Image modeling with combined optimization techniques for image semantic annotation

Dong Yang · Ping Guo

Abstract Image semantic annotation can be viewed as a multi-class classification problem, which maps image features to semantic class labels, through the procedures of image modeling and image semantic mapping. Bayesian classifier is usually adopted for image semantic annotation which classifies image features into class labels. In order to improve the accuracy and efficiency of classifier in image annotation, we propose a combined optimization method which incorporates affinity propagation algorithm, optimizing training data algorithm, and modeling prior distribution with Gaussian mixture model to build Bayesian classifier. The experiment results illustrate that the classifier performance is improved for image semantic annotation with proposed method.

Keywords Image annotation · Training data selection · Gaussian mixture model · Affinity propagation algorithm

1 Introduction

Image semantic annotation is the procedure that the database of images are annotated with semantic labels by a computer system automatically. Image semantic annotation can be viewed as a mapping procedure from image features to semantic labels, by the steps of image modeling and image semantic mapping. Usually, image features include low-level visual features (color, shape, texture, and topology), object-level features and 3-dimension scene features. Semantic labels include feature semantics, object semantics, scene semantics, behavior semantics, and emotion semantics [2].

For image semantic annotation, there are two categories of methods. When each semantic label is viewed as a variable, the mapping is a image semantic joint modeling problem, such as N-cut based method [3, 4], latent dirichlet allocation (LDA) method [5] and cross-media relevance models (CMRM) [6].

When each semantic label is considered as a class, the mapping can be viewed as a classification problem. Supervised OVA (one vs. all) adopted two-class classifiers to learn from positive and negative images, while the positive images have the given semantic label and the negative images do not have [7]. Luo and Savakis [8] have approached the scene classification using a divide-and-conquer strategy, a good first step of which is to consider only two classes such as indoor and outdoor images, while the latter may be further subdivided into city and landscape images.

Lienhart and Hartmann [9] proposed an image classification algorithm using Adaboost. Shao et al. [10] tried to find out appropriate visual features for semantic annotation and proposed an image classification system combining MPEG-7 visual descriptors and SVM, which is used to annotate city and natural scenes. Guo et al. [11] adopted SVM and Adaboost to learn the boundary of different classes. Basilil et al. [12] proposed an unsupervised image classification techniques combining different media-layer features using latent semantic analysis (LSA). Shen et al. [13] tried to extract similar regions from user-provided

D. Yang · P. Guo (✉)
Laboratory of Image Processing and Pattern Recognition,
Beijing Normal University, 100875 Beijing, China
e-mail: pguo@ieee.org

D. Yang
e-mail: d.yang@ieee.org

similar training images, and classify them by comparing the distance and spatial relationship between similar region pairs.

In automatic linguistic indexing of pictures (ALIP) algorithm, categorized images are used to train multiple statistical models each representing a major concept [14]. To measure the extent of association between an image and a concept, the likelihood of the occurrence of the image is computed. Each category is loosely defined and is represented by a set of words that characterize the category as a whole, but it may not accurately characterize each individual image. In ALIP method, a category classifier is built to determine the top several image categories to which the test image is most likely to belong. Then the annotations from those categories are pooled together, and the frequency of each candidate annotation is counted. The image annotation is determined based on the hypothesis test that a candidate annotation has occurred randomly in the list of candidate annotations.

In supervised multi-class labeling (SML) algorithm [15, 16], the distribution of feature points that are extracted from training images is approximated by a Gaussian mixture model (GMM) with expectation- maximization (EM) algorithm, then a multi-class Bayesian classifier is built by estimating the class-conditional distribution using a hierarchical structure of GMM from all training images associated with this class label.

The GMM estimation process in the previous methods, which uses a combination of C-means clustering [17] initialization and GMM parameter estimation through EM algorithm, works well for those image features with low dimensionality. However, for the high-dimensional feature cases, SML algorithm leads to expensive computation, and the mixture number is also not adaptively selected.

Affinity propagation (AP) clustering algorithm can be applied to identify a relatively small number of features called exemplars to represent the whole features [18, 19]. It seems able to produce a better fitness function than the mixture model with C-means methods, especially for high-dimensional data [20]. It takes the similarity measure of feature points as input parameters, therefore its computational complexity does not change as the data dimension changes (Fig. 16). However, the preference factor that affects the number of cluster need to be pre-assigned by user, which is difficult to be determined in practice. If we can develop an adaptive cluster number selection algorithm, it will bring the convenience to users.

In order to improve the accuracy and efficiency of image modeling algorithm, we intend to develop a novel method and expect to get a better result for image semantic annotation in this work.

The paper is organized as follows: In Sect. 2, we describe the framework of the image semantic annotation,

introduces AP algorithm, the image features and the feature distribution estimation, and the multi-class Bayesian classifier for image annotation. The proposed method is introduced in Sect. 3, including AP-based GMM estimation algorithm, training data optimization algorithm, and prior distribution modeling strategy. In Sect. 4, we illustrate the classifier performance improvement of the combined optimization method through the experiment results for image semantic annotation. The comparison for the proposed method versus SML is discussed in Sect. 5 and finally, the conclusion is presented in Sect. 6.

2 Methodology

The framework of the image semantic annotation is described as in Fig. 1. In this framework, low-level image features are extracted from sub-blocks of a given image, AP algorithm is applied to estimate the image feature distribution instead of a C-means and EM combination. Then a Bayesian classifier is built for image semantic annotation by the procedures of estimating class-conditional distribution and prior distribution. The class-conditional distribution among images in each class is modeled using a hierarchical structure of GMM [21]. A training data optimization algorithm is developed to improve the accuracy of the Bayesian classifier. In addition, the prior distribution is modeled using GMM with EM algorithm to raise the accuracy of image annotation.

2.1 AP algorithm

AP algorithm can be applied to identify a relatively small number of exemplars to represent the whole feature vectors. Each feature vector is viewed as a node in a network, and real-valued messages are recursively transmitted along edges of the network until a good set of exemplars and corresponding clusters emerges [18].

$$s(j, k) = - \parallel \mathbf{X}_j - \mathbf{X}_k \parallel^2, \tag{1}$$

$$r(j, k) \leftarrow s(j, k) - \max_{k' \neq k}\{a(j, k') + s(j, k')\}, \tag{2}$$

$$a(j, k) \leftarrow \min\left\{0, r(k, k) + \sum_{j' \neq j, j' \neq k} \max\{0, r(j', k)\}\right\}. \tag{3}$$

Where the similarity $s(j, k)$ indicates how well the node k is the exemplar of node j. The responsibility $r(j, k)$ and
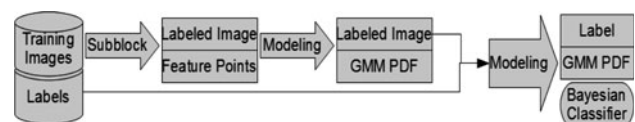

Fig. 1 The framework of the image semantic annotation

availability $a(j, k)$ are two kinds of messages exchanged between nodes.

Each message is set to λ times its value from the previous iteration plus $1 - \lambda$ times its prescribed updated value, where the damping factor λ usually is taken a default value 0.9.

The preference factor $s(k, k)$ of node k decides the degree that a node tends to select itself as the exemplar and affects the cluster number finally.

The selection of preference factor is a NP-complete (Non-determinstic Polynomial-complete) problem, usually the preference factor is pre-assigned a fixed value by users [19].

