FROM REGION FEATURES TO SEMANTIC LABELS: A PROBABILISTIC APPROACH

RUI LI, WEE KHENG LEOW

School of Computing, National University of Singapore
3 Science Drive 2, Singapore 117543
lir, leowwk@comp.nus.edu.sg

Content-based image retrieval has advanced from the initial stage of feature-based approach towards the semantic approach. Existing semantics-based methods typically classify an image or an image region to exactly one of several classes. Due to the presence of noise and ambiguity in images, it is practically impossible to derive classifiers that can accurately classify the images or regions into a large variety of classes. Therefore, some image retrieval methods have captured the uncertainty of region classification in the region labels and used image structures for disambiguation during image matching. This paper presents a novel method of semantic labeling that can assign multiple semantic labels to a region along with the confidence measures of the assignment. Unlike existing classification methods, it can learn to perform semantic labeling incrementally. Test results show that the method is effective and accurate in labeling a wide variety of regions.

1 Introduction

Early content-based image retrieval (CBIR) systems follow the paradigm of representing images by low-level features, such as color, texture, and shape. Image retrieval is performed by matching the features of the query image with those of the database images. This approach of matching global image features is effective for retrieving simple images and images that contain single distinct objects. CBIR systems such as QBIC [1], Virage [2], ImageRover [3], Photobook [4] and VisualSEEK [5] all belong to this category.

For more complicated images with multiple objects and regions, local features are extracted from segmented regions or fixed-sized blocks. Images are then retrieved by matching their region or block features with those in the query image. Such region-based CBIR systems include Netra [6], Blobworld [7], and SIMPLIcity [8].

All the above methods use low-level feature-based methods to match images. On the other hand, high-level semantic labels are more natural for human users. In recent years, CBIR research has shifted its focus towards bridging the gap between low-level features and high-level semantics. The general idea is to assign semantic labels to parts of an image or the whole image.

There are several approaches to semantic labeling. The first approach extracts global image features from the images and classifies the images into several coarse categories based on the global features. Examples of coarse categories include indoor vs. outdoor [9], city vs. landscape [10], textured vs. non-textured [8], etc.

The second approach extracts local features from the segmented regions (or fixed-sizes blocks) of images and classifies the regions into several semantic classes. Image matching is then performed based on the semantic labels of the image regions. The third approach classifies images into several categories based on the semantic labels of the image components. Campbell et al. [11] and Town and Sinclair [12] focused on the second approach while Fung and Loe [13] explored both approaches.

All the above methods perform crisp semantic classification, i.e., classification of a region or an image into exactly one semantic class. The shortcoming of this approach is that, due to the presence of noise and ambiguity in the images, it is practically impossible to derive very accurate classifiers for a large variety of region classes, let alone a large variety of image classes. The methods reported in the literature typically have been shown to work on only about 10 or fewer region and image classes (e.g., [8,9,10,11,12,13]).

To explore the semantic approach more fully, methods have been proposed to match images using both fuzzy region labels and image structures [14,15,16]. In particular, the methods of [15] acknowledge the difficulty of accurately classifying regions to semantic classes, and so represent each region by multiple fuzzy semantic labels. Disambiguation of the fuzzy labels is performed during image matching where image structures are used to constrain the matching between the query example and the images. Similarly, attributed relational graph matching method [14,16] can also use image structures for image matching. Test results demonstrated that these methods are able to make fine discrimination between images that share common parts.

In all the above semantics-based methods, labeling of image regions or whole images is the most important part of the methods. This paper proposes a novel method of performing semantic labeling of image regions (or fixed-sized blocks). It focuses on the task of analyzing the features in a region and assigning possibly multiple semantic labels to the region together with confidence measures of classifying the regions to the corresponding classes. The semantic labeling method adopts a probabilistic approach to estimate the confidence of assigning a semantic label to a region (Section 3). It includes algorithms for incremental learning based on clustering, probability estimation, and region labeling (Section 4). Test results show that the method is effective and accurate in assigning semantic labels to image regions (Section 5). The application of the semantic labeling method to image retrieval that uses image structures is reported separately in other papers [15].

2 Related Work

Region labeling methods that are most related to our work include [11,12,13]. Campbell et al. [11] trained a neural network to classify regions based on features such as color, texture, shape, size, rotation, and centroid. The regions were classified into 11 categories, such as sky, vegetation, road, and pavement, that appear in normal road scenes, and the neural network achieved 83.9% accuracy.

Town and Sinclair [12] also applied a neural network to classify regions into semantic classes based on region features that include area, boundary length, color mean and covariance matrix, texture orientation and density, moment, etc. They obtained a classification accuracy of about 90% but the test was perform on only 11 classes that contain well-defined textures such as brick, cloud, fur, grass, road, and sand.

Fung and Loe [13] used supervised clustering of color features to group them into a large number of elementary clusters, which were further grouped into conglomerate clusters. Each conglomerate cluster was associated with a semantic region class. Fixed-sized image blocks were assigned to the clusters using k-nearest-neighbor algorithm, and were then assigned the semantic labels of the majority clusters. The number of region classes and accuracy of region classification were not reported in the paper.

