

Content-based image retrieval using student's t -mixture model and constrained multiview nonnegative matrix factorization

Hongqing Zhu¹ · Qunyi Xie¹

Received: 12 October 2016 / Revised: 18 May 2017 / Accepted: 10 July 2017 /
Published online: 27 July 2017
© Springer Science+Business Media, LLC 2017

Abstract The expensive and time-consuming effort required for archiving images is the main motive for developing an effective retrieval system. This paper addresses a competitive scheme for Content-Based Image Retrieval (CBIR) based on a constrained multiview Nonnegative Matrix Factorization (NMF) that has the ability to generate a sparse representation. The scheme blends multiple visual features, which can together reflect the content of images in terms of similarity metrics and the Frobenius norm. Then, the proposed method constructs a similarity-preserving matrix factorization via an improved NMF, where the structural constraint, $L_{1/2}$ -sparse constraint and fairness-preserving constraint are integrated into the objective function of conventional NMF. In this way, the structure and content of high-dimensional feature data source can be preserved in low-dimensional space. Another critical part of the proposed system is to establish Student's t -Mixture Model (SMM) based on a Markov Random Field (MRF), which can best manipulate the clustering of sparse representations according to the statistical properties of the image features. With this method, the task of image retrieval of the whole dataset is reduced to a nearest-neighbour search in a specific category containing the query image. Convergence of the proposed update rule, investigated in this study, is also verified by numerical simulations. Lastly, we conduct experiments on public datasets to compare the performance of the proposed algorithm with existing works in terms of Precision and Recall Rates. The encouraging results indicate the effectiveness of the proposed technique.

Keywords Image retrieval · Student's t -mixture model · nonnegative matrix factorization · sparse constraint · Markov random field · multiview

✉ Hongqing Zhu
hqzhu@ecust.edu.cn

¹ School of Information Science & Engineering, East China University of Science and Technology, No. 130 Mei Long Road, Shanghai 200237, China

1 Introduction

With the fast growth of internet images, Content-Based Image Retrieval (CBIR) techniques have become one of the most important themes of research in computer vision. A CBIR system looks for a subset of images that is visually similar to a given query image and displays the results retrieved from image repositories. In most CBIR systems, feature descriptors, which may cater to the purposes required by the user, play a vital role in reflecting the image content. Generally, the lower-level image features that have been widely adopted by CBIR systems consist mainly of colour, spatial position, texture, scene, shape of the object etc. Obviously, the use of a single feature would inevitably cause a poor retrieval response because it is hard to comprehensively describe an image with an individual feature. To overcome this issue, multiview feature fusion schemes are treated as an alternative to a single type of visual feature, and they have attracted increasing research attention [24, 37, 41]. The CBIR algorithms heavily rely on image descriptors and a good similarity measure between images. The data of similarity matrices are often high-dimensional and non-sparse. The Nonnegative Matrix Factorization (NMF) method and its variants have been demonstrated to be particularly successful in addressing dimensionality-reduction problems by offering a sparse description of the original-dimensional data [15]. In essence, NMF seeks two nonnegative matrices with lower ranks, which are, respectively, called as the basis matrix and coefficient matrix, so that their product provides a better estimate of the given matrix. The coefficient matrix with the blended multiview feature is very low-dimensional. At this time, any clustering algorithm can be implemented to this matrix to associate each image with a given cluster. The finite mixture model makes the model-based clustering strategy attractive for CBIR systems [44], to analyse the properties of the coefficient matrix of NMF in a probabilistic manner and to cluster features according to the parameters of the mixture model.

In this paper, we introduce a novel scheme, which incorporates constrained multiview NMF and Student's t -Mixture Model (SMM) based on a Markov random field (MRF), for image retrieval. This method is termed SCMN, which has the following characteristics.

First, the proposed framework represents an image in a multidimensional feature space. The extracted underlying features, including texture, colour, spatial information, rotation-invariance, and scene, are merged by a Gaussian-like heat kernel to obtain a similarity-preserving matrix. We realized that little attention has been paid to rotation-invariance in image retrieval applications. However, real imaging systems are generally imperfect, and the obtained image usually presents a degraded version of the original one. Therefore, we think that rotation-invariance, i.e., remaining invariant under rotation, is a useful feature that deserves attention.

Second, we develop a constrained multiview NMF scheme through incorporating multiple constraints into the original NMF for the description of image features in a sparse space. Specifically, we impose a structural constraint into the objective function to obtain properties such as preservation of local structure and apply it to guide the matrix factorization. In addition, the proposed model enforces the $L_{1/2}$ -sparse constraint on the coefficient matrix and attempts to utilize the sparsity property of feature space as much as possible. The $L_{1/2}$ constraint has been proven better than others, since it can exploit the inherent sparseness of the data [38]. The farness-preserving constraint is utilized by the proposed objective function to preserve the data distribution in objective space. Additionally, the paper discusses the convergence of the update rule in theory, to ensure that our objective function converges with local minima.

Third, the proposed SCMN utilizes a multivariable Student's t -mixture model based on Markov Random Field (SMM-MRF) to approximate different shapes of sparse features. The SMM is highly acclaimed for its accuracy and effectiveness in image clustering [29]. In our scheme, images with sparse features belonging to the same Student's t component are similar, and these images can be grouped in the same category. Another key idea behind the proposed scheme is to adequately consider the spatial information of multiview features because the MRF is incorporated into SMM. The model parameters are estimated by adopting Expectation Maximization (EM). With the label information of the query image using the mentioned SMM-MRF, we can obtain a subset of images that are visually similar to a given query in terms of the multivariable clustering results.

The rest of this paper is organized as follows. Section 2 presents the related works and background. Section 3 introduces the proposed approach, its update rules and some computational aspects. Next, the theoretical proof of convergence and the computational complexity are discussed in section 4. In section 5, a comparative study of several standard datasets is performed. The last section reports the concluding remarks.

2 Related works and background

2.1 Related works

Most image retrieval algorithms rely on low-level image features to compare images based on visual similarity. These low-level features, represented in visual content, are easily implemented and obtained. There are numerous low-level feature descriptors have been reported and adopted in early works [33]. Deselaers et al. [11] conducted the experiments on CBIR using a large number of different low-level image features. Different features have different effectiveness to describe the same category of images, therefore, feature fusion is needed to select different features to have a better combination of basic features. Recently, there have been increasing efforts in developing multiview feature fusion schemes. For example, Liu et al. [24] studied Multiview Alignment Hashing (MAH) for indexing images. To preserve the geometric structure of a motion image, Wang et al. developed a multiview Laplacian graph via a linear regression model, along with multiview spectral embedding [39] etc. In [3], An et al. developed discriminative image features with attribute information encoded to achieve more accurate image retrieval. The basic requirement of CBIR is to decrease the redundancy of multiview features and to explore them in a relative low-dimensional feature space. The popular strategy to address this problem is to utilize dimensionality-reduction techniques, typical methods, including Singular Value Decomposition (SVD) [20], multidimensional scaling [8], Principal Component Analysis (PCA) [43], Independent Component Analysis (ICA) [18], and NFM. In NMF model, both basic matrix and coefficient matrix are nonnegative, therefore, NMF is a suitable algorithm for applications like image processing where the data are non-negative by nature. In addition, the coefficient matrix of NMF can model the features of images as an additive combination of a set of basis vectors. Compared with ICA and PCA, another advantage of NMF is that it can be designed for capturing intrinsic structures from the sample data through introducing different constraints into the basic NFM algorithm. In practice, users do not usually satisfy the condition of sparse representation of features during matrix factorization. Recently, we

noticed that numerous specialized NMF-based methods have been recently introduced by modifying the objective functions or by enforcing additional constraints [42]. For example, FNMF, presented by Babae et al. [4], added a constraint to the classical NMF. This leads to the far point still being far in the new space. Rajabi and Ghassemian [35] introduced multilayer NMF, where a sparseness constraint was applied to improve the performance of NMF. Babae et al. [5] employed the Laplacian of neighbourhood graphs to develop a graph-regularized NMF algorithm, which can preserve the locality characteristics of the underlying image. There are some other fashionable variants of NMF, such as Liu's semi-supervised NMF [22], as well as Wang's $L_{1/2}$ -NMF [38].

Recently, some works based on finite mixture model have been reported in the field of image retrieval, to fit different shapes of feature data using a multivariable probability distribution [32]. Amin et al. [2] verified a Laplacian mixture model that may model the distribution of wavelet coefficients, and also showed its superiority for video retrieval. The Gaussian Mixture Model (GMM) is another efficient method, widely applied in clustering tasks. The GMM-KL framework, proposed by Greenspan and Pinhas [16], categorized medical images through GMM along with image-matching using the Kullback-Leibler (KL) measure. In addition, a probabilistic relevance feedback method presented by Marakakis et al. [26] for CBIR also employed GMM. Piatek and Smolka [32] claimed an image retrieval scheme using GMM and considered the spatio-chromatic similarity between two images more accurately. The merit of GMM is that it can efficiently model the uncertainty with a few parameters, in addition to being easy to implement.

The aforementioned survey helped us to present a new framework, as we will present in Section 3, is different in the way that we incorporated constrained multiview NMF and Student's t -mixture model based on MRF. Specifically, this paper presents a Student's t -distribution-based retrieval and similarity ranking in retrieval phase.

2.2 Background

This subsection begins by reviewing the techniques that are most related to the proposed scheme, namely, the classical SMM and NMF.

(1) Student's t -mixture model

Let x_i , with dimension \bar{D} , $i = (1, 2, \dots, N)$, denote an observation at the i -th pixel of an image. To partition an image consisting of N pixels into K labels, SMM assumes that each observation x_i is independent of the label. The density function at an observation x_i is defined by:

$$f(x_i|\Pi, \Xi) = \sum_{j=1}^K \pi_{ij} S(x_i|\Xi_j), \quad (1)$$

where $\Pi = \{\pi_{ij}\}$, $j = (1, 2, \dots, K)$, is the set of prior distributions, and the prior distribution π_{ij} of observation x_i belonging to the j th label should satisfy the following constraints:

$$\pi_{ij} \geq 0 \quad \text{and} \quad \sum_{j=1}^K \pi_{ij} = 1. \quad (2)$$

Each Student’s t -distribution $S(x_i|\Xi_j)$ has its own parameters $\Xi_j = \{\mu_j, \Sigma_j, \bar{\nu}_j\}$ defined by [31]

$$S(x_i|\Xi_j) = \frac{\Gamma\left(\frac{\bar{\nu}_j + \bar{D}}{2}\right) |\Sigma_j|^{-\frac{1}{2}}}{(\pi \bar{\nu}_j)^{\frac{\bar{D}}{2}} \Gamma\left(\frac{\bar{\nu}_j}{2}\right) \left[1 + \bar{\nu}_j^{-1} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j)\right]^{\frac{\bar{\nu}_j + \bar{D}}{2}}}, \tag{3}$$

where Σ_j is the covariance, $|\Sigma_j|$ denotes the determinant operator of Σ_j , the vector μ_j denotes the mean, $\bar{\nu}_j$ is the number of degrees of freedom, and $(\cdot)^T$ denotes the transpose of the matrix. The Gamma function, $\Gamma(t)$, is defined by

$$\Gamma(t) = \int_0^{+\infty} s^{t-1} e^{-s} ds = (t-1)\Gamma(t-1). \tag{4}$$

If t is an integer, then $\Gamma(t) = (t-1)!$. Gamma function can also be computed by Matlab function: Gamma (real). The log-likelihood function of the density function, $f(x_i|\Pi, \Xi)$, can be expressed as

$$L(\Xi) = \log \prod_{i=1}^N f(x_i|\Pi, \Xi). \tag{5}$$

Finally, the log-likelihood function (5) must be maximized to estimate the model parameters.

