Multi-Level Annotation of Natural Scenes Using Dominant Image Components and Semantic Concepts

Jianping Fan
Dept of Computer Science
UNC-Charlotte
Charlotte, NC 28223, USA
jfan@uncc.edu

Yuli Gao
Dept of Computer Science
UNC-Charlotte
Charlotte, NC 28223, USA
ygao@uncc.edu

Hangzai Luo
Dept of Computer Science
UNC-Charlotte
Charlotte, NC 28223, USA
hluo@uncc.edu

ABSTRACT

Automatic image annotation is a promising solution to enable semantic image retrieval via keywords. In this paper, we propose a multi-level approach to annotate the semantics of ***natural scenes*** by using both the dominant image components (salient objects) and the relevant semantic concepts. To achieve automatic image annotation at the content level, we use salient objects as the dominant image components for image content representation and feature extraction. To support automatic image annotation at the concept level, a novel image classification technique is developed to map the images into the most relevant semantic image concepts. In addition, Support Vector Machine (SVM) classifiers are used to learn the detection functions for the pre-defined salient objects and finite mixture models are used for semantic concept interpretation and modeling. An ***adaptive EM algorithm*** has been proposed to determine the optimal model structure and model parameters simultaneously. We have also demonstrated that our algorithms are very effective to enable multi-level annotation of ***natural scenes*** in a large-scale image dataset.

Categories and Subject Descriptors
I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis-*object recognition*, H.2.8 [**Database Management**]: Database Applications - *image databases.*

General Terms
Algorithms, Measurement, Experimentation

Keywords: Automatic image annotation, salient objects, adaptive EM algorithm.

1. INTRODUCTION

With the exponential growth in personal digital image libraries, the need of automatic or semi-automatic image annotation is becoming increasely important to enable more effective image retrieval via keywords [1-3]. The image similarity on semantics can be categorized into two major classes [33]: (a) similar image components (e.g., sky, grass [21-22]) or similar global visual properties (i.e., openness, naturalness [18]); (b) similar semantic concepts (e.g., garden, beach, mountain view) or similar abstract concepts (e.g., image events such as sailing, skiing [33]).

To support image retrieval via keyword, there is an urgent need of research efforts to develop new techniques for enabling automatic image annotation. However, the performance of most existing techniques for automatic image annotation largely depends on two issues: (1) The effectiveness of image patterns for image content representation and feature extraction; (2) The accuracy of the underlying image classification algorithms. Many semantic image classification techniques have been proposed in the literatures [4-20]. The limitation of pages does not allow us to survey all these works. Instead we try to emphasize some of these works that are most relevant to our proposed work.

To support more effective image content representation, Carson *et al.* proposed a blob-based framework [9], but this approach will bring additional burdens for users to select important regions for calculating the similarity between images. To measure the similarity between images effectively, Smith *et al.* proposed a system that converts the images into region strings [4]. To achieve automatic image annotation at the concept level, Gaussian mixture model (GMM) with a pre-defined model structure (a fixed number of mixture components) has been used for semantic concept interpretation and modeling by using blobs for image content representation and feature extraction [5-6,30]. However, different semantic concepts may have different model structures to interpret their contextual relationships with the most relevant image patterns, thus there is an urgent need to develop new techniques that are able to obtain the underlying optimal model structures automatically. Li *et al.* have also developed an integrated region matching technique for binary image classification [12]. One common weakness of these region-based image classification techniques is that homogeneous image regions have little correspondence with the semantic concepts; thus they are not effective for multi-class image classification. In addition, these region-based approaches suffer from the problem of over-detection of semantic concepts [21-22].

Without using image segmentation, image-based approaches are very attractive to enable a low-cost framework for feature extraction and image classification [13-16]. Support Vector Machine (SVM) classifiers were recently developed for semantic image classification because of their good per-

MM'04, October 10-16, 2004, New York, New York, USA.

formance in high-dimensional feature space [15-16,23]. Since only the global visual properties are used for image content representation [18], the image-based approaches do not work very well for the images that consist of individual objects, especially when the individual objects are used by human beings to interpret the semantics of images [21-22].

As mentioned above, the image semantics can be described in multiple levels (i.e., both the content level and the concept level). Thus a good image classification and annotation scheme should enable the annotation of both the dominant image components and the relevant semantic concepts. However, few existing work has achieved such multi-level annotation of images [28]. In addition, few existing work has provided an intuitive solution for semantic concept interpretation. With these as our motivation, we propose a novel framework to enable more effective interpretation of semantic concepts and multi-level annotation of ***natural scenes***.

In this paper, we have proposed a novel framework for multi-level image annotation by using salient objects for image content representation and finite mixture models for semantic concept modeling, where an adaptive EM algorithm has been developed for achieving optimal model selection and parameter estimation simultaneously.

This paper is organized as follows: Section 2 presents a learning-oriented technique for automatic salient object detection; Section 3 presents our semantic image classification and multi-level image annotation algorithm; Section 4 shows the experimental results to evaluate the performance of our techniques; We conclude in Section 5.

2. SALIENT OBJECT DETECTION

The salient objects are defined as the visually distinguishable image components [21-22] or the global visual properties of whole images that can be identified by using the spectrum templates in the frequency domain [18]. For example, the salient object "sky" is defined as the connected image regions with large sizes (i.e., dominant image regions) that are related to the human semantics "sky". The salient objects that are related to the global visual properties in the frequency domain can be obtained easily by using wavelet transformation [18]. In the following discussion, we will focus on modeling and detecting the salient objects that correspond to the visually distinguishable image components. In addition, the *basic vocabulary* of such salient objects can be obtained by using the taxonomy of the dominant image components of *natural scenes*.

We have already implemented 32 functions to detect 32 types of salient objects in natural scenes, and each function is able to detect a certain type of these salient objects in the basic vocabulary. Each detection function consists of three parts: (a) automatic image segmentation by using the mean shift technique [25]; (b) image region classification by using the SVM classifiers with an optimal model parameter search scheme; (c) label-based region aggregation for automatic salient object generation.

