kNN 그라프를 이용한 비음수 행렬 분해

2010

안수남

석 사 학 위 논 문

k-NN 그라프를 이용한 비음수 행렬 분해

안수남(Shounan An)

컴퓨터공학과

포항공과대학교 대학원

2010

k-NN 그라프를 이용한
비음수 행렬 분해

Nonnegative Matrix Factorization
Regularized by k-NN Graphs

Nonnegative Matrix Factorization Regularized by k-NN Graphs

by

Shounan An

Department of Computer Science and Engineering

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

A thesis submitted to the faculty of Pohang University of Science and Technology in partial fulfillment of the requirements for the degree of Master of Science in the Department of Computer Science and Engineering

Pohang, Korea

2009 . 12 . 15 .

Approved by

____

Major Advisor

k-NN 그라프를 이용한 비음수 행렬 분해

안수남

위 논문은 포항공대 대학원 석사 학위 논문으로 학위 논문 심사 위원회를 통과하였음을 인정합니다.

2009 년 12 월 15 일

학위논문 심사위원회 위원장 최 승 진 (인)

위 원 황 승 원 (인)

위 원 유 환 조 (인)

DCS 20079153 안수남, Shounan An, k-NN 그라프를 이용한 비음수 행렬 분해, Nonnegative Matrix Factorization Regularized by k-NN Graphs, Department of Computer Science and Engineering, 2010, 전체 페이지 수 p, Advisor: Seungjin Choi. Text in English

Abstract

Nonnegative matrix factorization (NMF) is a widely used feature extraction method. NMF decomposes a data matrix into a basis matrix and a feature matrix with all of these matrices allow to have only nonnegative elements. With only non-subtractive constraints, NMF learns a sparse basis matrix, result in part-based representation. The fundamental goal of feature extraction is to exploit the more compact and discriminative representation of input data for further processing such as classification or clustering. NMF is unsupervised feature extraction method, which assumes that data points are generated from Euclidean space. Based on this observation , we use label information to directly exploit the discriminative geometrical structure of data points to extract more discriminative and also respects manifold structure of data. In this paper, we propose a novel feature extraction (subspace learning) method named NMF regularized by k-NN graphs (KNMF). KNMF is based on two kinds of graphs: intra-class k-NN graph and inter-class k-NN graph. intra-class k-NN graph connects only the neighboring data points which belong to the same class of the given data point, while inter-class k-NN graph connects the neighboring data points which belong to different class of the given data point. By minimizing the local regions of intra-class neighborhood and maximizing the local regions of inter-class neighborhood, KNMF could exploit more discriminative hidden patterns of given data set, benefit to the following classification based on the extracted features. Experiments on several benchmark face recognition

datasets and document datasets confirmed the useful behavior of our proposed method in the task of feature extraction.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Classification is the classical supervised learning method. The fundamental goal of classification is that discriminate data points from different classes. Traditional machine learning algorithms are suitable for vector-based representation of data. Extract features which incorporates more discriminative information as new representation of original data is important. With this new representation of data we could do classification or clustering based on wether we have label information or not.

Nonnegative matrix factorization (NMF) is a widely used feature extraction method. NMF is used for extract meaningful features for EEG classification, gene classification, face recognition, document classification etc. We propose NMF regularized by k-NN graphs (KNMF) as feature extraction method which have discriminative power and also respects the geometrical structure of data in a unified way.

In this work, we focus on face recognition and document classification task to evaluate the useful behavior of our proposed method. For document data we make preprocessing process to raw document collection based on vector space model, one of most widely used information retrieval technique.

1.1 Datasets

In this section, we describe the datasets used in the numerical experiments.

1.1.1 Document datasets

- WebKB [1] is a dataset consisting of the 6,000 web pages from computer science department of four universities (Cornell, Texas, Washington, and Wisconsin). Each web page is labeled with one out of seven categories: student, professor, course, project, staff, department, and "other".
- The Cora [1] data set contains the abstracts and references of about 34,000 research papers from the computer science community. We use part of them to categorize into one of subfields of data structure (DS), hardware and architecture (HA), machine learning (ML), operating system (OS) and programing language (PL). We adopt the same subset of the data as that in [1] to test our method.

1.1.2 Face image datasets

There are 2 face image datasets used in face recognition task.

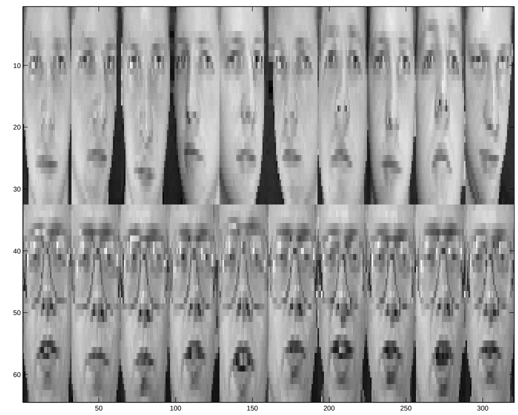
- The ORL [2] database contains 40 individuals and 10 different images for each individual, including variations in illuminations, facial expression (smiling/not smiling) and pose. Fig. 1 illustrates two sample subjects of the ORL database, along with variations in illumination, facial expression and pose. In our experiment, the images are resized into 32×32.

[1] http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/
[2] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

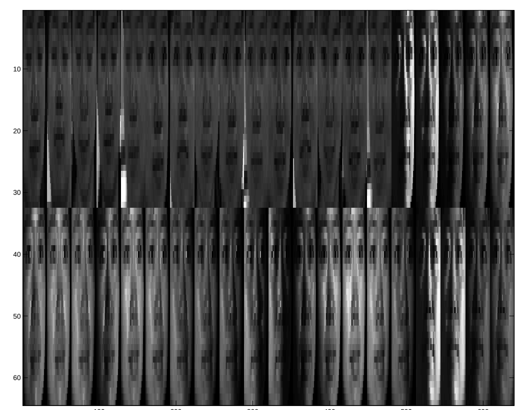
- The CMU PIE face database [2] contains 68 people with 41368 face images as a whole. The face images were captured by 13 synchronized cameras and 21 flashes, under varying poses, illuminations and expressions. In our experiments, five near frontal pose (C05, C07, C09, C27, C29) are selected under different illuminations and expressions which leaves us about 170 near frontal face images for each individual, and all the images were also resized to 32×32.

