# Hierarchical Classification for Automatic Image Annotation

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

In this paper, a hierarchical classification framework has been proposed for bridging the semantic gap effectively and achieving multi-level image annotation automatically. First, the semantic gap between the low-level computable visual features and the users' real information needs is partitioned into four smaller gaps, and multiple approaches are proposed to bridge these smaller gaps more effectively. To learn more reliable contextual relationships between the atomic image concepts and the co-appearances of salient objects, a ***multi-modal boosting*** algorithm is proposed. To enable hierarchical image classification and avoid inter-level error transmission, a ***hierarchical boosting*** algorithm is proposed by incorporating *concept ontology* and *multi-task learning* to achieve hierarchical image classifier training with automatic error recovery. To bridge the gap between the computable image concepts and the users' real information needs, a novel ***hyperbolic visualization*** framework is seamlessly incorporated to enable intuitive query specification and evaluation by acquainting the users with a good global view of large-scale image collections. Our experiments on large-scale image databases have also obtained very positive results.

**Categories and Subject Descriptors**
I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis-*object recognition.*

**General Terms**
Algorithms, Measurement, Experimentation

**Keywords**
Hierarchical Image Classification, Automatic Image Annotation, Concept Ontology, Multi-Modal Boosting, Hierarchical Boosting, Hyperbolic Visualization.

## 1. INTRODUCTION

As high-resolution digital cameras become more affordable and widespread, personal collections of digital images are growing exponentially. Thus, image classification becomes increasely important and necessary to support automatic image annotation and semantic image retrieval via keywords [1-2]. However, the performance of such keyword-based image retrieval approach largely depends on three inter-related issues: (1) representative visual features for characterizing diverse visual properties of images [3]; (2) effective algorithms for classifier training and automatic image annotation [4-5, 20-21]; (3) suitable system interfaces for intuitive query specification and evaluation [1-2].

To address the first issue, the underlying visual patterns for feature extraction should be able to characterize the image semantics at the object level effectively and efficiently [3-5]. To address the second issue, robust techniques for image classifier training are needed to bridge the semantic gap successfully [1-2, 20-21]. Because one single image may contain different meanings at multiple semantic levels, hierarchical image classification is strongly expected for achieving multi-level image annotations [20-21]. Hierarchical image classification can provide at least two advantages: (1) The classifiers for the high-level image concepts can effectively be learned by combining the classifiers for the relevant image concepts at the lower levels of the concept ontology (i.e., low-level image concepts with smaller within-concept variations of visual principles); (2) The computational complexity for training the classifiers for large amounts of image concepts can significantly be reduced through exploiting the strong correlations between the image concepts. The major problem with such hierarchical approach is that the classification errors may be transmitted among different concept levels (i.e., ***inter-level error transmission***) [20]. To address the third issue, there is an urgent need to develop new visualization framework, so that users can visually be acquainted with what keywords are used to annotate and index the images and can be used for query specification.



## 2. CONCEPT ONTOLOGY CONSTRUCTION

As mentioned above, classifying images into the most relevant image concepts at different semantic levels is one promising solution to enable automatic multi-level image annotation. Motivated by this observation, we have proposed a novel scheme by incorporating the *concept ontology* for image concept organization and hierarchical image classifier training and visualizing large-scale image collections.

Following the idea of *WordNet* [11], a hierarchical network is used as the representation of the concept ontology. In this network, each node represents either one image concept at one certain semantic level or one specific salient object class. We define the former nodes as the *concept nodes* because they represent the semantics of the whole *image*, and the

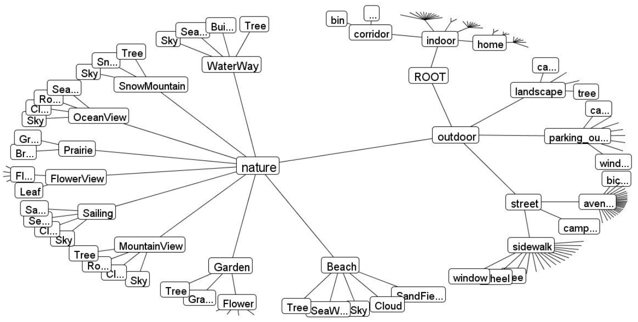


**Figure 1: Concept ontology for hierarchical organization of large amounts of image concepts.**

latter ones are defined as *content nodes* because they represent the semantics of salient objects which are the significant compounds of an image. The concept nodes at the first level of the concept ontology are defined as ***atomic image concept nodes***, which are used to represent the image concepts with the most specific subjects. They can further be assigned to the most relevant image concepts at the higher level of the concept ontology, which are used to interpret more general subjects of image contents with larger within-concept variations of visual properties.

In this paper, we have developed a semi-automatic scheme for concept ontology construction. We use our work on constructing the concept ontology for ***LabelMe***[1]as an example to depict our algorithm: (1) Labels in *LabelMe* contain text information of dominant salient objects as well as their contours and locations, but there are no explicit labels at the image concept levels [8]. Each image is stored in a folder with its name strongly indicating the concepts of the containing images. First, we remove all the most common stop words, date and other uninformative words automatically from the folders' names. Then the remaining meaningful words are separated automatically by using standard text analysis techniques to determine the basic vocabulary of image concepts (i.e., text terms for interpreting the relevant image concepts). (2) Latent Semantic Analysis (LSA) is used to group the similar text terms and identify the most significant image concepts [12]. The results of LSA are fuzzy clusters of text terms with same sense, where each cluster describes one significant image concept. (3) The contextual and logical relationships among these significant image concepts are obtained automatically, where both their joint probability and their contextual dependency are integrated to formulate a new measurement for determining the concept associations effectively.

