Multiview Spectral Embedding

Tian Xia, Dacheng Tao, *Member, IEEE*, Tao Mei, *Member, IEEE*, and Yongdong Zhang, *Member, IEEE*

Abstract—In computer vision and multimedia search, it is common to use multiple features from different views to represent an object. For example, to well characterize a natural scene image, it is essential to find a set of visual features to represent its color, texture, and shape information and encode each feature into a vector. Therefore, we have a set of vectors in different spaces to represent the image. Conventional spectral-embedding algorithms cannot deal with such datum directly, so we have to concatenate these vectors together as a new vector. This concatenation is not physically meaningful because each feature has a specific statistical property. Therefore, we develop a new spectral-embedding algorithm, namely, multiview spectral embedding (MSE), which can encode different features in different ways, to achieve a physically meaningful embedding. In particular, MSE finds a low-dimensional embedding wherein the distribution of each view is sufficiently smooth, and MSE explores the complementary property of different views. Because there is no closed-form solution for MSE, we derive an alternating optimization-based iterative algorithm to obtain the low-dimensional embedding. Empirical evaluations based on the applications of image retrieval, video annotation, and document clustering demonstrate the effectiveness of the proposed approach.

Index Terms—Dimensionality reduction, multiple views, spectral embedding.

I. Introduction

IN COMPUTER vision and multimedia search [5], [6], objects are usually represented in several different ways. This kind of data is termed as the multiview data. A typical example is a color image, which has different views from different modalities, e.g., color, texture, and shape. Different views form different feature spaces, which have particular statistical properties.

Because of the popularity of multiview data in practical applications, particularly in the multimedia domain, learning from multiview data, which is also known as multiple-view learning, has attracted more and more attentions. Although a great deal of efforts have been carried out on multiview data learning [1], including classification [21], clustering [4], [19], and feature selection [20], little progress has been made in dimensionality reduction, whereas it has many applications in multimedia [28], e.g., image retrieval and video annotation. Multimedia data generally have multiple modalities, and each modality is usually represented in a high-dimensional feature space which frequently leads to the "curse of dimensionality" problem. In this case, multiview dimensionality reduction provides an effective solution to solve or at least reduce this problem.

In this paper, we consider the problem of spectral embedding for multiple-view data based on our previous patch alignment framework [29]. The major challenge is learning a low-dimensional embedding to effectively explore the complementary nature of multiple views of a data set. The learned low-dimensional embedding should be better than a low-dimensional embedding learned by each single view of the data set.

Existing spectral-embedding algorithms assume that samples are drawn from a vector space and thus cannot deal with multiview data directly. A possible solution is to concatenate vectors from different views together as a new vector and then apply spectral-embedding algorithms directly on the concatenated vector. However, this concatenation is not physically meaningful because each view has a specific statistical property. This concatenation ignores the diversity of multiple views and thus cannot efficiently explore the complementary nature of different views. Another solution is the distributed spectral embedding (DSE) proposed in [3]. DSE performs a spectral-embedding algorithm on each view independently, and then based on the learned low-dimensional representations, it learns a common low-dimensional embedding which is "close" to each representation as much as possible. Although DSE allows selecting different spectral-embedding algorithms for different views, the original multiple-view data are invisible to the final learning process, and thus, it cannot well explore the complementary nature of different views. Moreover, its computational cost is dense because it conducts spectral-embedding algorithms for each view independently.

To effectively and efficiently learn the complementary nature of different views, we propose a new algorithm, i.e., multiview spectral embedding (MSE), which learns a low-dimensional and sufficiently smooth embedding over all views simultaneously. Empirical evaluations based on image retrieval, video

annotation, and document clustering show the effectiveness of the proposed approach.

The rest of this paper is organized as follows. In Section II, we provide a short review on related works. In Section III, we present the proposed MSE and the solution of MSE. Experimental results are shown in Section IV, and Section V concludes.

II. RELATED WORKS

In this section, first, we provide a short review on conventional spectral-embedding algorithms, which are all for single-view data. Although there are some previous works on multiview spectral methods, they are about multiview learning for clustering [4] and classification [21] but not for dimensionality reduction. As for spectral embedding for multiple-view data, only some preliminary effort [3] is known to us; thus, second, we give a brief introduction to the distributed method proposed in [3].

A. Spectral Embedding

The task of dimensionality reduction is to find a low-dimensional representation for high-dimensional observations. It generally falls into two classes: linear methods, e.g., principle component analysis and multidimensional scaling; and nonlinear methods, e.g., locally linear embedding (LLE) [8] and Laplacian eigenmaps (LE) [9].

Spectral methods for dimensionality reduction find the low-dimensional representations by using eigenvectors of specially constructed matrices [29]. Since, in the traditional problem setting of dimensionality reduction, it is assumed that the data are represented in a single vector space, the conventional spectral-embedding algorithms can all be regarded as methods with a single view.

Existing algorithms can be classified into two groups based on whether they are supervised or unsupervised. The focus of this paper is the latter. Representative algorithms include Isomap [7], LLE [8], LE [9], Hessian eigenmaps [10], local tangent space alignment [11], transductive component analysis [2], discriminative locality alignment [30], and DLLE [27]. They perform well for single-view data but cannot deal with multiview data directly.

B. Distributed Approach for Spectral Embedding With Multiple Views

As mentioned earlier, spectral embedding with multiple views is a new topic; it is first proposed in [3], and a distributed approach, i.e., DSE, is proposed in it. In the following is a brief summary of DSE.

Given a multiple-view datum with n objects having m views, i.e., a set of matrices $X = \{X^{(i)} \in \Re^{m_i \times n}\}_{i=1}^m$, each representation $X^{(i)}$ is a feature matrix from view i. DSE assumes that the low-dimensional embedding of each view $X^{(i)}$ is already known, i.e., $A = \{A^{(i)} \in \Re^{n \times k_i}\}_{i=1}^m$, $k_i < m_i$ ($1 \leq i \leq m$). DSE focuses on how to learn a consensus low-dimensional

embedding $B \in \Re^{n \times k}$ based on A; the objective function of DSE is defined as

$$\min_{B,P} \sum_{i=1}^{m} \left\| A^{(i)} - BP^{(i)} \right\|^2 \quad \text{s.t. } B^{\mathrm{T}}B = I \tag{1}$$

where $P = \{P^{(i)} \in \Re^{k \times k_i}\}_{i=1}^m$ is a set of mapping matrices. The global optimal solution to DSE is given by performing eigendecomposition of the matrix CC^{T}, $C = [A^{(1)}, \ldots, A^{(m)}]$.