The common shortcomings of the existing methods include the following:

- They classify region features in a Euclidean space that combine various feature types linearly. This vector space approach is convenient but requires the assumption that the features types are independent of each other and, thus, can be regarded as forming orthogonal dimensions of the vector space. This assumption is generally false and has been shown to lead to poorer classification and clustering results [17]. Further discussion of this issue is presented in Section 3.1.

- Existing methods have been demonstrated to work on only a small number of about 10 region classes. Moreover, images in the region classes usually have well-defined texture patterns.

- The classification methods used are not incremental. Addition of new training samples, feature types, and semantic classes entails the retraining of the entire collection of training samples.

In contrast, our method does not combine different feature types linearly in a Euclidean space. Instead, it adopts a probabilistic approach that captures the correlation between feature combinations and semantic classes. It adopts an incremental learning algorithm that can make use of the dissimilarity measure that is appropriate for each feature type. Finally, it has been tested on 30 semantic classes that span a wide variety of region types that are not restricted to images that contain well-defined textures.

3 Semantic Labeling

3.1 Overview

Semantic labeling can be framed as a problem of classifying an image region (or image patch, fixed-sized block) R to one of several semantic classes $C_i, i = 1, \ldots, m$. In practice, it is not possible to achieve perfect classification due to the presence of noise and ambiguity in the region features. A better approach is to represent the uncertainty of classification in the semantic labels and to defer the final decision to a latter stage at which the image structure can be taken into account using, for instance, the fuzzy conceptual graph matching method [15]. Let us denote by $Q_i(R)$ the confidence of classifying region R into semantic class C_i. Then, the semantic labeling problem can be defined as follows:

Semantic Labeling

Given a region R and M semantic classes $C_i, i = 1, \ldots, M$, compute the confidence $Q_i(R)$ that R belongs to C_i for each i.

A region R contains a set of features F_t of type $t = 1, \ldots, m$, each having a value v_t. Each feature F_t can contain more than one component, as for color histograms, Gabor texture features, wavelet features, etc. The symbols F_t and v_t denote the *whole* feature and feature value instead of the individual component.

The standard method of computing $Q_i(R) = Q_i(v_1, \ldots, v_m)$ is the vector space approach: Regard each set of feature values as a vector $\mathbf{v} = [v_1, \ldots, v_m]^T$ in a linear vector space, and estimate Q_i using function approximation. It is well-known that the various feature types are not independent of each other and the scales of the feature types are different. So, forming a linear vector space by assigning each feature type (or, more commonly, each feature component of each feature type) to an orthogonal dimension of the vector space is not expected to produce reliable results in general [18].

The usual method of dealing with this problem is to represent a region as a linear combination of feature values, and define $Q_i(R)$ as a function of the linear combination:

$$Q_i(R) = Q_i\left(\sum_t w_t v_t\right) = Q_i(\mathbf{w}^T\mathbf{v}) \tag{1}$$

for some weight vector $\mathbf{w} = [w_1, \ldots, w_m]^T$. Another method is to generalize the weighted sum to the quadratic form [19]:

$$Q_i(R) = Q_i\left((\mathbf{v} - \mu)^T \mathbf{X}^{-1}(\mathbf{v} - \mu)\right) \tag{2}$$

where μ is a $m\times 1$ weight matrix and $\mathbf{X}$ is a $m\times m$ weight matrix. The matrices μ and $\mathbf{X}$ can be taken as the mean vector and the covariance matrix of $\mathbf{v}$, but this would require the assumption of a linear vector space, which is not desirable as discussed above. So, more appropriate weight matrices should be obtained. The main difficulty is that it is not known *a priori* what are the appropriate values of the weights. It is also difficult to apply a learning method to obtain the weight values (as in [19]) because the desired values of $Q_i(R)$ are also unknown *a priori*, and they depend very much on the type of classifier used.

3.2 Probabilistic Labeling

To resolve the problems highlighted above, this paper presents a probabilistic method of computing $Q_i(R)$ by estimating the conditional probability $P(C_i \mid v_t)$. The approach encompasses the following characteristics:

- It can make use of the dissimilarity measures that are appropriate for the various types of features [17,20] instead of the Euclidean distance.
- It does not require the use of weights to combine the various feature types.
- It adopts a learning approach that can adapt incrementally to the inclusion of new training samples, feature types, and semantic classes.

The probabilistic labeling method consists of two stages: (1) semantic class learning and (2) region labeling.

Semantic Class Learning

The goal of the semantic class learning stage is to determine the conditional probabilities associated with each semantic class. It first clusters a set of training sample regions R_j, each assigned a pre-defined semantic class C_i, according to each feature type using the dissimilarity measure that is appropriate for the feature type (see Section 4 for details). This process produces a set of clusters Ω_{tk}, for each feature type t. After clustering, the conditional probability $P(C_i \mid \Omega_{tk})$ for semantic class C_i given cluster Ω_{tk} is estimated. Assuming that the distribution within each cluster is uniform, then $P(C_i \mid \Omega_{tk})$ can be estimated from the number of regions in the cluster:

$$P(C_i \mid \Omega_{tk}) = \frac{P(C_i, \Omega_{tk})}{P(\Omega_{tk})} = \frac{|C_i \cap \Omega_{tk}|}{|\Omega_{tk}|} \tag{3}$$

where $|\Omega_{tk}|$ denotes the number of regions in cluster Ω_{tk}, and $|C_i \cap \Omega_{tk}|$ the number of regions in Ω_{tk} that belong to semantic class C_i.

