Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Learning a hierarchical image manifold for Web image classification*

Rong ZHU[†1], Min YAO[†2], Li-hua YE[1], Jun-ying XUAN[1]

([1]*School of Information Engineering, Jiaxing University, Jiaxing 314001, China*)

([2]*School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*)

[†]E-mail: sikexing@163.com; myao@zju.edu.cn

Received Feb. 12, 2012; Revision accepted July 31, 2012; Crosschecked Sept. 11, 2012

Abstract: Image classification is an essential task in content-based image retrieval. However, due to the semantic gap between low-level visual features and high-level semantic concepts, and the diversification of Web images, the performance of traditional classification approaches is far from users' expectations. In an attempt to reduce the semantic gap and satisfy the urgent requirements for dimensionality reduction, high-quality retrieval results, and batch-based processing, we propose a hierarchical image manifold with novel distance measures for calculation. Assuming that the images in an image set describe the same or similar object but have various scenes, we formulate two kinds of manifolds, object manifold and scene manifold, at different levels of semantic granularity. Object manifold is developed for object-level classification using an algorithm named extended locally linear embedding (ELLE) based on intra- and inter-object difference measures. Scene manifold is built for scene-level classification using an algorithm named locally linear submanifold extraction (LLSE) by combining linear perturbation and region growing. Experimental results show that our method is effective in improving the performance of classifying Web images.

Key words: Web image classification, Manifold learning, Image manifold, Semantic granularity, Distance measure
doi:10.1631/jzus.C1200032 **Document code:** A **CLC number:** TP391

1 Introduction

Nowadays, computer vision research generally requires a large amount of digital images. The rapid development of image acquisition and processing technology has encouraged more and more digital images to be generated and further spread on the World Wide Web. Specifically, some online photo sharing platforms like Flickr (http://www.flickr.com) are reporting thousands of uploaded images per minute, and providing users more facilities for accessing, disseminating, and exchanging Web images (Ames and Naaman, 2007; Li LJ *et al.*, 2010; Sun *et al.*, 2011). Image classification plays an essential role in content-based image retrieval. However, Web images are usually difficult to classify, since most previous classification approaches result in reduced performance when they deal with large-scale and high-dimensional image data in real world applications. Therefore, Web image classification has been a promising yet challenging topic in computer vision research (Zhang, 2008; Zhou *et al.*, 2009; Luo *et al.*, 2010; Farajtabar *et al.*, 2011; Liu *et al.*, 2011; Parikh, 2011).

Image classification approaches can be roughly divided into four types (Lu and Weng, 2007): (1) Machine learning based: image classification can be seen as a process of finding the mappings between low-level visual features and high-level semantic concepts from the perspective of machine learning (Chang *et al.*, 2003; Kim DW *et al.*, 2007; Joshi *et al.*, 2009; Lin *et al.*, 2011). (2) Relevance feedback based:

* Project supported by the National High-Tech R & D Program (863) of China (No. 2009AA011900), the Zhejiang Provincial Natural Science Foundation of China (No. 2011Y1110960), and the Zhejiang Provincial Nonprofit Technology and Application Research Program of China (Nos. 2011C31045 and 2012C21020)

a certain relevance feedback scheme is applied to learn users' intentions during the classification process (Klaydios, 2004; Tao *et al.*, 2006; Cheng *et al.*, 2009; dos Santos *et al.*, 2011). (3) Ontology-based: domain-specific knowledge is used to create domain ontologies so that the semantic concepts of the images can be easily inferred (Jaimes and Smith, 2003; Vieux *et al.*, 2007; Chai *et al.*, 2009; Li XR *et al.*, 2010). (4) Object-based: according to human visual perception, the object region within an image that is important for describing image semantics is extracted, and the classification process is conducted based on the differences among objects. Some previous research shows that object-based classification has the advantage of simplifying the process of image classification and improving the classification accuracy, especially for those images having clear objects but a complex background (Luo *et al.*, 2004; Shao and Brady, 2006; Zeng *et al.*, 2009; El Sayad *et al.*, 2010). A common idea behind the above approaches is to shorten the well-known 'semantic gap' (Enser and Sandom, 2003; Wang CH *et al.*, 2008). Moreover, it is necessary to develop a novel classification method for Web images to satisfy some urgent requirements as follows:

1. Dimensionality reduction for high-dimensional visual features. The fusion of multiple visual features improves the classification performance, but it also brings high-dimensional visual features. Researchers have paid more attention to finding the mappings between visual features and semantic concepts, or to developing some effective learning machines for image representation. However, the side effect of 'dimensionality curse' (Bellman, 1961), which is caused by high-dimensional visual features, has often been ignored. Thus, most classifiers are not scalable enough for dealing with Web images. Consequently, dimensionality reduction for high-dimensional visual features should be regarded as a necessary step for image classification.

2. Users' demands for high-quality retrieval results. Finding users' intended items from the sea of Web images remains a difficult task. For example, to obtain good retrieval results in an image retrieval system, a certain relevance feedback scheme is applied to prompt users to make feedback on the returned results with several relevant levels. But this has proven tedious and time consuming (Rui *et al.*, 1998). On the other hand, people often have little patience to browse the images beyond the top three to five pages (Datta *et al.*, 2008). In addition, in many cases, users submit some special tags for image retrieval (e.g., 'a poodle', 'a brown poodle', or 'a brown poodle on lawn'). Clearly, global-based classification is not suitable for those requirements (Souvenir and Pless, 2005; Fan *et al.*, 2008).

3. Batch-based classification for unordered images. Nowadays, more and more users upload a group of images for sharing on the Internet. In many cases, the images in an image set share a common semantic concept (i.e., they describe the same or similar object but have various scenes). Obviously, if the subject of the classification is an image set, rather than just a single image (i.e., batch-based classification), the process of image classification will be greatly sped up. Moreover, different from the frames in a video sequence, although some similarities may exist, the images in an image set are generally unordered. Therefore, most frame-based approaches are not feasible for batch-based classification. The idea of batch-based classification first appeared in Wang RP *et al.* (2008). To the best of our knowledge, so far only a few studies have been published on this topic.

Manifold learning addresses the problem of seeking a low-dimensional manifold hidden in the high-dimensional data space, where the images in an image set are mapped into a nonlinear manifold while preserving their local similarities. Theoretically, given enough images, it is possible to reveal the intrinsic topological structure of the manifold, and thus much underlying useful information can be exploited without analyzing the high-dimensional image data. Previous work has shown that a nonlinear manifold provides a powerful structure for semantic representation, even if there is no convincing evidence of its accuracy. On the other hand, according to Seung and Lee (2000), the principle of manifold learning is more consistent with that of human visual perception.

In recent years, many state-of-the-art approaches based on manifold learning have been proposed, e.g., locally linear embedding (LLE) (Roweis and Saul, 2000), isometric mapping (Isomap) (Tenenbaum *et al.*, 2000), and Laplacian eigenmap (LE) (Belkin and Niyogi, 2001). These approaches have been widely used in various research fields, such as dimensionality reduction, face recognition, and image retrieval. In this paper, we try to develop a novel classification

method based on manifold learning. Specifically, we aim at constructing a hierarchical image manifold for representing and classifying Web images. However, since Web images are generally diversified, it is not reasonable to design the image manifold with a single framework. In addition, we can collect only a limited number of Web images in real world applications. Therefore, learning a hierarchal image manifold and then recovering its intrinsic topological structure, is not a trivial task.

The major contributions of our work are as follows: (1) Transform the classification in the high-dimensional data space to one on the low-dimensional image manifold with novel distance measures for calculation; (2) Construct a hierarchical image manifold in view of human visual perception by combing semantic granularity; (3) Formulate two kinds of manifold, object manifold and scene manifold. The former is built for object-level classification using an extended locally linear embedding (ELLE) algorithm, taking into account both intra- and inter-object distance measures, while the latter is set up for scene-level classification using a locally linear submanifold extraction (LLSE) algorithm, via linear perturbation and region growing.

Note that our method has some similarities with the method proposed by Wang RP *et al.* (2008); i.e., both aim at calculating the distances between two points lying on a nonlinear manifold. But they have great differences in their schemes for constructing a nonlinear manifold. In the method proposed by Wang RP *et al.* (2008), only one kind of manifold is defined for face recognition, and the manifold is built in a local coordinate system, while in our method, a hierarchical image model including an object manifold and a scene manifold is formulated for image classification, and these two kinds of manifold are set up in one global coordinate system. Moreover, we adopt the ideas of some previous subspace schemes and spectrum-based approaches to recover the intrinsic topological structure of the image manifold; thus, there may be some relationships between our method and others, including the methods proposed by Huang *et al.* (2003), Gao and Fan (2005), and Kim *et al.* (2010). At last, note that the proposed model is sensitive to the result of object extraction, and that it cannot deal with the landscape images without clear objects.

2 Image manifold construction

2.1 Motivation

Based on our observations, we found that the significant differences among Web images are caused mainly by the variations in imaging conditions (e.g., illumination, viewing direction, and camera features), the complexity of an object, and the changing scenes. For example, two images related to different semantic concepts (i.e., they describe different objects) are very similar in content (Figs. 1a and 1b). On the contrary, two images that are irrelevant based on global features may share a common semantic concept (i.e., they describe the same or similar object) (Figs. 1c and 1d). Fortunately, according to the photography principle (Patterson, 1986), there generally exists a particular area within an image where the eyes are attracted. In other words, an image can be segmented into two parts: the foreground (i.e., an object region, which represents the semantic concept of an image) and the background (i.e., a scene region, which represents its imaging environment). Therefore, developing a hierarchical image manifold by combining semantic granularity (e.g., object- and scene-oriented semantics) (Jaimes *et al.*, 1999) and classifying the images at different semantic levels (e.g., object-level and scene-level), the classification process will be simplified and the classification accuracy improved. Obviously, this strategy agrees with the coarse-to-fine processing of human visual perception for image understanding.

