DOI: 10.2298/CSIS100423035J

# Multi-Scale Image Semantic Recognition with Hierarchical Visual Vocabulary[1]

Xinghao Jiang[1,2], Tanfeng Sun[1,2], and GuangLei Fu[1]

[1] School of Information Security Engineering
Shanghai Jiao Tong University, Shanghai, 200240, China
{xhjiang, tfsun}@sjtu.edu.cn
[2] Key Lab. of Shanghai Information Security Management and Technology Research,
Shanghai, 200240, China

**Abstract.** Local features have been proved to be effective in image/video semantic analysis. The BOVW (bag of visual words) scheme can cluster local features to form the visual vocabulary which includes an amount of words, where each word is the center of one clustering feature. The vocabulary is used to recognize the image semantic. In this paper, a new scheme to construct semantic-binding hierarchical visual vocabulary is proposed. Some attributes and relationship of the semantic nodes in the model are discussed. The hierarchical semantic model is used to organize the multi-scale semantic into a level-by-level structure. Experiments are performed based on the LabelMe dataset, the performance of our scheme is evaluated and compared with the traditional BOVW scheme, experimental results demonstrate the efficiency and flexibility of our scheme.

**Keywords:** local feature, bag of visual words, image semantic analysis, visual vocabulary.

## 1. Introduction

With the rapid development of Internet and multimedia technology, explosively growing amount of images and videos can be acquired from the web or relevant database. The content-based image/video classification will play more and more important role in the field of images/videos processing. Human can easily figure out different genres of images/videos just by watching them. However, for the computer, it is a quite complicated work to automatically recognize the semantic of a image/video. How to use computer to analyze image semantic has been discussed and researched as a hot topic in this field. The research on image/video semantic analysis is closely

[1] Corresponding author: Dr. Tanfeng Sun.

connected with many applications, such as: content-based image or video retrieval system, the utility in intelligent traffic and safety surveillance, and so on.

A lot of work has been concentrated on some global features extracted from images such as color and texture [1, 2, 3]. An image can be represented by a global feature vector. Then the problem of analyzing image semantic is turned into the problem of supervised classifying. Support Vector Machine (SVM) can be used to judge whether an image belong to one semantic or the other based on amount of training features. Though the use of global feature need only cheap computing cost, its effectiveness is poor and reveals unsatisfied performance.

Local feature has been studied as an improvement on global feature. DoG (Difference-of-Gaussian) [4] is used to detect interest points from image and then SIFT (Scale Invariant Feature Transform) [4] is used to extract a vector of feature from each of those points. Feature is described by the pixel values around the interest point. In this way, an image can be represented as a collection of feature vectors. An easy way to analyze whether image includes some object is to match the feature collection of object image with the feature collection of testing image [5, 6]. Some matching structure can be used in this process to decrease the cost. But it is still not very efficient in recognizing multi-object or complicated semantic.

Recently a new model called BOVW (bag of visual words) which reflected on the BOW (bag of words) model in document retrieval has been discussed widely [7, 8, 9]. BOVW also takes advantage of local feature of image. Like the way BOW works, BOVW can be used to construct a visual vocabulary of an image. The building of visual vocabulary is done by clustering all the feature vectors extracted from the training images. Clustering generated a certain number of cluster centers in feature space. In this way, each cluster center is regarded as a word in visual vocabulary. Each feature vector extracted from image can find its nearest word in vocabulary. Then an image can be represented as a word vector in which each dimension number means that whether the image contains the word. The training image is used to train the SVM for the classifying task. Some details about the BOVW model such as weighting strategy, vocabulary size has been discussed in several papers [10, 11] as well. Though BOVW model has been proved to be more effective on problem of image objects or semantic analysis, it still has at least two drawbacks: 1) the features which are used to construct visual vocabulary have no semantic connection. This leads to the loss of semantic information of visual vocabulary, which the noise feature may have negative influence on the result of image analysis; 2) there is not an efficient structure which can fit large vocabulary. Small semantic analysis may be solved smoothly by the traditional BOVW model, but when there is need for complicated semantic recognition or multi-level semantic analysis, the traditional BOVW model is not enough.

Our new work aims at the drawbacks about the traditional BOVW model mentioned above. The work in [12] proposed the way of semantic-preserving BOVW model. Several codebooks which belong to certain semantic can be constructed firstly, then image can be analyzed by judging whether any feature extracted from it belonged to any codebook. In our method, a new hierarchical semantic model is proposed, which can be applied in complicated semantic analysis. Based on the hierarchical semantic model, the semantic-binding visual vocabulary tree can be constructed. We define some attributes and relationship of the semantic nodes in the model. The hierarchical semantic model is used to organize the multi-scale semantic into a level-by-level structure. Experiments demonstrate the performance of our scheme is efficiency and flexibility.

The experiments are performed based on the LabelMe image dataset from MIT [13] which contains 11,282 objects from 495 categories. The LabelMe dataset is an online interactive image database, from which users can obtain the annotation of objects in each image. The annotation is very useful for helping us to build the semantic-binding visual vocabulary tree.

The rest of this paper is organized as follows: In Sect. 2, some related work which would be involved in our method is introduced. Sect. 3 presents the definitions of semantic attributes and semantic relationship, and the scheme of building hierarchical semantic model. Sect. 4 discusses the construction of semantic-binding visual vocabulary tree. Sect. 5 gives the method to analyze image semantic using our model and vocabulary tree. Sect. 6 shows the experimental results and analysis. In Sect. 7, the conclusion and some future direction are presented.

## 2. Related Work

### 2.1. Sparse Image Interest Point Detecting

Sparse image interest point is compared to the dense image interest point which regards each point of image as the target. Ideal sparse image interest point is scale-invariant, affine-invariant and position-invariant. There are corner-like point detector such as Harris Laplas [14] and blob-like point detector such as LoG (Laplacian of Gaussian). Our work uses the DoG [4] in which the detection process involved not only the image itself but also the neighboring images in the scale space. DoG finds the interest point by determining whether it is a local maximum compared with 26 surrounding points (9 points in pre-scale image, 9 points in post-scale image and 8points in current image) and at the same time the point should be the maximum in it scale curve.

