Constrained Nonnegative Matrix Factorization for Image Representation

Haifeng Liu, *Member*, *IEEE*, Zhaohui Wu, *Senior Member*, *IEEE*, Xuelong Li, *Senior Member*, *IEEE*, Deng Cai, *Member*, *IEEE*, and Thomas S. Huang, *Life Fellow*, *IEEE*

Abstract—Nonnegative matrix factorization (NMF) is a popular technique for finding parts-based, linear representations of nonnegative data. It has been successfully applied in a wide range of applications such as pattern recognition, information retrieval, and computer vision. However, NMF is essentially an unsupervised method and cannot make use of label information. In this paper, we propose a novel semi-supervised matrix decomposition method, called *Constrained Nonnegative Matrix Factorization* (CNMF), which incorporates the label information as additional constraints. Specifically, we show how explicitly combining label information improves the discriminating power of the resulting matrix decomposition. We explore the proposed CNMF method with two cost function formulations and provide the corresponding update solutions for the optimization problems. Empirical experiments demonstrate the effectiveness of our novel algorithm in comparison to the state-of-the-art approaches through a set of evaluations based on real-world applications.

Index Terms—Nonnegative matrix factorization, semi-supervised learning, dimension reduction, clustering.

1 Introduction

A fundamental problem in many data analysis tasks is to find a suitable representation of the data [1], [2], [3], [4], [5], [6], [7], [8]. A useful representation typically makes latent structure in the data explicit so further process can be applied. Matrix factorization techniques have been receiving more and more attention as fundamental tools for such data representation. A number of different methods of doing so have been developed by using different criteria. The most popular techniques include Principal Component Analysis (PCA) [9], Singular Value Decomposition (SVD) [10], and Vector Quantization [11]. Central to the matrix factorization is finding two or more matrix factors whose product is a good approximation to the original matrix. In real applications, the dimension of the decomposed matrix factors is usually much smaller than that of the original matrix. This gives rise to compact representation of the data points, which can facilitate other learning tasks such as clustering and classification.

Among matrix factorization methods, Nonnegative Matrix Factorization (NMF) [2], [3] specializes in that it enforces the constraint that the factor matrices must be nonnegative, i.e., all elements must be equal to or greater than zero. This nonnegativity constraint leads NMF to a parts-based representation of the object in the sense that it only allows additive, not subtractive, combination of the original data. Therefore, it is an ideal dimensionality reduction algorithm for image processing, face recognition [2], [12], and document clustering [13], [14], where it is natural to consider the object as a combination of parts to form a whole.

NMF is an unsupervised learning algorithm. That is, NMF is inapplicable to many real-world problems where limited knowledge from domain experts is available. However, many machine learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy [15], [16], [17]. The cost associated with the labeling process may render a fully labeled training set infeasible, whereas acquisition of a small set of labeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value. Therefore, It would be of great benefit to extend the usage of NMF to a semi-supervised manner.

Recently, Cai et al. [1] proposed a Graph regularized NMF (GNMF) approach to encode the geometrical information of the data space. GNMF constructs a nearest neighbor graph to model the local manifold structure. When label information is available, it can be naturally incorporated into the graph structure. Specifically, if two data points share the same label, a large weight can be assigned to the edge connecting them. If two data points have different labels, the corresponding weight is set to be 0. This gives rise to semi-supervised GNMF. The major disadvantage of

- *H. Liu and Z. Wu are with the College of Computer Science, Zhejiang University, 38 ZheDa Road, Hangzhou, Zhejiang 310027, China. E-mail: {haifengliu, wzh}@zju.edu.cn.*
- *X. Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P.R. China. E-mail: xuelong_li@opt.ac.cn.*
- *D. Cai is with the State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou, Zhejiang 310058, China. E-mail: dengcai@cad.zju.edu.cn.*
- *T.S. Huang is with the Department of Electrical and Computer Engineering, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, 2039 Beckman Institute, 405 North Mathews Avenue, Urbana, IL 61801. E-mail: huang@ifp.uiuc.edu.*

Manuscript received 18 Sept. 2010; revised 21 June 2011; accepted 25 Aug. 2011; published online 1 Nov. 2011.
Recommended for acceptance by F. Kahl.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-09-0716.
Digital Object Identifier no. 10.1109/TPAMI.2011.217.

this approach is that there is no theoretical guarantee that data points from the same class will be mapped together in the new representation space, and it remains unclear how to select the weights in a principled manner.

In this paper, we propose a novel matrix decomposition method, called *Constrained Nonnegative Matrix Factorization* (CNMF), which takes the label information as additional *hard* constraints. The central idea of our approach is that the data points from the same class should be merged together in the new representation space. Thus, the obtained parts-based representation has the consistent label with the original data, and therefore can have more discriminating power. Another advantage of our approach is that it is parameter free, which avoids the cost of tuning parameters in order to get the best result. It makes our algorithm applicable to many real-world applications easily and efficiently. We also discuss how to solve the corresponding optimization problem efficiently. And the convergence proof of our optimization scheme is provided. The contributions of this paper are:

1. The standard NMF is an unsupervised learning algorithm which is not able to incorporate the label information. In this paper, we extend it to a semi-supervised learning algorithm. Moreover, our approach takes the label information as *hard* constraints; thus, the data points sharing the same label have the same coordinate in the new representation space. This way, the learned representations can have more discriminating power.
2. Earlier studies [18] showed that NMF and Probabilistic Latent Semantic Analysis (PLSA) are both instances of multinomial PCA. Especially, PLSA solves the problem of NMF with KL divergence [19], [20]. To further explore this aspect, we apply CNMF to KL divergence formulation and provide the update rules solving the optimization problem.
3. Unlike the semi-supervised GNMF, one advantage of our approach is that it is parameter free. Therefore, there is no cost of tuning parameters in order to get the best result. Thus, CNMF is applicable to many real-world applications very easily and efficiently. Moreover, experimental results show that our CNMF algorithm can improve the clustering performance greatly.
4. To our knowledge, there is no method to directly obtain the solution of NMF problem; the state-of-the-art algorithms are using update rules to iteratively get the optimal of the objective function. Therefore, the algorithm efficiency is very important for the real application. In this paper, we qualitatively analyze the computational complexities of our algorithms and experimentally test the convergence rate to demonstrate the algorithm efficiency quantitatively.

