International Journal of Intelligence Science, **2012, 2, 55-62**
http://dx.doi.org/10.4236/ijis.2012.23008 Published Online July 2012 (http://www.SciRP.org/journal/ijis)

Combining Generative/Discriminative Learning for Automatic Image Annotation and Retrieval

Zhixin Li[1], Zhenjun Tang[1], Weizhong Zhao[2], Zhiqing Li[2]
[1]College of Computer Science and Information Technology, Guangxi Normal University, Guilin, China
[2]College of Information Engineering, Xiangtan University, Xiangtan, China
Email: {lizx, zjtang}@gxnu.edu.cn, {zhaoweizhong, lizhiqingchina}@gmail.com

Received April 24, 2012; revised May 30, 2012; accepted June 9, 2012

ABSTRACT

In order to bridge the semantic gap exists in image retrieval, this paper propose an approach combining generative and discriminative learning to accomplish the task of automatic image annotation and retrieval. We firstly present continuous probabilistic latent semantic analysis (PLSA) to model continuous quantity. Furthermore, we propose a hybrid framework which employs continuous PLSA to model visual features of images in generative learning stage and uses ensembles of classifier chains to classify the multi-label data in discriminative learning stage. Since the framework combines the advantages of generative and discriminative learning, it can predict semantic annotation precisely for unseen images. Finally, we conduct a series of experiments on a standard Corel dataset. The experiment results show that our approach outperforms many state-of-the-art approaches.

Keywords: Automatic Image Annotation; Continuous PLSA; Semantic Gap; Hybrid Approach; Image Retrieval

1. Introduction

As an important research issue, Content-based image retrieval (CBIR) searches relative images of given example in visual level. Under this paradigm, various low-level visual features are extracted from each image in the database and image retrieval is formulated as searching for the best database match to the feature vector extracted from the query image. Although this process is accomplished quickly and automatically, the results are seldom semantically relative to the query example due to the notorious *semantic gap* [1]. As a result, automatic image annotation has emerged as a crucial problem for semantic image retrieval [2,3].

The state-of-the-art techniques of automatic image annotation can be categorized into two different schools of thought. The first one is based on discriminative model. It defines auto-annotation as a traditional supervised classification problem [4-8], which treats each semantic concept as an independent class and creates different classifiers for different concepts. This approach computes similarity at the visual level and annotates a new image by propagating the corresponding words. The second perspective takes a different stand. It is based on generative model and treats image and text as equivalent data. It attempts to discover the correlation between visual features and textual words on an unsupervised basis by estimating the joint distribution of features and words. Thus, it poses annotation as statistical inference in a graphical model. Under this perspective, images are treated as bags of words and features, each of which are assumed generated by a hidden variable. Various approaches differ in the definition of the states of the hidden variable: Some associate them with images in the database [9-11], while others associate them with image clusters [12,13] or latent aspects (topics) [14-16]. These two kinds of approaches have their own advantages and disadvantages. Although hybrid Approach has been used in scene classification [17], this paper will show that it is feasible and effective to combine the advantages of these two formulations.

As a latent aspect model, PLSA [18] and *latent Dirichlet allocation* (LDA)[19] have been successfully applied to annotate and retrieve images. PLSA-WORDS [15] is a representative approach, which achieves the annotation task by constraining the latent space to ensure its consistency in words. However, since traditional PLSA can only handle discrete quantity (such as textual words), this approach quantizes feature vectors into discrete visual words for PLSA modeling. Therefore, its annotation performance is sensitive to the clustering granularity. In the area of automatic image annotation, it is generally believed that using continuous feature vectors will give rise to better performance [7,10,14,16]. In order to model image data precisely, it is required to deal with continuous quantity using PLSA.

This paper presents continuous PLSA, which assumes that feature vectors of images are governed by a Gaussian distribution under a given latent aspect other than a multinomial one. In addition, corresponding EM algorithm is derived to estimate the parameters. Then, each image can be treated as a mixture of Gaussians under this model. Furthermore, we propose a hybrid framework to learn semantic classes of images. The framework employs continuous PLSA to model visual features of images in generative learning stage, and uses ensembles of classifier chains [20] to classify the multi-label data in discriminative learning stage. We compare our approach with some state-of-the-art approaches on a standard Corel dataset and the experiment results show that our approach performs more effectively and precisely.