### 2.2. Local Feature Extraction

Local feature extraction computes on the surrounding pixel values of the interest point and puts out a vector representing local feature. SIFT (scale invariant feature transform) [4] has been regarded as an excellent local feature in image analysis compared to other versions of local feature. In our work, we adopt SIFT to extract local features from images. For each interest point in image, SIFT choose 16 areas around it. The direction of every point in area is calculated out by its surrounding pixels. For all points in each area, all the directions are clustered into 8 bins. In this way, each area has eight numbers meaning the histogram of its direction. For this interest point, the 8-bin histogram of direction from 16 areas forms its final 128-dimensional local feature vector.

### 2.3. Distance metric learning with contextual constraints

Clustering features in traditional BOVW model use the Euclidean distance to calculate the distance between two feature vectors. Though computing Euclidean distance needs less cost, it lost the contextual information of feature. It is because the Euclidean distance does not take the semantic class of a feature into consideration. Distance metric learning can help to solve this problem. Distance metric learning takes advantage of the contextual constraints of feature. The so-called contextual constraint is the class information of feature [15, 16]. With the contextual constraints, a matrix A can be acquired through training. Then the distance of any two feature vectors can be calculated by the Mahalanobis distance as follows:

$$d_{x.y} = (x-y)^T \times A \times (x-y) \quad (1)$$

where $x$ and $y$ are two feature vectors, $A$ is learnt distance metric.

In our work we used the DCA (Discriminative Component Analysis) distance metric learning [17]. The basic idea of DCA is to learn an optimal data transformation that leads to the optimal distance metric by both maximizing the total variance between the discriminative data class and minimizing the total variance of data instances in the same class. DCA firstly calculates two covariance matrices $C_b$ and $C_w$ which describe the total variance between data of the discriminative class and the total variance of data among the same class respectively. The two matrices are computed as follows:

$$C_b = \frac{1}{n_b} \sum_{j=1}^{n} \sum_{i \in D} (m_j - m_i)(m_j - m_i)^T \quad (2)$$

$$C_w = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ji}-m_j)(x_{ji}-m_j)^T \qquad (3)$$

In formulas (2) and (3), $n_b = \sum_{j=1}^{n} | D_j |$ , $m_j$ is the mean vector of the *j*-th class, $x_{ji}$ is the *i*-th data instance in *j*-th class, and $D_j$ is the discriminative set in which each element is one of n class that has at least one negative constraint to the *j*-th class. DCA resolves the learning task by the optimization as follows:

$$J(A) = \arg\max_A \frac{\left|A^T C_b A\right|}{\left|A^T C_w A\right|} \qquad (4)$$

In formula (4), $A$ is the optimal transformation matrix to be learned. When $A$ is obtained, the optimal Mahalanobis matrix can be calculated by $M = AA^T$.

## 3. Hierarchical Semantic Model

A lot of work has been done on understanding image semantic by different kinds of image features, either global or local feature. Less attention is paid to the semantic itself. Our work will give out a new semantic model first which contains some attributes and the relationship between semantics.

### 3.1. Definitions of Semantic Attributes and Semantic Relationship

We have given some brief introduction about the model in our early work [18]. Hierarchical Semantic Model is used to construct all the image semantics in a semantic space. The constructing process is to place an image semantic into the semantic space and to make correct relationship with other image semantics. When the construction process is done, the semantic model is a multi-layer structure. The upper layer is for bigger image semantic and the lower layer is for smaller image semantic. The ‘big’ and ‘small’ are just comparative terms. Fox example, semantic of ‘wheel’ is a small semantic when it is compared to semantic of ‘car’. But semantic ‘car’ is not big enough if you take semantic of ‘street’ into consideration. Actually what is more important is not whether a semantic is big or small, but is the relationship between different semantics. Just like the example mentioned above, ‘car’ should have ‘wheels’, and probably there are many cars on the ‘street’.

First we give out some definition of semantic attributes here. We classify all the semantic into two classes. One class is called the 'combination semantic' and the other is called 'singleton semantic'. Some notations are be used here: 1) $\sigma$ stands for the scale or granularity of the semantic; 2) $\sum$ stands for the combination of several semantic; 3) $\bigcup$ stands for the union of several semantic or semantic set.

**Definition 1**. Singleton semantic: singleton semantic describes some simple semantic which has no necessity to be destructed again. An example of singleton semantic is semantic of 'car'. You still can destruct the semantic of 'car' into semantic of 'wheel' or semantic of 'windscreen'. But 'wheel' and 'windscreen' is too simple to form a visual vocabulary individually. This is also what 'no necessity' stands for.

**Definition 2**. Combination semantic: combination semantic describes some comparatively complicated semantic which are formed by the combination or union of several smaller semantics. An example of combination semantic is semantic of 'street'. Semantic of 'street' can be composed of the semantic of 'road' and semantic of 'car' or semantic of 'house' and so on. If $S$ stands for combination semantic, $s$ stands for the semantic which $\sigma(s) < \sigma(S)$, then $S = \bigcup_{i=1}^{n} ss_i, ss_i = \sum_{k=1}^{m} s_k$ .

Then we will discuss the definition of relationship between semantics. There are mainly two kinds of relationship between image semantics: relationship of combination and relation of belonging-to.

**Definition 3**. Relationship of combination: relationship of combination describes relationship between some smaller semantic and some bigger semantic. All smaller semantic make up the bigger semantic. An example of this kind of relationship is semantic of 'street' (bigger semantic) and semantic of 'car', semantic of 'road', semantic of 'house' (three smaller semantics). Those three smaller semantic form the bigger semantic of 'street'. If $S$ stands for up-level semantic and $s$ stands for the down-level semantic, then the relation ship of combination can be described by $S = \sum_{i=1}^{n} s_i$ .

**Definition 4**. Relationship of belonging to: relationship of belonging to also describes relationship between some smaller semantic and some bigger semantic. The difference from the relationship of combination is that bigger semantic does not need all the smaller semantic. An example is semantic of 'vehicle' (bigger semantic) and smaller semantic of 'car' and smaller semantic of 'truck'. Semantic of 'car' belongs to semantic of 'vehicle' no matter whether there exists semantic of 'truck'. If $S$ stands for up-level semantic and $s$ stands for the down-level semantic, then the relation ship of combination can be described by $S = \bigcup_{i=1}^{n} s_i$ .

The following three more definitions are for the relational attributes of the semantic.

**Definition 5**. Relationship of mutual exclusion: relationship of mutual exclusion describes the relation of two semantics which can’t be co-existed. An example is semantic of ‘street’ and semantic of ‘classroom’.