The joint probability $\rho(C_i, C_j)$, between the text terms for interpreting the corresponding image concept $C_i$ and $C_j$, is directly obtained from the relevant image annotations:

$$\rho(C_i, C_j) = \log \frac{P(C_i, C_j)}{P(C_i)P(C_j)} \quad (1)$$

where $P(C_i, C_j)$ is the frequency for the co-occurrence of the relevant text terms $C_i$ and $C_j$, $P(C_i)$ and $P(C_j)$ are the frequencies for the individual occurrences of the text terms $C_i$ and $C_j$.

WordNet is used as the priority set to accurately quantify the contextual dependency $\pi(C_i, C_j)$ [11]:

$$\pi(C_i, C_j) = -\log \frac{length(C_i, C_j)}{2D} \quad (2)$$

where $length(C_i, C_j)$ is the length of the shortest path between two text terms $C_i$ and $C_j$ on the WordNet, and $D$ is the maximum depth of the WordNet.

The association between the given image concepts $C_i$ and $C_j$ is then determined by:

$$\phi(C_i, C_j) = \rho(C_i, C_j)\pi(C_i, C_j) \quad (3)$$

The value of $\phi(C_i, C_j)$ increases with the strength of semantic relationship between $C_i$ and $C_j$. Thus each image concept is automatically linked with the most relevant image concepts with the highest value of the association $\phi(\cdot, \cdot)$. (4) The concept ontology that is learned automatically is further evaluated and modified to express the real contextual relationships among the image concepts more precisely. One concept ontology for our test image sets is given in Fig. 1.

After the concept ontology is constructed, it is further incorporated to enable more effective image labeling and reduce the hand-labeling cost for image classifier training. Thus only the training images for the atomic image concepts at the first level of the concept ontology are labeled manually. Because the contextual and logical relationships between the atomic image concepts and the image concepts at the higher levels of the concept ontology are accurately characterized by the underlying concept ontology, the keywords for interpreting the relevant image concepts at the higher semantic levels can be propagated automatically and be added as the labels for the training images. Incorporating the concept ontology for automatic label propagation can reduce the hand-labeling cost significantly.

# 3. BRIDGING SEMANTIC GAP

The CBIR community has long struggled to *bridge the semantic gap* from successful low-level feature extraction to high-level human interpretation of image semantics, thus bridging the semantic gap is of crucial importance for achieving more effective image retrieval [1-2]. Our essential goal for image analysis is to provide more precise image content representation that allows more effective solutions for image classification, indexing and retrieval by bridging the semantic gap. In this paper, we have developed a number of comprehensive techniques to bridge the semantic gap by: (a) using salient objects to achieve more precise image content representation, and the salient objects are defined as the salient image components that are roughly related to the real world physical objects in an image [5]; (b) developing new machine learning tools to incorporate concept ontology and multi-task learning for exploiting the strong correlations between the image concepts to boost hierarchical image classifier training; (c) incorporating hyperbolic visualization to bridge the gap between the computable image concepts and the users' real needs by visually acquainting the users with a good global view of large-scale image collections.

To enable computational interpretation of image semantics, a hierarchical scheme is proposed to bridge the semantic gap (between the users' real needs and the low-level computable visual features) in four steps as shown in Fig. 2: (1) The gap between the salient image components (i.e., real world physical objects in an image) and the low-level computable visual features (i.e., **Gap 1**) is bridged by

[1]The database can be downloaded from the following URL: *http://people.csail.mit.edu/brussell/research/LabelMe*