III. MSE

In this section, we introduce a new spectral-embedding algorithm, i.e., MSE, which finds a low-dimensional and sufficiently smooth embedding over all views simultaneously. To better present the technique details of the proposed MSE, we provide important notations used in the rest of this paper. Capital letters, e.g., X, represent matrices or sets, and $[X]_{ij}$ is the (i,j)th entry of X. Lower case letters, e.g., x, represent vectors, and $(x)_i$ is the ith element of x. Superscript (i), e.g., $X^{(i)}$ and $x^{(i)}$, represents data from the ith view.

Based on the aforementioned notations, MSE can be described as follows according to our previous patch alignment framework [29]. Given a multiview data set with n objects and m representations, i.e., a set of matrices $X = \{X^{(i)} \in \Re^{m_i \times n}\}_{i=1}^m$, wherein $X^{(i)}$ is the feature matrix for the ith view representation, MSE finds a low-dimensional and sufficiently smooth embedding of X, i.e., $Y \in \Re^{d \times n}$, wherein $d < m_i$ ($1 \leq i \leq m$) and d is a predefined number according to different applications.

Fig. 1 shows the working principle of MSE. MSE first builds a patch for a sample on a view. Based on the patches from different views, the part optimization can be performed to get the optimal low-dimensional embedding for each view. Afterward, all low-dimensional embeddings from different patches are unified as a whole one by global coordinate alignment. Finally, the solution of MSE is derived by using the alternating optimization.

A. Part Optimization

Given the ith view $X^{(i)} = [x_1^{(i)}, \ldots, x_n^{(i)}] \in \Re^{m_i \times n}$, consider an arbitrary point $x_j^{(i)}$ and its k related ones in the same view (e.g., nearest neighbors) $x_{j_1}^{(i)}, \ldots, x_{j_k}^{(i)}$; the patch of $x_j^{(i)}$ is defined as $X_j^{(i)} = [x_j^{(i)}, x_{j_1}^{(i)}, \ldots, x_{j_k}^{(i)}] \in \Re^{m_i \times (k+1)}$. For $X_j^{(i)}$, there is a part mapping $f_j^{(i)} : X_j^{(i)} \rightarrow Y_j^{(i)}$, wherein $Y_j^{(i)} = [y_j^{(i)}, y_{j_1}^{(i)}, \ldots, y_{j_k}^{(i)}] \in \Re^{d \times (k+1)}$. To preserve the locality in the projected low-dimensional space, the part optimization for the jth patch on the ith view is

$$\arg\min_{Y_j^{(i)}} \sum_{l=1}^{k} \left\| y_j^{(i)} - y_{j_l}^{(i)} \right\|^2 \left(w_j^{(i)} \right)_l \tag{2}$$
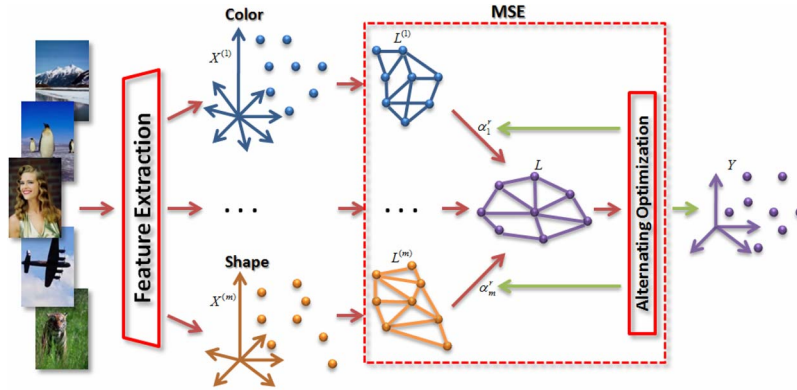
Fig. 1. Working flow of MSE: MSE first builds a patch for a sample on a view. Based on the patches from different views, the part optimization can be performed to get the optimal low-dimensional embedding for each view. Then, all low-dimensional embeddings from different patches are unified together as a whole one by global coordinate alignment. Finally, the solution of MSE is derived by using the alternating optimization.

where $w_j^{(i)}$ is a k-dimensional column vector weighted by $(w_j^{(i)})_l = \exp(-\|x_j^{(i)} - x_{j_l}^{(i)}\|^2/t)$. Therefore, (2) can be reformulated to

$$
\arg\min_{Y_j^{(i)}} \mathrm{tr}\left(\left[\begin{array}{c} \left(y_j^{(i)} - y_{j_1}^{(i)}\right)^{\mathrm{T}} \\ \cdots \\ \left(y_j^{(i)} - y_{j_k}^{(i)}\right)^{\mathrm{T}} \end{array}\right]\right.
$$

$$
\left. \times \left[y_j^{(i)} - y_{j_1}^{(i)}, \ldots, y_j^{(i)} - y_{j_k}^{(i)}\right] \mathrm{diag}\left(w_j^{(i)}\right)\right)
$$

$$
= \arg\min_{Y_j^{(i)}} \mathrm{tr}\left(Y_j^{(i)} \left[\begin{array}{c} -e_k^{\mathrm{T}} \\ I_k \end{array}\right] \mathrm{diag}\left(w_j^{(i)}\right)\right.
$$

$$
\left. \times \left[-e_k \quad I_k\right] \left(Y_j^{(i)}\right)^{\mathrm{T}}\right)
$$

$$
= \arg\min_{Y_j^{(i)}} \mathrm{tr}\left(Y_j^{(i)} L_j^{(i)} \left(Y_j^{(i)}\right)^{\mathrm{T}}\right) \tag{3}
$$

where $e_k = [1, \ldots, 1]^{\mathrm{T}}$, I_k is a $k \times k$ identity matrix, and $\mathrm{tr}(\cdot)$ is the trace operator. $L_j^{(i)} \in \Re^{(k+1)\times(k+1)}$ encodes the objective function for the jth patch on the ith view. According to (3), it is

$$
L_j^{(i)} = \left[\begin{array}{c} -e_k^{\mathrm{T}} \\ I_k \end{array}\right] \mathrm{diag}\left(w_j^{(i)}\right) \left[-e_k \quad I_k\right]
$$

$$
= \left[\begin{array}{cc} \sum_{l=1}^{k} \left(w_j^{(i)}\right)_l & -\left(w_j^{(i)}\right)^{\mathrm{T}} \\ -w_j^{(i)} & \mathrm{diag}\left(w_j^{(i)}\right) \end{array}\right]. \tag{4}
$$

