Pattern Recognition Letters

Authorship Confirmation

Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.

As corresponding author I, Mohammadreaza Babaee, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature Mohammadreza Babaee Date: 20/01/2016

List any pre-prints:

Relevant Conference publication(s) (submitted, accepted, or published):

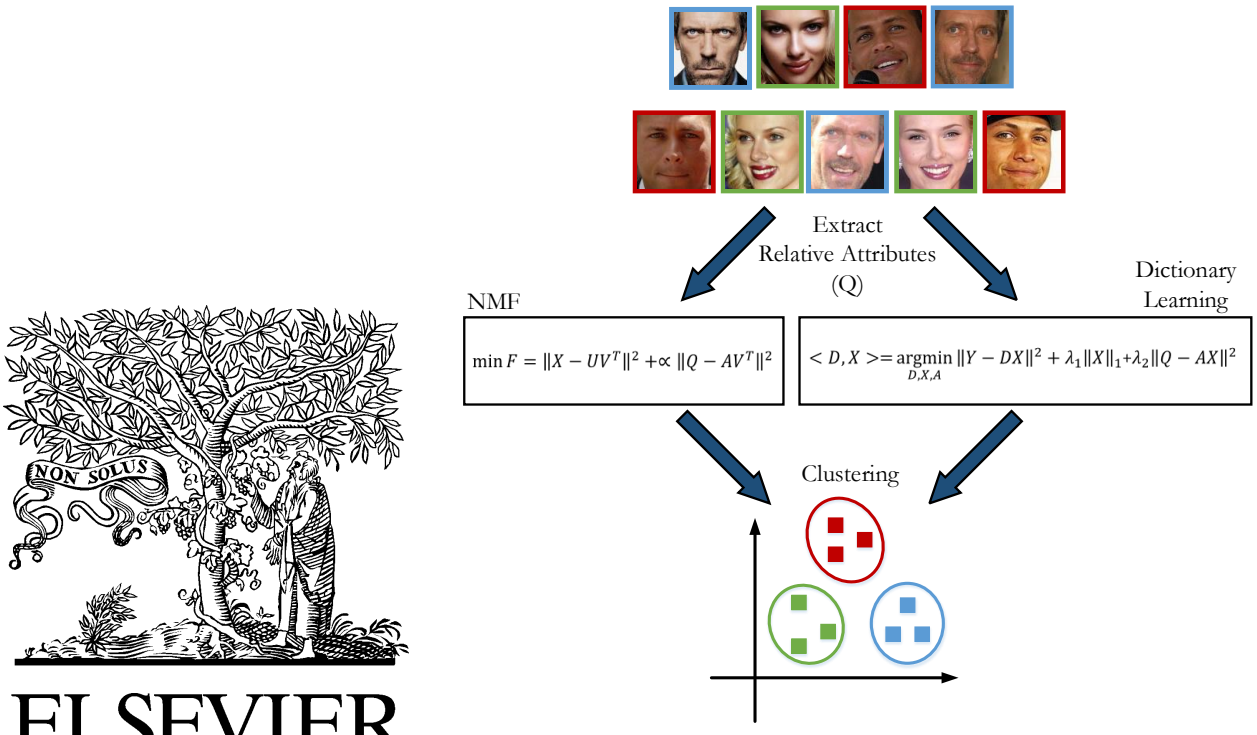
Justification for re-publication:

Graphical Abstract (Optional)

To create your abstract, please type over the instructions in the template box below. Fonts or abstract dimensions should not be changed or altered.

Nonlinear Subspace Clustering Using Curvature Constrained Distances

Author's names here



ELSEVIER

Research Highlights (Required)

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- We proposed new algorithms for dictionary learning and non-negative matrix factorization.
- We utilize relative attributes as semantic information to enhance the discriminative property of obtained signals.
- We apply our method to several datasets to show the efficiency of proposed algorithms.
- k-means clustering is used to measure the efficiency of new features
- Experimental results show the efficiency of relative attributes instead of binary labels in providing discriminative dictionary or subspace

Pattern Recognition Letters
journal homepage: www.elsevier.com

Toward Semantic Attributes in Dictionary Learning and Non-negative Matrix Factorization

Mohammadreza Babaee[a,**], Thomas Wolf[a], Gerhard Rigoll[a]

[a]*Institute for Human-Machine Communication, Technische Universität München, Munich, Germany*

ABSTRACT

Binary label information is widely used semantic information in discriminative dictionary learning and non-negative matrix factorization. A Discriminative Dictionary Learning (DDL) algorithm uses the label of some data samples to enhance the discriminative property of sparse signals. A discriminative Non-negative Matrix Factorization (NMF) utilizes label information in learning discriminative bases. All these technique are using binary label information as semantic information. In contrast to such binary attributes or labels, relative attributes contain richer semantic information where the data is annotated with the strength of the attributes. In this paper, we utilize the relative attributes of training data in non-negative matrix factorization and dictionary learning. Precisely, we learn rank functions (one for each predefined attribute) to rank the images based on predefined semantic attributes. The strength of each attribute in a data sample is used as semantic information. To assess the quality of the obtained signals, we apply k-means clustering and measure the performance for clustering. Experimental results conducted on three datasets, namely PubFig (16), OSR (24) and Shoes (15) confirm that the proposed approach outperforms the state-of-the-art discriminative algorithms.

1. Introduction

Image content representation plays a key role in computer vision and pattern recognition. The idea is to transform an image from its original representation into a new representation suitable for a desired task (e.g. classification). Modern techniques such as Bag-of-Words models of local features (e.g., SIFT (21), Weber (8), Gabor (19)) represent an image by a very high-dimensional feature vector. Although this representation leads to relatively high accuracy in visual recognition and search, it increases the computation time and, consequently, is improper for real-time applications. Therefore, developing new algorithms that generate a compact and informative representation of image content is highly needed. Perhaps, the most common way to tackle this problem is learning a subspace of the original feature space and using this representation for the recognition tasks.