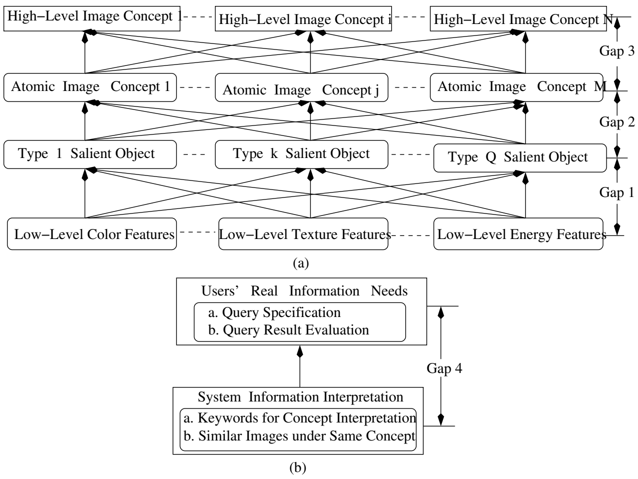


**Figure 2: The flowchart for bridging the semantic gap hierarchically.**

using salient objects [5] for image content representation. The salient objects are defined as the salient image components that capture the most significant perceptual properties linked to the semantic meaning of the corresponding physical objects in an image. Using salient objects for image content representation can provide at least four significant benefits: (a) The salient objects can effectively characterize the most significant perceptual properties of the relevant real world physical objects in an image [5], thus they can be used as ***building blocks*** to increase the expressiveness of intermediate image semantics. (b) The salient objects are not necessarily the accurate segmentation of the real world physical objects in an image [5], thus both the computational cost and the detection error rate are reduced significantly. (c) It is able to achieve a good balance between the computational complexity, the detection accuracy, and the effectiveness for interpreting the intermediate image semantics at the object level. (d) Similar images are not necessarily similar in all their salient image components, thus partitioning the images into a set of salient objects can also support partial image matching and achieve more effective image classification and retrieval. (2) The gap between the atomic image concepts and the salient objects (i.e., **Gap 2**) is bridged by using *multi-modal boosting* to exploit the strong correlations (i.e., contextual relationships) between the atomic image concepts and the co-appearances of the relevant salient objects. For example, the appearance of the atomic image concept "beach" is strongly related to the co-appearances of the salient objects, such as "sand field," "water," "tree," and "sky." (3) The gap between the high-level image concepts and the atomic image concepts (i.e., **Gap 3**) is bridged by incorporating concept ontology and multi-task learning to exploit their strong inter-concept correlations to boost hierarchical image classifier training. (4) The gap between the computable image concepts for image semantics interpretation and the users' real needs (i.e., **Gap 4**) is bridged by using hyperbolic visualization to visually acquaint the users with what keywords are used to annotate, index and access the images.

After the salient objects are extracted, the original images are decomposed into a set of salient objects. It is well-known that the diverse visual similarity between the images can be characterized more effectively and efficiently by using different types of visual features, and thus 83-dimensional visual features are extracted to characterize the diverse visual properties of images. These 83-dimensional visual features are automatically partitioned into ***9*** homogeneous feature subsets as shown in Fig. 3: 3-dimensional R,G,B average color; 4-dimensional R,G,B color variance; 3-dimensional L,U,V average color; 4-dimensional L,U,V color variance; 2-dimensional average & standard deviation of Gabor filter bank channel energy; 30-dimensional Gabor average channel energy; 30-dimensional Gabor channel energy deviation; 2-dimensional Tamura texture features (coarse & contrast), and 5-dimensional angel histogram derived from Tamura texture. Because each feature subset is used to characterize certain visual property of images, more suitable kernel function can be selected.

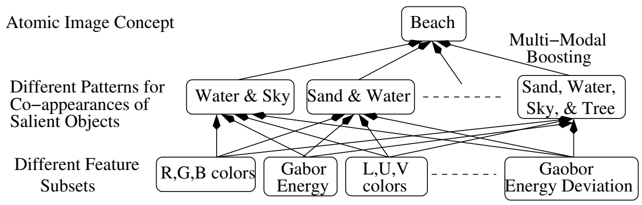


**Figure 3: The contextual relationship between the atomic image concept and the co-appearances of salient objects and their low-level visual features.**

# 4. IMAGE CLASSIFIER TRAINING

We have proposed a novel scheme by incorporating concept ontology for hierarchical image classifier training.

## 4.1 Multi-Modal Boosting

Because of the diversity of the visual similarity between the semantically-similar images, we have proposed a new framework to interpret the contextual relationships between the atomic image concepts and the co-appearances of the relevant salient objects. As shown in Fig. 3, the co-appearances of multiple salient objects are used to interpret the appearances of the relevant atomic image concepts, and such contextual relationships are well defined by our concept ontology. Due to the diversity and richness of the patterns of the co-appearances of salient objects (i.e., different images may consist of different numbers and types of salient objects), it is very attractive to learn the classifiers for all these potential co-appearance patterns and integrate them for achieving more reliable image classification.

If one certain atomic image concept $C_j$ is relevant to $n$ classes of salient objects, the total number $M$ of such co-appearance patterns of all these $n$ salient object classes can be determined by:

$$M = \sum_{i=2}^{n} C_n^i = 2^n - n - 1 \tag{4}$$

For each of these $M$ co-appearance patterns, one SVM image classifier can be trained to interpret the contextual relationship between the given atomic image concept $C_j$ and the relevant co-appearance pattern. In addition, the visual property for each of these $M$ co-appearance patterns is further characterized by **9** homogeneous feature subsets.

For one certain co-appearance pattern, the weak classifiers for all these 9 homogeneous feature subsets are integrated to boost the corresponding ensemble classifier [6]:

$$f_{C_j}^i(X) = sign\left\{\sum_{t=1}^{T}\sum_{j=1}^{9}\alpha_t^j f_t^j(X)\right\}, \quad \sum_{t=1}^{T}\sum_{j=1}^{9}\alpha_t^j = 1 \quad (5)$$

where $f_t^j(X)$ is the weak classifier for the $j$th homogeneous feature subset at the $t$th boosting iteration, and $T = 50$ is the total number of boosting iterations.

For the given atomic image concept $C_j$, its ensemble classifier is determined by a *multi-modal boosting* of the classifiers for all these $M$ potential co-appearance patterns:

$$f_{C_j}(X) = \sum_{i=1}^{M} p_i(X) f_{C_j}^i(X) \quad (6)$$

where $p_i(X)$ is the posterior distribution of the $i$th classifier $f_{C_j}^i(X)$ to be combined, and $p_i(X)$ is determined by [18]:

$$p_i(X) = \frac{exp(f_{C_j}^i(X))}{\sum_{i=1}^{M} exp(f_{C_j}^i(X))} \quad (7)$$

By incorporating high-dimensional multi-modal visual features and various co-appearance patterns for image semantics characterization, our multi-modal boosting technique is able to handle the huge diversity of within-concept visual properties, and it can further provide a natural way to select the most representative feature subsets and the most suitable kernel functions for each aotmic image concept.

