m-SNE: Multiview Stochastic Neighbor Embedding

Bo Xie, Yang Mu, Dacheng Tao, *Member, IEEE*, and Kaiqi Huang, *Senior Member, IEEE*

***Abstract*—Dimension reduction has been widely used in real-world applications such as image retrieval and document classification. In many scenarios, different features (or multiview data) can be obtained, and how to duly utilize them is a challenge. It is not appropriate for the conventional concatenating strategy to arrange features of different views into a long vector. That is because each view has its specific statistical property and physical interpretation. Even worse, the performance of the concatenating strategy will deteriorate if some views are corrupted by noise. In this paper, we propose a multiview stochastic neighbor embedding (m-SNE) that systematically integrates heterogeneous features into a unified representation for subsequent processing based on a probabilistic framework. Compared with conventional strategies, our approach can automatically learn a combination coefficient for each view adapted to its contribution to the data embedding. This combination coefficient plays an important role in utilizing the complementary information in multiview data. Also, our algorithm for learning the combination coefficient converges at a rate of $O(1/k^2)$, which is the optimal rate for smooth problems. Experiments on synthetic and real data sets suggest the effectiveness and robustness of m-SNE for data visualization, image retrieval, object categorization, and scene recognition.**

***Index Terms*—Dimension reduction, image retrieval, multiview learning, stochastic neighbor embedding.**

I. Introduction

In MANY COMPUTER vision and information retrieval applications, data often lie in a high-dimensional space. Direct manipulation in this space is difficult because of the so-called "curse of dimensionality." Many algorithms have been proposed to find a low-dimensional embedding of the original high-dimensional data. For example, principal component analysis [1] finds an orthogonal subspace to encode the data variance. Popular nonlinear embedding methods, such as ISOMAP [2], Laplacian eigenmaps [3], local linear embedding [4], and stochastic neighbor embedding (SNE) [5], preserve pairwise distances in the low-dimensional space by considering different types of geometric properties. Some algorithms utilize discriminative information provided in the labels for more effective classification tasks [6]–[8]. Methods that preserve locality [9]–[13] have also been applied to areas where data lie on a manifold. In addition, manifold assumption can be combined with sparse coding formulation [14], maximum margin [15], or various other assumptions [16]–[20] to perform dimension reduction.

However, these approaches are only applicable to single-view data, while in many real-world scenarios, multiple views are present and complementary to each other. A view of data refers to a type of feature that summarizes a specific characteristic of the data. (Note that feature in this paper refers to a multidimensional vector; in data mining community, feature may refer to individual dimension of the multidimensional vector.) For example, an image can be characterized by color histograms, as well as a collection of scale-invariant feature transform (SIFT) descriptors. Each view represents the data partially. Therefore, combining multiple views can better represent the data for subsequent utilizations, e.g., retrieval and recognition.

One way of incorporating multiview data is simply concatenating them into a long vector and applying the conventional dimension reduction techniques. This strategy has three major problems: 1) Different statistical properties are not duly considered; 2) the complementary information of different features is not well explored; and 3) the performance of concatenation will easily deteriorate if one or more views are corrupted by noise. For example, the bag-of-words feature is a histogram and the value of a specific bin refers to the counts of a specific word, while a wavelet feature is made up of responses of wavelet filters. It makes little sense to directly concatenate these two types of features into a long vector because they have different physical meanings and statistical properties, e.g., different means and variances. Although normalizing can alleviate the problem, there are currently no proper methods to find optimal combination coefficients for combining these views.

We propose to unify different features under a probabilistic framework. Under each view, we can construct a probability distribution from pairwise distances as that used in SNE. Unifying heterogeneous features in such a probabilistic strategy is meaningful because data with different properties, e.g., mean and variance, can be appropriately combined in a probability space. In our method, probability distributions encode the similarities between data points, which are the most important factors for embedding. In addition, features are effectively combined at this level. In this sense, unifying features under a probabilistic framework is a more principled approach because it circumvents the problems of incomparable scales and different representations. Moreover, our algorithm can automatically learn a weighting for each view; therefore, discriminative features are promoted, and random noise is suppressed. At

Manuscript received July 5, 2010; revised October 10, 2010 and December 19, 2010; accepted January 4, 2011. Date of publication February 4, 2011; date of current version July 20, 2011. This paper was recommended by Associate Editor E. Santos, Jr.

B. Xie is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: bxie@ee.cuhk.hk).

Y. Mu is with the Department of Computer Science, University of Massachusetts, Boston, MA 02125 USA (e-mail: yangmu@cs.umb.edu).

D. Tao is with the Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney, NSW 2700, Australia (e-mail: dacheng.tao@uts.edu.au).

K. Huang is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China (e-mail: kqhuang@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2011.2106208

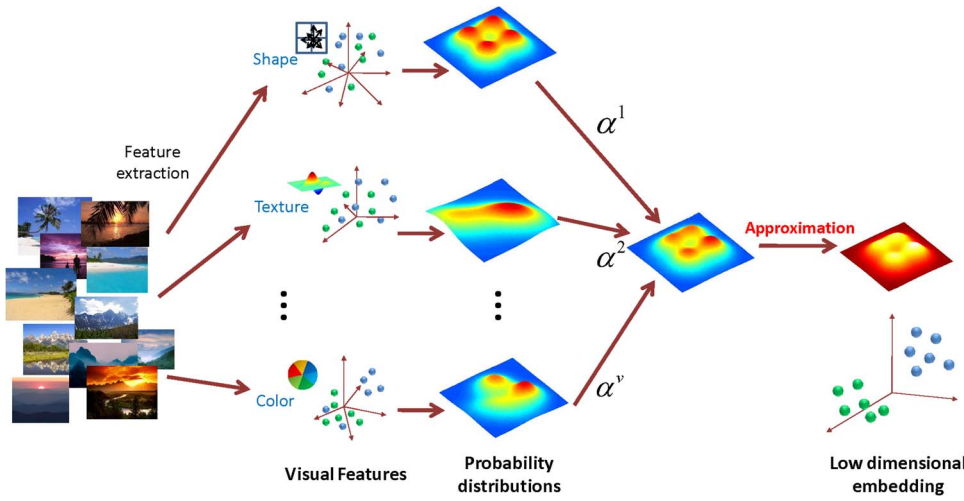


Fig. 1. System diagram for the proposed method. Different features (views) of the data are combined in the probabilistic space. The final low-dimensional embedding is learned by alternating between t-SNE and solving optimal combination coefficients (best viewed in color).

last, our algorithm for learning the weighting converges at an optimal rate of $O(1/k^2)$, where k is the number of iterations, by adopting Nesterov's accelerated first-order method [21], [22].