**Definition 6**. Required semantic: When several small semantics combine into a bigger semantic, some small semantic must be in this combination and such kind of semantic is called ‘required’. For example, semantic of ‘street’ can be combined by semantics of ‘road’, ’car’, ’house’, ’walking person’ and so on. ‘road’, ‘house’ should be two required semantic. Actually whether a semantic is required is closely connected to the application demand. Details will be explained latter.

**Definition 7**. Optional semantic: this is compared to the definition of required semantic. That is the semantic which may or may not exist in the combination of a bigger semantic.

Those definitions of semantic attributes and semantic relationship are used in the construction of hierarchical semantic model which would be discussed in the following part. Semantic has its own structure and order and our work does take advantage of such kind of order and structure to recognize image semantic.

### 3.2. Construction of Hierarchical Semantic Model

One different point from the traditional BOVW is that our visual vocabulary tree is bound to certain semantics. In other words, the vocabulary tree must be constructed to match with a semantic model. This also means that before construction of a useful semantic-binding visual vocabulary tree, a hierarchical semantic model should be constructed first. The model we discussed above is an abstract model. If we want to apply this model into practice, we should connect it with some concrete semantics.

For construction of our model, the main task is to decompose a bigger semantic into several smaller semantic iteratively until it comes to some level of singleton semantic. As discussed before, it is not necessary to decompose singleton semantic any more. When all the decomposition has been finished, we shall define the attributes of every semantic node in the hierarchical model and the relationship between any two semantic which are connected with each other. The definitions of attributes and relationship have been introduced in Sect. 3.1. When all the work has been done, a hierarchical semantic model has been successfully constructed. Fig. 1 describes this model.

Fig.1 shows an abstract model of hierarchical semantic. In Fig. 1, top level semantic is the biggest semantic in this model. And there are three combination semantics which maintain the ‘relationship of belonging to’ with its upper level semantic. For each combination semantic, there are two singleton semantics below them, and the singleton semantics combine into the upper level semantic with the relationship of ‘combination’. The number list like 1, 1 or 1,1,1 in figure just labels the position of relevant semantic node.

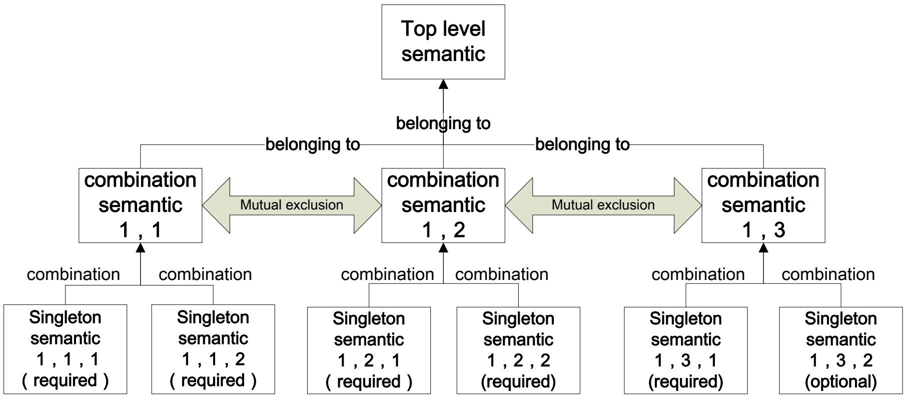
Top level semantic
belonging to
belonging to
belonging to
combination semantic 1 , 1
Mutual exclusion
combination semantic 1 , 2
Mutual exclusion
combination semantic 1 , 3
combination
combination
combination
combination
combination
combination
Singleton semantic 1 , 1 , 1 ( required )
Singleton semantic 1 , 1 , 2 ( required )
Singleton semantic 1 , 2 , 1 ( required )
Singleton semantic 1 , 2 , 2 (required)
Singleton semantic 1 , 3 , 1 (required)
Singleton semantic 1 , 3 , 2 (optional)

**Fig.1.** The hierarchical semantic model

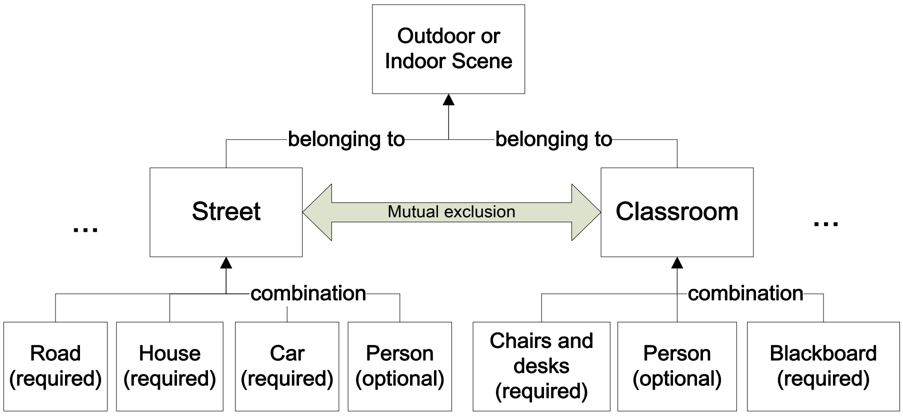

**Fig.2.** An example of hierarchical semantic model with concrete classification demand

When we put the hierarchical semantic model into practice, concrete classification demand should be taken into account. Fig. 2 shows an example of hierarchical semantic model with application demand, we should think about what semantics need to be recognized and analyzed, what singleton semantics are required and what else are optional. In Fig.2, each semantic node is connected to a concrete semantic. The top semantic is ‘outdoor’ or ‘indoor’ scene. And there are two combination semantics which are ‘street’ and ‘classroom’ belonging to this top semantic. Each of the combination semantic is made up of several singleton semantics. Semantic of ‘street’ combined by singleton semantic of ‘road’, ‘house’, ‘car’ and ‘walking person’ and semantic of ‘classroom combined by singleton semantic of ‘chairs and desks’, ’students’ and ‘blackboard’.

We can see that the singleton semantic of ‘person’ is included both in semantic of ‘street’ and of ‘classroom’ in Fig.2. This sometimes happens especially in some large semantic space situation. Singleton semantic is just like part which always used to make up the large up-level semantic. So the same singleton semantic is very likely to be used in several different large semantic.