For several years, the representation of images with visual attributes has been studied intensively by researchers in the fields of clustering, classification, object recognition, and face verification. Farhadi *et al.* (10) proposed a shift from naming images to describing images. Instead of a naming an animal a "dog" it can be described as a "spotty dog". This means a shift from traditional approaches, where each instance was labeled with one label, to a model with more semantic information. This information can be crucial to model and learn inter- and intra-class relations. Kumar *et al.* (17) have developed an attribute classifier which focuses on the similarity regions in an image, associates classed depending on them. In Silberer *et al.* (28) images were described by attributes of 8 different categories, such as *shape_size*, *color_patterns* and *structure*. Generally, these attributes are observable semantic cues, which can be learned from low-level features. For example, "smiling" and "dry" can be considered as attributes of a face or a scene, respectively. Recently, it has been proposed that *relative attributes* provide a richer source of semantic information in images (25),(14) than binary attributes. They depict the strength of attributes in an image and can be predicted by pre-learned rank functions. For each attribute, a single rank function, which is a rank-SVM, is learned from a set of training data (25). In this work, we use predicted relative attributes, as discriminant constraints to guide a NMF to generate a new subspace of images. More precisely, the relative attributes are embedded in a regularizer coupled with the NMF objective function. We call our proposed

**Corresponding author
e-mail: reza.babaee@tum.de (Mohammadreza Babaee)

method Attribute constrained NMF (ANMF).
Also we present a Dictionary Learning approach, utilizing relative attributes to find a discriminative sparse representation for images. In Dictionary Learning we consider a set of n input signals $\mathbf{Y} \in \mathbb{R}^{p \times n}$ and the goal is to find a dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_k] \in \mathbb{R}^{p \times k}$ and sparse representations $\mathbf{X} \in \mathbb{R}^{k \times n}$ such that $\mathbf{Y} \cong \mathbf{DX}$, where the term *over-complete* indicates $k > n$. Dictionaries can either be predefined as in the form of wavelets (23), or be learned from observations (1; 32; 7). Additionally, many approaches have been developed to impose discriminative capabilities onto the dictionary learning process. Those methods often use binary label information to acquire discriminative behavior. In this work, we present an approach that utilizes relative attributes instead of binary labels to enhance the discriminative property of the dictionary. Just as previous discriminative dictionary learning approaches use binary label information to enhance their discriminative capabilities, we incorporate relative attributes into the dictionary learning process as semantic information.
The rest of the paper is organized as follows. In section 2 related work in the field of dictionary learning and non-negative matrix factorization is presented. Section 3 gives an in-depth explanation of the problem solved by the dictionary learning approach. In sections 4 and 5 the details for the dictionary learning and non-matrix factorization algorithms are given. Afterwards in Section 6 the concluded experiments are described together with the obtained results. The report concludes with a discussion and summary in Section 8.

2. Related Work

The first approaches in the field of reconstructive dictionary learning are the K-SVD algorithm (1) and the Method of Optimal Direction (MOD) (9), where no semantic information is used in the learning process. An additional example for the usage of sparse representation is the Sparse Representation based Classification (SRC) (29) where the dictionary is built directly from the training data.

Another large field in dictionary learning is called Discriminative Dictionary Learning (DDL), where either the discriminative property of the signal reconstruction residual, or the discriminative property of the sparse representation itself is enhanced. Approaches with a focus on the reconstruction residual are the work of Ramirez *et al.* (26), which includes a structured incoherence term to find independent sub-directories for each class, and the work of Gao *et al.* (11), where sub-dictionaries for the different classes are learned as well as a shared dictionary over all classes.

Methods aiming at finding discriminative coding vectors learn the dictionary and a classifier simultaneously. In the work of Zhang *et al.* (32), the K-SVD algorithm is extended by a linear classifier. Jiang *et al.* (13) included an additional discriminative regularizer to come up with the so called Label Consistent K-SVD (LC-KSVD) algorithm. Both of these algorithms show good results for classification and face recognition tasks. The approach of Yang *et al.* (31) combines the two types of DDL by taking the discriminative capabilities of the reconstruction residual and the sparse representation into account. Therefore, class specific sub-dictionaries are learned while maintaining discriminative coding vectors by applying the Fisher discrimination criterion. In the recent work of Cai *et al.* (7) a method called Support Vector Guided Dictionary Learning (SVGDL) is presented, where the discrimination term consists of a weighted summation over squared distances between the pairs of coding vectors. The algorithm automatically assigns non-zero weights to critical vector pairs (the support vectors) leading to a generalized good performance in pattern recognition tasks.

Inspired by the part-based perception behavior of the human brain (i.e., combining the perceptions of an object to perceive it as a whole) non-negative Matrix Factorization (NMF) is a widely used matrix factorization method (6; 20; 18). A parts-based representation can be achieved by applying a non-negativity constraint to the matrix factors, only allowing additive combinations of original data.
In order to find a robust data representaion, further methods are needed besides the non-negativity constraint (20). One approach is to focus on preserving the intrinsic geometry of the data space by defining new objective functions. Cai (6) proposed the GNMF, constructing a nearest neighbor graph and encoding the geometrical information of the data space, and therefore considering the local invariance. Another approach is the CNMF, introduced by Liu and Wu (30) who constrain the NMF to only use the prior annotation of the data, enforcing similar encoding for points from the same class. In the work of Gu and Zhou (12) local linear embedding assumptions are used to propose the so called NPNMF. A new constraint was presented allowing each data point to be presented by its neighbors.

3. Background

For the general problem formulation we assume $\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, ...\mathbf{y_n}]$ to be the set of p-dimensional input signals, each belonging to one of C (hidden) classes, $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}]$ to be their corresponding k-dimensional sparse representation and $\mathbf{D} \in \mathbb{R}^{n \times k}$ to be the dictionary. As a consequence, the standard dictionary learning method is defined by

$$< \mathbf{D}, \mathbf{X} >= \arg\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{DX}\|_2^2 + \lambda_1 \|\mathbf{X}\|_1, \qquad (1)$$

with the regularization parameter λ_1. In order to take the relative attributes into account the objective function has to be extended with an additional term $\mathcal{L}(\mathbf{X})$.

$$< \mathbf{D}, \mathbf{X} >= \arg\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{DX}\|_2^2 + \lambda_1 \|\mathbf{X}\|_1 + \lambda_2 \mathcal{L}(\mathbf{X}) \qquad (2)$$

As additional information, the strength of M predefined attributes, the so called relative attributes (25), for the input signals are available.