(2) Overview of standard NMF

NMF attempts to seek two low-rank nonnegative matrices, $U = [U_{id}] \in \mathbb{R}_+^{M \times \bar{D}}$ and $V = [V_{dj}] \in \mathbb{R}_+^{\bar{D} \times N}$, to approximately describe an observation matrix, $X = [X_{ij}] \in \mathbb{R}_+^{M \times N}$. Mathematically, the standard NMF can be formulated as:

$$X \approx UV, \tag{6}$$

where V is customarily called the coefficient matrix of X projected on the basis matrix U . In practice, the inner dimension \bar{D} is always chosen such that $\bar{D} \ll \min(M, N)$. Obviously, this factorization leads to a compressed representation of the original matrix X . To convert the NMF process into an optimization problem, the Euclidean metric between X and UV has been popularly utilized.

$$\underset{U, V}{\operatorname{argmin}} \|X - UV\|_F^2 \quad \text{s.t. } U, V \geq 0 \tag{7}$$

where the operator $\|\cdot\|_F$ denotes the Frobenius norm. Thus far, a variety of strategies have been developed to search for a local minimum solution [19]. Lee and Seung [21] introduced the

well-known multiplicative update rule (MUR) to solve the optimization problem. When one of the matrixes is fixed, another can be updated in terms of the following expressions.

$$U_{id} \leftarrow U_{id} \frac{(XV^T)_{id}}{(UVV^T)_{id}}, \tag{8}$$

$$V_{dj} \leftarrow V_{dj} \frac{(U^T X)_{dj}}{(U^T UV)_{dj}}. \tag{9}$$

3 Proposed framework

This section addresses the proposed SCMN, which consists of four modules: underlying visual features, constrained multiview NMF, SMM based on MRF clustering, and similarity ranking. Fig. 1 illustrates the overall flowchart of the SCMN scheme involving the learning and retrieval phases. The learning phase consists of three main parts, which are feature extraction & fusion, proposed NMF and SMM-MRF clustering. The proposed NMF and SMM-MRF clustering are the crucial steps in our image retrieval system for retrieval result refinement. The retrieval phase contains three major components: (1) similarity measurement, (2) sparse query, (3) probability-based retrieval and similarity ranking. The main contribution of retrieval phase consists in the later one component. The following subsections present the implementation details of each part illustrated in Fig.1.

3.1 Feature extraction and fusion

There are several feature descriptors that can achieve the indexing purpose. Using them simultaneously will increase the memory burden on the computer. It has been proved through

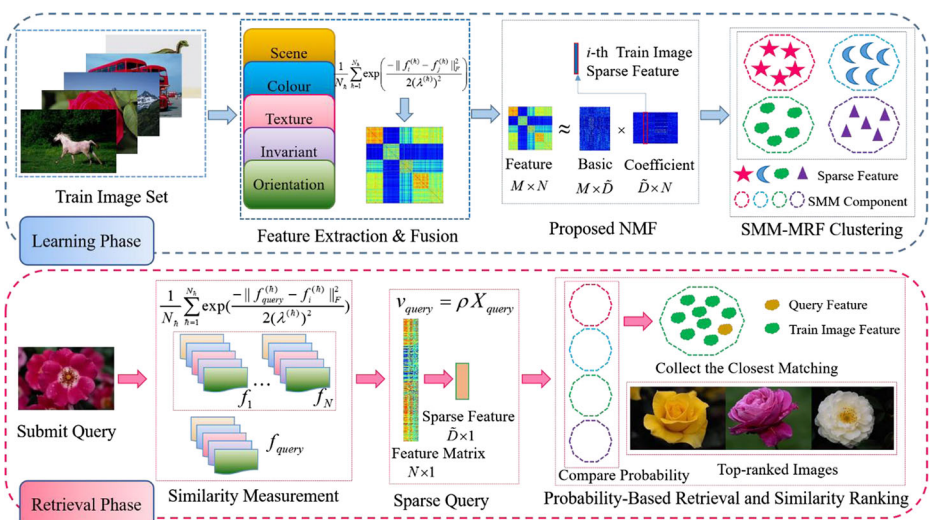


Fig. 1 Flowchart of the proposed framework

experimental results [12] that visual features, such as colour, spatial location, invariance, scene, and texture, are strongly related to human perception and are important to convey the information related to the image content. Therefore, this paper adopts these visual features and then merges them to describe the content of images.

The Histogram of Oriented Gradients (HOG) [10] is one of the most effective feature descriptors, which extracts the salient orientation information for each object. The HOG feature is less sensitive to changes in illumination, but it provides a good description of local information by calculating the gradients in local cells in eight directions of an image. Specifically, the grey-scaled image with Gamma correction is divided into 3×3 blocks, consisting of 4×4 local cells. With this method, the HOG feature is obtained with 1152-dimensionality.

Images generally contain rich scene information. They can provide high-level context to navigate the proposed algorithm for a more accurate retrieval. To extract the scene feature, the energy spectrum of the grey-scaled image is first filtered through 32 Gabor filters in 4 frequency bands with 8 orientations. Because the energy spectrum provides a scene representation invariant with respect to object arrangement and object identities. Each filtered spectrum image is then divided into several 4×4 grid sub-regions [30]. Thus, the considered image can be described by a vector with 512 dimensions, referred to as Gist.

Textural characteristics play an important role in the description of objects because real-world images usually have their own texture. Generally, there are two types of texture feature extraction algorithms such as statistical method, structure method. The former comprises Markov random field, co-occurrence matrix, and the latter contains SIFT descriptor [25], SURF descriptor, and LBP, etc. Recently, local image feature extraction algorithms create a centre of attention in recent years as they are tolerant to occlusion and distortion. It turns out that Local Binary Pattern (LBP) is one of the best local feature operators for the description of texture [1]. Using a string of binary numbers, LBP labels the grey value of each pixel by thresholding a 3×3 neighbourhood. Then, the histogram of labels serves as the texture descriptor. This paper adopts a 512-dimensional LBP feature.

In a CBIR system, the rotation-invariance feature, which is independent of the angle of the object, is rarely considered. However, most actual objects vary in orientation. This study has chosen six rotation-invariants by computing the magnitudes of Zernike moments of image [45] to characterize the rotation-invariance of objects. The 2D Zernike moment, A_{nm} , of order n with repetition m is defined using polar coordinates (r, θ) inside the unit circle as [13]

$$A_{nm} = \frac{n + 1}{\pi} \int_0^{2\pi} \int_0^1 R_{nm}(r) \exp(-jm\theta) f(r, \theta) r dr d\theta, \quad |m| \leq n \text{ and } n - |m| \text{ being even}, \quad (10)$$

here $R_{nm}(r)$ is the real-valued Zernike radial polynomials defined as

$$R_{nm}(r) = \sum_{k=0}^{(n-|m|)/2} \frac{(-1)^k (n-k)!}{k! \left(\frac{n+|m|}{2} - k\right)! \left(\frac{n-|m|}{2} - k\right)!} r^{n-2k}. \quad (11)$$

If we let the Zernike moments of an image $f(r, \theta)$ and its rotated version $f'(r, \theta)$ be A_{nm} and A'_{nm} , respectively, where $f'(r \cos \theta, r \sin \theta) = f(r \cos(\theta - \phi), r \sin(\theta - \phi))$, and θ is rotation angle. Thus, we have

$$\begin{aligned}
 A'_{nm} &= \frac{n+1}{\pi} \int_0^{2\pi} \int_0^1 R_{nm}(r) \exp(-jm\theta) f'(r\cos\theta, r\sin\theta) r dr d\theta \\
 &= \frac{n+1}{\pi} \int_0^{2\pi} \int_0^1 R_{nm}(r) \exp(-jm(\theta' + \phi)) f(r\cos\theta', r\sin\theta') r dr d\theta' \quad (12) \\
 &= \exp(-jm\phi) A_{nm}
 \end{aligned}$$

where, $\theta' = \theta - \phi$, therefore, the magnitude of Zernike moments $|A_{nm}|$ is invariant to rotation changes. Therefore, it can be taken as a rotation invariant feature of the underlying image. According to the constraints imposed on parameter n and m given in (10), we have used the following several invariants in all experiments.

$$inv^\alpha = \{|A_{22}^\alpha|, |A_{31}^\alpha|, |A_{33}^\alpha|, |A_{42}^\alpha|, |A_{44}^\alpha|, |A_{51}^\alpha|\} \quad (13)$$

The inv^f and inv^g denote rotation feature vectors of image f and g .

Almost all natural images include colour information. Therefore, it is extremely important for a retrieval scheme to select a colour descriptor. Various colour features are available for image retrieval including colour moments, colour coherence vector (CCV) [27], colour histogram (ColourHist), etc. Colour coherence vector is a more complex method than ColourHist. It classifies each pixel as either coherent or incoherent. The proposed method adopts the colour histogram, it corresponds to colour features and denotes the colour histogram of an image. ColourHist considers the colour similarity information by spreading each pixel’s total membership value to all the histogram bins. ColourHist is easy to obtain, and it is invariant to the rotation and translation of image content. According to [17], the proposed method computes a 64-bin histogram of each RGB channel and lists them together, leading to a 192-dimensional ColourHist.

Keeping the notations consistent, let $F^{(\hat{h})} = [f_1^{(\hat{h})}, f_2^{(\hat{h})}, \dots, f_N^{(\hat{h})}] \in \mathbb{R}^{D_h \times N}$ represent the training dataset, where N is the number of images, and D_h denotes the feature dimension of the considered images at the \hat{h} -th view. There are several possibilities for modelling the similarity of the matrix measure [16]. We choose Gaussian-like heat kernel functions to measure the closeness of two images, $f_i^{(\hat{h})}$ and $f_j^{(\hat{h})}$ because a heat kernel function can present a specific connection to the Laplace operator on differentiable functions [7].

$$\varpi_{ij}^{(\hat{h})} = \exp\left(\frac{-\|f_i^{(\hat{h})} - f_j^{(\hat{h})}\|_F^2}{2(\lambda^{(\hat{h})})^2}\right), \quad i, j \in [1, N], \quad (14)$$

where $\lambda^{(\hat{h})}$ is a scalable parameter. Then, the multiview feature matrix used for matrix factorization can be merged as

$$X = \frac{1}{N_h} \sum_{\hat{h}=1}^{N_h} \varpi_{ij}^{(\hat{h})} \quad (15)$$

In the current scheme, the structural constraint is achieved through a weight matrix, W , as

$$W = \frac{1}{N_h} \sum_{\hat{h}=1}^{N_h} \left(\frac{\varpi^{(\hat{h})} - I_N}{\sum_{i \neq j} \varpi_{ij}^{(\hat{h})}} \right), \quad (16)$$

where N_h is the view number. There are five features are used, thus, $N_h = 5$. I_N is a unit matrix with size $N \times N$. Equation 16 indicates that all feature vectors are incorporated into the weight

matrix, W , to achieve multiview CBIR in a real sense. It is not hard to see from (16) that the similarity matrix, W , is symmetric.

3.2 Constrained multiview NMF

The goal of this subsection is to exploit a novel, constrained NMF to preserve some of the properties of sparse features. This can be accomplished by incorporating manifold constraints about U and V into the original NMF. Considering the inherent geometric structure of each object, it becomes natural to impose structure regularization so that visually similar images are placed together.

Consequently, we can construct the following optimization problem by incorporating the structural constraint $Tr(VLV^T)$.

$$\underset{U,V}{\operatorname{arg\,min}} \|X-UV\|_F^2 + \lambda_2 Tr(VLV^T) \quad \text{s.t. } U, V \geq 0, \quad (17)$$

where $Tr(\cdot)$ denotes the trace operation, L is a Laplacian matrix $L = D-W$, D stands for a diagonal matrix whose elements correspond to $D_{jj} = \sum_i W_{ij}$, and $\lambda_2 \in \mathbb{R}^+$ is introduced to balance the reconstruction error and impact of the latter constraint.

We know the sparsity constraint from previous studies [34, 40], where the $L_{1/2}$ constraint indicated potential advantages for sparsity-promoting solutions. This characteristic inspired us to develop the $L_{1/2}$ -constrained multiview NMF, by incorporating a sparsity constraint of the basis matrix, $\|U\|_{1/2} = \sum_{i,d}^{M,D} (U_{id})^{1/2}$, into the conventional NMF.

$$\underset{U,V}{\operatorname{arg\,min}} \|X-UV\|_F^2 + 2\lambda_1 \|U\|_{1/2} \quad \text{s.t. } U, V \geq 0, \quad (18)$$

where the parameter $\lambda_1 \in \mathbb{R}^+$ balances the impact of the sparseness constraint so that the inherent sparseness property of the minimization problem is sufficiently exploited.

To emphasize the constraint of the spacing location, we expect that two closely spaced data in one space will also remain very close in another space. Here, we consider the spacing constraint term, $\exp[-\beta Tr(V\tilde{L}V^T)]$, introduced in [2],

$$\underset{U,V}{\operatorname{arg\,min}} \|X-UV\|_F^2 + 2\lambda_3 \exp[-\beta Tr(V\tilde{L}V^T)] \quad \text{s.t. } U, V \geq 0, \quad (19)$$

where $\tilde{L} = \tilde{D} - \tilde{W}$ and \tilde{D} is a diagonal matrix, whose entries are column sums of the Laplace operator, \tilde{W} , $\tilde{D}_{jj} = \sum_i \tilde{W}_{ij}$. The parameter β controls the overall contribution of the fairness property, and we chose the value $\beta = 0.01$ for all datasets. The parameter $\lambda_3 \in \mathbb{R}^+$ balances the contribution of the constraint in the objective function.