The remainder of the paper is organized as follows. In Section II, we briefly review on SNE. In Sections III and IV, we propose our multiview SNE (m-SNE) algorithm and provide its solution, as well as analytical properties. Experimental results are shown in Section V, and we conclude in Section VI.

II. SNE and Its Variants

The main idea of SNE is to construct probability distributions from pairwise distances wherein larger distances correspond to smaller probabilities and vice versa. Then, low-dimensional embedding is obtained by minimizing the Kullback–Leibler (KL) divergence of the two probability distributions.

The t-distributed SNE (t-SNE) [23] extends SNE in two ways: 1) A symmetric joint probability distribution is defined instead of the original conditional distribution, and this modification leads to simpler gradient computation in optimization, and 2) A t-distribution with degree one is used for low-dimensional data. This not only accelerates computation by avoiding costly exponential calculation but also alleviates the crowding problem because of the heavy-tail nature of t-distribution. Heavy-tailed SNE [24] generalizes the t-distribution to any heavy-tailed distribution and uses fixed-point optimization instead of gradient descent.

Formally, suppose that we have high-dimensional data points $\{x_i \in \mathbb{R}^d\}_{i=1}^n$. Moreover, the normalized pairwise distances, which can be considered as a joint probability distribution over sample pairs, are represented in a symmetric matrix $P \in \mathbb{R}_+^{n\times n}$, where $p_{ii} = 0$ and $\sum_{i,j} p_{ij} = 1$. Similarly, in the low-dimensional embedding, we define the probability distribution Q, with each element

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1 + \|y_k - y_l\|^2\right)^{-1}} \tag{1}$$

where $y_i \in \mathbb{R}^r$ is the low-dimensional data corresponding to x_i.

To find the optimal low-dimensional embedding, the KL divergence between the two distributions over all data points

$$f = KL(P|Q) = \sum_{i\neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2}$$

is minimized.

The objective function is not convex, and gradient descent can be used to find a local solution. The gradient with respect to a low-dimensional data point is

$$\frac{\partial f}{\partial y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + \|y_i - y_j\|^2\right)^{-1}. \tag{3}$$

Aside from the basic gradient descent, a number of strategies are employed to improve optimization [23].

III. m-SNE

In this section, we generalize SNE to handle multiview data. The probabilistic nature of SNE offers a clean formulation for combining different features (Fig. 1). In optimization, we alternatively solve SNE and learn combination coefficients, i.e., weightings for different views. Here, we base our algorithm on the variant t-SNE for speed benefits.

A. Multiview Dimension Reduction

We assume that the final probability distribution on the high-dimensional space is a convex combination of all the different views, i.e.,

$$p_{ij} = \sum_{t=1}^{v} \alpha^t p_{ij}^t \tag{4}$$

where α^t is the combination coefficient for view t and p_{ij}^t is the probability distribution under view t. The coefficient vector $\alpha = [\alpha^1, \cdots, \alpha^v]^T$ lies on a simplex in $\mathbb{R}^v$, denoted as $\alpha \in \Delta^v$. This is the same as $\alpha^t \geq 0$, $t = 1, \cdots, v$ and $\sum_{t=1}^v \alpha^t = 1$.

Obviously, p_{ij} is a probability distribution since $\sum_{i\neq j} p_{ij} = \sum_t \alpha^t \sum_{i\neq j} p_{ij}^t = \sum_t \alpha^t = 1$.

The optimization is then over both y_i and α. We adopt alternating optimization to solve this problem. In every round, we first fix α and use t-SNE to find low-dimensional embedding. After that, we fix y_i and optimize over α.

One caveat with this approach is that all the coefficients will concentrate on a single view that performs the best while the contributions of other views vanish. We tackle this problem by adding an l_2 norm regularization term to balance the coefficients over all views. The new objective function for learning optimal combination coefficients is

$$g(\alpha) = \sum_{i\neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} + \lambda \|\alpha\|^2 \tag{5}$$

where λ is a tradeoff coefficient.

Fortunately, this problem is convex and can be solved with a convergence rate of $O(1/k^2)$, where k is the number of iterations.

Proposition 1: The objective function $g(\alpha)$ is convex.

Proof: It is known that the KL divergence is convex in p_{ij} [25]. Since p_{ij} is linear in α and the l_2 norm is a convex function of α, this implies that the optimization problem is convex in α. ■

Algorithm 1 Multiview Stochastic Neighbor Embedding (m-SNE)

Input: high-dimensional data of different views $[x_1^1, \cdots, x_n^1], [x_1^2, \cdots, x_n^2], \cdots, [x_1^v, \cdots, x_n^v]$

Output: low-dimensional data $[y_1, \cdots, y_n]$

Initialize $\alpha = [\alpha^1, \cdots, \alpha^v] = [(1/v), \cdots, (1/v)]$

Compute similarity probability matrices of different views: $P^t \in \mathbb{R}^{n\times n}$, $t = 1, \cdots, v$ as in [23]

Repeat for m rounds:

1) Compute the combined probability matrix P using (4).
2) Compute low-dimensional data $[y_1, \cdots, y_n]$ using (3).
3) Compute combination coefficients α by minimizing (5) using Algorithm 2.

IV. Nesterov's Accelerated First-Order Method

Recently, Nesterov [21], [26], [27] has proposed an accelerated first-order method for convex optimization problems that converges at a rate of $O(1/k^2)$, where k is the number of iterations. We adopt this method to optimize α with fixed y_i.