3.1. Relative Attributes

The idea in learning relative attributes, assuming there are M attributes $\mathcal{A} = \{a_m\}$, is to learn M ranking functions $\mathbf{w_m}$

for $m = 1..M$. Therefore, the predicted relative attributes are computed by

$$r_m(\mathbf{x_i}) = \mathbf{w_m}^\intercal \mathbf{x_i}, \tag{3}$$

such that the maximum number of the following constraints is satisfied:

$$\forall (i,j) \in \mathcal{O}_m : \mathbf{w_m}^\intercal \mathbf{x_i} > \mathbf{w_m}^\intercal \mathbf{x_j}, \tag{4}$$

$$\forall (i,j) \in \mathcal{S}_m : \mathbf{w_m}^\intercal \mathbf{x_i} \approx \mathbf{w_m}^\intercal \mathbf{x_j} \tag{5}$$

whereby $\mathcal{O}_m = \{(i,j)\}$ is a set of ordered signal pairs with signal i having a stronger presence of attribute a_m than signal j and $\mathcal{S}_m = \{(i,j)\}$ being a set of un-ordered pairs where signal i and j have about the same presence of attribute a_m. It is possible to approximate this objective with the introduction of non-negative slack variables, similar to an SVM classifier:

$$\min \left(\frac{1}{2} \|\mathbf{w_m}^\intercal\| + c \left(\sum \xi_{ij} + \sum \gamma_{ij} \right) \right) \tag{6}$$

$$\text{s.t. } \mathbf{w_m}^\intercal (\mathbf{x_i} - \mathbf{x_j}) \geq 1 - \xi_{ij};\ \forall (i,j) \in \mathcal{O}_m \tag{7}$$

$$|\mathbf{w_m}^\intercal (\mathbf{x_i} - \mathbf{x_j})| \leq \gamma_{ij};\ \forall (i,j) \in \mathcal{S}_m \tag{8}$$

The work of Parikh *et al.* (25) provides us with a convenient *RankSVM* function that returns the ranking vector $\mathbf{w_m}$ for a set of input images and their relative ordering. This information can further be used in the objective function in Eq. (2).

4. Attributes Constrained Dictionary Learning

The *RankSVM* function maps the original input signal ($\mathbf{y_i}$) to a point (q_i) in a so-called relative attribute space. Additionally, we assume that there exists a linear transformation (i.e., $\mathbf{A}$) that maps the sparse signal ($\mathbf{x_i}$) to the point q_i (see Figure 1 and Eq. (9)). First, we define the matrix $\mathbf{Q} \in \mathbb{R}^{n \times M}$ with the elements $q_{im} = r_m(\mathbf{y_i})$ that contains the strength of the (relative) attributes of all signals in $\mathbf{Y}$. In order to find the transformation of $\mathbf{Y}$ into $\mathbf{Q}$, we apply the *RankSVM* function known from (25) onto the original input signal and obtain the weighting matrix $W = [\mathbf{w_1}^\intercal; \mathbf{w_2}^\intercal; ...; \mathbf{w_M}^\intercal]$.

$$\arg\min_{\mathbf{A}} \|\mathbf{Q} - \mathbf{AX}\|_2^2 = \arg\min_{\mathbf{A}} \|\mathbf{WY} - \mathbf{AX}\|_2^2. \tag{9}$$

The objective is finding a matrix $\mathbf{A}$, which transform the sparse representation of the signals into their corresponding relative attribute representations $\mathbf{Q}$ with a minimum distance between $\mathbf{w_m}^\intercal \mathbf{y_i}$ and $\mathbf{a_m}^\intercal \mathbf{x_i}$. By using Eq. (9) in Eq. (2) as a loss term we get the formulation

$$< \mathbf{D}, \mathbf{X} >= \arg\min_{\mathbf{D},\mathbf{X},\mathbf{A}} \|\mathbf{Y} - \mathbf{DX}\|_2^2 + \lambda_1 \|\mathbf{X}\|_1 + \lambda_2 \|\mathbf{WY} - \mathbf{AX}\|_2^2. \tag{10}$$

From the first part of the equation we can see that $\mathbf{Y} \cong \mathbf{DX}$. If $\mathbf{Y}$ in the loss term for the relative attributes is approximated by $\mathbf{DX}$ then the equation becomes

$$< \mathbf{D}, \mathbf{X} >= \arg\min_{\mathbf{D},\mathbf{X},\mathbf{A}} \|\mathbf{Y} - \mathbf{DX}\|_2^2 + \lambda_1 \|\mathbf{X}\|_1 + \lambda_2 \|\mathbf{WDX} - \mathbf{AX}\|_2^2. \tag{11}$$

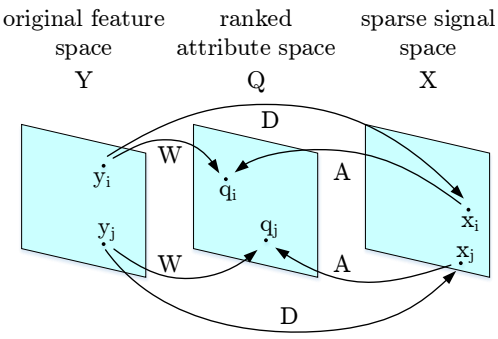


Fig. 1. Illustration of signal transformations. The goal is to transform $\mathbf{x_i}$ and $\mathbf{x_j}$ as close as possible to q_i and q_j.

The third term of Eq. (11) is minimized if $\mathbf{A} = \mathbf{WD}$. This information can be used to eliminate $\mathbf{A}$ from Eq. (10) to arrive at the final objective function

$$< \mathbf{D}, \mathbf{X} >= \arg\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{DX}\|_2^2 + \lambda_1 \|\mathbf{X}\|_1 + \lambda_2 \|\mathbf{W}(\mathbf{Y} - \mathbf{DX})\|_2^2. \tag{12}$$

Additionally, we can replace the term $\|\mathbf{X}\|_1$ with $\|\mathbf{X}\|_2^2$ since the goal is to learn a discriminative dictionary and not to obtain sparse signals (as in (7)). However, once the dictionary is learned, the sparse representation is obtained by the orthogonal matching pursuit (27). Finally, we end up with the following optimization problem:

$$< \mathbf{D}, \mathbf{X} >= \arg\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{DX}\|_2^2 + \lambda_1 \|\mathbf{X}\|_2^2 + \lambda_2 \|\mathbf{W}(\mathbf{Y} - \mathbf{DX})\|_2^2. \tag{13}$$

Since this equation is not a jointly convex optimization problem $\mathbf{X}$ and $\mathbf{D}$ are optimized sequentially. The update rules for $\mathbf{D}$ and $\mathbf{X}$ are found by deriving the objective function and setting the derivatives to zero.