We use our detection function of the salient object "grass" as an example to show how we can design our detection functions. As shown in Fig. 1, image regions with homogeneous color or texture are first obtained by using the mean shift techniques [25]. Since the visual properties of a certain type of salient object may look different at different lighting and capturing conditions [21], using only one image is insufficient to represent its visual characteristics. Thus this automatic image segmentation procedure is performed on a set of training images which consist of the salient object "grass".

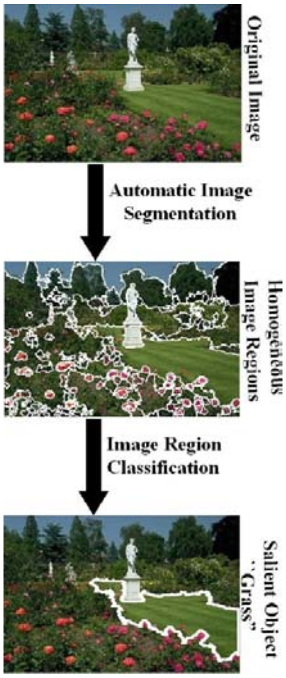
Original Image
Automatic Image Segmentation
Homogeneous Image Regions
Image Region Classification
Salient Object "Grass"

Figure 1: The flowchart for our automatic detection functions of salient objects.

The homogeneous regions in the training images, that are related to the salient object "grass", are selected and labeled as the training samples by human interaction. Region-based low-level visual features, such as 1-dimensional coverage ratio (i.e., density ratio) for a coarse shape representation, 6-dimensional region locations (i.e., 2-dimensions for region center and 4-dimensions to indicate the rectangular box for a coarse shape representation), 7-dimensional LUV dominant colors and color variances, 14-dimensional Tamura texture, and 28-dimensional wavelet texture features, are extracted for characterizing the visual properties of these labeled image regions that are explicitly related to the salient object "grass". The 6-dimensional region locations are used to determine the spatial contexts among different types of salient objects to avoid the wrong detection of the visual similar salient objects such as "beach sand" and "road sand".

We use *one-against-all* rule to label the training samples $\Omega_{g_j} = \{X_l, L_j(X_l)|l = 1, \cdots, N\}$: positive samples for the specific salient object "grass" and negative samples. Each labeled training sample is a pair $(X_l, L_j(X_l))$ that consists of a set of region-based low-level visual features X_l and the semantic label $L_j(X_l)$ for the corresponding labeled homogeneous image region.

The image region classifier is learned from these available labeled training samples. We use the well-known SVM classifiers for binary image region classification [23]. Consider a binary classification problem with linearly separable sample set $\Omega_{g_j} = \{X_l, L_j(X_l)|l = 1, \cdots, N\}$, where the semantic label $L_j(X_l)$ for the labeled homogeneous image region with the visual feature X_l is either $+1$ or -1. For the positive samples X_l with $L_j(X_l) = +1$, there exists the transformation parameters Λ and b such that $\Lambda \cdot X_l + b > +1$. Similarly, for negative samples X_l with $L_j(X_l) = -1$, we

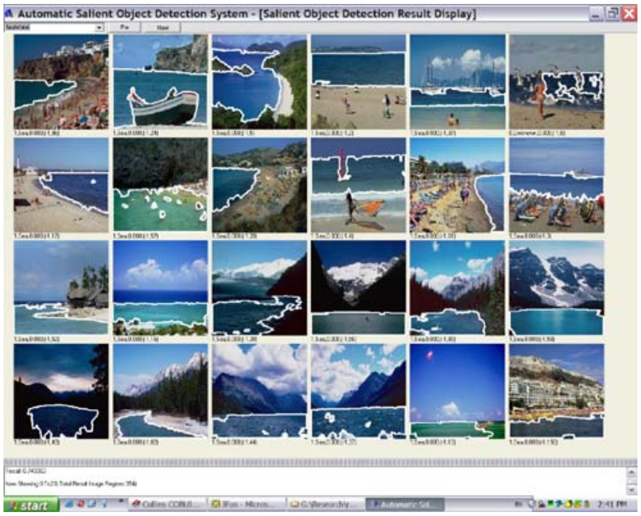
Automatic Salient Object Detection System - [Salient Object Detection Result Display]
start

Figure 2: The detection results of the salient object "water".

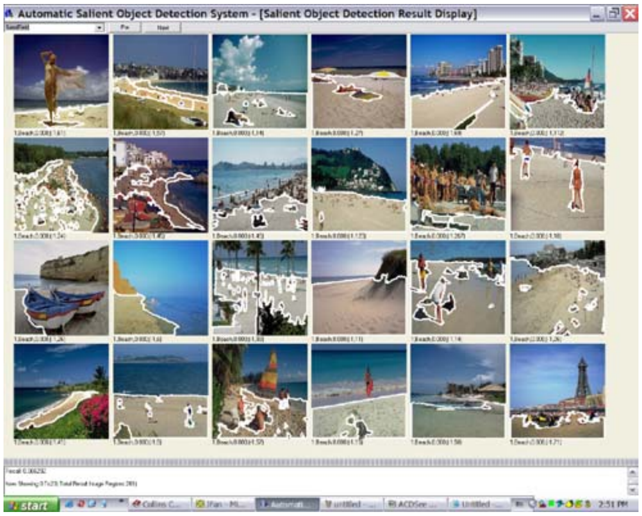
Automatic Salient Object Detection System - [Salient Object Detection Result Display]
start

Figure 3: The detection results of the salient object "sand field".

have $\Lambda \cdot X_l + b < -1$. The margin between these two supporting planes will be $2/||\Lambda||^2$. The SVM classifier is then designed for maximizing the margin with the constraints $\Lambda \cdot X_l + b > +1$ for the positive samples and $\Lambda \cdot X_l + b < -1$ for the negative samples.