Nesterov's method requires that the gradient of objective function is Lipschitz continuous, i.e., $\|\nabla g(\alpha_1) - \nabla g(\alpha_2)\| \leq L_g \|\alpha_1 - \alpha_2\|$. This condition is equivalent to

$$g(\alpha_1) \leq g(\alpha_2) + \langle \nabla g(\alpha_2), \alpha_1 - \alpha_2 \rangle + \frac{L_g}{2} \|\alpha_1 - \alpha_2\|^2 \tag{6}$$

where $\alpha_1 \in \Delta^v$ and $\alpha_2 \in \Delta^v$ are two arbitrary data points and L_g is the Lipschitz constant of $g(\alpha)$.

Here, we prove that the derivative of our objective function $\nabla g(\alpha)$ is Lipschitz continuous. Since $\nabla g(\alpha)$ is smooth, this reduces to all elements of Hessian that are bounded for all $\alpha \in \Delta^v$. Denote $c_{ij} = [p_{ij}^1, \cdots, p_{ij}^v]^{\mathrm{T}}$, and calculate the Hessian to be

$$H(g) = \sum_{i\neq j} \frac{c_{ij} c_{ij}^{\mathrm{T}}}{c_{ij}^{\mathrm{T}} \alpha} + 2\lambda I. \tag{7}$$

Proposition 2: The Hessian of $g(\alpha)$ is bounded.

Proof: It is sufficient to prove that $\sum_{i\neq j} (c_{ij} c_{ij}^{\mathrm{T}} / c_{ij}^{\mathrm{T}} \alpha)$ is bounded. Denote $e = \min_{\alpha\in\Delta^v} \min_{i\neq j} c_{ij}^{\mathrm{T}} \alpha$, and from definition, we have $c_{ij} > 0$; thus, $e > 0$. All entries in $c_{ij} c_{ij}^{\mathrm{T}}$ are positive and bounded, and the number of samples is finite; therefore, $\forall i \neq j$, $(H(g))_{ij} < M$, where M is a large positive constant. This completes the proof. ■

A. Accelerated First-Order Method

Now, we are ready to introduce Nesterov's accelerated first-order method. In every iteration round, we construct an estimate function

$$\psi_k(\alpha) = \frac{1}{2}\|\alpha - \alpha_0\|^2 + \sum_{i=0}^{k} b_i \left[g(\alpha_i) + \langle \nabla g(\alpha_i), \alpha - \alpha_i \rangle \right] \tag{8}$$

where $k \geq 0$, $\alpha_0 \in \Delta^v$ is the initial guess point, and b_i is some scaling coefficient.

Also, we construct a second-order approximation to the original function

$$m_L(y; \alpha) = g(y) + \langle \nabla g(y), \alpha - y \rangle + \frac{L}{2} \|\alpha - y\|^2 \tag{9}$$

$$T_L(y) = \arg\min_{\alpha\in\Delta^v} m_L(y; \alpha) \tag{10}$$

where y is any point and L is a positive constant.

Intuitively, the first function captures the history information about the objective function, and the second function is related to the gradient information of the current iteration. The two optimization problems have quadratic objective functions with linear constraints and can be conveniently solved using standard optimization toolkit.

We design the algorithm to maintain the following two relations in every iteration round:

$$\mathcal{R}_k^1 : B_k g(\alpha_k) \leq \psi_k^* \equiv \min_{\alpha} \psi_k(\alpha) \tag{11}$$

$$\mathcal{R}_k^2 : \psi_k(\alpha) \leq B_k g(\alpha) + \frac{1}{2}\|\alpha - \alpha_0\|^2 \tag{12}$$

where $k \geq 0$, $B_0 = 0$, and $B_k = B_{k-1} + b_k$.

These two relations lead to the following convergence rate of the minimizing sequence:

$$g(\alpha_k) - g(\alpha^*) \leq \frac{\|\alpha^* - \alpha_0\|^2}{2B_k} \tag{13}$$

where $k \geq 1$ and α^* is the optimal solution to the problem.

Detailed algorithm is given in Algorithm 2.

Algorithm 2 Accelerated First-Order Method for Combination Coefficient

Initialize $B_0 = 0$, $\alpha_0, L_0 > 0$, $\psi_0(\alpha) = (1/2)\|\alpha - \alpha_0\|^2$

Iterate:

1) Set $L = L_k$.
2) Repeat
   a) Find b by solving the equation
   $$Lb^2 = 2(B_k + b) \tag{14}$$
   b) Set $z_k = \arg\min_\alpha \psi_k(\alpha)$ and
   $$y = \frac{B_k\alpha_k + bz_k}{B_k + b} \tag{15}$$
   c) If $\langle \nabla g(T_L(y)), y - T_L(y)\rangle \le (1/L)\|\nabla g(T_L(y))\|^2$ Then $L := \gamma L$ Else Break
3) Set $y_k := y$, $M_k := L$, $b_{k+1} := b$, $L_{k+1} := L$ and update
$$\alpha_{k+1} = T_{M_k}(y_k) \tag{16}$$
$$\psi_{k+1}(\alpha) = \psi_k(\alpha) + b_{k+1}\left[g(\alpha_{k+1}) + \langle \nabla g(\alpha_{k+1}), \alpha - \alpha_{k+1}\rangle\right] \tag{17}$$

In our algorithm, the stopping criterion is

$$\langle \nabla g\left(T_L(y)\right), y - T_L(y)\rangle \le \frac{1}{L}\left\|\nabla g\left(T_L(y)\right)\right\|^2. \tag{18}$$

In the next lemma, we show a property of this condition.

Lemma 1: The condition (18) is satisfied for any $L \ge L_g$.