In the previous work, we found that there exists an approximately linear relationship between logarithm of the absolute value of the preference factor and logarithm of the cluster number (Fig. 2) [1]. Therefore, the preference factor can be approached with cluster number on consideration of this relationship in the AP algorithm.

$$\ln\left(|\text{pref}|\right) = a \ln\left(cn\right) + b, \tag{4}$$

$$\text{pref} = cn^a + e^b, \tag{5}$$

where the pref stands for the preference, and the cn expresses the cluster number.

2.2 Image feature extraction

MPEG-7 provides many useful low-level visual descriptors, including color layout descriptor (CLD) [22]. Rather than processing the whole image, we extract CLD from image sub-blocks.

In YCbCr color space, each image is divided into 8×8 sub-blocks, and the adjacent blocks overlap 1 pixel. For each sub-block, discrete cosine transform (DCT) is applied on each color channel, the coefficient values are quantized and zigzag scanned from left top to right down, because the information concentrates on the left top of the block. The 64 values are selected as the feature vector of this color

Fig. 2 Relationship between preference factor and cluster number

channel. Then the three vectors from Y, Cb, Cr color channels are concatenated as the 192-dimensional feature vector of the sub-block.

2.3 Image feature distribution estimation

Gaussian mixture model is an efficient representation for image feature distribution, the GMM estimation in the previous SML method takes a combination of clustering initialization with C-means and Gaussian mixture estimation through EM algorithm. For high-dimensional feature cases, it leads to expensive computation for SML algorithm.

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n | \mu_n, \Sigma_n)}{\sum_{j=1}^{K} \pi_j N(x_n | \mu_n, \Sigma_n)} \tag{6}$$

where $\gamma(z_{nk})$ evaluates the responsibilities associated with data x_n and class k using the current parameter values, in **E** step.

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \tag{7}$$

$$\Sigma_n^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(x_n - \mu_k^{\text{new}}\right) \left(x_n - \mu_k^{\text{new}}\right)^T \tag{8}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}, \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{9}$$

The parameters $\mu_k^{\text{new}}, \Sigma_n^{\text{new}}$ and π_k^{new} are re-estimated using the current responsibilities, in **M** step.

2.4 Hierarchical structure of GMM

A hierarchical structure of GMM was developed to estimate the parent GMM distribution from many children GMM distributions [15, 21]. The Gaussian mixture estimation through EM algorithm becomes more and more time-consuming, when the mixture number or the feature data point increases. Therefore, if there are too many feature points to be used for GMM estimation directly, we can divide the feature points into several groups, building the group-level GMMs first and finally, estimate a high-level GMM with group-level GMMs. Similarly, for the class-conditional distribution estimation, if there are too many image class to be deal with directly, we can take the strategy in the same manner.

Assuming that for each image j in class i, the GMM of image feature distribution has K components; this leads to a class-conditional distribution of D_i K components with parameters

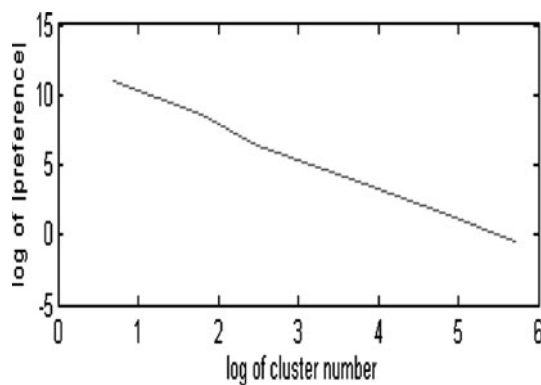$$\left\{ \pi_j^k, \mu_j^k, \Sigma_j^k \right\}, \quad j = 1, \ldots, D_i, \quad k = 1, \ldots, K \tag{10}$$

The class-conditional distribution is approximated with an M-component Gaussian mixture.

$$\left\{\pi_c^m, \mu_c^m, \Sigma_c^m\right\}, \quad m = 1, \ldots, M \tag{11}$$

$$h_{jk}^m = \frac{\left[N\left(\mu_j^k \middle| \mu_c^m, \Sigma_c^m\right) e^{-\frac{1}{2}\text{trace}\left\{\left(\Sigma_c^m\right)^{-1}\Sigma_j^k\right\}}\right]^{\pi_j^k} \pi_c^m}{\sum_l \left[N\left(\mu_j^k \middle| \mu_c^l, \Sigma_c^l\right) e^{-\frac{1}{2}\text{trace}\left\{\left(\Sigma_c^l\right)^{-1}\Sigma_j^k\right\}}\right]^{\pi_j^k} \pi_c^l} \tag{12}$$

where h_{jk}^m evaluates the responsibilities using the current parameter values, in E step.

$$\left(\mu_c^m\right)^{\text{new}} = \sum_{jk} \omega_{jk}^m \mu_j^k, \quad \omega_{jk}^m = \frac{h_{jk}^m \pi_j^k}{\sum_{jk} h_{jk}^m \pi_j^k} \tag{13}$$

$$\left(\Sigma_c^m\right)^{\text{new}} = \sum_{jk} \omega_{jk}^m \left[\Sigma_j^k + \left(\mu_j^k - \mu_c^m\right)\left(\mu_j^k - \mu_c^m\right)^T\right] \tag{14}$$

$$\left(\pi_c^m\right)^{\text{new}} = \frac{\sum_{jk} h_{jk}^m}{D_i K} \tag{15}$$

The parameters $\left(\mu_c^m\right)^{\text{new}}$, $\left(\Sigma_c^m\right)^{\text{new}}$ and $\left(\pi_c^m\right)^{\text{new}}$ are re-estimated using the current responsibilities, in M step.

2.5 Multi-class Bayesian classifier training

Bayesian classifier is adopted for image semantic annotation, which classifies image features into class labels. When training a Bayesian classifier, the training data selection is a challenging problem, not only image feature selection, but also the class label encoding and grouping problem need to be considered [23].

For image semantic annotation, the classifier input data are those feature points extracted from training images, and these training images are manually pre-annotated with class labels. It is noted that the class label is assigned as the output of the classifier. In the training image dataset, it is manually annotated with each image. For a test input image, the output class is labeled by the Bayesian decision rule.

2.5.1 Encoding of class label

The class labels of an image can be represented as a binary vector in which either 1 or 0 of each bit indicates whether or not correspondent label exists in the image, as illustrated in Fig. 3.

We name images those have only one class label as the single-class images, and images those have more than one class labels as the multiple-class images, or images with multiple-class labels. It is noted that each feature point is only considered to belong to a single-class. For example, all images in Fig. 3 are multiple-class images.

2.5.2 Selection of training data

SML algorithm selects the whole feature points that are associated with a given single-class label as the training data of this class. Then all the training data are used to train the Bayesian classifier, where class-conditional distributions of feature points are learned from image feature distributions using a hierarchical structure of GMM. Large training dataset is usually divided into subsets. The hierarchical structure of GMM is adopted to simplify the calculation, in which subsets of training data are used to estimate GMM distributions, then an overall GMM distribution is approached with these GMM distributions of subsets.

Specially, the image feature distribution is on the bottom level of the hierarchical structure. Because the class labels are associated with each image, the feature points of one image usually are not divided anymore.

2.5.3 Hierarchical selection of multiple-class label

Instead of selecting feature points from images of a given class as the training data, we use the hierarchical selection of input training data according to multi-class label. We select those images with same multiple-class to consist of a group, then select corresponding groups to estimate the class-conditional distribution of the given class. As illustrated in Fig. 3, the whole images are divided into three groups with multiple-class label **ABC**, **AB** and **AC**, respectively. Taking class **C** as an example, those groups which contain label **C**, such as multiple-class label **ABC** and **AC**, are selected to estimate the class-conditional distribution of class **C**.

