

# A Hierarchical Image Annotation Method Based on SVM and Semi-supervised EM

GAO Yan-Yu<sup>1,2</sup> YIN Yi-Xin<sup>1,2</sup> UOZUMI Takashi<sup>3</sup>

**Abstract** Automatic image annotation, which aims at automatically identifying and then assigning semantic keywords to the meaningful objects in a digital image, is not a very difficult task for human but has been regarded as a difficult and challenging problem to machines. In this paper, we present a hierarchical annotation scheme considering that generally human's visual identification to a scenery object is a rough-to-fine hierarchical process. First, the input image is segmented into multiple regions and each segmented region is roughly labeled with a general keyword using the multi-classification support vector machine. Since the results of rough annotation affect fine annotation directly, we construct the statistical contextual relationship to revise the improper labels and improve the accuracy of rough annotation. To obtain reasonable fine annotation for those roughly classified regions, we propose an active semi-supervised expectation-maximization algorithm, which can not only find the representative pattern of each fine class but also classify the roughly labeled regions into corresponded fine classes. Finally, the contextual relationship is applied again to revise the improper fine labels. To illustrate the effectiveness of the presented approaches, a prototype image annotation system is developed, the preliminary results of which showed that the hierarchical annotation scheme is effective.

**Key words** Hierarchical image annotation, support vector machine (SVM), semi-supervised expectation-maximization, coexistence, relative location relationship

**DOI** 10.3724/SP.J.1004.2010.00960

In recent years, a variety of image auto-annotation systems have been proposed with the development of artificial intelligence and statistical learning theory. These auto-annotation methods can be classified into three categories: 1) image-based auto-annotation<sup>[1]</sup> which considers the whole image as an individual visual pattern and uses visual features of the whole image to infer its semantic contents; 2) blob-based (region-based) auto-annotation<sup>[2]</sup> which takes the homogeneous image region or connected homogeneous image regions with the same visual attributes as the annotating object and extracts its visual features for blob understanding; 3) salient-based auto-annotation<sup>[3]</sup> which considers the salient regions as annotating objects and extracts their visual features for image understanding. Among these annotation methods, blob-based auto-annotation received more attention. One of its first attempts was reported by Mori et al.<sup>[2]</sup>, who estimated the co-occurrence probabilities between words and image regions created by a regular grid and used the probabilities to predict image contents. Jeon et al.<sup>[4]</sup> assumed image annotation as a kind of cross-lingual retrieval problem and built a cross-media relevance model (CMRM) to do image annotation. Their experiments showed that the CMRM performs better than the models proposed in [2] on the same image set. Graph models and word correlation are well considered in recent years. Liu et al.<sup>[5]</sup> proposed a unified framework for image annotation, which consisted of an image-based graph learning process and a word-based graph learning process. Their experiments demonstrated that their framework outperforms the CMRM and other recently proposed methods. Syncrizing image segmentation and region recognition for image annotation is another attractive aspect in more recent years. Kokkinos et al.<sup>[6]</sup> integrated image segmentation and object recognition in the framework of the expectation-maximization (EM) algo-

rithm and adopted the active appearance models to model objects. Their experiments on faces and cars show that the synergy scheme is not only faster than the other three methods but also insensitive to occlusion. There is also some interesting research based on authors' prudential observation. Jung<sup>[7]</sup> proposed an ontology-based semantic annotation method to improve the understandability of images from heterogeneous information sources considering that user's search context should be predicted by his annotated resources.

The above-mentioned methods made great efforts in improving annotation precision and speed as well as enlarging the scope of annotation objects. However, besides these aspects, the following issues are also important and worth in-depth study.

1) Object recognition by image analysis is easily affected by impersonal elements, such as lighting conditions, and subjective elements, such as photographing angles and distance. Extracting visual features that are insensitive to orientation, scale, and lighting can decrease these influences.

2) Human's understanding to an image is usually a rough-to-fine hierarchical process<sup>[8]</sup>. Therefore, if an image auto-annotation system could similarly perform, it may achieve more definite annotation results.

3) Different people may give different labels to the same scenery object because of the lingual diversities and cognition differences. Thus, forming a set of consistent annotation glossary is a necessary pre-step of auto-annotation.

Focusing on these issues, in this paper, we construct a blob-based automatic annotation system for outdoor scenery images. The system mainly consists of five parts: image segmentation, rough annotation, rough correction, fine annotation, and fine correction (as shown in Fig. 1). In the first part, the input image is segmented into multiple regions with an assumption that each region contains no more than a single object. Meanwhile, a questionnaire is performed, by which a set of uniform and widely acceptable keywords are selected. These keywords are organized into two hierarchies to describe objects of scenery images in general or in detail. Corresponding to these hierarchical keywords, two sets of visual features are defined. With the rough features, the segmented region is classified into a ro-

Manuscript received April 25, 2008; accepted December 1, 2009  
Supported by National Natural Science Foundation of China (60374032)

1. School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, P. R. China 2. Key Laboratory of Advanced Control of Iron and Steel Process (Ministry of Education), University of Science and Technology Beijing, Beijing 100083, P. R. China 3. Department of Computer Science and Systems Engineering, Muroran Institute of Technology, Muroran 050-8585, Japan