This paper is structured as follows: In Section 2, we briefly review the background of NMF and the related work. Section 3 introduces the idea of constrained NMF. The detailed algorithms and theoretical proof of the convergence of the algorithms in two formulations are provided in Sections 4 and 5. Section 6 discusses the computational complexity of our algorithms. Finally, Section 7 presents the experimental results and Section 8 concludes the paper.

2 Related Work

Factorization of matrices is generally nonunique, and a number of different methods of doing so have been developed by incorporating different constraints. PCA [9] and SVD [10] decompose the matrix as the linear combination of principle components. Hoyer [21] and Dueck et al. [22] compute sparse matrix factorization.

NMF differs from these methods in that it enforces the constraint that the elements of the factor matrices must be nonnegative. Suppose we have n data points $\{\mathbf{x}_i\}_{i=1}^n$. Each data point $\mathbf{x}_i \in \mathbb{R}^m$ is m-dimensional and is represented by a vector. The vectors are placed in the columns and the whole data set is represented by a matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$. NMF aims to find two nonnegative matrix factors $\mathbf{U}$ and $\mathbf{V}$ where the product of the two factors is an approximation of the original matrix, represented as

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^T.$$

The approximation is quantified by a cost function which can be constructed by some distance measures. One simple measure is the square of the euclidean distance (also known as the *Frobenius norm*) between two matrices [23]. Then, the goal of NMF can be restated as follows: to factor $\mathbf{X}$ into an $m \times k$ matrix $\mathbf{U}$ and a $k \times n$ matrix $\mathbf{V}^T$ such that the following objective function is minimized:

$$\mathcal{O}_F = \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|^2.$$

The other measure as described in [24] is referred to as the "divergence" of $\mathbf{X}$ from $\mathbf{Y}$:

$$\begin{aligned}\mathcal{O}_{KL} &= D(\mathbf{X}\|\mathbf{Y}) = D(\mathbf{X}\|\mathbf{U}\mathbf{V}^T) \\ &= \sum_{i,j}\left(x_{ij}\log\frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij}\right),\end{aligned}$$

where $\mathbf{Y} = [y_{ij}] = \mathbf{U}\mathbf{V}^T$. This measurement is not symmetric. The distance from $\mathbf{X}$ to $\mathbf{Y}$ is not necessarily the same as the distance from $\mathbf{Y}$ to $\mathbf{X}$. It reduces to the Kullback-Leibler divergence, or relative entropy, when $\sum_{ij} X_{ij} = 1$ and $\sum_{ij} Y_{ij} = 1$.

Both these objective functions are not convex in both variables $\mathbf{U}$ and $\mathbf{V}$. Thus, it is hard to find the global minima for either $\mathcal{O}_F$ or $\mathcal{O}_{KL}$. Lee and Seung proposed an iterative update algorithm [24] to find the locally optimal solution for the above optimization problem.

In the NMF factorization, each column vector of $\mathbf{U}$, $\mathbf{u}_i$, can be regarded as a basis and each data point $\mathbf{x}_i$ is approximated by a linear combination of these k bases, weighted by the components of $\mathbf{V}$. In other words, NMF maps each datum $\mathbf{x}_i$ to $\mathbf{v}_i$ from m-dimensional space to k-dimensional space. The new representation space is spanned by the k bases $\mathbf{u}_i$. In real applications such as image processing [2], face recognition [12], [25], and document clustering [13], [14], we usually set $k \ll m$ and $k \ll n$. Then, the high-dimensional data can be represented by a set of low-dimensional vectors in the hope that the basis vectors can discover the latent semantic structure among the data set. Differently from other dimension reduction algorithms such as PCA, LDA, and LPP [4], the nonnegative constraints on $\mathbf{U}$ and $\mathbf{V}$ only permit the additive combination of the basis vectors, which is the reason why NMF is considered as the parts-based representation.

Ding et al. [26] proposed a Semi-nonnegative Matrix Factorization algorithm where only one matrix factor is restricted to contain nonnegative entries, while it relaxes the constraint on the basis vectors.

The initial papers by Lee and Seung [2], [24] proposed NMF mainly for parts-based decomposition of images. Later on, it is shown that some types of NMF are instances of a more general probabilistic model called "multinomial PCA" [18]. Especially when NMF is obtained by minimizing the Kullback-Leibler divergence, it is equivalent to probabilistic latent semantic analysis [19]. This method is commonly used for analyzing and clustering textual data.

Buntine [18] and Ding et al. [27] also show that when the Frobenius norm is used as a divergence, NMF is equivalent to a relaxed form of K-means clustering: One matrix factor contains cluster centroids and the other contains cluster membership indicators. This also justifies the use of NMF for data clustering.

Other than the multiplicative update method to find the matrix factors, others have suggested gradient descent algorithms [28] to solve the optimization problem. The algorithm is also called alternating nonnegative least squares or "projected gradient." Lin [29] shows that the project gradient method converges faster than the multiplicative update method. Heiler et al. [30] derive optimization schemes for NMF based on sequential quadratic and second order cone programming. A key advantage of this approach is that NMF can be extended by incorporating prior knowledge in the form of additional constraints within the same optimization framework. The additional constraint is to restrict, for each class and for each labeled point that belongs to it, its coefficient to a cone around the class center. This idea is different from our paper, which takes the label information as additional *hard* constraints that the data points from the same class should be merged together in the new representation space.