To combine multiple feature types, we can determine the cluster combinations $\Psi(\tau, \kappa, n) = \{\Omega_{\tau(1),\kappa(1)}, \ldots, \Omega_{\tau(n),\kappa(n)}\}$ that have high probabilities of associating to some semantic classes C_i:

$$\begin{aligned} P(C_i \mid \Psi(\tau, \kappa, n)) &= P(C_i \mid \Omega_{\tau(1),\kappa(1)}, \ldots, \Omega_{\tau(n),\kappa(n)}) \\ &= \frac{\left| C_i \cap \bigcap_l \Omega_{\tau(l),\kappa(l)} \right|}{\left| \bigcap_l \Omega_{\tau(l),\kappa(l)} \right|} . \end{aligned} \tag{4}$$

The functions $\tau(l)$, $l = 1, \ldots, n$, denote a combination of feature types and $\kappa(l)$ a combination of cluster indices.

In practice, a semantic class is often associated with a specific combination of feature types. So, it is necessary to compute only the conditional probabilities $P(C_i \mid \Psi(\tau, \kappa, n))$ that are significantly larger than zero; the conditional probabilities associated with the other cluster combinations can be regarded as zero. The cluster combination $\Psi(\tau, \kappa, n)$, the associated semantic classes C_i, and the corresponding conditional probability values $P(C_i \mid \Psi(\tau, \kappa, n))$ are stored for region labeling.

Region Labeling

After the learning stage, a region R can be labeled by determining the associated semantic classes. Given the region R which contains a set of feature values v_t, the clusters that are nearest to the feature values v_t, for each feature type t, are determined. Next, the nearest clusters found are matched with the stored cluster combinations, obtained during the learning stage, that are associated with some semantic classes. The confidence measure $Q_i(R)$ can now be computed from the conditional probabilities of the matching cluster combinations $\Psi(\tau, \kappa, n)$:

$$Q_i(R) = \max_{\tau,\kappa,n} P(C_i \,|\, \Psi(\tau, \kappa, n))\,. \tag{5}$$

Note that $Q_i(R)$ as defined in Eq. 5 is no longer a conditional probability:

- The C_i's in Eq. 5 may be conditioned on different sets of feature types.
- While $\sum_i P(C_i \,|\, \Psi(\tau, \kappa, n)) = 1$ for each cluster combination $\Psi(\tau, \kappa, n)$, the sum $\sum_i Q_i(R) \neq 1$ in general.

Nevertheless, $Q_i(R)$ is well-founded on probability theory and is, thus, a good measure of the confidence that region R belongs to class C_i.

Discussion

The semantic labeling method presented above computes $Q_i(R)$ from the probabilities conditioned on the clusters nearest to the feature values v_t of R instead of the probabilities conditioned on the feature values v_t. In principle, it is possible to estimate $P(C_i \,|\, v_t)$ using more sophisticated methods such as Gaussian mixture, e.g.,

$$P(C_i \,|\, v_t) = \sum_k w_k \exp\left(-\frac{\|v_t - \mu_{tk}\|^2}{2\,\sigma_{tk}^2}\right) \tag{6}$$

where w_k is a weighting factor that can be optimized during learning, μ_{tk} and σ_{tk} are derived from the location and radius of cluster Ω_{tk}, and $\|v_t - \mu_{tk}\|$ is an appropriate dissimilarity measure. However, estimation of probability density that is conditioned on several feature types is more complex. Let $\mathbf{v} = [v_1, \ldots, v_m]^T$ denote a vector that combines the feature values v_t, $t = 1, \ldots, m$. Then,

$$P(C_i \,|\, \mathbf{v}) = \sum_k w_k \exp\left(-\frac{1}{2}(\mathbf{v} - \mu)^T \mathbf{X}^{-1} (\mathbf{v} - \mu)\right) \tag{7}$$

where $\mu = [\mu_{1,\kappa(1)}, \ldots, \mu_{m,\kappa(m)}]^T$ is the vector of the centroids of the clusters in which the feature values v_t lie, and $\mathbf{X}$ is an appropriate weight matrix. That is, this method defines the conditional probability in terms of a quadratic form of $\mathbf{v}$, which is undesirable (as discussed in Section 3.1). Therefore, in our current formulation, uniform probability distribution is assumed within each cluster.

4 Semantic Labeling Algorithms

The semantic labeling method described in Section 3 consists of several main algorithms: region clustering, probability estimation, region labeling, and region classification. The first two algorithms are used for semantic class learning. The region classification algorithm is a special case of region labeling.