Given the training set $\Omega_{g_j} = \{X_l, L_j(X_l) | l = 1, \cdots, N\}$, the margin maximization procedure is then transformed into the following optimization problem:

$$\underset{\Lambda, b, \xi}{arg\ \ min} \quad \frac{1}{2}\Lambda^T \cdot \Lambda + C \sum_{l=1}^{N} \xi_l \tag{1}$$

$$L_j(\Lambda \cdot \Phi(X_l) + b) \geq 1 - \xi_l$$

where $\xi_l \geq 0$ represents the training error rate, $C > 0$ is the penalty parameter to adjust the training error rate and the regularization term $\frac{\Lambda^T \cdot \Lambda}{2}$, $\Phi(X_l)$ is the function that maps X_l into higher-dimensional space (i.e., feature dimensions plus the dimension of response) and the kernel function is defined as $\kappa(X_i, X_j) = \Phi(X_i)^T \Phi(X_j)$. In our current implementation, we select radial basis function (RBF), $\kappa(X_i, X_j) = exp(-\gamma||X_i - X_j||^2), \gamma > 0$.

We have developed an efficient search algorithm to determine the optimal model parameters (C, γ) for the SVM classifiers: (a) The labeled image regions are partitioned into ν subsets in equal size, where $\nu - 1$ subsets are used for classifier training and the remaining one is used for classifier validation. (b) Our feature set for image region representation is first normalized to avoid the features in greater numeric ranges that dominate those in smaller numeric ranges. Because inner product is usually used to calculate the kernel values, this normalization procedure is able to avoid the numerical problem. (c) The numeric ranges for the parameters C and γ are exponentially partitioned into small pieces with M pairs. For each pair, $\nu - 1$ subsets are used to train the classifier model. When the M classifier models are available, cross-validation is then used to determine the underlying optimal parameter pair (C, γ). (d) Given the optimal parameter pair (C, γ), the final classifier model (i.e., support vectors) is trained again by using the whole training data set. (e) The spatial contexts among different types of salient objects (i.e., coherence among different types of salient objects) have also been used to cope well with the wrong detection problem for the visual similar salient objects [21].

Table 1: The average performance of some detection functions (precision ρ versus recall ϱ).

salient objects	brown horse	grass	purple flower
ρ	95.%6	92.9%	96.1%
ϱ	100%	94.8%	95.2%
salient objects	red flower	rock	sand field
ρ	87.8%	98.7%	98.8%
ϱ	86.4%	100%	96.6%
salient objects	water	human skin	sky
ρ	86.7%	86.2%	87.6%
ϱ	89.5%	85.4%	94.5%
salient objects	snow	sunset/sunrise	waterfall
ρ	86.7%	92.5%	88.5%
ϱ	87.5%	95.2%	87.1%
salient objects	yellow flower	forest	sail cloth
ρ	87.4%	85.4%	96.3%
ϱ	89.3%	84.8%	94.9%
salient objects	elephant	cat	zebra
ρ	85.3%	90.5%	87.2%
ϱ	88.7%	87.5%	85.4%

Some results for our detection functions are shown in Fig. 2 and Fig. 3. From these experimental results, one can find that the salient objects are more representative than the homogeneous image regions and the major visual properties of the dominant image components are maintained by using the salient objects for image content representation. Thus using the salient objects for feature extraction can enhance the quality of features and result in more effective semantic image classification. In addition, the salient objects can be visually distinguishable, and they are also semantic to human beings. Thus the keywords for interpreting the salient objects can also be used to achieve the annotations of the images at the content level. The average performance for some detection functions is given in Table 1.

It is worth noting that the procedure for salient object detection is automatic and the human interaction is only involved in the procedure to label the training samples (i.e., homogeneous image regions) for learning the detection functions. After the salient objects are extracted automatically from the images, a set of visual features are then calculated to characterize their visual properties. These visual features include 1-dimensional coverage ratio (i.e., density ratio) for a coarse shape representation, 6-dimensional object locations (i.e., 2-dimensions for object center and 4-dimensions to indicate the rectangular box for a coarse shape representation of salient object), 7-dimensional LUV dominant colors and color variances, 14-dimensional Tamura texture, and 28-dimensional wavelet texture features.

3. SEMANTIC IMAGE CLASSIFICATION AND MULTI-LEVEL ANNOTATION

In order to achieve the annotations of the images at the concept level, we have also proposed a novel semantic image classification technique. To exploit the contextual relationships between the semantic concepts and the relevant salient objects [32], we use the finite mixture model (FMM) to approximate the class distribution of the salient objects that are relevant to a specific semantic concept C_j:

$$P(X, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}) = \sum_{i=1}^{\kappa} P(X|S_i, \theta_{s_i})\omega_{s_i} \qquad (2)$$

where $P(X|S_i, \theta_{s_i})$ is the ith multivariate mixture component with n independent means and a common $n \times n$ covariance matrix, κ indicates the optimal number of mixture components, $\Theta_{c_j} = \{\theta_{s_i}, i = 1, \cdots, \kappa\}$ is the set of the parameters for these mixture components, $\omega_{c_j} = \{\omega_{s_i}, i = 1, \cdots, \kappa\}$ is the set of the relative weights among these mixture compoments, and X is the n-dimensional visual features that are used for representing the relevant salient objects.

The visual properties of a certain type of the salient objects may look different at different lighting and capturing conditions. For example, the salient object “sky” consists of various appearances, such as “blue sky pattern”, “white(clear) sky pattern”, “cloudy sky pattern”, and “sunset/sunrise sky pattern”, which have very different properties on color and texture under different viewing conditions [21]. Thus, the data distribution for a certain type of salient object is approximated by using multiple mixture components to accomodate the variability of the same type of salient object (i.e., presence/absence of distinctive parts, variability in overall shape, changing of visual properties due to lighting conditions, viewpoints etc).