Therefore, the part optimization for $X_j^{(i)}$ is

$$
\arg\min_{Y_j^{(i)}} \mathrm{tr}\left(Y_j^{(i)} L_j^{(i)} \left(Y_j^{(i)}\right)^{\mathrm{T}}\right). \tag{5}
$$

Based on the locality information encoded in $L_j^{(i)}$, (5) finds a sufficiently smooth low-dimensional embedding $Y_j^{(i)}$ by preserving the intrinsic structure of the jth patch on the ith view.

Because of the complementary property of multiple views to each other, different views definitely have different contributions to the final low-dimensional embedding. In order to well explore the complementary property of different views, a set of nonnegative weights $\alpha = [\alpha_1, \ldots, \alpha_m]$ is imposed on part optimizations of different views independently. The larger α_i is, the more important role the view $X_j^{(i)}$ plays in learning to obtain the low-dimensional embedding $Y_j^{(i)}$. By summing over all views, the multiview part optimization for the jth patch is

$$
\arg\min_{Y=\left\{Y_j^{(i)}\right\}_{i=1}^{m}, \alpha} \sum_{i=1}^{m} \alpha_i \mathrm{tr}\left(Y_j^{(i)} L_j^{(i)} \left(Y_j^{(i)}\right)^{\mathrm{T}}\right). \tag{6}
$$

B. Global Coordinate Alignment

For each patch $X_j^{(i)}$, there is a low-dimensional embedding $Y_j^{(i)}$. All $Y_j^{(i)}$ can be unified together as a whole one by assuming that the coordinate for $Y_j^{(i)} = [y_j^{(i)}, y_{j_1}^{(i)}, \ldots, y_{j_k}^{(i)}]$ is selected from the global coordinate $Y = [y_1, \ldots, y_n]$, i.e., $Y_j^{(i)} = Y S_j^{(i)}$, wherein $S_j^{(i)} \in \Re^{n\times(k+1)}$ is the selection matrix to encode the spatial relationship of samples in a patch in the original high-dimensional space. That is, low-dimensional embeddings in different views are consistent with each other globally. Therefore, (6) can be equivalently rewritten as

$$
\arg\min_{Y,\alpha} \sum_{i=1}^{m} \alpha_i \mathrm{tr}\left(Y S_j^{(i)} L_j^{(i)} \left(S_j^{(i)}\right)^{\mathrm{T}} Y^{\mathrm{T}}\right). \tag{7}
$$

By summing over all part optimizations defined by (7), the global coordinate alignment is given by

$$
\arg\min_{Y,\alpha} \sum_{j=1}^{n} \sum_{i=1}^{m} \alpha_i \mathrm{tr}\left(Y S_j^{(i)} L_j^{(i)} \left(S_j^{(i)}\right)^{\mathrm{T}} Y^{\mathrm{T}}\right)
$$

$$
= \arg\min_{Y,\alpha} \sum_{i=1}^{m} \alpha_i \mathrm{tr}\left(Y L^{(i)} Y^{\mathrm{T}}\right) \tag{8}
$$

where $L^{(i)} \in \Re^{n\times n}$ is the alignment matrix for the ith view, and it is defined as

$$
L^{(i)} = \sum_{j=1}^{n} S_j^{(i)} L_j^{(i)} \left(S_j^{(i)}\right)^{\mathrm{T}}. \tag{9}
$$

Putting (4) into (9), we have

$$L^{(i)} = D^{(i)} - W^{(i)} \qquad (10)$$

where $W^{(i)} \in \Re^{n \times n}$ and $[W^{(i)}]_{pq} = \exp(-\|x_p^{(i)} - x_q^{(i)}\|^2/t)$ if $x_p^{(i)}$ is among the k-nearest neighbors of $x_q^{(i)}$ or vice versa; $[W^{(i)}]_{pq} = 0$ otherwise. In addition, $D^{(i)}$ is diagonal, and $[D^{(i)}]_{jj} = \sum_l [W^{(i)}]_{jl}$. Therefore, $L^{(i)}$ is an unnormalized graph Laplacian matrix [12]. In MSE, we adopt a normalized graph Laplacian matrix by performing a normalization on $L^{(i)}$

$$L_n^{(i)} = \left(D^{(i)}\right)^{-1/2} L^{(i)} \left(D^{(i)}\right)^{-1/2}$$

$$= I - \left(D^{(i)}\right)^{-1/2} W^{(i)} \left(D^{(i)}\right)^{-1/2} \qquad (11)$$

where $L_n^{(i)}$ is symmetric and positive semidefinite; the proof is given in Appendix A. The constraint $YY^T = I$ is imposed on (8) to uniquely determine the low-dimensional embedding Y, i.e.,

$$\arg\min_{Y,\alpha} \sum_{i=1}^{m} \alpha_i \text{tr}\left(Y L_n^{(i)} Y^T\right)$$

$$\text{s.t. } YY^T = I; \quad \sum_{i=1}^{m} \alpha_i = 1, \qquad \alpha_i \geq 0. \qquad (12)$$

The solution to α in (12) is $\alpha_k = 1$ corresponding to the minimum $\text{tr}(Y L^{(i)} Y^T)$ over different views, and $\alpha_k = 0$ otherwise. This solution means that only one view is finally selected by this method. Therefore, the performance of this method is equivalent to the one from the best view. This solution does not meet our objective on exploring the complementary property of multiple views to get a better embedding than based on a single view.