$$O = \|\mathbf{Y} - \mathbf{DX}\|_2^2 + \lambda_1 \|\mathbf{X}\|_2^2 + \lambda_2 \|\mathbf{W}(\mathbf{Y} - \mathbf{DX})\|_2^2 \tag{14}$$

$$\begin{aligned} \frac{\partial O}{\partial \mathbf{D}} &= -2(\mathbf{Y} - \mathbf{DX})\mathbf{X}^\intercal - 2\lambda_2 \mathbf{W}^\intercal(\mathbf{WY} - \mathbf{WDX})\mathbf{X}^\intercal = 0 \\ &= (\mathbf{Y} - \mathbf{DX}) + \lambda_2 \mathbf{W}^\intercal \mathbf{W}(\mathbf{Y} - \mathbf{DX}) = 0 \\ &= (\mathbf{I} + \lambda_2 \mathbf{W}^\intercal \mathbf{W})(\mathbf{Y} - \mathbf{DX}) = 0 \\ \Rightarrow \mathbf{D} &= \mathbf{Y}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \end{aligned} \tag{15}$$

$$\begin{aligned} \frac{\partial O}{\partial \mathbf{X}} &= -2\mathbf{D}^\intercal(\mathbf{Y} - \mathbf{DX}) + 2\lambda_1 \mathbf{X} - 2\lambda_2 \mathbf{D}^\intercal \mathbf{W}^\intercal(\mathbf{WY} - \mathbf{WDX}) = 0 \\ &= (\mathbf{D}^\intercal \mathbf{D} + \lambda_1 \mathbf{I} + \lambda_2 \mathbf{D}^\intercal \mathbf{W}^\intercal \mathbf{WD})\mathbf{X} - \mathbf{D}^\intercal \mathbf{Y} - \lambda_2 \mathbf{D}^\intercal \mathbf{W}^\intercal \mathbf{Y} = 0. \end{aligned} \tag{16}$$

Therefore, we have

$$\mathbf{X} = (\mathbf{D}^\intercal\mathbf{D} + \lambda_1\mathbf{I} + \lambda_2\mathbf{D}^\intercal\mathbf{W}^\intercal\mathbf{W}\mathbf{D})^{-1}(\mathbf{D}^\intercal\mathbf{Y} + \lambda_2\mathbf{D}^\intercal\mathbf{W}^\intercal\mathbf{Y}). \quad (17)$$

The complete algorithm works as follows. Initially, the *RankSVM* (25) function is used to learn the ranking matrix $\mathbf{W}$ from the original input data $\mathbf{Y}$ and their relative ordering (i.e., sets O_m, $\mathcal{S}_m$). The initial dictionary $\mathbf{D}$ and the sparse representation of the data is obtained by first building a dictionary from random chosen input signals and then applying the KSVD-algorithm (1). Afterward, the dictionary and the sparse representation are optimized alternately until convergence. In order to avoid scaling issues the dictionary is L_2 normalized column-wise. The structure of the algorithm can be seen in Algorithm 1.

Algorithm 1 Relative Attribute Guided Dictionary Learning

Require: Original signal $\mathbf{Y}$, sets of ordered (O_m) and un-orderd images ($\mathcal{S}_m$)
Ensure: Dictionary $\mathbf{D}$

```
W ← RankSVM(Y, O_m, S_m)
D_init ← rndperm(Y)
D, X ← KSVD(D_init, Y)
for i = 0 to numIter do
    D ← Y(X^T X)^{-1} X^T
    D ← normcol(D)
    X ← (D^T D + λ1 I + λ2 D^T W^T W D)^{-1} (D^T Y − λ2 D^T W^T Y)
end for
```

5. Attributes Guided Non-negative Matrix Factorization

We assume that $\mathbf{X} \in \mathbb{R}^{D\times N}$ denotes N data points (e.g., images) represented by D dimensional low-level feature vectors. The NMF decomposes the non-negative matrix $\mathbf{X}$ into two non-negative matrices $\mathbf{U} \in \mathbb{R}^{D\times K}$ and $\mathbf{V} \in \mathbb{R}^{N\times K}$ such that the multiplication of $\mathbf{U}$ and $\mathbf{V}$ approximates the original matrix $\mathbf{X}$ (3; 4). Here, $\mathbf{U}$ represents the bases and $\mathbf{V}$ contains the coefficients, which are considered as new representation of the original data. The NMF objective function is:

$$\begin{aligned} F &= \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^\mathrm{T}\right\|_F^2 \\ \text{s.t.}\quad & \mathbf{U} = [u_{ik}] \geq 0 \\ & \mathbf{V} = [v_{jk}] \geq 0. \end{aligned} \quad (18)$$

Additionally, we assume that M semantic attributes have been predefined for the data and the relative attributes of each image are available. Precisely, the matrix of relative attributes, $\mathbf{Q} \in \mathbb{R}^{M\times N}$, has been learned by some ranking function (e,.g, rankSVM). Intuitively, those images which own similar relative attributes have similar semantic contents and therefore belong to the same semantic class. This concept can be formulated as a regularizer to be added to the main NMF objective function. This information can be formulated in a regularization term as

$$R = \alpha\left\|\mathbf{Q} - \mathbf{A}\mathbf{V}^\mathrm{T}\right\|^2, \quad (19)$$

where $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_N]^\mathrm{T} \in \mathbb{R}^{N\times K}$ and the matrix $\mathbf{A} \in \mathbb{R}^{M\times K}$. The matrix $\mathbf{A}$ linearly transforms and scales the vectors in the new representation in order to obtain the best fit for the matrix $\mathbf{Q}$. The matrix $\mathbf{A}$ is allowed to take negative values and is computed as part of the NMF minimization. We arrive at the following minimization problem:

$$\begin{aligned} \min F &= \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^\mathrm{T}\right\|^2 + \alpha\left\|\mathbf{Q} - \mathbf{A}\mathbf{V}^\mathrm{T}\right\|^2 \\ \text{s.t.}\quad & \mathbf{U} = [u_{ik}] \geq 0 \\ & \mathbf{V} = [v_{jk}] \geq 0. \end{aligned} \quad (20)$$