1.2 Outline

The rest of thesis is organized as follows. Chapter 2 explains KNMF in detail. Chapter 3 draws a conclusion of this work.



(a) ORL face images.



(b) PIE face images.

Fig. 1.1: Sample face images from the ORL and PIE database. (a) shows ORL face images, for each people there are 10 different illumination, facial expressions and poses for each subject. Most of all them are near front poses. (b) shows PIE face images. There are 170 face images for each people. Most of them are also near frontal poses.

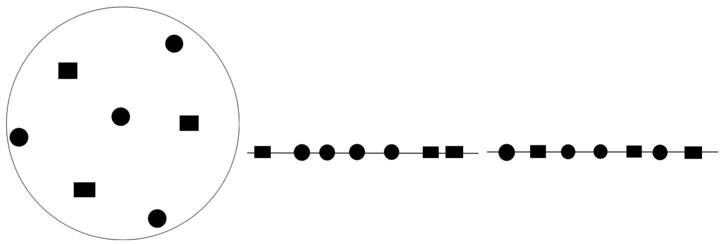
CHAPTER 2

Nonnegative matrix factorization regularized by k-NN graphs

NMF decompose a nonnegative data matrix into a basis matrix and a feature matrix with all of these matrices allow to have only nonnegative elements [3]. The nonnegative constraints in NMF require only additive, not subjective combination of basis vectors lead to a part-based representation of original data matrix [3]. Recently there are various NMF-type algorithms proposed. [4] gives sparseness constraints into NMF. [5] gives three constraints to both basis and feature matrix to learn more spatially localized features of given face images. While the above algorithms are unsupervised methods, supervised variations of NMF were proposed for improving the discriminative power of NMF [6], which adds fisher criterion in the subspace spanned by the basis matrix. Very recently, [7] proposed semi-supervised NMF (SSNMF) which jointly incorporate the data matrix and the (partial) class label matrix into NMF to handle the case when just a portion of data are labeled, result in semi-supervised feature extraction.

However, all of the feature extraction methods base on NMF assume that data are generated from Euclidean space. Recent research on manifold learning have shown that many of real world data likely to be generated from the low-dimensional intrinsic manifold embedded in high-dimensional input space [8, 9]. The labels of data points vary smoothly in regions of high data density, while the labels of data points vary more in regions of low data density. How to respect the intrinsic manifold structure of data in feature extraction is still an open problem. The limitation of NMF based on Euclidean space was first pointed out at [10]. GNMF [10] incorporates intrinsic manifold structure of data as regularization term into NMF. However, GNMF is used as clustering method and could not extract discriminative features for classification problem.

The final goal of learning (classification or clustering) is the predicted labels of different classes or clusters contrast to each other. Therefore, a good feature extraction method should benefit classifier assigns labels vary smoothly intra-class, while the labels predicted for inter-class diverse to each other. In this paper, we propose a novel feature extraction method named NMF regularized by k-NN graphs (KNMF). The basic idea is discriminate different classes directly by shrinking the local regions of intra-class neighborhood and expanding the local regions of inter-class neighborhood. The intrinsic manifold structure of data points in the same class are well approximated by constructing intra-class k-NN graph, while the intrinsic manifold structure of data points belong to different classes are well approximated by constructing inter-class k-NN graph. The intra-class manifold compactness and inter-class manifold contrastness are incorporated into NMF as regularization for feature extraction. Intuitive example of the useful behavior of our proposed method is shown (Fig.1). Efficient multiplicative update rules are derived which is easy to implement and the convergence is guaranteed theoretically. Experiments on several benchmark face recognition datasets and document datasets confirm the useful behavior of our proposed method in the task of feature extraction.



(a) Original input space. (b) Subspace learned by KNMF. (c) Subspace learned by GNMF.

Fig. 2.1: Illustration of the useful behaviors of KNMF. (a) data points in original input space, (b) projected data points in a line (1-D subspace) based on KNMF, (c) projected data points in a line (1-D subspace) based on GNMF.

We organize the paper as follows: Section 2.1 briefly reviews NMF; Section 2.2 shows the proposed DGNMF algorithm in detail; Section 2.3 give the semi-supervised setting of DGNMF; Section 2.4 presents experimental results of face recognition and document classification; Section 2.5 draws the summary of this work.

2.1 Nonnegative matrix factorization

NMF decompose a nonnegative data matrix into a basis matrix and a feature matrix with constraints that these matrices only have nonnegative elements. Suppose we have a nonnegative data matrix $\boldsymbol{X} = [\mathbf{x}_i, ..., \mathbf{x}_n] \in R^{d\times n}$, we want to decompose $\boldsymbol{X}$ into basis matrix $\boldsymbol{U} \in R^{d\times q}$ and feature matrix $\boldsymbol{V} \in R^{n\times q}$ which minimize

$$J_{NMF} = ||\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^\top||_F^2, \tag{2.1}$$

where $||.||_F$ is matrix Frobenius norm, both basis matrix $\boldsymbol{U}$ and feature matrix $\boldsymbol{V}$ are constrained to be nonnegative that $\boldsymbol{U} \geq 0$ and $\boldsymbol{V} \geq 0$.

Multiplicative updates for $\boldsymbol{U}$ and $\boldsymbol{V}$ which minimize (1) are derived as follows. Suppose that the gradient of an error function has a decomposition that is of the form

$$\triangledown \mathcal{J} = [\triangledown \mathcal{J}]^{+} - [\triangledown \mathcal{J}]^{-},$$

where $[\triangledown \mathcal{J}]^{+} > 0$ and $[\triangledown \mathcal{J}]^{-} > 0$. Then the multiplicative updates for the parameters Θ has the form

$$\Theta \leftarrow \Theta \odot (\frac{[\triangledown \mathcal{J}]^{-}}{[\triangledown \mathcal{J}]^{+}}). \tag{2.2}$$

It can be easily seen that the multiplicative update (2) preserves the nonnegativity of the parameter Θ, while $\triangledown \mathcal{J} = 0$ when the convergence is achieved.