In this paper, we adopt a trick utilized in [13] to avoid this phenomenon, i.e., we set $\alpha_i \leftarrow \alpha_i^r$ with $r > 1$. In this condition, $\sum_{i=1}^{m} \alpha_i^r$ achieves its minimum when $\alpha_i = 1/m$ with respect to $\sum_{i=1}^{m} \alpha_i = 1$, $\alpha_i > 0$. Similar α_i for different views will be obtained by setting $r > 1$, so each view has a particular contribution to the final low-dimensional embedding Y. Therefore, the new objective function is defined as

$$\arg\min_{Y,\alpha} \sum_{i=1}^{m} \alpha_i^r \text{tr}\left(Y L_n^{(i)} Y^T\right)$$

$$\text{s.t. } YY^T = I; \quad \sum_{i=1}^{m} \alpha_i = 1, \qquad \alpha_i \geq 0 \qquad (13)$$

where $r > 1$. According to (13) and related discussions, MSE finds a low-dimensional sufficiently smooth embedding Y by preserving the locality of each view simultaneously. The solution of (13) is given in the next section.

C. Alternating Optimization

In this section, we derive the solution of MSE defined in (13), which is a nonlinearly constrained nonconvex optimization problem. To the best of our knowledge, there is no direct way to find its global optimal solution. In this paper, we derive an iterative algorithm by using the alternating optimization [14] to obtain a local optimal solution. The alternating optimization iteratively updates Y and α in an alternating fashion.

First, we fix Y to update α. By using a Lagrange multiplier λ to take the constraint $\sum_{i=1}^{m} \alpha_i = 1$ into consideration, we get the Lagrange function

$$L(\alpha, \lambda) = \sum_{i=1}^{m} \alpha_i^r \text{tr}\left(Y L_n^{(i)} Y^T\right) - \lambda \left(\sum_{i=1}^{m} \alpha_i - 1\right). \qquad (14)$$

By setting the derivative of $L(\alpha, \lambda)$ with respect to α_i and λ to zero, we have

$$\begin{cases} \frac{\partial L(\alpha,\lambda)}{\partial \alpha_i} = r\alpha_i^{r-1} \text{tr}\left(Y L_n^{(i)} Y^T\right) - \lambda = 0, & i = 1,\ldots,m \\ \frac{\partial L(\alpha,\lambda)}{\partial \lambda} = \sum_{i=1}^{m} \alpha_i - 1 = 0. \end{cases} \qquad (15)$$

Therefore, α_i can be obtained

$$\alpha_i = \frac{\left(1/\text{tr}\left(Y L_n^{(i)} Y^T\right)\right)^{1/(r-1)}}{\sum_{i=1}^{m} \left(1/\text{tr}\left(Y L_n^{(i)} Y^T\right)\right)^{1/(r-1)}}. \qquad (16)$$

The alignment matrix $L_n^{(i)}$ is positive semidefinite, so we have $\alpha_i \geq 0$ naturally. When Y is fixed, (16) gives the global optimal α.

According to (16), we have the following understanding for r in controlling α_i. If $r \to \infty$, different α_i will be close to each other. If $r \to 1$, only $\alpha_i = 1$ corresponding to the minimum $\text{tr}(Y L_n^{(i)} Y^T)$ over different views, and $\alpha_i = 0$ otherwise. Therefore, the selection of r should be based on the complementary property of all views. Rich complementary prefers large r; otherwise, r should be small. The effect of parameter r is discussed in our experiments in Section IV-E.

Second, we fix α to update Y. The optimization problem in (13) is equivalent to

$$\min_{Y} \text{tr}(Y L Y^T) \quad \text{s.t } YY^T = I \qquad (17)$$

where $L = \sum_{i=1}^{m} \alpha_i^r L_n^{(i)}$. Because $L_n^{(i)}$ is symmetric, L is symmetric. Based on the Ky-Fan theorem (the details of the Ky-Fan theorem are given in Appendix B), (17) has a global optimal solution Y when α is fixed. The optimal Y is given as the eigenvectors associated with the smallest d eigenvalues of the matrix L.

According to the aforementioned descriptions, we can form an alternating optimization procedure, depicted in Algorithm 1, to obtain a local optimal solution of MSE.

Algorithm 1: MSE Algorithm
Input: A multiple-view datum $X = \{X^{(i)} \in \Re^{m_i \times n}\}_{i=1}^{m}$, the dimension of the low-dimensional embedding d ($d < m_i, 1 \leq i \leq m$), and $r > 1$.
Output: A spectral embedding $Y \in \Re^{d \times n}$.
Method:
1: Calculate $L_n^{(i)}$ for each view according to the patch alignment.
2: Initialize $\alpha = [1/m, \ldots, 1/m]$.

3: **Repeat**

4: $Y = U^{\mathrm{T}}$, where $U = [u_1, \ldots, u_d]$ are the eigenvectors associated with the smallest d eigenvalues of the matrix L defined in (17).

5: $\alpha_i = ((1/\mathrm{tr}(Y L^{(i)} Y^{\mathrm{T}}))^{1/(r-1)}/(\sum_{i=1}^m (1/\mathrm{tr}(Y L^{(i)} Y^{\mathrm{T}}))^{1/(r-1)})$

6: **Until** convergence

The algorithm converges because the objective function $\sum_{i=1}^m \alpha_i^r \mathrm{tr}(Y L_n^{(i)} Y^{\mathrm{T}})$ reduces with the increasing of the iteration numbers. In particular, with fixed α, the optimal Y can reduce the value of the objective function, and with fixed Y, the optimal α will also reduce the value of the objective function.

D. Time-Complexity Analysis

The time complexity of MSE contains two parts. One is for the constructions of alignment matrices for different views, i.e., the computation of $L_n^{(i)}$. From (11), the time complexity of this part is $O((\sum_{i=1}^m m_i) \times n^2)$. The other is for the alternating optimization. The update of Y has the time complexity of the eigenvalue decomposition of an $n \times n$ matrix. It is $O(n^3)$. The update of α has the time complexity of $O((m + d) \times n^2)$. Therefore, the entire time complexity of MSE is $O((n^3 + (m + d) \times n^2) \times T + (\sum_{i=1}^m m_i) \times n^2)$, where T is the number of training iterations and T is around three (always less than five) in all experiments.