To our knowledge, most of the existing NMF variants and extensions focus on the matrix factorization and fail to take into account the label information as hard constraints. In this paper, we propose a novel semi-supervised matrix decomposition method, which takes the label information as additional constraints. Thus, the new representations of the data points can have more discriminating power.

3 Semi-Supervised NMF with Hard Constraints

NMF is an unsupervised learning algorithm. It cannot be applied directly to the situation when the label information is available. In this section, we introduce a novel matrix decomposition method, called *Constrained Nonnegative Matrix Factorization*, which takes the label information as additional constraints. This method can guarantee that the data points sharing the same label are mapped into the same class in the low-dimensional space.

Consider a data set consisting of n data points $\{\mathbf{x}_i\}_{i=1}^n$, among which the label information is available for the first l data points $\mathbf{x}_1, \ldots, \mathbf{x}_l$, and the rest of the $n-l$ data points $\mathbf{x}_{l+1}, \ldots, \mathbf{x}_n$ are unlabeled.

Suppose there are c classes. Each data point from $\mathbf{x}_1, \ldots, \mathbf{x}_l$ is labeled with one class. We first build an $l \times c$ indicator matrix $\mathbf{C}$ where $c_{i,j} = 1$ if $\mathbf{x}_i$ is labeled with the jth class; $c_{i,j} = 0$ otherwise. With the indicator matrix $\mathbf{C}$, we define a label constraint matrix $\mathbf{A}$ as follows:

$$\mathbf{A} = \begin{pmatrix} \mathbf{C}_{l\times c} & 0 \\ 0 & \mathbf{I}_{n-l} \end{pmatrix},$$

where $\mathbf{I}_{n-l}$ is an $(n-l)\times(n-l)$ identity matrix. For example, consider n data points, among which $\mathbf{x}_1$, $\mathbf{x}_2$ are labeled with class I, $\mathbf{x}_3$, $\mathbf{x}_4$ are labeled with class II, $\mathbf{x}_5$ is labeled with class III, and the other $n-5$ data points are unlabeled. The label constraint matrix $\mathbf{A}$ based on this example can be represented as follows:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \mathbf{I}_{n-5} \end{pmatrix}.$$

Recall that NMF maps each data point $\mathbf{x}_i$ to $\mathbf{v}_i$ from m-dimensional space to k-dimensional space. To incorporate the label information, we can impose the label constraints by introducing an auxiliary matrix $\mathbf{Z}$:

$$\mathbf{V} = \mathbf{AZ}.$$

From the above equation, it is easy to check that if $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same label, then $\mathbf{v}_i = \mathbf{v}_j$. With the label constraints, original NMF is extended to a semi-supervised learning algorithm (CNMF) by finding two nonnegative matrix factors $\mathbf{U}$ and $\mathbf{Z}$, where the product of the factors $\mathbf{U}$, $\mathbf{Z}$, and $\mathbf{A}$ is an approximation of the original matrix represented as

$$\mathbf{X} \approx \mathbf{U}(\mathbf{AZ})^T.$$

4 Algorithm Minimizing the F-Norm Cost

4.1 The Updating Algorithm

Using the *Frobenius norm* as the cost function, our CNMF algorithm with the label constraints reduces to minimize the following objective function:

$$\mathcal{O}_F = \|\mathbf{X} - \mathbf{UZ}^T\mathbf{A}^T\|, \tag{1}$$

with the constraint that $u_{i,j}$ and $z_{i,j}$ are nonnegative.

The objective function of CNMF in (1) is not convex in both variables $\mathbf{U}$ and $\mathbf{Z}$. It is thus unrealistic to find the global minima for $\mathcal{O}_F$. In the following, we describe an iterative updating algorithm to obtain the local optima of $\mathcal{O}_F$.

Using the matrix property $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$, the objective function $\mathcal{O}_F$ can be rewritten as

$$\begin{aligned}\mathcal{O}_F &= \mathrm{Tr}((\mathbf{X} - \mathbf{UZ}^T\mathbf{A}^T)(\mathbf{X} - \mathbf{UZ}^T\mathbf{A}^T)^T) \\ &= \mathrm{Tr}(\mathbf{XX}^T) - 2\mathrm{Tr}(\mathbf{XAZU}^T) + \mathrm{Tr}(\mathbf{UZ}^T\mathbf{A}^T\mathbf{AZU}^T).\end{aligned}$$

Let α_{ij} and β_{ij} be the Lagrange multiplier for constraints $u_{ij} \geq 0$ and $z_{ij} \geq 0$, respectively, and $\boldsymbol{\alpha} = [\alpha_{ij}]$, $\boldsymbol{\beta} = [\beta_{ij}]$. The Lagrange function $\mathcal{L}$ is

$$\mathcal{L} = \mathcal{O}_F + \mathrm{Tr}(\boldsymbol{\alpha}U^T) + \mathrm{Tr}(\beta Z^T).$$

Requiring that the derivatives of $\mathcal{L}$ with respect to $\mathbf{U}$ and $\mathbf{Z}$ vanish, we have

$$\frac{\partial \mathcal{L}}{\partial \mathbf{U}} = -2\mathbf{XAZ} + 2\mathbf{UZ}^T\mathbf{A}^T\mathbf{AZ} + \alpha = 0,$$
$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}} = -2\mathbf{A}^T\mathbf{X}^T\mathbf{U} + 2\mathbf{A}^T\mathbf{AZU}^T\mathbf{U} + \beta = 0.$$

Using the Kuhn-Tucker condition $\alpha_{ij}u_{ij} = 0$ and $\beta_{ij}z_{ij} = 0$, we get the following equations for u_{ij} and z_{ij}:

$$(\mathbf{XAZ})_{ij}u_{ij} - (\mathbf{UZ}^T\mathbf{A}^T\mathbf{AZ})_{ij}u_{ij} = 0,$$
$$(\mathbf{A}^T\mathbf{X}^T\mathbf{U})_{ij}z_{ij} - (\mathbf{A}^T\mathbf{AZU}^T\mathbf{U})_{ij}z_{ij} = 0.$$