4.1 Region Clustering

Training sample regions R_j are clustered according to individual feature type t to obtain clusters Ω_{tk}. An adaptive k-means clustering algorithm [17] is used so that the appropriate number of clusters can be determined automatically. This is achieved by setting the maximum radius r of a cluster and the nominal separation s between clusters. The maximum radius r, for feature type t, is computed by measuring the average distance between the samples in a class:

$$r = \frac{1}{M} \sum_i \frac{1}{N(i)} \sum_{R_j, R_k \in C_i} d(v_{tj}, v_{tk}) \tag{8}$$

where M is the number of classes, $N(i)$ is the number of sample pairs in class C_i, and v_{tj} and v_{tk} are the type-t feature values of region R_j and R_k in class C_i. The nominal separation s is set at $1.5R$, which has been found in [17] to produce clusters with an optimal amount of overlaps between them. The same algorithm is applied to different feature types separately. The algorithm can be summarized as follows:

Adaptive Clustering

Repeat
- For each feature value v_t of each region,
  - Find the nearest cluster Ω_{tk} to v_t.
  - If no cluster is found or distance $d(\Omega_{tk}, v_t) \geq s$,
    - create a new cluster with feature v_t;
  - Else, if $d(\Omega_{tk}, v_t) \leq r$,
    - add feature value v_t to cluster Ω_{tk}.
- For each cluster Ω_{ti},
  - If cluster Ω_{ti} has at least N_m feature values,
    - update centroid of cluster Ω_{ti};
  - Else, remove cluster Ω_{ti}.

The centroid of cluster Ω_{ti} is a generalized mean of the feature values in the cluster (see Section 5 for more discussion). The function $d(\Omega_{tk}, v_t)$ is a dissimilarity measure appropriate for the feature type t [17,20].

4.2 Probability Estimation

After clustering the regions according to individual feature types, the conditional probability $P(C_i \,|\, \Omega_{tk})$ of each cluster Ω_{tk} is estimated. In addition, the conditional probabilities $P(C_i \,|\, \Psi(\tau, \kappa, n))$ of various cluster combinations $\Psi(\tau, \kappa, n)$ are also estimated.

In order to assess the accuracy of the semantic labeling approach, the conditional probability for all possible combinations of 1 to 4 features are computed in the current implementation so that the combinations with the highest probabilities can be identified. In actual applications, a cluster selection procedure can be performed to select candidate cluster combinations that are likely to yield significant probabilities. This method would remove the need to compute the probabilities for all possible combinations. The cluster combinations $\Psi(\tau, \kappa, n)$, their associated semantic classes C_i, and the corresponding conditional probabilities $P(C_i \,|\, \Psi(\tau, \kappa, n))$ that are significantly larger than zero are stored for region labeling.

4.3 Region Labeling and Classification

The region labeling algorithm determines the combinations of feature values of a region that are associated with some semantic classes with high probabilities. These semantic classes and the probabilities are assigned to the region as its semantic labels. In general, a region can be assigned multiple semantic classes.

Region Labeling

Given a region R

- Find the nearest cluster of each feature value v_t of R.
- Find the combinations Ψ_j of these clusters that match the stored cluster combinations.
- Retrieve the classes C_i and probabilities $P(C_i \,|\, \Psi_j)$ associated with the matching cluster combinations Ψ_j.
- For each retrieved class C_i,
  - Find the Ψ_k with the largest probability:
    $P(C_i \,|\, \Psi_k) = \max_j P(C_i \,|\, \Psi_j)$
  - Assign C_i and $P(C_i \,|\, \Psi_k)$ to R,
    i.e., $Q_i(R) = P(C_i \,|\, \Psi_k)$.

For the purpose of assessing the effectiveness of the semantic labeling method, region classification is performed to assign the label of the semantic class with the highest confidence to a region.

Region Classification
Given a region R
  Perform region labeling.
  Find the C_k with the largest confidence:
    $Q_k(R) = \max_i Q_i(R)$
  If $Q_k(R) >$ threshold Γ,
    assign C_k to R;
  else assign "unknown class" to R.

In some applications, it may be better to assign the "unknown class" label to a region when the probability is low than to assign it a wrong label. The threshold Γ can be determined empirically (see Section 5).

5 Performance Evaluation

Extensive tests were performed to evaluate the following aspects of the semantic labeling method:

- Is the confidence value estimated by the method a reliable measure of classification accuracy?
- Can the method improve the confidence value by combining the feature types that are most salient for classifying a region?

5.1 Test Setup

A wide variety of 30 semantic classes (Fig. 1) were randomly identified by browsing the images in the Corel 50,000 photo collection. For each class, 250 image blocks of size 64×64 pixels were cropped from the images. Out of the 250 blocks, 200 were randomly selected for semantic class learning and the remaining 50 for region classification test. In total, 6,000 blocks were used for training and 1500 blocks for testing. During semantic class learning, combinations of 1 to 4 feature types were considered, and the conditional probabilities for all these combinations were computed.

