix property $Tr(\boldsymbol{A}\boldsymbol{B}) = Tr(\boldsymbol{B}\boldsymbol{A})$ and $Tr(\boldsymbol{B}) = Tr(\boldsymbol{B}^{\mathrm{T}})$.

4.1 Multiplicative Update Rules Formulation

Given $\boldsymbol{\Phi} = [\phi_{ik}] \in \mathbb{R}^{m\times t}$ and $\boldsymbol{\Psi} = [\varphi_{jk}] \in \mathbb{R}^{n\times t}$, denote ϕ_{ik} and φ_{jk} as the Lagrange multipliers for constraint $w_{ik} \geqslant 0$ and $s_{jk} \geqslant 0$. Thus, the Lagrange $\boldsymbol{L}$ is as follows:

$$\begin{aligned} \boldsymbol{L} = Tr(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}) + Tr(\boldsymbol{W}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}) - 2Tr(\boldsymbol{X}\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}) + \\ \lambda Tr(\boldsymbol{D}\boldsymbol{S}^{\mathrm{T}}) + Tr(\boldsymbol{\Phi}\boldsymbol{W}^{\mathrm{T}}) + Tr(\boldsymbol{\Psi}\boldsymbol{S}^{\mathrm{T}}). \end{aligned} \tag{6}$$

With respect to $\boldsymbol{W}$ and $\boldsymbol{S}$, the partial derivatives of $\boldsymbol{L}$ are

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{W}} = -2\boldsymbol{X}\boldsymbol{S} + 2\boldsymbol{W}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S} + \boldsymbol{\Phi}, \tag{7}$$

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{S}} = -2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W} + 2\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + \lambda\boldsymbol{D} + \boldsymbol{\Psi}. \tag{8}$$

By utilizing the KKT conditions $\phi_{ik}w_{ik} = 0$ and $\varphi_{jk}s_{jk} = 0$, we obtain the following equations for w_{ik} and s_{jk}.

$$-(\boldsymbol{X}\boldsymbol{S})_{ik}w_{ik} + (\boldsymbol{W}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S})_{ik}w_{ik} = 0, \tag{9}$$

$$-(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W})_{jk}s_{jk} + (\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})_{jk}s_{jk} + \frac{1}{2}(\lambda\boldsymbol{D})_{jk}s_{jk} = 0. \tag{10}$$

(9) and (10) lead to the following update rules:

$$w_{ik} \leftarrow w_{ik}\frac{(\boldsymbol{X}\boldsymbol{S})_{ik}}{(\boldsymbol{W}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S})_{ik}}, \tag{11}$$

$$s_{jk} \leftarrow s_{jk}\frac{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W})_{jk}}{\Big(\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + (\lambda\boldsymbol{D})/2\Big)_{jk}}. \tag{12}$$

Regarding the update rules (11) and (12), we have the following theorem.

Theorem 1. *The objective function O_F of cdNMF in* (6) *is nonincreasing under the update rules in* (11) *and* (12). *The objective function is invariant under these updates if and only if $\boldsymbol{W}$ and $\boldsymbol{S}$ are at a stationary point.*

Theorem 1 guarantees the convergence under the update rules of $\boldsymbol{W}$ and $\boldsymbol{S}$, i.e., (11) and (12), and the final solution will be a local optimum. The proof of Theorem 1 is given in the following.

4.2 Proof of Convergence

In order to prove Theorem 1, the cost function O_F of cdNMF should be demonstrated to be nonincreasing under the update steps in (11) and (12). Meanwhile we has exactly the same update formula for $\boldsymbol{W}$ in (11) as the original NMF[15]. In addition, (12) is only related to $\boldsymbol{S}$. Thus, we just consider that O_F is nonincreasing under the second update step in (12).

To prove the convergence of O_F with (12), we employ the following property of an auxiliary function similar to that used in the Expectation Maximization algorithm[16].

Lemma 1. *If G is an auxiliary function of F, i.e., $G(s,s') \geqslant F(s)$ and $G(s,s) = F(s)$, then F is nonincreasing under the update.*

$$s^{(q+1)} = \arg\min_{s} G(s, s^{(q)}). \tag{13}$$

Proof.

$$F(s^{(q+1)}) \leqslant G(s^{(q+1)}, s^{(q)}) \leqslant G(s^{(q)}, s^{(q)}) = F(s^{(q)}).$$

Notice that $F(s^{(q+1)}) = F(s^{(q)})$ holds only if $s^{(q)}$ is a local minimum of $G(s, s^{(q)})$. □

Now we will show that the update step for $\boldsymbol{S}$ in (12) is exactly the update in (13) with a proper auxiliary function G.

We rewrite the objective function O_F of cdNMF in (6) as follows:

$$\begin{aligned} O_F &= ||\boldsymbol{X} - \boldsymbol{W}\boldsymbol{S}^{\mathrm{T}}||_F^2 + \lambda \sum\nolimits_j \boldsymbol{d}_{b_j}^{\mathrm{T}} \boldsymbol{s}_j \\ &= \sum_{i,j} \Big(x_{ij} - \sum\nolimits_k w_{ik} s_{jk} \Big)^2 + \lambda \sum\nolimits_{j,k} d_{jk} s_{jk}. \end{aligned} \tag{14}$$

Since the update is essentially element-wise, we use F_{jk} to denote the part of O_F which is only relevant to the element s_{jk} in $\boldsymbol{S}$. Thus, we have

$$\begin{aligned} F'_{jk} &= \Big(\frac{\partial O_F}{\partial \boldsymbol{S}} \Big)_{jk} \\ &= (-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W} + 2\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + \lambda \boldsymbol{D})_{jk} \end{aligned} \tag{15}$$

and

$$F''_{jk} = (2\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})_{kk}. \tag{16}$$

Lemma 2. *Function*

$$\begin{aligned} G(s, s_{jk}^{(q)}) = {} & F_{jk}(s_{jk}^{(q)}) + F'_{jk}(s_{jk}^{(q)})(s - s_{jk}^{(q)}) + \\ & \frac{(\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})_{jk} + \frac{1}{2}\lambda(\boldsymbol{D})_{jk}}{s_{jk}^{(q)}} (s - s_{jk}^{(q)})^2 \end{aligned} \tag{17}$$

is an auxiliary function for F_{jk}, the part of O_F which is only relevant to s_{jk}.