Derivative of the objective function (1) with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$ are given by

$$\begin{aligned} \triangledown_U \mathcal{J} &= [\triangledown_U \mathcal{J}]^{+} - [\triangledown_U \mathcal{J}]^{-} \\ &= [\boldsymbol{U}\boldsymbol{V}^{\top}]\boldsymbol{V} - [\boldsymbol{X}]\boldsymbol{V} \\ \triangledown_V \mathcal{J} &= [\triangledown_V \mathcal{J}]^{+} - [\triangledown_V \mathcal{J}]^{-} \\ &= [\boldsymbol{V}\boldsymbol{U}\boldsymbol{U}^{\top}] - [\boldsymbol{X}^{\top}\boldsymbol{U}]. \end{aligned}$$

With these gradient calculations, invoking (2) yields multiplicative updates for NMF:

$$\boldsymbol{U} \leftarrow \boldsymbol{U} \odot \frac{\boldsymbol{X}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{V}^{\top}\boldsymbol{V}} \tag{2.3}$$

$$\boldsymbol{V} \leftarrow \boldsymbol{V} \odot \frac{\boldsymbol{X}^{\top}\boldsymbol{U}}{\boldsymbol{V}\boldsymbol{U}^{\top}\boldsymbol{U}}, \tag{2.4}$$

where $\odot$ denotes the Hadamard product (elementwise product) and $\frac{\boldsymbol{U}}{\boldsymbol{V}}$ represents the elementwise division, i.e. $[\frac{\boldsymbol{U}}{\boldsymbol{V}}]_{ij} = \frac{\boldsymbol{U}_{ij}}{\boldsymbol{V}_{ij}}$. $\boldsymbol{u}_i \in R^d$ is a basis vector of basis matrix $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_q]$ and $\boldsymbol{v}_i \in R^q$ is a feature vector of feature matrix $\boldsymbol{V} = [\boldsymbol{v}_1^{\top}, \boldsymbol{v}_2^{\top}, ..., \boldsymbol{v}_n^{\top}]^{\top}$.

2.2 NMF regularized by k-NN graphs

The fundamental goal of learning algorithm (classification or clustering) is that the predicted labels of different classes or clusters contrast to each other. [9]

proposed a nice framework taking data manifold smoothness account. The main motivation of GNMF [10] is based on the assumption that if the data points are close in input space, then the projected data points in subspace need to be close to each either. GNMF incorporates manifold smoothness as regularization term into NMF. However, these geometrical regularization does not directly discriminate different classes, hence they might not exploit the discriminative hidden patterns of data. We directly construct intra-class nearest neighbor graph which exploit the geometrical structure of inter-class data points, while the geometrical structure of inter-class data points are well approximated by constructing between-class nearest neighbor graph. By shrinking the regions of intra-class neighborhood and expanding the regions of inter-class neighborhood, we could discriminate different classes directly.

2.2.1 Intra-class and inter-class k-NN graph

We approximate the geometry of manifold by a neighborhood graph of data points (sampled from the underlying manifold) as in [8]. With $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n] \in R^{d \times n}$, where d the dimension of input data, while n is the number of data points. For each data point $\boldsymbol{x}_i$, we use $\mathcal{N}_k^w(\boldsymbol{x}_i)$ to denote the set contains its within-class k nearest neighbors.

In order to construct a intra-class k-NN graph, we define the corresponding adjacency matrix by

$$W_{ij}^w = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \in \mathcal{N}_k^w(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in \mathcal{N}_k^w(\boldsymbol{x}_i) \\ 0 & \text{otherwise.} \end{cases}$$

We define row(or column) sum of $\boldsymbol{W}^w$ as $\boldsymbol{D}^w$, and corresponding unnormalized graph laplacian $\boldsymbol{L}^w = \boldsymbol{D}^w - \boldsymbol{W}^w$, which is a discrete approximation to the Laplace-Beltrami operator on the within-class data manifold [8].

Different type of weights could be assigned to the adjacency matrix other than binary weights. We use cosine similarity between two documents represented as vector space model to assign the weights of the adjacency matrix in document classification. The cosine similarity is measured as follows:

$$W_{ij}^{w}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\frac{\boldsymbol{x}_i^{\top}\boldsymbol{x}_j}{||\boldsymbol{x}_i||\cdot||\boldsymbol{x}_j||}.$$

The range of cosine similarity is [0, 1]. If the two documents are exactly same, then $W_{ij}^{w}(\boldsymbol{x}_i,\boldsymbol{x}_j)=1$. On the other hand, $W_{ij}^{w}(\boldsymbol{x}_i,\boldsymbol{x}_j)=0$ means the two documents $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ shares no common word.

Similar to intra–class k-NN graph, for each data point $\boldsymbol{x}_i$, we use $\mathcal{N}_k^b(\boldsymbol{x}_i)$ to denote the set contains its inter-class k nearest neighbors. Then, for a inter-class k-NN graph, we define the corresponding adjacency matrix by

$$W_{ij}^{b}=\begin{cases}1 & \text{if } \boldsymbol{x}_i\in\mathcal{N}_k^b(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j\in\mathcal{N}_k^b(\boldsymbol{x}_i)\\ 0 & \text{otherwise.}\end{cases}$$

We define row(or column) sum of $\boldsymbol{W}^b$ as $\boldsymbol{D}^b$, and corresponding unnormalized graph laplacian $\boldsymbol{L}^b=\boldsymbol{D}^b-\boldsymbol{W}^b$.

2.2.2 Algorithm

We extract the features discriminate different classes directly by shrinking the local regions of intra-class neighborhood and expanding the local regions of inter-class neighborhood. Minimize the following term, archive the goal:

$$\begin{aligned}
\boldsymbol{G} &= \frac{1}{2}\sum_{i=1}^{N}\Big(\sum_{\boldsymbol{x}_j\in\mathcal{N}_k^w(\boldsymbol{x}_i)}||\boldsymbol{v}_i-\boldsymbol{v}_j||^2-\sum_{\boldsymbol{x}_j\in\mathcal{N}_j^b(\boldsymbol{x}_i)}||\boldsymbol{v}_i-\boldsymbol{v}_j||^2\Big)\\
&= \frac{1}{2}\sum_{i,j=1}^{N}\left(||\boldsymbol{v}_i-\boldsymbol{v}_j||^2W_{ij}^w-||\boldsymbol{v}_i-\boldsymbol{v}_j||^2W_{ij}^b\right)\\
&= \Big(\sum_{i=1}^{N}\boldsymbol{v}_i^{\top}\boldsymbol{v}_iD_{ii}^w-\sum_{i,j=1}^{N}\boldsymbol{v}_i^{\top}\boldsymbol{v}_jW_{ij}^w\Big)\\
&- \Big(\sum_{i=1}^{N}\boldsymbol{v}_i^{\top}\boldsymbol{v}_iD_{ii}^b-\sum_{i,j=1}^{N}\boldsymbol{v}_i^{\top}\boldsymbol{v}_jW_{ij}^b\Big)\\
&= tr(\boldsymbol{V}^{\top}\boldsymbol{L}^w\boldsymbol{V})-tr(\boldsymbol{V}^{\top}\boldsymbol{L}^b\boldsymbol{V})\\
&= tr[\boldsymbol{V}^{\top}(\boldsymbol{L}^w-\boldsymbol{L}^b)\boldsymbol{V}],
\end{aligned}$$