Hierarchical selection of training data group with multiple-class label can also be used for Bayesian classifier training as well as the selection strategy in SML method. For example, in Fig. 3, the class-conditional distribution of class **C** can be estimated from the distributions of multiple-class **ABC** and **AC**. Meanwhile, the image set 1, 2, 3, 8, 9

Fig. 3 Three labels **A**, **B**, and **C**, and images represented using one-dimension feature points

	Image 1	Image 2	Image 3	Image 4	Image 5	Image 6	Image 7	Image 8	Image 9
Label A	◯₁	◯₁	◯₁	◯₁	◯₁	◯₁	◯₁	◯₁	◯₁
Label B	◯₁	◯₁	◯₁	◯₁	◯₁	◯₁	◯₁	0	0
Label C	◯₁	◯₁	◯₁	0	0	0	0	◯₁	◯₁

can also be divided into two subsets, 1, 2, 3, and 8, 9, where the distribution of multiple-class **ABC** is estimated from images 1, 2, 3, and the distribution of multiple-class **AC** is estimated from images 8, 9.

In SML method, it is assumed that training dataset is large enough. When estimating the Bayesian classifier, the negative samples present in positive bags tend to spread all over the feature space, and the positive samples are much more likely to be concentrated within a small region [15]. Therefore, the negative samples have a small weight in contribution to the estimation of class-conditional distribution, and approximately obey uniform distribution as illustrated in Fig. 4a.

A simple training data selection method is applied in the SML algorithm, which only considers the distribution of single-label case. This may lead to a low accuracy for the Bayesian classifier. As we know, there are two kinds of feature points in the images associated with a given class, feature points of the given class and feature points of other classes. And it can be explained with following figures. In Fig. 3, images 1, 2, 3, 8, 9 are used to estimate the class-conditional distribution of class **C** in SML algorithm. However, the feature points belonging to class **A** and **B** in these images are also densely distributed in feature space. These points have a large weight in the class-conditional distribution of class **C**. For the small dataset as shown in Fig. 3, the class-conditional distribution of class **C** is illustrated in Fig. 4b. Obviously, SML method does not work well for a small dataset case.

In fact, considering the distribution of images with multiple-class label is more reasonable than only considering the distribution of single-class case.

The distribution estimation of multiple-class has the advantage that it is not affected by the feature points of other classes, because there is no feature points of other classes in each group of a multiple-class. For example, in Fig. 3, the distribution of multiple-class **AC** is estimated from images 8, 9, and there are no feature points belonging to class **B** in images 8, 9.

For the Bayesian classifier in the small dataset case, the hierarchical selection of training data on considering multiple-class label has an advantage over that of SML algorithm. Therefore, in Sect. 3.3, we consider to use this training data optimization method to improve the accuracy of the Bayesian classifier.

2.6 Image annotation

The image annotation can be implemented by the Bayesian decision rule. For a given class, the probability that an image to be annotated belongs to this class is the product of the probabilities that the image features belong to this class.

$$\lg(P(\mathbf{I}|s_i)) = \sum_{\mathbf{X} \in \mathbf{I}} \lg(P(\mathbf{X}|s_i)) \tag{16}$$

By introducing a set of class-conditional distributions, the semantic annotation results for this image can be obtained with the labels whose posterior probabilities $P_{s|\mathbf{I}}(s_i|\mathbf{I})$ are the top several large values.

3 Image annotation analysis

3.1 Feature distribution estimation using AP-Normfit

Mixture model parameters can be estimated in the maximum likelihood manner [24], and we can apply AP algorithm to cluster image feature points into several groups. The mean and covariance matrix parameters of each component model are estimated with corresponding feature points of each group, then each normal distribution model of GMM is obtained with maximum likelihood estimate formula. This method is called as AP-Normfit.

$$\{\mathbf{E}_c\} = \mathrm{AP}(\{\mathbf{X}\}, p, \lambda), \quad c = 1, \ldots, C. \tag{17}$$

$$\mu_c = \mathbf{E}_c, \quad \Sigma_c = \mathrm{cov}(\{\mathbf{X}_{\mathbf{E}_c}\}), \quad \omega_c = \mathrm{num}(\{\mathbf{X}_{\mathbf{E}_c}\}), \tag{18}$$

$$P(\mathbf{X}|\mathbf{I}) = \sum_{c=1}^{C} \omega_c\, G(\mu_c, \Sigma_c), \tag{19}$$

where $\{\mathbf{E}_c\}$ are the exemplars extracted from feature points $\{\mathbf{X}\}$, and $\{\mathbf{X}_{\mathbf{E}c}\}$ are the feature points whose exemplar is $\mathbf{E}_c$. Then, the image feature distribution $P(\mathbf{X}|\mathbf{I})$ of image $\mathbf{I}$ is approximated with GMM.

We find that the results obtained with time-saving AP-Normfit method are approximate same with those results obtained with the time-consuming EM algorithm (Sects. 4.3, 4.4). In addition, AP algorithm is not sensitive to initial value of exemplars, while C-means algorithm strongly relies on initialization of mean values in order to get the better clustering results.

3.2 Cluster number selection algorithm

In AP algorithm, the preference factor that affects the number of cluster need to be pre-assigned by user, which is difficult to be determined in practice [25].
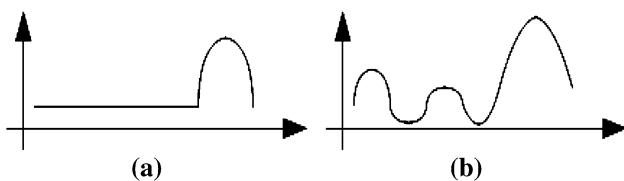

Fig. 4 Comparison of class-conditional distributions of class **C** with large and small datasets. **a** For large dataset. **b** For small dataset

Fig. 5 The framework of semantic-based cluster number selection algorithm

The cluster number can be inferred from the semantic information of the image. We tried to find out the relationship between cluster numbers and semantic labels from a subset of training images, as illustrated in Fig. 5.

The feature points associated with same semantic label are assumed to have similar distributions in different images; therefore, the semantic label is considered to produce same cluster number in feature space of different images. The cluster number of each image can be estimated using AP algorithm. Then, if cluster number of each label is viewed as a variable, these variables can form a linear equation for each image.

For each image, the real cluster number is estimated using C-means + EM algorithm with different cluster number. According to the minimal sum of distance within clusters, the clustering result with the minimal sum is selected, and it is chosen as the estimated cluster number of this image.

$$d = \sum_{c=1,\ldots,C} \sum_{x_i \in c, x_j \in c} \|x_i, x_j\|, \qquad (20)$$

where d is the sum of distance within clusters, when the cluster number increases, d will approach a stable value (Fig. 6).

A small set of typical training images is selected, then we get a system of linear equations from these images. The cluster number of each label can be estimated by solving the system of linear equations.

$$cn(\mathbf{I}) = \sum_{s_i \in \text{labels}(\mathbf{I})} cn(s_i), \qquad (21)$$

where $cn(\mathbf{I})$ is the cluster number of image $\mathbf{I}$, $cn(s)$ is the cluster number of the semantic label s, and labels($\mathbf{I}$) are the semantic labels of image $\mathbf{I}$.

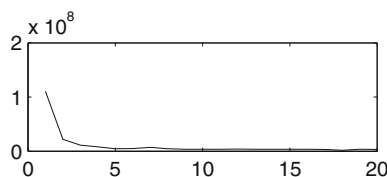For other training images, the cluster numbers can be simply estimated according to their semantic labels.