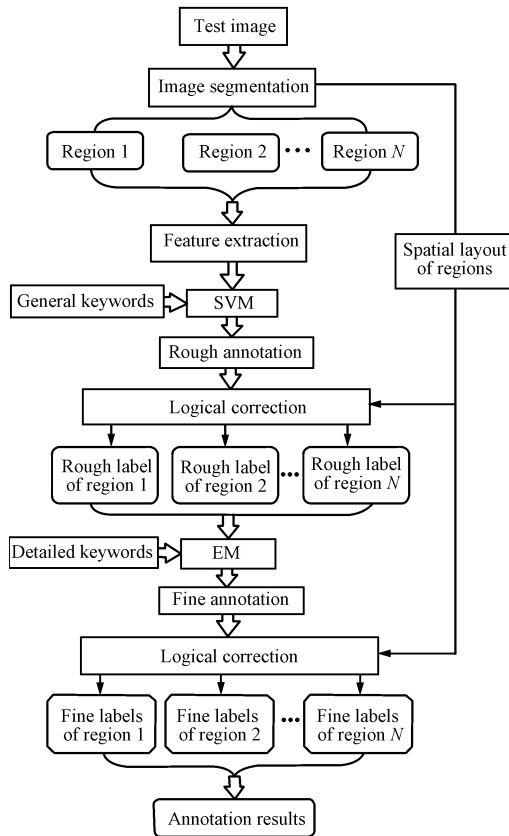


Fig. 1 Schematic illustration of the proposed automatic image annotation system

ough semantic category by support vector machines (SVMs) under the consideration that the precision of rough annotation would directly affect subsequent fine annotation, while SVM along with the supervised learning method can obtain high annotation precision. Then, an active semi-supervised EM algorithm is proposed to set the representative patterns of detailed keywords and to annotate the roughly labeled regions with detailed keywords according to their fine features. In order to further reduce annotation errors, statis-

tical information of coexistence and relative location relationship is calculated based on the training samples, which is helpful in judging or even revising rough and fine annotation results.

The rest of the paper is organized as follows. Section 1 explains image segmentation, visual feature extraction, and annotation keywords selection. The hierarchical annotation scheme is explained in Section 2. Annotation revision with contextual relationship is described in Section 3. Section 4 reports the experimental results and provides some analysis. Section 5 concludes with several remarks on further work.

## 1 Preprocessing

In this section, we introduce three preprocessing steps: semantic keywords selection, image segmentation, and visual feature extraction.

### 1.1 Semantic keywords selection

Semantic keywords selection has a critical effect on reducing lingual diversity and normalizing cognitive variety in image annotation. However, only a few researchers have paid attention to this issue till now.

In order to obtain a set of comprehensive and coherent keywords, we performed a small-scale subjective experiment, where totally eight subjects (4 males, 4 females) with normal color vision and normal or corrected-to-normal vision participated. Each subject was asked to watch a set of natural scenery images selected from the Corel stock CDs and write down names of observed objects. To obtain as many names of scenery objects as possible and lighten the intensity of questionnaire, two rules were followed: 1) Each image set contained 100 images and no identical images could be found in any two image sets; 2) Images in each set should involve as many as possible themes. After analyzing all names written down, we deleted the rarely used synonyms. Then, by referring the WordNet lexical database, we added scores of commonly heard object names. Finally, we obtained 87 words, some of which were very general and applicable to describe a large range of scenery objects (e.g. stone), but others were only fit for describing specific objects (e.g. pebble). We grouped these words into two hierarchies (as shown in Table 1) and referred them as general keywords and detailed keywords, respectively.

Table 1 Semantic keywords organized in two hierarchies

General keywords	Detailed keywords	Features for differentiation
Water	Sea, river, lake, waterfall, rain, etc.	Color, texture (directionality)
Stone	Pebble, boulder, stonewall, reef, flagstone, etc.	Color, texture, shape
Ground	Soil, dry-land, mud, sandbeach, desert, snow-field, etc.	Texture, color
Mountain	Hill, cliff, ice-mountain, snow-mountain, barren-mountain, green-mountain, volcano, etc.	Color, texture, shape
Sky	Blue-sky, white-clouds, storm-sky, dim-sky, sunrise/sunset, etc.	Color, texture
Grass	Withered-grass, green-grass, crop, bamboo, reed, etc.	Color, texture
Tree	Bush, stub, shrub, pine, green-tree, defoliated-tree, autumn-tree, etc.	Color, texture
Flower	Flowerbed, red-flower, yellow-flower, purple-flower, pink-flower, varicolor-flower, etc.	Color, texture, shape
Road	Lane, tunnel, highway, railway, stairway, flagging, etc.	Color, texture
Building	Arena, temple, castle, office-building, woody-building, fence, sculpture, stairs, etc.	Color, texture, shape
Vehicle	Car, airplane, balloon, truck, bus, ship, train, etc.	Shape, color
Animals	Human, tiger, dolphin, elephant, horse, bear, penguin, etc.	Shape, color

## 1.2 Image segmentation

Image segmentation is the process that groups image pixels together based on color distribution, region edge, and spatial location, so that the segmented regions have a strong correlation with the real-world objects. The existent image segmentation techniques can be classified into four classes<sup>[9–10]</sup>: 1) pixel-based approach, which groups pixels into different regions according to the low-level visual features; 2) edge-based approach, which first detects local discontinuities and then uses the edge information to separate the image into regions sequentially or in a parallel way; 3) region-based approach, which starts with a seed pixel (or a group of pixels) and then grows or splits the seed until the original image is composed of homogeneous regions only; 4) hybrid approach.

In this paper, we adopt the Gaussian mixture model (GMM)-based pixel clustering technique<sup>[10]</sup>, which applies the color or intensity feature as well as the pixel location information to determine the segments. Experimental results in [10] proved that this technique can obtain better segmentation performance than other methods. Since the time cost of the GMM-based clustering method is directly proportional to the number of iterations, we did some segmentation tests on several scenery images with 50, 40, 30, and 20 iterations. The segmentation results showed that the segments formed after 20 iterations were almost as same as those formed after 50 iterations, but the time cost was greatly less than that with 50 times.

## 1.3 Visual feature extraction

During the past semicentury, color and texture features have received deeper investigation and gained wider application. In this paper, we combine color moments and Gabor wavelet texture<sup>[11]</sup> to describe each region roughly and use the two features as well as Tamura directionality<sup>[12]</sup> to describe each region finely.