Proof. Since $G(s,s) = F_{jk}(s)$ is obvious, we only need to show that $G(s, s_{jk}^{(q)}) \geqslant F_{jk}(s)$. To do this, we compare the Taylor series expansion of $F_{jk}(s)$:

$$\begin{aligned} F_{jk}(s) = {} & F_{jk}(s_{jk}^{(q)}) + F'_{jk}(s_{jk}^{(q)})(s - s_{jk}^{(q)}) + \\ & (\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})_{kk}(s - s_{jk}^{(q)})^2 \end{aligned} \tag{18}$$

with (17) to find that $G(s, s_{jk}^{(q)}) \geqslant F_{jk}(s)$ is equivalent to

$$\frac{(\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})_{jk} + \frac{1}{2}\lambda(\boldsymbol{D})_{jk}}{s_{jk}^{(q)}} \geqslant (\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})_{kk}. \tag{19}$$

It is easy to check that

$$(\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})_{jk} = \sum_{l=1}^{t} s_{jl}^{(q)} (\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})_{lk} \geqslant s_{jk}^{(q)} (\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})_{kk} \tag{20}$$

and $\lambda(D)_{jk} \geqslant 0$. Therefore, (19) holds and we have $G(s, s_{jk}^{(q)}) \geqslant F_{jk}(s)$. □

Now we can prove the convergence of Theorem 1.

Proof of Theorem 1. Replacing $G(s, s_{jk}^{(q)})$ in (13) by (17) leads to the update rule

$$\begin{aligned} s_{jk}^{(q+1)} &= s_{jk}^{(q)} - s_{jk}^{(q)} \frac{F'_{jk}(s_{jk}^{(q)})}{2(\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})_{jk} + \lambda(\boldsymbol{D})_{jk}} \\ &= s_{jk}^{(q)} \frac{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W})_{jk}}{\Big(\boldsymbol{S}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + \frac{1}{2}\lambda \boldsymbol{D}\Big)_{jk}}. \end{aligned}$$

According to Lemma 1, F_{jk} is nonincreasing under this update rule. □

5 Class-Driven NMF Based on KL-Divergence Cost

Based on the KL-divergence, we incorporate the inhomogeneous representation cost into the cost function (3), as follows

$$O_{\mathrm{KL}} = \sum_{i,j}\Big(x_{ij}\log\frac{x_{ij}}{\sum\limits_k w_{ik}s_{jk}} - x_{ij} + \sum_k w_{ik}s_{jk} + \sum_k \gamma d_{jk}s_{jk}\Big), \tag{21}$$

where $\gamma \geqslant 0$ is the regularization parameter.

5.1 Multiplicative Update Rules Formulation

Since the objective function O_{KL} of cdNMF in (21) is convex only with respect to $\boldsymbol{W}$ or $\boldsymbol{S}$ respectively, it is unrealistic to find an algorithm to achieve the global minimum of O_{KL}. Similar to the original NMF[15], we also obtain two update rules which can lead to a local minimum:

$$w_{ik} \leftarrow w_{ik}\frac{\sum\limits_j \Big(x_{ij}s_{jk}\Big/\sum\limits_k w_{ik}s_{jk}\Big)}{\sum\limits_j s_{jk}}, \tag{22}$$

$$s_{jk} \leftarrow s_{jk}\frac{\sum\limits_i \Big(x_{ij}w_{ik}\Big/\sum\limits_k w_{ik}s_{jk}\Big)}{\sum\limits_i w_{ik} + \gamma d_{jk}}. \tag{23}$$

Regarding the update rules (22) and (23), we have the following theorem.

Theorem 2. *The objective function O_{KL} of cdNMF in* (21) *is nonincreasing under the update rules in* (22) *and* (23)*. The objective function is invariant under these updates if and only if $\boldsymbol{W}$ and $\boldsymbol{S}$ are at a stationary point.*

Theorem 2 guarantees the convergence of O_{KL} under the update rules of $\boldsymbol{W}$ and $\boldsymbol{S}$ in (22) and (23), and the final solution will be a local optimum. The proof of Theorem 2 is given in the following.

5.2 Proof of Convergence

Similar to the proof of Theorem 1, we only need to prove that O_{KL} is nonincreasing under the update step in (23) for Theorem 2. In addition, the update step in (23) is exactly the update in (13) with a proper auxiliary function.