Fig. 1 Some samples of Web images
(a) ‘Woman’; (b) ‘Dog’; (c) ‘Tulip’; (d) ‘Tulip’

2.2 Object manifold definition

Assume there is a group of images X in the high-dimensional data space $\mathbb{R}^D$, $X=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N | \boldsymbol{x}_i=[x_i^1, x_i^2, \ldots, x_i^D]^{\mathrm{T}}, i=1, 2, \ldots, N\}$, where N is the number of images in X. Suppose that X consists of C image sets, i.e., $X=\{X_1, X_2, \ldots, X_C\}$, and that each set X_i ($i=1, 2, \ldots, C$) includes N_i images describing the same or similar object but having different scenes, where $N=\sum_{i=1}^{C} N_i$.

Definition 1 An object manifold M_{O} is constructed for X, defined as a combination of multiple nonlinear manifolds, i.e., $M_{\mathrm{O}}=\{M_{\mathrm{O},1}, M_{\mathrm{O},2}, \ldots, M_{\mathrm{O},C}\}$. Each nonlinear manifold $M_{\mathrm{O},i}$ ($i=1, 2, \ldots, C$) is corresponding to X_i, and any two nonlinear manifolds are disjoint, i.e., $M_{\mathrm{O},i} \cap M_{\mathrm{O},j}=\varnothing$ ($i, j=1, 2, ..., C$; $i \neq j$).

We attempt to construct a suitable object manifold and to find a good embedding for X in the low-dimensional space, not only preserving the local similarities of the images in X, but also maintaining the whole topological structure of X. Some approaches to tackling the problem of learning multiple nonlinear manifolds have been presented over the last few years. There are usually two solutions: (1) Each nonlinear manifold is mapped separately into a local coordinate system (Saul and Roweis, 2003; Lu and Weng, 2007); (2) All nonlinear manifolds are projected into one global coordinate system (Yang, 2002; de Ridder *et al.*, 2003; Wu and Chan, 2004; Pillati and Viroli, 2005; Jun and Ghosh, 2010). Although the former has proven effective in learning multiple nonlinear manifolds, it lacks a uniform coordinate system to represent images. We propose a novel algorithm named extended locally linear embedding (ELLE) using path-based clustering (Fischer and Buhmann, 2003) based on intra- and inter-object distance measures.

Definition 2 Suppose that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are two images belonging to the ith image set X_i in the high-dimensional data space $\mathbb{R}^D$, and that they are mapped into one nonlinear manifold $M_{\mathrm{O},i}$. Assuming that there is a set of paths P between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, $P(\boldsymbol{x}_1, \boldsymbol{x}_2)=\{p_1(\boldsymbol{x}_1, \boldsymbol{x}_2), p_2(\boldsymbol{x}_1, \boldsymbol{x}_2), \ldots, p_l(\boldsymbol{x}_1, \boldsymbol{x}_2)\}$, and that there exist q_k small edge elements on each path $p_k(\boldsymbol{x}_1, \boldsymbol{x}_2)$ ($k=1, 2, \ldots, l$), the distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is defined as

$$d_P(\boldsymbol{x}_1, \boldsymbol{x}_2)=\min\left\{\max_{1\le e\le q_1+1} d_e, \max_{1\le e\le q_2+1} d_e, \cdots, \max_{1\le e\le q_l+1} d_e\right\}, \tag{1}$$

where e denotes each edge element on one path, and d_e denotes the weight of edge e. Thus, the intra-object distance measure between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ can be formulated as

$$d_{\mathrm{intra}}(\boldsymbol{x}_1, \boldsymbol{x}_2)=d_P(\boldsymbol{x}_1, \boldsymbol{x}_2). \tag{2}$$

Definition 3 Suppose that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are two images belonging to the ith image set X_i and the jth image set X_j respectively in the high-dimensional data space $\mathbb{R}^D$, and that they are mapped into the nonlinear manifold $M_{\mathrm{O},i}$ and the nonlinear manifold $M_{\mathrm{O},j}$, respectively. Assume that $\boldsymbol{x}_i'$ and $\boldsymbol{x}_j'$ are two boundary points taking the largest Euclidean distance between X_i and X_j, as follows:

$$d_{\mathrm{E}}(\boldsymbol{x}_i', \boldsymbol{x}_j')=\max_{\boldsymbol{x}_i'\in X_i, \boldsymbol{x}_j'\in X_j} d_{\mathrm{E}}\{X_i, X_j\}. \tag{3}$$

Then the inter-object distance measure between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ can be formulated as

$$\begin{aligned} d_{\mathrm{inter}}(\boldsymbol{x}_1, \boldsymbol{x}_2)&=d_P(\boldsymbol{x}_1, \boldsymbol{x}_i')+d_{\mathrm{E}}(\boldsymbol{x}_i', \boldsymbol{x}_j')+d_P(\boldsymbol{x}_j', \boldsymbol{x}_2),\\ &i, j=1, 2, ..., C;\ i\neq j. \end{aligned} \tag{4}$$

The main steps of ELLE can be summarized as follows.

Step 1: Suppose M_{O} is a d-dimensional object manifold constructed for X, $d<<D$. For each image $\boldsymbol{x}_i$ ($i=1, 2, \ldots, N$) in X, compute the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ ($j=1, 2, \ldots, N$; $i\neq j$) based on the intra- and inter-object distance measures (Eqs. (1)–(4)). If there exist s_i images that satisfy a certain condition (k-nearest neighbor or r-radius), these images are chosen as $\boldsymbol{x}_i$'s nearest neighbors.

Step 2: Calculate the reconstruction weight $w_{i,j}$ for $\boldsymbol{x}_i$ using the s_i nearest neighbors. Minimize the reconstruction error function as follows:

$$\begin{aligned} \varepsilon(\boldsymbol{W})&=\arg\min\sum_{i=1}^{N}\left|\boldsymbol{x}_i-\sum_{j=1}^{s_i} w_{i,j}\boldsymbol{x}_{i,j}\right|^2\\ &=\arg\min\sum_{i=1}^{N}\sum_{j=1}^{s_i}\sum_{k=1}^{s_i} w_{i,j}w_{i,k}\boldsymbol{Q}^i\\ &\text{s.t.}\quad \sum_{j=1}^{s_i} w_{i,j}=1, \end{aligned} \tag{5}$$

where $\boldsymbol{W}=(w_{i,j})_{N\times N}$ is a reconstruction weight matrix, and $\boldsymbol{Q}^i$ is a symmetric, semi-positive covariance

matrix, i.e., $\boldsymbol{Q}^i=(\boldsymbol{x}_i-\boldsymbol{x}_{i,j})(\boldsymbol{x}_i-\boldsymbol{x}_{i,k})$, where $\boldsymbol{x}_{i,j}$ and $\boldsymbol{x}_{i,k}$ are the nearest neighbors of $\boldsymbol{x}_i$, $\boldsymbol{x}_{i,j}\neq\boldsymbol{x}_{i,k}$ ($1\leq j, k\leq s_i; j\neq k$).

Step 3: Assume $Y=\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$ is the embedding of $X=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ on M_{O}. To obtain Y, minimize the cost function as follows:

$$\begin{aligned}&\varepsilon(Y)=\arg\min\sum_{i=1}^{N}\left|\boldsymbol{y}_i-\sum_{j=1}^{s_i}w_{i,j}\boldsymbol{y}_{i,j}\right|^2\\&\text{s.t.}\quad\sum_{i=1}^{N}\boldsymbol{y}_i=\boldsymbol{0},\quad\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_i\boldsymbol{y}_{i,j}{}^{\mathrm{T}}=\boldsymbol{I},\end{aligned}\tag{6}$$

where $\boldsymbol{y}_{i,j}$ ($1\leq j\leq s_i$) is $\boldsymbol{y}_i$'s nearest neighbor. Finally, Eq. (6) can be calculated by finding the d eigenvectors with the smallest (nonzero) eigenvalues of the cost matrix $\boldsymbol{M}=(\boldsymbol{I}-\boldsymbol{W})^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{W})$.

The flow of ELLE is shown in Algorithm 1.

Algorithm 1 Extended locally linear embedding (ELLE)

Input: a group of images X in the high-dimensional data space $\mathbb{R}^D$, $X=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N|\boldsymbol{x}_i=[x_i^1, x_i^2, \ldots, x_i^D]^{\mathrm{T}}, i=1, 2, ..., N\}$; C image sets X_i ($i=1, 2, \ldots, C$), $X=\{X_1, X_2, \ldots, X_C\}$; an object manifold M_{O} constructed for X; the dimensionality of M_{O}, d ($d<<D$).

Output: the embedding Y for X on M_{O}, $Y=\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N|\boldsymbol{y}_i=[y_i^1, y_i^2, \ldots, y_i^d]^{\mathrm{T}}, i=1, 2, ..., N\}$.

```
For each x_i∈X do {
  For each x_j∈X do {
    If (x_i∈X_i && x_j∈X_i) then
      Compute the distance between x_i and x_j:
        d_{x_i,x_j} = d_intra(x_i, x_j);          // Eqs. (1) and (2)
    If (x_i∈X_i && x_j∈X_j) then
      Compute the distance between x_i and x_j:
        d_{x_i,x_j} = d_inter(x_i, x_j); }        // Eqs. (3) and (4)
  Select the s_i nearest neighbors of x_i according to d_{x_i,x_j};
  Calculate the reconstruction weight w_{i,j};
  Establish the reconstruction weight matrix W;   // Eq. (5)
  Construct the cost matrix M according to
    M=(I−W)^T(I−W);
  Find the d eigenvectors with the smallest (nonzero)
    eigenvalues of M; }                           // Eq. (6)
Return {y_1, y_2, ..., y_N}.
```

In terms of the computational complexity of ELLE, we consider only the additional complexity generated by the calculation of the intra- and inter-object distance measures between pairs of the images. The processing cost is computed as follows:

$$\sum_{i=1}^{N}\sum_{j=1}^{N-1}1=N(N-1)=N^2-N.\tag{7}$$

Therefore, ELLE has an additional complexity of $O(N^2)$.