These equations lead to the following updating rules:

$$u_{ij} \quad u_{ij}\frac{(\mathbf{XAZ})_{ij}}{(\mathbf{UZ}^T\mathbf{A}^T\mathbf{AZ})_{ij}}, \tag{2}$$

$$z_{ij} \quad z_{ij}\frac{(\mathbf{A}^T\mathbf{X}^T\mathbf{U})_{ij}}{(\mathbf{A}^T\mathbf{AZU}^T\mathbf{U})_{ij}}. \tag{3}$$

We have the following theorem regarding the above iterative updating rules:

Theorem 1. *The objective function $\mathcal{O}_F$ in (1) is nonincreasing under the update rules in (2) and (3). The objective function is invariant under these updates if and only if* $\mathbf{U}$ *and* $\mathbf{Z}$ *are at a stationary point.*

Theorem 1 guarantees the convergence of the iterations in (2) and (3) and therefore the final solution will be a local optimum. In the following, we will give the proof of Theorem 1.

4.2 Proof of Convergence

To prove Theorem 1, we use the following property of an auxiliary function as used in the Expectation-Maximization algorithm [31], [32].

Lemma 2. *If there exists an auxiliary function G for $F(x)$ which satisfies the conditions $G(x, x') \geq F(x)$ and $G(x, x) = F(x)$, then F is nonincreasing under the update*

$$x^{t+1} = \arg\min_x G(x, x'). \tag{4}$$

The equality $F(x^{t+1}) = F(x^t)$ holds only if x^t is a local minimum of $G(x, x^t)$. By iterating the updates in (4), the sequence of estimates will converge to a local minimum $x_{min} = \arg\min_x F(x)$. We will show this by defining an appropriate auxiliary function for the objective function in (1).

First, we prove the convergence of the update rule in (3). For any element z_{ab} in $\mathbf{Z}$, let $F_{z_{ab}}$ denote the part of $\mathcal{O}$ relevant to z_{ab}. Since the update is essentially element wise, it is sufficient to show that each $F_{z_{ab}}$ is nonincreasing under the update step of (3). We prove this by defining the auxiliary function regarding z_{ab} as follows:

Lemma 3. *Let F' denote the first order derivative with respective to* $\mathbf{Z}$*. The function*

$$G(z, z_{ab}^t) = F_{z_{ab}}(z_{ab}^t) + F'_{z_{ab}}(z_{ab}^t)(z - z_{ab}^t) + \frac{(\mathbf{A}^T\mathbf{AZU}^T\mathbf{U})_{ab}}{z_{ab}^t}(z - z_{ab}^t)^2 \tag{5}$$

is an auxiliary function for $F_{z_{ab}}$, which is the part of $\mathcal{O}_F$ that is only relevant to z_{ab}.

Proof. Obviously, $G(z, z) = F_{z_{ab}}(z)$. According to the definition of auxiliary function, we only need to show that $G(z, z_{ab}^t) \geq F_{z_{ab}}(z)$. In order to do this, we compare $G(z, z_{ab}^t)$ in (5) with the Taylor series expansion of $F_{z_{ab}}(z)$:

$$F_{z_{ab}}(z) = F_{z_{ab}}(z_{ab}^t) + F'_{z_{ab}}(z - z_{ab}^t) + \frac{1}{2}F''_{z_{ab}}(z - z_{ab}^t)^2, \tag{6}$$

where F'' is the second order derivative with respect to $\mathbf{Z}$. It is easy to check that

$$F'_{z_{ab}} = \left(\frac{\partial \mathcal{O}}{\partial \mathbf{Z}}\right)_{ab} = (-2\mathbf{A}^T\mathbf{X}^T\mathbf{U} + 2\mathbf{A}^T\mathbf{AZU}^T\mathbf{U})_{ab}$$

$$F''_{z_{ab}} = 2(\mathbf{A}^T\mathbf{A})_{aa}(\mathbf{U}^T\mathbf{U})_{bb}. \tag{7}$$

Putting (7) into (6) and comparing with (5), we can see that, instead of showing $G(z, z_{ab}^t) \geq F_{z_{ab}}(z)$, it is equivalent to prove

$$\frac{(\mathbf{A}^T\mathbf{AZU}^T\mathbf{U})_{ab}}{z_{ab}^t} \geq \frac{1}{2}F''_{z_{ab}} = (\mathbf{A}^T\mathbf{A})_{aa}(\mathbf{U}^T\mathbf{U})_{bb}.$$

To prove the above inequality, we have

$$\begin{aligned}(\mathbf{A}^T\mathbf{AZU}^T\mathbf{U})_{ab} &= \sum_{l=1}^{k}(\mathbf{A}^T\mathbf{AZ})_{al}(\mathbf{U}^T\mathbf{U})_{lb} \\ &\geq (\mathbf{A}^T\mathbf{AZ})_{ab}(\mathbf{U}^T\mathbf{U})_{bb} \\ &\geq \sum_{l=1}^{k}(\mathbf{A}^T\mathbf{A})_{al}z_{lb}^t(\mathbf{U}^T\mathbf{U})_{bb} \\ &\geq z_{ab}^t(\mathbf{A}^T\mathbf{A})_{aa}(\mathbf{U}^T\mathbf{U})_{bb}.\end{aligned}$$

□

Next, we define an auxiliary function for the update rule in (2). Similarly, let $F_{u_{ab}}$ denote the part of $\mathcal{O}_F$ relevant to u_{ab}. Then, the auxiliary function regarding u_{ab} is defined as follows:

Lemma 4. *The function*

$$G(u, u_{ab}^t) = F_{u_{ab}}(u_{ab}^t) + F'_{u_{ab}}(u_{ab}^t)(u - u_{ab}^t) + \frac{(\mathbf{UZ}^T\mathbf{A}^T\mathbf{AZ})_{ab}}{u_{ab}^t}(u - u_{ab}^t)^2 \tag{8}$$

is an auxiliary function for $F_{u_{ab}}$, which is the part of $\mathcal{O}_F$ that is only relevant to u_{ab}.