where $tr(\cdot)$ is trace of a matrix. The intra-class manifold compactness and inter-class manifold contrastness are incorporated into NMF as regularization for feature extraction. The objective function of KNMF is:

$$\begin{aligned} \mathcal{J}_{KNMF} &= ||\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^\top||_F^2 + \alpha \boldsymbol{G} && (2.5) \\ \mathcal{J}_{KNMF} &= ||\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^\top||_F^2 + \alpha tr[\boldsymbol{V}^\top(\boldsymbol{L}^w - \boldsymbol{L}^b)\boldsymbol{V}] && (2.6) \\ s.t. &\quad \boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0, \end{aligned}$$

where α is the regularization parameter. Both of the edge weights of intra-class k-NN graph and inter-class k-NN graph is nonnegative, i.e. $[W]_{ij}^w \geq 0$, $[W]_{ij}^b \geq 0$, then the elements of degree matrix is also nonnegative for $[D]_{ii}^w = \sum_j W_{ij}^w \geq 0$ and $[D]_{ii}^b = \sum_j W_{ij}^b \geq 0$.

With similar methods to derive the multiplicative update rules for NMF, we could easily derive the update rule for KNMF as follows:

$$\begin{aligned} \boldsymbol{U} &\leftarrow \boldsymbol{U} \odot \frac{\boldsymbol{X}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{V}} && (2.7) \\ \boldsymbol{V} &\leftarrow \boldsymbol{V} \odot \frac{\boldsymbol{X}^\top\boldsymbol{U} + \alpha(\boldsymbol{D}^b + \boldsymbol{W}^w)\boldsymbol{V}}{\boldsymbol{V}\boldsymbol{U}\boldsymbol{U}^\top + \alpha(\boldsymbol{D}^w + \boldsymbol{W}^b)\boldsymbol{V}}. && (2.8) \end{aligned}$$

2.3 Semi-supervised extension

When we face situation that only a portion of data is labeled, and the remaining data is unlabeled, how could we extract discriminative features? We proposed a semi-supervised setting of KNMF. Semi-supervised learning [11], which can exploit the large amount of unlabeled samples to improve classification, is successfully used in document classification. In this section, we will present a semi-supervised version of KNMF (SSKNMF).

KNMF considers finding the most discriminative features purely on the labeled training set. While text classification frees organizations from the need of manually organizing document bases, it still needs professionals to label a large enough training data set for learning a classifier, which requires much expensive human labor and much time. Furthermore, compared with the large amount of documents increasing every day, the labeled samples are always insufficient. That is, in reality, it is possible to acquire a large set of unlabeled data rather than labeled data. To address this problem, semi-supervised learning, which aims to learn from partially labeled data, provides a solution. Thus, we extend KNMF to incorporate the unlabeled data. In general, semi-supervised learning can be categorized into two classes:

- transductive learning: to estimate the labels of the given unlabeled data;
- inductive learning: to induce a decision function which has a low error rate on the whole sample space.

Our method belongs to inductive learning. Other semi-supervised inductive methods used for text classification is semi-supervised discriminant analysis (SDA) [12].

Semi-supervised learning is formulated as follows. Given a point set $\mathcal{X} = \boldsymbol{x}_1, ..., \boldsymbol{x}_l, \boldsymbol{x}_{l+1}, ..., \boldsymbol{x}_n \in R^d$, and a label set $\mathcal{L} = 1, ..., c$, the first l points $\boldsymbol{x}_i, 1 \leq i \leq l$ are labeled as $y_i \in \mathcal{L}$ and the remaining points $\boldsymbol{x}_u, l+1 \leq u \leq n$ are unlabeled. Based on the assumption that nearby points are likely to have the same embedding, which means they are likely to be the features belong to the same class. Hence, for all unlabeled and labeled data, we construct the k nearest neighbor graph as follows:

$$W_{ij}^{(2)} = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \in \mathcal{N}_k(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in \mathcal{N}_k(\boldsymbol{x}_i) \\ 0 & \text{otherwise} \end{cases}$$

as in GNMF. The size of $\boldsymbol{W}^{(2)}$ is $n \times n$. We define the graph laplacian matrix of adjacency matrix of $\boldsymbol{W}^{(2)}$ as $\boldsymbol{L}^{(2)} = \boldsymbol{D}^{(2)} - \boldsymbol{W}^{(2)}$, where $\boldsymbol{D}^{(2)}$ is the row (or column) sum of $\boldsymbol{W}^{(2)}$. We incorporates this manifold smoothness regularization term into SSKNMF to make use the unlabeled data for exploiting more discriminative and geometrical structure of both labeled and unlabeled data.