IV. EXPERIMENTAL RESULTS

Images and videos are usually represented by multiview features, and each view is represented in a high-dimensional space. In this paper, we compare the effectiveness of the proposed MSE with the conventional feature concatenation-based spectral embedding (CSE), the DSE [3], the average performance of the single-view-based spectral embedding (ASE), and the best performance of the single-view-based spectral embedding (BSE) in both image retrieval and video annotation. We also show the performance comparison in the application of document clustering, which shows the results for multiview data with poor complementary. Finally, we give our analysis on parameter r and the dimensionality of the low-dimensional embedding d.

In ASE, BSE, CSE, and DSE, the LE is adopted. The CSE is performed based on the Gaussian normalized concatenated vector from different views. For all experiments, the value of k for patch construction in MSE and that of k-nearest neighbor construction in LE are fixed at 30. In LE and MSE, an unweighted graph as defined in [9] is adopted.

A. Toy Data Set Test

In this section, we use a toy data set to illustrate the effectiveness of MSE in comparing with CSE-LE and DSE-LE. The toy data set, which is a subset of Corel image gallery, consists of three semantic categories, i.e., bus, ship, and train. Each category includes 100 images. Fig. 2 shows some example images. For each image, we extract two kinds of low-level visual features, i.e., a 64-D HSV color histogram (HSVCH)
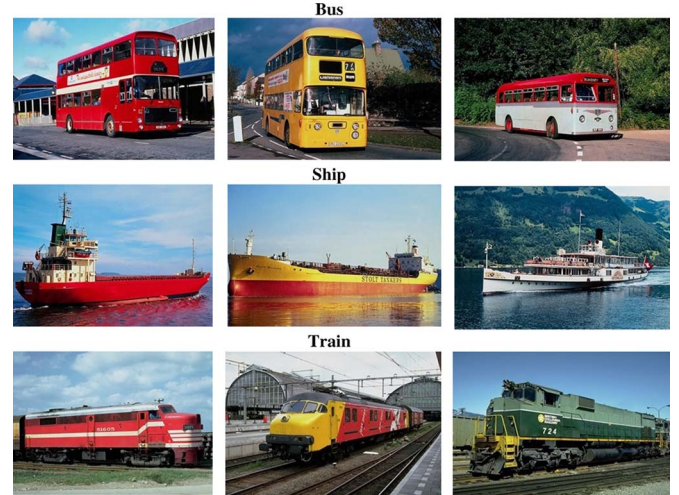

Fig. 2. Example images in the toy data set.

and a 75-D edge directional histogram (EDH), to represent two different views. Because the two views for an image are generally complementary to each other, we empirically set r in MSE as five.

Fig. 3(a) and (b) shows the low-dimensional embeddings obtained by LE performed on HSVCH and EDH independently, and Fig. 3(c)–(e) shows the embeddings obtained by CSE, DSE, and MSE, respectively. The results shown in Fig. 3(a)–(d) demonstrate that the existing algorithms merge different categories in the low-dimensional space. On the contrary, the proposed MSE can well separate different categories because MSE takes the complementary property of different views into consideration for embedding.

B. Image Retrieval

In this section, we show the effectiveness of MSE in image retrieval. The procedure for performance evaluation is as follows: 1) The low-dimensional embedding of an image retrieval data set is learned by an embedding algorithm, e.g., MSE, and 2) based on a low-dimensional embedding, a standard image retrieval procedure is conducted for all images in the data set. In detail, for each category, one image is selected as a query, and then, all the other images in the data set (including other categories) are ranked according to the Euclidean distance to the query computed in the low-dimensional embedding. The retrieval performance is evaluated through the average precision (AP) based on the top N images. The mean AP (MAP) is computed by averaging all APs for different categories. In this paper, we use MAP to compare different embedding algorithms.

Corel 2000 and Caltech256-2045 are utilized independently for an image-retrieval test. Corel 2000 is a subset of the Corel photo gallery. There are 20 image categories, and each category contains 100 images. These 20 categories are the following: balloon, beach, bird, bobsled, bonsai, building, bus, butterfly, car, cat, cougar, dessert, dog, eagle, elephant, firework, fitness, flag, foliage, and fox. Fig. 2 shows some example images from Corel 2000. Caltech256-2045 is a subset of the Caltech256 data set [15]; it contains 2045 images from 15 categories, including
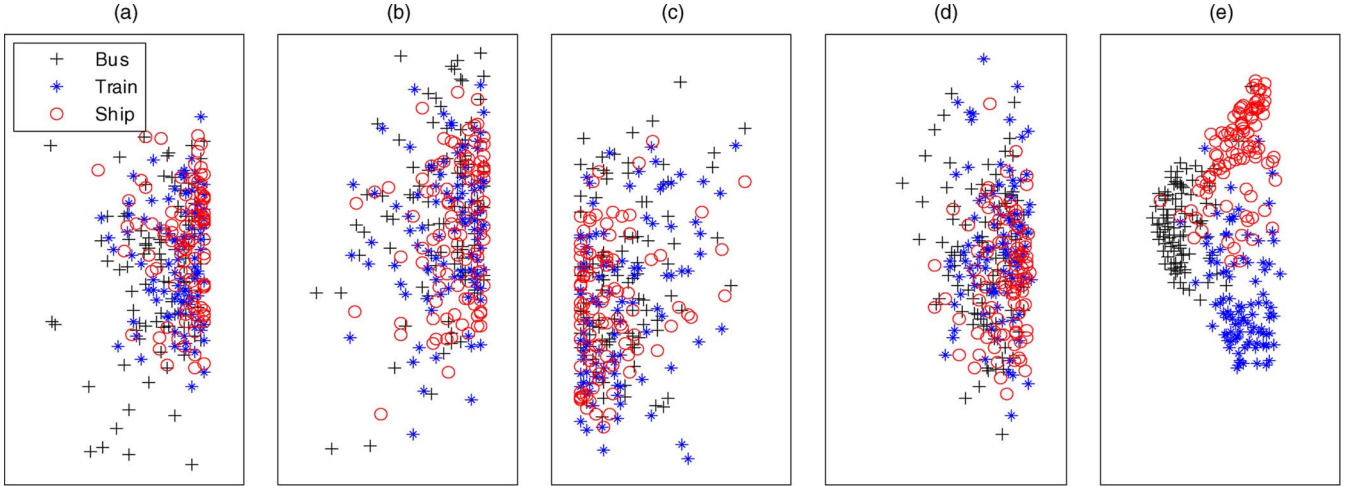
Fig. 3. Low-dimensional embeddings of different spectral-embedding algorithms. (a) LE on HSVCH. (b) LE on EDH. (c) CSE. (d) DSE. (e) MSE.