With the above considerations, the new objective function can be obtained mathematically, as follows.

$$\underset{U,V}{\operatorname{arg\,min}} \|X-UV\|_F^2 + 2\lambda_1 \|U\|_{1/2} + \lambda_2 Tr(VLV^T) \\ + 2\lambda_3 \exp[-\beta Tr(V\tilde{L}V^T)] \quad \text{s.t. } U, V \geq 0, \quad (20)$$

Obviously, it is difficult to explore a closed-form solution for the optimization problem with respect to (20). Alternatively, resorting to a multiplicative update scheme, the matrices U and V

can be alternately obtained. Considering the non-negativity of the two matrices U and V , and assuming $\Phi = [\Phi_{id}] \in \mathbb{R}_+^{M \times \bar{D}}$ and $\Psi = [\Psi_{dj}] \in \mathbb{R}_+^{\bar{D} \times N}$, we then formulate the Lagrange function \mathcal{L} as follows

$$\begin{aligned} \mathcal{L} &= \|X - UV\|_F^2 + 2\lambda_1 \|U\|_{1/2} + \lambda_2 \text{Tr}(VLV^T) + 2\lambda_3 \exp[-\beta \text{Tr}(V\tilde{L}V^T)] + \text{Tr}(\Phi U^T) + \text{Tr}(\Psi V^T) \\ &= \text{Tr}(XX^T) + \text{Tr}(UVV^T U^T) - 2\text{Tr}(XV^T U^T) + 2\lambda_1 \|U\|_{1/2} + \lambda_2 \text{Tr}(VLV^T) \\ &\quad + 2\lambda_3 \exp[-\beta \text{Tr}(V\tilde{L}V^T)] + \text{Tr}(\Phi U^T) + \text{Tr}(\Psi V^T) \end{aligned} \tag{21}$$

where Φ_{id} and Ψ_{dj} are the Lagrange multipliers. The partial derivatives of \mathcal{L} with respect to U, V are

$$\frac{\partial \mathcal{L}}{\partial U} = 2UVV^T - 2XV^T + \lambda_1 U^{-1/2} + \Phi = 0, \tag{22}$$

$$\frac{\partial \mathcal{L}}{\partial V} = 2U^T UV - 2U^T X + 2\lambda_2 VL + 2\lambda_3 \exp[-\beta \text{Tr}(V\tilde{L}V^T)](-2\beta V\tilde{L}) + \Psi = 0. \tag{23}$$

Using the Karush-Kuhn-Tucker (KKT) conditions [6], where $\Phi_{id} U_{id} = 0, \forall i, d$, the following equation for U_{id} can be obtained.

$$\left(2UVV^T + \lambda_1 U^{-1/2}\right)_{id} \cdot U_{id} - (2XV^T)_{id} \cdot U_{id} + \Phi_{id} \cdot U_{id} = 0. \tag{24}$$

Transposition and division leads to the update rule for matrix U

$$U_{id} \leftarrow U_{id} \frac{(XV^T)_{id}}{(UVV^T + (\lambda_1/2)U^{-1/2})_{id}}. \tag{25}$$

In the same manner, based on the KKT conditions, $\Psi_{dj} V_{dj} = 0, \forall d, j$, the following equation was obtained by multiplying both sides with V_{dj} .

$$\begin{aligned} &\left(2U^T UV + 2\lambda_2 V \cdot D + 4\lambda_3 \beta \cdot \exp[-\beta \text{Tr}(V\tilde{L}V^T)] \cdot (V\tilde{W})_{dj}\right) \cdot V_{dj} \\ &\quad - \left(2U^T X + 2\lambda_2 V \cdot W + 4\lambda_3 \beta \cdot \exp[-\beta \text{Tr}(V\tilde{L}V^T)] \cdot (V\tilde{D})_{dj}\right) \cdot V_{dj} + \Psi_{dj} \cdot V_{dj} = 0 \end{aligned} \tag{26}$$

Solving the above equation leads to the resulted multiplicative update rules for V_{dj}

$$V_{dj} \leftarrow V_{dj} \frac{(U^T X + \lambda_2 VW + 2\lambda_3 \beta \cdot \exp[-\beta \text{Tr}(V\tilde{L}V^T)] \cdot (V\tilde{D}))_{dj}}{(U^T UV + \lambda_2 VD + 2\lambda_3 \beta \cdot \exp[-\beta \text{Tr}(V\tilde{L}V^T)] \cdot (V\tilde{W}))_{dj}}. \tag{27}$$

In this way, U and V can be updated iteratively until the objective function (20) converges, or the predefined iterations are achieved. With regard to convergent properties of the proposed update schemes, we have the following theorem:

Theorem 1: The objective function in (20) is non-increasing under the update rules in (25) and (27).

This theorem can ensure that the proposed objective function (20) converges to a local minimum, and its proof is presented in section 4.

3.3 SMM-based clustering

NMF has been widely applied because of its capabilities in keeping the intrinsic features of low-dimensional space. To be clear, each column of the coefficient matrix V indicates the characteristics of the associated training sample. Therefore, any one clustering algorithm can be implemented to V , to associate each sample with a given cluster. This paper is particularly interested in a statistical method to label the data $V = [v_1, v_2, \dots, v_N]^T$. In other words, the data of V can be modelled using a mixture of Student’s t -distribution. This is because, in theory, the arbitrarily shaped distribution can be approximated using a mixture of probability density functions, provided that this mixture has numerous components. Note that v_N is a \bar{D} -dimensional feature vector. Labels are denoted by $(\Xi_1, \Xi_2, \dots, \Xi_K)$. We define the posterior probability density function to the partition matrix, Π , of N columns into K labels.

$$p(\Pi, \Xi|V) \propto p(V|\Pi, \Xi)p(\Pi). \tag{28}$$

Thus, the new joint conditional probability density of the data (the column of coefficient matrix V) can be represented via multivariable SMM in the form:

$$p(V|\Pi, \Xi) = \prod_{i=1}^N \left[\sum_{j=1}^K \pi_{ij} S(v_i|\Xi_j) \right], \tag{29}$$

where $S(v_i|\Xi_j)$ is the multivariable Student’s t -distribution declared by (3), and π_{ij} is the prior probability that x_i belongs to the label Ξ_j . Considering the spatial information between neighbouring columns of the coefficient matrix V , we have introduced the MRF in the form

$$p(\Pi) = \exp(-\bar{U}(\Pi)), \tag{30}$$

where $\bar{U}(\Pi)$ is the smooth prior. By combining (28), (29), and (30), the log-likelihood of (28) can be expressed by the following formula.

$$\begin{aligned} L(\Pi, \Xi|V) &= \log p(\Pi, \Xi|V) \\ &= \sum_{i=1}^N \log \left\{ \sum_{j=1}^K \pi_{ij} S(v_i|\Xi_j) \right\} - \bar{U}(\Pi). \end{aligned} \tag{31}$$

We choose the smooth prior $\bar{U}(\Pi)$ in terms of Nguyen and Wu’s work [28]

$$\bar{U}(\Pi) = - \sum_{i=1}^N \sum_{j=1}^K \exp \left(\gamma \sum_{m \in N_i} (z_{mj} + \pi_{mj}) \right) \log \pi_{ij}, \tag{32}$$

where N_i is the size of the window; in this paper, e.g., $N_i = 9$ for a 3×3 window. The parameter γ controls the impact of smoothing. Generally, it has a value in the range of [0.5, 3]; in our

experiment, it has been set to 2.5 ($\gamma = 2.5$). This smoothing function acts as a linear filter for smoothing images contaminated by noise. The smooth prior is only modelled as a combination of the posteriors z_{mj} and priors π_{mj} in the previous step and is therefore easy to implement. Consequently, according to (31)–(32), the log-likelihood of (31) can be stated by

$$L(\Pi, \Xi|V) = \sum_{i=1}^N \log \left\{ \sum_{j=1}^K \pi_{ij} S(v_i|\Xi_j) \right\} + \sum_{i=1}^N \sum_{j=1}^K \exp \left(\gamma \sum_{m \in N_i} (z_{mj} + \pi_{mj}) \right) \log \pi_{ij}. \tag{33}$$

Next, we apply Jensen’s inequality in the form of $\log \left(\sum_{j=1}^K z_{ij} \varsigma \right) \geq \sum_{j=1}^K z_{ij} \log(\varsigma)$ to modify the above expression. Thus, maximizing the log-likelihood function $L(\Pi, \Xi|V)$ results in an increase in the value of the following objective function.

$$J(\Pi, \Xi|V) = \sum_{i=1}^N \sum_{j=1}^K z_{ij} \{ \log \pi_{ij} + \log S(v_i|\Xi_j) \} + \sum_{i=1}^N \sum_{j=1}^K \exp \left(\gamma \sum_{m \in N_i} (z_{mj} + \pi_{mj}) \right) \log \pi_{ij}. \tag{34}$$

After considering Bayesian theory, the posterior probability z_{ij} in (34) at the current iteration step is

$$z_{ij}^{(t+1)} = \frac{\pi_{ij}^{(t)} S(v_i|\Xi_j^{(t)})}{\sum_{m=1}^K \pi_{im}^{(t)} S(v_i|\Xi_m^{(t)})}. \tag{35}$$

Next, we need to maximize the function (34). For this, the EM algorithm is implemented by taking the derivative of $J(\Pi, \Xi|V)$ with respect to each parameter in set $\{\Pi, \Xi\}$, and then equating its value to zero. Thus, the solution $\partial J(\Pi, \Xi|V)/\partial \mu_j = 0$ yields the estimates of mean μ_j at the $(t + 1)$ set by

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^N z_{ij}^{(t)} u_{ij}^{(t)} v_i}{\sum_{i=1}^N z_{ij}^{(t)} u_{ij}^{(t)}}, \tag{36}$$

where the numerical solution of $u_{ij}^{(t)}$ is denoted as

$$u_{ij}^{(t)} = \frac{\bar{v}_j^{(t)} + \bar{D}}{\bar{v}_j^{(t)} + (v_i - \mu_j^{(t)})^T \Sigma_j^{-1(t)} (v_i - \mu_j^{(t)})}. \tag{37}$$

Similar to the computation of the mean μ_j , let $\partial J(\Pi, \Xi|V)/\partial \Sigma_j = 0$. Then, the estimation of covariance Σ_j is formulated as

$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^N z_{ij}^{(t)} u_{ij}^{(t)} (v_i - \mu_j^{(t)}) (v_i - \mu_j^{(t)})^T}{\sum_{i=1}^N z_{ij}^{(t)}}. \tag{38}$$

Using the constraint $\sum_{j=1}^K \pi_{ij} = 1$, the solution to $\partial J(\Pi, \Xi|V)/\partial \pi_{ij} = 0$ enables the following iterative expression for prior probability π_{ij}

$$\pi_{ij}^{(t+1)} = \frac{z_{ij}^{(t)} + \exp\left(\gamma \sum_{m \in N_i} (z_{mj}^{(t)} + \pi_{mj}^{(t)})\right)}{1 + \sum_{h=1}^K \exp\left(\gamma \sum_{m \in N_i} (z_{mh}^{(t)} + \pi_{mh}^{(t)})\right)}. \tag{39}$$

Finally, we consider the estimates of the degrees of freedom, \bar{v}_j , which are obtained through the derivation of $J(\Pi, \Xi | V)$, with \bar{v}_j at the $(t + 1)$ iteration step given by

$$\log\left(\frac{\bar{v}_j^{(t+1)}}{2}\right) - \psi\left(\frac{\bar{v}_j^{(t+1)}}{2}\right) + 1 - \log\left(\frac{\bar{v}_j^{(t)} + \bar{D}}{2}\right) + \psi\left(\frac{\bar{v}_j^{(t)} + \bar{D}}{2}\right) + \frac{\sum_{j=1}^N z_{ij}^{(t)} (\log u_{ij}^{(t)} - u_{ij}^{(t)})}{\sum_{j=1}^N z_{ij}^{(t)}} = 0, \tag{40}$$

where $\psi(x) = \{\partial \Gamma(x) / \partial(x)\} / \Gamma(x)$ is the digamma function. Then, the final clustering results can be obtained in terms of the posterior probability, and it follows that

$$v_i \in \Xi_j : \text{if } z_{ij} \geq z_{im}, \quad j, m = 1, 2, \dots, K \tag{41}$$

3.4 Similarity ranking

The computation of similarity is still an important problem in CBIR systems. Different similarity measurements lead to different results. In this study, to retrieve a list of “similar” images, we first calculate previously defined image features, and the similarity measurements between the query image and images under consideration are then obtained by

$$X_{\text{query}} = \frac{1}{N_h} \sum_{h=1}^{N_h} \exp\left(\frac{-\|f_{\text{query}}^{(h)} - f_i^{(h)}\|_F^2}{2(\lambda^{(h)})^2}\right), \quad i \in [1, N], \tag{42}$$

where $f_{\text{query}}^{(h)}$ denotes the h -th feature of the query image while $f_i^{(h)}$ refers to the same type of feature of each sample in the training database. We generated a linear projection matrix, ρ , which is defined as

$$\rho = (U^T U)^{-1} U^T. \tag{43}$$

To capture the sparse representation of the query image, we project the kernel matrix (42) into low-dimensional space in the hope that new matrix, v_{query} , can perform a parts-based preservation of the latent feature information of X_{query} . This can be achieved via the projection matrix, ρ , formulated as

$$v_{\text{query}} = \rho X_{\text{query}}. \tag{44}$$

To rank the training images that are strongly related to the query image, the probability that the query image falls into the correct SMM component should be calculated in advance. After obtaining all probabilities, $\{S_1(v_{\text{query}} | \Xi_1), S_2(v_{\text{query}} | \Xi_2), \dots, S_K(v_{\text{query}} | \Xi_K)\}$, for a given query image, the task of the retrieval system is to collect the most closely matching samples from a

collection in the database. For those that belong to one component of SMM, the similarity is compared by ranking the probabilities, $\{S_j(v_i | \Xi_j) | v_i \in \Xi_j\}$, $i = 1, \dots, N, j = 1, \dots, K$ of participants in descending order. With this rule, all training samples can be sorted in a reverse order. This ensures maximal matches for any pre-defined query image using this queue. The returned image with a retrieval length, ℓ , which is in the form of a percentage, can be regarded as the $N \times \ell$ images nearest to the query image. Algorithm 1 describes the general procedure, which we refer to as SCMN.