For the labeled data points we construct the intra-class k-NN graph and the inter-class k-NN graph as in section 2.2. We could just make one adjacency matrix to incorporate both the intra-class adjacency matrix and inter-class adjacency matrix as follows:

$$W_{ij}^{l} = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \in \mathcal{N}_k^w(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in \mathcal{N}_k^w(\boldsymbol{x}_i) \\ -1 & \text{if } \boldsymbol{x}_i \in \mathcal{N}_k^b(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in \mathcal{N}_k^b(\boldsymbol{x}_i) \\ 0 & \text{otherwise} \end{cases}$$

where the size of $\boldsymbol{W}^l$ is $l \times l$. Based on this, we define an adjacency matrix for only labeled data:

$$\boldsymbol{W}^1 = \begin{bmatrix} \boldsymbol{W}^{(1)} & \underline{0} \\ \underline{0} & \underline{0}. \end{bmatrix}$$

We also define the graph laplacian matrix of adjacency matrix of $\boldsymbol{W}^{(1)}$ as $\boldsymbol{L}^{(1)} = \boldsymbol{D}^{(1)} - \boldsymbol{W}^{(1)}$, where $\boldsymbol{D}^{(1)}$ is the row (or column) sum of $\boldsymbol{W}^{(1)}$. The objective function of SSKNMF is as follows:

$$\begin{aligned} \mathcal{J}_{SSKNMF} &= ||\boldsymbol{X} - \underline{\boldsymbol{U}}\boldsymbol{V}^\top||_F^2 + \alpha tr[\boldsymbol{V}^\top(\boldsymbol{L}^{(1)} + \boldsymbol{L}^{(2)})\boldsymbol{V}] \\ & s.t. \quad \boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0, \end{aligned} \tag{2.9}$$

Based on similar method for deriving the update rule of SSKNMF, we could easily derive the following multiplicative update rule:

$$\boldsymbol{U} \leftarrow \boldsymbol{U} \odot \frac{\boldsymbol{X}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{V}} \tag{2.10}$$

$$\boldsymbol{V} \leftarrow \boldsymbol{V} \odot \frac{\boldsymbol{X}^\top\boldsymbol{U} + \alpha(\boldsymbol{W}^{(1)} + \boldsymbol{W}^{(2)})\boldsymbol{V}}{\boldsymbol{V}\boldsymbol{U}\boldsymbol{U}^\top + \alpha(\boldsymbol{D}^{(1)} + \boldsymbol{D}^{(2)})\boldsymbol{V}}. \tag{2.11}$$

2.4 Experiments

In this section, we evaluate the performance of the proposed method on face recognition and document classification.

Table 2.1 Characteristics of ORL and PIE

Data sets	#data/class	#attributes	#classes
ORL	10	1024	40
PIE	170	1024	68

2.4.1 Face recognition

In face recognition task, we compare our method with Principal Component Analysis (PCA) [13], Linear Discriminant Analysis (LDA) [13], NMF [3], LNMF [5], FNMF [6], GNMF [10].

Data sets

In our experiments, we use two benchmark face recognition data sets which are widely used in subspace learning for face recognition task. The characteristics of two face recognition databases are briefly summarized in Table 1.

The ORL face database [1] provides totally 40 people (classes) with 10 images for each people, which were taken at different times, varying the lightings, facial expressions and facial details. The size of original images (with 256 gray levels) are 92×112, which are resized to 32×32 for efficiency.

The CMU PIE face database [2] contains 68 people with 41368 face images as a whole. The face images were captured by 13 synchronized cameras and 21 flashes, under varying poses, illuminations and expressions. In our experiments, five near frontal pose (C05, C07, C09, C27, C29) are selected under different illuminations and expressions which leaves us about 170 near frontal face images for each individual, and all the images were also resized to 32×32.

[1] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

Experiment settings

Here we briefly explain the experiment setting and parameter selection. For each data set, we randomly divide it into training sets $\boldsymbol{X}_{train}$ and testing sets $\boldsymbol{X}_{test}$. We learn the feature matrix of $\boldsymbol{X}_{test}$ as $\boldsymbol{V}_{test} = \boldsymbol{U}^{\dagger}\boldsymbol{X}_{test}$, where $\boldsymbol{U}^{\dagger} = (\boldsymbol{U}^{\top}\boldsymbol{U})^{-1}\boldsymbol{U}^{\top}$ and $\boldsymbol{U}$ is learned from $\boldsymbol{X}_{train}$. Then we evaluate the face recognition accuracy on $\boldsymbol{V}_{test}$. More specifically, for each people in the ORL database, $t = 2, 3, 4$ images are randomly selected as $\boldsymbol{X}_{train}$, while for each people in the PIE database, $t = 10, 20, 50$ images are randomly selected as $\boldsymbol{X}_{train}$ and the rest are assigned to $\boldsymbol{X}_{test}$. As mentioned in [13], for LDA we first use PCA to reduce the size of dimensions to $n - c$ and then perform LDA except for the case of PIE database with $p = 20, 50$. Because in these two cases, $n - c$ is larger than the original size of dimension d, therefore we directly apply LDA to the dataset. The neighborhood size k of intra-class k-NN graph and inter-class k-NN graph was set to be the same. k is set by searching the grid { 1,2, . . . ,10 }. The regularization parameter α is set by searching the grid { 0.01,0.1,1,10,100 }. KNMF is robust to different values of k and α.

Face recognition results

The best average performance of 20 trials is reported with the corresponding size of dimension of features in (Tab.2.2) and (Tab.2.3) respectively, where the value in each entry represents the average recognition accuracy of 20 independent trials, and the number in brackets is the corresponding projection dimensionality.

2.4.2 Document classification

In this section, we compare the good behavior of our proposed KNMF with PCA, LDA, NMF, GNMF in document classification task. Also we choose directly apply support vector machine (SVM) on the term-document matrix. Linear

