Improvement of Image Modeling with Affinity Propagation Algorithm for Semantic Image Annotation

Dong Yang and Ping Guo

Image Processing and Pattern Recognition Laboratory
Beijing Normal University
Beijing 100875, China
d.yang@ieee.org, pguo@ieee.org

Abstract. Semantic image annotation can be viewed as a classification problem, which maps image features to semantic labels, through the procedures of image modeling and image-semantic mapping. In order to improve the performance of image modeling, we propose a novel method which is based on affinity propagation (AP) algorithm. For a given image, low-level image features are extracted from image sub-blocks, and the image feature distribution can be modeled by a mixture of Gaussian components. An adaptive mixture component number selection algorithm which is related to the image semantic information is also developed. The AP algorithm is adopted to improve the efficiency and accuracy of the distribution estimation. For a given label, the overall distribution is modeled, and the mixture component number is selected according to the mixture exemplars extracted from all images and the average value of the preference parameter. The experiment results illustrate that the proposed algorithm has the higher efficiency and accuracy compared with C-means and expectation-maximization (EM) algorithm combination.

Keywords: image annotation, clustering, image modeling, affinity propagation algorithm.

1 Introduction

Automatic semantic image annotation is the process that the database of images are annotated with semantic labels by a computer system automatically. Semantic image annotation can be viewed as a mapping procedure from image features to semantic labels, by the steps of image modeling and image-semantic mapping. Image features include low-level visual features (color, shape, texture, topology), object-level features and 3-dimension scene features. While semantic labels include feature semantics, object semantics, scene semantics, behavior semantics and emotion semantics [1]. The low-level visual features have been successfully used in content based image retrieval (CBIR) [2]. However, high-level image features and semantic labels used in semantic based image retrieval (SBIR) [3] make the retrieval process more flexible.

C.S. Leung, M. Lee, and J.H. Chan (Eds.): ICONIP 2009, Part I, LNCS 5863, pp. 778–787, 2009.

To bridge the semantic gap between low-level image features and high-level semantic labels, we should focus on two key steps: image modeling and image-semantic mapping.

For image modeling, low-level image features are extracted from image sub-blocks, then the image feature distribution is represented by the Gaussian mixture model (GMM), for example, the model parameters are computed by C-means and EM algorithm combination [4][5].

For image-semantic mapping, there are two categories of methods. If each semantic label is considered as a class, the mapping can be viewed as a semantic classification problem, such as earlier indoor-outdoor [10], blobworld [4] and supervised multiclass labeling (SML) [5][11] problems. If each semantic word is viewed as a variable, the mapping is a image-semantic joint modeling problem, such as N-cut based method [3], latent dirichlet allocation (LDA) method[12] and cross-media relevance models(CMRM) [13]. Besides, relevance feedback methods integrate users' feedbacks to retrieve images [14].

When image-semantic mapping is viewed as the classification problem, semantic labels are considered as predefined classes, and the mapping is taken as supervised classification. 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[15]. Luo and Savakis [10] 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. SML [5] [11] adopted a multiclass Bayesian classifier to classify the images with multiple semantic labels, and assumed that the labels have independent distributions although each image has multiple labels. EM algorithm was adopted to iteratively estimate the distribution parameters. However, this is a computational expensive process. Affinity propagation (AP) clustering algorithm is to identify a relatively small number of features, called exemplars to represent the whole features [7] [8]. It seems to produce a better fitness function than mixture modeling with C-means methods [9].

In this work, we intend to apply AP algorithm to find out how to fast estimate the image density distribution model parameters, and how to efficiently produce image annotation results more precisely.

2 Methods

The framework of the proposed method is shown as Figure 1.

In Figure 1, rectangles represent objects, while rounded corner rectangles represent methods. To bridge the semantic gap between image features and semantic labels is the central target of semantic image annotation. And image modeling (modeling of one image), image-semantic mapping (modeling of images and supervised classification) are the three key steps.