Algorithm 1: Image retrieval through merging SMM-MRF and constrained multiview NMF

Input: The inner dimension \bar{D} ; the regularized parameters $(\lambda_1, \lambda_2, \lambda_3)$; $\gamma = 2.5$, $\beta = 0.01$, the components K of SMM, image dataset.

Output: Set of images similar to the query image.

- 1: Calculate all visual features for each training sample, then obtain Gaussian-like heat kernel weighting ϖ and similarity measurement matrix W in terms of (14) and (16), respectively;
 - 2: Compute the symmetric matrix $L = D - W$ and $\tilde{L} = \tilde{D} - \tilde{W}$;
 - 3: Initialize the basic matrix U and the coefficient matrix V by randomly choosing entries in $[0, 1]$;
 - 4: **Iterate**
 - 5: Update the basic matrix U using (25) and normalize each column of it;
 - 6: Update the coefficient matrix V using (27);
 - 7: **Until** convergence criterion satisfied or iteration terminated;
 - 8: Initialize the SMM parameters: covariance Σ_j , prior probability π_{ij} , freedom of degree $\bar{\nu}_j$, and mean μ_j ;
 - 9 (E-step): Estimate the posterior probability z_{ij} using (35);
 - 10 (M-step): Update the covariance Σ_j , the freedom of degree $\bar{\nu}_j$, the prior probability π_{ij} , and the mean μ_j in terms of (38), (40), (39), and (36), respectively;
 - 11: Repeat steps 7-10 until reaching the maximum number of iterations; subclusters can be obtained by (41);
 - 12: Calculate the similarity matrix for each given query image using (42), map this matrix using (43) for part-based representation.
 - 13: Calculate $S_j(v_{query} | \Xi_j)$ for every query image and then collect the most closely matching samples from a collection, by ranking the probabilities, $\{S_j(v_i | \Xi_j) | v_i \in \Xi_j\}$.
-

4 Convergence and complexity analysis

4.1 Convergence analysis

To prove Theorem 1, we need to prove that the following objective function $F(U, V)$ is non-increasing under the update formulae (25) and (27).

$$F(U, V) = \|X - UV\|_F^2 + 2\lambda_1 \|U\|_{1/2} + \lambda_2 \text{Tr}(VLV^T) + 2\lambda_3 \exp[-\beta \text{Tr}(V\tilde{L}V^T)]. \quad (45)$$

The proof of convergence will make use of an auxiliary function, which has the following characteristics.

Lemma 1. If $h(x, x')$ is an auxiliary function of $F(x)$ and the condition $h(x, x') \geq F(x)$ and $h(x, x) = F(x)$ are satisfied for any given x, x' , then, F will be convergent under the update

$$x^{(t+1)} = \underset{x}{\text{argmin}} h(x, x^{(t)}). \quad (46)$$

Proof:

The known conditions obviously lead to the following expression.

$$F(x^{(t+1)}) \leq h(x^{(t+1)}, x^{(t)}) \leq h(x^{(t)}, x^{(t)}) = F(x^{(t)}). \quad (47)$$

Therefore, the equality, $F(x^{(t+1)}) = F(x^{(t)})$, holds only if $x^{(t)}$ is the local minimum of $h(x, x^{(t)})$.

Since the update schemes defined by (25) and (27) are element-wise in nature, letting U be a constant, it is enough to verify that $F(U, V) = F(U)$ is non-increasing for any element U_{id} in U . To achieve this, we define the following auxiliary function with respect to $U_{id}^{(t)}$

$$h(u, U_{id}^{(t)}) = F(U_{id}) + F'(U_{id}^{(t)}) \cdot (u - U_{id}^{(t)}) + \frac{(UVV^T + \frac{\lambda_1}{2}U^{-\frac{1}{2}})_{id}}{U_{id}^{(t)}} \cdot (u - U_{id}^{(t)})^2. \quad (48)$$

Observing the above expression; it is not hard to find that $h(U_{id}^{(t)}, U_{id}^{(t)}) = F(U_{id}^{(t)})$. Thus, the problem is equivalent to proving that $h(u, U_{id}^{(t)}) \geq F_{id}(u)$. We first compute the Taylor series expansion of $F(u)$ as

$$\begin{aligned} F(u) &= F(U_{id}) + F'(U_{id}^{(t)}) \cdot (u - U_{id}^{(t)}) + \frac{1}{2} F''(U_{id}^{(t)}) (u - U_{id}^{(t)})^2 \\ &= F(U_{id}) + F'(U_{id}^{(t)}) \cdot (u - U_{id}^{(t)}) + \left((VV^T)_{dd} - \left(\frac{\lambda_1}{4} U^{-\frac{3}{2}} \right)_{id} \right) (u - U_{id}^{(t)})^2 \end{aligned} \quad (49)$$

where $F'(U_{id})$ and $F''(U_{id})$ are the corresponding first-order and second-order derivatives of the objective function (41), relevant to the variable U_{id} .

$$F'(U_{id}) = \left(2UVV^T - 2XV^T + \lambda_1 U^{-1/2} \right)_{id}. \quad (50)$$

$$F''(U_{id}) = (2VV^T)_{dd} - \left(\frac{\lambda_1}{2}U^{-3/2}\right)_{id}. \tag{51}$$

It is easy to verify that

$$(UVV^T)_{id} = \sum_l U_{il}^{(t)}(VV^T)_{ld} \geq U_{id}^{(t)} \cdot (VV^T)_{dd}. \tag{52}$$

Additionally, it is easy to see

$$\left(\frac{\lambda_1}{2}U^{-\frac{1}{2}}\right)_{id} \geq \left(\frac{\lambda_1}{2}U_{id}^{-\frac{1}{2}}\right) \cdot \left(-\frac{1}{2}\right) = -\frac{\lambda_1}{4}U_{id}^{-\frac{3}{2}} \cdot U_{id}^{(t)}. \tag{53}$$

Comparing the Taylor series expansion of $F(u)$ to the auxiliary function (48), and combining (52) and (53) leads to the following inequality.

$$h(u, U_{id}^{(t)}) \geq F(u). \tag{54}$$

Substituting $h(u, U_{id}^{(t)})$ of (48) into (46), we obtain

$$U_{id}^{(t+1)} = \underset{u}{\operatorname{argmin}} h(u, U_{id}^{(t)}). \tag{55}$$

The first-order derivative of $h(u, U_{id}^{(t)})$ with respect to u is

$$\frac{\partial h(u, U_{id}^{(t)})}{\partial u} = F'(U_{id}^{(t)}) + \frac{2(UVV^T + \frac{\lambda_1}{2}U^{-\frac{1}{2}})_{id}}{U_{id}^{(t)}} \cdot (u - U_{id}^{(t)}) = 0. \tag{56}$$

Using (50), the above expression reduces to

$$\left(2UVV^T - 2XV^T + \lambda_1 U^{-\frac{1}{2}}\right)_{id} \cdot U_{id}^{(t)} + \left(2UVV^T + \lambda_1 U^{-\frac{1}{2}}\right)_{id} \cdot (u - U_{id}^{(t)}) = 0. \tag{57}$$

From (57), we can conclude the value of u by

$$u = \frac{(XV^T)_{id}}{(UVV^T + (\lambda_1/2)U^{-1/2})_{id}} \cdot U_{id}^{(t)} = U_{id}^{(t+1)}. \tag{58}$$

Finally, according to (48), (54), and (58), we can derive

$$F(U_{id}^{(t+1)}) \leq h(U_{id}^{(t+1)}, U_{id}^{(t)}) \leq h(U_{id}^{(t)}, U_{id}^{(t)}) = F(U_{id}^{(t)}). \tag{59}$$

Similar to the auxiliary function modeled for U_{id} , we present another auxiliary function for G_{dj} by

$$h(v, V_{dj}^{(t)}) = F(V_{dj}^{(t)}) + F'(V_{dj}^{(t)})(v - V_{dj}^{(t)}) + \frac{(U^TUV + \lambda_2VD + 2\lambda_3\beta\exp[-\beta\operatorname{Tr}(V\tilde{L}V^T)] \cdot (V\tilde{W}))_{dj}}{V_{dj}^{(t)}}(v - V_{dj}^{(t)})^2. \tag{60}$$

Since it is obvious that $h(V_{dj}^{(t)}, V_{dj}^{(t)}) = F(V_{dj}^{(t)})$, the Taylor series expansion of $F(v)$ is then utilized to prove the inequality $h(g, V_{dj}^{(t)}) \geq F_{dj}(v)$, expressed as

$$\begin{aligned}
 F(v) &= F(V_{dj}^{(t)}) + F'(V_{dj}^{(t)}) \cdot (v - V_{dj}^{(t)}) + \frac{1}{2} F''(V_{dj}^{(t)}) (v - V_{dj}^{(t)})^2 = F(V_{dj}^{(t)}) + F'(V_{dj}^{(t)}) \cdot (v - V_{dj}^{(t)}) \\
 &+ \left[(U^T U)_{dd} + (\lambda_2 L)_{jj} + \lambda_3 \exp[-\beta Tr(V \tilde{L} V^T)]_{dd} \times (4\beta^2 (\tilde{V} \tilde{L}))_{dj}^2 - (2\beta \tilde{L})_{jj} \right] \cdot (v - V_{dj}^{(t)})^2 \tag{61}
 \end{aligned}$$

In the above, the partial derivatives of F with respect to V_{dj} can be calculated as follows:

$$F'(V_{dj}) = (2U^T UV - 2U^T X + 2\lambda_2 VL + 2\lambda_3 \cdot \exp[-\beta Tr(V \tilde{L} V^T)]) \cdot (-2\beta V \tilde{L})_{dj}, \tag{62}$$

$$F''(V_{dj}) = (2U^T U)_{dd} + (2\lambda_2 L)_{jj} + 2\lambda_3 \cdot \exp[-\beta Tr(V \tilde{L} V^T)]_{dd} \cdot (4\beta^2 \cdot (\tilde{V} \tilde{L}))_{dj}^2 - (2\beta \cdot \tilde{L})_{jj} \tag{63}$$

It is found that the following inequalities are established

$$(U^T UV)_{dj} = \sum_i (U^T U)_{di} \cdot V_{ij} \geq (U^T U)_{dd} \cdot V_{dj}^{(t)} \tag{64}$$

$$(\lambda_2 VD)_{dj} = \sum_i \lambda_2 \cdot V_{di} (D)_{ij} \geq \lambda_2 \cdot V_{dj}^{(t)} (D)_{jj} \geq \lambda_2 \cdot V_{dj}^{(t)} (D - W)_{jj} = \lambda_2 \cdot V_{dj}^{(t)} (L)_{jj} \tag{65}$$

Based on this analysis, it easy to observe that while parameter β takes a small enough value, we have

$$(V \tilde{W})_{dj} \geq 2\beta (\tilde{V} \tilde{L})_{dj}^2 - (\tilde{L})_{jj} \tag{66}$$

Consequently, we derive

$$h(v, V_{dj}^{(t)}) \geq F(v) \tag{67}$$

Likewise, substituting $h(v, V_{jd}^{(t)})$ into (46), the update scheme for V in (27) can be obtained as a local optimum of the auxiliary function (60):

$$V_{dj}^{(t+1)} = \underset{v}{\operatorname{argmin}} h(v, V_{dj}^{(t)}) \tag{68}$$

This is because the derivation of $h(v, V_{jd}^{(t)})$ with respect to the variable v is as follows.