Obviously, ELLE has the capability of shortening the distance between two points lying on one nonlinear manifold (which are connected by some small edge elements), while enlarging the distance between two points lying on different nonlinear manifolds (which are linked by a long path). For simplification, assume that an image in the high-dimensional data space is corresponding to a point on the low-dimensional manifold. The advantages of ELLE can be summarized as follows: (1) It conforms to the local consistency of LLE. That is, the nearby images in the high-dimensional data space remain nearby and have high affinity on the object manifold. (2) It conforms to the global consistency of the clustering method. That is, the images describing the same or similar semantic concept (they belong to the same image set) are gathered together on the object manifold, whereas the images related to different semantic concepts (they belong to different image sets) are far away from each other on the object manifold. (3) Since neither the Euclidean distance nor the geodesic distance can be used to deal with the nonlinear manifold with a complex structure, especially when the structure of the nonlinear manifold with curved surfaces bears folding or bending (like a spiral) (Zhu and Yao, 2009), path-based clustering, a psychophysically plausible similarity measure, is applied to define novel distance measures, i.e., intra- and inter-object distance measures. Note that if the intrinsic dimensionality of each nonlinear manifold is different, or two nonlinear manifolds intersect or overlap, the overall topological structure of the object manifold may not be well recovered in one global coordinate system.

2.3 Scene manifold definition

Definition 4 A scene manifold $M_{\mathrm{S},i}$ is constructed for X_i ($i=1, 2, \ldots, C$), defined as an integration of a series of locally linear submanifolds, i.e., $M_{\mathrm{S},i}=\{S_{i,1}, S_{i,2}, \ldots, S_{i,C'}\}$ ($i=1, 2, \ldots, C$), where C' is the number of locally linear submanifolds. Each locally linear submanifold $S_{i,j}$ ($i=1, 2, \ldots, C$; $j=1, 2, \ldots, C'$) is corresponding to the images having the same or similar scene, and any two locally linear submanifolds

do not overlap, i.e., $S_{i,l} \cap S_{i,m}=\varnothing$ (i=1, 2, ..., C; l, m=1, 2, …, C'; $l \neq m$).

Different from the object manifold in which each nonlinear manifold is constructed for the images having the same semantic concept, a scene manifold is built for the images sharing one object semantic but having different scene semantics. Moreover, the number of possible scenes is unknown. Therefore, it is hard to reveal the distinct topological structure of the scene manifold. Fortunately, according to previous work, a nonlinear manifold can be viewed as an integration of a series of local regions (i.e., locally linear submanifolds). Therefore, if the topological structures of these locally linear submanifolds can be well found, the embeddings of the images will also be well recovered on the scene manifold.

To automatically extract a series of locally linear submanifolds from a nonlinear manifold, many approaches have been presented (Souvenir and Pless, 2005; Fan and Yeung, 2006; Zhai *et al.*, 2008). According to geometric intuition (Carlsson *et al.*, 2008; Briggs *et al.*, 2009), we propose a novel algorithm named locally linear submanifold extraction (LLSE) by combining linear perturbation and region growing. Here linear perturbation acts as a constraint condition, which is defined as the deviation of the Euclidean distance and the geodesic distance (Wang RP *et al.*, 2008). Specifically, each linear region is gradually stemmed from a seed point using a region growing method until the constraint condition is broken. At last, the maximum linear region can be obtained; it is just a locally linear submanifold. The above process is repeated until the points on the scene manifold are used up. Fig. 2 shows a scene manifold having three locally linear submanifolds.

The flow of ELLE is shown in Algorithm 2.

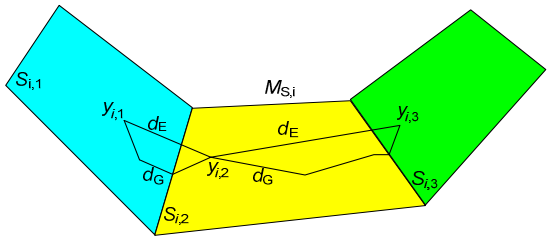

Fig. 2 A scene manifold having three locally linear submanifolds, i.e., $M_{S,i}=\{S_{i,1}, S_{i,2}, S_{i,3}\}$

$\boldsymbol{y}_{i,1}$, $\boldsymbol{y}_{i,2}$, and $\boldsymbol{y}_{i,3}$ are the embeddings for $\boldsymbol{x}_{i,1}$, $\boldsymbol{x}_{i,2}$, and $\boldsymbol{x}_{i,3}$ on $M_{S,i}$, respectively; d_E denotes the Euclidean distance; d_G denotes the geodesic distance

Algorithm 2 Locally linear submanifold extraction (LLSE)

Input: the ith image set X_i in the high-dimensional data space $\mathbb{R}^D$, $X_i=\{\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \ldots, \boldsymbol{x}_{i,N_i} | \boldsymbol{x}_{i,j}=[x_{i,j}^1, x_{i,j}^2, \ldots, x_{i,j}^D]^T, j=1, 2, ..., N_i\}$; a scene manifold $M_{S,i}$ constructed for X_i; the dimensionality of $M_{S,i}$, d ($d \ll D$); the embedding Y_i for X_i on $M_{S,i}$, $Y_i=\{\boldsymbol{y}_{i,1}, \boldsymbol{y}_{i,2}, \ldots, \boldsymbol{y}_{i,N_i} | \boldsymbol{y}_{i,j}=[y_{i,j}^1, y_{i,j}^2, \ldots, y_{i,j}^d]^T, j=1, 2, ..., N_i\}$.

Output: a series of locally linear submanifolds, $S_{i,1}, S_{i,2}, \ldots, S_{i,C'}$ (C' is the number of submanifolds).

```
1  l=1, S_{i,l}=∅;
2  While Y_i≠∅ do {
3     Randomly select a point, denoted as y_{i,l,1},
        which acts as a seed point for region growing;
4     Insert y_{i,l,1} into S_{i,l}: S_{i,l}=S_{i,l}∪{y_{i,l,1}};
5     Delete y_{i,l,1} from Y_i: Y_i=Y_i\{y_{i,l,1}};
6     k=1;
7     For each y_{i,j}∈Y_i do {
8        If d_G(y_{i,l,k}, y_{i,j})−d_E(y_{i,l,k}, y_{i,j})≤Th then {
           // Th denotes a threshold
9           Insert y_{i,j} into S_{i,l}: S_{i,l}=S_{i,l}∪{y_{i,j}};
10          Delete y_{i,j} from Y_i: Y_i=Y_i\{y_{i,j}};
11          k=k+1;
12          y_{i,j} is denoted as y_{i,l,k}; }}
13    l=l+1, S_{i,l}=∅; }
14 Return {S_{i,1}, S_{i,2}, …, S_{i,C'}}.
```

Similar to the evaluation of the processing cost of ELLE, in LLSE, only the additional complexity generated by the distance calculation should be considered. Therefore, LLSE also has an additional complexity of $O(N_i^2)$.

Note that although both the object manifold and scene manifold depend on the calculation of distance measures, there still exist great differences between them, mainly in the following two aspects: (1) The object manifold is built for a group of image sets with various semantic concepts, and each nonlinear manifold is corresponding to an image set having one semantic concept; thus, based on prior knowledge, the distance between two points on one nonlinear manifold can be shortened, whereas the distance between two points from different nonlinear manifolds can be enlarged, which results in good performance in multiple manifold learning. The scene manifold is set up for an image set having several scenes, and since the number of scenes is unknown, no prior knowledge can be used to define the distance measure. (2) For the object manifold, ELLE is conducted in the high-dimensional data space, and the finding of each nonlinear manifold can be regarded as a process of reducing the dimensionality of image data. For the

scene manifold, LLSE is conducted in the low-dimensional space, and extracting each locally linear submanifold can be viewed as a process of performing a region growing method.

Finally, a hierarchical image manifold (IM) is represented as a tree structure which can be described by Fig. 3. At the object semantic level, an object manifold M_O is divided into a set of nonlinear manifolds, $\{M_{O,1}, M_{O,2}, \ldots, M_{O,C}\}$. Each nonlinear manifold $M_{O,i}$ (i=1, 2, …, C) is corresponding to a scene manifold $M_{S,i}$ (i=1, 2, …, C). At the scene semantic level, a scene manifold $M_{S,i}$ is further divided into a series of locally linear submanifolds, $\{S_{i,1}, S_{i,2}, \ldots, S_{i,C_i'}\}$. Note that in this model the object manifold is constructed earlier than the scene manifold, since scene semantic is more abstract than object semantic from the perspective of human cognition. Moreover, its object semantic of an image is usually more important than its scene semantic. Although the above hierarchical image manifold is proposed for the classification of Web images, considering only the users' urgent demands and the characteristics of Web images, it is not enough to classify the images in real world applications. Actually, the proposed model can be further extended. For example, at the object semantic level, when the number of the images in an image set is very large, or the described object has a diversified form, the object manifold will be composed of more levels and each object class can be further specified; e.g., an object class 'tiger' is specified to 'brown tiger' and 'white tiger'.