The optimal model structure and parameters $(\hat{\kappa}, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j})$ for a specific semantic concept are determined by:

$$\left(\hat{\kappa}, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j}\right) = \underset{\kappa, \omega_{c_j}, \Theta_{c_j}}{arg \;\; max} \left\{L(C_j, X, \kappa, \omega_{c_j}, \Theta_{c_j})\right\} \qquad (3)$$

where $L(C_j, X, \kappa, \omega_{c_j}, \Theta_{c_j}) = -\log P(X, C_j, \Theta_{c_j}) + \log p(\Theta_{c_j})$ is the objective function, $-\log P(X, C_j, \kappa, \omega_{c_j}, \Theta_{c_j})$ is the likelihood function as described in Eq. (2), and $\log p(\Theta_{c_j})$ $= -\frac{n+\kappa+3}{2}\sum_{l=1}^{\kappa} \log \frac{N\omega_l}{12} - \frac{\kappa}{2}\log\frac{N}{12} - \frac{\kappa(N+1)}{2}$ is the minimum description length (MDL) term to penalize the complex models [36], N is the total number of training samples and n is the dimensions of visual features. The estimate of maximum likelihood can be achieved by using the EM algorithm [26,30-31]. However, the EM algorithm and its recent variants are very sensitive to initial configuration and usually get stuck at the local extrema.

Without a good organization of the distribution of mixture components, the *local extrema* will arise when there are too many mixture components in one sample area and too few in another. To escape the *local extrema*, SMEM and competitive EM (CEM) algorithms have been proposed by performing automatic merging and splitting of mixture components [34-35]. Without changing the total number of mixture components (model structure), the SMEM algorithm cannot provide an effective solution to re-organize the distribution of mixture components because moving the mixture components from one sample area to another is very difficult. With a simple annihilation mechanism, the CEM algorithm cannot obtain the optimal model structure accurately and it may result in low prediction accuracy. In addition, negative samples are not used for classifier training, thus the learned independent classifiers are unsuitable for multi-class image classification.

To enable more effective model selection and parameter estimation, we have proposed an ***adaptive EM algorithm*** by taking the following steps:

Step 1: To avoid the initialization problem, our adaptive EM algorithm starts from a reasonably large value of κ_j to capture the essential structure of the data. To escape the local extrema, our adaptive EM algorithm re-organizes the distribution of the mixture components and modifies the optimal number of mixture components by performing automatic ***merging***, ***splitting*** and ***elimination***.

Our adaptive EM algorithm uses symmetric *Jensen-Shannon* (JS) *divergence* $JS(P(X_{c_h}|C_h, \theta_l), P(X_{c_h}|C_h, \theta_k))$ to measure the divergence between two mixture components $P(X_{c_h}|C_h, \theta_l)$ and $P(X_{c_h}|C_h, \theta_k)$ for the same concept model C_h [30]:

$$\begin{aligned}&JS(P(X_{c_h}|C_h,\theta_l), P(X_{c_h}|C_h,\theta_k)) = \\ &\pi_1 KL(P(X_{c_h}|C_h,\theta_l), \pi_1 P(X_{c_h}|C_h,\theta_l) + \pi_2 P(X_{c_h}|C_h,\theta_k)) + \\ &\pi_2 KL(P(X_{c_h}|C_h,\theta_k), \pi_1 P(X_{c_h}|C_h,\theta_l) + \pi_2 P(X_{c_h}|C_h,\theta_k))\end{aligned} \qquad (4)$$

where $KL(P(X_{c_h}|C_h, \theta_l), \pi_1 P(X_{c_h}|C_h, \theta_l) + \pi_2 P(X_{c_h}|C_h, \theta_k))$ and $KL(P(X_{c_h}|C_h, \theta_k), \pi_1 P(X_{c_h}|C_h, \theta_l) + \pi_2 P(X_{c_h}|C_h, \theta_k))$ are the Kullback-Leibler divergences between $P(X_{c_h}|C_h, \theta_l)$ and $P(X_{c_h}|C_h, \theta_k)$, $\pi_1 + \pi_2 = 1$, $\pi_1, \pi_2 \geq 0$.

If $JS(P(X, C_j|S_l, \mu_{s_l}, \sigma_{s_l}), P(X, C_j|S_k, \mu_{s_k}, \sigma_{s_k}))$ is small, these two mixture components provide strongly overlapped densities and overpopulate the relevant sample regions, thus they can be merged as one single mixture component. In addition, the *local JS divergence* $JS(P(X, C_j|S_{lk}, \mu_{s_{lk}}, \sigma_{s_{lk}}), P(X, C_j|\Theta))$ is used to measure the divergence between the merged mixture component $P(X, C_j|S_{lk}, \mu_{s_{lk}}, \sigma_{s_{lk}})$ and the local sample density $P(X, C_j|\Theta)$. To detect the best candidates for ***merging***, our adaptive EM algorithm calculates the local JS divergences for $\frac{\kappa_j(\kappa_j-1)}{2}$ pairs of the mixture components that could be merged. The pair with the minimum value of the local JS divergence is selected as the best candidate for potential merging. The local sample density $P(X|\theta_{s_{ij}})$ is modified as the empirical distribution weighted by the posterior probability and defined as:

$$P(X|\theta_{s_{ij}}) = \frac{\sum_{i=1}^{N} \delta(X - X_i) P(S_i|X, C_{ij}, \theta_{s_{ij}})}{P(S_i|X, C_{ij}, \theta_{s_{ij}})} \qquad (5)$$

where $P(S_i|X, C_{ij}, \theta_{s_{ij}})$ is the posterior probability.