5.1. Update rules

For the derivation of the update rules we expand the objective to

$$\begin{aligned} O =& \mathrm{Tr}\left(\mathbf{X}\mathbf{X}^\mathrm{T}\right) - 2\mathrm{Tr}\left(\mathbf{X}\mathbf{V}\mathbf{U}^\mathrm{T}\right) + \mathrm{Tr}\left(\mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{U}^T\right) \\ &+ \alpha\mathrm{Tr}\left(\mathbf{Q}\mathbf{Q}^\mathrm{T}\right) - \alpha 2\mathrm{Tr}\left(\mathbf{Q}\mathbf{V}\mathbf{A}^\mathrm{T}\right) + \alpha\mathrm{Tr}\left(\mathbf{A}\mathbf{V}^T\mathbf{V}\mathbf{A}^T\right) \end{aligned} \quad (21)$$

and introduce Lagrange multipliers $\mathbf{\Phi} = [\phi_{ik}]$, $\mathbf{\Psi} = [\psi_{jk}]$ for the constraints $[u_{ik}] \geq 0$, $[v_{jk}] \geq 0$ respectively. Adding the Lagrange multipliers and ignoring the constant terms leads to the Lagrangian:

$$\begin{aligned} \mathcal{L} =& -2\mathrm{Tr}\left(\mathbf{X}\mathbf{V}\mathbf{U}^\mathrm{T}\right) + \mathrm{Tr}\left(\mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{U}^T\right) + \mathrm{Tr}\left(\mathbf{\Phi}\mathbf{U}\right) \\ &+ \mathrm{Tr}\left(\mathbf{\Psi}\mathbf{V}\right) - \alpha 2\mathrm{Tr}\left(\mathbf{Q}\mathbf{V}\mathbf{A}^\mathrm{T}\right) + \alpha\mathrm{Tr}\left(\mathbf{A}\mathbf{V}^T\mathbf{V}\mathbf{A}^T\right). \end{aligned} \quad (22)$$

The partial derivatives of $\mathcal{L}$ with respect to $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{A}$ are:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{U}} = -2\mathbf{X}\mathbf{V} + 2\mathbf{U}\mathbf{V}^\mathrm{T}\mathbf{V} + \mathbf{\Phi} \quad (23)$$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{V}} = -2\mathbf{X}^T\mathbf{U} + 2\mathbf{V}\mathbf{U}^T\mathbf{U} - \alpha 2\mathbf{Q}^T\mathbf{A} + \alpha 2\mathbf{V}\mathbf{A}^T\mathbf{A} + \mathbf{\Psi} \quad (24)$$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{A}} = -2\mathbf{Q}\mathbf{V} + 2\mathbf{A}\mathbf{V}^\mathrm{T}\mathbf{V} \quad (25)$$

For the derivation of the update rules for $\mathbf{U}$ and $\mathbf{V}$ we apply the KKT-conditions $\phi_{ik}u_{ik} = 0$, $\psi_{jk}v_{jk} = 0$ (5). For $\mathbf{A}$ the update rules can be derived directly by setting its derivative of the Lagrangian to 0. Thus, we arrive at the following equations:

$$u_{ik} \leftarrow u_{ik}\frac{[\mathbf{X}\mathbf{V}]_{ik}}{[\mathbf{U}\mathbf{V}^\mathrm{T}\mathbf{V}]_{ik}} \quad (26)$$

$$v_{jk} \leftarrow v_{jk}\frac{[\mathbf{X}^\mathrm{T}\mathbf{U} + \alpha(\mathbf{V}\mathbf{A}^\mathrm{T}\mathbf{A})^- + \alpha(\mathbf{Q}^\mathrm{T}\mathbf{A})^+]_{jk}}{[\mathbf{V}\mathbf{U}^\mathrm{T}\mathbf{U} + \alpha(\mathbf{V}\mathbf{A}^\mathrm{T}\mathbf{A})^+ + \alpha(\mathbf{Q}^\mathrm{T}\mathbf{A})^-]_{jk}} \quad (27)$$

$$\mathbf{A} \leftarrow \mathbf{Q}\mathbf{V}(\mathbf{V}^\mathrm{T}\mathbf{V})^{-1} \quad (28)$$

where for a matrix $\mathbf{M}$ we define $\mathbf{M}^+$, $\mathbf{M}^-$ as $\mathbf{M}^+ = (|\mathbf{M}| + \mathbf{M})/2$ and $\mathbf{M}^- = (|\mathbf{M}| - \mathbf{M})/2$. The newly introduced terms depend only on the variables $\mathbf{V}$ and $\mathbf{A}$.

6. Experiment 1

6.1. Datasets

In order to assess the quality of the learned dictionary, we purpose a clustering task for three public available datasets, namely Public Figure Face (PubFig) (16), Outdoor Scene

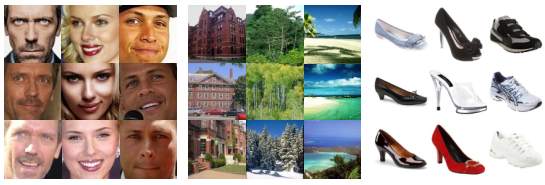


Fig. 2. Example images from the PubFig, OSR and Shoes datasets.

Recognition (OSR) (24) and Shoes (15). Some sample images of each dataset are presented in Figure 2. The conducted tests lead to the used parameters $\lambda_1 = 0.01$ and $\lambda_2 = 1$ for all experiments and datasets.

a) The subset of the **PubFig** dataset is the same as in (25) containing 772 images from 8 different identities defined by the 512 dimensional GIST (24) features and is split into 241 training images and 531 test images.
b) The **OSR** set consist of 2688 images from 8 categories described again by the 512 dimensional GIST (24) features split into 240 training and 2488 testing images.
c) In the **Shoes** dataset there are 14658 images from 10 different types. Out of this set 240 images were used for training and 1579 for testing. The images are described by 960 dimensional GIST (24) features.

6.2. *Finding the optimal λ*

Additionally tests were conducted to find the optimal values for λ_1 and λ_2. Therefore, different fixed values were chosen λ_1 while iterating over candidates for λ_2. The tests lead us to $\lambda_1 = 0.01$ and $\lambda_2 = 1$ for the Pubfig dataset, $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$ for the OSR dataset and $\lambda_2 = 0.001$ and $\lambda_2 = 0.1$ for the Shoes dataset.

6.3. *Evaluation Metrics*

After the dictionary is learned, the goal is to acquire a sparse representation of the input data and to quantify their separability. The sparse representation is obtained by solving the error-constrained sparse coding problem, given by Eq. (29), with the help of the OMP-Box Matlab toolbox (27), where the reconstruction error from the training phase is chosen as ε.