Fig. 4. Example images in Caltech256-2045.

AK-47, American flag, backpack, baseball bat, baseball glove, baseball hoop, bat, bathtub, bear, bear mug, billiards, binoculars, birdbath, blimp, and bonsai. Fig. 4 shows some example images from Caltech256-2045.

For each image, we extract five kinds of low-level visual features to represent five different views. These five features are color moment, color correlogram, HSVCH, edge directional histogram, and wavelet texture. The color moment consists of 5×5 blockwise Lab color moments, and each block is represented by three statistical moments: mean, variance, and skewness over three color channels. Therefore, the color moment is 225-D. The color correlogram is extracted according to [16] and is 114-D. The HSVCH contains 64 bins. A 75-D edge directional histogram is extracted for edge representation. A 128-D wavelet texture feature is extracted for texture representation. Table I summarizes the dimensionality for every modality. Because different views for an image are generally complementary to each other, we empirically set r in MSE as five. The dimensionality of the low-dimensional embedding d is set as 30.

The results for the performance comparison on Corel 2000 and Caltech256-2045 are shown in Fig. 6. MSE achieves the best performance on both two data sets.

C. Video Annotation

In this section, we show the effectiveness of MSE in video annotation. We adopt a standard video annotation testbed, the TRECVID 2008 training set [17], which consists of 39 674 shots from 20 concepts. A key frame is extracted from each shot for representation. We select ten concepts for performance evaluation. They are classroom, bridge, emergency vehicle, dog, airplane flying, kitchen, bus, harbor, telephone, and demonstration or pretest. We use all positive samples of the ten selected concepts to construct our data set, which has 1179 key frames. Fig. 5 shows some example key frames from TRECVID 2008. Similar to the performance-evaluation procedure used in the image retrieval, MAP is applied to measure the performance of an embedding algorithm. Because different views for a key frame are generally complementary to each other, we empirically set r in MSE as five. The dimensionality of the low-dimensional embedding d is set as 30.

For each key frame, we extract five kinds of low-level visual features, which are also used for image retrieval as shown in Table I, to represent five views. Fig. 6 shows that the proposed MSE achieves the best performance on this data set.

D. Document Clustering

In the aforementioned experiments, the data sets are multimedia data (e.g., an image or a video) which generally have a complementary among different features; however, the complementary is not very rich since the different features of a sample are relatively related to each other. In this section, we consider a text data set to show the performance on multiview data with very rich complementary, in which the different views of a sample are nearly independent. We utilize the application of document clustering to perform evaluation. The procedure for evaluation is as follows: 1) The low-dimensional embedding of a text data set is learned by an embedding algorithm,

TABLE I
DIMENSIONALITY OF FIVE FEATURES FOR IMAGE RETRIEVAL

Modality	Colour Moment	Colour Correlogram	HSV Histogram	Edge Distribution Histogram	Wavelet Texture
Dimension	225	114	64	75	128

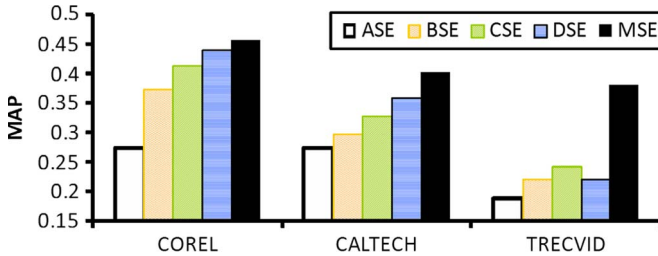Fig. 5.   Example images in TRECVID 2008.

Fig. 6.   Performance comparison on the three data sets measured by MAP of top 100 samples.

e.g., MSE, and 2) based on a low-dimensional embedding, a K-means algorithm [24] is performed to cluster the data set. As for cluster evaluation, following [25], we use the Rand index (RI) defined as

$$RI = \frac{\#CD}{N(N-1)/2} \qquad (18)$$

where N denotes the number of samples, $\#CD$ denotes the number of correct decisions, and a decision is considered correct if the clustering algorithm agrees with the real clustering. In our experiments, K-means adopts random sampling to select the initial cluster centroids, and we compute the average RI for evaluation after having performed K-means 50 times.

TABLE II
PERFORMANCE COMPARISON ON THE 20-NEWSGROUP DATA SET
MEASURED BY RI

	Non-DR	LE on View 1	LE on View 2	CSE	DSE	MSE
RI	0.527	0.521	0.492	**0.623**	0.594	**0.629**

The text data set in this experiment is a modified 20-newsgroup data set downloaded from [22]; it is a tiny version of the 20-newsgroup data set [23] with binary occurrence data for 100 selected keywords across 16 242 postings. Although the data are organized into 20 different newsgroups, we can separate them into four clusters according to the highest level of the topic naming, i.e., comp, rec, sci, and talk. We randomly select 500 samples from each cluster, and thus, our data set consists of 2000 samples from four clusters; each sample has a dimensionality of 100.

We generate two views on the text data set following the same way in [19]; we randomly draw 50 words from the 100 keywords and regard the term vector on the 50 words as view 1 and the term vector on the remaining 50 words as view 2. The two views are nearly independent since they are represented in two separated term spaces. Because the two views for a document have rich complementary among them, we empirically set r in MSE as nine. The dimensionality of the low-dimensional embedding d is set as ten.

Table II shows the results for performance comparison. "Non-DR" denotes the result of performing K-means on the original 100-D data; it is the result without performing dimensionality reduction. "LE on View 1" denotes the result by LE performed on view 1. We can see that MSE achieves the best performance. Compared with the multimedia data set, the performance of MSE has limited advantage on the 20-newsgroup data set, e.g., the RI of CSE is very close to that of MSE. That is because the two views of the text data set are nearly independent, and each single view has very low performance for clustering (i.e., 0.521 and 0.492); in such a case, it is very difficult to handle the disagreements from the two views.