Proof: Denote $T = T_L(y)$. We use the first-order optimality condition (4.2) in [26]

$$\nabla g(T) = L(y - T) + \nabla g(T) - \nabla g(y) \tag{19}$$

Multiplying both sides by $y - T$ leads to

$$\begin{aligned}\langle \nabla g(T), y - T\rangle &= L\|y - T\|^2 + \langle \nabla g(T) - \nabla g(y), y - T\rangle \\ &= \frac{1}{L}\Big[\|\nabla g(T)\|^2 - \|\nabla g(T) - \nabla g(y)\|^2 \\ &\qquad + 2L\langle \nabla g(y) - \nabla g(T), y - T\rangle\Big] \\ &\quad - \langle \nabla g(y) - \nabla g(T), y - T\rangle \\ &= \frac{1}{L}\|\nabla g(T)\|^2 - \frac{1}{L}\|\nabla g(T) - \nabla g(y)\|^2 \\ &\quad + \langle \nabla g(y) - \nabla g(T), y - T\rangle. \end{aligned} \tag{20}$$

Since $\nabla g(\alpha)$ is Lipschitz continuous with constant L_g, the condition is always satisfied for any $L \ge L_g$. Therefore, we can guarantee that

$$L_k \le M_k \le \gamma L_g. \tag{21}$$

■

This varying step size approach avoids costly computation of the Lipschitz constant and can adjust the step size according to local variation.

B. Convergence Analysis

We first show that the sequences $\{\alpha_k\}$, $\{y_k\}$, $\{z_k\}$, and $\{b_k\}$ generated by Algorithm 2 satisfy the two relations in every iteration round. Then, we prove that the coefficients $\{B_k\}$ are $O(k^2)$. Combining these results with (13), we show that the convergence rate of our algorithm is $O(1/k^2)$.

Lemma 2: The sequences $\{\alpha_k\}$, $\{B_k\}$, and $\{\psi_k\}$ generated by Algorithm 1 satisfy relations for all $k \ge 0$.

The proof employs similar strategies as in [26] and is given in the Appendix.

Next, we show in the following lemma that the scaling coefficients $\{B_k\}$ grow at a quadratic order.

Lemma 3: The scaling coefficients grow as follows:

$$B_k \ge \frac{k^2}{2\gamma L_g}. \tag{22}$$

Proof: Using the updating rule (14) in Algorithm 2, we have

$$\begin{aligned} B_{k+1} &= \frac{M_k}{2}(B_{k+1} - B_k)^2 \\ &= \frac{M_k}{2}\left(B_{k+1}^{1/2} - B_k^{1/2}\right)^2\left(B_{k+1}^{1/2} + B_k^{1/2}\right)^2 \\ &\le 2B_{k+1}M_k\left(B_{k+1}^{1/2} - B_k^{1/2}\right)^2 \\ &\le 2B_{k+1}\gamma L_g\left(B_{k+1}^{1/2} - B_k^{1/2}\right)^2. \end{aligned} \tag{23}$$

Thus, for any $k \ge 0$, we obtain

$$B_k^{1/2} \ge \frac{k}{\sqrt{2\gamma L_g}}. \tag{24}$$

This implies the lemma. ■

Finally, by combining (13) and (22), we arrive at the following theorem.

Theorem 1: Let $\nabla g(\alpha)$ be Lipschitz-continuous with constant L_g and $0 \le L_0 \le L_g$. Moreover, the rate of convergence of Algorithm 1 is given as follows:

$$g(\alpha_k) - g(\alpha^*) \le \frac{\gamma L_g\|\alpha^* - \alpha_0\|^2}{k^2}, \quad k \ge 0. \tag{25}$$

V. Experiment

In this section, we first demonstrate the effectiveness of the proposed algorithm on a toy data set. Afterward, we present experimental results on real data sets in image retrieval, object categorization, and scene recognition.

A. Toy Data set

We have designed a toy data set by using a subset of the USPS digits [28] (digits 1 through 5). We generated four different views from these data by mixing some classes via Fischer linear discriminative analysis (FLDA) [29]. FLDA embeds data

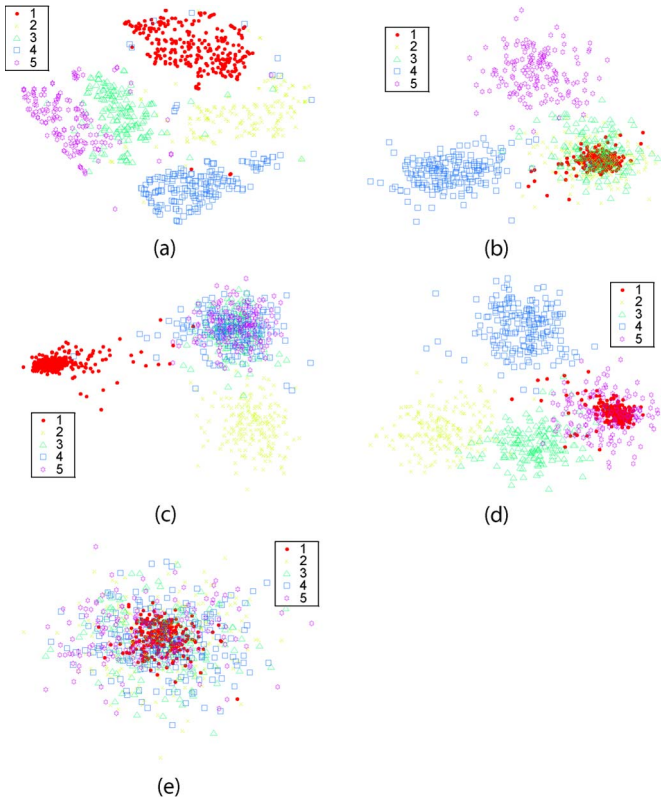


Fig. 2. Toy data set. (a) USPS digits 1–5. (b) View 1; digits 1, 2, and 3 are mixed. (c) View 2; digits 3, 4, and 5 are mixed. (d) View 3; digits 1 and 5 are mixed. (e) View 4; digits are randomly mixed.

in a low-dimensional space such that the trace of within-class scatter matrix is minimized while the trace of between-class scatter matrix is maximized. In detail, in the first view, digit classes 1, 2, and 3 were considered as the same; in the second view, classes 3, 4, and 5 were relabeled into the same class, and for the third view, labels of classes 1 and 5 were made the same. Then, FLDA was performed with the new labels to map the original 256-D data into a 2-D space. The fourth view was generated similarly but the class labels were random. Thus, this view is considered as noise. We have also increased the magnitude of the noise to illustrate better the problem that different features may contribute the classification differently. In Fig. 2, we plot the original data (by using t-SNE) and the four generated views.