Fig. 6 Selecting best cluster number using sum of distance within clusters

Table 1 Label number and cluster number

Label number	1	2	3	4	5
Cluster number	3	4	5	6	6

For VOC dataset [26], we found that the best cluster number is five for almost all class labels. The mapping table of cluster number and label number is as Table 1.

3.3 Class-conditional distribution estimation

As mentioned in Sect. 2.5, SML adopts a simple training data selection method, which only considers the distribution of single-label case.

There are two kinds of feature points in the training images of a given class, feature points of this class and feature points of other classes. The hierarchical selection of training data by multiple-class has the advantage that it is not affected by feature points of other classes.

Using the hierarchical selection of training data, we introduce a training data optimization algorithm to improve the accuracy of the Bayesian classifier (Fig. 7).

In the proposed method, additional feature points are added to increase the feature points of the given class. Therefore, there are three kinds of feature points: original feature points of this class, original feature points of other classes, and added feature points of this class.

The method can be described as follows (Fig. 7):

1. Hierarchical selection of training data of multiple-class is applied, and the distributions of each multiple-class are estimated.
2. For the training data of a given class, we do not decrease the feature points of other classes, but select points to increase the feature points of this class.
3. To improve the generalization ability, we do not add the selected feature points directly but generate random feature points around them.
4. The distribution of the generated feature points is modeled with the GMM. And the class-conditional distribution is computed based on both distribution of multiple-class and distribution of random generated feature points.

3.3.1 Feature point selection

In order to select feature points of the given class from each training image, two Bayesian classifiers are built to

Fig. 7 The framework of the training data optimization algorithm

classify the whole feature points in this image using two distributions of multiple-class. The first multiple-class label contains the given single-class label, and classifier uses distribution of the first multiple-class to classify the whole feature points into two categories: feature points of the first multiple-class and feature points of other classes. The second multiple-class label contains the same labels of the first multiple-class, except the given single-class label. And the second classifier uses distribution of the second multiple-class to classify the feature points of the first multiple-class into two categories: feature points of the second multiple-class, and feature points of the given class.

Taking in Fig. 8 as example to explain this method, we illustrate the class-conditional distribution estimation of class **C**. Feature points of class **C** will be selected from each training image of class **C**, which are images 1, 2, 3, 8, 9. In image 1, two multiple-class labels **ABC** and **AB** are selected to build the classifiers. The first multiple-class label **ABC** contains class **C**. By Bayesian decision rule, the first classifier of the multiple-class label **ABC** is built, and it is used to classify the whole feature points of image 1 into two categories, feature points of **A**, **B**, **C**, and feature points of other classes. The second multiple-class label **AB** contains the same labels of the first multiple-class, except the given class **C**. The second classifier of the multiple-class label **AB** is built in the same manner, which is used to classify the feature points of **A**, **B**, **C** into two categories, feature points of **A**, **B** and feature points of **C**. For same feature point **X**, this two-class Bayesian classifier can use the class-conditional distribution of multiple-class to replace the posterior distribution of the multiple-class.

For the first classifier, feature points of **A**, **B**, **C** are selected when the first distribution of multiple-class **A**, **B**, **C** is larger than τ_{76}; for the second classifier, feature points of **C** are selected when the second distribution of multiple-class **A**, **B** is smaller than τ_{67}.

We define two thresholds and combine the classifiers together as follows.

$$\mathbf{X}|[0\,0\,1]_+ = \{\mathbf{X}|P(\mathbf{X}|[1\,1\,1]) > \tau_{76}$$
$$> \tau_{67} > P(\mathbf{X}|[1\,1\,0]) \quad (22)$$
$$\mathbf{X} \in \mathbf{I}, \quad \text{labels}(\mathbf{I}) = [1\,1\,1]\}$$

where τ_{76} and τ_{67} are thresholds, which are small values approximating 0.

The combination of the above pair of classifiers is adopted to select feature points for class **C** from all training images, 1, 2, 3, 8, 9.

It is noted that two multiple-class labels **A B C** and **A B**, whose binary vectors are [1 1 1] and [1 1 0], have a Hamming distance of 1, and the different position of them is class **C**. It seems that any pair of multiple-class labels whose Hamming distance is 1 and the difference is class **C**, can be used to select feature points for class **C**.

In Fig. 3, there is only one pair of multiple-class labels whose difference is class **C**. If there are images with single-class [1 0 0], then there is another pair of multiple-class labels whose difference is class **C**, as described in formula (23).

$$\mathbf{X}|[0\,0\,1]_{2+} = \{\mathbf{X}|P(\mathbf{X}|[1\,0\,1]) > \tau_{54}$$
$$> \tau_{45} > P(\mathbf{X}|[1\,0\,0]),$$
$$\mathbf{X} \in \mathbf{I}, \quad \text{labels}(\mathbf{I}) = [1\,0\,1]\} \quad (23)$$

All the selected feature points are added as described in formula (24).

$$P(\mathbf{X}|[0\,0\,1]) = f_H(P(\mathbf{X}|[1\,0\,1]), P(\mathbf{X}|[1\,1\,1]),$$
$$P(\mathbf{X}'|[0\,0\,1]_+), P(\mathbf{X}'|[0\,0\,1]_{2+})), \quad (24)$$

where f_H is the hierarchical algorithm, $\mathbf{X}'$ represents the generated feature points, $P(\mathbf{X}'|[0\,0\,1]_+)$ and $P(\mathbf{X}'|[0\,0\,1]_{2+})$ are the distribution of generated feature points.

Figure 9 provides an example of selecting feature points. The feature points associated with different labels are selected separately. Figure 9b–d illustrate the feature points belonging to class 1, 3 and 5, which are sky, plant, and other objects, respectively.

3.3.2 Feature point generation

As mentioned above, we do not add the selected feature points to original data set directly but generate artificial feature points around them randomly.

3.3.3 Class-conditional distribution estimation

The distribution of the randomly generated feature points is modeled by GMM distributions. Usually feature points selected from each pair of multiple-class labels are modeled with single Gaussian model distribution.

And the class-conditional distribution is estimated based on both the distributions of multiple-class and the distributions of random generated feature points using the hierarchical structure of GMM.

3.4 Modeling prior distribution with GMM

In order to improve the accuracy of image annotation, the prior distribution is modeled using GMM (Fig. 10).

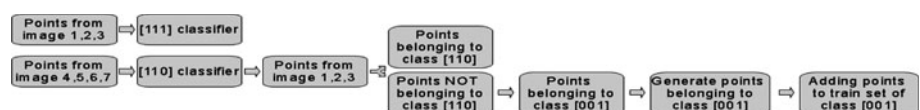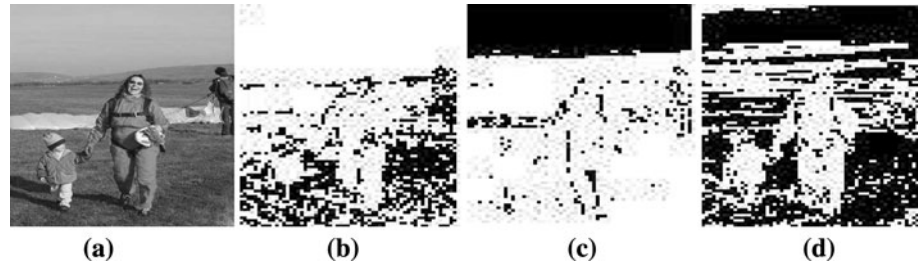**Fig. 8** Selected feature points to be generated

Fig. 9 Feature points are selected to improve the accuracy of Bayesian classifier. **a** Original image. **b** Points of 1st label "sky". **c** Points of 3rd label "plant". **d** Points of 5th label "others"

(a)   **(b)**   **(c)**   **(d)**

For prior distribution modeling problem, traditional an approximate method is used, which only considers the number of images with the given labels (formula 25).