For each image block, four different types of features were extracted:

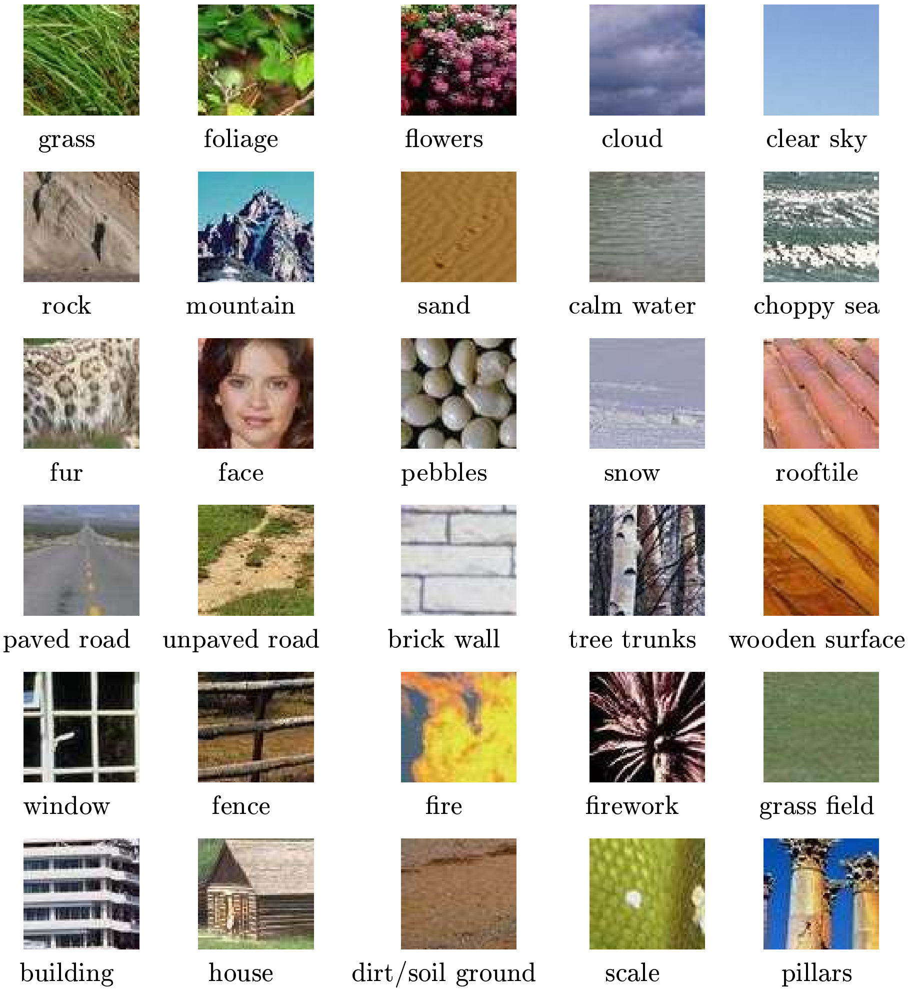

Figure 1. Sample images of the semantic classes used in the tests.

1. Adaptive color histograms:
   It has been shown in [17] that adaptive color histograms have better overall performance, in terms of good accuracy, small number of bins, no empty bin and efficient computation compared to normal fixed-binning

histograms. The weighted correlation [17] is used to compute the dissimilarity between two adaptive histograms. Mean histogram [18] is used to compute cluster centroids.

2. MRSAR texture features:
   Liu and Picard showed that multiresolution simultaneous autoregressive (MRSAR) model is good for capturing the characteristics of random textures [21]. The usual dissimilarity measure for MRSAR is the Mahalanobis distance, and the Euclidean mean is used to compute cluster centroids.

3. Gabor texture features:
   The Gabor texture features and weighted-mean-variance (WMV) as defined by Ma and Manjunath [22] have been shown to produce good texture discrimination, particularly for structured and oriented textures. WMV is a good dissimilarity measure for Gabor texture features. For computing cluster centroids, Euclidean mean is used.

4. Edge histograms:
   Normalized edge direction and magnitude histograms as given in [23,24] are extracted from the images. For these features, Euclidean distance and Euclidean mean are used for region clustering.

5.2 Clustering Results

The region clustering algorithm produced 137 color clusters, 37 MRSAR clusters, 31 Gabor clusters, and 28 edge clusters. It produced more color clusters than other clusters because there were more color variations than texture and edge variations in the images. Most of the clusters had low probabilities $P(C_i \,|\, \Omega_{tk})$ when there were considered individually (Table 1). This shows that individual features were not discriminative. However, the probabilities improved significantly when they were combined appropriately (Table 1). These results are analyzed in more details in the following sections.

5.3 Salient Features

Salient features are features that are highly correlated with a semantic class. During semantic class learning, cluster combinations with high probability of associating with various semantic classes are identified. The feature values of the cluster centroids constitute the salient features of the classes.

Table 1. Salient features. Columns 2–5 give the average confidence measures of the semantic classes using single features (Color (**C**), MRSAR (**M**), Gabor (**G**) and Edge (**E**)) . Numbers in bold are the highest average confidence among the four feature types. Column 6 lists the salient featur e pairs (S.F.) and column 7 lists the corresponding improved average confidence (Conf.) using sal ient feature pairs.