At the same time, our adaptive EM algorithm also calculates the *local JS divergence* $JS(P(X, C_j|S_l, \mu_{s_l}, \sigma_{s_l}), P(X, C_j|\Theta))$ to measure the divergence between the lth mixture component $P(X, C_j|S_l, \mu_{s_l}, \sigma_{s_l})$ and the local sample density $P(X, C_j|\Theta)$. If the local JS divergence for a certain mixture component $P(X, C_j|S_l, \mu_{s_l}, \sigma_{s_l})$ is big, the relevant sample region is underpopulated and the elongated mixture component $P(X, C_j|S_l, \mu_{s_l}, \sigma_{s_l})$ is selected as the best candidate to be ***split*** into two more representative mixture components for expressive concept interpretation.

In order to achieve discriminative classifier training, the classifiers for multiple semantic concepts are trained simultaneously, where the positive samples for one specific semantic concept can be the negative samples for other semantic concepts. To control the overlapping of the class distributions for different semantic concepts, our adaptive EM algorithm calculates the JS divergence $JS(P(X, C_j|S_l, \mu_{s_l}, \sigma_{s_l}), P(X, C_i|S_m, \mu_{s_m}, \sigma_{s_m}))$ between two mixture components from the class distributions of two different semantic concepts C_j and C_i. If the JS divergence between these two mixture components $P(X, C_j|S_l, \mu_{s_l}, \sigma_{s_l})$ and $P(X, C_i|S_m, \mu_{s_m}, \sigma_{s_m})$ is small, these two mixture components are overlapped in feature space and thus they are selected as the best candidate to be removed from the concept models such that discriminative classifier training can be achieved. By removing the overlapping mixture components (i.e., ***elimination***), our classifier training technique is able to maximize the margin between multiple classifiers for different semantic concepts and thus it will result in higher prediction power.

Step 2: To optimize the operations of merging, splitting and elimination, their probabilities are defined as

$$J_{merge}(l, k, \Theta) = \frac{JS(P(X, C_j|S_{lk}, \mu_{s_{lk}}, \sigma_{s_{lk}}), P(X, C_j|\Theta))}{\Phi(\Theta)} \tag{6}$$

$$J_{death}(l, m, \Theta) = \frac{JS(P(X, C_j|S_l, \mu_{s_l}, \sigma_{s_l}), P(X, C_i|S_m, \mu_{s_m}, \sigma_{s_m}))}{\Phi(\Theta)} \tag{7}$$

$$J_{split}(l, \Theta) = \frac{\Phi(\Theta)}{JS(P(X, C_j|S_l, \mu_{s_l}, \sigma_{s_l}), P(X, C_j|\Theta))} \tag{8}$$

where $\Phi(\Theta)$ is a normalized factor to make:

$$\sum_{l=1}^{\kappa_j} J_{split}(l, \Theta) + \sum_{l=1}^{\kappa_j} \sum_{k=l+1}^{\kappa_j} J_{merge}(l, k, \Theta)$$

$$+ \sum_{l=1}^{\kappa_j} \sum_{m=l+1}^{\kappa_j} J_{death}(l, m, \Theta) = 1 \tag{9}$$

The acceptance probability to perform merging, splitting or elimination operation is defined by:

$$P_{accept} = min\left(exp\left[\frac{L(X, \Theta_1) - L(X, \Theta_2)}{\tau}\right], 1\right) \tag{9}$$

where $L(X, \Theta_1)$ and $L(X, \Theta_2)$ are the objective functions for the models Θ_1 and Θ_2 (i.e., before and after merging, splitting or elimination operation), τ is a constant that is determined by experiments.

Step 3: Given the finite mixture model with a certain number of mixture components (i.e., after merging, splitting or elimination operation), the EM iteration is performed to estimate their mixture parameters such as means and covariances and weights among different mixture components.

After the EM iteration procedure converges, a weak classifier is built. The performance of this weak classifier is obtained by testing a small number of labeled samples that are not used for classifier training. If the average performance of this weak classifier is good enough, $P(C_j|X, \Theta_{c_j}) \geq \delta_1$, go to step 4. Otherwise, go back step 1. δ_1 is set to 80% in our current experiments.

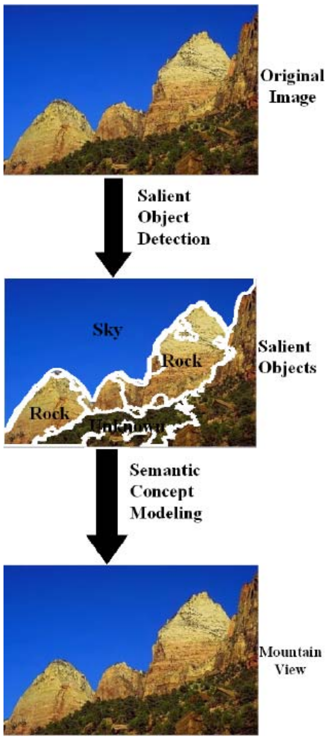

Figure 4: The flowchart for semantic image classification and annotation.

Step 4: Output the mixture model Θ_{c_j}.

By performing automatic merging, splitting and elimination of mixture components, our adaptive EM algorithm has the following advantages: (a) It does not require a careful initialization by starting with a reasonably large number of the mixture components; (b) It is able to take the advantage of the negative samples for discriminative classifier training; (c) It is able to escape the local extrema and enable a global solution by re-organizing the distribution of the mixture components in the feature space and modifying the optimal number of mixture components automatically.

Once the classifiers for the N_c pre-defined semantic concepts are in place, our system takes the following steps for semantic image classification as shown in Fig. 4: (1) Given a certain test image I_l, its salient objects are detected automatically. It is important to note that one certain test image may consist of multiple types of different salient objects in the basic vocabulary. Thus $I_l = \{S_1, \cdots, S_i, \cdots, S_n\}$. (2) The class distribution of these salient objects $I_l = \{S_1, \cdots, S_i, \cdots, S_n\}$ is then modeled as a finite mixture model $P(X, C_j|\kappa, \omega_{c_j}, \Theta_{c_j})$ [29-32]. (3) The test image I_l is finally classified into the best matching semantic image concept C_j with the maximun posterior probability.