For the situation of large semantic space, single hierarchical semantic model maybe is not enough to cover the whole semantic space. We can make extension for the model proposed above. Several models can be built with the certain semantic spaces, so a ‘forest’ can be formed. Each ‘tree’ of this ‘forest’ stands for a united sub semantic space and all ‘trees’ stand for the whole semantic space. One virtual root node can be made to take the charge of all the ‘trees’.

## 4. Construction of Semantic-binding Visual Vocabulary

The objective for building such a hierarchical semantic model is to make a template on which a visual vocabulary tree can be constructed. As we discussed in Sect. 3, the decomposition of semantic is a top-to-bottom process. On the contrary, the process of constructing a visual vocabulary tree is from bottom to top.

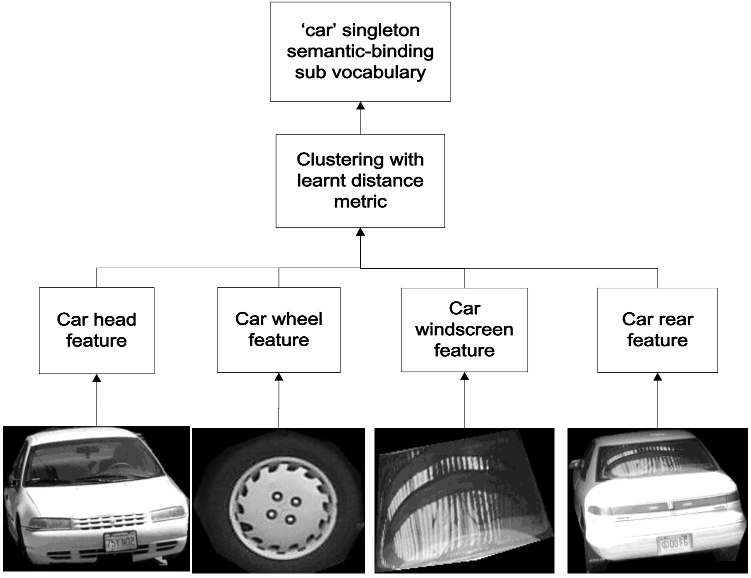
‘car’ singleton semantic-binding sub vocabulary
Clustering with learnt distance metric
Car head feature
Car wheel feature
Car windscreen feature
Car rear feature

**Fig.3.** Construction singleton semantic-binding sub vocabulary of ‘car’

Fig.1 and Fig.2 elaborate what the hierarchical semantic model looks like. Actually when we construct a vocabulary tree, each node in the tree is a sub vocabulary of which semantic is responding to that in the model. Followings are the main steps for constructing a semantic-binding visual vocabulary tree.

**Step 1**. First, we build those singleton semantic-binding sub vocabularies which are located in the bottom level in hierarchical semantic model. For each singleton semantic node, we collect the images that represent this semantic. SIFT feature is extracted from those images and distance metric is learnt by the contextual information of feature. The contextual information here means different feature comes from different class into one semantic. Take semantic of 'car' as an example. A 'car' can product features from its 'wheel' or its 'windscreen' or its 'body'. When we obtained the distance metric, we cluster features into vocabulary using the learnt metric to compute feature distance. Here k-means algorithm is adopted for clustering. Fig. 3 describes this process briefly, i.e. the features to train the vocabulary bound with semantic of 'car' are taken from different parts of the car, such as 'wheel', 'windscreen', 'car head' and 'car rear'. In this way, a singleton semantic binding vocabulary is successfully built.

Now we have a sub vocabulary of K words (if we set K as the number of clusters in clustering step), and we also should calculate out radius of each word as following formula:

$$r_i = \frac{\sum_{j=1}^{n_i} (|x_{ij} - c_i|_A)}{n_i} \tag{5}$$

In formula (5), $r_i$ is the word radius for the *i*-th word in a certain semantic binding vocabulary. $n_i$ is the number of the feature vectors belonging to this word. $c_i$ is the clustering center point vector and $x_{ij}$ is each feature vector. $A$ is the learnt distance matrix. We calculate the word radius by means of the sum of the distance between feature and center point. Since not all the features we take to train the semantic binding vocabulary is totally correct, there still may be some noise features, so averaging can decrease the negative influence from those noise features.

The radius of the whole vocabulary is calculated out as follows:

$$R = \max(|x_i - vc|_A) \tag{6}$$

In formula (6), *R* is the radius of whole vocabulary, $x_i$ is each feature in the vocabulary, $vc$ is the vocabulary center point which equals to the mean of all the word center points. And the longest distance between the features and vocabulary center points is defined as vocabulary radius.

**Step 2**. After all the singleton semantic vocabulary has been built, we can build the sub vocabulary responding to the up-level semantic in the model.

In this step, we do not need to extract new feature. The features used to build up-level semantic-binding sub vocabulary are formed by features of the semantics combined into it or the semantics belonging to it. And the contextual information here is down-level semantic which the feature comes from. With the feature and contextual information, we can acquire the learnt distance metric. And in the same way, up-level vocabulary can be obtained by clustering the features with the learnt distance metric. Fig. 4 shows us the process to construct the up-level semantic-binding vocabulary from down-level semantic binding vocabularies.

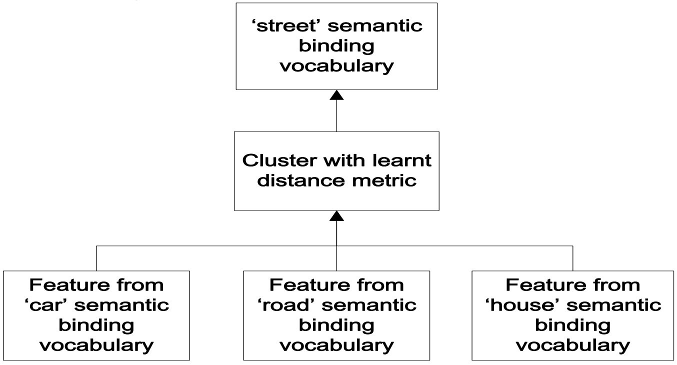
'street' semantic binding vocabulary
Cluster with learnt distance metric
Feature from 'car' semantic binding vocabulary
Feature from 'road' semantic binding vocabulary
Feature from 'house' semantic binding vocabulary

**Fig.4.** Construct the up-level vocabulary from down-level vocabularies

**Step 3**. We can build the other sub vocabulary iteratively from bottom semantic to top semantic just in the way Step 2 shows us.