### 1.3.1 Color moment feature

Considering that most segmented regions have homogeneous color distribution, we calculate the first-, second-, third-, and fourth-order color central moments for each region as its color feature. Here, we calculate the moment features in the CIE  $L^*a^*b^*$  color space because it reflects human-perceived color differences better than other spaces. The first and second moments measure the average and variance of color values in each color channels. The third central moment reflects the skewness of each color channel. If its value in a color channel is small, the color distribution of the color channel is symmetric. The fourth central moment, also called as Kurtosis, measures the peakedness of each color channel. If the Kurtosis of a color channel is big, color distribution of the color channel tends to have a distinct peak near the mean, declines rather rapidly, and has heavy tails. Otherwise, the color distribution tends to have a flat top near the mean. A distinct advantage of the color moment feature is that it represents various statistical color attributes in a compact size.

### 1.3.2 Gabor wavelet texture features

Gabor wavelets have achieved impressive results in analyzing texture of grayscale images because they provide the best trade-off between spatial resolution and frequency resolution<sup>[13]</sup>. Gabor wavelets can decompose images into components corresponding to different scales and orientations. In this paper, we convolve Gabor wavelets in 3 scales at 6 orientations ( $0, \pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6$ ) with each gray-scaled region and obtain 18 coefficients. The mean

and standard deviation of the magnitudes of these coefficients are combined to form the texture features of the region. Since Gabor wavelet transform is originally designed for rectangular images and the segmented regions are in arbitrary shapes, we perform the transform on the maximum rectangular region inside each segment.

### 1.3.3 Tamura directionality

Directionality refers to the placement rule of texture primitives. It is very important to differentiate fine semantic classes. We adopt the Tamura method to calculate the directionality of a region, which is represented by a 16-point histogram  $H_D$  corresponding to 16 direction scopes ( $(2k-1)\pi/32 \leq \theta < (2k+1)\pi/32$ , with  $k = 0, 1, \dots, 15$ ). If  $H_D$  is nearly flat, the region is regarded as having isotropic texture. Otherwise, the angle corresponding to the histogram peak is regarded as the texture direction of the region.

### 1.3.4 Features for rough and fine annotation

In this paper, we propose a parallel feature representation scheme and apply it in rough and fine annotation. In rough annotation, color moments and Gabor wavelet texture are reassembled into a group of feature vectors, each of which consists of a pair of Gabor wavelet texture extracted from one scale and one orientation as well as 12 color features. Each training sample is represented by eighteen 14-dimensional feature vectors, while each test region is represented by a 14-dimensional vector randomly selected from its feature group. If the test vector is nearest to one vector of a training feature group, the test region is regarded as belonging to the same class with the training sample. Such kind of feature representation can effectively decrease influences from inconsistent image scales and various photographing angles.

As for the fine annotation, each feature vector includes a pair of Gabor wavelet features extracted from one scale and one orientation, 12 color moments, and 16 directionality features. As same as the rough annotation, each training sample is represented by eighteen 30-dimensional feature vectors, while each test region is represented by a 30-dimensional vector randomly selected from the feature group. To eliminate the inconsistent ranges of various visual features, we normalize each feature vector to be of zero mean and unit variance.

## 2 Hierarchical image annotation

This section explains the idea and process of applying SVM technique for rough annotation and an active semi-supervised EM algorithm for fine annotation.

### 2.1 Rough annotation by SVM

Denoting the segmented regions as  $\mathbf{s} = \{s_1, s_2, \dots, s_\theta\}$ , the goal of rough annotation is to assign a set of general keywords  $\{g_1, g_2, \dots, g_\theta\}$  to these regions and hope that the assignment is as accurate as possible. This, indeed, can be viewed as a pattern-classification problem, namely, classify each region into one of the predefined rough semantic classes. To obtain high recognition precision, SVM is recommended, which seeks the optimal separating hyperplane between two classes by focusing on the training samples that lie on the class boundaries while discarding other training samples effectively<sup>[14]</sup>.

The simplest classification problem that SVM could deal with is the linearly separable binary classification. Given a set of training samples,  $\{(x_i, y_i)\}_{i=1}^N, x_i \in \mathbf{R}^d, y_i = \{-1, 1\}$ , SVM separates these samples into two classes by the optimal hyperplane  $w^T x + b = 0$ , which runs between the two

classes with the distance to the closest training samples in both classes as large as possible. Here,  $w$  is an adaptive weight vector and is normal to the hyperplane,  $b$  is a bias. The optimal hyperplane can be obtained by solving the optimization problem:  $\min\{\frac{1}{2}\|w\|^2\}$ , subject to the constraints of  $y_i(w^T x_i + b) \geq +1$ . By introducing the Lagrange multiplier  $\alpha_i$ , the quadratic optimization problem with linear inequality constraints can be converted into the following dual problem:

$$\max Q(a) = \sum_{i=1}^m \alpha_i - 0.5 \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (1)$$

where  $\alpha_i \geq 0$  and  $\sum_{i=1}^m y_i \alpha_i = 0$ . With the standard quadratic programming (QP) method, the optimum Lagrange multipliers  $\{\alpha_i\}_{i=1}^m$  as well as the discriminant function associated with the optimal hyperplane can be decided. As for the commonly faced non-separable case, the objective function becomes  $\min\{\frac{1}{2}\|w\|^2 + C(\sum_i \xi_i)\}$ , subject to the constraints of  $y_i(w^T x_i + b) \geq 1 - \xi_i$ , where  $\xi_i$  are positive slack variables and  $C$  is a user-specified positive parameter. Its dual problem is almost the same as the separable case except that the Lagrange multipliers  $\alpha_i$  have an upper bound of  $C$ , i.e.,  $0 \leq \alpha_i \leq C$ .