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \|\mathbf{X}\|_0 \quad \text{s.t.} \quad \mathbf{X} = \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_2^2 \leq \varepsilon, \qquad (29)$$

In order to quantify the clustering capabilities of the sparse representation, the k-means (22) algorithm is applied to $\hat{\mathbf{X}}$ and the accuracy (AC) and the normalized Mutual Information (nMI) metrics (2) are computed. The accuracy describes the percentage of correctly clustered data points for the best match of the clusters found by the k-means algorithm to the original label information provided by the dataset.

6.4. *Results*

As a benchmark for the results, different unsupervised and unsupervised (discriminative) dictionary learning techniques are used, namely (1) KSVD (1), (2) SRC (29) as unsupervised techniques and (3) LC-KSVD (32), (4) FDDL (31), (5) SVGDL (7) as supervised techniques. Additionally, the original features (O. Feat.) are clustered as well by the k-means algorithm to evaluate the additional value of using relative attributes as semantic information. The results were compared by their performance for full label information, varying dictionary sizes and a varying amount of training data. Table 1 shows the accuracy and normalized mutual information for all algorithms tested on the three datasets when using all training data, their label information and fixed a dictionary size of 130. Additionally, the average result is computed. One can see that although the proposed algorithm uses a different kind of semantic information, it reaches a comparable (for Shoes) up to better performance (for PubFig and OSR) in regards to other algorithms that use label information. Additionally, it clearly outperforms the unsupervised algorithms indicating the benefit of using relative attributes.

In Table 2 the runtime of the training phase of the algorithms is analyzed, where the numbers confirm that the proposed algorithm runs much faster than all the contestants. The experiments were conducted on an Asus N56VZ-S4044V Notebook with an Intel Core i7-3610QM processor and a clock speed of 2.3 GHz. Figure 3 shows the behavior of the algorithms for an increasing the dictionary size with all training data available. The dictionary sizes used were [16, 40, 80, 120, 160, 240] for the PubFig and OSR dataset and [20, 50, 100, 140] for the Shoes dataset, which corresponds to [2, 5, 10, 15, 20, 30] and [2, 5, 10, 14] atoms per class. The number of atoms per class are constrained by the partition of the data into training and testing (for the Shoes dataset one class only includes 14 training samples). One should notice that the FDDL algorithm cannot use all training data, since the dictionary size restricts the size of the training samples. Therefore, only in the last test case the algorithm uses the complete training information. The results show that for the proposed algorithm the accuracy increases with the dictionary size, up to values exceeding the compared algorithms. However, for the OSR and Shoes dataset and an increasing dictionary size the SVGDL and FDDL produce comparable results. Figure 4 illustrates the results when the amount of used training data is varied. In addition, the dictionary sizes were matched to the size of training data. Again, the proposed algorithm can exceed the results of the compared approaches up to a number of training samples in the OSR and Shoes dataset were the SVGDL and FDDL algorithms produce comparable results. The number of training samples per class were [2, 5, 10, 15, 20, 30] for the PubFig and OSR dataset and [2, 5, 10, 14] for the Shoes dataset.

7. Experiment 2

We apply the proposed method (ANMF), PCA, and NMF on the original representations of three datasets to generate (learn)

Clustering Results

Accuracy							
Method	O.Feat.	SRC	KSVD	LC-KSVD	FDDL	SVGDL	proposed
PubFig	0.324 ± 0.000	0.226 ± 0.000	0.310 ± 0.001	0.306 ± 0.001	0.584 ± 0.003	0.595 ± 0.002	**0.789 ± 0.001**
OSR	0.563 ± 0.000	0.239 ± 0.000	0.466 ± 0.001	0.500 ± 0.001	0.680 ± 0.000	0.662 ± 0.001	**0.731 ± 0.000**
Shoes	0.356 ± 0.000	0.198 ± 0.000	0.345 ± 0.001	0.302 ± 0.001	**0.498 ± 0.000**	0.481 ± 0.001	0.463 ± 0.000
Avg.	0.414	0.221	0.374	0.369	0.576	0.579	**0.661**
normalized Mutual Information							
PubFig	0.170 ± 0.000	0.062 ± 0.000	0.159 ± 0.001	0.161 ± 0.001	0.417 ± 0.001	0.448 ± 0.001	**0.600 ± 0.000**
OSR	0.433 ± 0.000	0.071 ± 0.000	0.334 ± 0.000	0.342 ± 0.000	0.498 ± 0.000	0.521 ± 0.000	**0.564 ± 0.000**
Shoes	0.322 ± 0.000	0.061 ± 0.000	0.261 ± 0.001	0.220 ± 0.001	**0.407 ± 0.001**	**0.407 ± 0.000**	0.394 ± 0.000
Avg.	0.308	0.065	0.251	0.241	0.441	0.459	**0.519**

Table 1. Accuracy and normalized Mutual Information for different dictionary learning algorithms applied to the datasets

Runtime (in seconds)

Method	O.Feat.	SRC	KSVD	LC-KSVD	FDDL	SVGDL	proposed
PubFig	-	-	3.910	5.652	33.170	8.130	**1.443**
OSR	-	-	3.803	5.467	32.492	7.628	**1.422**
Shoes	-	-	5.116	7.058	30.316	9.612	**2.381**
Avg.	-	-	4.276	6.059	31.993	8.457	**1.749**

Table 2. Runtime (in seconds) for different dictionary learning algorithms applied to the datasets.