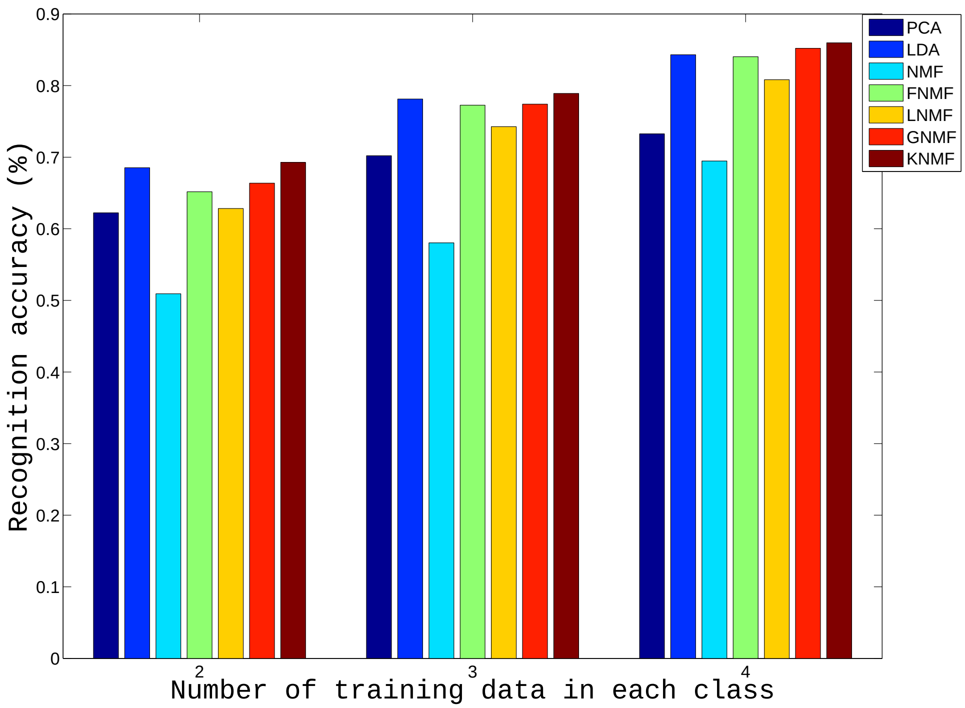


Fig. 2.2: Face recognition accuracy on ORL data.

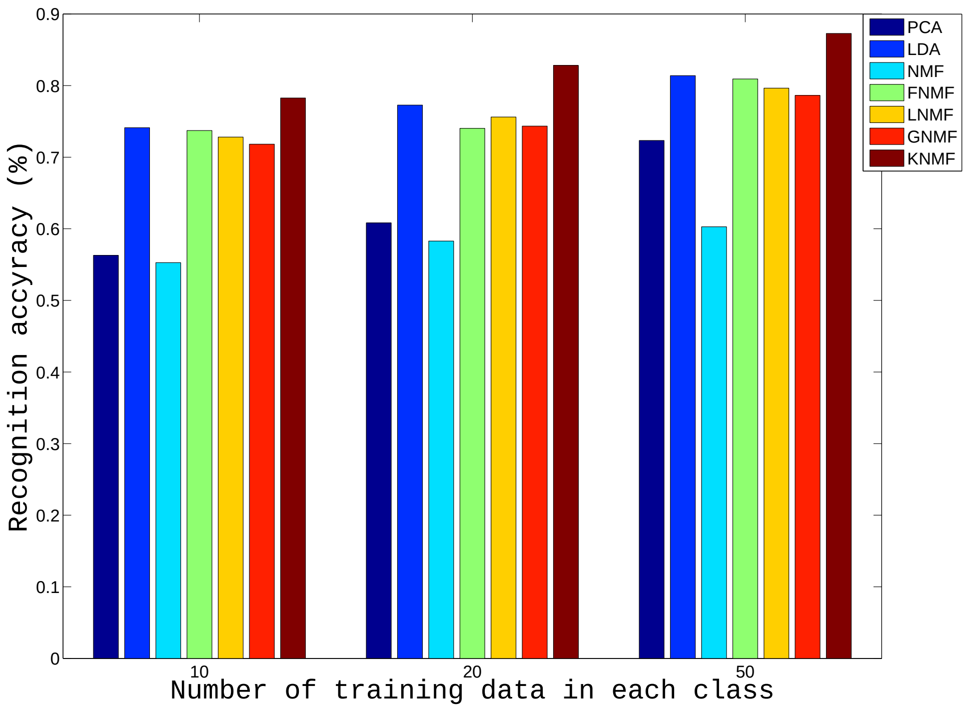


Fig. 2.3: Face recognition accuracy on PIE data.

Table 2.2 Face Recognition accuracy on the ORL data set. The number in brackets is the corresponding projection dimensionality.

Algorithm	2 Train	3 Train	4 Train
PCA	0.6223 (85)	0.7021 (135)	0.7326 (160)
LDA	0.6853 (34)	0.7813 (39)	0.8431 (39)
NMF	0.5093 (100)	0.5804 (120)	0.6947 (105)
FNMF	0.6518 (65)	0.7727 (95)	0.8403 (110)
LNMF	0.6284 (120)	0.7426 (165)	0.8082 (180)
GNMF	0.6638 (150)	0.7741 (145)	0.8521 (175)
DGNMF	**0.6927 (135)**	**0.7891 (125)**	**0.8598 (110)**

Table 2.3 Face Recognition accuracy on the PIE data set. The number in brackets is the corresponding projection dimensionality.

Algorithm	5 Train	10 Train	20 Train
PCA	0.4081 (180)	0.5629 (190)	0.6082 (185)
LDA	0.6327 (52)	0.7412 (67)	0.7728 (67)
NMF	0.5029 (120)	0.5527 (135)	0.5829 (145)
FNMF	0.6582 (155)	0.7372 (95)	0.7403 (175)
LNMF	0.6421 (115)	0.7281 (190)	0.7562 (165)
GNMF	0.6624 (165)	0.7182 (125)	0.7435 (110)
DGNMF	**0.6837 (190)**	**0.7827 (170)**	**0.8282 (175)**

SVM is used and we use the software of libSVM [2] After extract features, a linear support vector machine (SVM) is trained for classification based on the low-dimensional representation. We adopt 5-fold cross validation to evaluate our method. More specifically, we randomly split the data into five folds (subsets), and then repeat the test five times, in each one of which we use one fold for testing and the other four folds for training. We repeat the experiments 10 times and the average classification accuracies over the 10 repeats are adopted as the performance metric.

[2]http://www.csie.ntu.edu.tw/ cjlin/libsvm

Table 2.4 Characteristics of WebKB dataset

	#data	#attributes	#classes
Cornell	827	4134	7
Texas	814	4029	7
Washington	1166	4165	7
Wisconsin	1210	4189	6

Table 2.5 Characteristics of Cora dataset

	#data	#attributes	#classes
DS	751	6234	9
HA	400	3989	7
ML	1617	8329	7
PL	1575	7949	9

Data sets

We use WebKB, Cora data sets. Vector space model was applied to document data as representation method. We eliminate all the documents belong to multiple classes, because in this document classification we do not handle the multi-label problem. The characteristics of WebKB data set are provided (Table 2.4). The characteristics of Cora data set are provided (Table 2.5). Detail descriptions of datasets please refer to the Chap 1.