The linear SVM can be extended to solve the nonlinear classification problem by introducing the nonlinear operator  $\phi(x)$  to map the input pattern  $x$  into a higher dimensional space  $H$ . Accordingly, the inner products in the original space ( $x_i \cdot x_j$ ) defined in (1) would be replaced by  $(\phi(x_i) \cdot \phi(x_j))$ . Since the explicit computation of  $\phi(x_i)$  is expensive and sometimes unfeasible, the kernel function  $K(\cdot, \cdot)$  was introduced, which satisfies the conditions of Mercer's theorem and corresponds to some type of inner product in the high-dimensional space  $H$ , i.e.,  $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ . In this study, the nonlinear SVMs with the Gaussian radial basis function (RBF) kernel are used considering that the Gaussian RBF kernel usually yields excellent results compared with linear and polynomial kernels. The RBF kernel is defined as

$$K(x_i, y_j) = \exp(\gamma \|x_i - y_j\|^2), \quad \gamma > 0 \quad (2)$$

where  $\gamma$  shapes the kernel function. Since the hyperparameters  $C$  and  $\gamma$  decisively affect the classification performance ( $C$  controls the trade-off between low training error and large margin, while  $\gamma$  can alter the effectiveness of the eventual separating surface), so far many hyperparameter tuning methods have been proposed<sup>[15]</sup>, such as  $k$ -fold cross-validation, leave-one-out (LOO), Xi-alpha bound, and generalized approximate cross-validation (GACV), etc. We combine the 3-fold cross-validation and grid-search algorithm and apply them to the pre-labeled training images to find the best  $C$  and  $\gamma$ .

SVM was originally designed for binary classification. To make it competent for multiclass classification, three strategies have been proposed, namely one-against-all, one-against-one, and directed acyclic graph SVM (DAGSVM). We adopt the DAGSVM in rough annotation since the DAGSVM needs less testing time than the one-against-one SVM and has better generalization ability than the one-against-all SVM<sup>[16]</sup>.

## 2.2 Fine annotation by EM algorithm

To complete fine annotation, we first need to find out the representative pattern of each detailed keyword. Here, the unsupervised learning method — EM algorithm<sup>[17]</sup> is adopted, which assumes that the scenery objects of the

same rough category are generated by a multi-component Gaussian mixture model (GMM) with each Gaussian component corresponding to a fine semantic class.

Let  $X = \{x_i\}_{i=1}^N$  be  $N$  samples that belong to the same rough category and  $x_i = [x_i^1, \dots, x_i^d]^T$  represents  $d$ -dimensional feature vector of the  $i$ -th sample. The feature distribution of  $X$  is assumed to follow a  $K$ -component Gaussian mixture model:

$$p(x_i|\Theta) = \sum_{k=1}^K \alpha_k \frac{1}{(2\pi)^{\frac{d}{2}} \sigma_k^{\frac{d}{2}}} \times \exp \left[ -\frac{1}{2} (x_i - \mu_k)^T \sum_k^{-1} (x_i - \mu_k) \right] \quad (3)$$

where  $\alpha_k$  is the probability of choosing the  $k$ -th mixture component and  $\sum_{k=1}^K \alpha_k = 1$ ;  $\Theta = \{\mu_1, \dots, \mu_K, \sigma_1, \dots, \sigma_K, \alpha_1, \dots, \alpha_K\}$  are the parameters of Gaussian mixture model that we need to estimate.

Assume that  $Y = \{y_1, \dots, y_N\}$  represents the fine labels of image samples  $\{x_i\}_{i=1}^N$ . Each label  $y_i = [y_i^1, \dots, y_i^K]_{i=1}^N$  is a binary vector that indicates which component of the Gaussian mixture model produces the image sample  $x_i$ . If  $x_i$  is produced by the  $k$ -th component, then  $y_i^k = 1$  and  $y_i^j = 0$  for  $j \neq k$ , and  $k, j \in [1, K]$ . EM algorithm calculates the maximum likelihood (ML) estimation of labels  $\{y_i\}_{i=1}^N$  and parameters  $\Theta$  by alternately performing the expectation step (E-step) and the maximization step (M-step) until convergence. At the  $(t+1)$ -th iteration, the E-step calculates the expected value of the complete-data log-likelihood function

$$Q(\Theta|\Theta^{(t)}) = \sum_{i=1}^N \sum_{k=1}^K \log(\alpha_k p(x_i|\theta_k^{(t)})) P(y_i^k = 1|x_i; \Theta^{(t)}) \quad (4)$$

where  $P(y_i^k = 1|x_i; \Theta^{(t)})$  estimates the labels  $\{y_i\}_{i=1}^N$  according to the observed visual features and the model parameters in the  $t$ -th iteration  $\Theta^{(t)}$ :

$$P(y_i^k = 1|x_i; \Theta^{(t)}) = \frac{\alpha_k^t p(x_i|\theta_k^{(t)})}{\sum_{l=1}^K \alpha_l^t p(x_i|\theta_l^{(t)})} \quad (5)$$

In the M-step, the model parameters of the  $(t+1)$ -th iteration are estimated by maximizing  $Q(\Theta|\Theta^{(t)})$ .

$$\mu_k^{t+1} = \frac{\sum_{i=1}^N x_i P(y_i^k = 1|x_i; \Theta^{(t)})}{\sum_{i=1}^N P(y_i^k = 1|x_i; \Theta^{(t)})} \quad (6)$$

$$\sigma_k^{t+1} = \frac{\sum_{i=1}^N P(y_i^k = 1|x_i; \Theta^{(t)}) \{(x_i - \mu_k^{t+1})(x_i - \mu_k^{t+1})^T\}}{\sum_{i=1}^N P(y_i^k = 1|x_i; \Theta^{(t)})} \quad (7)$$

$$\alpha_k^{t+1} = \frac{1}{N} \sum_{i=1}^N P(y_i^k = 1|x_i; \Theta^{(t)}) \quad (8)$$

On convergence, the maximum likelihood estimation of labels  $\{y_i\}_{i=1}^N$  represent the fine class that the roughly labeled region belongs to.

In practice, three problems hampered the application of EM algorithm in fine annotation. First, the performance of EM algorithm depends strongly on the choice of the initial parameters  $\Theta^{(0)}$ . Second, a premise of EM algorithm is that the number of component densities  $K$  is known or pre-defined. In our case, although we have obtained plenty of detailed keywords, those words are not comprehensive yet and their number might be less than the real component number  $K$ . Third, the fine label of a region is directly affected by its rough label due to the hierarchical relationship between the general keywords and the detailed keywords. If a region has been labeled falsely with a general keyword, no matter how strong the fine classifier is, the region cannot be recognized correctly.