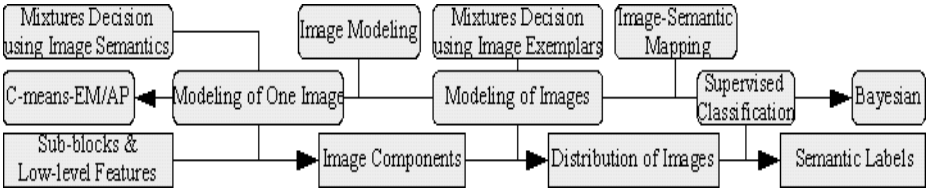


Fig. 1. Framework of the proposed method

2.1 Image Features

Considerable research efforts have been devoted to the low-level image features used in CBIR and SBIR. A localized color feature, which is the discrete cosine transform (DCT) coefficient vector of 8×8 image sub-blocks that overlap 6 pixels between adjacent blocks in YCbCr color space[5], is selected.

$$\mathbf{R}^c_{i,j} = \mathbf{I}^c(2i : 2i+7,\ 2j : 2j+7), \tag{1}$$

$$\mathbf{T}^c = \mathrm{DCT}(\mathbf{R}^c) \quad c = y, cb, cr \quad i, j = 0, 1, 2, 3..., \tag{2}$$

$$\mathbf{X} = [\mathbf{T}^y(:)',\ \mathbf{T}^{cb}(:)',\ \mathbf{T}^{cr}(:)']', \tag{3}$$

$$f(\mathbf{I}) = \{\mathbf{X}_{0,0}, \mathbf{X}_{0,1}, ..., \mathbf{X}_{1,0}, \mathbf{X}_{1,1}, ...\}. \tag{4}$$

Where $\mathbf{R}^c_{i,j}$ is the (i, j)-th sub-block of image $\mathbf{I}$, $\mathbf{T}^c$ is the DCT coefficient matrix of the sub-block $\mathbf{R}^c$, $\mathbf{X}_{i,j}(:)$ is the feature vector that concatenates feature vectors from three color channels, and $f(\mathbf{I})$ is the set of image feature vectors.

2.2 Modeling of One Image

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. It can be briefly described as following [7]:

$$s(i, k) = - \parallel \mathbf{X}_d - \mathbf{X}_k \parallel^2, \tag{5}$$

$$r(i, k) \leftarrow s(i, k) - \max_{k' \neq k} \{a(i, k') + s(i, k')\}, \tag{6}$$

$$a(i, k) \leftarrow \min\{0, r(k, k) + \sum_{i' \neq i, i' \neq k} \max\{0, r(i', k)\}\}. \tag{7}$$

Where the similarity $s(i, k)$ indicates how well the feature vector with index j is the exemplar of feature i. The responsibility $r(i, k)$ reflects the accumulated

evidence for how appropriate feature k is the exemplar of feature i, considering other potential exemplars of feature i. Availability $a(i,k)$ reflects the accumulated evidence for how appropriate it would be for feature i to choose feature k as its exemplar, considering the support from other feature vectors that feature k should be an exemplar. When the preference $s(k,k)$ grows big, each node tends to select itself as the exemplar, then the number of clusters will increase [7].

Clustering features and the mixture model. Considering the dimension and amount of image features, the Gaussian mixture representation is compact and robust. Instead of C-means and EM algorithm combination, we propose an AP-based algorithm for image modeling:

1) AP algorithm is adopted to cluster the feature vectors into several groups with corresponding exemplars;

2) For each group, these similar feature vectors are used to estimate the Gaussian distribution. The weight of each group is estimated according to the number of feature vectors in the group;

3) Each image is represented by the mixture model of these Gaussian distributions and weights.

$$\{\mathbf{e}_i\} = \mathrm{AP}(f(\mathbf{I}), p)\ ,\ i = 1..cn \tag{8}$$

$$\mu_i = \mathbf{e}_i, \quad \mathbf{\Sigma}_i = \mathrm{cov}(\mathbf{A}_i), \quad \omega_i = num(\mathbf{A}_i), \tag{9}$$

$$\mathbf{A}_i = \{\mathbf{x} | exemplar(\mathbf{x}) = \mathbf{e}_i\}, \tag{10}$$

$$P_{\mathbf{X}|\mathbf{I}} = \sum_{i=1..cn} \omega_i\ G(\mu_i, \mathbf{\Sigma}_i) \tag{11}$$

Where the parameter $f(\mathbf{I})$ is the set of image feature vectors. The preference parameter p can be estimated by the adaptive mixture component number selection algorithm described in the next section. And cn is the real number of the exemplars computed by AP algorithm. $\mathbf{A}_i$ means the set of feature vectors whose representation exemplar is $\mathbf{e}_i$, and $G(\mu_i, \mathbf{\Sigma}_i)$ means the Gaussian distribution with mean vector μ_i and covariance matrix $\mathbf{\Sigma}_i$. $P_{\mathbf{X}|\mathbf{I}}$ is the mixture model of image $\mathbf{I}$.