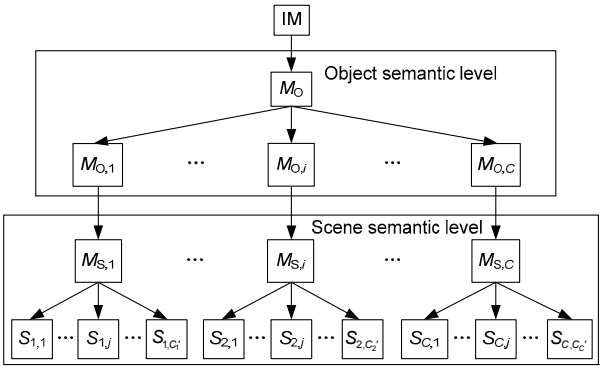

Fig. 3 A hierarchical image manifold (IM)
M_O is an object manifold; $M_{O,1}$, $M_{O,i}$, and $M_{O,C}$ are nonlinear manifolds; $M_{S,1}$, $M_{S,i}$, and $M_{S,C}$ are scene manifolds; $S_{1,1}$, $S_{1,j}$, $S_{1,C_1'}$, $S_{2,1}$, $S_{2,j}$, $S_{2,C_2'}$, $S_{C,1}$, $S_{C,j}$, and $S_{C,C_C'}$ are locally linear submanifolds

3 Image classification

3.1 Object-level classification

Given a test set T that includes N_t images in the high-dimensional data space $\mathbb{R}^D$, i.e., $T=\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{N_t} | \boldsymbol{t}_i=[t_i^1, t_i^2, \ldots, t_i^D]^T, i=1, 2, \ldots, N_t\}$, and supposing that an object class corresponds to one semantic concept (i.e., object semantic), the object class identity k_i of each image $\boldsymbol{t}_i$ (i=1, 2, …, N_t) can be obtained as

$$k_i = p(\boldsymbol{t}_i, M_O) = \arg\max p(\boldsymbol{t}_i, M_{O,j}), \quad i=1,2,\ldots,N_t;\ j=1,2,\ldots,C, \tag{8}$$

where M_O denotes an object manifold constructed for a training set X, $M_{O,j}$ is the jth nonlinear manifold in M_O, and $p(\boldsymbol{t}_i, M_O)$ denotes the probability of $\boldsymbol{t}_i$ lying on M_O. To be more exact, $p(\boldsymbol{t}_i, M_{O,j})$ denotes the probability of $\boldsymbol{t}_i$ lying on a certain nonlinear manifold $M_{O,j}$ (j=1, 2, …, C).

To calculate $p(\boldsymbol{t}_i, M_{O,j})$, first the embedding of Y for X on M_O is obtained by conducting ELLE. Then, the points in Y are gathered together into different clusters using k-means clustering, and a centroid $\boldsymbol{u}_{O,j}$ (j=1, 2, …, C) is calculated to represent each nonlinear manifold $M_{O,j}$. Meanwhile, an object manifold $M_{O,T}$ (i.e., a nonlinear manifold) is constructed for T by applying LLE, and the embedding $\boldsymbol{t}_i'$ (i=1, 2, …, N_t) of each $\boldsymbol{t}_i$ is obtained on $M_{O,T}$. A simple strategy to calculate $p(\boldsymbol{t}_i, M_{O,j})$ is to compute the distance between $\boldsymbol{t}_i'$ and $\boldsymbol{u}_{O,j}$, as follows:

$$p(\boldsymbol{t}_i, M_{O,j}) = \min d_{L2}(\boldsymbol{t}_i', \boldsymbol{u}_{O,j}), \quad i=1,2,\ldots,N_t;\ j=1,2,\ldots,C, \tag{9}$$

where $d_{L2}(\boldsymbol{t}_i', \boldsymbol{u}_{O,j})$ denotes the distance between $\boldsymbol{t}_i'$ and $\boldsymbol{u}_{O,j}$ using the L2 distance measure (de Juan and Bodenheimer, 2004).

Finally, the whole identity k_O of T can be determined via a majority voting based scheme. In fact, the images included in T usually describe the same or similar object (i.e., they belong to the same semantic concept); thus, the object class identity k_i of each image $\boldsymbol{t}_i$ does not need to be obtained, and then the implementation of object-level classification will be sped up. The whole identity k_O is determined as

$$\begin{aligned} k_O &= \arg\min d_{L2}(M_{O,T}, M_{O,j}) \\ &= \arg\min d_{L2}(\boldsymbol{u}_{O,T}, \boldsymbol{u}_{O,j}), \quad j=1,2,\ldots,C, \end{aligned} \tag{10}$$

where $\boldsymbol{u}_{O,T}$ denotes a centroid for $M_{O,T}$.

3.2 Scene-level classification

We suppose that the images within T are correctly classified into the ith object class during object-level classification, and $M_{S,T}$ is a scene manifold constructed for T. After performing LLSE, a series of locally linear submanifolds (i.e., $S_1, S_2, \ldots, S_{C''}$) can be extracted from $M_{S,T}$. Assume that $M_{S,i}$ is a scene manifold corresponding to $M_{O,i}$ (which is a nonlinear manifold constructed for the images included in the ith object class), and $S_{i,1}, S_{i,2}, \ldots, S_{i,C'}$ are several locally linear submanifolds generated from $M_{S,i}$. Then the scene semantic identity k_S of T can be determined by calculating the distances between all pairs of the submanifolds in $M_{S,T}$ and $M_{S,i}$:

$$k_S = \arg \min_{1 \le c_1 \le C''} \{ \min_{1 \le c_2 \le C'} d_S(S_{c_1}, S_{i,c_2}) \}. \quad (11)$$

To deal with the distance calculation between two subspaces, a feasible scheme is to apply certain exemplar-based or cluster mean based approaches (Wang L *et al.*, 2006; Kim TK *et al.*, 2007; Wang RP *et al.*, 2008), but when these subspaces have different dimensionalities, they may easily result in reduced performance. Therefore, the distance calculation between a d_1-dimensional submanifold S_1 and a d_2-dimensional submanifold S_2 is formulated as

$$d_S(S_1, S_2) = \left[\max(d_1, d_2) - \sum_{i=1}^{d_1} \sum_{j=2}^{d_2} (\boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{v}_j)^2 \right]^{1/2}, \quad (12)$$

where $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{d_1}$ and $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{d_2}$ are the orthonormal bases of S_1 and S_2, respectively.

Fig. 4 shows the process of classifying Web images based on the proposed hierarchical image manifold.

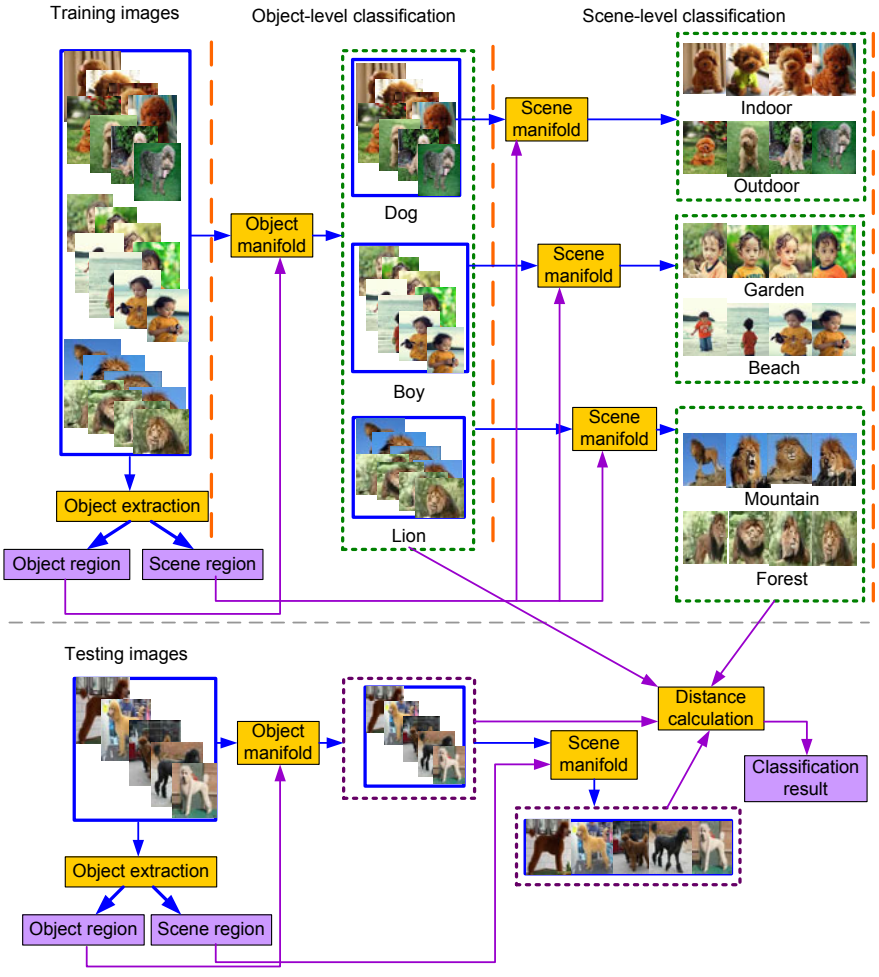

Fig. 4 The process of classifying Web images based on the proposed hierarchical image manifold

4 Experimental results

We designed some experiments to further evaluate the effectiveness of the proposed hierarchical image manifold. All images were in JPEG format with size of 126×189 or 189×126 pixels. Considering that object extraction is critical for object-level classification, object region within an image was segmented using an efficient method (Kang *et al.*, 2009), and then it was clipped out via a rectangular box.

As to the extraction of visual features, different from the images oriented to some special domains, such as medical images and remote sensing images, Web images are digital photos taken under natural scenes, and they usually have such characteristics as complex imaging conditions and rich content; thus, it is necessary to select some distinguishing features according to the requirements. Here a 514-dimensional visual feature was extracted (Vailaya *et al.*, 1998; Guo *et al.*, 2002; Rizon *et al.*, 2006). It includes color histogram (CH, 256), color coherence vector (CCV, 128), color moment (CM, 9), tamura feature (TF, 18), pyramidal wavelet transform (PWT, 24), edge direction histogram (EDH, 72), and shape invariant moment (SIM, 7). The CH, CCV, CM, TF, PWT, and SIM were used during object-level classification, while the CH, CCV, CM, TF, PWT, and EDH were used during scene-level classification. AdaBoost-based face detection (Viola and Jones, 2004) was applied to segment the face region within an image.