## 4.2 Hierarchical Boosting

Because of the inherent complexity of the task, automatic detection of the high-level image concepts with larger within-concept variations of visual properties is still beyond the ability of the state-of-the-art techniques [1-2]. The image concepts are dependent and such dependencies can be characterized effectively by the concept ontology. Unfortunately, most existing techniques for hierarchical image classifier training suffer from the problem of inter-level error transmission. Thus there is an urgent need to develop new scheme for hierarchical image classifier training with automatic error recovery. In this paper, we have proposed a novel algorithm by incorporating the concept ontology for hierarchical image classifier training. First, the concept ontology is used to identify the related tasks, e.g., training the classifiers for the sibling image concepts under the same parent node. Second, such task relatedness is used to determine the transferable knowledge and common features among the classifiers for the sibling image concepts to generalize their classifiers significantly from fewer training images. Because the classifiers for the sibling image concepts under the same parent node are used to characterize both their individual visual properties and the common visual properties for their parent node, their outputs are strongly correlated according to the new task (i.e., learning a biased classifier for their parent node).

For a given second-level image concept $C_k$, its child image concepts (i.e., the sibling atomic image concepts under $C_k$) are strongly correlated and share some common visual properties for their parent node $C_k$, thus *multi-task learning* can be used to train their classifiers simultaneously [9-10]. Because the related tasks are characterized effectively by the concept ontology, our hierarchical classifier training algorithm can provide a good environment to enable more effective multi-task learning. To integrate multi-task learning for SVM image classifier training, a *common regularization term* $W_0$ of the SVM image classifier is used to represent and quantify the transferable knowledge and common features among the SVM image classifiers for the sibling image concepts under the same parent node. The weak classifier for the atomic image concept $C_j$ can be defined as [10]:

$$f_{C_j}(X) = W_j^T X + b \quad (8)$$

where $W_j = W_0 + V_j$, $W_0$ is the common regularization term shared between the classifiers for the sibling atomic image concepts under the same parent node, and $V_j$ is the specific regularization term for the atomic image concept $C_j$.

Given the labeled training samples for $L$ sibling atomic image concepts under $C_k$: $\Omega = \{X_{ij},\ Y_{ij} | i = 1, \cdots, N;\ j = 1, \cdots, L\}$, training multiple classifiers for the sibling atomic image concepts under the same parent node $C_k$ is then transformed into a joint optimization problem:

$$min\left\{C\sum_{j=1}^{L}\sum_{i=1}^{N}\xi_{ij} + \beta_1\sum_{j=1}^{L}\|V_j\|^2 + \beta_2\|W_0\|^2\right\} \quad (9)$$

**subject to**:

$$\forall_{i=1}^{N}\forall_{j=1}^{L} : Y_{ij}(W_0 + V_j)\cdot X_{ij} + b \geq 1 - \xi_{ij}, \quad \xi_{ij} \geq 0$$

where $\xi_{ij} \geq 0$ represents the training error rate, $L > 0$ is the total number of atomic image concepts under the same parent node, $\beta_1$ and $\beta_2$ are positive regularization parameters, $C$ is the penalty term. The dual optimization problem for Eq. (9) is to determine the optimal $\alpha_{ij}^*$ by:

$$max\left\{\sum_{j=1}^{L}\sum_{i=1}^{N}\alpha_{ij} - \frac{1}{2}\sum_{j=1}^{L}\sum_{i=1}^{N}\sum_{h=1}^{L}\sum_{l=1}^{N}\alpha_{ih}Y_{ih}\alpha_{jl}Y_{jl}K_{jh}(X_{ih}, X_{jl})\right\} \quad (10)$$

**subject to:**

$$\forall_{i=1}^{N}\forall_{j=1}^{L} : 0 \leq \alpha_{ij} \leq C, \quad \sum_{j=1}^{L}\sum_{i=1}^{N}\alpha_{ij}Y_{ij} = 0$$

where $K_{jh}(\cdot,\cdot)$ is the underlying kernel function. By exploiting the transferable knowledge and common features for image classifier training, our multi-task learning algorithm is able to handle the inter-concept visual similarity effectively.

The common regularization term $W_0$ for the sibling atomic image concepts is further treated as a prior regularization term to bias the SVM classifier for their parent node. Setting such prior regularization term is able to exploit the inter-concept correlations between the SVM classifiers according to the new task and can reduce the training cost significantly. Based on such prior regularization term, a *biased classifier* for their parent node is trained effectively by using few new training images. Thus the biased classifier for their parent node $C_k$ is determined by:

$$min\left\{\frac{1}{2}\|W - W_0\|^2 + \alpha\sum_{l=1}^{m}[1 - Y_l(W^T\cdot X_l + b)]\right\} \quad (11)$$

where $W_0$ is the common regularization term for the sibling atomic image concepts under $C_k$, $(X_l, Y_l)$, $l = 1, \cdots, m$ are the new training samples for learning the biased classifier for $C_k$. The dual problem for Eq. (11) is solved by:

$$min\left\{\frac{1}{2}\sum_{l=1}^{m}\sum_{h=1}^{m}\alpha_l\alpha_h Y_l Y_h X_l^T X_h - \sum_{l=1}^{m}\alpha_l(1 - Y_l W_0^T X_l)\right\} \quad (12)$$