In order to decide the component number  $K$  and the initial parameters, and to decrease the negative influence from rough annotation, we propose a semi-supervised active EM algorithm that adopts the active learning technique to select the most informative samples to train the EM classifier and retrain the rough classifier. The whole algorithm consists of 10 steps.

**Step 1.** Prepare some training samples  $X_l$  that have been manually labeled with general and detailed keywords, and involve at least two samples for each fine class. Prepare a larger amount of segmented regions  $X_u = \{x_i\}_{i=1}^n$ , which have been roughly labeled by DAGSVM but not finely labeled. Prepare  $X_d$  for retraining DAGSVM classifier, and initialize it as  $X_d = \emptyset$ .

**Step 2.** Assuming that  $K$  fine semantic classes have been defined under the rough category  $C_r$ , we use  $X_l \in C_r$  to train the  $K$ -components GMM of  $C_r$  and obtain its parameters  $\Theta$  from (6) ~ (8).

**Step 3.** The trained GMM is used to classify the  $X_u$  that has been roughly labeled as  $C_r$ . Their posterior probabilities are  $P(y_i^k = 1|x_i; \Theta)$  for all  $k \in [1, K]$  by (5).

**Step 4.** If a region  $x_m \in X_u$  has  $\max_{k=1, \dots, K} P(y_i^k = 1|x_m; \Theta) < T_f$ , the region is regarded as unsure and requires manual annotation.

**Step 5.** If the rough annotation that is manually labeled to  $x_m$  is different from  $C_r$ , set  $X_d = X_d \cup x_m$ ,  $X_u = X_u - x_m$ , and  $X_l = X_l \cup x_m$ .

**Step 6.** If the manual rough label of  $x_m$  is  $C_r$ , but its fine label is different from currently existed detailed keywords,  $x_m$  is regarded as the representative sample of new fine-class. Set  $K = K + 1$ ,  $X_l = X_l \cup x_m$ , and  $X_u = X_u - x_m$ .

**Step 7.** If a region  $x_m \in X_u$  has  $\max_{k=1, \dots, K} P(y_i^k = 1|x_m; \Theta) \geq T_f$ , then  $x_m \in$  fine-class  $j$ , where  $P(y_i^j = 1|x_m; \Theta) = \max_{k=1, \dots, K} P(y_i^k = 1|x_m; \Theta)$ .

**Step 8.** Return to Step 2 until all regions in  $X_u \in C_r$  being finely classified.

**Step 9.** When all  $X_u$  are finely classified, we check the set  $X_d$ . If  $X_d \neq \emptyset$ ,  $X_d$  along with  $X_l$  are used to retrain the DAGSVM rough classifier.

**Step 10.** Return to Step 2 to retrain the GMM model of each rough category with renewed set  $X_l$  and  $K$ .

In the above-mentioned course, Step 1 is for initialization; Steps 2 ~ 8 train the fine classifier and complete fine annotation; Step 9 retrains the rough classifier; and Step 10 retrains the fine classifier.  $T_f$  is the threshold deciding the sample numbers of manual annotation. If  $T_f$  is too small, not enough informative images could be selected for manual annotation. If  $T_f$  is too large, too many images would be regarded as informative and too much manual annotation is required. The  $T_f$  for different rough categories may be different and commonly complex rough classes need larger

$T_f$  so as to obtain a preferable accuracy with acceptable labor cost.

### 3 Annotation correction by contextual relationship

Due to the segmentation error, unusual lighting conditions, and similar appearance of different scenery objects, recognizing an isolated region is error-prone. Semantic context, although received comparatively little attention, can play an important role in reducing ambiguity and error<sup>[18]</sup>. Here, we investigate the coexistence relationship and relative location relationship, and use them to judge or even revise the improper annotation. Coexistence judges whether two objects could coexist in an image. For example, sea is more often associated with sandbeach but less often with desert. If an image was labeled as “sea, sky, desert”, we would like to revise it as “sea, sky, sandbeach”. Relative location judges the annotation of a region according to its relative location to its neighbors. For example, sea and blue sky sometimes have similar visual features. However, if the upper part of an image is labeled with sea and the lower part is labeled as sky, in common sense, we regard such annotation as illogical.

The 800 images that have been used in questionnaire in Subsection 1.1 are adopted here to extract the basic probability information of coexistence and relative location relationship. For any pair of rough categories  $RC_i$  and  $RC_j$ , we calculate their coexistence probability  $p(RC_i, RC_j|RC_j)$ . If the coexistence probability of two rough classes is not zero, we calculate their relative location probability  $p(L_{i,j}|RC_i, RC_j)$ , which reflects the probability of two rough categories in relative location  $L_{i,j}$ . In order to simplify the probability calculation, we only set two kinds of relative locations: upper and lower. If  $L_{i,j}$  is upper,  $p(L_{i,j}|RC_i, RC_j)$  represents the probability that  $RC_i$  is higher than  $RC_j$ . Then, we list all subclasses of the pair of coexistent rough classes and calculate the coexistence probability between any pair of subclasses  $p(FC_i, FC_j|FC_j)$ . If the coexistence probability of two fine classes is not zero, we calculate their relative location probability  $p(L_{i,j}|FC_i, FC_j)$ . The process of computing the contextual probability information is shown in Fig. 2.

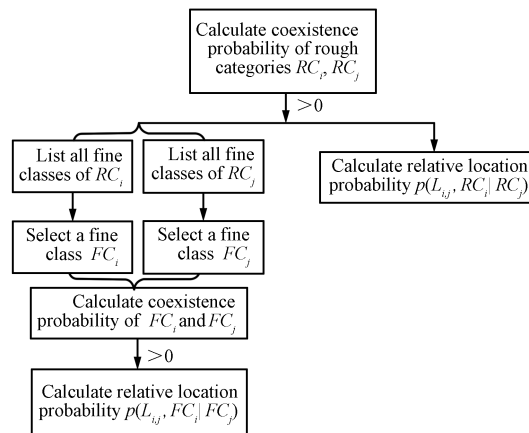


Fig. 2 The process of computing the probability information

Coexistence probabilities of five rough categories—stone, water, grass, sky, mountain, and five fine classes—desert, waterfall, green grass, blue sky, snow mountain are listed in Tables 2 and 3, respectively.