The above steps show how to build a semantic binding visual vocabulary tree. The process makes sure that each sub vocabulary in the tree is bound to a certain semantic. It can be easily proved because in the hierarchical model, the up-level semantic is formed by its down-level semantics. And in the process of building the vocabulary tree, all the features in the down-level semantics run into the up-level semantic eventually. According to the extent of semantic, those features belong to up-level semantic naturally.

Just like the traditional BOVW model, the vocabulary size is an important factor which influents performance greatly. We give out a solution to decide the size of each semantic binding sub vocabulary. The size of a vocabulary is proportional to its complexity, i.e. the more complex it is, the bigger size it has. We use a randomization method to know the complexity and to decide the size.

The features of each semantic binding vocabulary are taken from some different class. As we known, the feature of singleton semantic vocabulary is taken from different parts of the singleton semantic object, and the feature of combination semantic vocabulary is taken from different down-level semantic binding vocabulary. We define $\partial$ here, and 0<$\partial$<1. For a vocabulary which we want to decide its size, we take $\partial N_i$ feature vectors from each of its down-level vocabulary randomly. Where r is the number of down-level vocabularies, $N_i$ means the number of feature vectors of the *i*-th down-level vocabulary. So we can get $N = \sum_{i=1}^{r} \partial N$ feature vectors and we compute its complexity as follows:

$$D = \frac{\sum_{i=1}^{N} \sum_{j>i}^{j \le N} \left| x_i - x_j \right|_A}{N(N-1)} \quad (7)$$

$$C = \frac{D}{2R} \quad (8)$$

The formula (8) shows that we compute the complexity of the vocabulary by those randomly picked feature. A vocabulary is more complex when its average distance between feature vectors is bigger. Where $R$ is the radius of vocabulary, $D$ is the mean distance among those picked feature and $C$ is the complexity for the vocabulary (0< $C$ <1). We can use $C$ to decide the size of vocabulary as follows:

$$size = C \times SMAX \quad (9)$$

In formula (9), $SMAX$ stands for the largest size of all the vocabularies.

## 5. Analyzing Image Semantic with Semantic Binding Visual Vocabulary

Just like the traditional BOVW model, a built vocabulary is used to analyze the image features and to perform the final words histogram. In our method, we take the similar way to analyze image semantic. The big difference in our method is that we do not use SVM to classify which semantic image should belong to. Instead we analyze the semantic of image by what sub vocabulary or its semantic the image possessed. The followings are the steps to analyze image semantic based on the semantic binding visual vocabulary:

**Step 1**. Detect the interest points in test image by DoG method, and the SIFT feature vector is extracted for each interest point.

**Step 2**. For each SIFT feature vector extracted from test image, we match it with sub vocabularies in visual vocabulary tree from top to bottom. For a certain sub vocabulary, the method to judge whether a feature belong to this vocabulary is described as follows:

$$\bigcup_{\substack{c_i \in vocabulary \\ f_j \in vocabulary}} \delta(\left| f_j - c_i \right|_A < r_i) = 1 \tag{10}$$

In formula (10), $\delta$ is the function: when its input argument is true, then its output is 1; when argument is false, then its output is 0. $f_j$ is SIFT feature extracted from image, $c_i$ is *i*-th word centers of the vocabulary, $r_i$ is the word radius for the *i*-th word, and $A$ is the learnt distance matrix for the vocabulary. If the result of formula (10) is 1, $f_j$ drops into the vocabulary.

Actually we can adopt two different strategies. Strategy 0: a feature can drop down into any number of vocabularies on each level. Under such condition, the above formula is used to judge which vocabularies the feature drops into; Strategy 1: a feature can only drop down into one vocabulary on each level. Under such condition, if the above formula shows that a feature may be in several vocabularies on each level, then the feature is discarded as an unstable feature. The comparison will be revealed in experiments (Sect. 6.3) on these two strategies.

**Step 3**. Match each SIFT feature with top semantic binding sub vocabulary in the way as Step 2. For the feature belonging to this sub vocabulary, we match it to each of the down-level sub vocabulary. This process works iteratively until the feature reaches the singleton semantic binding sub vocabulary or until the feature is discarded for it belongs to no sub vocabulary.

**Step 4**. Do step 3 on all the SIFT feature vectors extracted from test image. For each bottom-level singleton semantic binding sub vocabulary, we know if it contains any feature. If one singleton semantic contains feature extracted from test image, we say the image possesses this singleton semantic.

**Step 5**. Now we know what singleton semantic the test image possesses. Based on the hierarchical semantic model we have built, when we know what the down-level semantic the test image possesses, we can know what up-level semantic the test image possesses. We do this semantic aggregation work from bottom (singleton semantic) to top (the biggest semantic). After this process is finished, we can know what semantic the test image has in each level of the hierarchical semantic model. For the example in Fig. 2, If a test image has singleton semantic of ‘car’, ‘house’ and ‘road’, then we can say the image also has semantic of ‘street’. And further we can say the image has a scene semantic.

**Step 6**. When Step 5 is done, we can have the knowledge of the semantic of image from big to small scale. Here big scale means what combination semantic the image has, and the small scale means what the singleton semantic the image has. One point should be paid attention to here is that if a image gets the two semantics which maintain the relationship of mutual exclusion, then this image should not have either of the both semantics.

## 6. Experimental Results

We evaluate our model and algorithm from different scales of semantic on the test images. For one test image, we can also evaluate the performance of our work by the accuracy of analysis results on each level semantic. And then the performance of our scheme is evaluated and compared with the traditional BOVW scheme.

### 6.1. Dataset for Experiments

The experiments are carried out on the dataset provided by LabelMe project from MIT [13]. LabelMe is an image dataset in which each image has a responding annotation file. The annotation file annotates objects of different semantic in the image. So we can collect large number of training material of certain semantic from LabelMe dataset. LabelMe dataset also includes a wide range of image categories which totally covers 11,281 objects from 495 categories.

### 6.2. Experiment Setting

Two combination semantics are chosen for the experiments. One is outdoor semantic: the semantic of 'street'. The other is indoor semantic: the semantic of 'office'. Actually our model can be applied to any combination semantic as long as the semantic can be decomposed in the way we introduced in Sect. 3. The hierarchical semantic model for the experiments should be constructed first, as shown in Fig. 5.