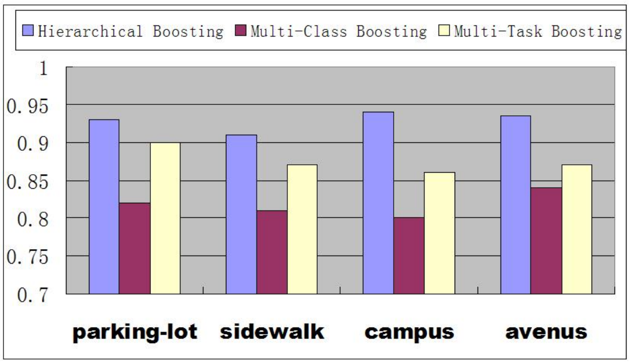


**Figure 4: The comparison results between our hierarchical boosting algorithm, multi-class boosting and multi-task boosting for four sibling image concepts.**

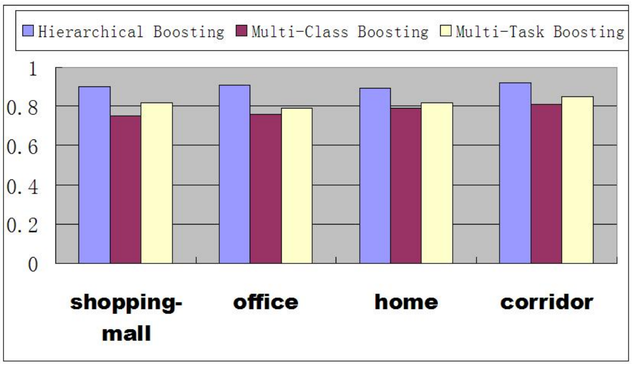


**Figure 5: The comparison results between our hierarchical boosting algorithm, multi-class boosting and multi-task boosting for four sibling image concepts.**

**subject to:**

$$\forall_{l=1}^{m} : 0 \leq \alpha_l \leq C, \qquad \sum_{l=1}^{m} \alpha_l Y_l = 0$$

The optimal solution of Eq. (12) satisfies:

$$W = W_0 + \sum_{l=1}^{m} \alpha_l Y_l X_l \tag{13}$$

Thus the bias classifier for the given second-level image concept $C_k$ is obtained as:

$$f_{C_k}(X) = W^T X + b \tag{14}$$

To learn the ensemble classifier for the given second-level image concept $C_k$, a novel ***hierarchical boosting*** scheme has been developed by combining its biased classifier with the classifiers for its child image concepts. Unfortunately, all the existing boosting techniques can only combine the weak classifiers that are learned in different ways (i.e., different input spaces) but for the same task [6], and they did not include the regularization between different tasks which is very essential for hierarchical image classifier training. We have developed a new scheme for multi-task classifier combination (called *hierarchical boosting*) that is able to integrate the classifiers trained for multiple tasks and leverage their distinct strengths and exploit the strong correlations of their outputs according to the new task. Our hierarchical boosting scheme can search an optimal combination of these multi-task classifiers by sharing their transferable knowledge and common features according to the new task (i.e., learning the ensemble classifier for their parent node $C_k$), and thus it is able to generalize the ensemble classifier significantly while reducing the computational complexity dramatically. For the given second-level image concept $C_k$, the final prediction of its ensemble classifier can be obtained by a *logistic boosting* of the predictions of its biased classifier and the classifiers for its child image concepts [18]:

$$H_{C_k}(X) = \sum_{h=1}^{L+1} p_h(C_h) f_{C_h}(X) \tag{15}$$

By exploiting the strong inter-concept correlations for hierarchical image classifier training, our hierarchical boosting algorithm is able to learn the classifiers for large amounts of image concepts simultaneously.

Our hierarchical boosting algorithm can significantly outperform the traditional techniques such as multi-class boosting and multi-task boosting [7-8]. The multi-class boosting techniques do not explicitly exploit the transferable knowledge and common features among the classifiers to enhance their classification performance [7]. The multi-task boosting algorithms have recently been proposed to enable multi-class object detection by sharing the common features among the classifiers [8]. Rather than incorporating the transferable knowledge and common features to learn a biased classifier, the ensemble classifier for each object class is simply composed by the classifiers that are trained for all the pair-wise object class combinations [8], thus the strong inter-concept correlations between the classifiers cannot be exploited effectively. On the other hand, our hierarchical boosting algorithm can integrate the transferable knowledge and common features to enhance all these single-task classifiers at the same semantic level simultaneously, exploit their strong inter-concept correlations to learn a biased classifier, and boost an ensemble classifier for their parent node with higher discrimination power. As shown in Fig. 4 and Fig. 5, one can find that our hierarchical boosting algorithm can significantly outperform the traditional techniques such as multi-class boosting and multi-task boosting.

# 5. MULTI-LEVEL IMAGE ANNOTATION

After our hierarchical image classifiers are available, they are used to classify the images into the most relevant image concepts at different semantic levels. We carry out our experimental studies of our proposed algorithms by using three public image databases: *LabelMe*, *Corel Images*, *Google Images*. For LabelMe image database, it contains more than 25,000 images and our experiments are done on a snapshot of this database downloaded at April 2006. For Corel image database, we have also included 2800 images for natural scenes. For Google Images, we have included 9000 natural images. Table 1 gives the average accuracy of our hierarchical image classifiers for some image concepts.