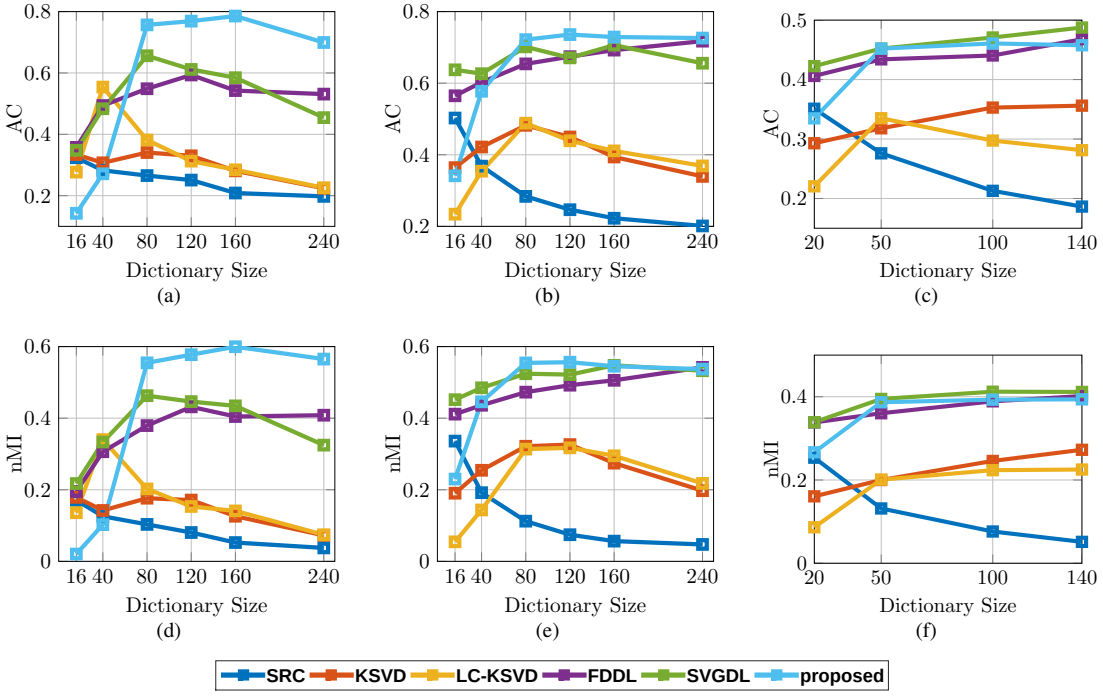


Fig. 3. Clustering results for all three datasets for increasing dictionary sizes. The first and second row represent the Accuracy (AC) and normalized MI (nMI), respectively. The first, second, and third columns are the results of the PubFig, OSR, and Shoes datasets, respectively.

different subspaces of the data. Then we apply k-means clustering on the new subspaces and also on the original data with k equal to the dimension of subspaces. We perform the experiments with k different classes, sampled from each dataset. In order to obtain representative results, we repeat the experiments 10 times for each k. The k-means runs 20 times per experiment and the best result is selected. For the subspace learning techniques (i.e., PCA, NMF, ANMF), we always set the new dimension equal to the number of classes. In ANMF, the regularization parameter is chosen by running a cross-validation on each dataset. For the OSR dataset, the PubFig dataset, and the Shoes dataset, this parameter was 10, 100, and 100, respectively.

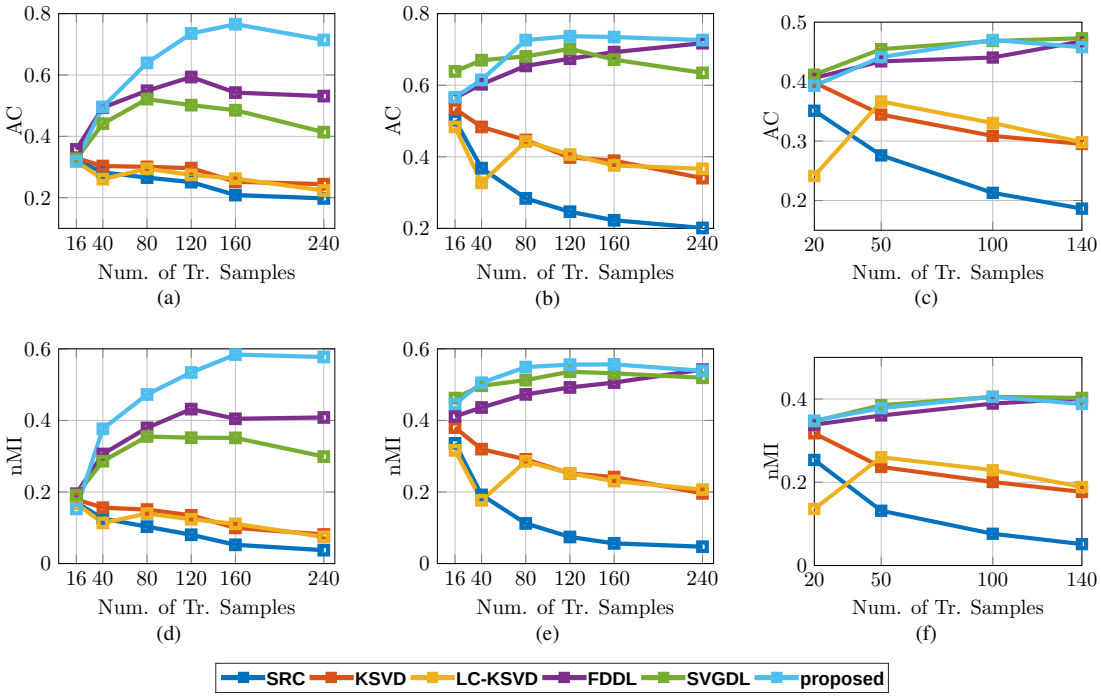


Fig. 4. Clustering results for all three datasets for increasing training data. The first and second row represent the Accuracy (AC) and normalized MI (nMI), respectively. The first, second, and third columns are the results of PubFig, OSR, and Shoes datasets, respectively.

7.1. Results

The clustering results on the learned subspaces are depicted in Fig. 5. Figs. 5(a) and 5(d) show the accuracy and normalized Mutual Information (nMI) of the clustering results for the PubFig dataset. The OSR results are depicted in Figs. 5(b) and 5(e) and the results for the Shoes dataset are represented in 5(c) and 5(f). It can be seen that the proposed method outperforms the other techniques significantly on both datasets. For the PubFig dataset we even achieve 75% − 85% accuracy. The algorithm converges quickly after 20 iterations and therefore can be considered computationally efficient. The experimental results confirm that the proposed method learns the bases with different semantic attributes successfully.

8. Conclusion

We have presented novel discriminative dictionary learning and non-negative matrix factorization algorithms that use relative attributes as semantic information instead of binary labels. In dictionary learning algorithm, we use the learned ranking functions in the learning process. The ranking functions transform the original features into a relative attribute space and therefore, we aim to transform the sparse signal linearly into this attribute space. This can be achieved by adding an additional loss term to the objective function of a standard dictionary learning problem. In matrix factorization, the proposed algorithm uses predicted relative attributes embedded in a regularizer coupled with the main objective function of NMF. Our experiments on three image datasets confirm that the obtained sparse and/or subspace representations are very discriminative and lead in better clustering results in comparison to other methods.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, 2006.

[2] M. Babaee, R. Bahmanyar, G. Rigoll, and M. Datcu. Farness preserving non-negative matrix factorization. In *IEEE International Conference on Image Processing (ICIP)*, pages 3023–3027, 2014.