We then applied our m-SNE algorithm to combine the four views and embedded the data into a 2-D space for visualization. For comparison, t-SNE by concatenating multiple views and distributed approach to spectral clustering (DSE) [30] were also conducted. Experimental results are shown in Fig. 3.

From the results, we can see that simple concatenation and DSE cannot duly combine data from multiple views. It was affected by the large magnitude noise, and thus, some classes were mixed up. Our m-SNE method correctly integrated the four views and produced a good low-dimensional embedding. The coefficient α also depicts different importance of these views: The first two views receive intermediate importance weightings because they provide intermediate amount of information, the third view has the largest weighting since it is least distorted, and the fourth view, which is random noise, is depressed in weighting.

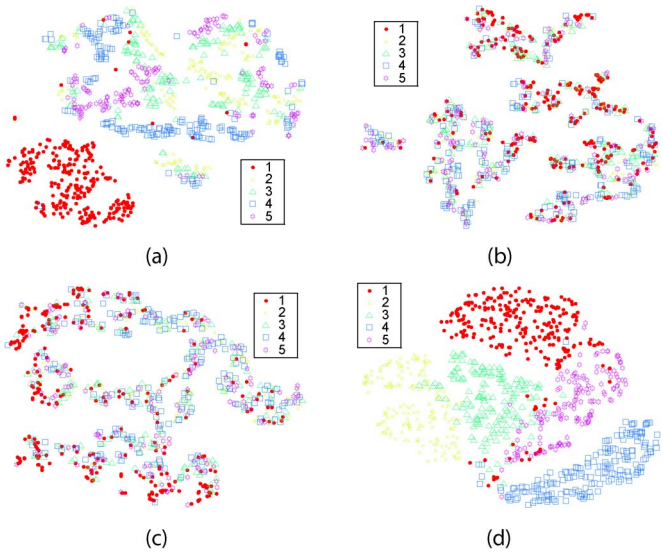


Fig. 3. Results by different algorithms on the toy data set. (a) Concatenation. (b) DSE with low-dimensional embedding by t-SNE. (c) DSE with low-dimensional embedding by Laplacian eigenmap. (d) m-SNE ($\alpha = [0.291, 0.201, 0.393, 0.115]$).

Fig. 4. Some sample images of NUS WIDE data set.

B. Image Retrieval

Next, we conducted an image retrieval experiment on the NUS WIDE data set [31]. The data set contains 269 648 images with a total of 81 concepts. It is a challenging data set since the concepts differ greatly and the images have large variances in scale, color, and shape (Fig. 4). Six low-level features are provided for these images, including 500-D bag of words based on SIFT descriptions, 64-D color histogram, 225-D blockwise color moments (CMs), 144-D color correlogram, 73-D edge direction histogram, and 128-D wavelet texture.

For evaluation, we used the precision curve against the number of images retrieved. Specifically, we randomly sampled ten query images per concept, and the averaged precision results were generated for each concept. To perform multiview image retrieval, each query image first retrieves 100 nearest neighbors for each view with Euclidean distance. Thus, each query has less than or equal to 700 retrieval candidates since some retrieval candidates may be the same for several views. In order to rank these retrieval candidates combining multiple views, we performed m-SNE with the query image and its retrieval candidates with parameter λ set to five. The final ranking