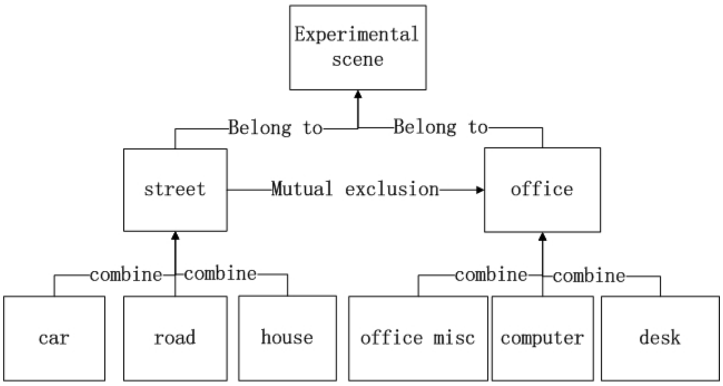

**Fig.5.** The hierarchical semantic model for our experiments

Fig. 5 shows the semantic model for our experiments. Our experiments aim at two combination semantic 'street' and 'office', both of them belong to the top-level semantic. The ‘street’ semantic and ‘office’ semantic have relationship of mutual exclusion between them. For ‘street’ semantic, it is formed by ‘car’, ‘street’ and ‘house’ semantic. All of the three singleton semantics are ‘required’ to their up-level semantic. For ‘office’ semantic, it is formed by ‘office miscellaneous’, ‘computer’ and ‘desk’ semantic. And all of those three semantics are also ‘required’ to their up-level semantic.

The semantic binding visual vocabulary tree is constructed based on the hierarchical semantic model in Fig. 5. Some training images of certain singleton semantic are collected from labelMe dataset first. Fig. 6 describes this process.

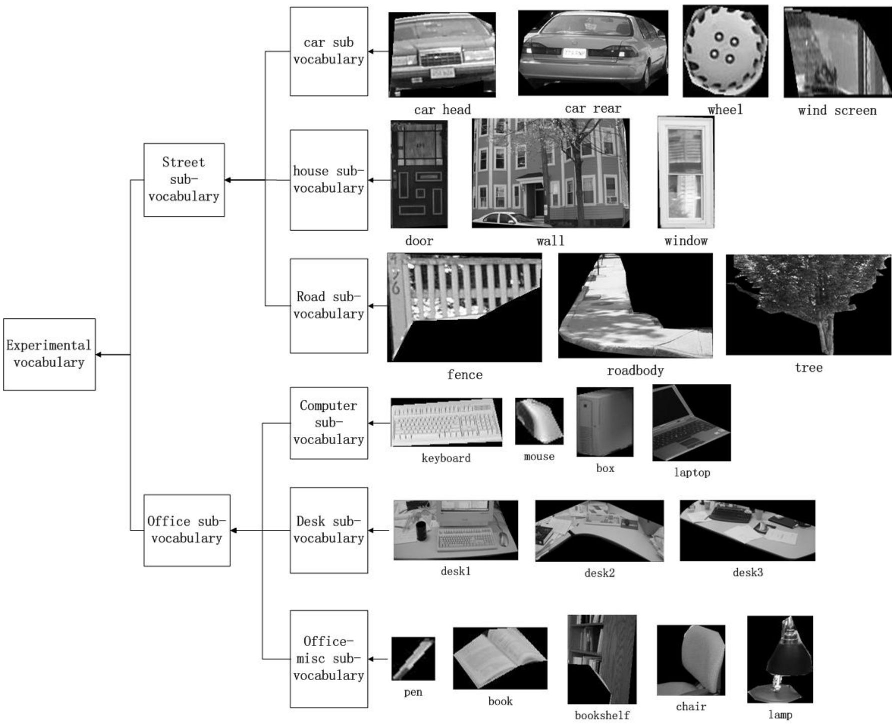

**Fig.6.** Construction for semantic binding vocabulary tree for experiments

Fig. 6 shows that each bottom-level semantic-binding sub vocabulary is trained by the image of responding semantics. Actually the image shown in Fig. 6 for each singleton semantic vocabulary is from different parts of the object. Just as discussed in Sect. 4, different parts can provide us with the contextual information which is useful in distance metric learning. In the experiment, we learn the distance matrix for each of the six singleton semantic binding vocabulary with the contextual information. Three of the singleton semantic binding vocabularies (‘car’, ‘road’, ‘house’) aggregate to their up-level vocabulary (‘street’), and the other three (‘office misc’, ‘desk’, ‘computer’) aggregate to the other up-level vocabulary (‘office’). The aggregation process has been discussed in details in Sect. 4.

In our experiment, the size of each sub vocabulary is determined in the vocabulary tree according to the complexity of the vocabulary itself. The method has been discussed in Sect. 4, and Table 1 describes the details of the size of each sub vocabulary.

**Table 1.** The size of each sub vocabulary in the vocabulary tree of our experiment

| | | | | | |
|---|---|---|---|---|---|
| Top semantic vocabulary (size): 500 | | | | | |
| ‘street’ vocabulary (size): 400 | | | ‘office’ vocabulary (size): 300 | | |
| ‘car’:<br>200 | ‘house’<br>: 300 | ‘road’<br>: 300 | ‘office misc’<br>: 300 | ‘computer’<br>: 200 | ‘desk’<br>: 200 |

For each test image, we extract the SIFT features for each interest point which is detected by DoG method from the image. For each SIFT feature, we analyze it with the vocabulary tree from top to bottom. In Sect. 5, we proposed two strategies in judging whether a feature drops in a vocabulary. One is that in each level a feature can drop in several sub vocabulary, the other is that in each level a feature can only drop in one sub vocabulary and otherwise the feature is discard. We will give a performance comparison in the experiment results.

When all the features from test image are analyzed by the whole vocabulary tree, we recognize the image semantic from the bottom singleton semantic, and gradually to up-level combination semantic. To evaluate our model and method in details, we can give the performance for each level semantic in our experimental model. For each semantic node in the hierarchical model, we evaluate the accuracy of our method as follows:

$$precision_i = \frac{N_{correct}}{N_{test}} \tag{11}$$

$$recall_i = \frac{N_{correct}}{N_{truth}} \tag{12}$$

In formula (11) and (12), $i$ means the *i*-th semantic node in the experimental model, $N_{correct}$ stands for the number of test images which are recognized correctly, $N_{test}$ is the number of total test images, $N_{truth}$ is the number of test images which really have the certain semantic. We give the evaluation for accuracy of each semantic node by two strategies in experiment results.