In our hierarchical image classification scheme, the initial classification of a test image is critical because the classifiers at the subsequent levels cannot recover from the misclassification of the test image that may occur at a higher concept level, and this misclassification can be propagated to the terminal concept node (i.e., inter-level error transmission) [20]. We have integrated two innovative solutions seamlessly to address such inter-level error transmission problem: (1) en-

**Table 1: The classification accuracy (i.e., precision/recall) for some image concepts.**

| concepts | mountain view | beach | garden |
|---|---|---|---|
| | 80.6% /85.6% | 90.4% /90.6% | 89.5% /88.3% |
| concepts | sailing | skiing | desert |
| | 85.8% /84.6% | 83.6% /84.2% | 79.5% /74.7% |
| concepts | ocean view | waterway | prairie |
| | 82.3% /81.5% | 85.4% /85.7% | 80.5% /82.4% |
| concepts | shopping | office | bathroom |
| | 83.6% /84.2% | 89.8% /88.7% | 86.5% /87.3% |
| concepts | sidewalk | corridor | kitchen |
| | 85.4% /84.9% | 86.4% /85.8% | 89.6% /90.2% |
| concepts | home | avenus | campus |
| | 90.5% /91.2% | 92.3% /95.2% | 90.2% /91.5% |

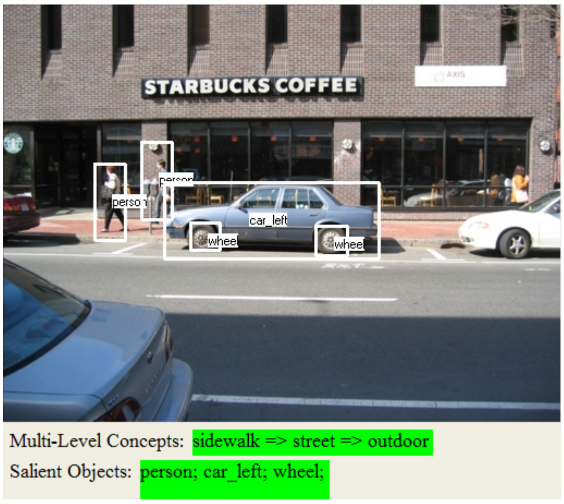


**Figure 6: Multi-level image annotation results both at the object level and at multiple concept levels.**

hancing the classifiers for the image concepts at the higher levels of the concept ontology, so that they can have higher discrimination power (i.e., classifying the images more accurately); (2) integrating new protocols that are able to detect such misclassification path early and take appropriate actions for automatic error recovery.

Three significant respects of our hierarchical image classification scheme are able to address the inter-level error transmission problem effectively: (a) The transferable knowledge and common features can be shared among the classifiers for the sibling image concepts at the same semantic level of the concept ontology to maximize their margins and enhance their discrimination power significantly, so that the test images can be classified more accurately at the beginning, i.e., image concepts at the higher levels of the concept ontology. By exploiting the strong correlations between the classifiers for their child image concepts, our hierarchical boosting scheme is able to learn more reliable ensemble classifiers for the high-level image concepts. (b) The classification decision for each test image is determined by a voting from multiple multi-task classifiers for the sibling image concepts at the same semantic level to make their errors to be transparent. (c) An *overall probability* is calculated to determine the best path for hierarchical image classification. For a given test image, an optimal classification path should provide maximum value of the overall probability among all the possible classification paths. The overall probability $h(C_k)$ for the optimal classification path (from one certain higher level image concept $C_k$ to the most relevant lower level image concept $C_j$) is defined as [17]:

$$h(C_k) = p(C_k) + g(C_j), \quad g(C_j) = max\{p(C_i)|i = 1, \cdots, L\} \tag{16}$$

where $p(C_k)$ is the posterior probability for the given test image to be classified into the current image concept $C_k$ at the higher level of the concept ontology, $p(C_i)$ is the posterior probability for the given test image to be classified into the child image concept $C_i$ of $C_k$, $g(C_j)$ is the maximum posterior probability for the given test image to be classified into the most relevant child concept node $C_j$. Thus a good path should achieve higher classification accuracy for both the high-level concept node and the most relevant child concept node. By using the overall probability, it is able for us to detect the incorrect classification path early and take appropriate actions for automatic error recovery.

It is important to note that once a test image is classified, the keywords for interpreting the salient object classes (what are visible in the images) and the relevant image concepts (what the images are about and what can be evoked by the visible salient objects) at different levels of the concept ontology become the keywords for interpreting its semantics more sufficiently. Our **multi-level image annotation** scheme is very attractive to support semantic image retrieval via keywords such that the naive users can have more flexibility to specify their query concepts via various keywords at different semantic levels. Our experimental results on hierarchical image classification and multi-level annotation are given in Fig. 6. From our experimental results, one can find that our proposed hierarchical image classification scheme is able to achieve more sufficient image annotations.

## 6. HYPERBOLIC IMAGE VISUALIZATION

For naive users to harvest the research achievements of CBIR community, it is very important to develop more comprehensive framework for intuitive query specification and evaluation, but it is also a problem without a good solution so far. The problem, in essence, is also about how to present a good global view of large-scale image collections to users [13-16], so that users can easily specify their queries. Therefore, there is a great need to generate the overall information of large-scale image collections conceptually and incorporate the concept ontology to organize and visualize such concept-oriented overall information more effectively and intuitively.

To achieve multi-modal representation of concept ontology, each concept node on the concept ontology is jointly characterized by: *keyword* to interpret its semantics, *most representative images* to display its concept-oriented summary, *decision principles* (i.e., support vectors and importance factors for classifier combination) to characterize its feature-based principal properties, and contextual relationships between the relevant image concepts.

We have developed a novel scheme to generate the concept-oriented summarization of large-scale image collections. For one given image concept on the concept ontology, three types of images are automatically selected to generate its concept-oriented summary: (a) Images which locate on the decision boundaries of the SVM image classifier; (b) Images which have higher confidence scores in the classification procedure; (c) Images which locate at the centers of dense areas and can be used to represent large amounts of semantically similar images for the given image concept. To obtain such representative images, kernel PCA is used to cluster the semantically similar images for the given image concept into multiple significant groups [19].