Lemma 3. *Function*

$$\begin{aligned} &G(\boldsymbol{S}, \boldsymbol{S}^{(q)}) \\ &= \sum_{i,j}\Big(x_{ij}\log x_{ij} - x_{ij} + \sum_k w_{ik}s_{jk} + \sum_k \gamma d_{jk}s_{jk}\Big) - \\ &\sum_{i,j,k}\Big(x_{ij}\frac{w_{ik}s_{jk}^{(q)}}{\sum\limits_k w_{ik}s_{jk}^{(q)}}\Big(\log w_{ik}s_{jk} - \log\frac{w_{ik}s_{jk}^{(q)}}{\sum\limits_k w_{ik}s_{jk}^{(q)}}\Big)\Big) \end{aligned} \tag{24}$$

is an auxiliary function for

$$F(\boldsymbol{S}) = \sum_{i,j}\Big(x_{ij}\log\frac{x_{ij}}{\sum\limits_k w_{ik}s_{jk}} - x_{ij} + \sum_k w_{ik}s_{jk} + \sum_k \gamma d_{jk}s_{jk}\Big).$$

Proof. Since $G(\boldsymbol{S}, \boldsymbol{S}) = F(\boldsymbol{S})$ is obvious, we need only show that $G(\boldsymbol{S}, \boldsymbol{S}^{(q)}) \geqslant F(\boldsymbol{S})$. To do this, we utilize the convexity of the log function to derive the inequality:

$$-\log\Big(\sum_k w_{ik}s_{jk}\Big) \leqslant -\sum_k\Big(\alpha_k\log\frac{w_{ik}s_{jk}}{\alpha_k}\Big),$$

which holds for all non-negative α_k that sum to unity.

Setting

$$\alpha_k = \frac{w_{ik}s_{jk}^{(q)}}{\sum\limits_k w_{ik}s_{jk}^{(q)}},$$

we have

$$-\log\Big(\sum_k w_{ik}s_{jk}\Big) \leqslant -\sum_k\Big(\frac{w_{ik}s_{jk}^{(q)}}{\sum\limits_k w_{ik}s_{jk}^{(q)}}\Big(\log w_{ik}s_{jk} - \log\frac{w_{ik}s_{jk}^{(q)}}{\sum\limits_k w_{ik}s_{jk}^{(q)}}\Big)\Big). \tag{25}$$

From this inequality, we further get $G(\boldsymbol{S}, \boldsymbol{S}^{(q)}) \geqslant F(\boldsymbol{S})$. □

Proof of Theorem 2. Respect to $\boldsymbol{S}$, the minimum of $G(\boldsymbol{S}, \boldsymbol{S}^{(q)})$ is determined by setting the gradient of $\boldsymbol{G}$ to zero:

$$\sum_i w_{ik} - \sum_i x_{ij}\frac{w_{ik}s_{jk}^{(q)}}{\sum\limits_k w_{ik}s_{jk}^{(q)}}\frac{1}{s_{jk}} + \gamma d_{jk} = 0, \tag{26}$$

where $1 \leqslant j \leqslant n$ and $1 \leqslant k \leqslant t$. Thus, the update rule of (13) takes the form

$$s_{jk}^{(q+1)} = s_{jk}^{(q)}\frac{\sum\limits_i \Big(x_{ij}w_{ik}\Big/\sum\limits_k w_{ik}s_{jk}^{(q)}\Big)}{\sum\limits_i w_{ik} + \gamma d_{jk}}. \tag{27}$$

Since $G(\boldsymbol{S}, \boldsymbol{S}^{(q)})$ is an auxiliary function, $F(\boldsymbol{S})$ is nonincreasing under this update. Rewritten in matrix form, this is equivalent to the update rule in (23). □

6 Experiments

In this section, we will firstly introduce the experimental setup and evaluation metrics. Then, we evaluate the performance of proposed cdNMF model for data clustering on three public databases: Yale

Face①, Caltech 101[17] and AT&T ORL② in comparison with CF[9], NMF[15], GNMF[10], semi-supervised GNMF (sGNMF)[13] and CNMF[13] (including CNMF_F and CNMF_KL). Finally, we analyze the computational complexity of cdNMF, and experimentally show the speed of its convergence.

Following the same setting in CNMF, N categories will be randomly picked up from the dataset by fixing the cluster number N. All of these images are mixed as the collection $\boldsymbol{X}$ for clustering. In addition, 10% images are randomly selected from each category in $\boldsymbol{X}$ as training data for the semi-supervised algorithms (such as CNMF and our cdNMF). To obtain the new representation $\boldsymbol{S}$, we set the dimensionality of the new space to be the same as the number of clusters N. Then K-means is applied to $\boldsymbol{S}$ for clustering. The above process is repeated 10 times, and the average clustering performance is given as the final result. For convenience, we term our cdNMF based on Frobenius norm cost as cdNMF_F and KL-divergence cost as cdNMF_KL, respectively.

6.1 Evaluation Metrics

The clustering results are usually evaluated by comparing the cluster label of each sample with its label provided by the database. Similar to [13], two standard clustering metrics, the accuracy (AC) and normalized mutual information metric (NMI), are utilized to measure the clustering performance. Given a dataset with n images, for each image x_i, let e_i and r_i be the cluster label and the label provided by the database, respectively. The metric AC is defined as follows.

$$AC = \frac{\sum_{i=1}^{n} \delta(r_i, map(e_i))}{n}, \tag{28}$$

where $\delta(x, y)$ is the delta function, which equals 1 if $x = y$ and equals 0 otherwise, and $map(e_i)$ is the mapping function that maps each cluster label e_i to the best label from the database. The best mapping can be found by employing the Kuhn-Munkres algorithm[18].