### 6.3. Experiment results

In the experiments, totally 1000 test images are chosen from LableMe dataset, in which 500 images of them match the semantics in the experimental model and the other 500 do not match. 'Match' here means the image can match any node semantic in the experimental model. We hope such kind of composition can make the testing more standard and convincing.

We evaluated the performance of our method on every semantic node in experiment hierarchical semantic model with two feature dropping strategies (discussed in Sect. 5). Strategy 0 means that a feature vector can drop into any sub vocabulary in one level, Strategy 1 means that a feature vector can drop into only one sub vocabulary in the level (the multi-dropping feature be discard as unstable feature). Table 2 and Table 3 show the accuracy of precision and accuracy of recall about each semantic. There are 6 singleton semantics ('car', 'house', 'road', 'office misc', 'computer', 'desk') and 2 combination semantics ('street', 'office').

**Table 2.** Accuracy of each semantic node in experiment model with strategy 0

| Feature dropping strategy 0 | | | | |
|---|---|---|---|---|
| semantic | 'car' | 'house' | 'road' | '**street**' |
| precision | 0.74 | 0.82 | 0.73 | 0.70 |
| recall | 0.79 | 0.85 | 0.75 | 0.72 |
| semantic | 'misc' | 'computer' | 'desk' | '**office**' |
| precision | 0.79 | 0.78 | 0.82 | 0.78 |
| recall | 0.82 | 0.83 | 0.73 | 0.69 |

**Table 3.** Accuracy of each semantic node in experiment model with strategy 1

| Feature dropping strategy 1 | | | | |
|---|---|---|---|---|
| semantic | 'car' | 'house' | 'road' | **'street'** |
| precision | 0.72 | 0.72 | 0.79 | 0.66 |
| recall | 0.77 | 0.76 | 0.75 | 0.67 |
| semantic | 'misc' | 'computer' | 'desk' | '**office**' |
| precision | 0.64 | 0.66 | 0.68 | 0.64 |
| recall | 0.68 | 0.71 | 0.69 | 0.61 |

From Table 2 and Table 3, we can observe that the accuracy (precision or recall) of the singleton semantic is higher than that of the combination semantic. In Table 2 , the precisions of 'car', 'house', 'road' are 0.74,0.82 and 0.73, and all of them are higher than the precision of 'street' (0.70). The precisions of 'office misc', 'computer', 'desk' are 0.79, 0.78 and 0.82, and all of them are also higher than that of 'office' (0.78). So is the recall. The reason is the confirmation of a combination semantic needs all the confirmation of its down-level required semantic. In our experiment model, if an image possesses semantic 'street', it must possess semantic 'car', 'house' and 'road'. This leads to the combination semantic ('street' or 'office') with lower accuracy than the singleton semantic. Compare the results of Table 2 with Table 3, the accuracy of semantic node in Table 3 (with Strategy 1) is lower than that of the responding semantic node in Table 2 (with Strategy 0). The precision of semantic 'car' in Table 2 is 0.74 and that in Table 3 decreased to 0.72. The precision of semantic 'office' in Table 2 is 0.78 and that in Table 3 decreased to 0.64. The reason is that Strategy 1 has more limits than Strategy 0, so the unstable feature should be discarded according to Strategy 1 (discussed in Sect. 5), thus the number of features dropping into bottom-level sub vocabulary would decrease. Fig. 7 shows the accuracy results with comparison in chart.

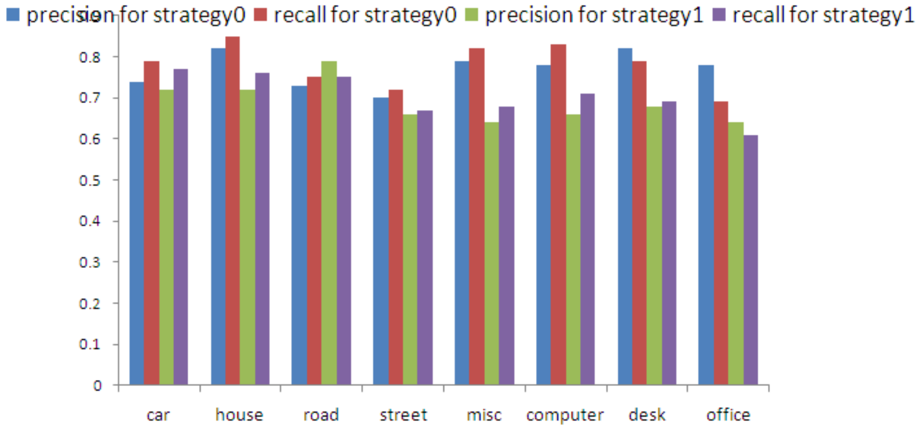

**Fig. 7.** Accuracy on each semantic node with two strategies

In Fig. 7, X-axis is semantic and Y-axis is accuracy. We can clearly see that the accuracy on singleton semantic is to some extent satisfactory. But the accuracy on combination semantic, especially with strategy 1, has still space to make improvement. Actually the dataset for training and distance metric learning used in vocabulary construction are two important factors in the running of the whole model. In our experimental results, semantic 'house' and 'desk' get higher accuracy (0.82 and 0.82 individually) since the training dataset for those two sub vocabulary has larger complexity than others. This may inspire us that the complexity and discriminative of the training data can impact the effect of the vocabulary. And the DCA (Discriminative Component Analysis) we used in our model can be replaced by a better distance metric learning algorithm which adapts to our method. This will be studied in our future work.

In order to compare the performance of our model with that of traditional BOVW model, we perform the comparison experiments with BOVW model on the same dataset as well. We constructed the codebooks of BOVW model on the training data from images of semantic 'street' or semantic 'office'.

SIFT features are extracted from all of those training images. And all the features are clustered into 500 clusters. Then a codebook of 500-word size is generated. KNN (K-Nearest Neighbor) algorithm is used to find the nearest word in the codebook. By using the trained SVM, the test images between the semantic 'street' and 'office' can be classified. Table 4 shows the performance comparison between our model and BOVW model.

**Table 4.** Comparison between our model and BOVW model

| Method | Semantic 'street' | Semantic 'office' |
|---|---|---|
| BOVW Model (precision) | 0.62 | 0.55 |
| Our method with Strategy 0 (precision) | 0.70 | 0.78 |
| Our method with Strategy 1 (precision) | 0.66 | 0.64 |

Table 4 shows that even in Strategy 1, the accuracy of our method is still higher than that of traditional BOVW model (semantic 'street' is 4 percentage points higher and semantic 'office' is 9 percentage points higher), which reveal that our model can work effectively in image semantic recognition.