$P_W(w_i)$ is computed from the training set as the proportion of images containing annotation w

$$P(s_i) = \text{num}(\mathbf{I}_{s_i}). \tag{25}$$

It may be more reasonable if the prior distribution of different classes is determined by incorporating the information from the distribution of image feature points.

During the iteration process of computing the prior distribution, for all classes s_i:

$$L_{s_i}(\mathbf{X}) = P(s_i|\mathbf{X}) = \frac{P(\mathbf{X}|s_i)P(s_i)}{P(\mathbf{X})}. \tag{26}$$

where L_{s_i} is the evidence that the feature point $\mathbf{X}$ belongs to class label s_i. The old value of prior distribution $P(s_i)$ is taken from the previous iteration.

Then for each image, the prior distribution $P(L_{s_i}|\mathbf{I})$ is estimated using the sum of evidence from each feature point.

$$P(L_{s_i}|\mathbf{I}) = \sum_{\mathbf{X} \in \mathbf{I}} L_{s_i}(\mathbf{X}). \tag{27}$$

For the whole image dataset, the overall GMM distribution $P(\mathbf{X}|s_i)$ of each class s_i and $P(\mathbf{X})$ for all training data are approached using the hierarchical structure of GMM, as illustrated in the previous sections.

Then the prior probability of each class is estimated using the estimation value from each image.

$$P'(s_i) = \sum P(L_{s_i}|\mathbf{I}), \tag{28}$$

$$P(s_i) = \frac{P'(s_i)}{\sum P'(s_i)}. \tag{29}$$

The above steps are iteratively performed until the difference of the prior probabilities between two adjacent iterations is less than a predefined threshold for each class.

4 Experiments

In this section, we conduct experiment studies extensively to illustrate the classifier performance improvement of image modeling and image semantic annotation.

4.1 Dataset

This work is focused on the image semantic annotation for the case of multiple objects, and small sample number with high-dimension setting. The image semantic annotation is comparatively easy to extend to the case of large sample number. And some strategies such as hierarchical classification can be adopted when processing a dataset with more than five labels (Fig. 11).

The images are selected from database [26]. We selected a subset of 276 images which contain five labels: sky, land, plant, water, and other objects.

The image numbers of these five classes are not balanced. Figure 12a illustrates the number of images associated with two classes. Figure 12b illustrates the number of images associated with multiple-class labels, and the binary vectors of the multiple-class label are converted to decimal numbers.

4.2 Experiment procedure

The algorithm is briefly described as follows:

1. Divide image into sub-blocks and extract image features.
2. Select a small set of typical images and estimate the cluster number of each label using cluster number selection algorithm.
3. Estimate the relationship between cluster number and AP parameter [1].
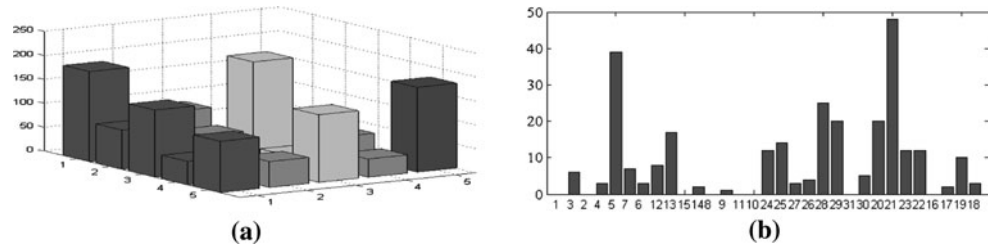4. Estimate the feature distribution using AP-Normfit algorithm.

Fig. 10 The framework of the prior distribution modeling strategy

Fig. 11 A few images in dataset

Fig. 12 Dataset and class labels. **a** The number of images associated with two classes. **b** The number of images associated with multiple-class label

5. Build the training dataset of each label using the training data optimization algorithm.

6. Estimate the class-conditional distribution using the training dataset of each label.

7. Classify each sub-block of images using Bayesian classifier.

8. Optimize the prior distribution iteratively and re-train the Bayesian classifier.

9. Classify testing images using Bayesian classifier.

The experiment procedure is described as follows:

1. The half of dataset is for training, and the other half is for testing. Six different divisions of dataset are selected.

2. For feature clustering of all images, sum of distance within clusters is computed, using sixteen different initializations.

3. For feature distribution estimation of all images, sum of distance between histograms of estimated and original distributions is computed, using sixteen different initializations. And the average time consumption is recorded, when different max iteration numbers are limited.

4. For class-conditional distribution estimation of each class, sum of distance between histograms of estimated and original distributions is computed.

5. For image annotation of each class, recall and precise factors are computed, averaging from all six divisions of datasets.

The algorithm is described in pseudo code as in Algorithms 1 and 2.

4.3 Evaluating features and clustering algorithms

The results of C-means clustering using pixel-based 3-Dimension feature (Fig. 13b) and block-based 18-Dimension feature (Fig. 13c) are compared. After clustering, each feature point is assigned with a class label. The class label of all feature points is expressed as a gray image. It seems that the block-based features are better than pixel-based features. For example, the block-based features have better results on the regions of sky and ground.

Based on block-based features, C-means algorithm (Fig. 13c), hierarchical clustering algorithm [27] (Fig. 13d) and AP algorithm (Fig. 13e) are compared. In Fig. 14, using normalized sum of distance within clusters, results of evaluating these three algorithms are shown. The quantitative results illustrate that AP algorithm has better clustering result than other two algorithms. It is known that clustering results of C-means algorithm relies on the initial value; therefore, average value of multi-runs should be used for C-means algorithm.

4.4 Evaluating GMM estimation algorithms

Figure 16a shows the results of evaluating three EM-based methods and one non-EM method: C-means and EM combination, Hierarchical clustering and EM combination, AP and EM combination, and AP-Normfit algorithms, respectively. AP-Normfit algorithm is a combination of AP algorithm and normal distribution fitting (AP-Normfit).

Algorithm 1 Cluster number selection and AP-based GMM estimation

---

Input: Training images $\mathbf{I}_i$, $i = 1 \dots N$,
image labels $s_k \in \mathrm{labels}(\mathbf{I}_i)$, $k \in [1\dots K]$,
and test image $\mathbf{I}_0$.

Output labels($\mathbf{I}_0$)

1 Select a small set of training images ($\mathbf{I}_j$, $j = 1\dots N_1$) to estimate the best cluster
  number of each label.

2 $[CN_{\min} \dots CN_{\max}]$ is the sequence of possible cluster number, for example [2... 20].

3 **for** $j = 1$; $j \leq N_1$; $j + +$ **do**

4   **for** $cn = CN_{\min}$; $cn \leq CN_{\max}$; $cn + +$ **do**

5      $CMeansEM(\mathbf{I}_j, cn)$;

6   d_{cn} is the sum of distance within clusters.

7   **end**

8   Select the best cn according to d_{cn} (as Fig. 6).

9   Form a linear equation: $cn(\mathbf{I}) = \sum\limits_{s_k \in \mathrm{labels}(\mathbf{I})} cn(s_k)$.

10 **end**

11 Solve the above system of linear equations to find the best $cn(s_k)$.

12 Select a small set of training images ($\mathbf{I}_j$, $j = 1\dots N_2$) to estimate the relationship of
  cluster number and AP parameter p [1].

13 **for** $j = 1$; $j \leq N_2$; $j + +$ **do**

14    $(p_{\min}, p_{\max})$ is the range of p for image $\mathbf{I}_j$.