Let $\boldsymbol{C}$ denote the set of clusters obtained from the ground truth and $\tilde{\boldsymbol{C}}$ obtained from our algorithm. Their mutual information metric $MI(\boldsymbol{C}, \tilde{\boldsymbol{C}})$ is defined as follows.

$$MI(\boldsymbol{C}, \tilde{\boldsymbol{C}}) = \sum_{c_i \in \boldsymbol{C}, \tilde{c}_j \in \tilde{\boldsymbol{C}}} p(c_i, \tilde{c}_j) \times \log \frac{p(c_i, \tilde{c}_j)}{p(c_i) \times p(\tilde{c}_j)}, \tag{29}$$

where $p(c_i)$ and $p(\tilde{c}_j)$ are the probabilities that an image arbitrarily selected from the dataset belongs to the cluster c_i and $\tilde{c}_j$, respectively, and $p(c_i, \tilde{c}_j)$ is the joint probability that the arbitrarily selected image belongs to the cluster c_i as well as $\tilde{c}_j$ at the same time. In our experiments, we use the normalized mutual information NMI as follows.

$$NMI(\boldsymbol{C}, \tilde{\boldsymbol{C}}) = \frac{MI(\boldsymbol{C}, \tilde{\boldsymbol{C}})}{\max(H(\boldsymbol{C}), H(\tilde{\boldsymbol{C}}))}, \tag{30}$$

where $H(\boldsymbol{C})$ and $H(\tilde{\boldsymbol{C}})$ are the entropies of $\boldsymbol{C}$ and $\tilde{\boldsymbol{C}}$, respectively. Note that $NMI(\boldsymbol{C}, \tilde{\boldsymbol{C}})$ ranges from 0 to 1. $NMI = 1$ if the two sets of clusters are identical, and $NMI = 0$ if the two sets are independent.

6.2 Clustering on Yale Face Database

The Yale Face database consists of 165 grayscale images with 15 subjects. Each subject has 11 images, which are different facial expressions or configurations: center-light, w/glasses (with glasses), happy, left-light, w/no glasses (without glasses), normal, right-light, sad, sleepy, surprised, and wink. Following the same preprocessing[13], each image is represented by a 1 024-dimensional vector in the original data space. Fig.1 shows the effectiveness of the proposed cdNMF. Especially, cdNMF_KL outperforms all the other algorithms all the way. The detailed results are described in Table 1.

6.3 Clustering on Caltech 101 Database

Caltech 101 dataset contains 9 144 images which are divided among 101 object classes and one background class including animals, vehicles, etc. Following the same experimental setup[13], we choose the 10 largest categories as our experimental data which consists of 3 044 images in total, and extract the SIFT descriptors to form a 500-dimensional frequency histogram for each image. Fig.2 and Table 2 show the clustering results on the Caltech 101. Specifically, cdNMF_KL achieves the best performance.

6.4 Clustering on AT&T ORL Database

The ORL database contains 10 different images each of which contains 40 distinct subjects, thus 400 images in total. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). Fig.3 and Table 3 show the effectiveness of the proposed cdNMF on ORL dataset.

It is easy to find that our proposed cdNMF_KL has the best performance compared with other algorithms such as CNMF_KL on both Yale Face and Caltech 101

①http://cvc.yale.edu/projects/yalefaces/yalefaces.html, Aug. 2013.

②http://www.uk.research.att.com/facedatabase.html, Aug. 2013.

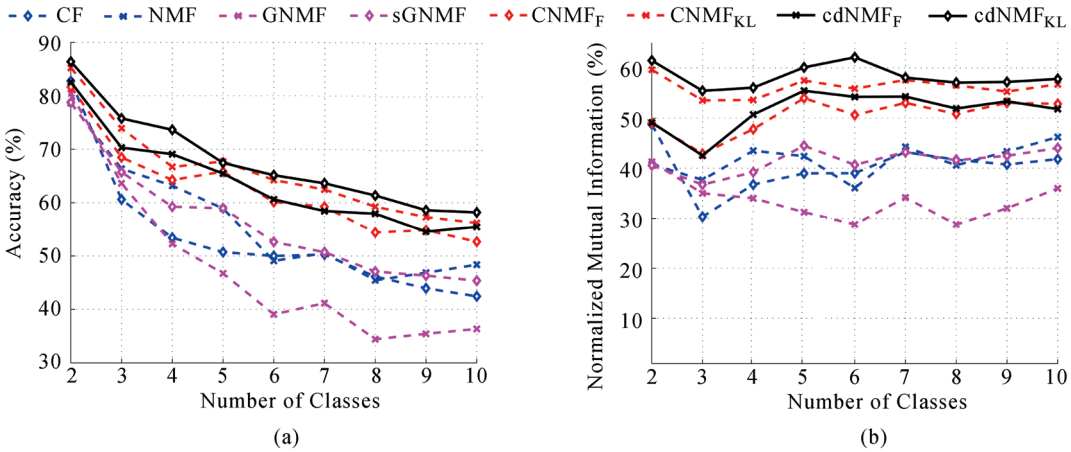


Fig.1. Clustering results on Yale face database. (a) Accuracy. (b) Normalized mutual information.