Table 2 Coexistence probabilities of five rough categories

$p(RC_i, RC_j   RC_j)$	$RC_{j=1}$	$RC_{j=2}$	$RC_{j=3}$	$RC_{j=4}$	$RC_{j=5}$
$RC_{i=1}$	—	0.232	0.133	0.113	0.063
$RC_{i=2}$	0.509	—	0.217	0.306	0.303
$RC_{i=3}$	0.4	0.353	—	0.452	0.493
$RC_{i=4}$	0.527	0.554	0.584	—	0.951
$RC_{i=5}$	0.136	0.201	0.248	0.308	—

Notes.  $RC_1$  stands for stone,  $RC_2$  stands for water,  $RC_3$  stands for grass,  $RC_4$  stands for sky, and  $RC_5$  stands for mountain.

Table 3 Coexistence probabilities of five fine classes

$p(FC_i, FC_j   FC_j)$	$FC_{j=1}$	$FC_{j=2}$	$FC_{j=3}$	$FC_{j=4}$	$FC_{j=5}$
$FC_{i=1}$	—	0	0.014	0.076	0
$FC_{i=2}$	0	—	0.053	0.007	0
$FC_{i=3}$	0.154	0.533	—	0.348	0.25
$FC_{i=4}$	0.731	0.067	0.372	—	0.659
$FC_{i=5}$	0	0	0.029	0.113	—

Notes.  $FC_1$  stands for desert,  $FC_2$  stands for waterfall,  $FC_3$  stands for green grass,  $FC_4$  stands for blue sky, and  $FC_5$  stands for snow mountain.

According to the statistical data, if the coexistence or relative location probability of two keywords is “0”, and they are labeled in an image or with the relative location, we will discard or substitute one of the two labels.

## 4 Experiments and results

A prototype system has been developed using Matlab platform on a Pentium 2.0GHz PC running Windows XP operating system. The 800 images that have been used in Subsection 1.1 and Section 3, as well as another 500 scenery images selected from the Corel stock photo library constitute our image database, which involve various themes and contents, such as field, waterfall, sunset, flowers, and desert. We manually cropped 390 single-object regions from 300 images and labeled them with general and detailed keywords (as shown in Table 4). Since rough classes “vehicle” and “animal” as well as their subclasses are mainly dependent on shape feature for classification and recognition, we disregard the two classes in this study and only make training samples for the rest 10 rough classes, in which each subclass has at least 2 samples. To obtain exact visual description, those regions are required to be cropped as large as possible. The visual features of these samples are then used to pre-train the DAGSVM rough classifier and to find the best parameters of  $C$  and  $\gamma$  by the 3-fold cross-validation and grid-search algorithm. Here,  $C$  and  $\gamma$  are selected from exponentially growing sequences  $C \in [2^{-3}, 2^{-1}, \dots, 2^{11}]$  and  $\gamma \in [2^{-11}, 2^{-9}, \dots, 2^3]$ , respectively. The experimental results show that the pair  $[2^7, 2^{-11}]$  gives the maximum 3-fold cross-validation accuracy (as shown in Table 5). Therefore, it is selected as the optimal parameters and used in our system. The 390 regional images also form set  $X_l$  for pretraining the EM fine classifier.

Then, the rest 1000 images are divided into two parts evenly. One part is used for validation, and the other part is used for testing. In the validation phase, 500 images are divided into five sets randomly. Each set includes 100 images. We adopt the 5-fold cross-validation to estimate annotation accuracy of the hierarchical annotation scheme. First, all images are segmented by spatially constrained mixture model. Although different images may have different numbers of segments, we set their initial region number as  $c = 5$

uniformly. After segmentation, if a region is smaller than a given threshold  $T_{\min}$ , it will be disregarded in respect that visual features of a small region cannot represent the scenery object well and easily engender recognition error. Here, we set  $T_{\min}$  as 2.5% of the image area empirically. Second, for segmented regions of 100 training images of a training-test partition, we calculate their rough visual features and use the pre-trained DAGSVM to find their rough labels. After logical correction, the regions labeled with the same general keyword  $C_r$  form the set  $X_u \in C_r$ . Following Steps 3 ~ 8 of the semi-supervised active EM algorithm, we obtain the fine labels of these regions. Following Steps 9 and 10, the DAGSVM rough classifier and the EM fine classifier are retrained. Third, we apply the retrained DAGSVM and EM classifiers to annotate the 400 test images. After logical correction, we compute their rough and fine annotation accuracies. In the same way, the annotation accuracies of the other 4 training-test partitions are calculated. It is worth noting that in this experiment the retrained DAGSVM rough classifier in the preceding training-test trial is regarded as the pretrained DAGSVM classifier of the next trial, meanwhile the renewed  $X_l$  and  $K$  in the preceding training-test trial are regarded as the initial  $X_l$  and  $K$  of the next trial. The rough annotation accuracies of the 5 trials and their average accuracy are 80.62%, 81.36%, 81.97%, 82.38%, 82.74%, and 81.814%, respectively. The fine annotation accuracies of the 5 trials and their average accuracy are 65.28%, 66.87%, 67.52%, 68.24%, 68.61%, and 67.304%, respectively. In the testing phase, the renewed DAGSVM and EM classifiers updated in the last training-test trial are applied to annotate 500 test images. Three test images, their segmentation results, step by step automatic annotation, and manual labels are listed in Table 6.