Our current experiments focus on generating 15 basic semantic concepts, such as “beach”, “garden”, “mountain view”, “sailing” and “skiing”, which are widely distributed in natural scenes. It is important to note that once an unlabeled test image is classified into a certain semantic concept, the text keywords that are used for interpreting the relevant semantic concept and the relevant salient objects become the text keywords for annotating the multi-level semantics of

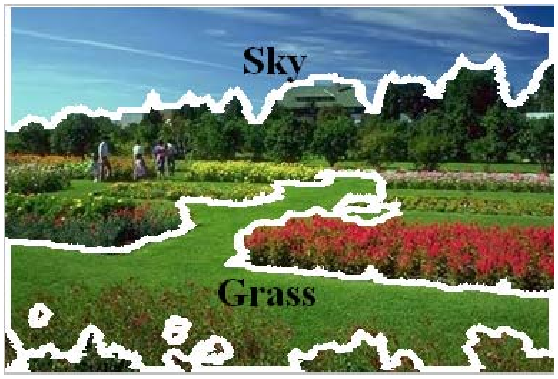

Figure 5: The result for our multi-level image annotation system, where the image annotation includes the keywords for the salient objects "sky", "grass", "forest", "flowers" and the semantic concept "garden".

the corresponding image. The text keywords for interpreting the salient objects (i.e., dominant image components) provide the annotations of the images at the content level. The text keywords for interpreting the relevant semantic concepts provide the annotations of the images at the concept level. Thus our multi-level image annotation framework can support more expressive interpretations of natural scenes as shown in Fig. 5 and Fig. 6. In addition, our multi-level image annotation technique will be very attractive to enable semantic image retrieval such that naive users will have more flexibility to specify their query concepts via various keywords at different semantic levels.

4. PERFORMANCE EVALUATION

Our experiments are conducted on two image databases: a photography database that is obtained from Google search engine and a Corel image database. The photography database consists of 35000 digital pictures. The Corel image database includes more than 125,000 pictures with different image concepts. These images (total 160,000) are classified into 15 pre-defined classes of semantic concepts and one additional category for **outliers**. Our training sets for 15 semantic concepts consist of 1,800 labeled samples, where each semantic concept has 120 positive labeled samples.

Our algorithm and system evaluation works focus on: (1) By using the same classifier, evaluating the performance differences of two image content representation frameworks: *salient objects* versus *image blobs*. (2) Under the same image content representation framework (i.e., using salient objects), comparing the performance differences between our proposed classifiers and the well-known SVM classifiers.

The *benchmark metric* for classifier evaluation includes *classification precision* α and *classification recall* β. They are defined as:

$$\alpha = \frac{\pi}{\pi + \tau}, \qquad \beta = \frac{\pi}{\pi + \mu} \tag{10}$$

where π is the set of true positive samples that are related to the corresponding semantic concept and are classified correctly, τ is the set of true negative samples that are irrelevant to the corresponding semantic concept and are classified incorrectly, μ is the set of false positive samples that are related to the corresponding semantic concept but are mis-classified.

Table 2: The classification performance (i.e., average precision versus average recall) comparison for our classifiers.

	concept	mountain view	beach	garden
salient	$\overline{\rho}$	81.7%	80.5%	80.6%
objects	$\overline{\varrho}$	84.3%%	84.7%	90.6%
image	$\overline{\rho}$	78.5%	74.6%	73.3%
blobs	$\overline{\varrho}$	75.5%	75.9%	78.2%
	concept	sailing	skiing	desert
salient	$\overline{\rho}$	87.6%	85.4%	89.6%
objects	$\overline{\varrho}$	85.5%	83.7%	82.8%
image	$\overline{\rho}$	79.5%	79.3%	76.6%
blobs	$\overline{\varrho}$	77.3%	78.2%	78.5%

As mentioned above, two key issues may affect the performance of the classifiers: (a) the performance of our detection functions of salient objects; (b) the performance of the underlying classifier training techniques. Thus the real impact for semantic image classification comes from these two key issues, the *average precision* $\overline{\rho}$ and *average recall* $\overline{\varrho}$ are then defined as:

$$\overline{\rho} = \rho \times \alpha, \qquad \overline{\varrho} = \varrho \times \beta \tag{11}$$

where ρ and ϱ are the precision and recall for our detection functions of the relevant salient objects, α and β are the classification precision and recall for the classifiers.

To obtain the real impact of using the salient objects for semantic image classification, we compared the performance differences of the same semantic image classifier by using image blobs and salient objects. The average performance differences are given in Table 2 and Table 3 for some semantic concepts. For the SVM approach, the search scheme as introduced in Section 2 is used to obtain the optimal model parameters. The average performances are obtained by averaging *precision* and *recall* over 125,000 Corel images and 35,000 photographs. One can find that using the salient objects for image content characterization has improved the accuracy of the semantic image classifiers significantly (i.e., both the finite mixture models and the SVM classifiers). It is worth noting that the average performance results $\overline{\rho}$ and $\overline{\varrho}$ shown in Table 2 and Table 3 have already included the potential detection errors that are induced by our detection functions of the relevant salient objects. In addition, the problem of over-detection of semantic concepts is also avoided as shown in Fig. 5 and Fig. 6.

By using the salient objects for image content representation and feature extraction, the performance comparison between our classifiers and the SVM classifiers is given in Fig. 7, where each point indicates the values of precision or recall for both SVM and our classifier. The experimental results are obtained for 15 semantic concepts from the same test dataset. By determining the optimal model structure and re-organizing the distributions of mixture components, our proposed classifiers are very competitive with the SVM classifiers. In addition, another advantage of our classifiers is that the models for semantic concept modeling are interpretable.