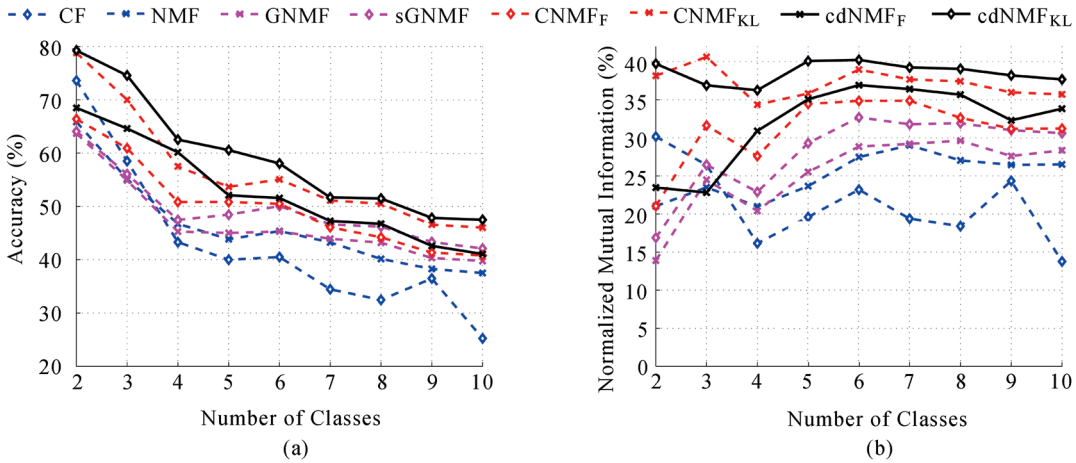


Fig.2. Clustering results on the Caltech 101 database. (a) Accuracy. (b) Normalized mutual information.

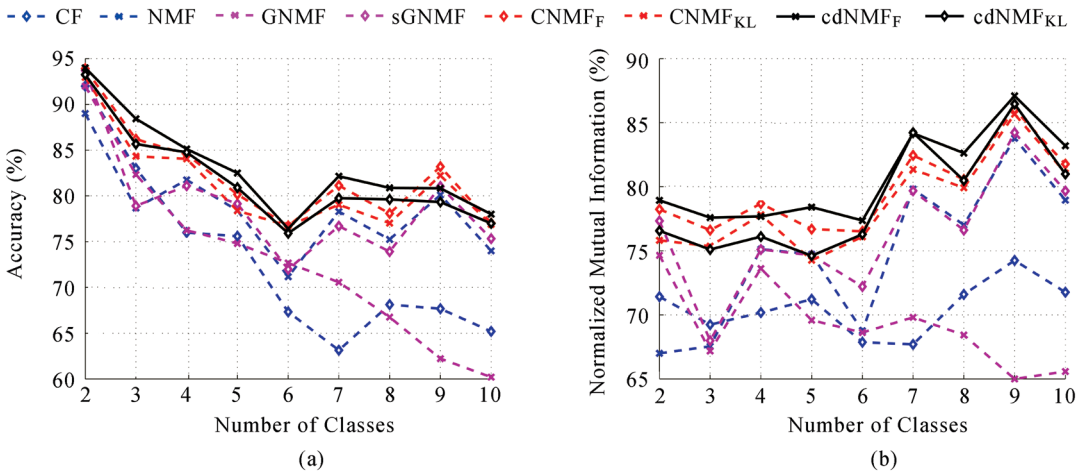


Fig.3. Clustering results on ORL database. (a) Accuracy. (b) Normalized mutual information.

Table 1. Clustering Results on Yale Face Database

N	Accuracy (%)								Normalized Mutual Information (%)							
	CF	NMF	GNMF	sGNMF	$CNMF_F$	$CNMF_{KL}$	$cdNMF_F$	$cdNMF_{KL}$	CF	NMF	GNMF	sGNMF	$CNMF_F$	$CNMF_{KL}$	$cdNMF_F$	$cdNMF_{KL}$
2	82.73	78.64	80.45	78.73	81.64	85.23	82.55	**86.36**	48.74	40.76	41.27	40.56	48.99	59.66	49.11	**61.49**
3	60.61	66.36	63.64	65.73	68.48	73.91	70.30	**75.76**	30.32	37.69	35.00	36.73	42.90	53.52	42.54	**55.43**
4	53.41	63.18	52.27	59.25	64.25	66.70	69.09	**73.64**	36.74	43.50	34.00	39.22	47.80	53.60	50.71	**56.06**
5	50.73	58.91	46.73	58.96	65.82	67.71	65.46	**67.45**	38.98	42.39	31.23	44.50	53.99	57.49	55.45	**60.11**
6	50.00	49.09	39.09	52.64	60.12	64.26	60.61	**65.15**	39.01	36.07	28.78	40.68	50.63	55.88	54.24	**62.12**
7	50.26	50.52	41.17	50.71	59.25	62.47	58.44	**63.64**	43.21	44.25	34.16	43.23	53.03	57.60	54.30	**58.08**
8	46.14	45.45	34.43	47.08	54.40	59.26	57.95	**61.36**	41.69	40.59	28.75	41.61	50.83	56.52	51.91	**57.08**
9	43.94	46.87	35.45	46.33	54.85	57.30	54.55	**58.59**	40.75	43.34	32.01	42.48	53.05	55.30	53.34	**57.20**
10	42.45	48.36	36.36	45.37	52.68	56.20	55.45	**58.18**	41.81	46.22	35.99	44.03	52.84	56.73	51.81	**57.81**
Avg.	53.36	56.38	47.73	56.09	62.39	65.89	63.82	**67.79**	40.14	41.64	33.47	41.45	50.45	56.26	51.49	**58.37**