Our approach for concept ontology visualization exploits hyperbolic geometry [16]. The hyperbolic geometry is particularly well suited to graph-based layout of large-scale concept ontology, and it has "more space" than Euclidean geometry. The essence of our approach is to project the concept ontology onto a hyperbolic plane according to the contextual relationships between the image concepts, and layout the concept ontology by mapping the relevant concept nodes onto a circular display region. Thus our concept ontology visualization framework takes the following steps: (a) The image concept nodes on the concept ontology are projected to a hyperbolic plane according to their contextual relationships, and such projection can accurately preserve the original contextual relationships between the image concept nodes. (b) Poincaré disk model [16] is used to map the concept nodes on the hyperbolic plane to a 2D display coordinate.

Each concept node on the graph is assigned a location $z = (x, y)$ within the unit disk, which represents the Poincaré coordinates of the corresponding image concept node. By treating the location of the image concept node as a complex number, we can define such a mapping as the linear fractional transformation [16]:

$$z_t = \frac{\theta z + P}{1 + \bar{P}\theta z} \tag{17}$$

where $P$ and $\theta$ are the complex numbers, $|P| < 1$ and $|\theta| = 1$, and $\bar{P}$ is the complex conjugate of $P$. This transformation indicates a rotation by $\theta$ around the origin following by moving the origin to $P$ (and $-P$ to the origin).

After the hyperbolic visualization of the concept ontology is available, it can be used to enable interactive exploration and navigation of large-scale image collections at the concept level via *change of focus*. The *change of focus* is implemented by changing the mapping of the image concept nodes from the hyperbolic plane to the unit disk for display, and the positions of the image concept nodes in the hyerbolic plane need not to be altered during focus manipulation. Users can change their focus of image concepts by clicking on any visible image concept node to bring it into focus at the center, or by dragging any visible image concept node interactively to any other location without losing the contextual relationships between the image concept nodes, where the rest of the layout of the concept ontology transforms appropriately. Thus our hyperbolic framework for concept ontology visualization has demonstrated the remarkable capabilities for interactively exploring large-scale image collections at the concept level. By supporting change of focus, our hyperbolic visualization framework can theoretically display unlimited number of image concepts in a 2D unit disk.

Moving the focus point over the display disk unit is equivalent to translating the concept ontology on the hyperbolic plane, such change of focus can provide a mechanism for controlling which portion of the concept ontology receives the most space and changing the relative amount of the image concept nodes for current focus. Through such change of focus on the display disk unit for concept ontology visualization and manipulation, it is able for users to interactively explore and navigate large-scale image archives at the concept level. Therefore, users can always see the details of the regions of interest by changing the focus. Different views of the layout results of our concept ontology visualization are given in Fig. 7. By changing the focus points, our hyperbolic framework for concept ontology visualization can provide an effective solution for interactive exploration of large-scale image collections at the concept level. In addition, users can visually be acquainted with what keywords are used to annotate and index the images, and thus they can easily and intuitively specify their queries with better knowledge of large-scale image collections.

Because the concept ontology is used to represent the abstract information of large-scale image collections, the amounts of the returned images for such keyword-based queries may be very large and it is too expensive for users to look for some particular images effectively. Our solution for this problem is to project the returned images onto a 2D visualization space according to their kernel-based visual similarity distances. In addition, the visual similarity relationships between the returned images are explained more intuitively by using hyperbolic visualization, thus users can easily judge the relevance between the returned images and interactively browse large amounts of returned images according to their visual similarity. Such ***interactive exploration*** can also allow users to zoom into regions of interest, obtain some additional images of interest that may not be found by using 1D or 2D top list (some interesting images which are loosely relevant to users' queries), and enable fortunate discoveries of unexpected images by accident.

For a given returned image $I$, we set $\rho$ to be the hyperbolic distance between $I$ and the center of the hyperbolic plane for image projection, and $r$ be the distance between $I$ and the center of the display unit circle. The relationship between their derivatives is described by:

$$d\rho = \frac{2}{1 - r^2} \cdot dr \tag{18}$$

An example of this hyperbolic visualization is shown in Fig. 8, where the returned images for the query "nature scene" are layouted according to the global color histogram by using Kernel PCA projection [19]. One can observe that such 2D hyperbolic visualization of the returned images can provide more intuitive interpretation of their visual similarity, where the similar images are closer according to their kernel-based visual similarity. Therefore, users are allowed to manipulate not only the images, but also their visual similarity relationships. By allowing users to zoom into regions of interest via changing the focus, our algorithm can easily support visualization of large amounts of returned images.