15   Select several values p from $(p_{\min}, p_{\max})$.

16   Compute the result cluster number by:

17    $AP(\mathbf{I}_j, p)$;

18 **end**

19 Fit the relationship $p = map\ (cn)$.

20 Use the whole training images ($\mathbf{I}_i$, $i = 1\dots N$) to estimate the feature distribution.

21 **for** $i = 1$; $i \leq N$; $i + +$ **do**

22   Compute the best cluster number for image $\mathbf{I}_i$: $cn(\mathbf{I}) = \sum\limits_{s_k \in \mathrm{labels}(\mathbf{I})} cn(s_k)$.

23   Compute the AP parameter: $p = map\ (cn)$.

24   Compute the feature distribution:

25    $GMM(\mathbf{I}) = APNormfit(\mathbf{I}, p)$.

26 **end**

In order to compare the results of the distribution estimators, the distance between estimated distribution and the original distribution is studied. In the studies, Same amount of feature points are randomly generated from these estimated distributions, and the dimension of the original and simulated feature points are reduced to 2-D simultaneously. Under this situation, the 2-D histograms are compared (Fig. 15).

The quantitative comparative results are shown in Fig. 16a. The AP-EM algorithm achieves the best performance, and the AP-Normfit algorithm is the second, but the advantage of AP-EM over AP-Normfit is not very obvious.

Compared these algorithms in time performance (Fig. 16b), the time consumption of EM-based algorithm increases greatly when the feature dimension increases. Because AP algorithm processes the similarity of feature data instead of feature data themselves, the time consumption of AP-based method does not change when the feature dimension increases. Therefore, the proposed AP-based method is more fit for high-dimension feature dimension than EM-based method.

We can find that even without incorporating EM algorithm, AP algorithm still gives a good image feature distribution estimation in the cost of less computation time in contrast to other two algorithms.

4.5 Evaluating the class-conditional distribution estimation algorithms

As we known, the selection of training data is a challenging problem for images with multiple-class labels. The SML algorithm adopts a simple training data selection method, which only considers the distribution of single-class case, and assumed that the feature points belonging to other classes tend to be uniformly and sparsely distributed in feature space, which seems not suitable to the small dataset situation, so the training data optimization algorithm is

---

Algorithm 2 Training data optimization and prior distribution modeling

---

1 Each image belongs to multiple-class $[s_1...s_K]$ ($s_k \in \{0, 1\}$, $k = 1...K$), select those $[s_1... s_K]$ that are annotated to enough many images, KK is the number of selected multiple-classes.

2 Compute class-conditional distribution for each selected multiple-class $[s_1 ...s_K]$.

3 **for** $kk = 1$; $kk \leq KK$; $kk + +$ **do**

4   Select the training images ($\{\mathbf{I}_m\}$) that belongs to multiple-class $[s_1 ... s_K]$,

5   Compute the class-conditional distribution using Hierarchical GMM [21]:

6 $GMM([s_1...s_K]) = HierGMM(\{GMM(\mathbf{I}_m)\})$.

7 end

8 Training dataset $dataset(s_k)$ is initialized with feature points $\mathbf{X}$ from images with label s_k.

9 Optimize the training dataset for each label s_k.

10 for $k = 1$; $k \leq K$; $k + +$ **do**

11   Select the training images ($\mathbf{I}_m$), ($\mathbf{I}_n$), that ($\mathbf{I}_n$) doesn't contain label s_k and ($\mathbf{I}_m$) contains all labels of ($\mathbf{I}_n$) and label s_k.

12   Select optimized training data for label s_k from ($\mathbf{I}_m$), ($\mathbf{I}_n$) and corresponding $GMM([s_1 ...s_K])$ (as Fig. 9).

13   Feature points generation randomly.

14   Add these feature points to training dataset $dataset(s_k)$.

15 end

16 Select a small set of training images ($\mathbf{I}_j$, $j = 1...N_3$) to estimate the prior distribution $P(s_k)$ for each label s_k, $k = 1...K$.

17 Set the initial value $P(s_k) = num(\mathbf{I}_{s_k})$.

18 while $|P(s_k)_{\text{new}} - P(s_k)| \leq threshold$ **do**

19   Train the Bayesian classifier for each label s_k, to classify each feature point $\mathbf{X}$ in each training image $\mathbf{I}$.

20   **for** $k = 1$; $k \leq K$; $k + +$ **do**

21 $GMM(s_k) = HierGMM(\{GMM(\mathbf{I}_j)\})$.

22 $P(\mathbf{X}|s_k) = GMM(s_k)$.

23 $L_{s_k}(\mathbf{X}) = P(s_k|\mathbf{X}) = \frac{P(\mathbf{X}|s_k)P(s_k)}{P(\mathbf{X})}$.

24 $P(L_{s_k}|\mathbf{I}) = \sum_{\mathbf{X}\in\mathbf{I}} L_{s_k}(\mathbf{X})$.

25 $P'(s_k) = \sum P(L_{s_k}|\mathbf{I})$.

26   **end**

27 $P(s_k)_{\text{new}} = \frac{P'(s_k)}{\sum P'(s_k)}$

28 end

29 Use the whole training images ($\mathbf{I}_i$, $i = 1...N$) to train the Bayesian classifier for each label s_k.

30 for $k = 1$; $k \leq K$; $k + +$ **do**

31   Select feature point $\mathbf{X}$ from optimized training dataset $dataset(s_k)$.

32 $P(\mathbf{X}|s_k) = GMM(\mathbf{X})$.

33 end

34 For test image $\mathbf{I}_0$,

35 for $k = 1$; $k \leq K$; $k + +$ **do**

36 $\lg(P(\mathbf{I}_0|s_k)) = \sum_{\mathbf{X}\in\mathbf{I}} \lg(P(\mathbf{X}|s_k))$.

37 $P(s_k|\mathbf{I}_0) = \frac{P(\mathbf{I}_0|s_k)P(s_k)}{P(\mathbf{I}_0)}$.

38 end

39 labels($\mathbf{I}_0$) = $\{s_k\}$, the corresponding $P(s_k|\mathbf{I}_0)$ are the top several large values.
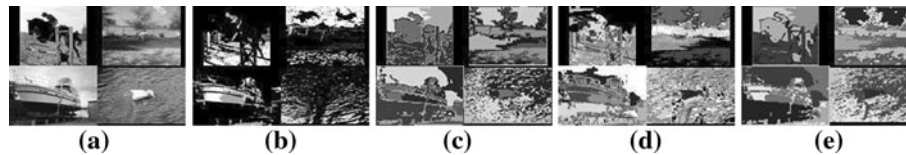

Fig. 13 Comparison of pixel-based and block-based features (**b** vs. **c**); Comparison of C-means, hierarchical clustering and AP algorithms (**c–e**). **a** is the original image, **b–e** are the gray images changed from clustering results. **a** RGB, Image; **b** YCbCr, C-means; **c** Block, C-means; **d** Block, Hier; **e** Block, AP

developed to improve the accuracy of the Bayesian classifier.

For each class label, the class-conditional distributions estimated with training data optimization algorithm are compared to those estimated with SML algorithm based on the 2-D histogram of original and estimated distributions. The proposed algorithm is better than SML algorithm in approximating the original distribution (Fig. 17).

From Fig. 16a, it can be seen that the image feature distribution has a better approximation to the original distribution than that of class-conditional distribution.