$$\frac{\partial h(v, V_{dj}^{(t)})}{\partial v} = F'(V_{dj}^{(t)}) + \frac{2}{G_{dj}^{(t)}} (B^T BV + \lambda_2 VD + 2\lambda_3 \beta \cdot \exp[-\beta Tr(V \tilde{L} V^T)] \cdot (VW_2))_{dj} \cdot (v - V_{dj}^{(t)}) \tag{69}$$

The above equation can be simplified to

$$\begin{aligned}
 &(2U^T UV - 2U^T X + 2\lambda_2 VL + 2\lambda_3 \exp[-\beta Tr(V \tilde{L} V^T)]) \cdot (-2\beta V \tilde{L})_{dj} V_{dj}^{(t)} \\
 &+ (2U^T UV + 2\lambda_2 VD + 2\lambda_3 \exp[-\beta Tr(V \tilde{L} V^T)]) \cdot (2\beta V \tilde{W})_{dj} (v - V_{dj}^{(t)}) = 0 \tag{70}
 \end{aligned}$$

Thus, we obtain the following equation:

$$\begin{aligned}
 v &= \frac{(U^T X + \lambda_2 VW + 2\lambda_3 \beta \cdot \exp[-\beta \cdot Tr(V \tilde{L} V^T)] \cdot (V \tilde{D}))_{dj} \cdot V_{dj}^{(t)}}{(U^T UV + \lambda_2 VD + 2\lambda_3 \beta \cdot \exp[-\beta \cdot Tr(V \tilde{L} V^T)] \cdot (V \tilde{W}))_{dj}} \cdot V_{dj}^{(t)} \\
 &= V_{dj}^{(t+1)} \tag{71}
 \end{aligned}$$

This leads to the following inequality.

$$h\left(V_{dj}^{(t+1)} \cdot V_{dj}^{(t)}\right) \leq h\left(V_{dj}^{(t)} \cdot V_{dj}^{(t)}\right) \quad (72)$$

The comparison of the above inequality collectively and using (60), (67) and (72) yields

$$F\left(V_{dj}^{(t+1)}\right) \leq h\left(V_{dj}^{(t+1)} \cdot V_{dj}^{(t)}\right) \leq h\left(V_{dj}^{(t)} \cdot V_{dj}^{(t)}\right) = F\left(V_{dj}^{(t)}\right) \quad (73)$$

Finally, according to (59) and (73), we obtain

$$F\left(U_{id}^{(t+1)}, V_{dj}^{(t+1)}\right) \leq F\left(U_{id}^{(t)} \cdot V_{dj}^{(t+1)}\right) \leq F\left(U_{id}^{(t)} \cdot V_{dj}^{(t)}\right) \quad (74)$$

The convergence of Theorem 1 is proved.

4.2 Complexity analysis

The cost of the proposed SCMN learning phase mainly contains two parts. The first part is for the constructions of heat kernel function ϖ and Laplacian matrix L . If we use the big \mathcal{O} notation to represent the complexity of the algorithm, the time complexity of this part is $\mathcal{O}(2(\sum_{i=1}^n D_h)N^2)$. The second part is for the matrix factorization. The main computational costs are in the update steps of the matrices U and V . Therefore, in view of the update schemes summarized in the above section, we count the number of floating-point operations, including addition/subtraction (Fladd), multiplication (Flmlt), and division (Fldiv). Table 1 lists the floating-point arithmetic operations involved in updating each matrix. Consequently, assuming that the multiplicative update rule terminates after t iterations, the total cost of the SCMN is $\mathcal{O}(t(MN\bar{D})^2)$. Table 2 gives the computational complexity of the proposed SCMN and compares it to the standard NMF. From Table 2, we can draw conclusion that the SCMN is moderately more expensive than the classical NMF for a single update. The complexity of SCMN is mainly caused by the application of sparsity constraints.

5 Experimental results

In this section, to assess the performance of the SCMN, we report a series of retrieval experiments to compare the proposed method against other NMF-based algorithms and discuss its computational complexity and convergence rate. All experiments are run on a personal computer with an Intel (R) Core (TM) i7–2600 3.4GHz CPU and 8GB RAM. Numerical simulations have been carried out in Matlab 7.11.0 (2010b) in a Windows environment.

5.1 Data corpora

This study conducts experiments over four publicly available datasets: (1) Caltech101¹; (2) Corel 1 K²; (3) Corel 5 K²; and (4) WdcImageData.³ These datasets are diverse enough to cope with different themes of image retrieval tasks.

¹ http://www.vision.caltech.edu/Image_Datasets/Caltech101

² <http://wang.ist.psu.edu/docs/related/>

³ <http://imagedatabase.cs.washington.edu/>

Table 1 Floating-point computational times for multiplication of matrices

Matrices	Fladd	Fmlt
$U^T X$	$MN\bar{D}$	$MN\bar{D}$
$U^T UV$	$(M + N)\bar{D}^2$	$(M + N)\bar{D}^2$
$Tr(V\bar{L}V^T)$	$2N^2\bar{D}$	$2N^2\bar{D}$
VW	$MN\bar{D}$	$MN\bar{D}$
VD	$MN\bar{D}$	$MN\bar{D}$
$V^{-1/2}$	$(N\bar{D})^2$	$(N\bar{D})^2$

The Caltech101 benchmark dataset contains 9146 images of variable size, with 101 different object categories and another additional background category. In our experiment, 500 images from the Caltech101 collection are selected to form ten subsets for training, and another 30 images constitute a test set. All images are resized so that each side of an image is 128 pixels and is in RGB colour.

The Corel 1 K dataset consists of 1000 colour images in the JPEG format. Our training set contains five different categories with 100 samples per category, including flowers, buses, mountains, horses, and dinosaurs. The query images come from the test set, which consists of 40 images. All images have the same resolution, either 256×384 or 384×256 , in the range 0 to 255 in each of the R, G and B colour channels.

The Corel 5 K image dataset is composed of 5000 colour images in 50 categories, with two sizes, 192×128 and 128×192 . This is a relatively larger image dataset including diverse contents, such as animals, airplanes, trees, and stained glass. In our retrieval experiment, the training set contains 360 images of 192×128 pixels for the convenience of feature extraction. These are classified into four categories. Another 32 images are selected randomly as query images for test purposes.

This paper also applies the WdcImageData dataset, which consists of 1333 colour images of 22 categories, to evaluate the validity of the proposed SCMN. The images are very loosely grouped by category, including trees, people, sea, animals, buildings, and so on. In this work, the training set contains 300 images divided into six categories and for each category, there are unequal numbers of samples with a size of 756×504 pixels in the JPEG format. The test set used is a collection of 36 randomly selected samples.

5.2 Evaluation metrics

For an overall evaluation of performance, the two most commonly used metrics Precision (PR) and Recall Rate (RR) [41] are implemented to measure the accuracy of image retrieval. Precision measures the effectiveness of the underlying method to

Table 2 Floating-point computational times for a single iteration in SCMN and NMF

	SCMN	NMF
Fladd	$4MN\bar{D} + 2(M + N)\bar{D}^2 + 4(MN\bar{D})^2 + (N\bar{D})^2$	$2MN\bar{D} + 2(M + N)\bar{D}^2$
Fmlt	$4MN\bar{D} + 2(M + N)\bar{D}^2 + 4(MN\bar{D})^2 + (N\bar{D})^2$	$2MN\bar{D} + 2(M + N)\bar{D}^2 + (M + N)\bar{D}$
Fldiv	$(M + N)\bar{D}$	$(M + N)\bar{D}$
Overall	$\mathcal{O}((MN\bar{D})^2)$	$\mathcal{O}(MN\bar{D})$

retrieve only images that are relevant. It is defined as the ratio of relevant images to all retrieved images

$$Precision = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}}$$

Recall computes the ratio of the retrieved, relevant images to all the relevant images in the dataset. It is used for assessing the capabilities of the algorithm to retrieve all images that are relevant and is defined by

$$Recall = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images in the dataset}}$$

5.3 Parameter setting

Before evaluating the SCMN, an analysis of the empirical parameter sensitivity is necessary. We will investigate the impact of inner-dimensional \bar{D} and three other parameters, $\lambda_1, \lambda_2,$ and $\lambda_3,$ to find the optimal response under a wide range of parameter values.

The initial check is carried out to ensure that the response of \bar{D} works on the proposed framework. Thus, \bar{D} is changed while the other parameters are set to 0.1 for all datasets. We set the retrieval length to a fixed value of 0.1 in all experiment. Generally, the top 10 most similar targeted images are adopted by the most image retrieval system. In this paper, we wish there were more retrieval images to involve in assessing. If the retrieval length is selected to be 0.1, for $N = 200$ and retrieval length $\ell = 0.1,$ the returned image set has 20 images. Thus, we set the retrieval length to 0.1 and run the SCMN with varying values of $\bar{D}.$ It is found that our method seems to work well in practice when \bar{D} is equal to 8, 16, 64, and 32 for Caltech-101, Corel 1 K, Corel 5 K, and WdcImageData, respectively. Table 3 presents the PR and RR responses of the SCMN. Next, we fix \bar{D} for different datasets, based on the above analysis, and vary the parameter λ_1 to find the best performance for each dataset. The comparative results are illustrated in Table 4, where bold values indicate the best result of the column. Again, another parameter λ_2 is used to assess the effect of the regularization term. In the same way, setting \bar{D}, λ_1 in terms of the

Table 3 Comparison retrieval response in the light of \bar{D} by fixing retrieval length ($\ell = 0.1$), $\lambda_1 = 0.1, \lambda_2 = 0.1,$ and $\lambda_3 = 0.1$ for all datasets

\bar{D}	Caltech-101		\bar{D}	Corel 1 K		\bar{D}	Corel 5 K		\bar{D}	WdcImageData	
	RR	PR		RR	PR		RR	PR		RR	PR
4	0.2031	0.2072	8	0.4556	0.9128	8	0.2704	0.6829	4	0.3423	0.5316
8*	0.3669	0.3710	16*	0.4730	0.9463	16	0.3015	0.7675	8	0.3711	0.6061
16	0.3164	0.3159	32	0.4717	0.9457	32	0.3144	0.7996	16	0.3690	0.6124
32	0.3485	0.3507	64	0.4590	0.9213	64*	0.3155	0.8042	32*	0.3767	0.6212
64	0.2932	0.2942	96	0.4496	0.9000	96	0.2797	0.7096	64	0.3625	0.5884
96	0.252	0.2377	128	0.4463	0.8970	128	0.3115	0.7932	96	0.3521	0.5821

*Bold values indicate the best result of the column

Table 4 Comparison retrieval response in the light of λ_1 by fixing retrieval length ($\ell = 0.1$), $\lambda_2 = 0.1$, and $\lambda_3 = 0.1$ for all datasets. $\bar{D} = 8, 16, 64, 32$ for Caltech-101, Corel 1 K, Corel 5 K, and WdcImageData, respectively

λ_1	Caltech-101		λ_1	Corel 1 K		λ_1	Corel 5 K		λ_1	WdcImageData	
	RR	PR		RR	PR		RR	PR		RR	PR
10^{-5}	0.3438	0.3449	10^{-5}	0.4730	0.9463	10^{-5}	0.3135	0.7987	10^{-5}	0.3911	0.6326
10^{-3}	0.3399	0.3377	10^{-3}	0.4750	0.9500	10^{-3}	0.3075	0.7776	10^{-3}	0.3887	0.6376
10^{-2}	0.3568	0.3594	10^{-2}	0.4857	0.9707	10^{-2}	0.3225	0.8189	10^{-2}	0.3995	0.6679
10^{-1*}	0.3581	0.3695	10^{-1}	0.4714	0.9415	10^{-1}	0.3292	0.8364	10^{-1}	0.3509	0.5593
1	0.3039	0.3043	1	0.4590	0.9213	1	0.3002	0.7629	1	0.3371	0.5745
10	0.2045	0.2072	10	0.4309	0.8659	10	0.3005	0.7601	10	0.3287	0.5253

*Bold values indicate the best result of the column

favourable results indicated by the above analysis, we repeat the image retrieval operation using SCMN, where the value of λ_2 is increased from 10^{-5} to 10, in increments of factors of 10. Table 5 demonstrates the obtained performance. Finally, to seek a suitable value for λ_3 , we follow a similar operation by fixing the parameters \bar{D} , λ_1 , and λ_2 and varying only the value of λ_3 . Table 6 shows that the SCMN is not very sensitive to the parameter λ_3 . In fact, one can observe that this also holds true with other parameters. Therefore, unless otherwise specified, in our algorithm, λ_1 , λ_2 , and λ_3 are set equal to 0.1 for all experiments. In contrast, inner-dimensional \bar{D} makes this significantly different. The experiment should therefore be adjusted in accordance with specific datasets.