An adaptive mixture component number selection algorithm. There have been several mixture component number selection principles, such as fixed number [5] and the minimum description length principle [4], or more general criterion [6]. We found that the mixture model of clustering features can be referred from the semantic information of the image. That is to say, instead of fixed or homogeneous component number, we develop an adaptive mixture component number selection method incorporating with the semantic labels of the image and corresponding label attributes.

$$cn(\mathbf{I}) = \sum_{s_i \in labels(\mathbf{I})} cn(s_i). \tag{12}$$

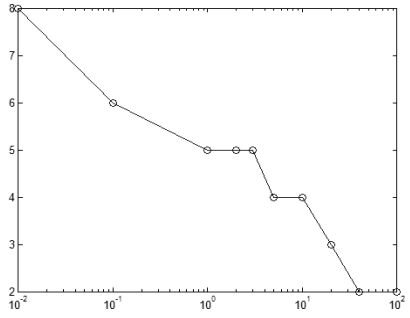
Where $cn(\mathbf{I})$ is the mixture component number of image $\mathbf{I}$, $labels(\mathbf{I})$ are the semantic labels of image $\mathbf{I}$. $cn(s)$ is the empirical approximate mixture number of the semantic label s, which can be estimated in advance. The mixture numbers of some semantic labels are shown as in Table 1.

Table 1. Cluster number of several semantic classes

sky	plant	aeroplane	land	animal
1	1	1,2	2,3	2,3

In AP algorithm, the mixture component number is a variable that relies on the preference parameter. We find that there is a similar mapping relationship between preference and mixture number. As it is shown in Figure 2, from 20p to 100p all can lead to a two-class clustering result. This illustrates that there is a wide range of preference value that can produce a steady clustering result.

(a) original image



(b) cluster number vs. logarithm of the preference

(c) 0.01 (d) 0.1 (e) 1 (f) 2 (g) 3 (h) 5 (i) 10 (j) 20 (k) 40 (l) 100

Fig. 2. Preference values influencing the clustering result. The original image and result images with increasing preference values(0.01p, 0.1p, ...), where p is the median similarity.

An empirical map between mixture number and preference value can be built up in advance. Taking the aeroplane picture 2(a) as an example, this picture has two labels: sky and aeroplane. By looking up the empirical approximate mixture number table, this picture might contain two or three clusters totally. Then by looking up the preference and mixture number map, the preference value might be 20 to 40. We can select a average value 30 as the preference for AP algorithm.

$$p = \text{map}(cn(\mathbf{I})) \tag{13}$$

The preference and mixture number map can be built up in the training process.

2.3 Modeling of Images

The goal of modeling images is to find the prior distribution and the class-conditional distribution in feature space, which can be computed from images with the given label.

Hierarchical distribution estimation. For a given label, images with this label contain two categories of features: features that belong to this label and features that do not belong to this class. There is an assumption that the former features tend to cluster together, while the latter features tend to spread over the entire feature space [5]. We believe that this assumption is reasonable when the number of samples of each label is large and balanced enough.

However, it is too expensive to estimate the distribution from all images at the same time. A hierarchical mechanism is adopted based on a mixture hierarchy where children densities consist of different combinations of subsets of the parents' components. A general description of bottom-up propagating parameters in two consecutive levels is given using EM algorithm [16].

The semantic label distribution can be estimated using the mixture model of images.

$$\mathbf{A}_s = \{P_{\mathbf{X}|\mathbf{I}} | s \in labels(\mathbf{I})\}, \tag{14}$$

$$P_{\mathbf{X}|s} = \text{H}(\mathbf{A}_s) = \sum_{i=1,\ldots,cn(s)} \omega_i \; G(\mu_i, \mathbf{\Sigma}_i)). \tag{15}$$

Where $P_{\mathbf{X}|\mathbf{I}}$ is the mixture model of image $\mathbf{I}$ computed in the section 2.2, and $P_{\mathbf{X}|s}$ is the distribution of label s. $\mathbf{A}_s$ represents the distribution of images with label s, and the function $H(\cdot)$ is the hierarchical distribution estimation algorithm. Fixed cluster number is required when applying $H(\cdot)$ to build mixture model, therefore we need to find out the largest number of clusters from all images, and supply null components to those mixture models that have less number of clusters.

A class-level mixture component number selection algorithm. The mixture component number of class-conditional distribution can be inferred from the exemplars of this class, because the number of the exemplars is relatively smaller than that of all feature vectors. For the hierarchical distribution estimation, this selected number adapts to the real distribution than that of fixed number.

$$cn(s) = cn(\mathbf{A}), \quad \mathbf{A} = \{exemplars(\mathbf{I}) | s \in labels(\mathbf{I})\}. \tag{16}$$

Where $cn(s)$ is the mixture number of label s, and $s \in labels(\mathbf{I})$ means all images with label s.