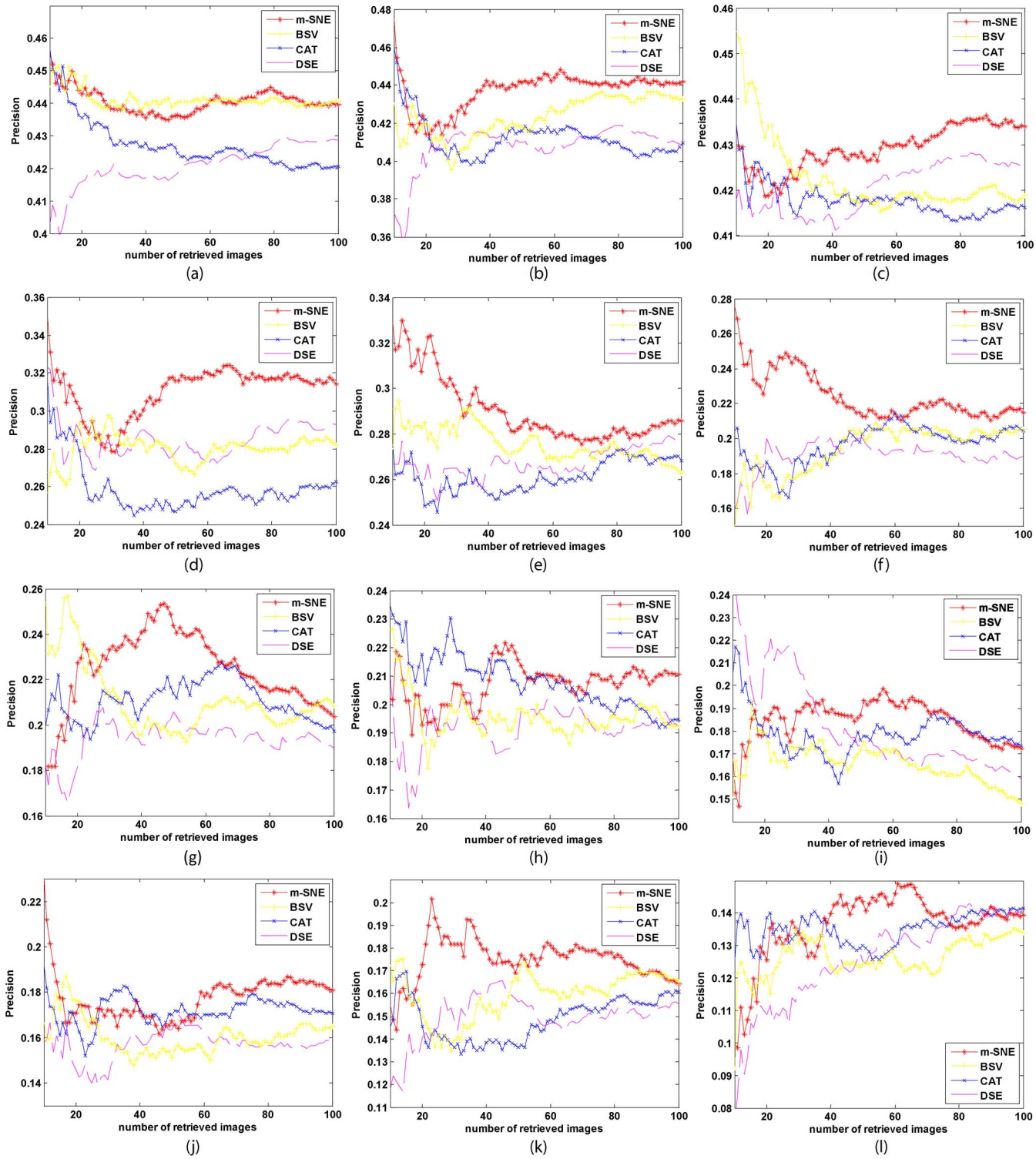


Fig. 5. Retrieval results on NUS WIDE data set. Precision results for m-SNE, BSV, CAT, and DSE are shown for 12 concepts. (a) Sky. (b) Reflection. (c) Water. (d) Boats. (e) Plants. (f) Surf. (g) Fox. (h) Whales. (i) Garden. (j) Toy. (k) Statue. (l) Person.

was calculated based on the candidates' Euclidean distances to the query image in the low-dimensional embedded space (30-D in the experiment). We also conducted experiments by using concatenated features (CAT) and DSE. For comparison, precision results of 12 concepts (due to limitation of space, we do not show all the results) for best single view (BSV), CAT, DSE, and m-SNE are shown in Fig. 5. We can see that m-SNE performs better than or comparable to others.

Fig. 6. Some sample images of Caltech 256 data set.

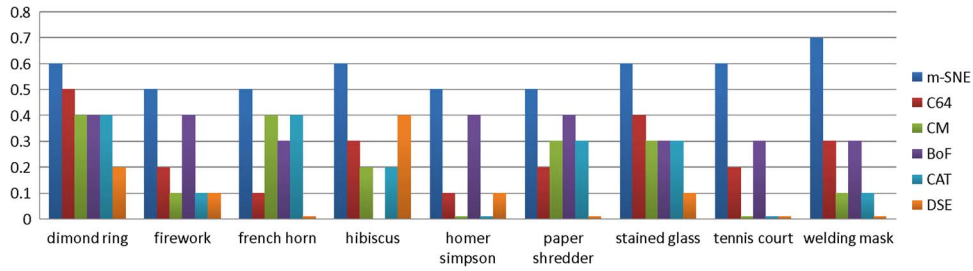


Fig. 7. Comparison of accuracies on nine categories.

C. Object Categorization

The Caltech 256 data set [32] contains 29 780 images from 256 different categories. There are many variations in the objects' scale, shape, and location (Fig. 6). We adopted three different kinds of features: C64 feature [33] (C64), CM, and bag of features (BoF). C64 is a 64-D low-level feature that considers color and texture. Some 225-D CMs consists of 5×5 blocks, with each block represented by mean, variance, and skewness over three Lab color channels. BoF is calculated by extracting local descriptors and aggregating them through spatial pyramid into a high-dimensional vector for each image [34]. We learned a 500-item visual dictionary and used 1×1, 2×2, and 3×1 spatial pyramid arrangements, resulting in 8000-D features.

Since SNE is slow to deal with large-scale data sets, we used a different experiment setting other than that used in [32]. For each category, we randomly selected ten test images. The rest made up the training set. A nearest neighbor classifier was used. In detail, each test image was classified as the same label as the nearest one in the training set. In order to decide between multiple views, we applied m-SNE on each test image and the nearest neighbors under different views and learned a 30-D embedding with parameter λ set to five. Thus, the nearest neighbor combining multiple views was the image with the smallest Euclidean distance to the test image in the low-dimensional embedding. Classification accuracy per category was averaged among the ten test images. For comparison, we also performed experiments with each single view, CAT, and DSE.

In 142 out of 256 categories, m-SNE achieved the highest accuracies among other approaches. C64 single view was the highest in 12 categories, 27 categories for CM, 68 for BoF, 2 for CAT, and 5 for DSE. In Fig. 7, we demonstrate the performance of these six approaches on nine categories.

D. Scene Recognition

Indoor scene data set [35] is a challenging collection of 67 categories of indoor scenes. There are a total of 15 620 images and at least 100 images for each category. These categories further constitute five larger categories of store, home, public spaces, leisure, and working place. The scene images have a lot of clutter and vary greatly in scale, configuration, and viewpoints (Fig. 8).

We used the same subset of the data set as in [35], with 80 images for training and 20 for testing in each category. The

Fig. 8. Some sample images of indoor scene data set.

proposed m-SNE was performed on the combined training and testing set to learn a 30-D embedding with parameter λ set to five. Then, the nearest neighbor rule was used in computing the labels of the test samples. Average accuracies were calculated with the 20 testing images for each category. Following similar setting, we conducted experiments by using each single-view feature, CAT, and DSE.

Finally, we report accuracy results for the average performance of every single view (ASV), one single view that performs the best compared to other views (BSV), CAT, DSE, and m-SNE. Results for the top 20 scenes over all 67 scenes are summarized in Table I. In the experiment, BSV is color moment which encodes the arrangement of color information in the image. However, it does not always perform well in all conditions; for some scenes, it is below average performance. Also, we can see that CAT may be contaminated by noise because it assigns the same weight for all views. DSE has very poor performance mainly because the radius parameter of graph Laplacian is too sensitive. m-SNE achieved better performance most of the time compared with the other methods.