A good classifier should well distinguish the images from one class to the others. That is, the number of the images that are correctly classified into the *i*th class (they belong to the *i*th class) is as large as possible, while the number of the images that are wrongly classified into the *i*th class (they actually do not belong to the *i*th class) is as small as possible. Therefore, considering that there is still no uniform evaluation criterion that can be applied for Web image classification, we use two metrics, object classification accuracy (OCA) and false positive rate (FPR), to quantitatively evaluate the performance of the proposed ELLE. They are defined as follows:

$$\mathrm{OCA}_i = N_{i\_\mathrm{p}} / N_i, \tag{13}$$

$$\mathrm{FPR}_i = N_{\bar{i}\_\mathrm{f}} / N_{\bar{i}}, \tag{14}$$

where N_{i_p} denotes the number of the images that are correctly classified into the *i*th class, N_i denotes the number of the images included in the *i*th class, $N_{\bar{i}\_\mathrm{f}}$ denotes the number of the images that are wrongly classified into the *i*th class (they actually do not belong to the *i*th class), and $N_{\bar{i}}$ denotes the number of the images not included in the *i*th class.

All experiments were conducted five times. In each collection, half of the images were used for training and the other half for testing. The hardware platform is a T2350 with 1.86 GHz CPU and 2 GB main memory.

Experiment 1 The collection consists of 1620 images downloaded from the publically available image dataset MIRFlickr (http://press.liacs.mirflickr) (Huiskes and Lew, 2008). These images were divided into six sets according to their tags (each set is corresponding to one class): 'Baby' (200), 'Bird' (200), 'Car' (220), 'Dog' (300), 'Flower' (400), and 'Food' (300). Fig. 5 shows some samples of the images in the collection.

We compared the performance of ELLE(path-based clustering) with other manifold learning-based ones, i.e., LLE(Euclidean distance), LLE(geodesic distance), and Isomap(geodesic distance) (Wu and Chan, 2004). Here some important parameters were set as follows: k=8, d=20. Note that both k and d are flexible parameters. Thus, the changes of their values may affect the result of ELLE. An optimum value relies on experience in most cases; e.g., the value of d is determined by preserving 95% data variances. Figs. 6 and 7 show the comparison of the classification performances among the four algorithms.

On one hand, among the six classes, both 'Baby' and 'Flower' achieve higher OCA. Since the images belonging to 'Baby' usually have a face region, AdaBoost-based face detection can play an important role in the recognition of the images belonging to this class. The images in 'Flower' have distinct visual characteristics, such as regular shapes; thus, the images belonging to this class can be easily distinguished from the images in other classes. Since the images included in 'Food' often have the diversity of color and texture features, and many images in 'Bird' are grayscale ones, the OCA of these two classes is lower than that of the other four classes. On the other hand, the FPR of 'Baby' is the lowest, and 'Car' comes second, while 'Food' yields relatively poor performance, i.e., the highest FPR, since some images

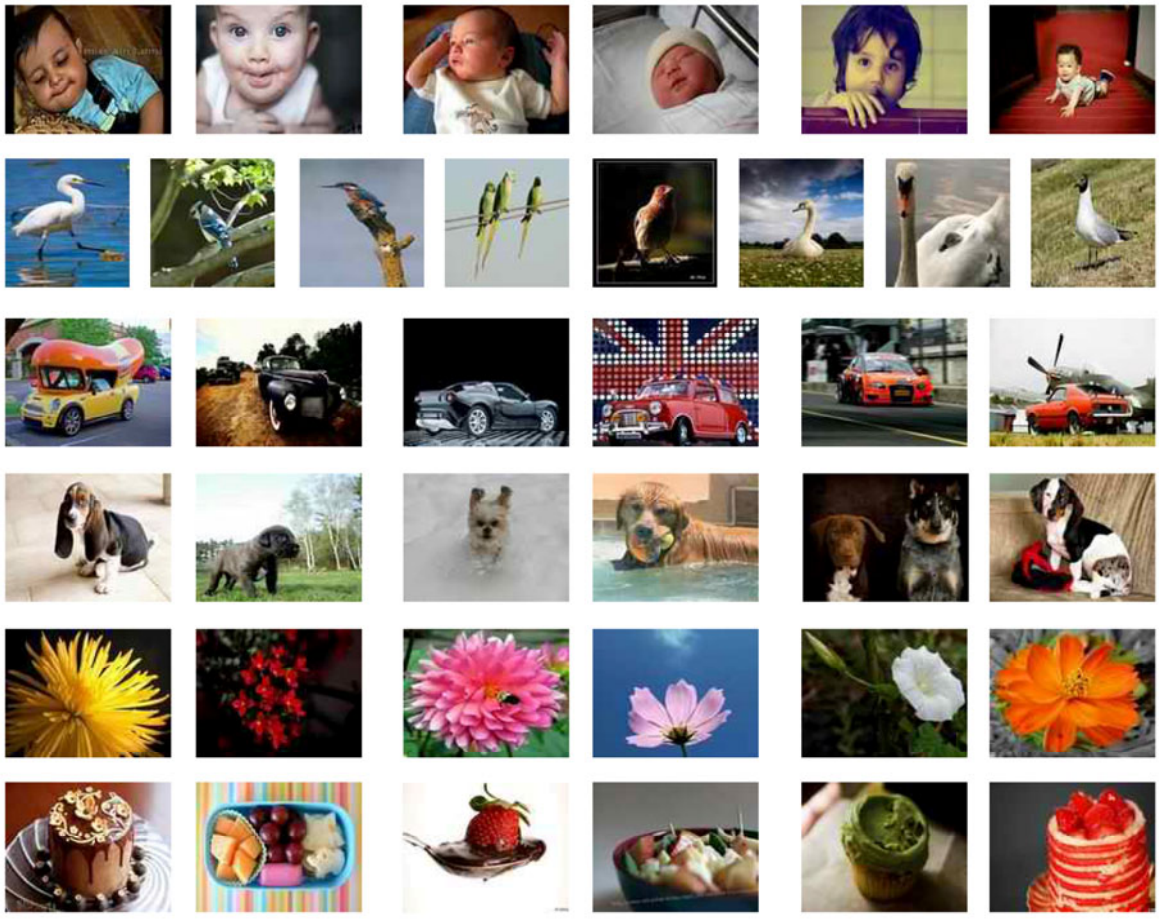

Fig. 5 Some samples of the images in the collection (each row corresponds to one class and they are 'Baby', 'Bird', 'Car', 'Dog', 'Flower', and 'Food' from top to bottom)

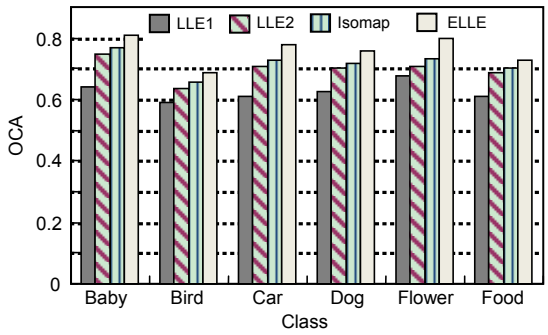

Fig. 6 Comparison of object classification accuracy (OCA) among the four algorithms

LLE1: LLE(Euclidean distance); LLE2: LLE(geodesic distance); Isomap: Isomap(geodesic distance); ELLE: ELLE (path-based clustering)

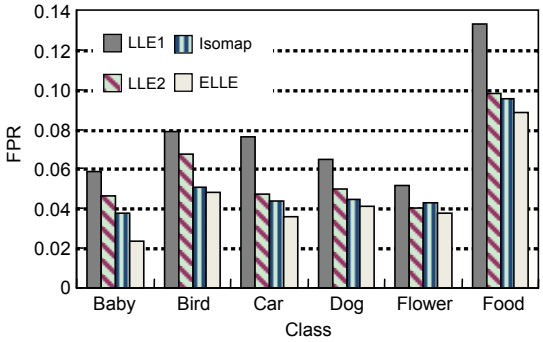

Fig. 7 Comparison of false positive rate (FPR) among the four algorithms

LLE1: LLE(Euclidean distance); LLE2: LLE(geodesic distance); Isomap: Isomap(geodesic distance); ELLE: ELLE (path-based clustering)

belonging to 'Flower' and 'Car' are easily misrecognized as 'Food'. Besides, some phenomena, such as unclear objects, fuzzy content, and inaccurate annotations, may also influence the performance of image classification. For example, the main object of an image is 'boy', not 'bird' (Fig. 8a); an image does not have a distinguishing object (Fig. 8b); the area 'bird' within an image is a decorative pattern (Fig. 8c).

(a) (b) (c)

Fig. 8 Some images belonging to 'Bird': (a) an image with the main object being a 'boy'; (b) an image without a distinguishing object; (c) an image with a decorative pattern 'bird'

As shown in Figs. 6 and 7, LLE(Euclidean distance) obtains poor performance, especially in dealing with the images in 'Bird' and 'Food' that usually contain larger variations; thus, Euclidean distance based algorithms may not work well for complex images in real world applications. The other three algorithms, LLE(geodesic distance), Isomap(geodesic distance), and ELLE(path-based clustering), obtain different results due to their respective properties. LLE(geodesic distance) results in the lowest OCA and the highest FPR. By integrating the properties of LLE and the clustering method in a novel way, the proposed ELLE achieves good performance.

The reasons for the better classification results may be as follows: (1) Different from LLE(Euclidean distance), the other three algorithms calculate the distance between two images based on either the geodesic distance or path-based clustering, and thus they can well recover the distinct topological structure of a nonlinear manifold. (2) Both LLE(Euclidean distance) and LLE(geodesic distance) fail to solve the problem of learning multiple nonlinear manifolds, so they construct each nonlinear manifold in different coordinate systems. In contrast, Isomap(geodesic distance) and ELLE(path-based clustering) tackle the consistency problem in multi-manifold learning; thus, both of them can obtain good low-dimensional embeddings for the images in a global coordinate system. (3) Since path-based clustering is applied to formulate novel distance measures for calculation, different from Isomap(geodesic distance), ELLE(path-based clustering) can deal with the nonlinear manifold with a complex structure.

Experiment 2 Considering that a large number of Web images are uploaded on various websites every day, and that the images in the online datasets are constantly updated, we invited three undergraduate students to gather two collections for use in the following experiments. When collecting images, they took several semantic concepts as tags and input them into the search textbox separately. In real world applications, it is hard to create a large-scale collection with semantic consistency. Therefore, they did not remove irrelevant images but directly gathered the images from a few return pages. Some inaccurate tags, however, may reduce the performance of image classification to a certain extent.