5.4 Retrieval results

This subsection will evaluate the response of our SCMN system for retrieving images, with experiments conducted using the parameters discussed above. The test selects as many different categories as possible, especially categories 10, 5, 4, and 6 for Caltech-101, Corel 1 K, Corel 5 K, and WdcImageData, respectively. Taking into account that the SCMN merges with method of multiview NMF and the finite mixture model with the MRF, this study compares the performance of SCMN with four algorithms which are

Table 5 Comparison retrieval response in the light of λ_2 by fixing retrieval length ($\ell = 0.1$), $\lambda_3 = 0.1$ for all datasets, $\lambda_1 = 0.1, 0.01, 0.1, 0.01$, and $\bar{D} = 8, 16, 64, 32$ for Caltech-101, Corel 1 K, Corel 5 K, and WdcImageData, respectively

λ_2	Caltech-101		λ_2	Corel 1 K		λ_2	Corel 5 K		λ_2	WdcImageData	
	RR	PR		RR	PR		RR	PR		RR	PR
10^{-5}	0.3463	0.3333	10^{-5}	0.4340	0.8720	10^{-5}	0.2886	0.7316	10^{-5}	0.3810	0.6048
10^{-3*}	0.3793	0.3797	10^{-3}	0.4739	0.9457	10^{-3}	0.3146	0.7996	10^{-3}	0.3439	0.5581
10^{-2}	0.3179	0.3174	10^{-2}	0.4341	0.8689	10^{-2}	0.3278	0.8327	10^{-2}	0.3842	0.6237
10^{-1}	0.3580	0.3662	10^{-1}	0.4812	0.9622	10^{-1}	0.3283	0.8346	10^{-1}	0.3938	0.6584
1	0.2921	0.2884	1	0.4460	0.8963	1	0.3177	0.8051	1	0.3676	0.6351
10	0.1226	0.1130	10	0.1576	0.2994	10	0.2704	0.6838	10	0.1760	0.2841

*Bold values indicate the best result of the column

Table 6 Comparison retrieval response in the light of λ_3 by fixing retrieval length ($\ell = 0.1$) for all datasets, $\lambda_1 = 0.1, 0.01, 0.1, 0.01, \lambda_2 = 0.001, 0.1, 0.1, 0.1$, and $D = 8, 16, 64, 32$ for Caltech-101, Corel 1 K, Corel 5 K, and WdcImageData, respectively

λ_3	Caltech-101		λ_3	Corel 1 K		λ_3	Corel 5 K		λ_3	WdcImageData	
	RR	PR		RR	PR		RR	PR		RR	PR
10^{-5}	0.3233	0.3145	10^{-5}	0.4472	0.9032	10^{-5}	0.3115	0.7932	10^{-5}	0.3504	0.5631
10^{-3}	0.3087	0.3116	10^{-3}	0.4589	0.9189	10^{-3}	0.3012	0.7656	10^{-3}	0.3662	0.6275
10^{-2}	0.3154	0.3130	10^{-2}	0.4879	0.9750	10^{-2}	0.3125	0.7960	10^{-2}	0.3709	0.6035
10^{-1} *	0.3676	0.3693	10^{-1}	0.4865	0.9665	10^{-1}	0.3166	0.8270	10^{-1}	0.3914	0.6427
1	0.3541	0.3594	1	0.3575	0.7232	1	0.3137	0.7996	1	0.3765	0.6237
10	0.2665	0.2696	10	0.4582	0.9177	10	0.3104	0.7904	10	0.3378	0.5429

*Bold values indicate the best result of the column

related to multiview NMF and a statistics-based model, such as MAH [8], JNMF [23], KL-GMM [9], and MNFCM. MNFCM is similar to Liu’s JNMF, the different is that the FCM is employed to cluster the data of NMF’s coefficient matrix. For the sake of randomizing experiments, in this study, multiple runs, each with a different query image, are performed and the average values of PR and RR are recorded. The run times should be decided by the number of query images in the test set. Fig. 2 summarizes the PR and RR versus the retrieval length achieved by each examined algorithm. From this figure, it can be seen that the MAH and SCMN algorithms achieve an obviously better response across all retrieval exams. This is attributed to the involvement of multiview features, which are crucial for an image retrieval system. Considering the multiple constraints of the objective function, the SCMN achieves the highest precision. In contrast, KL-GMM and MNFCM achieve inferior responses in terms of PR. Additionally, Fig. 2 presents the RR of all methods at varying retrieval lengths, showing that a multiview scheme usually achieves better results than a single-view framework. Due to the utilization of the SMM-MRF scheme, the SCMN is successful in clustering sparse features so as to return more relevant images. This is also demonstrated by the plots of RR. It should be noted that for the Caltech101 dataset, all algorithms present relatively poor behaviour. This is reasonable, and one reason for this is that images in this dataset have a complex construction and are rich in information in most cases. It is thus relatively difficult to represent them. Another reason is that for the retrieval experiment on Caltech101, we have selected more categories which are challenging disadvantages in most retrieval schemes. Despite this, the SCMN still shows an encouraging retrieval response. Table 7 shows the top-returned results indexed from the respective categories corresponding to the query images.

5.5 Convergence study

Convergence determines the merits of the algorithm as well as its execution time. Since the proposed SCMN is an iteration strategy, we need to discuss the behaviour of the update rules to minimize the objective function and to compare it with the classical NMF.

Fig. 2 PP and RR of the obtained results using different approaches. The first to the last row represent the Caltech-101, Corel 1 K, Corel 5 K, and WdcImageData datasets, respectively. The left column is Precision, and the right one is the Recall Rate

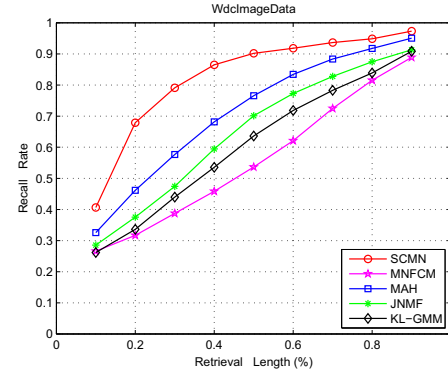
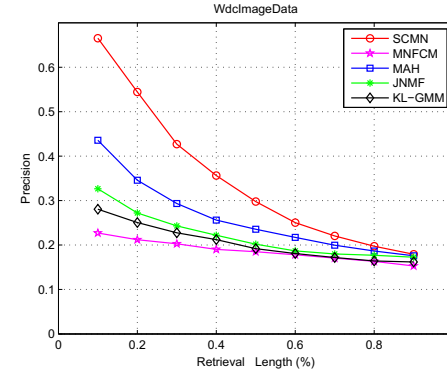
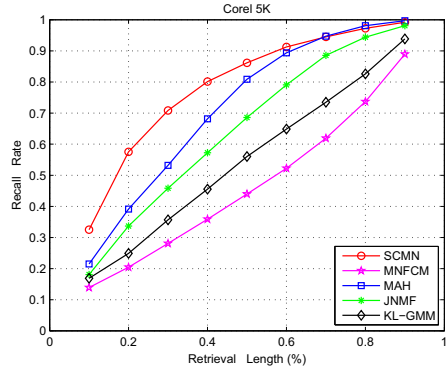
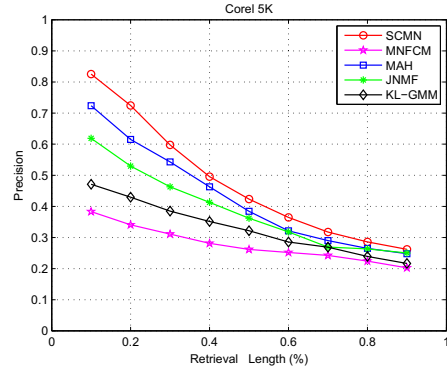
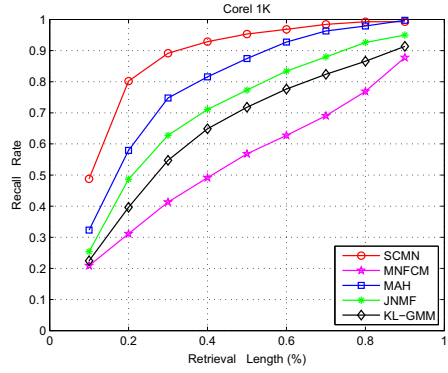
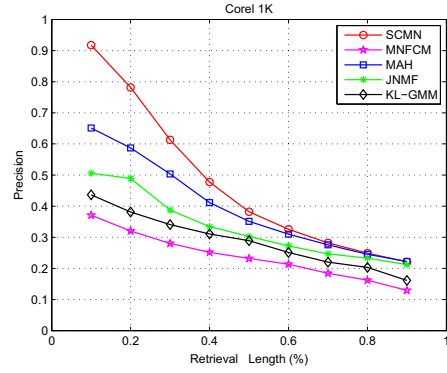
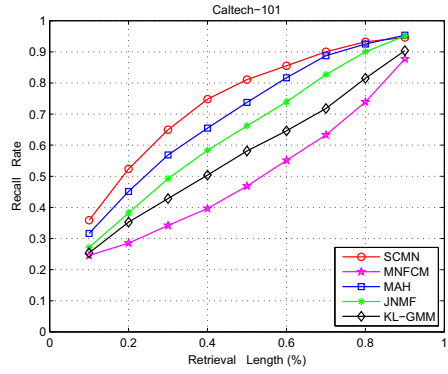
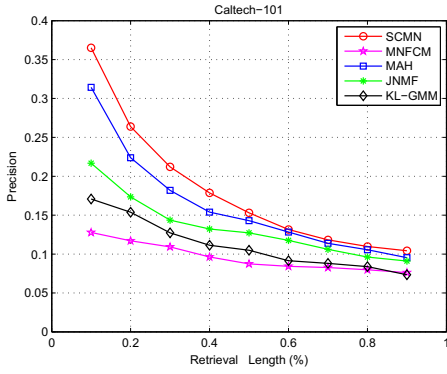


Table 7 Query images (at the first column) and the first top retrieval results obtained with different methods









































































































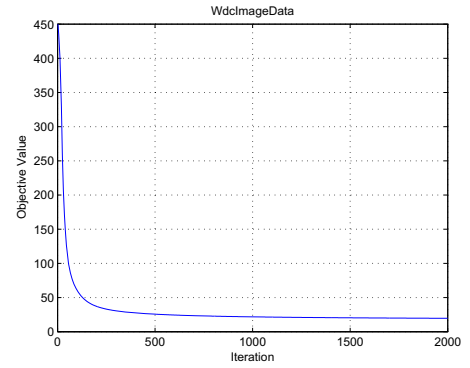
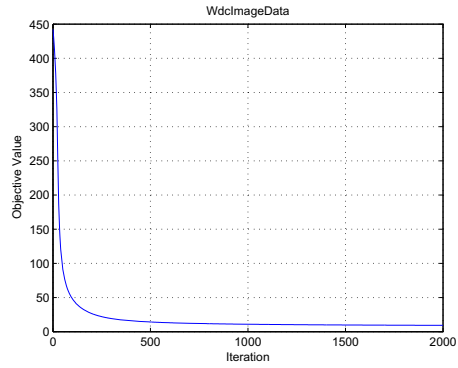
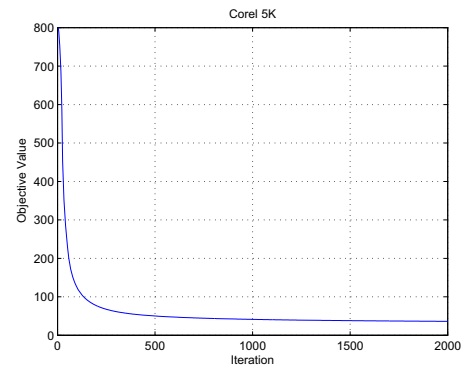
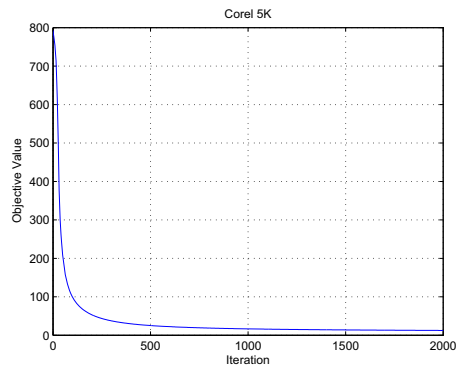
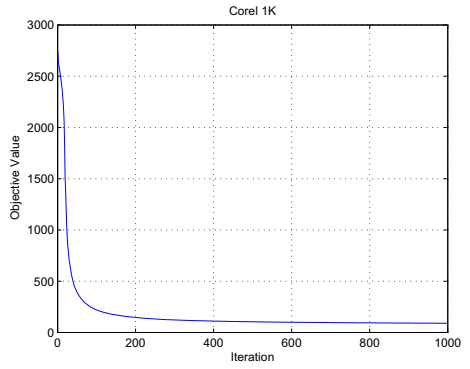
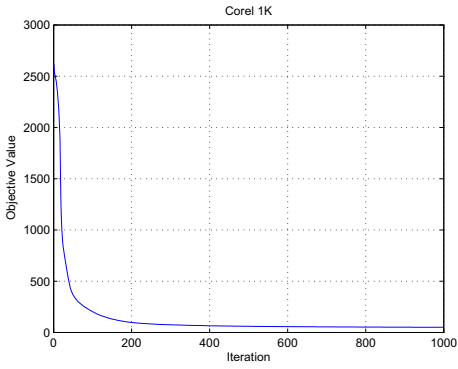
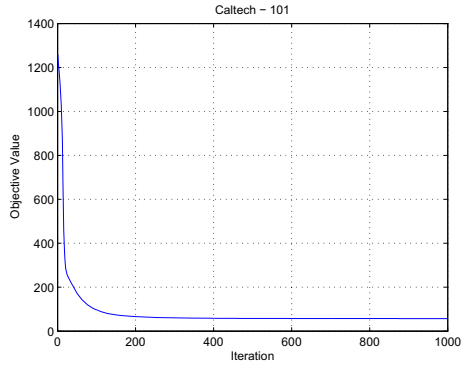
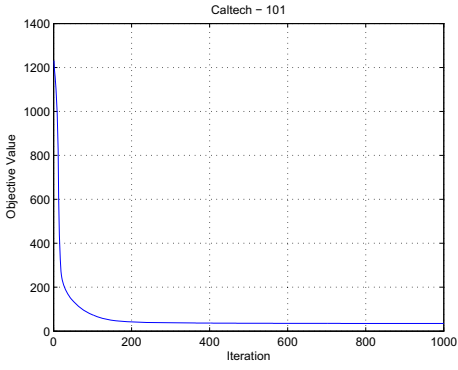
	KL-GMM		JNMF		MAH		MNFCM		SCMN	
										