ID	Class	C	M	G	E	S. F.	Conf.
1	grass	**0.229**	0.096	0.058	0.081	color, Gabor	0.746
2	foliage	**0.321**	0.085	0.123	0.132	color, MRSAR	0.801
3	flower	**0.191**	0.097	0.098	0.156	color, edge	0.796
4	cloud	0.285	0.339	0.365	**0.381**	color, Gabor	0.741
5	clear sky	0.452	0.737	0.758	**0.847**	color, Gabor	0.860
6	rock	**0.082**	0.082	0.048	0.050	color, MRSAR	0.755
7	mountain	0.098	0.039	0.038	**0.120**	color, edge	0.697
8	sand	0.109	**0.144**	0.041	0.054	color, MRSAR	0.724
9	calm water	**0.139**	0.042	0.039	0.087	color, MRSAR	0.708
10	choppy sea	**0.159**	0.043	0.042	0.062	color, MRSAR	0.733
11	fur	0.082	**0.132**	0.034	0.048	color, MRSAR	0.740
12	face	**0.229**	0.099	0.115	0.101	color, edge	0.755
13	pebbles	0.094	0.108	**0.141**	0.055	MRSAR, edge	0.697
14	snow	0.202	0.053	0.045	0.078	color, MRSAR	0.442
15	roof tiles	0.094	0.078	**0.252**	0.079	color, Gabor	0.786
16	paved road	**0.116**	0.050	0.045	0.095	color, edge	0.715
17	unpaved road	**0.095**	0.059	0.046	0.066	color, edge	0.698
18	brick wall	0.088	0.059	**0.239**	0.181	color, Gabor	0.635
19	tree trunks	0.090	0.052	0.051	**0.157**	color, edge	0.736
20	wooden surface	**0.155**	0.102	0.046	0.103	color, MRSAR	0.768
21	window	0.077	0.052	0.048	**0.197**	color, edge	0.742
22	fence	0.089	0.054	0.046	**0.166**	color, edge	0.750
23	fire	**0.141**	0.090	0.076	0.068	color, MRSAR	0.768
24	firework	**0.131**	0.047	0.048	0.120	color, MRSAR	0.785
25	grass field	**0.236**	0.035	0.028	0.069	color, MRSAR	0.704
26	building	0.079	0.055	0.047	**0.154**	color, edge	0.765
27	house	0.072	0.051	0.042	**0.116**	color, edge	0.755
28	dirt ground	**0.150**	0.054	0.051	0.081	color, MRSAR	0.658
29	scale	0.141	0.104	**0.152**	0.057	color, Gabor	0.756
30	pillars	**0.143**	0.072	0.087	0.121	color, edge	0.697

Table 1 tabulates the confidence measures of a semantic class C_i averaged over all samples that belong to C_i, i.e.,

$$Q_i(\Psi(\tau,n)) = \frac{1}{|C_i|}\sum_{\kappa}|C_i \cap \Psi(\tau,\kappa,n)|\ P(C_i \mid \Psi(\tau,\kappa,n))\,. \qquad (9)$$

The average confidence gives an overall assessment of how strongly a feature type correlates with a semantic class. With a single feature, almost all semantic classes have very low confidence values. This confirms the expected results that single features are not enough to identify the semantic classes of image regions. An interesting surprise is that MRSAR, Gabor, and edge histograms are highly correlated with clear sky. This is due to the fact that clear sky regions have almost no texture and no edge whereas all other image blocks have some textures and edges. Therefore, the learning method can associate not only the *presence* but also the *absence* of features to semantic classes. Whatever the case may be, the learning method always chooses the one with the highest confidence.

Table 1 also shows that using a combination of only two feature types can already improve the mean confidence values of all the semantic classes significantly. The mean confidence values of all classes except snow are above 0.6, and the overall average is 0.731. Increasing the number of feature types to 3 or 4 did not produce higher confidence values. So, for our data set of 30 semantic classes, combinations of two feature types are enough.

It is interesting to see that a salient pair of features may not be individually salient. For example, for the grass images, the Gabor feature is less salient than MRSAR. Nevertheless, the combination of color and Gabor is the most salient pair for grass. Another interesting example is the class of pebbles. For this class, MRSAR and edges constitute the salient pair, but individually both features are less salient than Gabor. Color is not found to be a salient feature because the pebble images in the training set contain large variations of colors.

The above results are consistent with those of Szummer and Picard for indoor vs. outdoor classification of images [9]. They observed that a pair of features is more accurate for image classification than a single feature. Moreover, combining two weak features consistently produced more accurate classification than a single good feature.