The first collection was built to evaluate the effectiveness of ELLE during object-level classification, with 2100 images in nine sets. They are 'Airplane' (300), 'Buddha' (200), 'Butterfly' (250), 'Camera' (200), 'Elephant' (300), 'Face' (200), 'Lotus' (250), 'Pigeon' (200), and 'Starfish' (200). Fig. 9 shows some samples of the images in the first collection. The second collection was set up to evaluate the effectiveness of LLSE during scene-level classification, with 1010 images in five sets. They are 'Bird' (234 images: 'Ground' 48, 'Roof' 40, 'Sky' 50, 'Tree' 54, and 'Water' 42), 'Butterfly' (148 images in four scenes: 'Garden' 40, 'Grass' 32, 'Ground' 38, and 'Sky' 38), 'Dog' (216 images in five scenes: 'Grass' 50, 'Indoor' 48, 'Snowfield' 34, 'Street' 42, and 'Water' 42), 'Girl' (226 images in five scenes: 'Beach' 52, 'Garden' 40, 'Grass' 50, 'Indoor' 42, and 'Plaza' 42), and 'Lion' (186 images in five scenes: 'Brushwood' 36, 'Forest' 30, 'Mountain' 38, 'Snowfield' 40, and 'Zoo' 42). Fig. 10 shows some samples of 'Lion' in the second collection.

As can be seen from Table 1, among the nine classes, 'Camera', 'Lotus', and 'Starfish' yield satisfying results due to the fact that the images belonging to these classes usually have clear objects. The images included in 'Buddha' and 'Elephant' have diversity in object shapes, and thus their OCA is lower than that of the other classes. In addition, Table 1 shows that the images belonging to 'Face' can be more easily distinguished using AdaBoost-based face detection to recognize the face region within an image.

LLE(Euclidean distance) obtains the poorest performance among the four algorithms; e.g., for 'Buddha', it has the lowest OCA. Obviously, the proposed ELLE achieves the best results.

The following is the evaluation of the results of LLSE during scene-level classification. After conducting ELLE, the images within the second collection were correctly classified into the corresponding object

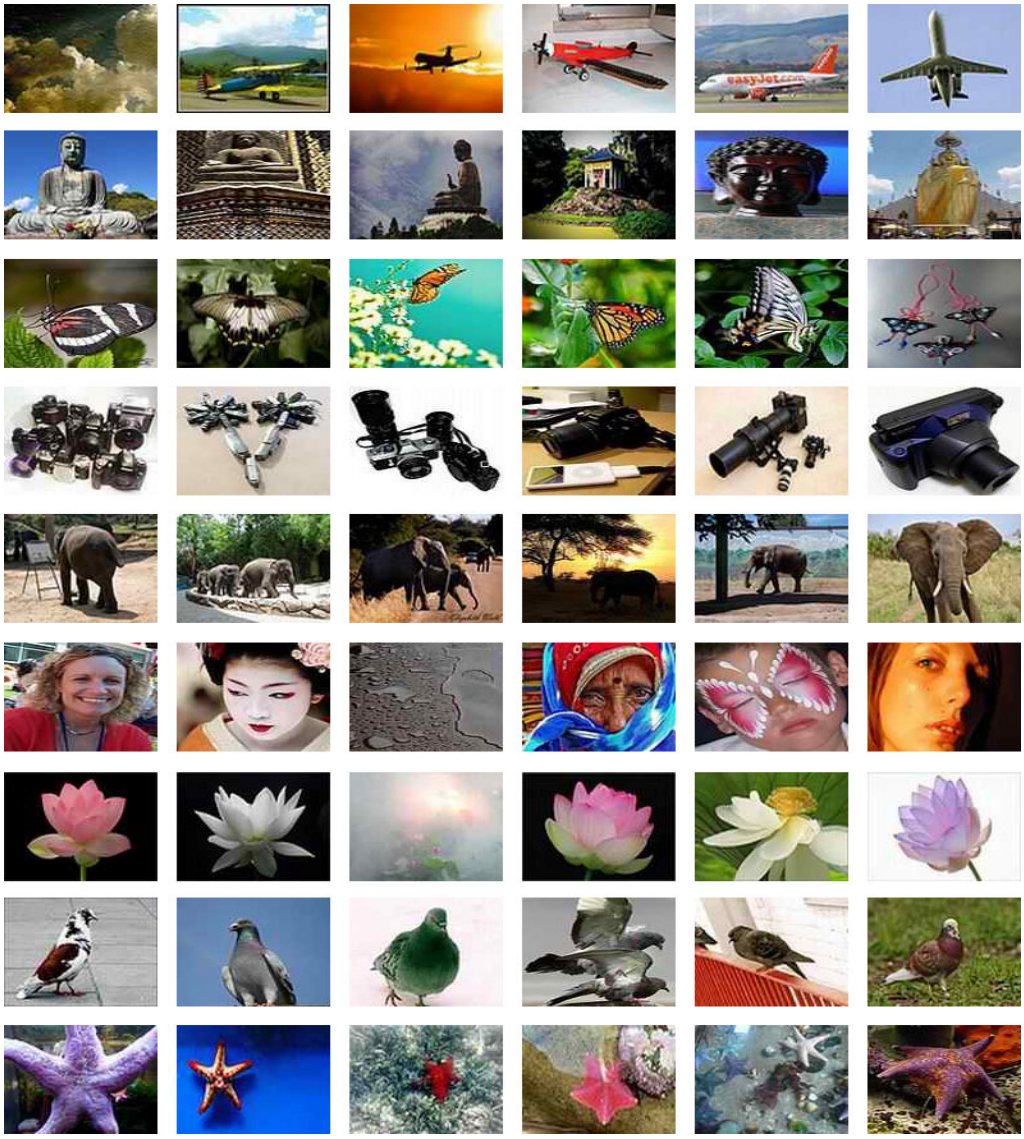

Fig. 9 Some samples of the images in the first collection (each row corresponds to one class and they are 'Airplane', 'Buddha', 'Butterfly', 'Camera', 'Elephant', 'Face', 'Lotus', 'Pigeon', and 'Starfish' from top to bottom)

Table 1 Comparison of object classification accuracy (OCA) and false positive rate (FPR) among the four algorithms

Class	OCA				FPR			
	LLE1	LLE2	Isomap	ELLE	LLE1	LLE2	Isomap	ELLE
Airplane	0.663	0.710	0.764	0.817	0.017	0.014	0.012	0.009
Buddha	0.655	0.690	0.725	0.784	0.052	0.043	0.036	0.027
Butterfly	0.742	0.760	0.794	0.852	0.042	0.036	0.029	0.023
Camera	0.795	0.816	0.830	0.885	0.040	0.033	0.027	0.021
Elephant	0.710	0.747	0.800	0.813	0.064	0.052	0.043	0.033
Face	0.750	0.778	0.828	0.870	0.021	0.018	0.015	0.011
Lotus	0.752	0.812	0.854	0.889	0.024	0.019	0.016	0.012
Pigeon	0.705	0.760	0.815	0.864	0.026	0.021	0.017	0.014
Starfish	0.800	0.825	0.856	0.890	0.038	0.031	0.026	0.020

LLE1: LLE(Euclidean distance); LLE2: LLE(geodesic distance); Isomap: Isomap(geodesic distance); ELLE: ELLE(path-based clustering)

Fig. 10 Some samples of 'Lion' in the second collection (each row corresponds to one scene and they are 'Brushwood', 'Forest', 'Mountain', 'Snowfield', and 'Zoo' from top to bottom)

classes. Two other algorithms, connected component (CC) (Zhai *et al.*, 2008) and manifold-manifold distance (MMD) (Wang RP *et al.*, 2008), were introduced for comparison. The parameters were set according to experience as follows: Th=1.2; $8 \le d \le 14$; each scene manifold was divided into 4 to 7 locally linear submanifolds. To quantitatively evaluate the effectiveness of LLSE, scene classification accuracy (SCA) and false positive rate (FPR) were used to evaluate the classification performance:

$$\mathrm{SCA}_{i,j} = N_{i,j\_\mathrm{p}} / N_{i,j}, \tag{15}$$

$$\mathrm{FPR}_{i,j} = N_{i,\bar{j}\_\mathrm{f}} / N_{i,\bar{j}}, \tag{16}$$

where N_{i,j_p} denotes the number of the images that are correctly classified into the ith class and jth scene, $N_{i,j}$ denotes the number of the images included in the ith class and jth scene, $N_{i,\bar{j}\_\mathrm{f}}$ denotes the number of the images that are wrongly classified into the ith class and jth scene (they belong to the ith class but are not included in the jth scene), and $N_{i,\bar{j}}$ denotes the number of the images that belong to the ith class but are not included in the jth scene.

Table 2 shows that, compared with CC, both MMD and LLSE obtain better results. For example, for 'Lion', the SCA of CC for each scene is lower than that of MMD and LLSE, and the FPR of CC is higher than that of MMD and LLSE. The reasons are as follows: different from MMD and LLSE, CC applies the exemplar-based scheme to calculate the distance between two locally linear submanifolds, but it is difficult to select a suitable exemplar in real world applications, and thus CC obtains the lowest SCA and the highest FPR for each scene. LLSE is superior to MMD when calculating the distance between two locally linear submanifolds with different dimensionalities.