Table 4 Cropped samples of rough classes ground, sky, and water

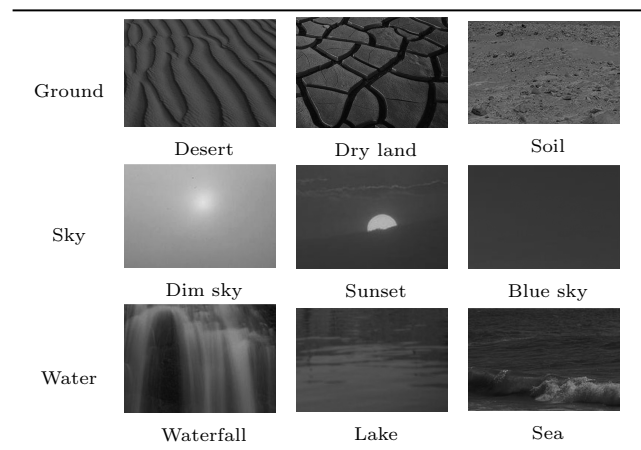








Table 5 The results of grid-search using 3-fold cross-validation

$C$	$\gamma$							
	$2^{-11}$	$2^{-9}$	$2^{-7}$	$2^{-5}$	$2^{-3}$	$2^{-1}$	$2^1$	$2^3$
$2^{-3}$	68.57	68.57	68.57	68.57	68.57	56.33	39.2	26.2
$2^{-1}$	69.23	70.2	71.93	73.47	66.97	58.67	39.2	27.83
$2^1$	70.57	72.63	73.87	68.13	66.2	52.97	40.73	27.83
$2^3$	75.03	76.37	74.6	67.5	64.67	52.97	40.73	27.83
$2^5$	78.5	77.03	71.6	67.13	65.53	52.97	40.73	27.83
$2^7$	80.17	72.43	67.03	66.87	65.53	52.97	40.73	27.83
$2^9$	76.93	68.27	66.97	66.4	65.53	52.97	40.73	27.83
$2^{11}$	67.47	66.53	66.2	65.63	65.63	52.97	40.73	27.83

Table 6 Automatic and manual annotation results of three scenery images

Original images			
Segmented images			
SVM rough	Sky, tree, mountain	Mountain, water, road, tree	Water, grass, stone, ground
Rough-label correction	Sky, tree, mountain	Mountain, sky, road, tree	Water, grass, stone, ground
EM fine	Blue sky, stub, snow mountain	Hill, blue sky, highway, green tree	Waterfall, green grass, pebble, soil
Fine-label correction	Blue sky, stub, snow mountain	Hill, blue sky, highway, green tree	Waterfall, green grass, pebble
Manual annotation	Blue sky, pine, snow mountain, snow field	Hill, blue sky, white clouds, highway, green grass, greentree	Waterfall, green grass, green mountain, pebble, boulder

From Table 6, we find that although automatic annotation is not as detailed as manual annotation, it is usually exact, especially the rough annotation. However, the beneficial effect of logical correction is not prominent. It corrects some errors in annotation results, but sometimes it deletes the correct annotation. To our knowledge, one possible reason is that the context relationship defined in this paper is incomplete and greatly dependent on the training samples. The more the training images are, the more representative the context relationship is.

We also developed the retrieval function upon the annotation system, which brings forward images with at least one region having been labeled with the query keyword. Retrieval precisions  $P_r = N_c/N_r$  and recalls  $R_e = N_c/N_e$  of 6 pairs of general and detailed keywords (sky, blue sky, tree, bush, grass, withered grass, ground, desert, water, waterfall, mountain, snow mountain) are calculated, where  $N_c$  is the number of correctly retrieved images,  $N_r$  is the number of images that have been labeled with the keyword by the annotation system,  $N_e$  is the number of images that contain the keyword in their actual annotation (manual annotation). The recall-precision (R-P) curves of these general and detailed keywords are shown in Figs. 3 and 4.

Fig. 3 shows that the retrieval precisions of general keywords sky, tree, and grass are much higher than that of mountain under the same recall rate. We also did retrieval tests for road, stone, flower, and building. However, their R-P curves are similar to or even lower than that of mountain. We boil down the reasons into two sides. Firstly, shape feature that plays an important role in describing objects such as buildings and flowers, has not been introduced in this study considering that exact shape description always has low generality and is fragile to uncomplete profile, while rough shape description usually has low differentiating ability. Second, some subclasses of different rough classes have similar appearance, which increases the difficulty of rough annotation. For example, "flagstone" in stone class looks like "flagging" in road class; "stairway" in road class looks similar to "stairs" in building class; the

mountain covered with trees is easily mistaken as trees only, while the barren mountain is easily misunderstood as ground. Besides, because of the annotation errors, the recall rates of R-P curves of all general keywords cannot reach to 1.0. Here, we set the maximum recall of the 6 general keywords to 0.9 equally.

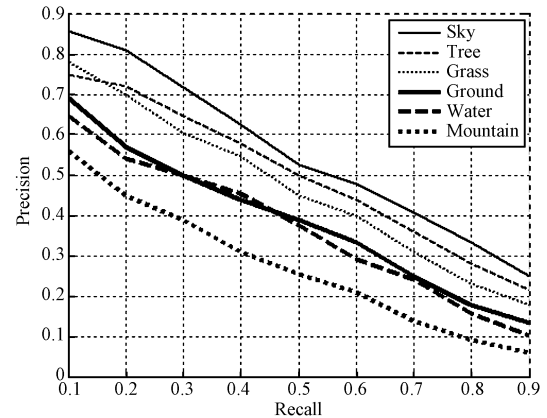


Fig. 3 The recall-precision curves of 6 general keywords

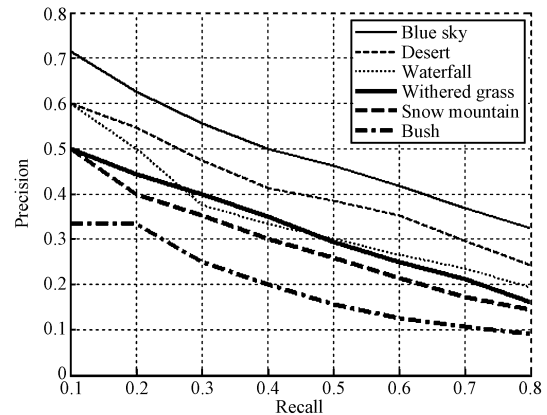


Fig. 4 The recall-precision curves of 6 detailed keywords

As for the detailed keywords retrieval, since under the hierarchical annotation structure the available keywords of fine annotation are restricted to the rough annotation results, the maximum recall rates of many detailed keywords are less than that of their corresponding general keywords. Despite this, the R-P curves of some detailed keywords are higher than those of their corresponding general keywords, which indicates that our system is valuable in real applications as people usually hope to obtain detailed and exact image description.