The proof of Lemma 4 is essentially similar to the proof of Lemma 3 and is omitted here due to space limitation. With the above lemmas, now we give the proof of Theorem 1.

Proof of Theorem 1. Putting $G(z, z_{ab}^t)$ of (5) into (4), we get

$$z_{ab}^{t+1} = \arg\min_z G(z, z_{ab}^t) = z_{ab}^t\frac{(\mathbf{A}^T\mathbf{X}^T\mathbf{U})_{ab}}{(\mathbf{A}^T\mathbf{AZU}^T\mathbf{U})_{ab}}.$$

□

Since (5) is an auxiliary function, $F_{z_{ab}}$ is nonincreasing under this update rule, according to Lemma 3.

TABLE 1
Computational Operation Counts for Each Iteration in NMF and CNMF

	F-norm fomulation			
	addition	multiplication	division	overall
NMF	$2mnk + 2(m+n)k^2$	$2mnk + 2(m+n)k^2 + (m+n)k$	$(m+n)k$	$O(mnk)$
CNMF	$(2m+2n-l+c)(n-l+c)k$ $+2(m+n)k^2$	$(2m+2n-l+c)(n-l+c)k$ $+2(m+n)k^2 + (m+n-l+c)k$	$(m+n-l+c)k$	$O(mnk)$
	KL Divergence fomulation			
	addition	multiplication	division	overall
NMF	$4mnk + (m+n)k$	$4mnk + (m+n)k$	$2mn + (m+n)k$	$O(mnk)$
CNMF_{KL}	$2n(n-l+c)k + 4mnk$ $+2n(n-l+c) + (m+n)k$	$2n(n-l+c)k + 4mnk$ $+(m+n)k + 2(n-l+c)k$	$2mn$ $+(m+n-l+c)k$	$O(n(m+n)k)$

Although, for each update step, the cost for the CNMF in KL divergence formulation is a little more expensive than that of NMF, the overall algorithm complexity of CNMF_{KL} may not be slower since the number of iterations to converge is different. We will experimentally study the convergence rate for these algorithms in the next section. Suppose the multiplicative updates for NMF, CNMF, and CNMF_{KL} stop after t_1, t_2, and t_3 iterations, individually, the overall computational complexity for these algorithms will be $O(t_1mnk)$, $O(t_2mnk)$, and $O(t_3n(m+n)k)$.

7 Experimental Results

In this section, we investigate the use of our proposed CNMF algorithm for data clustering. Several experiments are carried out to show the effectiveness of our algorithm for image clustering.

7.1 Evaluation Metrics

We use two metrics to evaluate the clustering performance [1], [13]. The result is evaluated by comparing the cluster label of each sample with the label provided by the data set. One metric is accuracy (AC), which is used to measure the percentage of correct labels obtained. Given a data set containing n images, for each sample image, let l_i be the cluster label we obtain by applying different algorithms and r_i be the label provided by the data set. The accuracy is defined as

$$AC = \frac{\sum_{i=1}^{n} \delta(r_i, map(l_i))}{n}, \tag{20}$$

where $\delta(x, y)$ is the delta function that equals 1 if $x = y$ and equals 0 otherwise, and $map(l_i)$ is the mapping function that maps each cluster label l_i to the equivalent label from the data set. The best mapping can be found by using the Kuhn-Munkres algorithm [34].

The second metric is the normalized mutual information ($\widehat{MI}$). In clustering applications, mutual information is used to measure how similar two sets of clusters are. Given two sets of image clusters $\mathcal{C}$ and $\mathcal{C}'$, their mutual information metric $MI(\mathcal{C}, \mathcal{C}')$ is defined as

$$MI(\mathcal{C}, \mathcal{C}') = \sum_{c_i \in \mathcal{C}, c'_j \in \mathcal{C}'} p(c_i, c'_j) \cdot \log \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)}, \tag{21}$$

where $p(c_i)$, $p(c'_j)$ denote the probabilities that an image arbitrarily selected from the data set belongs to the clusters c_i and c'_j, respectively, and $p(c_i, c'_j)$ denotes the joint probability that this arbitrarily selected image belongs to the cluster c_i as well as c'_j at the same time. $MI(\mathcal{C}, \mathcal{C}')$ takes values between zero and $\max(H(\mathcal{C}), H(\mathcal{C}'))$, where $H(\mathcal{C})$ and $H(\mathcal{C}')$ are the entropies of $\mathcal{C}$ and $\mathcal{C}'$, respectively. It reaches the maximum $\max(H(\mathcal{C}), H(\mathcal{C}'))$ when the two sets of image clusters are identical and it becomes zero when the two sets are completely independent. One important character of $MI(\mathcal{C}, \mathcal{C}')$ is that the value keeps the same for all kinds of permutations. In our experiments, we use the normalized metric $\widehat{MI}(\mathcal{C}, \mathcal{C}')$, which takes values between 0 and 1:

$$\widehat{MI}(\mathcal{C}, \mathcal{C}') = \frac{MI(\mathcal{C}, \mathcal{C}')}{\max(H(\mathcal{C}), H(\mathcal{C}'))}. \tag{22}$$

7.2 Performance Evaluation and Comparisons

To show the data clustering performance, we compare our algorithm with other related methods on four data sets. The algorithms that we evaluated are listed below:

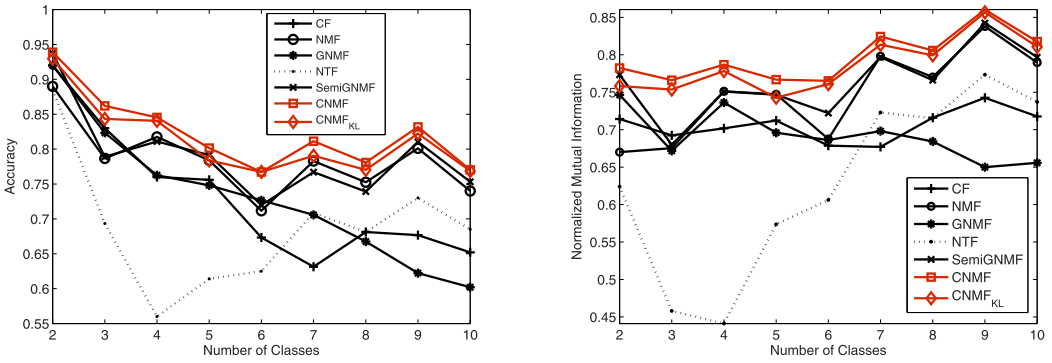
- Our proposed Constrained Nonnegative Matrix Factorization algorithm minimizing the F-norm cost.
- Our proposed Constrained Nonnegative Matrix Factorization algorithm minimizing the KL divergence cost (**CNMF_{KL}**).
- Nonnegative Matrix Factorization-based clustering. We implemented a normalized cut weighted version of NMF as suggested in [13].
- Nonnegative Tensor Factorization (**NTF**) [35]. NTF is an extension of NMF to tensor data. In NTF, each face image is represented as a second order tensor, rather than a vector.

TABLE 2
Parameters Used in Complexity Analysis

Parameters	Description
m	number of features for each data point
n	number of data points
k	number of factors
l	number of labeled data points
c	number of classes

TABLE 3
Statistics of the Data Sets

dataset	size(N)	dimensionality(M)	of classes(K)
ORL	400	1024	40
Yale	165	1024	15
Corel	4970	500	50
Caltech-101	3044	500	10



(a) Accuracy vs. number of clusters (b) Mutual Information vs. number of clusters

Fig. 1. Clustering performance on the AT&T ORL database.

- Graph regularized Nonnegative Matrix Factorization [1] which encodes the geometrical information of the data space into matrix factorization.
- Semi-supervised Graph regularized Nonnegative Matrix Factorization on Manifold (**SemiGNMF**) [1]. This method incorporates the label information into the graph structure by modifying the weight matrix.
- Concept Factorization-based clustering (**CF**) [36].

We evaluate the clustering performance on four image data sets. These data sets contain a number of categories of images. The important statistics of these data sets are summarized in Table 3. We will describe the details of the data sets individually later. For each data set, the evaluations are conducted with different numbers of clusters varying from 2 to 10. For the fixed cluster number k, we run the experiments as follows:

1. We randomly choose k categories from the data set, and mix the images of these k categories as the collection $\mathbf{X}$ for clustering. For the semi-supervised algorithms (**CNMF**, **CNMF**$_{KL}$, and **SemiGNMF**), we randomly picked up 10 percent images from each category in $\mathbf{X}$ and use their category number as the available label information. The exception is the ORL database. There are only 10 images for each category in ORL and 10 percent is just one image. One label is meaningless for CNMF since this algorithm maps the images with the same label onto the same point. Thus, for ORL, we randomly choose two images from each category to provide the label information.
2. On the clustering set, we apply different matrix factorization algorithms as listed above to obtain new data representations $\mathbf{V}$. We set the dimensionality of the new space to be the same as the number of clusters (the same technique is used in spectral clustering). Therefore, this step maps the data from the original space to a low-dimensional (k-dimensional) space.
3. K-means is then applied to the new data representation $\mathbf{V}$ for images clustering. K-means is repeated 20 times with different initial points and the best result in terms of the cost function of K-means is recorded.
4. We compare the obtained clusters with the original image category to compute the accuracy and normalized mutual information.

The above process is repeated 10 times and the average clustering performance is recorded as the final result. As we mentioned before, there is no parameter in our approach. For other algorithms, the parameters are set to be the values that each algorithm can achieve its best results.

7.2.1 ORL Database

The AT&T ORL database[1] consists of 10 different images for each of 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). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position.

In all the experiments, images are preprocessed so that faces are located. Original images are first normalized in scale and orientation such that the two eyes are aligned at the same position. Then, the facial areas were cropped into the final images for clustering. Each image is 32×32 pixels with 256 gray levels per pixel.

Fig. 1 shows the plots of accuracy and normalized mutual information versus the number of clusters. Our proposed CNMF and CNMF_{KL} algorithms consistently outperform all the other algorithms. CF, NMF, GNMF, and SemiGNMF perform comparably to one another. In this data set, CNMF achieves the best and CNMF_{KL} achieves the second best performance.

We also randomly select 25 subjects with 10 images for each subject from the ORL database and run the algorithms on the total 250 images. Fig. 2 shows the effects. The first image contains the 25 subjects and the other three are the basis vectors obtained by NMF, CNMF, and CNMF_{KL}.

7.2.2 Yale Database

The Yale database[2] contains 165 grayscale images of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink. We do the same preprocessing for this data set as for the ORL data set. Thus, each image is also represented by a 1,024-dimensional vector in image space.

1. http://www.uk.research.att.com/facedatabase.html.
2. http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

The clustering results are summarized in Table 6. Graphical plots are shown in Fig. 4. As we can see, NMF, GNMF, and SemiGNMF have similar results. SemiGNMF has not shown much advantage over NMF and GNMF although SemiGNMF makes use of label information. In comparison, CNMF and CNMF_{KL} greatly outperform the above three methods for all the cases and CNMF_{KL} achieves the best performance. From Table 6, we can see that comparing to the best algorithm other than our proposed CNMF algorithms, i.e., SemiGNMF, CNMF_{KL} achieves 7.46 percent improvement and CNMF achieves 4.41 percent improvement in accuracy. For normalized mutual information, CNMF_{KL} achieves 8.38 percent improvement and CNMF achieves 4.81 percent improvement.