Besides the higher accuracy, our scheme can understand and analyze image semantic in a much more flexible way. Our scheme can analyze images and get the semantic recognition on each semantic node. But for BOVW it is necessary to do the classification for each semantic.

## 7. Conclusion

In this paper, a hierarchical semantic model is proposed. The hierarchical semantic model is used to organize the multi-scale semantic into a level-by-level structure. The attributes and relationship of the semantic node in the model are defined first. Those definitions are very useful in constructing the hierarchical image semantic model. We also discuss how to construct a semantic-binding hierarchical visual vocabulary tree based on the built hierarchical semantic model. Each sub vocabulary node in the tree is bound to a certain semantic. The semantic binding vocabulary helps to filter out the noise feature and refine the performance. Then the procedure of analyzing image semantic with semantic binding visual vocabulary is described in detail. And two feature dropping strategies are discussed. Experiments are performed based on the LabelMe dataset, the performance of our scheme is evaluated and compared with the traditional BOVW scheme. The experimental results demonstrate the efficiency and flexibility of our scheme. Our model can help to understand and analyze image semantic in a flexible multi-resolution way, and get the semantic recognition on each semantic node. But for traditional BOVW model, it is necessary to do the classification for each semantic. Our future work will focus on improving the performance of our method, choosing the proper distance learning metric algorithm and the applying in image retrieval system.

**Acknowledgments.** This work was supported by the National Natural Science Foundation of China (No. 60802057, 61071153), Sponsored by Shanghai Rising-Star Program (10QA1403700), Program for New Century Excellent Talents in University (NCET-10-0569).

## Reference

1. Niblack, W., Barber, R., Equitz, W., (ed.): The QBIC project: Querying images by content using color, texture, and shape. In Proceedings of SPIE on Storage and Retrieval for Image and Video Databases, 173-187. (1993)
2. Carneiro, G., and Vasconcelos, N.: Formulating semantic image annotation as a supervised learning problem. In Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 163-168. (2005)
3. Fan, J., Gao, Y., Luo, H.: Multi-level annotation of natural scenes using dominant image components and semantic concepts. In Proceedings of the 12th annual ACM international conference on Multimedia, 540-547. (2004)
4. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, Vol. 60, 91-110. (2004)
5. Barnard, K., Duygulu, P., Forsyth, D., Freitas, N.D., (ed.): Matching words and pictures. Journal of Machine learning Research, 1107-1135. (2003)
6. Jin R., Chai, J.Y., Si, L., Effective automatic image annotation via a coherent language model and active learning. In Proceedings of the 12th annual ACM international conference on Multimedia, 892-899. (2004)
7. Wu, L., Hu, Y., Li, M., Yu, N., Hua, X.S.: Scale-Invariant visual language modeling for object categorization. IEEE Transactions on Multimedia, Vol.11, No. 2, 286-294. (2009)
8. Tirilly, P., Claveau, V., Gros, P.: Language modeling for bag-of-visual words image categorization. In Proceedings of 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 249-258. (2008)
9. Yang, L., Jin, R., Sukthankar, R., Jurie, F.: Unifying discriminative visual codebook generation with classifier training for object category recognition. In Proceedings of 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1-8. (2008)
10. Jiang, Y.G., Ngo, C.W., Yang, J.: Toward optimal bag-of-features for object categorization and semantic video retrieval. In Proceedings of the 6th ACM international conference on Image and video retrieval, 494-501. (2007)
11. Yang, J., Jiang, Y.G., Hauptmann, A.G., Ngo, C.W.: Evaluating bag-of-visual-words representations in scene classification. In Proceedings of the international workshop on Workshop on multimedia information retrieval, 197-206. (2007)
12. Wu, L., Hoi, S.C.H., Yu, N.H.: Semantics-preserving bag-of-words models for efficient image annotation. In Proceedings of the First ACM workshop on Large-scale multimedia retrieval and mining. 19-26. (2009)
13. Russel, B.C, Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: A database and web-based tool for image annotation. International Journal of Computer Vision, Vol. 77(1-3), 157-173. (2008)

14. Leibe, B., Schiele, B.: Scale-Invariant Object Categorization Using a Scale-Adaptive Mean-Shift Search. Pattern Recognition. Lecture Notes in Computer Science, Vol. 3175, 145-153, Springer-Verlag, Berlin Heidelberg, New York. (2004)
15. Yang, L., Jin, R.: Distance metric learning: A comprehensive survey. [Online]. Available: http://www.cse.msu.edu/~yangliu1/frame_ survey_v2.pdf. (2006)
16. Xing, E., Ng, A., Jordan, M., Russell, S.: Distance metric learning with application to clustering with side-information. In Proceedings of the Conference on Advances in Neural Information Processing Systems, 505-512. (2003)
17. Hoi, S.C.H.,, Liu, W., Lyu, M.R., Ma, W.Y.: Learning Distance Metrics with Contextual Constraints for Image Retrieval. In Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2072-2078. (2006)
18. Sun, T.F., Jiang, X.H., Fu, G.L., Li, R.J., Feng, B.: Image Semantic Recognition Scheme with Semantic-binding Hierarchical Visual Vocabulary Model, In Proceedings of the 3rd International Congress on Image and Signal Processing, 1576-1582. (2010)

**Xinghao Jiang** received the PhD. degree in Electronic Science and Technology from Zhejiang University, Hangzhou, PR China in 2003. He is an associate professor of the School of Information Security Engineering at Shanghai Jiao Tong University, Shanghai, PR China. His current research interests include multimedia security and image retrieval, cyber information security, information hiding and watermarking, multimedia content management and rights management.

**Tanfeng sun** received the PhD. degree in Information and Communication Engineering from Jilin University, Jili, PR China in 2003. He is a lecturer of the School of Information Security Engineering at Shanghai Jiao Tong University, Shanghai, PR China. His current research interests include multimedia security and image retrieval, information hiding and watermarking.

**Guanglei Fu** received the M.S. degree from Shanghai Jiao Tong University, Shanghai, PR China in 2011. His current research interests include image and video processing, retrieval and classification, multimedia content protection.

*Received: April 23, 2010; Accepted: February 20, 2011.*