The rest of the paper is organized as follows. Section 2 presents the continuous PLSA model and derives corresponding EM algorithm. Section 3 proposes a hybrid framework and describes the training and annotation procedure. Experiment results are reported and analyzed in Section 4. Finally, the overall conclusions of this work are presented in Section 5.

2. Related Work

Various approaches based on discriminative model have been proposed for semantic image annotation and retrieval. A representative work is automatic linguistic indexing of pictures (ALIP) proposed by Li and Wang [4]. ALIP uses two-dimensional multi-resolution hidden Markov models (2D MHMMs) to capture spatial dependencies of visual features of given semantic categories. Besides, the content-based soft annotation (CBSA) system proposed by Chang *et al.* [5] is based on binary classifyers trained for each word and it indexes a new image with the output of each classifier. They experiment with two learning methods, Bayes point machines (BPMs) and support vector machines (SVMs), and compare their class prediction accuracy. Cusano *et al.* [6] annotate images by a classification system based on a multi-class SVM. They claim that there system could be applied in the management of large image and video databases. Caneiro *et al.* [7] propose supervised multiclass labeling (SML), which employs optimal principle of minimum probability of error and treats annotation as a multiclass classification problem where each of the semantic concepts of interest defines an image class. At annotation stage, these classes all directly compete for the image to annotate. Therefore, this approach no longer suffers a sequence of independent binary tests. Afterwards, Wang *et al.* [8] present a multi-label sparse coding (MSC) framework for feature extraction and classification within the context of automatic image annotation. This method propagates the multi-labels of the training images to the query image with the sparse $\ell 1$ reconstruction coefficients.

Most Approaches based on generative model detect semantic concepts from images by learning the correlation between visual features and textual words. Duygulu *et al.* [12] propose machine translation models, in which the words and blobs are considered as two equivalent languages. After training, the translation model can translate blobs into words, that is, it can attach words to a new image region. Barnard *et al.* [13] discuss several models to represent the joint distribution of words and blobs. Once the joint distribution has been learned, the annotation problem is converted into a likelihood problem relating blobs to words. However, the performance of these models is strongly affected by the quality of image segmentation. Similarly, Blei *et al.* [14] employ correspondence latent Dirichlet allocation (LDA) model to build a language-based correspondence between words and images. The model can be viewed in terms of a generative process that first generates the region descriptions and subsequently generates the caption words. In addition to this, Jeon *et al.* [9] propose cross-media relevance models (CMRM) to annotate images, assuming that the blobs and words are mutually independent given a specific image. Lavrenko *et al.* [10] propose similar continuous-space relevance model (CRM), in which the word probabilities are estimated using multinomial distribution and the blob feature probabilities using a non-parametric kernel density estimate. Compared with CMRM, CRM directly models continuous feature, therefore it does not rely on clustering and consequently does not suffer from the granularity issues. Feng *et al.* [11] propose multiple Bernoulli relevance model (MBRM), in which a multiple Bernoulli distribution is used to generate words instead of the multinomial one as in CRM. Besides, Monay *et al.* [15] propose a new way of modeling multi-modal co-occurrences. This approach constrains the definition of latent space to ensure its consistency in semantic terms (words), while retaining the ability to jointly model visual information. Zhang *et al.* [16] propose a probabilistic semantic model in which the visual features and the textual words are connected via a hidden layer which constitutes the semantic concepts to be discovered. Liu *et al.* [21] propose a graph learning framework for image annotation. A nearest spanning chain method is proposed to construct the image-based graph, whose edge-weights are derived from the chain-wise statistical information instead of the traditional pair-wise similarities. Furthermore, the word-based graph learning is developed to refine the relationships between images and words to get final annotations for each image.

In our previous works [22-24], we propose PLSA-FUSION and GM-PLSA. PLSA-FUSION employs two PLSA models to capture semantic information from visual and textual modalities respectively, while GM-PLSA improves the learning procedure by modeling visual features directly. However, when we use PLSA to model the textual words, each image has a very sparse histogram because there are only 4 or 5 words related to an image. Therefore, both PLSA-FUSION and GM-PLSA adopt asymmetric learning approach to learn the correlation between visual and textual modalities. Although the efficiency and accuracy of asymmetric learning approach are quite good, we believe that using multi-label learning to learn semantic concepts is more appropriate to solve the problem caused by sparse representation of textual words. In the experiment section, we will see that our hybrid approach indeed acquire higher accuracy and superior effectiveness.