It can also be observed that the images in different scenes have different classification performances. For example, for 'Lion', the SCA is the best in the scene 'Forest' mainly because the images in this scene class have distinct color and texture features, whereas the SCA is the poorest for 'Zoo', since the images belonging to this scene class usually have a complex background. Specifically, many images are easily misrecognized as 'Brushwood' due to the color of the wall. Among the five scenes, the FPR for 'Snowfield' is the lowest as the images belonging to this scene class usually contain a large white region. In contrast, the FPR for 'Mountain' is the poorest, since some images have parts that are visually similar

Table 2 Comparison of scene classification accuracy (SCA) and false positive rate (FPR) among the three algorithms

Class	Scene	SCA			FPR		
		CC	MMD	LLSE	CC	MMD	LLSE
Bird	Ground	0.771	0.917	0.978	0.065	0.021	0.011
	Roof	0.625	0.875	0.950	0.077	0.025	0.010
	Sky	0.700	0.900	0.960	0.076	0.025	0.010
	Tree	0.852	0.944	0.981	0.033	0.011	0.005
	Water	0.643	0.881	0.952	0.094	0.031	0.015
Butterfly	Garden	0.675	0.825	0.925	0.111	0.058	0.025
	Grass	0.625	0.813	0.906	0.129	0.068	0.029
	Ground	0.789	0.884	0.947	0.073	0.036	0.016
	Sky	0.815	0.895	0.972	0.045	0.020	0.010
Dog	Grass	0.860	0.920	0.960	0.042	0.017	0.009
	Indoor	0.688	0.854	0.917	0.071	0.029	0.015
	Snowfield	0.824	0.912	0.941	0.060	0.024	0.013
	Street	0.667	0.857	0.928	0.080	0.031	0.018
	Water	0.809	0.905	0.952	0.057	0.023	0.012
Girl	Beach	0.692	0.865	0.923	0.098	0.041	0.023
	Garden	0.725	0.875	0.925	0.032	0.013	0.007
	Grass	0.820	0.920	0.960	0.034	0.014	0.008
	Indoor	0.571	0.809	0.881	0.108	0.046	0.025
	Plaza	0.595	0.833	0.904	0.107	0.046	0.025
Lion	Brushwood	0.722	0.917	0.944	0.060	0.020	0.013
	Forest	0.800	0.933	0.967	0.064	0.022	0.012
	Mountain	0.763	0.921	0.947	0.101	0.034	0.020
	Snowfield	0.725	0.900	0.950	0.041	0.014	0.007
	Zoo	0.595	0.857	0.928	0.090	0.031	0.014

CC: connected component; MMD: manifold-manifold distance; LLSE: locally linear submanifold extraction

to the ones in ‘Snowfield’ and ‘Zoo’; i.e., the color of the upper part within an image is blue and the color of its bottom part is brown. For ‘Bird’, more satisfactory results were achieved for ‘Tree’ as the images belonging to this scene class have significant differences compared with the images in ‘Ground’, ‘Roof’, ‘Sky’, and ‘Water’, although there exist large variances among the images in ‘Tree’, such as the shape of a tree’s leaves and the color of the flowers. In addition, the images in ‘Roof’, ‘Sky’, and ‘Water’ may be easily confused with each other, and they have lower SCA and higher FPR, since many images in the three scene classes have similarities in color features. For ‘Girl’, among five scenes, the SCA is the highest for ‘Grass’, followed by ‘Garden’ and ‘Beach’, and the poorest for ‘Indoor’ and ‘Plaza’. However, there are many similarities in the images belonging to ‘Indoor’ and ‘Plaza’, and thus both of them obtain poor FPR. For ‘Butterfly’, ‘Sky’ achieves the highest SCA and the lowest FPR, as the images in this scene class have distinct visual characteristics and they can be easily distinguished from the ones included in the other three scene classes. For ‘Dog’, compared with ‘Indoor’, ‘Snowfield’, ‘Street’, and ‘Water’, the images belonging to ‘Grass’ have distinct color features, so satisfactory classification performance can be obtained. Many images in ‘Indoor’ and ‘Street’ contain similar backgrounds, and thus they obtain relatively low classification results.

Although our method has some robustness with respect to the classification errors during object-level classification, there still exist some limitations: (1) The proposed algorithms may not hold for a variety of Web images, such as the scenery images with no specific object. In such cases, Web image classification can be viewed as a global-based one, and the

classifier can be designed on a novel distance measure based on double manifold learning. (2) We did not analyze the effectiveness of the extracted features; instead, we directly adopted some visual features for the classification based on previous research. Therefore, when we classified two objects with similar appearance, such as 'Tiger' and 'Lion', much prior knowledge should be used in the classification process. (3) The proposed hierarchical image manifold was constructed at different semantic levels, designed according to human cognition, and thus our method may have subjectivity to some extent.

5 Conclusions

In this paper, under the assumption that the images in an image set are usually related to the same or similar object but with various scenes, we propose a novel method based on manifold learning to learn a hierarchical image manifold for Web image classification. To achieve good classification performance and effectively reduce the computational complexity, a coarse-to-fine processing strategy is applied to develop the image manifold at the different levels of semantic granularity. That is, two kinds of manifold (object manifold and scene manifold) are constructed using extended locally linear embedding (ELLE) and locally linear submanifold extraction (LLSE), respectively, considering the diversification of Web images. In our method, object-level classification and scene-level classification are viewed as two important parts in one framework, and each part can be further divided into several smaller ones. Therefore, our method is extensible and flexible enough for the classification of Web images.

Our future work will focus on the following aspects: (1) Develop a hierarchical image manifold to tackle the problem of object-based image classification due to the complexity of Web images. A more reasonable manifold learning based method should be presented for global-based image classification. (2) Propose a solution to disjoint multiple nonlinear manifolds, and thus a more sophisticated scheme should be applied to handle the intersected manifolds or the ones with imperfect structures. (3) Calculate the distance between two images on the image manifold based on a specific pairwise similarity measure. This process may be time consuming, and thus more precise calculation methods or parallel execution schemes should be exploited. (4) The proposed hierarchical image model is constructed considering only object semantics and scene semantics in high-level image semantics. Some novel classification models based on more abstract semantic concepts, e.g., behavioral semantics and emotional semantics, should be developed. (5) In this study the experiments were conducted on small collections with limited semantic concepts; some larger collections and more semantic concepts should be used for further experiments.

References

Ames, M., Naaman, M., 2007. Why We Tag: Motivations for Annotation in Mobile and Online Media. SIGCHI Conf. on Human Factors in Computing, p.971-980.

Belkin, M., Niyogi, P., 2001. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. Advances in Neural Information Processing Systems 14. MIT Press, p.585-591.

Bellman, R.E., 1961. Adaptive Control Processes: a Guided Tour. Princeton University Press, New Jersey.

Briggs, F., Raich, R., Fern, X.Z., 2009. Audio Classification of Bird Species: a Statistical Manifold Approach. Ninth IEEE Int. Conf. on Data Mining, p.51-60. [doi:10.1109/ICDM.2009.65]

Carlsson, G., Ishkhanov, T., de Silva, V., Zomorodian, A., 2008. On the local behavior of spaces of natural images. *Int. J. Comput. Vis.*, **76**(1):1-12. [doi:10.1007/s11263-007-0056-x]

Chai, Y.M., Zhu, X.Y., Zhou, S., Bian, Y.T., Bu, F., Li, W., Zhu, J., 2009. Ontology-Based Digital Photo Annotation Using Multi-source Information. IEEE Int. Conf. on Computational Intelligence for Measurement Systems and Applications, p.38-41. [doi:10.1109/CIMSA.2009.5069914]

Chang, E., Goh, K., Sychay, G., Wu, G., 2003. CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines. *IEEE Trans. Circ. Syst. Video Technol.*, **13**(1):26-38. [doi:10.1109/TCSVT.2002.808079]

Cheng, E., Jing, F., Zhang, L., 2009. A unified relevance feedback framework for Web image retrieval. *IEEE Trans. Image Process.*, **18**(6):1350-1357. [doi:10.1109/TIP.2009.2017128]

Datta, R., Joshi, D., Li, J., Wang, J.Z., 2008. Image retrieval: ideas, influences, and trends of the new age. *ACM Comput. Surv.*, **40**(2):1-60. [doi:10.1145/1348246.1348248]

de Juan, C., Bodenheimer, B., 2004. Cartoon Textures. Proc. ACM SIGGRAPH/Eurographics Symp. on Computer Animation, p.267-276. [doi:10.1145/1028523.1028559]

de Ridder, D., Kouropteva, O., Okun, O., Pietikainen, M., Duin, R.P.W., 2003. Supervised locally linear embedding. *LNCS*, **2714**:175. [doi:10.1007/3-540-44989-2_40]

dos Santos, J.A., Ferreira, C.D., Torres, R.S., Goncalves, M.A., Lamparelli, R.A.C., 2011. A relevance feedback method based on genetic programming for classification of remote sensing images. *Inform. Sci.*, **181**(13):2671-2684. [doi:10.1016/j.ins.2010.02.003]

El Sayad, I., Martinet, J., Urruty, T., Amir, S., Dieraba, C., 2010. Effective Object-Based Image Retrieval Using Higher-Level Visual Representation. Int. Conf. on Machine and Web Intelligence, p.218-224. [doi:10.1109/ICMWI.2010.5648110]

Enser, P., Sandom, C., 2003. Towards a Comprehensive Survey of the Semantic Gap in Visual Image Retrieval. Int. Conf. on Image and Video Retrieval, p.291-299. [doi:10.1007/3-540-45113-7_29]

Fan, J.P., Gao, Y.L., Luo, H.Z., Jain, R., 2008. Mining multilevel image semantic via hierarchical classification. *IEEE Trans. Multimedia*, **10**(2):167-187. [doi:10.1109/TMM.2007.911775]

Fan, W., Yeung, D.Y., 2006. Locally Linear Models on Faces Appearance Manifolds with Application to Dual-Subspace Based Classification. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, **2**:1384-1390. [doi:10.1109/CVPR.2006.178]

Farajtabar, M., Rabbiee, H.R., Shaban, A., Soltani-Farani, A., 2011. Efficient Iterative Semi-supervised Classification on Manifold. IEEE 11th Int. Conf. on Data Mining Workshops, p.228-235. [doi:10.1109/ICDMW.2011.181]

Fischer, B., Buhmann, J.M., 2003. Path-based clustering for grouping of smooth curves and texture segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **25**(4):513-518. [doi:10.1109/TPAMI.2003.1190577]

Gao, Y., Fan, J.P., 2005. Semantic Image Classification with Hierarchical Feature Subset Selection. Proc. 7th ACM SIGMM Int. Workshop on Multimedia Information Retrieval, p.135-142. [doi:10.1145/1101826.1101850]

Guo, G.D., Jain, A.K., Ma, W.Y., Zhang, H.J., 2002. Learning similarity measure for natural image retrieval with relevance feedback. *IEEE Trans. Neur. Networks*, **13**(4):811-820. [doi:10.1109/TNN.2002.1021882]

Huang, J., Kumar, S.R., Zabih, R., 2003. Automatic hierarchical color image classification. *EURASIP J. Appl. Signal Process.*, (2):151-159. [doi:10.1155/S1110865703211161]

Huiskes, M.J., Lew, M.S., 2008. The MIR Flickr Retrieval Evaluation. Proc. 1st ACM Int. Conf. on Multimedia Information Retrieval, p.39-43. [doi:10.1145/1460096.1460104]

Jaimes, A., Smith, J.R., 2003. Semi-automatic, Data-Driven Construction of Multimedia Ontologies. Proc. Int. Conf. on Multimedia and Expo, **1**:781-784. [doi:10.1109/ICME.2003.1221034]

Jaimes, A., Jaimes, R., Chang, S.F., 1999. Model-Based Classification of Visual Information for Content-Based Retrieval. Conf. on Storage and Retrieval for Image and Video Databases, p.402-414.