7.2.4 Caltech-101 Database

Caltech-101 is a data set of digital images created by Caltech University containing 101 object categories. Each category contains about 40 to 800 images. Size of each image is roughly 300 x 200 pixels. In our experiment, we select the 10 largest categories, except the BACKGROUND_GOOGLE category. In total, the subset for our test contains 3,044 images. We do the same preprocessing for each image as for the Corel data set. We first extract SIFT descriptors, and then generate codewords as the features for each image. For the Caltech-101 data set, the number of SIFT descriptors is 555,292, and we also generate 500 codewords. Finally, each image in Caltech-101 is represented by a 500-dimensional frequency histogram.

The clustering results are summarized in Table 7 and graphical plots are shown in Fig. 5. On this data set, CNMF_{KL} still shows the best performance. CNMF obtains the second best for the normalized mutual information measurement and most cases for the accuracy measurement. In general, our methods demonstrate much better effectiveness in clustering. As shown in Table 7, comparing to SemiGNMF, CNMF_{KL} improves 7.2 percent in accuracy. For normalized mutual information, CNMF_{KL} improves 9.01 percent and CNMF improves 2.87 percent.

7.3 Convergence Study

We use iterative update rules to obtain the local optima of the objective function of CNMF no matter the cost measurement is in Frobenius norm or KL divergence. In the previous sections, we have proven the convergence of our update rules and analyze the computational complexity. Here, we experimentally show the speed of convergence of our algorithm.

We compare the convergence speed of the original NMF algorithm, our CNMF minimizing the F-norm cost (CNMF), and CNMF minimizing the KL divergence cost (CNMF_{KL}). Fig. 6 shows the convergence rate of all three algorithms on the four image databases. For each figure, the x-axis is the number of iterations and the y-axis is the value of objective function. We can see that both CNMF and CNMF_{KL} algorithms converge very fast. For the ORL, Corel, and Caltech-101 databases, all three algorithms converge within 20 iterations. For the Yale database, they converge within 100 iterations. Especially, we notice that CNMF_{KL} converges within less than 10 iterations for both Yale database and Corel database, demonstrating the extraordinary performance. This also verifies the statement we made in Section 6 that although CNMF_{KL} need more operations for each update step, the overall cost for the whole process is not expensive since it may converge faster than NMF.

8 Conclusions

In this paper, we have presented a novel matrix factorization method, called Constrained Nonnegative Matrix Factorization, which makes use of both labeled and unlabeled data points. CNMF imposes the label information to the objective function as hard constraints. This way, the new representations of the data points can have more discriminating power. We showed the CNMF approach in two formulations and proposed update algorithms for both optimization problems. The experimental results on four standard image databases have demonstrated the effectiveness of our approach. Both CNMF and CNMF_{KL} outperform other existing algorithms. Especially, the CNMF approach in the KL-divergence form shows remarkable advantage over the other algorithms on all cases. Moreover, our algorithm is parameter free. Thus, our proposed CNMF can be easily applied to a wide range of practical problems.

Acknowledgments

This work was supported by the National Basic Research Program of China (973 Program) under Grant 2012CB316400, National Natural Science Foundation of China (Grant No.: 61125106, 91120302, 61072093), and the Qian Jiang Talent Program of Zhejiang Province under Grant 2011R10055.

References

[1] D. Cai, X. He, X. Wu, and J. Han, "Non-Negative Matrix Factorization on Manifold," *Proc. Eighth IEEE Int'l Conf. Data Mining,* pp. 63-72, 2008.
[2] D.D. Lee and H.S. Seung, "Learning the Parts of Objects by Non-Negative Matrix Factorization," *Nature,* vol. 401, pp. 788-791, 1999.
[3] T. Li and C. Ding, "The Relationships among Various Nonnegative Matrix Factorization Methods for Clustering," *Proc. Sixth IEEE Int'l Conf. Data Mining,* pp. 362-371, 2006.
[4] X. He and P. Niyogi, "Locality Preserving Projections," *Advances in Neural Information Processing Systems,* vol. 16, MIT Press, 2003.
[5] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face Recognition Using Laplacianfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 3, pp. 328-340, Mar. 2005.
[6] W. Liu, D. Tao, and J. Liu, "Transductive Component Analysis," *Proc. IEEE Int'l Conf. Data Mining,* pp. 433-442, 2008.
[7] D. Tao, X. Li, X. Wu, and S.J. Maybank, "Geometric Mean for Subspace Selection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 31, no. 2, pp. 260-274, Feb. 2009.
[8] D. Tao, X. Li, X. Wu, and S.J. Maybank, "General Averaged Divergence Analysis," *Proc. IEEE Int'l Conf. Data Mining,* 2007.
[9] I. Jolliffe, *Principle Component Analysis.* Springer-Verlag, 1986.
[10] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification.* Wiley-Interscience, 2000.
[11] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression.* Kluwer Academic Press, 1992.
[12] S. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning Spatially Localized, Parts-Based Representation," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* pp. 207-212, 2001.
[13] W. Xu, X. Liu, and Y. Gong, "Document Clustering Based on Non-Negative Matrix Factorization," *Proc. Ann. ACM SIGIR Conf. Research and Development in Information Retrieval,* 2003.