										
										
										
										
										
										
										
										
										

Fig. 3 illustrates the variations of the objective function in the implementation process. It can be observed that the solution is very close to the local minimum after 200 iterations for the Caltech-101 and Corel 1 K datasets. For the other two datasets, the objective function converges with more updates and stabilizes at approximately 250 iterations. It is worth noting that the convergence level of the SCMN is nearly as fast as the classical NMF in all cases.

5.6 Time comparison

In the last experiment, we empirically compare the time needed for implementing the aforementioned retrieval tasks. By running all algorithms twenty times with a different query each time, we obtain the comparative results illustrated in Fig. 4. For the proposed SCMN, the learning phase would take a relatively long time. We believe this is because of the feature-extraction step of our method, which ultimately results in the increase of the computational time occupied in the learning phase. For the task of retrieving identical images, in contrast, the KL-GMM is time-saving, for training as well as retrieval phases. Additionally, as expected, the feature-extraction step is also required according to the

Fig. 3 Objective function value versus the number of iterations. The first to the last row represent Caltech-101, Corel 1 K, Corel 5 K, and WdeImageData, respectively; the left column is the NMF method, and the right one is the proposed SCMN



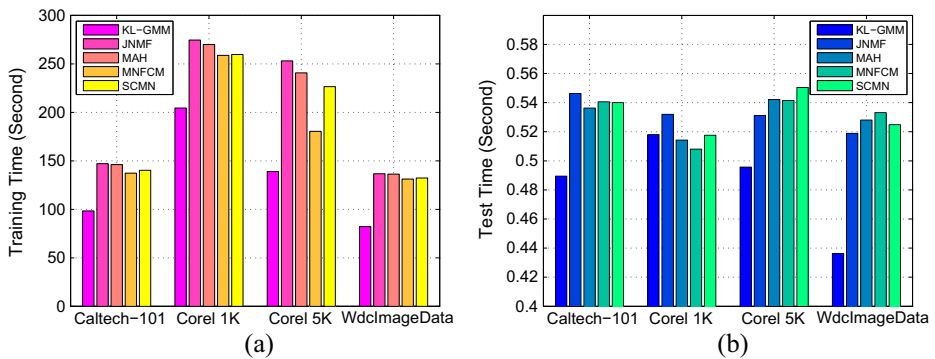


Fig. 4 Comparison of average running time for different retrieval systems. (a) Training time; (b) test time

mechanisms of JNMF, MAH and MNFCM. Additionally, considerable execution time required for the JNMF, MAH, and MNFCM methods in this figure, which seems to confirm this conclusion. For the average CPU time in the retrieval phase, no significant difference could be observed for all participants. Thus, the proposed SCMN achieves an acceptable time complexity as the baseline method.

5.7 Evaluation of SCMN on clinical MR images

Finally, to evaluate the reliability of the SCMN not only for the four public databases, but also for other types of images, we provided an additional experiment to assess the performance of the proposed method on magnetic resonance (MR) images in terms of average retrieval accuracy. One public dataset used in current experiment consists of 900 standard clinical MR images, taken from SPM12 website.⁴ For the purpose of illustration, Appendix provides a pseudocode for our image retrieval algorithm. Generally, model parameters should be specified by users primarily based on experiment. This paper provides a statistical method to select a proper parameter set for SCMN by using 10-fold cross validation (CV) [36]. The capability of CV to perform estimation or evaluation enables CV to conduct our model parameter selection. In [14], it was once reported to determine the number of neighbors of classification. In 10-fold CV, a labeled dataset S (900 standard clinical MR images) is partitioned into 10 equally sized subsets. The proposed method has four parameters: the inner dimension \bar{D} , three regularization factors ($\lambda_1, \lambda_2, \lambda_3$). For simplicity, we discuss the impact of two important parameters \bar{D} and λ_2 on the performance of our method. Two other parameters λ_1, λ_3 are set a fixed positive value 0.1 and iteratively changed parameters \bar{D} and λ_2 to reach a better chosen. In our experiment, the following values for inner dimension \bar{D} are considered: 16, 32, 64 and 96. Another regularization parameter λ_2 varies from 10^{-3} to 1 increased by 10 times, forming a set of parameter $M = \{M_1, M_2, \dots, M_{16}\}$ for the proposed model SCMN. For every selection M_i , an iterative process is then conducted. In each iteration, one different subset is selected as a test set, and the remaining nine subsets are the training data. The retrieval Precision and Recall Rate are obtained by the average of the accuracies of these

⁴ <http://www.fil.ion.ucl.ac.uk/spm>

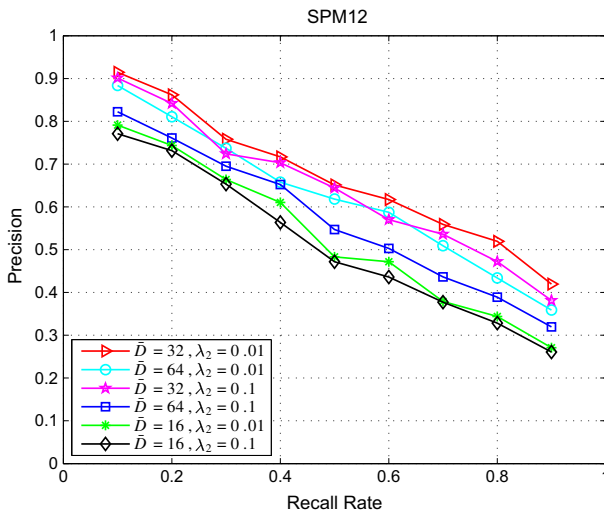


Fig. 5 Precision vs. Recall Rate for the different parameters, using 10-fold cross validation

10 classifiers. Pick the model M_i with the best image retrieval results. We evaluated the performance with different values for \bar{D} and λ_2 , and some performance curves are illustrated in Fig.5. This figure tell us that the best image retrieval performance is obtained with $\bar{D} = 32, \lambda_2 = 0.01$. Also, the values $\bar{D} = 32, \lambda_2 = 0.01, \lambda_1 = 0.1, \lambda_3 = 0.1$ are selected as the best parameter combination for SCMN. In the following experiment, the proposed SCMN is tested on SPM12 dataset for validation. The training set is having three different categories with 300 images per category, including sagittal plane (188×68), coronal plane (156×68), and horizontal plane (188×156). For each category, 10 images are randomly sampled as query images. In total, there are 30 query images. Table 8 provides some sample retrieval results for SPM12 dataset. Precision and Recall for the presented SCMN and other algorithms are calculated and demonstrated through graphs. Fig.6 shows variation in Precision and Recall Rate with number of

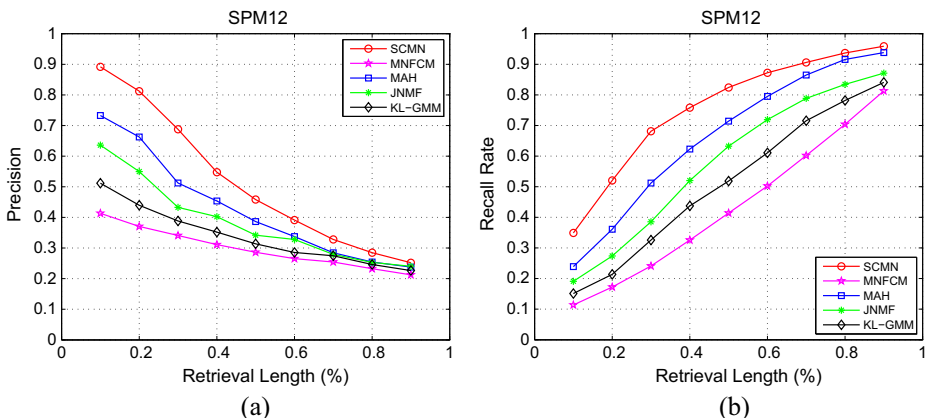

































Fig. 6 Comparison of retrieval performance for SPM12 dataset. (a) Precision; (b) Recall Rate

Table 8 Query images (at the first column) and the first top retrieval results obtained with different methods

	KL-GMM		JNMF		MAH		MNFCM		SCMN	
										
										
										

images retrieved. The results mostly confirmed our theoretical expectation. The SCMN method yields more accurate performance than all other methods in comparison.

6 Conclusions

This article addressed a CBIR scheme by merging the constrained NMF with multiple visual features and a MRF-based SMM approach. The following main advantages were revealed: (1) The algorithms embedded some additional constraints, such as local geometric structure and spacing location, to develop a novel objective function. This resulted in a relatively better performance according to the measures of PR and RR, obtained for the image retrieval task, while maintaining an acceptable computational cost. (2) Using similarity metrics based on the Frobenius norm, the proposed method fused multiple distinct features so that the coefficient matrix of the SCMN preserved the features of the underlying images in the low-dimensional space and ensured that the proposed framework produced result images as relevant as possible to the test image. (3) According to the Bayesian theorem, the study successfully incorporated MRF into SMM, which contributed to alleviating the disturbance of noise in the training process, thus improving the robustness of the algorithm. Optimization was performed using an EM algorithm to estimate the parameters of SMM-MRF. (4) Finally, the rule of convergence of updates was proved theoretically, and the complexity of the algorithm was also discussed. Overall, the experimental results indicate that the SCMN is stable for the test images chosen and exhibits a better retrieval response than other competing algorithms.

Future extensions of this work may aim to bridge the semantic gap between low level features and user preferences, and to investigate the possibility that the mixed visual features involve semantic information. We are likely to develop other effective techniques to bridge the semantic gap.

Acknowledgements The authors would like to thank the anonymous reviewers and the associate editor for their insightful comments that significantly improved the quality of this paper. This work was supported by the National Nature Science Foundation of China under Grant 61371150.