[3] M. Babaee, S. Tsoukalas, M. Babaee, G. Rigoll, and M. Datcu. Discriminative nonnegative matrix factorization for dimensionality reduction. *Neurocomputing*, 173:212–223, 2016.

[4] M. Babaee, S. Tsoukalas, G. Rigoll, and M. Datcu. Immersive visualization of visual data using nonnegative matrix factorization. *Neurocomputing*, 173:245–255, 2016.

[5] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2009.

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

[7] Sijia Cai, Wangmeng Zuo, Lei Zhang, Xiangchu Feng, and Ping Wang. Support vector guided dictionary learning. In *Computer Vision–ECCV 2014*, pages 624–639. Springer, 2014.

[8] J. Chen, S. Shan, G. Zhao, X. Chen, W. Gao, and M. Pietikainen. A robust descriptor based on weber's law. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1–7, 2008.

[9] K. Engan, S. O. Aase, and J. Husoy. Frame based signal compression using method of optimal directions (mod). In *Circuits and Systems, 1999. ISCAS'99. Proceedings of the 1999 IEEE International Symposium on*, volume 4, pages 1–4. IEEE, 1999.

[10] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1778–1785, 2009.

[11] S. Gao, I.-H. Tsang, and Y. Ma. Learning category-specific dictionary and shared dictionary for fine-grained image categorization. *Image Processing, IEEE Transactions on*, 23(2):623–634, 2014.

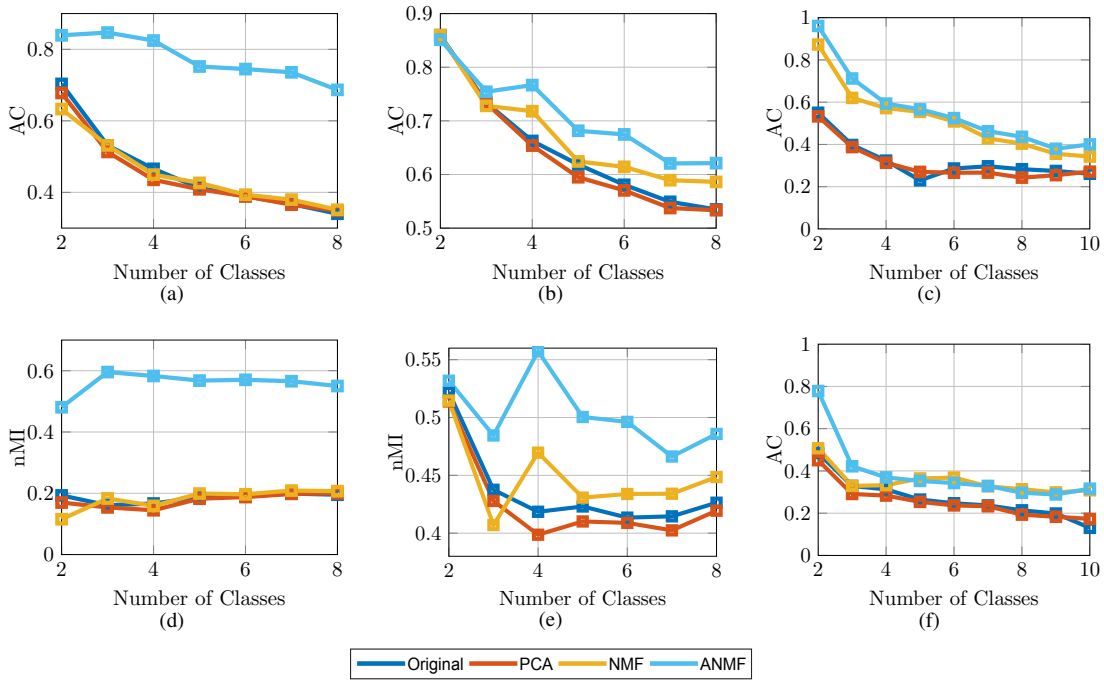


Fig. 5. Clustering results of the new representation computed by PCA, NMF, ANMF, and original data evaluated by accuracy (AC) and normalized mutual information (nMI). (a) and (d) show the AC and nMI for the PubFig dataset, respectively. (b) and (e) show the AC and nMI for the OSR dataset, respectively. (c) and (f) show the AC and nMI of the Shoes dataset, respectively.

[12] Quanquan Gu and Jie Zhou. Neighborhood preserving nonnegative matrix factorization. In *BMVC*, pages 1–10, 2009.

[13] Zhuolin Jiang, Zhe Lin, and Larry S Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1697–1704. IEEE, 2011.

[14] A. Kovashka and K. Grauman. Attribute pivots for guiding relevance feedback in image search. In *Computer Vision (ICCV), IEEE International Conference on*, pages 297–304. IEEE, 2013.

[15] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image Search with Relative Attribute Feedback. In *Computer Vision and Pattern Recognition (CVPR), IEEE International Conference on*, June 2012.

[16] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2009.

[17] N. Kumar, A. C Berg, P.N. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 12th IEEE International Conference on*, pages 365–372, 2009.

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

[19] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *Image processing, IEEE Transactions on*, 11(4):467–476, 2002.

[20] Haifeng Liu, Zheng Yang, Zhaohui Wu, and Xuelong Li. A-optimal non-negative projection for image representation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1592–1599. IEEE, 2012.

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[22] J. MacQueen and et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297. Oakland, CA, USA., 1967.

[23] S. Mallat. *A wavelet tour of signal processing*. Academic press, 1999.

[24] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[25] D. Parikh and K. Grauman. Relative attributes. In *Computer Vision (ICCV), IEEE International Conference on*, pages 503–510. IEEE, 2011.

[26] I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 3501–3508. IEEE, 2010.

[27] R. Rubinstein, Mi. Zibulevsky, and M. Elad. Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit. *CS Technion*, 40(8):1–15, 2008.

[28] Carina Silberer, Vittorio Ferrari, and Mirella Lapata. Models of semantic representation with visual attributes. In *ACL (1)*, pages 572–582, 2013.

[29] J. Wright, A. Y Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

[30] Haifeng Liu Zhaohui Wu. Non-negative matrix factorization with constraints. 2010.

[31] M. Yang, D. Zhang, and X. Feng. Fisher discrimination dictionary learning for sparse representation. In *Computer Vision (ICCV), IEEE International Conference on*, pages 543–550. IEEE, 2011.

[32] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 2691–2698. IEEE, 2010.