Table 2. Clustering Results on Caltech 101 Database

N	Accuracy (%)								Normalized Mutual Information (%)							
	CF	NMF	GNMF	sGNMF	$CNMF_F$	$CNMF_{KL}$	$cdNMF_F$	$cdNMF_{KL}$	CF	NMF	GNMF	sGNMF	$CNMF_F$	$CNMF_{KL}$	$cdNMF_F$	$cdNMF_{KL}$
2	73.60	65.85	63.70	64.07	66.39	78.80	68.49	**79.26**	30.16	21.03	13.89	16.89	21.05	38.17	23.48	**39.73**
3	58.50	55.13	54.87	56.06	60.91	69.98	64.62	**74.56**	26.45	23.50	24.49	26.47	31.62	**40.63**	22.82	36.89
4	43.30	46.73	45.28	47.45	50.82	57.50	60.15	**62.52**	16.15	21.00	20.41	22.90	27.60	34.36	30.92	**36.26**
5	40.00	43.86	45.02	48.46	50.83	53.67	52.10	**60.56**	19.60	23.66	25.53	29.31	34.46	35.82	35.07	**40.08**
6	40.50	45.37	45.30	50.07	50.49	55.05	51.56	**58.05**	23.20	27.48	28.86	32.68	34.85	38.97	36.93	**40.23**
7	34.46	43.26	43.93	46.65	46.09	51.12	47.26	**51.68**	19.38	29.02	29.21	31.78	34.88	37.67	36.40	**39.24**
8	32.46	40.15	43.20	46.19	44.21	50.50	46.74	**51.49**	18.40	27.04	29.63	31.92	32.63	37.43	35.67	**39.07**
9	36.48	38.28	40.33	43.34	41.34	46.58	42.60	**47.84**	24.34	26.46	27.61	31.01	31.20	35.97	32.31	**38.20**
10	25.24	37.51	39.76	42.11	40.80	46.02	41.08	**47.48**	13.77	26.51	28.37	30.62	31.21	35.71	33.83	**37.67**
Avg.	42.73	46.24	46.82	49.38	50.21	56.58	52.73	**59.27**	21.27	25.08	25.33	28.18	31.05	37.19	31.94	**38.60**

Table 3. Clustering Results on AT&T ORL Database

N	Accuracy (%)								Normalized Mutual Information (%)							
	CF	NMF	GNMF	sGNMF	$CNMF_F$	$CNMF_{KL}$	$cdNMF_F$	$cdNMF_{KL}$	CF	NMF	GNMF	sGNMF	$CNMF_F$	$CNMF_{KL}$	$cdNMF_F$	$cdNMF_{KL}$
2	92.00	89.00	92.00	93.70	93.90	92.95	**94.00**	93.25	71.42	67.00	74.65	77.33	78.24	75.82	**78.94**	76.56
3	83.00	78.67	82.33	78.90	86.20	84.33	**88.43**	85.67	69.24	67.54	67.18	67.99	76.61	75.36	**77.59**	75.10
4	76.00	81.75	76.25	81.10	84.55	84.05	**85.13**	84.76	70.18	75.13	73.62	75.10	**78.69**	77.82	77.68	76.11
5	75.60	78.40	74.80	79.14	80.16	78.38	**82.50**	80.90	71.22	74.73	69.59	74.66	76.69	74.27	**78.41**	74.62
6	67.33	71.17	72.67	71.98	76.72	**76.73**	76.42	75.92	67.87	68.75	68.62	72.21	76.52	76.09	**77.37**	76.29
7	63.14	78.29	70.57	76.70	81.14	79.04	**82.15**	79.75	67.70	79.82	69.81	79.69	82.45	81.35	84.17	**84.22**
8	68.13	75.25	66.75	73.92	78.08	77.03	**80.88**	79.62	71.60	77.02	68.43	76.61	80.57	79.92	**82.63**	80.49
9	67.67	80.11	62.22	81.04	**83.19**	82.23	80.83	79.33	74.25	83.79	65.00	84.23	86.03	85.67	**87.09**	86.46
10	65.20	74.00	60.20	75.33	77.03	76.88	**78.00**	77.00	71.77	78.96	65.58	79.67	81.77	81.07	**83.20**	80.99
Avg.	73.12	78.51	73.09	79.09	82.33	81.29	**83.15**	81.80	70.58	74.75	69.16	76.39	79.73	78.60	**80.79**	78.98

databases. Meanwhile, cdNMF_F shows better performance than CNMF_F on these databases. In addition, cdNMF_F achieves the best performance on ORL database, and cdNMF_KL demonstrates better performance than CNMF_KL. It is because that cdNMF can learn a set of discriminative basis vectors which are enforced to represent better for their own classes but worse for the others. Thus, data points with the same class label will have similar representations, and consequently the obtained new representations can have more discriminative power than CNMF which just simply embeds the label information in the representations. Note that Frobenius norm cost for cdNMF shows better performance than KL-divergence cost on ORL. This result is similar to CNMF that Frobenius norm is more appropriate for ORL data than KL-divergence.

6.5 Parameter Selection

In the experiments, the tuning parameters in cdNMF, i.e., λ for cdNMF_F and γ for cdNMF_KL, are verified by cross validation to avoid over-fitting. Fig.4 shows how the performance of cdNMF_F and cdNMF_KL varies with the parameter λ and γ respectively on Yale database for cluster number N=10. Thus, we set $\lambda = 1$ and $\gamma = 10$ on Yale database. Due to the space limitation, we have not showed the results on ORL dataset and Caltech dataset whose are similar. Specifically, we empirically set $\lambda = 0.1$ and $\gamma = 10$ for the ORL data, and $\lambda = 10$ and $\gamma = 1$ for Caltech data. It is easy to find that cdNMF is stable with respect to the parameters λ and γ.

6.6 Computational Complexity Analysis and Convergence Study

Based on the Frobenius norm and KL-divergence cost respectively, we utilize iterative update rules to minimize the objective functions of the proposed cdNMF. To analyze the computational complexity, we count the number of operations (addition, multiplication and division) based on the updating rules. Specifically, the overall costs of cdNMF_F and cdNMF_KL are the same as the original NMF for each update step (i.e., $O(mnt)$). In CNMF algorithm, the overall cost of CNMF_F is $O(mnt)$, while that of CNMF_KL becomes $O(n(m+n)t)$.