We have also tested the convergence of our adaptive EM algorithm experimentally. As shown in Fig. 8, one can find that the classifier's performance increases incrementally before our adaptive EM algorithm converges to the underly optimal model. After our adaptive EM algorithm converges to

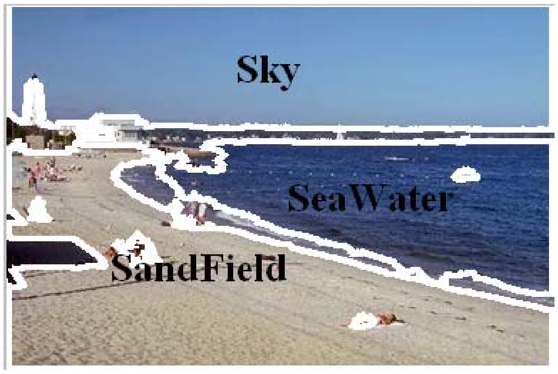

Figure 6: The result for our multi-level image annotation system, where the image annotation includes the keywords for the salient objects "sky", "sea water", "sand field" and the semantic concept "beach".

Table 3: The classification performance (i.e., average precision versus average recall) comparison for the SVM classifiers.

	concept	mountain view	beach	garden
salient	$\overline{\rho}$	81.2%	81.1%	79.3%
objects	$\overline{\varrho}$	80.5%	82.3%	84.2%
image	$\overline{\rho}$	80.1%	75.4%	74.7%
blobs	$\overline{\varrho}$	76.6%	76.3%	79.4%
	concept	sailing	skiing	desert
salient	$\overline{\rho}$	85.5%	84.6%	85.8%
objects	$\overline{\varrho}$	86.3%	87.3%	88.3%
image	$\overline{\rho}$	81.2%	78.9%	80.2%
blobs	$\overline{\varrho}$	75.6%	79.4%	81.7%

the underlying optimal model, merging more mixture components in the finite mixture models decreases the classifier's performance. The EM algorithm is guaranteed to converge only to the local extrema and does not guarantee the global solution. On the other hand, our adaptive EM algorithm is able to avoid the local extrema by involving an automatic merging, splitting and elimination procedure. Thus it can support the optimal global solution as shown in Fig. 8.

Some results of our multi-level image annotation system are given in Fig. 9 and Fig. 10, where the keywords for automatic image annotation include the multi-level keywords for interpreting both the visually distinguishable salient objects and the relevant semantic concepts.

5. CONCLUSION AND FUTURE WORKS

This paper has proposed a novel framework to enable more effective semantic image classification and automatic annotation. Based on a novel semantic-sensitive image content representation and semantic image classification framework, our multi-level image annotation system has achieved very good performance.

It is worth noting that the proposed automatic salient object detection and semantic image classification techniques can also be used for other image domains when the labeled training samples are available.

Through a density-based approach, finite mixture model has also provided a natural way to deal with the problem of semi-supervised classifier training with labeled and unlabeled samples [31]. Thus our adaptive EM algorithm is also very attractive for integrating large-scale unlabeled training

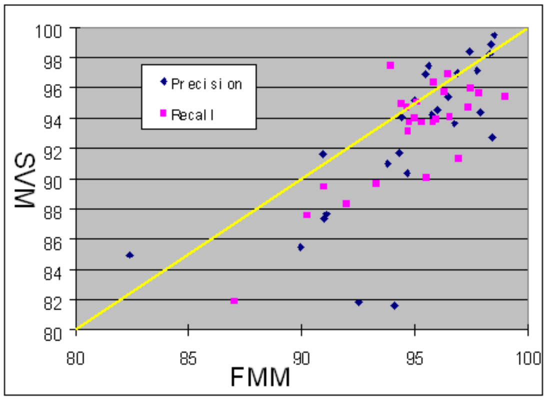

Figure 7: The performance comparison (i.e., classification precision versus classification recall) between finite mixture model and SVM approach.

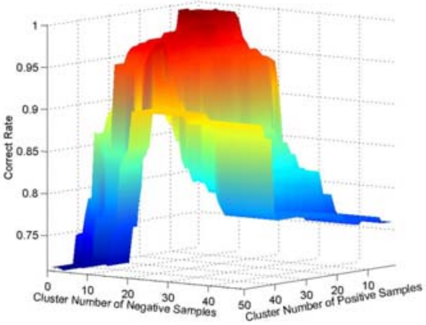

Figure 8: The classifier performance versus different number of mixture components, where the optimal number of mixture components for the semantic concept "garden" is $\kappa = 36$.

samples with a limited number of labeled training samples to enable more effective semi-supervised classifier training.

6. REFERENCES

[1] J.R. Smith and S.F. Chang, "Visually searching the web for content", *IEEE Multimedia*, 1997.

[2] E. Chang, "Statistical learning for effective visual information retrieval", Proc. ICIP, 2003.

[3] X. He, W.-Y. Ma, O. King, M. Li and H.J. Zhang, "Learning and inferring a semantic space from user's relevance feedback", ACM MM, 2002.

[4] J.R. Smith and C.S. Li, "Image classification and querying using composite region templates", *Computer Vision and Image Understanding*, vol.75, 1999. images with words", MISRM, 1999.

[5] P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary", ECCV, 2002.

[6] K. Branard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, M.I. Jordan, "Matching words and pictures", *Journal of Machine Learning Research*, vol.3, pp.1107-1135, 2003. models", ACM SIGIR, pp.119-126, 2003.

[7] M. Szummer and R.W. Picard, "Indoor-outdoor image classification", Proc. ICAIVL, 1998.

[8] R. Schettini, A. Valsasna, C. Brambilla, M. De Ponti, "A indoor/outdoor/close-up photo classifier", Proc.