Sensitivity of parameters

We examine the sensitivity of KNMF to k in k nearest neighbor graph; the size of feature (the number of low dimension) q; the regularization parameter α. We use the HA data in Cora dataset for illustration. Figure 2.5 illustrates the accuracy of KNMF when k, α and q take different values. From Figure 2.5(a), we can see that the smaller the $alpha$, the worse the performance will be. Larger α means that the discriminative manifold smoothness regularization plays a more significant role. Hence, we can conclude that the discriminative manifold smoothness regularization is very important. We could also see that

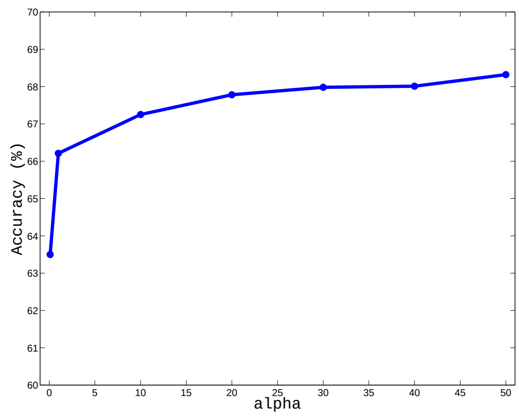
Table 2.6 Classification accuracy average of 10 runs on WebKB dataset

	DS	HA	ML	OS	PL
Baseline	46.87	60.15	64.16	64.12	48.78
PCA	47.30	61.50	60.12	65.83	50.76
LDA	49.48	63.32	67.34	**73.28**	50.19
NMF	43.92	56.77	66.90	71.88	47.11
GNMF	49.12	63.87	69.45	69.84	50.38
DGNMF	**53.74**	**68.32**	**71.12**	67.23	**54.28**

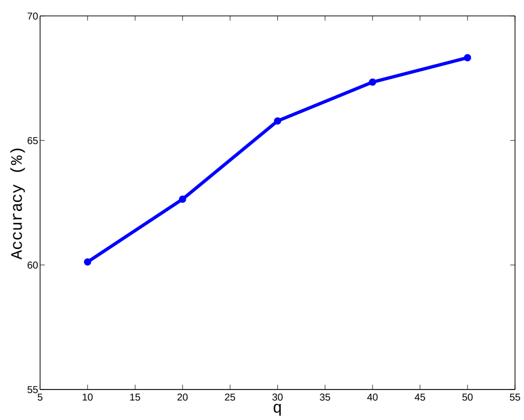
KNMF is not sensitive to the value of $alpha$. From Figure 2.5(b), we can see that as q increases, the accuracy also increases. But larger q will incur higher computation cost. Hence, in real applications, we need to consider the tradeoff between computation cost and accuracy. From Figure 2.5(c), we could also see that KNMF is robust to different value of k, the size of neighborhood of intra-class and inter-class.

Classification results

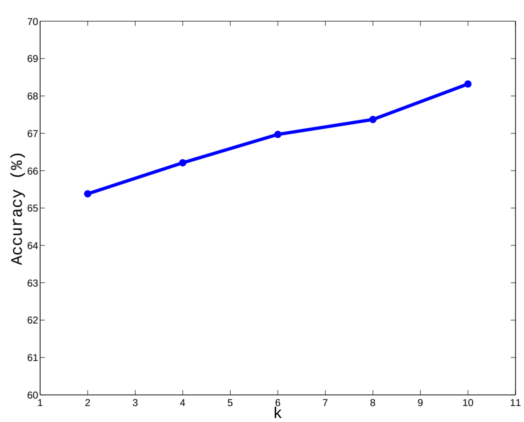
From table 2.6 and table 2.7 we can see that KNMF does outperform other methods in most datasets. Even though the LDA method uses label information either, KNMF can still give much better result in most datasets, showing that they are indeed very effective. PCA aims at optimal representation of the original data in the lower dimensional space in the means square error sense. Comparing with PCA (LSI), we can see that the good performance of KNMF. GNMF, KNMF all performed dramatically better than NMF, we can see that the manifold smoothness regularization to NMF indeed conformed that data points are lie on low-dimensional manifold embedded in high dimensional input space. Comparing KNMF to GNMF, we can see that the intra-class k-NN graph and inter-class k-NN graph indeed extract more discriminative features than GNMF which improved the classification accuracy further.



(a) Effect of α



(b) Effect of q



(c) Effect of k

Fig. 2.4: Sensitivity of parameters of DGNMF.

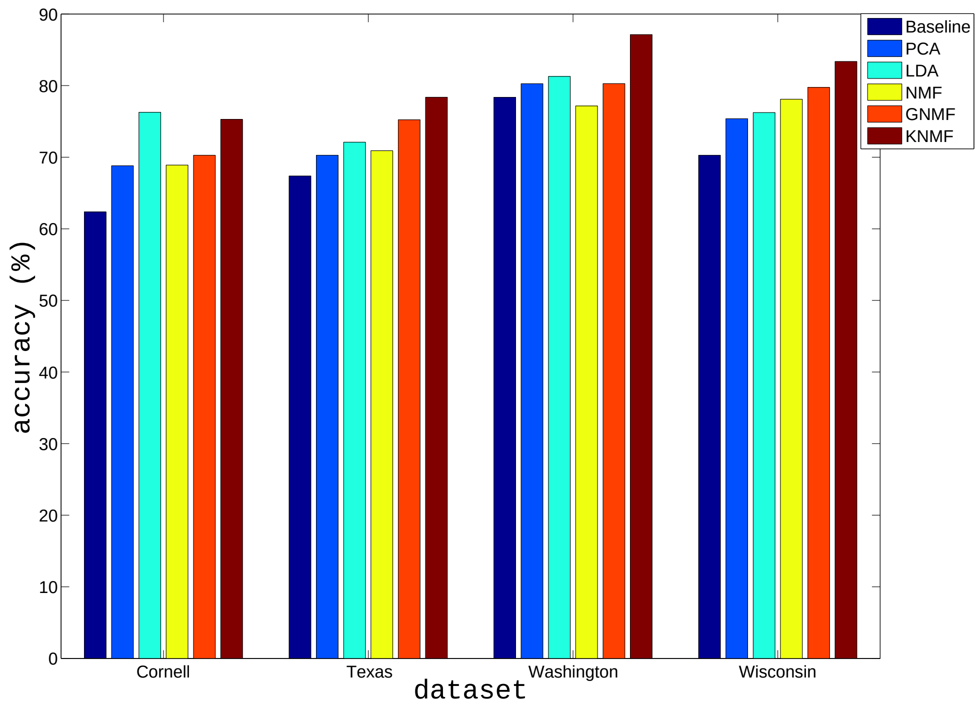


Fig. 2.5: Classification accuracy on WebKB dataset.