In addition, we have theoretically proved the convergence of cdNMF. And now we experimentally show the speed of convergence of cdNMF in comparison with NMF in Fig.5. Note that we set the cluster number $N = 10$. Fig.5 demonstrates that our cdNMF converges as fast as NMF within 200 iterations.

7 Discussion and Conclusions

In this paper, we proposed a class-driven NMF method for image representation. To utilize the label information of training data, we associate class labels with basis vectors by introducing an inhomogeneous representation cost constraint. This constraint leads to learn a set of discriminative basis vectors which are enforced to represent better for their own classes but worse for the others. Therefore, the data samples in the same class will have similar representations. Then the new representations can have more discriminating power than CNMF which just simply embeds the label information in the representations. The experiments conducted on standardized datasets have demonstrated the effectiveness of the proposed method. However, the inhomogeneous constraint only focuses on minimizing the inhomogeneous coefficients while fails to consider maximizing the homogeneous ones, which is not sufficient to learn an optimal structured basis. Studying such basis for enhancing cdNMF is our future work.

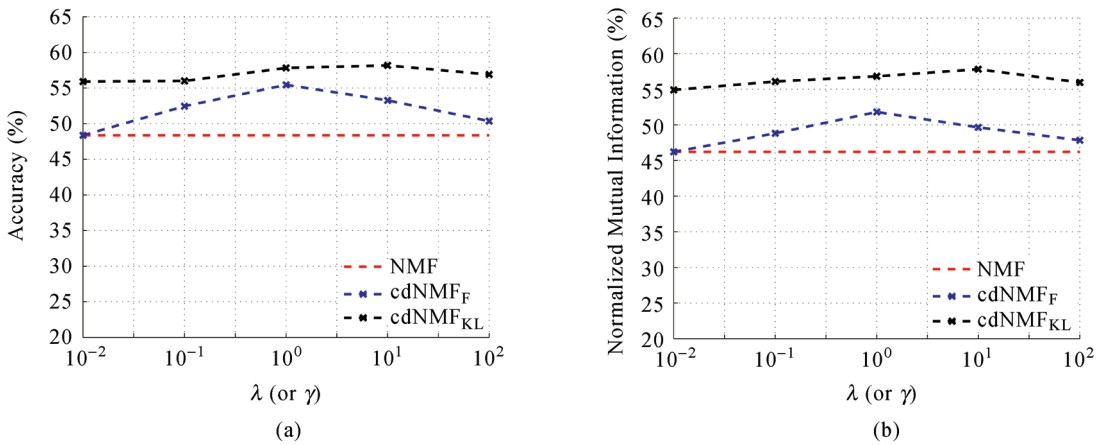


Fig.4. Performance of cdNMF vs parameters λ and γ. In particular, λ is the tuning parameter in cdNMF_F and γ is in cdNMF_KL. In addition, the test values for both λ and γ are $\{10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$. (a) Accuracy. (b) Normalized mutual information.