4.6 Annotation result analysis

For a given semantic class, we assumed that there are w_H human annotated images in the test set and the system automatic annotates number is w_{Auto}, of which w_C are correct, the recall and precision are defined as following:

$$\text{recall} = \frac{w_C}{w_H}, \tag{30}$$

$$\text{precise} = \frac{w_C}{w_{Auto}} \tag{31}$$

When compared with those of SML algorithm, the proposed method can further improve the accuracy of both recall and precision factors (Fig. 18).

The recall values for the 4-th class are rather high, which illustrates that for most images the 4-th class is annotated correctly. While the precision values for the 4-th class are rather low, which illustrates that too many images are annotated with the 4-th class. In other words, the 4-th class is not properly defined.

5 Discussions

In this section, the proposed method is discussed with other methods on considering time complexity and robustness under the situation of small dataset.

For image feature distribution estimation, AP-Normfit algorithm can generate reasonable clustering results than other clustering algorithms (Fig. 16a). AP algorithm considers the dynamic relationship between each pair of feature points during clustering, while the hierarchical
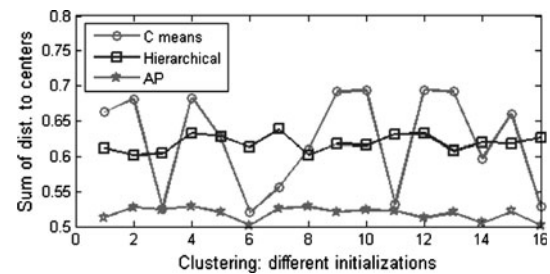

Fig. 14 Evaluating clustering results using the criterion of sum of distance within clusters

clustering algorithm seeks to build a hierarchy of clusters, and the results of C-means algorithm are related with the initial selection of cluster centers (Fig. 16a).

If we combine the AP initialization and EM algorithm (AP-EM), it can further improve the image feature distribution estimation (Fig. 16a). But the time complexity will increase much with compared AP-Normfit algorithm only, as illustrated in Fig. 16b–e.

The proposed method can improve the accuracy of image annotation, especially for the situation of small dataset, compared with other methods.

The ALIP method, SML method, and the proposed method are all applied under the framework of supervised classification. For SML method and the proposed method, the image feature distribution estimation of each image can be viewed as an unsupervised process, and the class-conditional distribution estimation of each class can be viewed as a supervised process. The main advantage of the proposed method is that it applies a novel training data optimization algorithm for image modeling.

In SML algorithm, the distribution of single-class label case is mainly concerned; it assumes that the feature points of other classes tend to be uniformly and sparsely distributed in feature space, which seems work for large image dataset. For a small dataset, it maybe not accurate in image modeling if ignoring the influence of those feature points of other classes.

Not only for small dataset, but also for dataset with unbalanced class labels, the SML method might not work well. As illustrated in Fig. 12, the number of images associated with each class label are not balanced, and the number of class labels associated with each image are not
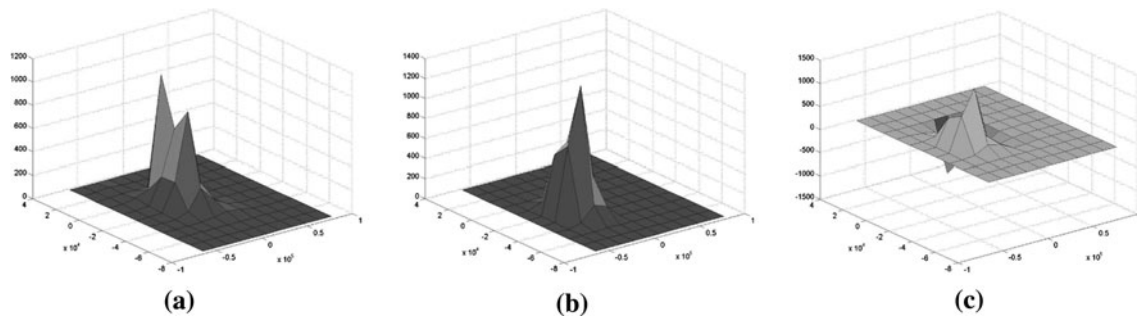
Fig. 15 Comparison of approximation level of the estimated distribution and the original distribution. **a** 2-D histogram of points from original distribution. **b** 2-D histogram of points from estimated distribution. **c** Difference of the above two histograms
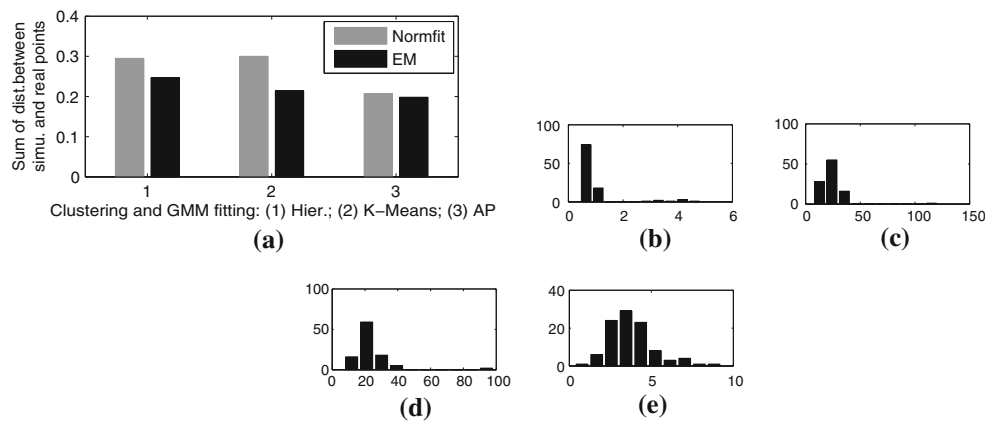

Fig. 16 Evaluating GMM estimation algorithms using the criterion of sum of distance between original and estimated distributions and the time performance. **a** Evaluates the estimation accuracy. **b–d** Illustrates the histogram of time consumption for EM-based algorithms, when the data dimension increases, the time consumption increases greatly. **e** Illustrate that for AP-based algorithm, because AP takes the similarity matrix as input parameter, it doesn't influenced by the data dimension. **a** Approximation level; **b** time of EM dim = 30; **c** time of EM dim = 96; **d** time of EM dim = 192; **e** time of AP any dim

balanced too. This can be explained with following examples: the number of images of multiple-class 1 and 2 is approximately equal to the number of images with single-class 2. In SML algorithm, we have to use the images with multiple-class 1 and 2 and the images with single-class 2 to estimate the class-conditional distribution of class 2. As illustrated in formula (24), it is the sets of feature points $\mathbf{X}|[1\ 0\ 1]$ and $\mathbf{X}|[1\ 1\ 1]$ that are used to estimate the class-conditional distribution. For unbalanced dataset, SML algorithm does not consider the situation that the positive information might be not enough compared with the negative information, therefore, does not make full use of the positive and negative information for image modeling.