Joshi, A.J., Porikli, F., Papanikolopoulos, N., 2009. Multi-class Active Learning for Image Classification. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Poster Session 5.

Jun, G., Ghosh, J., 2010. Nearest-Manifold Classification with Gaussian Processes. 20th Int. Conf. on Pattern Recognition, p.914-917. [doi:10.1109/ICPR.2010.230]

Kang, S.D., Park, S.S., Yoo, H.W., Shin, Y.G., Jang, D.S., 2009. Development of expert system for extraction of the objects of interest. *Exp. Syst. Appl.*, **36**(3):7210-7218. [doi:10.1016/j.eswa.2008.09.062]

Kim, B.S., Park, J.Y., Mohan, A., Gilbert, A., Savarese, S., 2010. Hierarchical Classification of Images by Sparse Approximation. Proc. British Machine Vision Conf., p.106.1-106.11. [doi:10.5244/C.25.106]

Kim, D.W., Song, J.H., Lee, J.H., Choi, B.G., 2007. Support vector machine learning for region-based image retrieval with relevance feedback. *ETRI J.*, **29**(5):700-702. [doi:10.4218/etrij.07.0207.0037]

Kim, T.K., Kittle, J., Cipolla, R., 2007. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. Pattern Anal. Mach. Intell.*, **29**(6):1005-1008. [doi:10.1109/TPAMI.2007.1037]

Klaydios, K., 2004. Relevance Feedback Methods for Web Image. PhD Thesis, Technical University of Crete, Chania, Greece.

Li, L.J., Wang, C., Lim, Y.W., Blei, D.M., Li, F.F., 2010. Building and Using a Semantivisual Image Hierarchy. IEEE Conf. on Computer Vision and Pattern Recognition, p.336-3343. [doi:10.1109/CVPR.2010.5540027]

Li, X.R., Snoek, C.G.M., Worring, M., 2010. Unsupervised Multi-feature Tag Relevance Learning for Social Image Retrieval. Proc. ACM Int. Conf. on Image and Video Retrieval, p.10-17. [doi:10.1145/1816041.1816044]

Lin, Y.Q., Lv, F.J., Zhu, S.H., Yang, M., Cour, T., Yu, K., Cao, L.L., Huang, T., 2011. Large-Scale Image Classification: Fast Feature Extraction and SVM Training. IEEE Conf. on Computer Vision and Pattern Recognition, p.1689-1696. [doi:10.1109/CVPR.2011.5995477]

Liu, D., Yang, S.C., Mu, Y.D., Hua, X.S., Zhang, H.J., 2011. Towards Optimal Discriminating Order for Multiclass Classification. IEEE 11th Int. Conf. on Data Mining, p.388-397. [doi:10.1109/ICDM.2011.147]

Lu, D., Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. *Int. J. Remote Sens.*, **28**(5):823-870. [doi:10.1080/01431160600746456]

Luo, D.J., Huang, H., Ding, C., 2010. Discriminative High Order SVD: Adaptive Tensor Subspace Selection for Image Classification, Clustering, and Retrieval. IEEE Int. Conf. on Computer Vision, p.1443-1448. [doi:10.1109/ICCV.2011.6126400]

Luo, J.B., Singhal, A., Etz, S.P., Gray, R.T., 2004. A computational approach to determination of main subject regions in photographic images. *Image Vis. Comput.*, **22**(3):227-241. [doi:10.1016/j.imavis.2003.09.012]

Parikh, D., 2011. Recognizing Jumbled Images: the Role of Local and Global Information in Image Classification. IEEE Int. Conf. on Computer Vision, p.519-526. [doi:10.1109/ICCV.2011.6126283]

Patterson, F., 1986. Photography and the Art of Seeing. Baker & Taylor Books, Charlotte, North Carolina.

Pillati, M., Viroli, C., 2005. Supervised Locally Linear Embedding for Classification: an Application to Gene Expression Data Analysis. Annual Conf. of the German Classification Society, p.15-18.

Rizon, M., Yazid, H., Saad, P., Shakaff, A.Y.M., Saad, A.R., Mamat, M.R., Yaacob, S., Desa, H., Karthigayan, M., 2006. Object detection using geometric invariant moment. *Am. J. Appl. Sci.*, **3**(6):1876-1878. [doi:10.3844/ajassp.2006.1876.1878]

Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensional reduction by locally linear embedding. *Science*, **290**(5500):2323-2326. [doi:10.1126/science.290.5500.2323]

Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S., 1998. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Trans. Circ. Video Technol.*, **8**(5):644-655. [doi:10.1109/76.718510]

Saul, L.K., Roweis, S.T., 2003. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res.*, **4**:119-155. [doi:10.1162/153244304322972667]

Seung, H.S., Lee, D.D., 2000. The manifold ways of perception. *Science*, **290**(5500):2268-2269. [doi:10.1126/science.290.5500.2268]

Shao, L., Brady, M., 2006. Specific object retrieval based on salient regions. *Pattern Recogn.*, **39**(10):1932-1948. [doi:10.1016/j.patcog.2006.04.010]

Souvenir, R., Pless, R., 2005. Manifold Clustering. Int. Conf. on Computer Vision, p.648-653.

Sun, A., Bhowmick, S.S., Nguyen, K.T.N., Bai, G., 2011. Tag-based social image retrieval: an empirical evaluation. *J. Am. Soc. Inform. Sci. Technol.*, **62**(12):2364-2381. [doi:10.1002/asi.21659]

Tao, D., Tang, X., Li, X., Rui, Y., 2006. Direct kernel biased discriminant analysis: a new content-based image retrieval relevance feedback algorithm. *IEEE Trans. Multimedia*, **8**(4):716-727. [doi:10.1109/TMM.2005.861375]

Tenenbaum, J.B., Silva, V.D., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**(5500):2319-2323. [doi:10.1126/science.290.5500.2319]

Vailaya, A., Jain, A., Zhang, H.J., 1998. On image classification: city images vs. landscapes. *Pattern Recogn.*, **31**(12):1921-1935. [doi:10.1016/S0031-3203(98)00079-X]

Vieux, R., Domenger, J.P., Benois-Pineau, J., Braquelaire, A., 2007. Image Classification with User Defined Ontology. 15th European Signal Processing Conf., p.723-727.

Viola, P., Jones, M., 2004. Robust real-time face detection. *Int. J. Comput. Vis.*, **57**(2):137-154. [doi:10.1023/B:VISI.0000013087.49260.fb]

Wang, C.H., Zhang, L., Zhang, H.J., 2008. Learning to Reduce the Semantic Gap in Web Image Retrieval and Annotation. Proc. 31st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, p.355-362. [doi:10.1145/1390334.1390396]

Wang, L., Wang, X., Feng, J., 2006. Subspace distance analysis with application to adaptive Bayesian algorithm for face recognition. *Pattern Recogn.*, **39**(3):456-464. [doi:10.1016/j.patcog.2005.08.015]

Wang, R.P., Shan, S.G., Chen, X.L., Gao, W., 2008. Manifold-Manifold Distance with Application to Face Recognition Based on Image Set. IEEE Conf. on Computer Vision and Pattern Recognition, p.1-8. [doi:10.1109/CVPR.2008.4587719]

Wu, Y., Chan, K.L., 2004. An Extended Isomap Algorithm for Learning Multi-class Manifold. Int. Conf. on Machine Learning and Cybernetics, **6**:3429-3433.

Yang, M.H., 2002. Extended Isomap for Pattern Classification. Proc. AAAI/AAI, p.224-229.

Zeng, Z.Y., Yao, Z.Q., Liu, S.G., 2009. An Efficient and Effective Image Representation for Region-Based Image Retrieval. Proc. 2nd Int. Conf. on Interaction Sciences: Information Technology, Culture and Human, p.429-434. [doi:10.1145/1655925.1656004]

Zhai, S.D., Luo, B., Zhang, C.Y., 2008. Video abstraction based on manifold learning and mixture model. *J. Image Graph.*, **13**(4):735-740 (in Chinese).

Zhang, Y.J., 2008. Image Classification and Retrieval with Mining Technologies. *In*: Song, M., Wu, Y.F.B. (Eds.), Handbook of Research on Text and Web Mining Technologies, Chapter VI, p.96-110. [doi:10.4018/978-1-59904-990-8.ch006]

Zhou, X., Cui, N., Li, Z., Liang, F., Huang, T.S., 2009. Hierarchical Gaussianization for Image Classification. IEEE 12th Int. Conf. on Computer Vision, p.1971-1977. [doi:10.1109/ICCV.2009.5459435]

Zhu, R., Yao, M., 2009. Image feature optimization based on nonlinear dimensionality reduction. *J. Zhejiang Univ.-Sci. A*, **10**(12):1720-1737. [doi:10.1631/jzus.A0920310]