5.4 Classification Accuracy

To test whether the confidence measure estimated by the method correlate with classification accuracy, a region classification test was performed on 1,500 testing image regions, 50 per semantic class. The region classification method described in Chapter 3 was executed at various threshold values. Figure 2 plots the classification accuracy vs. the maximum confidence of a region R_i, i.e., $Q_M(R_j) = \underset{i}{Max} Q_i(R_j)$. To compute the classification accuracy, a recursive algorithm was applied to group the samples into groups containing samples with similar $Q_M(R_j)$.Test results in Figure 2 shows that above confidence value of 0.75, the accuracy is above 0.9.

Regions labeled with semantic classes having low confidence values are ambiguous. In applications where image structures are used for image matching, such as [14,15,16], image structures provide additional information that can be used for disambiguation. So, regions labeled with "unknown class" should

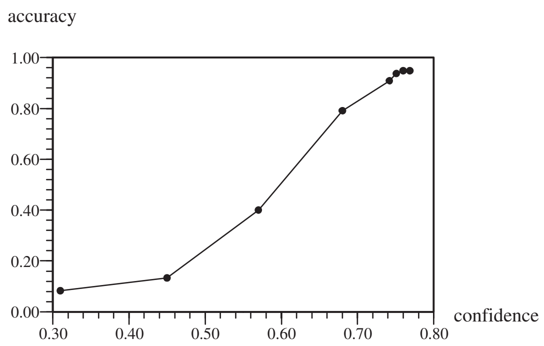

Figure 2. Region classification accuracy at various confidence levels.

Table 2. Classification Accuracy Comparison (SVM vs. probabilistic labeling).

SVM Labeling	Probabilistic Labeling	
	with rejection	without rejection
61.6%	70%	91%

be regarded as ambiguous and should not be classified prematurely by the semantic labeling algorithm. If these regions are rejected and not classified due to low confidence in classification, then the classification accuracy of the remaining regions improves significantly from 0.70 to 0.91 (Fig. 3). The amount of rejection for threshold value of 0.75 is 23%. The confusion matrix in Fig. 4 further confirms the improved classification accuracy.

To compare the performance of our labeling method with traditional approach, support vector machine (SVM) was used for the region classification problem. For probabilistic labeling method without rejection of low confidence samples, a classification accuracy of 70% was achieved (Table 2). With rejection, the accuracy increased to 90%. The usual SVM implementation does not perform rejection, and is has the lowest classification accuracy of 61.6%. From the above test results, we can conclude that the confidence values estimated by the semantic labeling algorithm are very reliable: a confidence value greater or equal to 0.75 translates to a mean classification accuracy of 91% or higher. For regions with low confidence, multiple labels are assigned to them together with the corresponding confidence values. These information are much more valuable than single class labels for higher-level modules such as image retrieval and image classification. These modules can regard regions with high confidence to be correctly classified into the respective semantic

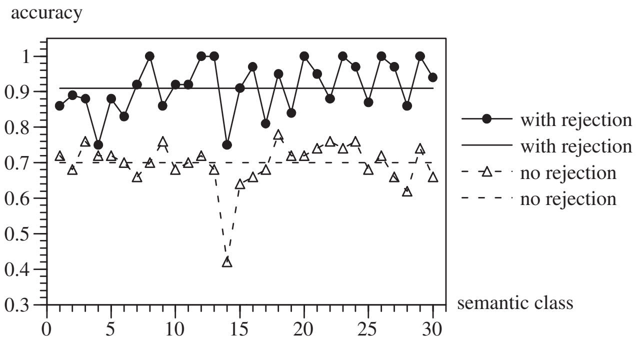

Figure 3. Region classification accuracy at threshold of 0.75 (solid lines: with rejection, dotted lines: without rejection, horizontal lines: mean accuracy).

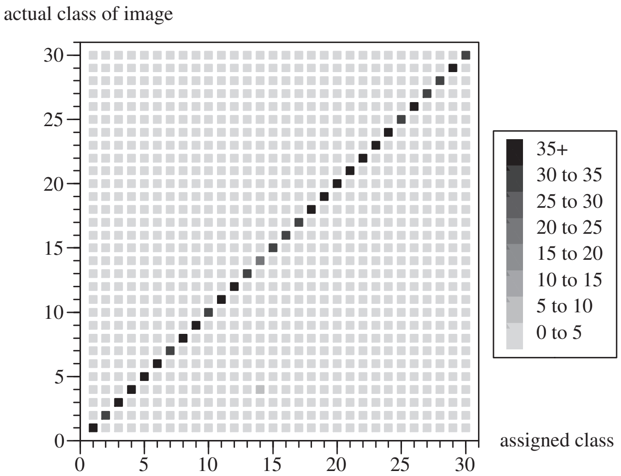

Figure 4. Confusion matrix of region classification at threshold of 0.75 (with rejection). Numbers along the axes denote the semantic class indices shown in Table 1.

classes indicated by their labels. On the other hand, the multiple labels and confidence measures of regions with low confidence can allow the modules to disambiguate between various possibilities using image structures, e.g., by applying fuzzy conceptual graph matching [15] or attributed relational graph matching [14,16]. This kind of disambiguation would be impossible without the multiple labels and confidence measures.