Appendix

In this appendix, we provide the implementation details of each part shown in Fig. 1.

```

/* Learning Phase */
/* First step: Feature extraction & fusion */
for i = 1 : NImages_Train_InFile
    Image = imread(Path_Image_Train);
    im_Gist = LMgist (Image)
    im_HOG = HOG_feature (Image);
    im_LBP = LBP_feature (Image);
    im_ColorHist = ColorHist_feature (Image);
    im_Zernike = Zernike_feature (Image);
end
for i = 1 : NTrainImages
    Ki ( 1 ) = exp ( - ( norm ( im_Gist ( i ) - im_GIST ) ) ^ 2 / ( 2 × λ ( 1 ) ^ 2 ) );
    Ki ( 2 ) = exp ( - ( norm ( im_HOG ( i ) - im_HOG ) ) ^ 2 / ( 2 × λ ( 2 ) ^ 2 ) );
    Ki ( 3 ) = exp ( - ( norm ( im_LBP ( i ) - im_LBP ) ) ^ 2 / ( 2 × λ ( 3 ) ^ 2 ) );
    Ki ( 4 ) = exp ( - ( norm ( im_ColorHist ( i ) - im_ColorHist ) ) ^ 2 / ( 2 × λ ( 4 ) ^ 2 ) );
    Ki ( 5 ) = exp ( - ( norm ( im_Zerniket ( i ) - im_Zernike ) ) ^ 2 / ( 2 × λ ( 5 ) ^ 2 ) );
end
Wi ( 1 ) = ( Ki ( 1 ) - I ) / ( sum ( sum ( Ki ( 1 ) ) ) )
    :
Wi ( 5 ) = ( Ki ( 5 ) - I ) / ( sum ( sum ( Ki ( 5 ) ) ) )
X = [Ki ( 1 ) + Ki ( 2 ) + Ki ( 3 ) + Ki ( 4 ) + Ki ( 5 ) ] / 5
W1 = [Wi ( 1 ) + Wi ( 2 ) + Wi ( 3 ) + Wi ( 4 ) + Wi ( 5 ) ] / 5
L1 = D1 - W1

/* Second step: The proposed NMF, initialize inner dimension D, λ1, λ2, λ3, β */
U = rand ( NTrainImages, D ); V = rand ( D, NTrainImages );
while ( 1 )
    U = U × ( ( X × V' ) / ( U × V × V' + 0.5 × λ1 × U(-1/2) ) );
    V = V × ( ( U' × X + λ2 × V × W1 + 2 × λ3 × β × exp ( -beta × trace ( V × L2 × V' ) ) × ( V × D2 ) ) / ( U' × U
        × V + λ2 × V × D1 + 2 × λ3 × β × exp ( -beta × trace ( V × L2 × V' ) ) × ( V × W2 ) ) );
    if norm ( U × V - X ) < NMF_Threshold
        break;
    end
end

/* Third step: SMM-MRF Clustering for coefficient matrix V, initialize sigma, π, μ */
[π, μ, sigma, Z_ij_ij ] = SMM_MRF_MultDimData ( V, π, μ, sigma, NComponents )
for i = 1 : NTrainImages // NComponents is number of label
    ImageLabel_SMMComponent ( i ) = find ( Z_ij_ij ( i ) == max ( Z_ij_ij ( i ) ) )
    Image_probability ( i ) = Z_ij_ij ( i, ImageLabel_SMMComponent ( i ) )
end

```

```

/* Fourth step: Similarity measurement */
for TestImage = 1 to NImages_Test_InFile (TestFile)
    Image_test = imread (TestImage)
    im_Gist_test = LMgist (Image_test)
    im_HOG_test = HOG_feature (Image_test)
    im_LBP_test = LBP_feature (Image_test)
    im_ColorHist_test = ColorHist_feature (Image_test)
    im_Zernike_test = Zernike_feature (Image_test)
end
for i = 1 to NTrainImages
    Ki_Test (1) = exp ( - ( norm ( im_Gist_Test - im_GIST ( i ) ) )2 / ( 2 × λ ( 1 )2 ) )
    Ki_Test (2) = exp ( - ( norm ( im_HOG_Test - im_HOG ( i ) ) )2 / ( 2 × λ ( 2 )2 ) )
    Ki_Test (3) = exp ( - ( norm ( im_LBP_Test - im_LBP ( i ) ) )2 / ( 2 × λ ( 3 )2 ) )
    Ki_Test (4) = exp ( - ( norm ( im_ColorHist_Test - im_ColorHist ( i ) ) )2 / ( 2 × λ ( 4 )2 ) )
    Ki_Test (5) = exp ( - ( norm ( im_Zerniket_Test - im_Zernike ( i ) ) )2 / ( 2 × λ ( 5 )2 ) )
end

/* Fifth step: Sparse query step */
X_query = (Ki_Test (1) + Ki_Test (2) + Ki_Test (3) + Ki_Test (4) + Ki_Test (5) ) / 5
P = pinv ( U' × U ) × U'
V_query = P × X_query

/* Sixth step: Probability-based retrieval and similarity ranking */
for k = 1 to NComponents_SMM // NComponents_SMM = 3 in SPM12 dataset experiment
    x_data = V_query - μ ( :, k )
    SMM_Test ( 1, k ) = gamma ( ( v + D ) / 2 ) × sigma-1/2 / ( ( π × v )D/2 × gamma ( v / 2 ) × ( 1 + ( x_data - μ )2 / ( sigma × v ) )( v + D ) / 2 )
end
SMM_TestP = sortrows ([ 1 : NComponents_SMM; SMM_Test ], -2)
for k = 1 to NComponents_SMM
    ComponentLabel_TestImage = SMM_TestP ( k, 1 );
    [aa_Order_SimilarImagesInk] = find ( ImageLabel_SMMComponent == ComponentLabel_TestImage )
    temp = [Order_SimilarImagesInk; Image_probability ( Order_SimilarImagesInk )];
    temp = sortrows ( temp', -2 )
    Order_SimilarImages = [Order_SimilarImages, temp ( :, 1 )']
end
NSimilar = fix ( 0.1 × NTrainImages ); // select the top N × ℓ images
Similar_FilePosition = Position_Images ( 1, Order_SimilarImages ( 1, 1 : NSimilar ) )
Similar_ImagePosition = Position_Images ( 2, Order_SimilarImages ( 1, 1 : NSimilar ) )

```

References

1. Ahonen T, Hadid A, Pietikäinen M (2004) Face recognition with local binary patterns. *Lect Notes Comput Sci* 3021:469–481
2. Amin T, Zeytinoglu M, Guan L (2007) Application of Laplacian mixture model to image and video retrieval. *IEEE Trans Multimedia* 9(7):1416–1429
3. An L, Zou CJ, Zhang LY, Denney B (2016) Scalable attribute-driven face image retrieval. *Neurocomputing* 172:215–224
4. Babae M, Bahmanyar R, Rigoll G, Datcu M (2014) Farness preserving non-negative matrix factorization. In: *ICIP'14: International Conference on Image Processing* 3023–3027
5. Babae M, Tsoukalas S, Babae M, Rigoll R, Datcu M (2016) Discriminative nonnegative matrix factorization for dimensionality reduction. *Neurocomputing* 173:212–223
6. Boyd SP, Vandenberghe L (2004) *Convex optimization*. Cambridge University Press, United Kingdom
7. Cai D, He X, Han J, Huang TS (2011) Graph regularized non-negative matrix factorization for data representation. *IEEE Trans Pattern Anal Mach Intell* 33(8):1548–1560
8. Cox TE, Cox MA (2010) *Multidimensional scaling*. CRC Press, United States
9. Cui S, Datcu M (2015) Comparison of Kullback-Leibler divergence approximation methods between Gaussian mixture models for satellite image retrieval. *IEEE Geoscience and Remote Sensing Symposium* 3719–3722
10. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: *CVPR'05: IEEE Conference on Computer Vision and Pattern Recognition, San Diego 1* (12): 886–893
11. Deselaers T, Keysers D, Ney H (2008) Features for image retrieval: an experimental comparison. *Inf Retr* 11(2):77–107
12. Feng L, Bhanu B (2016) Semantic concept co-occurrence patterns for image annotation and retrieval. *IEEE Trans Pattern Anal Mach Intell* 38(4):785–799
13. Flusser J, Zitova B, Suk T (2009) *Moments and moment invariants in pattern recognition*. Wiley, New York
14. Gertheiss J, Tutz G (2009) Feature selection and weighting by nearest neighbor ensembles. *Chemom Intell Lab Syst* 99(2):30–38
15. Gillis N, Kuang D, Park H (2015) Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization. *IEEE Trans Geosci Remote Sens* 53(4):2066–2078
16. Greenspan H, Pinhas AT (2007) Medical image categorization and retrieval for PACS using the GMM-KL framework. *IEEE Trans Info Technol Biomed* 11(2):190–202
17. Han J, Ma KK (2002) Fuzzy color histogram and its use in color image retrieval. *IEEE Trans Image Process* 11(8):944–952
18. Hyvärinen A (2001) Independent component analysis. *Neural Comput Sur* 4:60–83
19. Kim H, Park H (2008) Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM J Matrix Anal Appl* 30(2):713–730
20. Klema VC, Laub AJ (1980) The singular value decomposition: Its computation and some applications. *IEEE Trans Autom Control* 25(2):164–176
21. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401: 788–791
22. Liu H, Wu Z, Cai D, Huang TS (2012) Constrained nonnegative matrix factorization for image representation. *IEEE Trans Softw Eng* 34(7):1299–1311
23. Liu J, Wang C, Gao J, Han J (2013) Multi-view clustering via joint nonnegative matrix factorization. In: *SDM'13: Proceeding of the 2013 SIAM International Conference on Data Mining* 252–260
24. Liu L, Yu M, Shao L (2015) Multiview alignment hashing for efficient image search. *IEEE Trans Image Process* 24(3):956–966
25. Lowe DG (2004) Distinctive image features from scale invariant key points. *Int J Comput Vis* 60(2):91–110
26. Marakakis A, Galatsanos N, Likas A, Stafylopatis A (2009) Probabilistic relevance feedback approach for content-based image retrieval based on Gaussian mixture models. *IET Image Process* 3(1):10–25
27. Mittal A, Sofat S (2013) A novel color coherence vector based obstacle detection algorithm for textured environments. *Int J Comput Theory Eng* 5(1):81–84
28. Nguyen TM, Jonathan Wu QM (2013) Fast and robust spatially constrained Gaussian mixture model for image segmentation. *IEEE Trans Circuits Syst Video Technol* 23(4):621–635
29. Nguyen TM, Jonathan Wu QM (2014) Bounded asymmetrical Student's-*t* mixture model. *IEEE Trans Cybern* 44(6):857–869
30. Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int J Comput Vis* 42(3):145–175
31. Peel D, McLachlan G (2000) Robust mixture modeling using the *t*-distribution. *Stat Comput* 10:335–344

32. Piatek ML, Smolka B (2013) Color image retrieval based on spatio-chromatic multichannel Gaussian mixture modelling. In: ISPA'13: 8th International Symposium on Image and Signal Processing and Analysis 130–135
33. Qi SY, Luo YP (2016) Object retrieval with image graph traversal-based re-ranking. *Signal Process Image Commun* 41:101–114
34. Qian Y, Jia S, Zhou J, Robles-Kelly A (2011) Hyperspectral unmixing via L1/2 sparsity-constrained nonnegative matrix factorization. *IEEE Trans Geosci Remote Sens* 49(11):4282–4297
35. Rajabi R, Ghassemian H (2015) Spectral unmixing of hyperspectra imagery using multilayer NMF. *IEEE Geosci Remote Sens Lett* 12(1):38–42
36. Shunfeng C, Michael P (2012) Using cross-validation for model parameter selection of sequential probability ratio test. *Expert Syst Appl* 39:8467–8473
37. Wang Z, Feng Y, Qi T, Yang X, Zhang JJ (2016) Adaptive multi-view feature selection for human motion retrieval. *Signal Process* 120:691–701
38. Wang W, Qian Y, Tang YY (2016) Hypergraph-regularized sparse NMF for hyperspectral unmixing. *IEEE J Sel Top Appl Earth Obs Remote Sens* 9(2):681–694
39. Xia T, Tao D, Mei T, Zhang YD (2010) Multiview spectral embedding. *IEEE Trans Syst Man Cybern B Cybern* 40(6):1438–1446
40. Xu Z, Chang X, Xu F, Zhang H (2012) L1/2 regularization: A thresholding representation theory and a fast solver. *IEEE Trans Neural Netw Learn Syst* 23(7):1013–1027
41. Yang WH, Liu GQ, Zhang L, Chen EH (2012) Multi-view learning with batch mode active selection for image retrieval. In: ICPR'12: 21st International Conference on Pattern Recognition 979–982
42. Yang SY, Zhang XT, Yao YG, Cheng SQ, Jiao LC (2015) Geometric nonnegative matrix factorization (GNMF) for hyperspectral unmixing. *IEEE J Sel Top Appl Earth Obs Remote Sens* 8(6):2696–2703
43. Yeung KY, Ruzzo WL (2001) Principal component analysis for clustering gene expression data. *Bioinformatics* 17(9):763–774
44. Zeng S, Huang R, Wang HB, Kang Z (2016) Image retrieval using spatiograms of colors quantized by Gaussian mixture models. *Neurocomputing* 171:673–684
45. Zhu HQ, Liu M, Ji H, Li Y (2010) Combined invariants to blur and rotation using Zernike moment descriptors. *Pattern Anal Applic* 13:309–319



Hongqing Zhu obtained her Ph.D. in 2000 from Shanghai Jiao Tong University. From 2003 to 2005, she was a postdoctoral fellow in the Department of Biology and Medical Engineering at Southeast University. Currently, she is a professor at East China University of Science and Technology. Her current research interests include machine learning, deep learning, big data, image reconstruction, image segmentation, image compression, and artificial intelligence.



Qunyi Xie is currently pursuing the Ph.D. degree with the Department of Electronics and Communication Engineering, East China University of Science and Technology, Shanghai, China, where he received the B.S. degree from the School of Information Science and Engineering, in 2014. His research interests are deep learning, big data, machine learning, artificial intelligence.