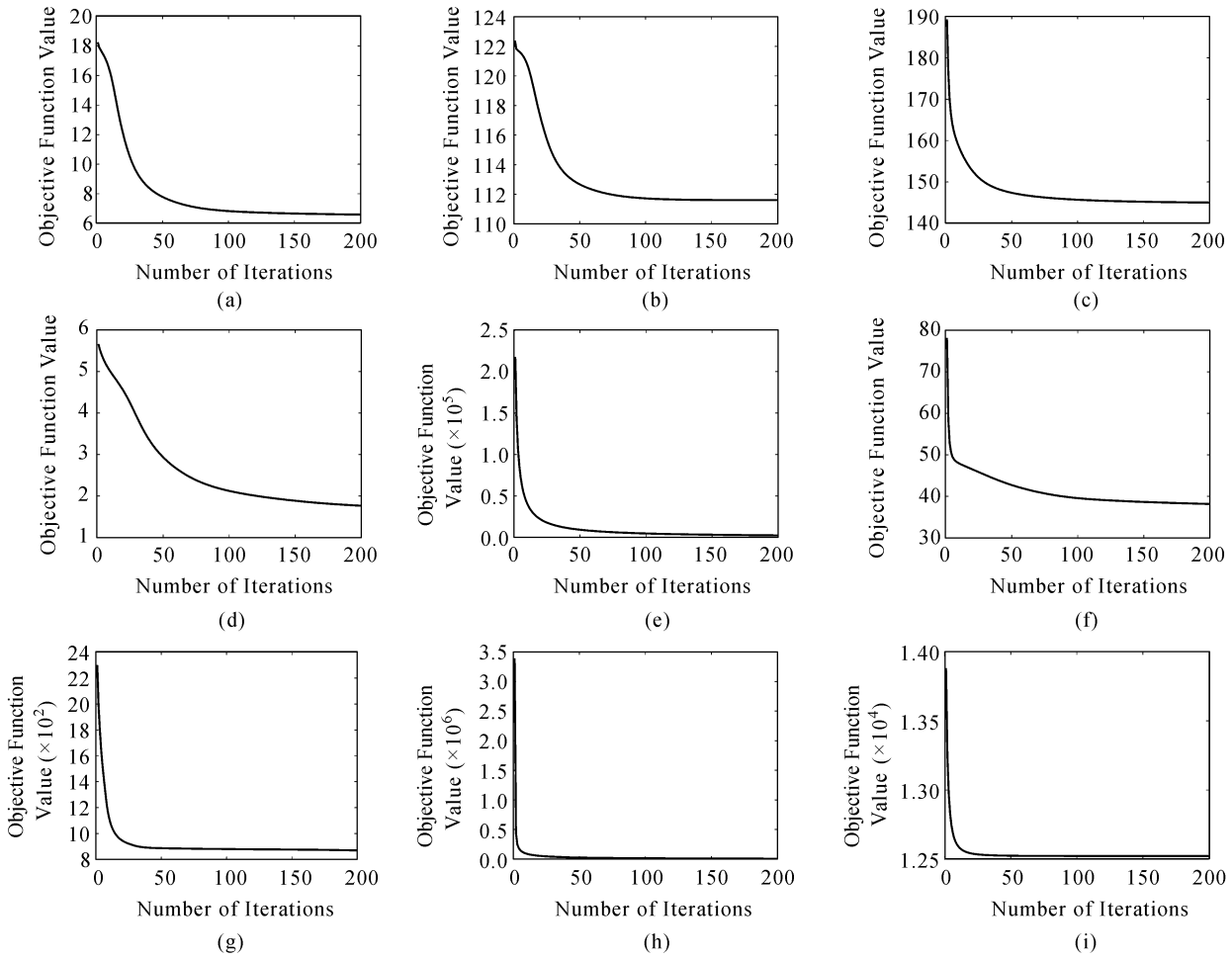


Fig.5. Convergence on Yale, ORL and Caltech database. (a) Yale-NMF. (b) Yale-cdNMF$_{\mathrm{F}}$. (c) Yale-cdNMF$_{\mathrm{KL}}$. (d) ORL-NMF. (e) ORL-cdNMF$_{\mathrm{F}}$. (f) ORL-cdNMF$_{\mathrm{KL}}$. (g) Caltech-NMF. (h) Caltech-cdNMF$_{\mathrm{F}}$. (i) Caltech-cdNMF$_{\mathrm{KL}}$.

References

[1] Duda R O, Hart P E, Stork D G. Pattern Classification (2nd edition). New York: John Wiley, 2001.

[2] Lin Q L, Sheng B, Shen Y, Xie Z F, Chen Z H, Ma L Z. Fast image correspondence with global structure projection. *Journal of Computer Science and Technology*, 2012, 27(6): 1281-1288.

[3] Ping Y, Tian Y J, Zhou Y J, Yang Y X. Convex decomposition based cluster labeling method for support vector clustering. *Journal of Computer Science and Technology*, 2012, 27(2): 428-442.

[4] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999, 401(6755): 788-791.

[5] Palmer S E. Hierarchical structure in perceptual representation. *Cognitive Psychology*, 1977, 9(4): 441-474.

[6] Wachsmuth E, Oram M W, Perrett D I. Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque. *Cerebral Cortex*, 1994, 4(5): 509-522.

[7] Li S Z, Hou X W, Zhang H J, Cheng Q S. Learning spatially localized, parts-based representation. In *CVPR*, December 2001, Volume 1, pp.207-212.

[8] Xu W, Liu X, Gong Y. Document clustering based on non-negative matrix factorization. In *Proc. the 26th ACM SIGIR*, July 28-August 1, 2003, pp.267-273.

[9] Xu W, Gong Y. Document clustering by concept factorization. In *Proc. the 27th ACM SIGIR*, July 2004, pp.202-209.

[10] Cai D, He X, Han J, Huang T S. Graph regularized non-negative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(8): 1548-1560.

[11] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 2006, 7: 2399-2434.

[12] Basu S, Banerjee A, Mooney R. Semi-supervised clustering by seeding. In *Proc. the 19th Int. Conf. Machine Learning*, July 2002, pp.19-26.

[13] Liu H F, Wu Z H, Li X L, Cai D, Huang T S. Constrained non-negative matrix factorization for image representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(7): 1299-1311.

[14] Yang J, Wang J, Huang T. Learning the sparse representation for classification. In *Proc. ICME*, July 2011.

[15] Lee D D, Seung H S. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 2000, 13: 556-562.

[16] Dempster A P, Laird N M, Rubin D B. Maximum likeli-

hood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* (*Methodological*), 1977, 39(1): 1-38.

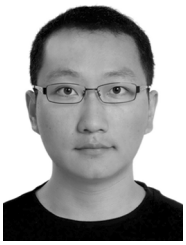
[17] Li F F, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 2007, 106(1): 59-70.

[18] Loász L, Plummer M. Matching Theory. American Mathematical Society, 2009.

Yan-Hui Xiao received the B.S. degree from Beijing Jiaotong University (BJTU), China, in 2007, and is currently a Ph.D. candidate of BJTU. His research interests include image representation, computer vision, and machine learning.

Zhen-Feng Zhu received the Ph.D. degree in pattern recognition and intelligence system from Institute of Automation, Chinese Academy of Sciences, Beijing, in 2005. He is currently an associate professor of the Institute of Information Science, Beijing Jiaotong University. He has been a visiting scholar at the Department of Computer Science and Engineering, Arizona State University, USA, during 2010. His research interests include image and video understanding, computer vision, and machine learning.

Yao Zhao received the B.S. degree from Fuzhou University, China, in 1989 and the Master's degree in engineering from the Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the Ph.D. degree in signal and information processing from the Institute of Information Science, Beijing Jiaotong University, in 1996. His research interests include image/video coding, fractals, digital watermarking, and content-based image retrieval.



Yun-Chao Wei received the B.S. and M.S. degrees from Hebei University of Economics and Business, Shijiazhuang, in 2009, and Beijing Jiaotong University (BJTU) in 2011, respectively. He is currently pursuing the Ph.D. degree in signal and information processing from BJTU. His research interests include machine learning and its application to computer vision and multimedia analysis, e.g., image annotation and cross-media retrieval.