Finally, we give some comments to the time cost and feasibility of the annotation scheme. It is obvious that the training phase of the scheme is quite onerous and time-consuming. It includes getting parameters of DAGSVM classifier with 390 manually cropped single-object regions, training EM and DAGSVM classifiers with 500 images by 5-fold cross validation, and calculating coexistence probability and relative location probability manually by using 800 images. However, its test process is comparatively fast. Usually it takes us 18s to segment a test image, 1.7s to roughly annotate it, and 1.5s more to get its fine annotation. Furthermore, the annotation time can be greatly reduced if the algorithm is implemented in C language.

## 5 Conclusions and discussions

Automatic image annotation has been investigated for many years. In order to clearly annotate scenery images, in this paper, a hierarchical annotation scheme was presented that consists of five steps: image segmentation, rough annotation, auto-correction of rough labels, fine annotation, and fine-labels correction. The scheme not only accords with human's rough-to-fine hierarchical understanding process, but also effectively decreases annotation errors due to lingual diversities and cognition differences as well as various image scales and photographing angles. Experiments have been performed on 1000 scenery images with some encouraging results being achieved.

In the future, we will investigate more visual features, especially shape description, and combine them with current features to describe regions and improve annotation accuracy. In addition, although we have paid attention to keywords selection and constructed a hierarchical relationship between general and detailed keywords, we found that the same general keyword may have different sets of subclass modes. For example, the rough category "flower" may include rose, tulip, chrysanthemum by species definition; or red flower, white flower, varicolored flower by color feature; or big flower, small flower, clustering flower based on shape feature. To enlarge the application range of the annotation scheme, we intend to apply the ontology technique to organize and describe hierarchical keywords.

## References

- 1 Carneiro G, Chan A B, Moreno P J, Vasconcelos N. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, **29**(3): 394–410
- 2 Mori Y, Takahashi H, Oka R. Image-to-word transformation based on dividing and vector quantizing images with words. In: Proceedings of the International Workshop on Multimedia Intelligence Storage and Retrieval Management. Orlando, USA: Springer, 1999. 1–9
- 3 Zhang Q N, Izquierdo E. Adaptive salient block-based image retrieval in multi-feature space. *Image Communication*, 2007, **22**(6): 591–603
- 4 Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. Toronto, Canada: ACM, 2003. 119–126
- 5 Liu J, Wang B, Lu H, Ma S. A graph-based image annotation framework. *Pattern Recognition Letters*, 2008, **29**(4): 407–415
- 6 Kokkinos I, Maragos P. Synergy between object recognition and image segmentation using the expectation-maximization algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(8): 1486–1501
- 7 Jung J J. Exploiting semantic annotation to supporting user browsing on the web. *Knowledge-Based Systems*, 2007, **20**(4): 373–381
- 8 Fan J P, Gao Y L, Luo H Z, Xu G Y. Statistical modeling and conceptualization of natural images. *Pattern Recognition*, 2005, **38**(6): 865–885
- 9 Spirkovska L. A Summary of Image Segmentation Techniques, NASA Technical Memorandum 104022, Ames Research Center, USA, 1993
- 10 Blekas K, Likas A, Galatsanos N, Lagaris I. A spatially-constrained mixture model for image segmentation. *IEEE Transactions on Neural Network*, 2005, **16**(2): 494–498
- 11 Manjunath B S, Ma W Y. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996, **18**(8): 837–842
- 12 Tamura H, Mori S, Yamawaki T. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 1978, **8**(6): 460–473
- 13 Liu C, Wechsler H. Independent component analysis of Gabor features for face recognition. *IEEE Transactions on Neural Networks*, 2003, **14**(4): 919–928
- 14 Vapnik V. *Statistical Learning Theory*. New York: Wiley, 1998
- 15 Duan K, Keerthi S S, Poo A N. Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing*, 2003, **51**: 41–59
- 16 Platt J C, Cristianini N, Shawe-Taylor J. Large margin DAGs for multiclass classification. *Advances in Neural Information Processing Systems*. Cambridge: MIT Press, 2000. 547–553
- 17 Bilmes J A. A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Technical Report ICSI-TR-97-021, University of Berkeley, USA, 1997
- 18 Rabinovich A, Vedaldi A, Galleguillos C, Wiewiora E, Belongie S. Objects in context. In: Proceedings of the 11 International Conference on Computer Vision. San Diego, USA: IEEE, 2007. 1–8



**GAO Yan-Yu** Received her Ph. D. degree from University of Science and Technology Beijing, China in 2004. From 2005 to 2007, she was at Satellite Venture Business Laboratory, Muroran Institute of Technology, Japan as a post-doctor fellow. Her research interest covers image understanding, computer vision, and Kansei engineering. Corresponding author of this paper. E-mail: gaoyy@ustb.edu.cn



**YIN Yi-Xin** Received his B. S., M. S., and Ph. D. degrees from University of Science and Technology Beijing (USTB), China in 1982, 1984, and 2002, respectively. He is currently a professor at the School of Information Engineering, USTB. His research interest covers intelligent control, adaptive control, Kansei engineering, and artificial life. E-mail: yyx@ies.ustb.edu.cn



**UOZUMI Takashi** Received his B. E. degree from Muroran Institute of Technology, Japan in 1973 and his Ph. D. degree from Hokkaido University, Japan in 1980. Since 1988, he has been an associate professor in the Department of Computer Science and System Engineering, Muroran Institute of Technology. His research interest covers evolutionary algorithms, image recognition, computer vision, and machine learning. E-mail: uozumi@epsilon2.csse.muroran-it.ac.jp