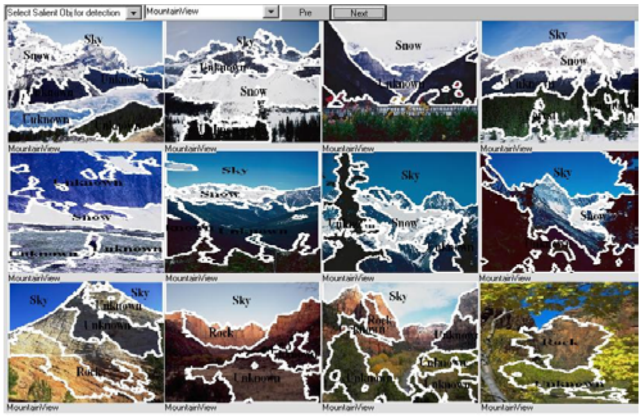

Figure 9: The semantic image classification and annotation results of the natural scenes that consist of the semantic concept "mountain view" and the most relevant salient objects.

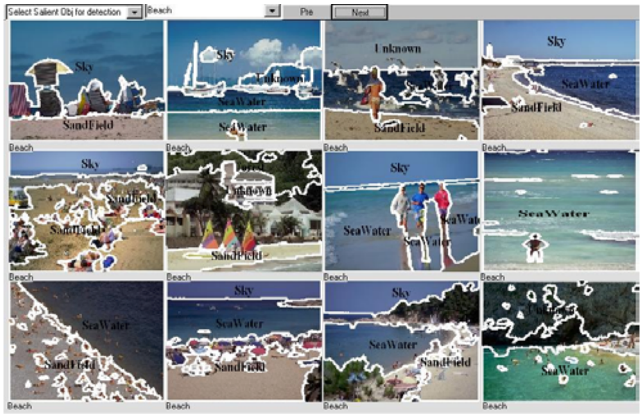

Figure 10: The semantic image classification and annotation results of the natural scenes that consist of the semantic concept "beach" and the most relevant salient objects.

Color Imaging, 2001.

[9] C. Carson, S. Belongie, H. Greenspan, J. Malik, "Region-based image querying", ICAIVL, 1997.

[10] J. Huang, S.R. Kumar and R. Zabih, "An automatic hierarchical image classification scheme", ACM MM, 1998.

[11] N. Campbell, B. Thomas, T. Troscianko, "Automatic segmentation and classification of outdoor images using neural networks", Intl. Journal of Neural Systems, vol.8, pp.137-144, 1997.

[12] J. Li, J.Z. Wang, and G. Wiederhold, "SIMPLIcity: Semantic-sensitive integrated matching for picture libraries", VISUAL, Lyon, France, 2000.

[13] A. Vailaya, M. Figueiredo, A.K. Jain, H.J. Zhang, "Image classification for content-based indexing", *IEEE Trans. on Image Processing*, vol.10, 2001.

[14] A. Hartmann, R. Lienhart, "Automatic classification of images on the web", Proc. SPIE, vol.4676, 2002.

[15] E. Chang, K. Goh, G. Sychay, G. Wu, "CBSA: Content-based annotation for multimodal image retrieval using Bayes point machines", *IEEE Trans. CSVT*, 2002.

[16] B. Li, K. Goh, E. Chang, "Confidence-based dynamic ensamble for image annotation and semantic discovery", ACM MM, 2003.

[17] A. Mojsilovic, J. Gomes, B. Rogowitz, "ISee: Perceptual features for image library navigation", Proc. SPIE, 2001.

[18] A.B. Torralba and A. Oliva, "Semantic organization of scenes using discriminant structural templates", Proc. of IEEE ICCV, 1999.

[19] J.R. Smith and S.-F. Chang, "Multi-stage classification of images from features and related text", Proc. DELOS, 1997.

[20] F. Money, D. Gatica-Perez, "On image auto-annotation with latent space model", ACM MM, 2003.

[21] J. Luo and S. Etz, "A physical model-based approach to detecting sky in photographic images", *IEEE Trans. on Image Processing*, vol.11, 2002.

[22] S.F. Chang, W. Chen, H. Sundaram, "Semantic visual template: Linking visual features to semantics", Proc. ICIP, 1998.

[23] S. Tong and E. Chang, "Support vector machine active learning for image retrieval", ACM MM, 2001.

[24] C. Zhang, T. Chen, "Indexing and retrieval of 3D models aided by active learning", ACM MM, 2001.

[25] D. Comanicu, P. Meer, "Mean shift: A robust approach toward feature space analysis", IEEE Trans. PAMI, vol.24, pp.603-619, 2002.

[26] Y. Wu, Q. Tian, T.S. Huang, "Discriminant-EM algorithm with application to image retrieval", Proc. CVPR, pp.222-227, 2000.

[27] J. Lin, "Divergence measures based on the Shannon entropy", *IEEE Trans. on IT*, vol.37, no.1, 1991.

[28] A.B. Benitez, J.R. Smith and S.-F. Chang, "MediaNet: A multimedia information network for knowledge representation", Proc. SPIE, vol.4210, 2000.

[29] H. Greenspan, J. Goldberger, A. Mayer, "Probabilistic space-time video modeling via piecewise GMM", *IEEE Trans. PAMI*, vol.26, no.3, 2004.

[30] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures", Proc. ICCV, pp.408-415, 2001.

[31] M.R. Naphade, X. Zhou, and T.S. Huang, "Image classification using a set of labeled and unlabeled images", Proc. SPIE, 2000.

[32] M.R. Naphade and T.S. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrival", *IEEE Trans. on Multimedia*, vol.3, pp.141-151, 2001.

[33] R. Oami, A. Benitez, S.-F. Chang, N. Dimitrova, "Understanding and modeling user interests in consumer videos", ICME, 2004.

[34] N. Ueda and R. Nakano, Z. Ghahramani, G. E. Hinton, "SMEM algorithm for mixture models", NIPS, 1998.

[35] B. Zhang, C. Zhang, X. Yi, "Competitive EM algorithm for finite mixture models", *Pattern Recognition*, vol.37, pp.131-144, 2004.

[36] M.A.T. Figueiredo and A.K. Jain, "Unsupervised learning of finite mixture models", *IEEE Trans. on PAMI*, vol.24, no.3, pp.318-396, 2002.