The algorithm is as follows:

1)For each image in the class, the exemplars and preference parameters are recorded after AP clustering (section 2.2).

2)The average value of the preference parameters is used in clustering the exemplars, in order to produce a proper mixture component number.

The image-level and class-level mixture component number selection algorithms are different:

1) For a given image, the former algorithm adaptively computes an component number instead of fixed number.And a mixture number and preference map is built up previously, because the AP algorithm requires preference parameter instead of component number.

2) For a given label, the latter algorithm adopts AP algorithm to compute the component number instead of fixed number, which is required as a parameter in hierarchical distribution estimation algorithm.

2.4 Supervised Classification

Under the framework of Bayesian classification, both the image annotation and retrieval can be implemented with a minimum probability of error principle. For a given class, the probability that a test image belongs to this class is the product of the class-conditional probabilities of the image components.

$$\lg(P_{\mathbf{I}|s}(\mathbf{I}|s_i)) = \sum_{\mathbf{X}\in\mathbf{I}} \lg(P_{\mathbf{X}|s}(\mathbf{X}|s_i)) \tag{17}$$

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

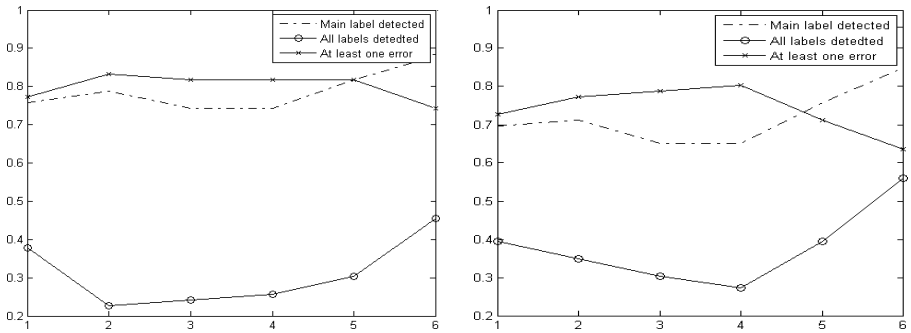
3 Experiments

In this section, we validate the efficiency and accuracy of the image modeling with AP algorithm through annotation results. The images are selected from database [17] and [18]. We selected a subset of outdoor images which contain five classes: aeroplane, sky, land, plant and animal, altogether 378 images are selected. Typically each image contains three or more classes. In order to speed up the processing, all images are resized to the small blocks with the size of from 200×200 to 300×300 pixels.

Table 2 illustrates the efficiency of the proposed algorithm.

The experiment procedure is described as follows:

1. Half of images are for training set, and half of those for testing. Six different sets of training images are selected and the adjacent two sets have five-sixth overlap.
2. For each training set, three factors are computed: a) percentage of some attractive label annotated; b) percentage of all labels annotated; and c) percentage of any wrong label annotated (Figure 3).



(a) C-means+EM (b) AP-based

Fig. 3. Two algorithms are compared using the percentage of images which are labeled correctly

Table 2. Time consumption of modeling one image

methods	25 loops max	50 loops max	100 loops max
C-means	25.8	43.3	46.5
C-means+EM	63.6	133.7	167.3
AP-based	17.8	45.4	73.6

3. For each label, recall and precise factors are computed, averaging from all six training sets (Figure 4).
4. The average time consumption is computed.

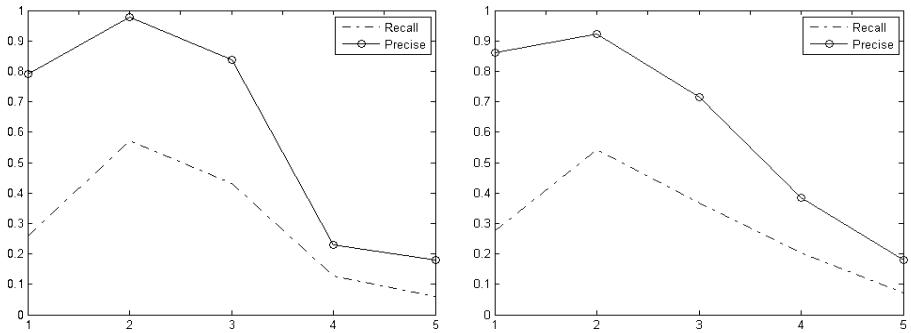
For a given semantic descriptor, assuming that there are w_H human annotated images in the test set and the system automatic annotates number is w_{Auto}, of which w_C are correct, recall and precision are given as following:

$$recall = \frac{w_C}{w_H}, \tag{18}$$

$$precise = \frac{w_C}{w_{Auto}} \tag{19}$$

The labels that are manually annotated might relate with obscure features of the image. Comparing the C-means and EM algorithm combination with the AP-based algorithm, we find that there are about 70 % of test images in which most attractive label is annotated, and about 30 % of images in which all labels are annotated. However, AP-based algorithm improves the percentage that all labels are annotated, and reduces the percentage that wrong label is annotated(Figure 3).