[14] F. Shahnaz, M.W. Berry, V. Pauca, and R.J. Plemmons, "Document Clustering Using Nonnegative Matrix Factorization," *Information Processing and Management,* vol. 42, pp. 373-386, 2006.
[15] M. Belkin, V. Sindhwani, and P. Niyogi, "Manifold Regularization: A Geometric Framework for Learning from Examples," *J. Machine Learning Research,* vol. 7, pp. 2399-2434, 2006.
[16] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf, "Learning with Local and Global Consistency," *Advances in Neural Information Processing Systems,* vol. 16, MIT Press, 2003.
[17] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions," *Proc. 20th Int'l Conf. Machine Learning,* 2003.
[18] W. Buntine, "Variational Extensions to EM and Multinomial PCA," *Proc. 13th European Conf. Machine Learning,* 2002.
[19] E. Gaussier and C. Goutte, "Relation between PLSA and NMF and Implications," *Proc. Ann. ACM SIGIR Conf. Research and Development in Information Retrieval,* pp. 601-602, 2005.
[20] C. Ding, T. Li, and W. Peng, "Nonnegative Matrix Factorization and Probabilistic Latent Semantic Indexing: Equivalence Chi-Square Statistics and a Hybrid Method," *Proc. AAAI Conf. Artificial Intelligence,* 2006.
[21] P.O. Hoyer, "Non-Negative Matrix Factorization with Sparseness Constraints," *J. Machine Learning Research,* vol. 5, pp. 1457-1469, 2004.
[22] D. Dueck, Q.D. Morris, and B.J. Frey, "Probabilistic Sparse Matrix Factorization," Technical Report PSI-2004-23, Univ. of Toronto, 2004.
[23] P. Paatero and U. Tapper, "Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values," *Environmetrics,* vol. 5, pp. 111-126, 1994.
[24] D.D. Lee and H.S. Seung, "Algorithms for Non-Negative Matrix Factorization," *Advances in Neural Information Processing Systems,* MIT Press, 2001.
[25] W. Liu, N. Zheng, and X. Lu, "Non-Negative Matrix Factorization for Visual Coding," *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing,* 2003.
[26] C. Ding, T. Li, and M.I. Jordan, "Convex and Semi-Nonnegative Matrix Factorizations," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 1, pp. 45-55, Jan. 2010.
[27] C. Ding, X. He, and H.D. Simon, "On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering," *Proc. SIAM Data Mining Conf.,* pp. 606-610, 2005.
[28] C.-J. Lin, "Projected Gradient Methods for Nonnegative Matrix Factorization," *Neural Computation,* vol. 19, no. 10, pp. 2756-2779, 2007.
[29] C.-J. Lin, "On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization," *IEEE Trans. Neural Networks,* vol. 18, no. 6, pp. 1589-1596, Nov. 2007.
[30] M. Heiler, C. Schnörr, P. Bennett, and E. Parrado-Hernandez, "Learning Sparse Representations by Non-Negative Matrix Factorization and Sequential Cone Programming," *J. Machine Learning Research,* vol. 7, pp. 1385-1407, 2006.
[31] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the Em Algorithm," *J. Royal Statistics Soc.,* vol. 39, pp. 1-38, 1977.
[32] L. Saul and F. Pereira, "Aggregate and Mixed-Order Markov Models for Statistical Language Processing," *Proc. Second Conf. Empirical Methods in Nature Language Processing,* pp. 81-89, 1977.
[33] J. Kivinen and M. Warmuth, "Additive versus Exponentiated Gradient Updates for Linear Prediction," *J. Information and Computation,* vol. 132, pp. 1-64, 1997.
[34] L. Lovasz and M. Plummer, *Matching Theory.* North Holland, 1986.
[35] A. Shashua and T. Hazan, "Non-Negative Tensor Factorization with Applications to Statistics and Computer Vision," *Proc. Int'l Conf. Machine Learning,* pp. 793-800, 2005.
[36] W. Xu and Y. Gong, "Document Clustering by Concept Factorization," *Proc. Ann. ACM SIGIR Conf. Research and Development in Information Retrieval,* 2004.
[37] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.
[38] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object Retrieval with Large Vocabularies and Fast Spatial Matching," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2007.

Haifeng Liu received the bachelor's degree in computer science from the Special Class for the Gifted Young at the University of Science and Technology of China. She received the PhD degree from the Department of Computer Science at the University of Toronto in 2009. She is an assistant professor in the College of Computer Science at Zhejiang University, China. Her research interests lie in the fields of machine learning, pattern recognition, web mining, and information dissemination. She is a member of the IEEE.

Zhaohui Wu received the PhD degrees in computer science from Zhejiang University, Hangzhou, China, and Kaiserslautern University, Germany, in 1993. He is currently a professor in the College of Computer Science and the vice principal at Zhejiang University. His research interests include distributed artificial intelligence, grid computing and systems, and embedded ubiquitous computing. He is a senior member of the IEEE and a member of the IEEE Computer Society.

Xuelong Li is a full professor with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences. He has nearly 60 IEEE journal papers in the areas of image processing and pattern recognition. He is an editor for four IEEE transactions and the Elsevier *Pattern Recognition* journal. He has served as a chair of conferences more than 40 times, e.g., PC cochair of ICME 2014, and as a PC member more than 200 times. He is a fellow of the British Computer Society, IAPR, IET, and the SPIE, and a senior member of the IEEE.

Deng Cai received the bachelor's and master's degrees in automation from Tsinghua University in 2000 and 2003, respectively. He received the PhD degree in computer science from the University of Illinois at Urbana Champaign in 2009. He is an associate professor in the State Key Lab of CAD&CG, College of Computer Science at Zhejiang University, China. His research interests include machine learning, data mining, and information retrieval. He is a member of the IEEE.

Thomas S. Huang received the ScD degree from the Massachusetts Institute of Technology in electrical engineering, and was on the faculty of MIT and Purdue University. He joined the University of Illinois at Urbana Champaign in 1980 and is currently the William L. Everitt Distinguished Professor of Electrical and Computer Engineering, a research professor in the Coordinated Science Laboratory, professor of the Center for Advanced Study, and cochair of the Human Computer Intelligent Interaction major research theme of the Beckman Institute for Advanced Science and Technology. He is a member of the National Academy of Engineering and has received numerous honors and awards, including the IEEE Jack S. Kilby Signal Processing Medal (with Ar. Netravali) and the King-Sun Fu Prize of the International Association of Pattern Recognition. He has published 21 books and more than 600 technical papers in network theory, digital holography, image and video compression, multimodal human computer interfaces, and multimedia databases. He is a life fellow of the IEEE.