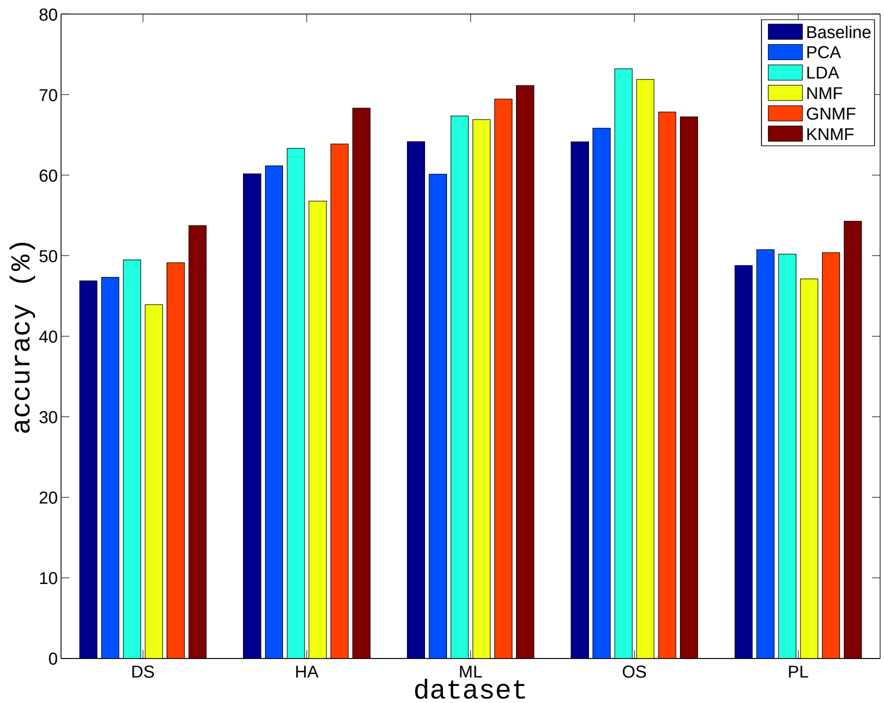


Fig. 2.6: Classification accuracy on Cora dataset.

Table 2.7 Classification accuracy accuracy average of 10 runs on Cora dataset

	Cornell	Texas	Washington	Wisconsin
Baseline	62.36	67.38	78.37	70.28
PCA	68.81	70.28	80.27	75.38
LDA	**76.28**	72.10	81.28	76.23
NMF	68.90	70.91	77.16	78.10
GNMF	70.28	75.23	80.28	79.75
DGNMF	75.28	**78.38**	**87.12**	**83.37**

2.5 Summary

We have presented a novel feature extraction method. The proposed method extracts more discriminative features by shrinking the local regions of intra-class neighborhood which is well approximated by constructing intra-class k-NN graph, and expanding the local regions of inter-class neighborhood which is well approximated by inter-class k-NN graph. Experiments of several bench mark face recognition datasets conform the useful behavior of our proposed method in the task of feature extraction.

CHAPTER 3

Conclusion

In this paper, we propose a novel feature extraction (subspace learning) method named nonnegative matrix factorization regularized by k-NN graphs (KNMF). We construct two types of neighborhood adjacency graph, where within-class nearest neighbors connects only the neighboring data points which belong to the same class of the given data point, while between-class nearest neighbors connects the neighboring data points which belong to different class of the given data point. Hence, by minimizing the local regions of within-class neighborhood and maximizing the local regions of between-class neighborhood, KNMF could exploit more discriminative hidden patterns of given data set, benefit to the following classification based on the extracted features. Experiments on several benchmark face recognition datasets and document datasets confirmed the useful behavior of our proposed method for feature extraction for face recognition and document classification.

Bibliography

[1] S. Zhu, K. Yu, Y. Chi, and Y. Gong, "Combining content and link for classification using matrix factorization," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, Amsterdam, The Netherlands, 2007.

[2] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, 2003.

[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[4] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[5] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. S. Cheng, "Learning spatially localized parts-based representation," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Kauai, Hawaii, 2001, pp. 207–212.

[6] Y. Wang, Y. Jia, C. Hu, and M. Turk, "Fisher non-negative matrix factorization for learning local features," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, Jeju Island, Korea, 2004.

[7] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," 2010, vol. 17, pp. 4–7.

[8] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, pp. 1373–1396, 2003.

[9] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[10] D. Cai, X. He, X. Wu, and J. Han, "Non-negative matrix factorization on manifold," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, Pisa, Italy, 2008, pp. 63–72.

[11] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2003, pp. 912–919.

[12] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proceedings of the International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, 2007.

[13] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, Oct. 1997.

감 사 의 글

First of all, I would like to thank my advisor Professor Seungjin Choi. For his kind advise and keep encouraging me to pursuit my own goal. Being honest, I am a little scared of my advisor, but I learned from him work harder and smarter is the only way to being a good researcher. Except research and study of knowledge, I learned lots of from my advisor of being a better person. Also, I would like to thank Professor Seungwon Hwang, Professor Huanjo Yu gave me kind reviews and comments to my thesis.

Thank every members in machine learning laboratory. Whenever I met problems, I just come to them and ask helps, without their kind helps I could not finish my study in POSTECH.

It is my honor to receive Asia-Fellowship from POSCO TJ-Park Foundation for two years study at POSTECH. Without this fellowship, I could not even start study in Korea. Thank you for providing financial support and kind caring about our POSCO-Asia fellows' life in Korea, even our future life.

Thank administration officers in IAO and department of computer science. They always help me and give me kind supports.

Being last, but the most important to me is my family. Dad and mom, your warm supports and real concerns about me is the best blessed gift to my life.

Curriculum Vitae

Name	:	Shounan An

Education

2003.9.1 2007.7.1	:	B.S. in Department of Computer Science and Information Technology, Southwest University
2007.9,1 2010.2(예정)	:	M.S. in Department of Computer Science and Engineering, POSTECH Thesis Title : **Nonnegative Matrix Factorization Regularized by k-NN Graphs** Advisor: Prof. Seungjin Choi