Figure 4 illustrates that the accuracy is improved with the proposed algorithm for three classes, while that for other two classes is near same with C-means and EM algorithm combination. From Figure 3 we can easily know that the recall or precise values are different when the classification model is built with different training sets, which means that the distributions of the labels in the database are



(a) C-means+EM (b) AP-based

Fig. 4. The comparison of the average recall and precise values, computed with one particular label among five labels

uneven. The recall or precise values of the five classes in Figure 4 are different, probably because the classes have large difference amount of information in those images. That is to say, the image database requires to be well organized in order to improve the annotation performance.

4 Conclusions

In this paper, we have investigated the improvement problem of image modeling with AP algorithm for semantic image annotation. The efficiency and accuracy of distribution estimation is improved when AP algorithm is adopted. For a given image, a mixture component number selection method is developed on considering the semantic labels. For a given label, the mixture component number is selected according to the average parameter value and the mixture exemplars extracted from all training data set. The experiment results show that the effectiveness of the developed number selection methods. When the algorithm developed from this study is applied to the automatic image annotation problem, it certainly can accelerate and optimize the image retrieval process.

Acknowledgement. The research work described in this paper was fully supported by the grants from the National Natural Science Foundation of China (Project No. 60675011, 90820010). Prof. Ping Guo is the author to whom all correspondence should be addressed.

References

1. Eakins, J.P.: Automatic Image Content Retrieval - Are We Getting Anywhere? In: Proc. of the Third International Conf. on Electronic Library and Visual Information Research, pp. 123–135 (1996)
2. Datt, R., Li, J., Wang, J.Z.: Content-based Image Retrieval: Approaches and Trends of The New Age. In: ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 253–262 (2005)

3. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., Jordan, M.I.: Matching Words and Pictures. J. of Machine Learning Research 3, 1107–1135 (2003)
4. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image Segmentation Using Expectation-maximization and Its Application to Image Querying. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(8), 1026–1038 (2002)
5. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised Learning of Semantic Classes for Image Annotation and Retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence 29, 394–410 (2007)
6. Guo, P., Chen, C.L.P., Lyu, M.R.: Cluster number selection for a small set of samples using the Bayesian ying-yang Model. IEEE Trans. Neural Network 13(3), 757–763 (2002)
7. Frey, B.J., Dueck, D.: Mixture Modeling by Affinity Propagation. In: Advances in Neural Information processing Systems, pp. 379–386 (2006)
8. Frey, B.J., Dueck, D.: Clustering by Passing Messages between Data Points. Science 315, 972–976 (2007)
9. Dueck, D., Frey, B.J.: Non-metric Affinity Propagation for Unsupervised Image Categorization. In: IEEE International Conf. on Computer Vision, pp. 1–8 (2007)
10. Luo, J., Savakis, A.: Indoor vs Outdoor Classification of Consumer Photographs Using Low-level and Semantic Features. In: International Conf. on Image Processing, pp. 745–748 (2001)
11. Vasconcelos, N.: Minimum Probability of Error Image Retrieval. IEEE Trans.on Signal Processing 52(8), 2322–2336 (2004)
12. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. of Machine Learning Research 3(5), 993–1022 (2003)
13. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic Image Annotation and Retrieval Using Cross-media Relevance Models. In: Annual ACM Conf. on Research and Development in Information Retrieval, pp. 119–126 (2003)
14. Zhou, X.S., Huang, T.S.: Relevance Feedback in Image Retrieval: A Comprehensive Review. Multimedia Systems 8(6), 536–544 (2003)
15. Vailaya, A., Figueiredo, M., Jain, A., Zhang, H.J.: Image classification for content-based indexing. IEEE Trans. on Image processing 10(1), 117–130 (2001)
16. Vasconcelos, N.: Image Indexing with Mixture Hierarchies. In: IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 10, p. I-3–I-10 (2001)
17. Visual Object Classes Challange, `http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008`
18. Caltech 256, `http://www.vision.caltech.edu/Image_Datasets/`