E. Effect of Parameter r

In this section, we investigate the effect of parameter r in MSE. Table III illustrates the performance variation of MSE with respect to r; the dimensionality of the low-dimensional embedding d is set as 30 for the Corel 2000 data set and as 10 for the 20-newsgroup data set. As discussed in Section III-C, rich complementary prefers large r. It is known that the 20-newsgroup data set has richer complementary among different views than Corel 2000 does. We can see that the

TABLE III
PERFORMANCE COMPARISON OF MSE WITH DIFFERENT r

	r=2	r=3	r=4	r=5	r=6	r=7	r=8	r=9	r=10
Corel-2000 (MAP)	0.386	0.443	0.448	**0.456**	**0.455**	**0.456**	0.451	0.451	0.451
20-newsgroup (RI)	0.597	0.536	0.594	0.58	0.601	0.61	0.617	**0.629**	0.611

TABLE IV
PERFORMANCE COMPARISON ON THE COREL 2000 DATA SET WITH
DIFFERENT d MEASURED BY RI

	BSE	ASE	CSE	DSE	MSE
d=10	0.414	0.257	0.393	0.417	**0.475**
d=20	0.376	0.279	0.435	0.397	**0.488**
d=30	0.376	0.278	0.428	0.431	**0.456**
d=40	0.386	0.284	0.429	0.422	**0.448**
d=50	0.375	0.285	**0.437**	0.42	**0.428**

20-newsgroup data set performs the best around $r = 9$; the Corel 2000 data set performs the best around $r = 6$.

F. Performance Under Different Values of d

In this section, we show the performance comparison under different dimensions of the low-dimensional embedding. Table IV illustrates the performance comparison on the Corel 2000 data set with respect to d; parameter r is empirically set as five in MSE. We can see that MSE achieves the best performance on all the settings of d except when $d = 50$, and also, we can find that the optimal value of d for the Corel 2000 data set is between 10 and 20.

V. CONCLUSION

In many applications, objects are represented by different views. However, existing spectral-embedding algorithms cannot deal with multiview data directly. In this paper, we have proposed a novel MSE, which learns a low-dimensional and sufficiently smooth embedding over all views simultaneously.

From the experimental results, we can conclude that MSE can effectively explore the complementary property of different views to obtain an effective low-dimensional embedding for multiview data sets. We also find the following: 1) MSE has promising results on multimedia data (e.g., an image or a video), which generally have complementary among different features, and 2) MSE achieves limited success in the case when different views are nearly independent, and each single view is not informative enough, e.g., the text data set in Section IV-D; in such a case, the performances of CSE and MSE are very close.

There are, however, some open problems in MSE: 1) One is how to select the optimal dimension of the low-dimensional embedding; from Table IV, we can find that there exists an optimal value of d for some data set, and MSE achieves limited success under inappropriate setting of d; and 2) MSE is a spectral method which needs to perform eigenspace decomposition of matrices of size $N \times N$ (N is the number of samples in the data set); this generally takes $O(N^3)$ time, and it will be very time consuming when N is very large. In the future, we will consider how to utilize the sampling technique [26] in MSE to handle a large-scale data set.

APPENDIX I

A. Proof of $L_n^{(i)}$ Being Symmetric and Positive Semidefinite

According to (11), the symmetry of $L_n^{(i)}$ follows directly from the symmetry of $W^{(i)}$ and $D^{(i)}$. As for positive semidefiniteness, given a vector $f \in \Re^n$, we have

$$f^{\mathrm{T}} L_n^{(i)} f = f^{\mathrm{T}} f - f^{\mathrm{T}} \left(D^{(i)} \right)^{-1/2} W^{(i)} \left(D^{(i)} \right)^{-1/2} f$$

$$= \sum_{k=1}^{n} f_k^2 - \sum_{p,q=1}^{n} \frac{f_p}{\sqrt{\left[D^{(i)} \right]_{pp}}} \frac{f_q}{\sqrt{\left[D^{(i)} \right]_{qq}}} \left[W^{(i)} \right]_{pq}$$

$$= \frac{1}{2} \left(\sum_{k=1}^{n} f_k^2 - 2 \sum_{p,q=1}^{n} \frac{f_p}{\sqrt{\left[D^{(i)} \right]_{pp}}} \right.$$

$$\left. \times \frac{f_q}{\sqrt{\left[D^{(i)} \right]_{qq}}} \left[W^{(i)} \right]_{pq} + \sum_{k=1}^{n} f_k^2 \right)$$

$$= \frac{1}{2} \sum_{p,q=1}^{n} \left[W^{(i)} \right]_{pq} \left(\frac{f_p}{\sqrt{\left[D^{(i)} \right]_{pp}}} - \frac{f_q}{\sqrt{\left[D^{(i)} \right]_{qq}}} \right)^2$$

$$\geq 0$$

thus, $L_n^{(i)}$ is positive semidefinite.

B. Ky-Fan Theorem

Let $M \in \Re^{n \times n}$ be a symmetric matrix with the smallest k eigenvalues $\lambda_1 \leq \cdots \leq \lambda_k$, and the corresponding eigenvectors $U = [u_1, \ldots, u_k]$. Then, $\sum_{i=1}^{k} \lambda_i = \min_{X \in \Re^{n \times k}, X^{\mathrm{T}} X = I_k} \mathrm{tr}(X^{\mathrm{T}} M X)$. Moreover, the optimal X is given by UQ, where Q is an arbitrary orthogonal matrix. Please refer to [18] for the detailed proof.
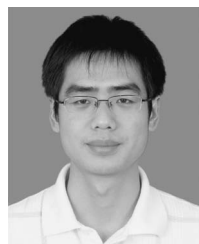
REFERENCES

[1] S. Ruping and T. Scheffer, in *Proc. Workshop Learning With Multiple Views, 22nd ICML*, Bonn, Germany, Aug. 2005.

[2] W. Liu, D. Tao, and J. Liu, "Transductive component analysis," in *Proc. IEEE Int. Conf. Data Mining*, 2008, pp. 433–442.