6 Conclusion

This paper presented a probabilistic approach for semantic labeling of image regions. Unlike existing methods that assign a single label to a region, our method assigns multiple labels to a region together with the confidence measures of classifying the region to the corresponding semantic classes. Moreover, it derives the confidence measures by clustering the regions according to each feature type separately and using the dissimilarity measure that is appropriate for each feature type. The probabilistic method is applied to combine feature types without arbitrary weights to measure the confidence of classifying regions using combined features. The learning algorithm is incremental in that adding new feature types and new semantic classes does not require re-training over the entire set of existing training samples.

Extensive performance tests have been conducted on a wide variety of 30 region types, which are not restricted to regions with distinct texture patterns. In total, 6,000 samples were used for training and 1,500 different samples were used for testing. Test results show that the learning method can determine the combinations of features that are most useful for identifying each semantic class of regions. Moreover, the confidence values estimated by the method is very reliable. In particular, it is found that regions with confidence measures of greater or equal to 0.75 can be classified into the correct semantic classes with an average accuracy of 91%. For regions with lower confidence values, the multiple semantic labels and the corresponding confidence values that they carry allow a higher-level algorithm, such as fuzzy conceptual graph matching and attributed relational graph matching, to disambiguate them using information about image structures. In summary, the semantic labeling method presented in this paper is expected to contribute significantly to bridging the gap between low-level features and high-level semantics for image retrieval.

7 Acknowledgments

This research is supported by NUS ARF R-252-000-072-112 and NSTB UPG/98/015.

References

1. J. Hafner, H. S. Sawhney, W. Esquitz, M. Flickner, and W. Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Trans. PAMI*, 17:729–736, 1995.

2. A. Gupta and R. Jain. Visual information retrieval. *Comm. of the ACM*, 40(5), 1997.
3. S. Sclaroff, L. Taycher, and M. La Cascia. Image-Rover: A content-based image browser for the world wide web. In *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1997.
4. A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Tools for content-based manipulation of image databases. *Int. J. Computer Vision*, 18(3):233–254, 1996.
5. J. R. Smith and S.-F. Chang. Single color extraction and image query. In *Proc. ICIP*, 1995.
6. W. Y. Ma and B. S. Manjunath. NeTra: A toolbox for nagivating large image databases. In *Proc. ICIP*, pages 568–571, 1997.
7. C. Carson, S. Belongie, H. Greenspan, and J. Malik. Region-based image querying. In *Proc. CVPR Workshop on Content-Based Access of Image and Video Libraries*, 1997.
8. J. Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics sensitive integrated matching for picture libraries. *IEEE Trans. on PAMI*, 23(9):947–963, 2001.
9. M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *Proc. ICCV Workshop on Content-based Access of Image and Video Databases*, pages 42–51, 1998.
10. A. Vailaya, A. Jain, and H. J. Zhang. On image classification: City images vs. landscapes. *Pattern Recognition*, 31:1921–1935, 1998.
11. N. W. Campbell, W. P. J. Mackeown, B. T. Thomas, and T. Troscianko. Interpreting image databases by region classification. *Pattern Recognition*, 30(4):555–563, 1997.
12. C. Town and D. Sinclair. Content based image retrieval using semantic visual categories. Technical Report 2000.14, AT&T Laboratories Cambridge, 2000.
13. C. Y. Fung and K. F. Loe. Learning primitive and scene semantics of images for classification and retrieval. In *Proc. ACM Multimedia*, pages II: 9–12, 1999.
14. S. Medasani and R. Krishnapuram. A fuzzy approach to content-based image retrieval. In *Proc. IEEE Conf. on Fuzzy Systems*, pages 1251–1257, 1999.
15. P. Mulhem, W. K. Leow, and Y. K. Lee. Fuzzy conceptual graph for matching images of natural scenes. In *Proc. Int. Joint Conf. on Artificial Intelligence*, pages 1397–1402, 2001.
16. G. M. Petrakis and C. Faloutsos. Similarity searching in large image databases. *IEEE Trans. on Knowledge and Data Engineering*, 9(3):435–

447, 1997.
17. W. K. Leow and R. Li. Adaptive binning and dissimilarity measure for image retrieval and classification. In *Proc. IEEE CVPR*, 2001.
18. W. K. Leow and R. Li. Clustering and classification of adaptive-binning histograms. *IEEE Trans. on PAMI (submitted).*
19. Y. Rui and T. Huang. Optimizing learning image retrieval. In *Proc. IEEE CVPR*, 2000.
20. J. Puzicha, J. M. Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity for color and texture. In *Proc. ICCV '99*, pages 1165–1172, 1999.
21. F. Liu and R. W. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. PAMI*, 18(7):722–733, 1996.
22. B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. PAMI*, 8(18):837–842, 1996.
23. Sami Brandt. Use of shape features in content-based image retrieval. Master's thesis, Helsinki University of Technology, Finland, 1999.
24. J. R. Smith, S. Basu, G. lyengar, C.-Y. Lin, M. Naphade, B. Tseng, S. Srinivasan, A. Amir, and D. ponceleon. Integrating features, models, and semantics for trec video retrieval. In *Proc. of The Tenth TREC*, 2001.