VI. Conclusion

In this paper, we have proposed m-SNE for learning low-dimensional data from multiview high-dimensional data. Compared with traditional methods, the algorithm operates on a probabilistic framework that meaningfully integrates different views. Our approach can automatically learn a combination coefficient of different views according to their contributions to the final embedding. Thus, this combination coefficient can exploit complementary information in different views and suppress noise at the same time. In optimizing over combination coefficients, we employ Nesterov's gradient accelerating scheme and achieve a convergence rate of $O(1/k^2)$.

TABLE I
RESULTS FOR THE TOP 20 SCENES

Scenes	ASV	BSV	CAT	DSE	mSNE	Scenes	ASV	BSV	CAT	DSE	mSNE
greenn house	30.00	50.00	15.00	5.00	**65.00**	winecellar	12.38	14.29	4.76	4.76	**23.81**
pantry	25.00	15.00	20.00	5.00	**40.00**	restr kitchen	7.83	8.70	0.00	4.35	**21.74**
tv studio	32.22	33.33	33.33	0.00	**38.89**	train station	16.00	10.00	15.00	5.00	**20.00**
casino	10.53	26.32	0.00	0.00	**36.84**	elevator	8.57	9.52	**19.05**	0.00	**19.05**
pool inside	18.00	20.00	5.00	0.00	**35.00**	office	10.48	14.29	0.00	0.00	**19.05**
toy store	24.55	22.73	4.55	18.18	**31.82**	warehouse	13.33	14.29	9.52	0.00	**19.05**
movietheater	19.00	20.00	10.00	5.00	**30.00**	florist	15.79	**31.58**	5.26	0.00	15.79
dining room	7.78	11.11	11.11	0.00	**27.78**	cloister	12.00	**20.00**	**20.00**	0.00	15.00
music studio	13.68	10.53	15.79	10.53	**26.32**	restaurant	8.00	**15.00**	0.00	10.00	**15.00**
mall	7.00	10.00	0.00	10.00	**25.00**	waiting room	5.71	9.52	**14.29**	0.00	**14.29**

APPENDIX
PROOF OF LEMMA 2

Proof: For some $k = 0$, the relations are trivially satisfied. Suppose $\mathcal{R}_k^1$ and $\mathcal{R}_k^2$ hold for some $k \geq 0$. We have

$$\begin{aligned}\psi_{k+1}(\alpha) &\leq B_k g(\alpha) + \frac{1}{2}\|\alpha - \alpha_0\|^2 \\ &\quad + b_{k+1}\left[g(\alpha_{k+1}) + \langle \nabla g(\alpha_{k+1}), \alpha - \alpha_{k+1}\rangle\right] \\ &\leq (B_k + b_{k+1}) g(\alpha) + \frac{1}{2}\|\alpha - \alpha_0\|^2 \end{aligned} \tag{26}$$

and this is $\mathcal{R}_k^2$.

Denote $z_k = \arg\min_\alpha \psi_k(\alpha)$. Since $\psi_k(\alpha)$ is strongly convex with convexity number 1 and because of $\mathcal{R}_k^1$, we have

$$\psi_k(\alpha) \geq \psi_k^* + \frac{1}{2}\|\alpha - z_k\|^2 \geq B_k g(\alpha_k) + \frac{1}{2}\|\alpha - z_k\|^2. \tag{27}$$

Therefore

$$\begin{aligned}\psi_{k+1}(\alpha) &= \psi_k(\alpha) + b_{k+1}\left[g(\alpha_{k+1}) + \langle \nabla g(\alpha_{k+1}), \alpha - \alpha_{k+1}\rangle\right] \\ &\geq B_k g(\alpha_k) + \frac{1}{2}\|\alpha - z_k\|^2 \\ &\quad + b_{k+1}\left[g(\alpha_{k+1}) + \langle \nabla g(\alpha_{k+1}), \alpha - \alpha_{k+1}\rangle\right] \\ &\geq (B_k + b_{k+1}) g(\alpha_{k+1}) \\ &\quad + B_k \langle \nabla g(\alpha_{k+1}), \alpha_k - \alpha_{k+1}\rangle \\ &\quad + b_{k+1} \langle \nabla g(\alpha_{k+1}), \alpha - \alpha_{k+1}\rangle \\ &\quad + \frac{1}{2}\|\alpha - z_k\|^2. \end{aligned} \tag{28}$$

According to the update rule (15) in the algorithm, this is equivalent to

$$\begin{aligned}\psi_{k+1}(\alpha) &\geq B_{k+1} g(\alpha_{k+1}) + \frac{1}{2}\|\alpha - z_k\|^2 \\ &\quad + \langle \nabla g(\alpha_{k+1}), B_{k+1} y_k - b_{k+1} z_k - B_k \alpha_{k+1}\rangle \\ &\quad + b_{k+1} \langle \nabla g(\alpha_{k+1}), \alpha - \alpha_{k+1}\rangle \\ &\geq B_{k+1} g(\alpha_{k+1}) + B_{k+1} \langle \nabla g(\alpha_{k+1}), y_k - \alpha_{k+1}\rangle \\ &\quad + b_{k+1} \langle \nabla g(\alpha_{k+1}), \alpha - z_k\rangle + \frac{1}{2}\|\alpha - z_k\|^2. \end{aligned} \tag{29}$$

Thus, we obtain the inequality

$$\begin{aligned}\psi_{k+1}^* = \min_\alpha \psi_{k+1}(\alpha) &\geq B_{k+1} g(\alpha_{k+1}) - \frac{b_{k+1}^2}{2}\|\nabla g(\alpha_{k+1})\|^2 \\ &\quad + B_{k+1}\langle \nabla g(\alpha_{k+1}), y_k - \alpha_{k+1}\rangle. \end{aligned} \tag{30}$$

From our termination criterion (18), we have

$$\langle \nabla g(\alpha_{k+1}), y_k - \alpha_{k+1}\rangle \geq \frac{1}{M_k}\|\nabla g(\alpha_{k+1})\|^2. \tag{31}$$

Note that we choose b_{k+1} by the following equation:

$$B_{k+1} = B_k + b_{k+1} = \frac{M_k b_{k+1}^2}{2}. \tag{32}$$

Thus, $\mathcal{R}_{k+1}^1$ is also satisfied. This completes the proof. ■

REFERENCES

[1] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 2002.