## 7. CONCLUSIONS

In this paper, we have proposed a novel algorithm for automatic multi-level image annotation via hierarchical classification. A novel multi-modal boosting algorithm is proposed to achieve more reliable interpretation of the contextual relationships between the atomic image concepts and the co-appearances of salient objects. To avoid the inter-level error transmission problem, a novel hierarchical boosting algorithm is proposed by incorporating concept ontology and multi-task learning to boost hierarchical image classifier training. Our hierarchical image classifier training algorithm is able to simultaneously learn the classifiers for large

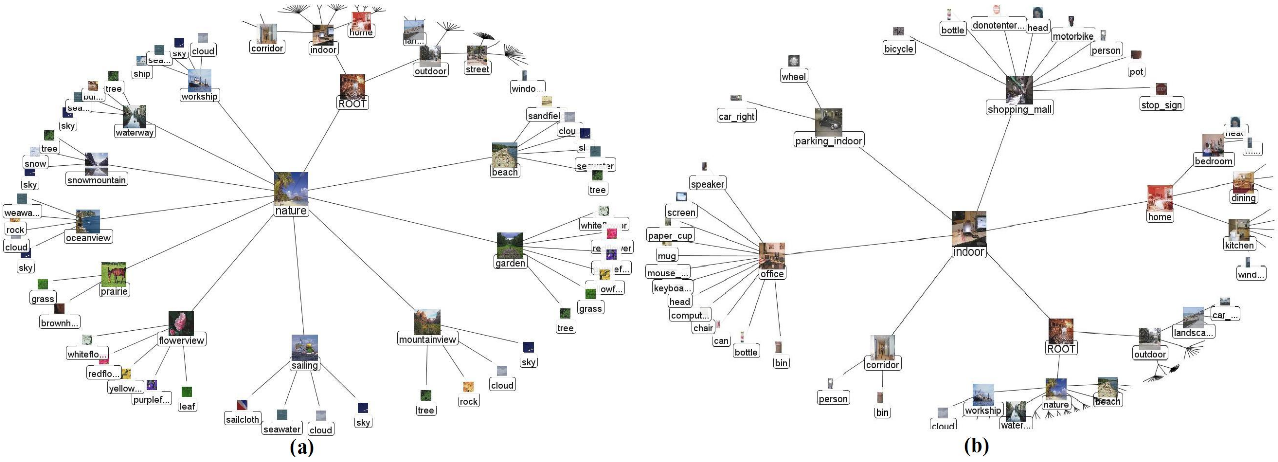


**Figure 7: Two different views of our hyperbolic visualization of large-scale concept ontology.**

amounts of image concepts with huge within-concept visual diversities and inter-concept visual similarities. A novel hyperbolic visualization framework is seamlessly incorporated to enable intuitive query specification and similarity-based evaluation of large amounts of returned images. Our experiments on large-scale image database have also obtained very positive results.

## 8. REFERENCES

[1] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, "Content-based image retrieval at the end of the early years", *IEEE Trans. on PAMI*, 2000.

[2] X-J. Wang, L. Zhang, F. Jing, W-Y. Ma, "AnnoSearch: Image auto-annotation by search", IEEE CVPR, 2006.

[3] W-Y. Ma, B. S. Manjunath, "Texture features and learning similarity", IEEE CVPR, pp.425-430, 1996.

[4] J. Jeon, R. Manmatha, "Using maximum entropy for automatic image annotation", CIVR, pp.24-32, 2004.

[5] **J. Fan**, Y. Gao, H. Luo, G. Xu, "Automatic image annotation by using concept-sensitive salient objects for image content representation", ACM SIGIR, 2004.

[6] Y. Freund, R.E. Schapire, "Experiments with a new boosting algorithm", Proc. ICML, pp.148-156, 1996.

[7] R.E. Schapire, Y. Singer, "BoosTexter: A boosting-based system for text categorization", *Machine Learning*, vol.39, pp.135-168, 2000.

[8] A. Torralba, K. P. Murphy, W. T. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection", IEEE CVPR, 2004.

[9] K. Yu, A. Schwaighofor, V. Tresp, W.-Y. Ma, H.J. Zhang, "Collaborative ensemble learning: Combining content-based information filtering via hierarchical Bayes", Proc. of Intl. Conf. on Uncertainty in Artificial Intelligence (UAI), 2003.

[10] T. Evgeniou, C.A. Micchelli, M. Pontil, "Learning multiple tasks with kernel methods", *Journal of Machine Learning Research*, vol.6, pp.615-637, 2005.

[11] G.A. Miller, "WordNet: A lexical database for English", *Comm. of ACM*, vol.38, no.11, 1999.

[12] S. Deerwester, S.T. Dumais and R. Harshman, "Indexing by latent semantic analysis", *Journal of the American Society of Information Science*, 1990.

**Figure 8: Hyperbolic visualization of the visual similarity of the returned images for query "nature scene".**

[13] Y. Rubner, C. Tomasi, L. Guibas, "A metric for distributions with applications to image databases", IEEE ICCV, pp.59-66, 1998.

[14] J.A. Walter, D. Webling, K. Essig, H. Ritter, "Interactive hyperbolic image browsing-towards an integrated multimedia navigator", KDD, 2006.

[15] G.P. Nguyen, M. Worring, "Similarity based visualization of image collections", AVIVDiLib, 2005.

[16] J. Lamping, R. Rao, "The hyperbolic browser: A focus+content technique for visualizing large hierarchies", *Journal of Visual Languages and Computing*, 1996.

[17] D.E. Knuth, *The Art of Computer Programming*, vol. 3: *Sorting and Searching*, Addison-Wesley, 1978.

[18] J. Friedman, T. Hastie, R. Tibshirani, "Additive logistic regression: a statistical view of boosting", *Annals of Statistics*, vol.28, no.2, pp.337-374, 2000.

[19] K. Grauman, T. Darrell, "The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features", MIT-CSAIL-TR-2006-20, MIT 2006.

[20] J. Fan, H. Luo, Y. Gao, M.-S. Hacid, "Mining image databases on semantics via statistical learning", KDD, 2005.

[21] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures", IEEE ICCV, 2001.