[3] B. Long, P. S. Yu, and Z. Zhang, "A general model for multiple view unsupervised learning," in *Proc. 8th SIAM Int. Conf. Data Mining*, Atlanta, GA, Apr. 2008, pp. 822–833.

[4] D. Zhou and C. Burges, "Spectral clustering and transductive learning with multiple views," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, Jun. 2007, pp. 1159–1166.

[5] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in

image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1088–1099, Jul. 2006.

[6] W. Bian and D. Tao, "Biased discriminant Euclidean embedding for content based image retrieval," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 545–554, Feb. 2010.

[7] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.

[8] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[9] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2002, pp. 585–591.

[10] D. L. Donoho and C. Grimes, "Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data," *Proc. Nat. Acad. Sci.*, vol. 100, no. 10, pp. 5591–5596, 2003.

[11] Z. Zhang and H. Zha, "Principle manifolds and nonlinear dimension reduction via local tangent space alignment," *SIAM J. Sci. Comput.*, vol. 26, no. 1, pp. 313–338, 2005.

[12] U. V. Luxburg, "A tutorial on spectral clustering," Max Planck Inst. Biol. Cybern., Tübingen, Germany, Tech. Rep. TR-149, Aug. 2006.

[13] M. Wang, X. S. Hua, X. Yuan, Y. Song, and L. R. Dai, "Optimizing multigraph learning: Towards a unified video annotation scheme," in *Proc. ACM Multimedia*, Augsburg, Germany, Sep. 2007, pp. 862–870.

[14] J. C. Bezdek and R. J. Hathaway, "Some notes on alternating optimization," in *Proc. AFSS Int. Conf. Fuzzy Syst.*, vol. 2275, *LNAI*, 2002, pp. 288–300.

[15] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. Technol., Pasadena, CA, Tech. Rep. TR-7694, Mar. 2007.

[16] F. Long, H. Zhang, and D. Feng, "Fundamental of content-based image retrieval," in *Multimedia Information Retrieval and Management*. New York: Springer-Verlag, 2002.

[17] TRECVID, Trec Video Retrieval Evaluation 2008. [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[18] R. Bhatia, *Matrix Analysis*. New York: Springer-Cerlag, 1997.

[19] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proc. Int. Conf. Data Mining*, 2004, pp. 19–26.

[20] Z. Zhao and H. Liu, "Multi-source feature selection via geometry-dependent covariance analysis," in *Proc. JMLR Workshop Conf.*, 2008, vol. 4, pp. 36–47.

[21] A. Zien and C. S. Ong, "Multiclass multiple kernel learning," in *Proc. 24th ICML*, Corvallis, OR, 2007, pp. 1191–1198.

[22] [Online]. Available: http://www.cs.toronto.edu/~roweis/data.html

[23] [Online]. Available: http://people.csail.mit.edu/jrennie/20Newsgroups/

[24] S. P. Lloyd, "Least square quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[25] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, Dec. 1971.

[26] A. Talwalkar, A. Kumar, and H. Rowley, "Large-scale manifold learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Anchorage, AK, Jun. 2008, pp. 1–8.

[27] X. Li, S. Lin, S. Yan, and D. Xu, "Discriminant locally linear embedding with high-order tensor data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 2, pp. 342–352, Apr. 2008.

[28] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, pp. 1–60, Apr. 2008.

[29] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1299–1313, Sep. 2009.

[30] T. Zhang, D. Tao, and J. Yang, "Discriminative locality alignment," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 725–738.

Dacheng Tao (M'07) received the B.Eng. degree from the University of Science and Technology of China, Hefei, China, the M.Phil. degree from the Chinese University of Hong Kong, Hong Kong, China, and the Ph.D degree from the University of London, London, U.K.

He is currently a Nanyang Assistant Professor with the School of Computer Engineering in the Nanyang Technological University. His research is mainly on applying statistics and mathematics for data analysis problems in computer vision, multimedia, machine learning, data mining, and video surveillance. He has authored/edited six books and eight journal special issues. He has published more than 100 scientific papers including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (T-PAMI) Neural Information Processing Systems, with best paper awards. One of his T-PAMI papers was listed as a "New Hot Paper" in ScienceWatch.com (Thomson Scientific). His H-Index in google scholar is 17+ and his Erdös number is 3.

Dr. Tao is an Associate Editor of IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING and the *Computational Statistics & Data Analysis* (Elsevier). He has (co)chaired for special sessions, invited sessions, workshops, panels, and conferences. He has served with more than 80 major international conferences and more than 30 prestigious international journals. He is a member of the IEEE Computer Society and IEEE Signal Processing Society.

Tao Mei (M'07) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively.

In 2006, he joined Microsoft Research Asia, Beijing, China, as a Researcher Staff Member. His current research interests include multimedia content analysis, computer vision, and Internet multimedia applications such as search, advertising, management, social network, and mobile applications. He is the author of one book, five book chapters, and over 80 journal and conference papers in these areas and is the holder of more than 20 filed international and U.S. patents or pending applications.

Dr. Mei is a member of the Association for Computing Machinery (ACM). He serves as an Editorial Board member for the *Journal of Multimedia* and a Guest Editor for IEEE MULTIMEDIA for the Special Issue on "Knowledge Discovery Over Community-Contributed Multimedia Data: Opportunities and Challenges," for ACM/Springer *Multimedia Systems* for the Special Issue on "Multimedia Intelligent Services and Technologies," and for the Elsevier *Journal of Visual Communication and Image Representation* for the Special Issue on "Large-Scale Image and Video Search: Challenges, Technologies, and Trends." He was the Principle Designer of the automatic video search system that achieved the best performance in the worldwide TRECVID evaluation in 2007. He was the recipient of the Best Paper and Best Demonstration Awards in the 2007 ACM International Conference on Multimedia, the Best Poster Paper Award in the 2008 IEEE International Workshop on Multimedia Signal Processing, and the Best Paper Award in the 2009 ACM International Conference on Multimedia.

Tian Xia received the B.E. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2004 and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2009.

He is currently with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests focus on multimedia retrieval and machine learning.

Yongdong Zhang (M'08) received the Ph.D. degree from Tianjin University, Tianjin, China, in 2002.

He is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests focus on image processing and video processing.