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

[3] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press, 2001, pp. 585–591.

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

[5] G. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press, 2003, pp. 833–840.

[6] T. Zhang, K. Huang, X. Li, J. Yang, and D. Tao, "Discriminative orthogonal neighborhood-preserving projections for classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 1, pp. 253–263, Feb. 2010.

[7] T. Zhang, B. Fang, Y. Y. Tang, Z. Shang, and B. Xu, "Generalized discriminant analysis: A matrix exponential approach," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 1, pp. 186–197, Feb. 2010.

[8] 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.

[9] C. Chen, J. Zhang, and R. Fleischer, "Distance approximating dimension reduction of Riemannian manifolds," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 1, pp. 208–217, Feb. 2010.

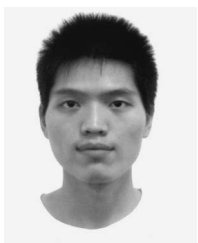
[10] X. He, "Laplacian regularized *d*-optimal design for active learning and its application to image retrieval," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 254–263, Jan. 2010.

[11] D. Cai, X. Wang, and X. He, "Probabilistic dyadic data analysis with local and global consistency," in *Proc. 26th Annu. ICML*, 2009, pp. 105–112.

[12] D. Cai, X. He, X. Wang, H. Bao, and J. Han, "Locality preserving nonnegative matrix factorization," in *Proc. 21st IJCAI*, 2009, pp. 1010–1015.

[13] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 6, pp. 1438–1446, Dec. 2010.

[14] T. Zhou, D. Tao, and X. Wu, "Manifold elastic net: A unified framework for sparse dimension reduction," *Data Mining Knowl. Discov.*, 2010, DOI: 10.1007/s10618-010-0182-x.
[15] W. Bian and D. Tao, "Max–min distance analysis by using sequential SDP relaxation for dimension reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, DOI: 10.1109/TPAMI.2010.189.
[16] S. Si, D. Tao, and B. Geng, "Bregman divergence based regularization for transfer subspace learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 929–942, Jul. 2010.
[17] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 174–184, Jan. 2010.
[18] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.
[19] 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.
[20] W. Bian and D. Tao, "Manifold regularization for SIR with rate root-n convergence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1–8.
[21] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Program.*, vol. 103, no. 1, pp. 127–152, May 2005.
[22] B. Xie, Y. Mu, and D. Tao, "m-SNE: Multiview stochastic neighbor embedding," in *Proc. Int. Conf. Neural Inf. Process.*, 2010, vol. 1, pp. 388–396.
[23] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[24] Z. Yang, I. King, Z. Xu, and E. Oja, "Heavy-tailed symmetric stochastic neighbor embedding," in *Proc. Adv. Neural Inf. Process. Syst. 22*, 2009, pp. 2169–2177.
[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge Univ. Press, 2004.
[26] Y. Nesterov, "Gradient methods for minimizing composite objective function," Center Oper. Res. Econometrics, Catholic Univ. Louvain (UCL), Louvain-la-Neuve, Belgium, Tech. Rep. 76, 2007.
[27] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. 26th Annu. ICML*, 2009, pp. 457–464.
[28] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, May 1994.
[29] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
[30] B. Long, P. S. Yu, and Z. M. Zhang, "A general model for multiple view unsupervised learning," in *Proc. SIAM Int. Conf. Data Mining*, Atlanta, GA, 2008, pp. 822–833.
[31] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world Web image database from National University of Singapore," in *Proc. ACM CIVR*, Santorini, Greece, Jul. 8–10, 2009.
[32] G. Griffin, A. Holub, and P. Perona, *Caltech-256 object category dataset*, California Inst. Technol., Pasadena, CA, Tech. Rep. 7694, 2007. [Online]. Available: http://authors.library.caltech.edu/7694.
[33] M. Li, "Texture moment for content-based image retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2007, pp. 508–511.
[34] J. Yang, K. Yu, Y. Gong, and T. S. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE CVPR*, 2009, pp. 1794–1801.
[35] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 413–420.



Bo Xie received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2009.

He is currently a Research Assistant with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong. From August 2009 to December 2010, he was a Research Assistant with the School of Computer Engineering at the Nanyang Technological University, Singapore. His current research interests include machine learning, computer vision, medical imaging, and convex optimization.

Yang Mu received the B.S degree from Jilin University, Changchun, China, in 2008. He is currently working toward the Ph.D. degree at the University of Massachusetts, Boston.

From December 2008 to June 2010, he was a Research Assistant with the School of Computer Engineering, Nanyang Technological University, Singapore. His current research interests include computer vision, data mining, and machine learning.

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

He is currently a Professor with the Centre for Quantum Computation and Information Systems and the Faculty of Engineering and Information Technology, University of Technology, Sydney, Broadway, Australia. He mainly applies statistics and mathematics for data analysis problems in data mining, computer vision, machine learning, multimedia, and video surveillance. He has authored and coauthored more than 100 scientific articles at top venues, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON IMAGE PROCESSING, and *Advances in Neural Information Processing Systems*, with best paper awards.

Kaiqi Huang (M'05–SM'09) received the M.Sc. degree from the Nanjing University of Science and Technology, Nanjing, China, and the Ph.D. degree from Southeast University, Nanjing.

During 2004 to 2005, he was a Postdoctoral Fellow with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China, where he has been an Associate Professor since 2005. He has published nearly 70 papers on major international journals and conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Pattern Recognition*, *Computer Vision and Image Understanding*, IEEE Computer Vision and Pattern Recognition (CVPR), and International Conference on Pattern Recognition. His current research interests include visual surveillance, image processing, and pattern recognition.

Dr. Huang is the Deputy Secretary of IEEE Beijing Section, the local Chair of the Fifth International Visual Surveillance Workshop in conjunction with the 2005 International Conference on Computer Vision, and a Committee Member of the Sixth and Seventh International Visual Surveillance Workshop in conjunction with the 2006 European Conference on Computer Vision, and CVPR'07.