A major assumption in SML method is that while the negative samples present in positive bags tend to spread all over the feature space, the positive samples are much more likely to be concentrated within a small region. However, for small dataset, the negative samples might not spread all over the feature space, they might also be concentrated within a small region.
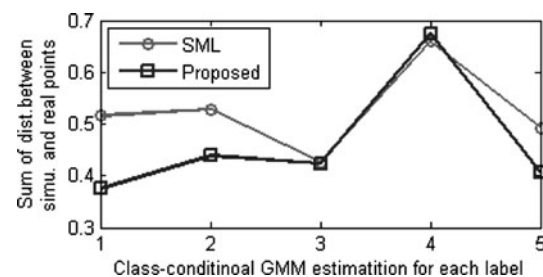

Fig. 17 Evaluating class-conditional GMM estimation algorithms using the criterion of sum of distance between original and estimated distributions

In ALIP method [14], the top several classified categories to which the test image is most likely to belong are selected, and the annotations from these categories are pooled into a list of candidate annotations. The frequency of each candidate annotation is counted. The candidate annotations are then ordered based on the hypothesis test that a candidate annotation has occurred randomly in the list of candidate annotations. A low probability that the

Fig. 18 a Recall and **b** precise
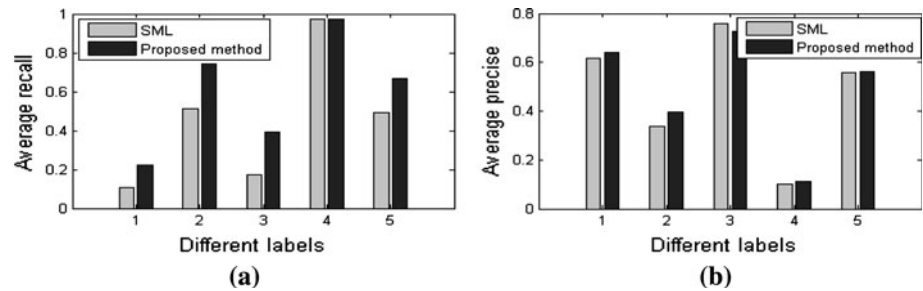
candidate word occurred randomly means the word has high significance as an annotation.

Although the ALIP method considers the information of multiple-class label, it uses the image category to represent a major concept. Each category is represented by a set of words that characterize the category as a whole but may not accurately characterize each individual image. For small dataset, the category classifier that is built using these categories may not accurately characterize these major concepts, which might decrease the accuracy of the image annotation results.

In fact, considering the information of multiple-class label is more reasonable than only considering the case of single-class. The proposed method tries to model distribution adopting all the existent information by grouping information with same multiple-class labels together. Based on the hierarchical selection algorithm of training data, the proposed method uses multi-class label information to generate extra feature points referring the original data point of the given class to improve the estimation accuracy of the class-conditional distribution (Fig. 17).

When using the training data optimization algorithm, the proposed algorithm can improves the robustness of the classifier. All feature points in one group of images with multi-class labels are used to estimate the distribution of this multi-class. In SML algorithm, training images are divided into two parts: images with this class label and images without this class label. All feature points in the former part of images are used to estimate the class-conditional distribution. If one training image is wrongly labeled with this class label, all feature points of this image are wrong samples.

In contrast to this method, the images in the class are divided into several groups according the properties with multi-class labels. In the case of if one training image is wrongly labeled with one single-class label, only the feature points associated with that wrong label become wrong samples instead of all feature points. Besides, the artificial true points are generated using multiple pairs of distributions of multi-class. For the classification problem, if we have the more the correct training data, we will get the better distribution estimation. Consequently, good classifier

would be obtained, and image annotation results would be more accuracy.

6 Conclusions

In this paper, a combined optimization method, which incorporates AP algorithm, training data optimization algorithm and prior distribution modeling strategy, is developed for image semantic annotation problem. When building the classifier, image feature distribution of each image is approximated with a Gaussian mixture model, and the class-conditional distribution among images in each class is modeled using a hierarchical structure of GMM. In the modeling process, AP algorithm is applied to improve the time performance in estimating image feature distribution. And the training data optimization algorithm is developed to improve the accuracy of the Bayesian classifier. In addition, the prior distribution modeling strategy is also developed to raise the accuracy of image annotation. Both the theoretical analysis and the experimental studies show that the proposed algorithm can improve the accuracy and efficiency of image modeling for image semantic annotation.

The proposed algorithm could be extended to the application for the case of incomplete class labels or complex background images in the further work. This will need a more sophisticated consideration of situation of less strict assumptions, utilization of the human visual cognition knowledge, and the development of the cognitive computing model for image semantic analysis.

References

1. Yang D, Guo P (2009) Improvement of image modeling with affinity propagation algorithm for image semantic annotation. In: Proceedings of international conference on neural information processing, Bangkok, pp 778–787
2. Eakins JP (1996) Automatic image content retrieval—are we getting anywhere? In: Proceedings of international conference on

electronic library and visual information research, London, pp 123–135

3. Barnard K, Duygulu P, Forsyth D, de Freitas N, Blei DM, Jordan MI (2003) Matching words and pictures. J Mach Learn Res 3:1107–1135

4. Carson C, Belongie S, Greenspan H, Malik J (2002) Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Trans Pattern Anal Mach Intell 24(8):1026–1038

5. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(5):993–1022

6. Jeon J, Lavrenko V, Manmatha R (2003) Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of annual ACM conference on research and development in information retrieval, Genevese, pp 119–126

7. Vailaya A, Figueiredo M, Jain A, Zhang HJ (2001) Image classification for content-based indexing. IEEE Trans Image Process 10(1):117–130

8. Luo J, Savakis A (2001) Indoor vs outdoor classification of consumer photographs using low-level and semantic features. In: Proceedings of international conference on image processing, Thessaloniki, pp 745–748

9. Lienhart R, Hartmann A (2002) Classifying images on the web automatically. J Electron Imaging 11(4):445–454

10. Shao W, Naghdy G, Phung SL (2007) Automatic image annotation for semantic image retrieval. Lecture Notes in Computer Science, vol 4781. Springer, Berlin, pp 369–378

11. Guo GD, Jain AK, Ma WY, Zhang JJ (2002) Learning similarity measure for natural image retrieval with relevance feedback. IEEE Trans Neural Netw 13(4):811–820

12. Basili1 R, Petitti R, Saracino D (2007) LSA-based automatic acquisition of semantic image descriptions, LNCS 4816. Springer, Berlin, pp 41–55

13. Shen XJ, Ju SG, Cho SY, Li F (2008) Mining user hidden semantics from image content for image retrieval. J Vis Commun Image 19:145–164

14. Li J, Wang J (2003) Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans Pattern Anal Mach Intell 25(9):1075–1088

15. Carneiro G, Chan AB, Moreno PJ, Vasconcelos N (2007) Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans Pattern Anal Mach Intell 29:394–410

16. Vasconcelos N (2004) Minimum probability of error image retrieval. IEEE Trans Signal Process 52(8):2322–2336

17. Lloyd SP (1957) Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd SP(1982) Least squares quantization in PCM. IEEE Trans Inf Theory 28(2):129–137

18. Frey BJ, Dueck D (2006) Mixture modeling by affinity propagation. In: Proceedings of advances in neural information processing systems, Vancouver, pp 379–386

19. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Sci Agric 315:972–976

20. Dueck D, Frey BJ (2007) Non-metric affinity propagation for unsupervised image categorization. In: Proceedings of IEEE international conference on computer vision, Rio De Janeiro, pp 1–8

21. Vasconcelos N (2001) Image indexing with mixture hierarchies. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition, Hawaii, pp I-3–I-10

22. MPEG-7, http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

23. Quelhas P, Monay F, Odobez JM, Gatica-Perez D, Tuytelaars T (2007) A thousand words in a scene. IEEE Trans Pattern Anal Mach Intell 29(9):1575–1589

24. Guo P, Jia YD, Lyu MR (2008) A study of regularized Gaussian classifier in high-dimension small sample set case based on MDL principle with application to spectrum recognition. Pattern Recognit 41:2842–2854

25. Guo P, Chen CLP, Lyu MR (2002) Cluster number selection for a small set of samples using the bayesian ying-yang model. IEEE Trans Neural Netw 13(3):757–763

26. Visual Object Classes Challenge, http://www.pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009

27. Johnson SC (1967) Hierarchical clustering schemes. Psychometrika 32(3):241–254