3. Continuous PLSA

Just like traditional PLSA, continuous PLSA is also a statistical latent class model which introduces a hidden variable (latent aspect) $z_k\left(k\in 1,\cdots,K\right)$ in the generative process of each element $x_i\left(j\in 1,\cdots,M\right)$ in a document $d_i\left(i\in 1,\cdots,N\right)$. However, given this unobservable variable z_k, continuous PLSA assumes that elements x_j are sampled from a multivariate Gaussian distribution, instead of a multinomial one in traditional PLSA. Using these definitions, continuous PLSA [23] assumes the following generative process:

1) Select a document d_i with probability $P(d_i)$;

2) Sample a latent aspect z_k with probability $P(z_k|d_i)$ from a multinomial distribution conditioned on d_i;

3) Sample $x_j \sim P\left(x_j\middle|z_k\right)$ from a multivariate Gaussian distribution $N\left(x\middle|\mu_k,\Sigma_k\right)$ conditioned on z_k.

Continuous PLSA has two underlying assumptions. First, the observation pairs (d_i, x_j) are generated independently. Second, the pairs of random variables (d_i, x_j) are conditionally independent given the latent aspect z_k. Thus, the joint probability of the observed variables is obtained by marginalizing over the latent aspect z_k,

$$P\left(d_x,x_j\right)=P\left(d_i\right)\sum_{k=1}^{K}P\left(z_x\middle|d_i\right)P\left(x_j\middle|z_x\right) \tag{1}$$

A representation of the model in terms of a graphical model is depicted in **Figure 1**.

The mixture of Gaussian is assumed for the conditional probability $P(.|z)$. In other words, the elements are generated from K Gaussian distributions, each one corresponding a z_k. For a specific latent aspect z_k, the condition probability distribution function of elements x_j is

$$P\left(x_j\middle|z_k\right)=\frac{1}{\left(2\pi\right)^{d/2}\left|\Sigma_{\boldsymbol{k}}\right|^{1/2}}\exp\left\{-\frac{1}{2}\left(x_j-\boldsymbol{\mu_k}\right)^T\Sigma_{\boldsymbol{k}}^{-1}\left(x_j-x_j\right)\right\} \tag{2}$$

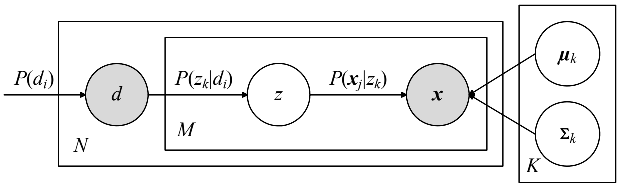


Figure 1. Graphical model representation of continuous PLSA.

where D is the dimension, μ_k is a D-dimensional mean vector and Σ_k is a D × D covariance matrix.

Following the maximum likelihood principle, $P(z_k|d_i)$ and $P(x_j|z_k)$ can be determined by maximization of the log-likelihood function

$$\begin{aligned}\mathcal{L}&=\sum_{i=1}^{N}\sum_{j=1}^{M}n\left(d_i,x_j\right)\log P\left(d_i,x_j\right)=\sum_{i=1}^{N}n\left(d_i\right)\log P\left(d_i\right)\\&+\sum_{i=1}^{N}\sum_{j=1}^{M}n\left(d_i,x_j\right)\log\sum_{k=1}^{K}P\left(z_x\middle|d_i\right)P\left(x_j\middle|z_k\right)\end{aligned} \tag{3}$$

where $n(d_i, x_j)$ denotes the number of element x_j in d_i.

The standard procedure for maximum likelihood estimation in latent variable models is the EM algorithm [25,26]. In the E-step, applying Bayes' theorem to (1), one can obtain

$$P\left(z_k\middle|d_i,x_i\right)=\frac{P\left(z_k\middle|d_i\right)P\left(x_j\middle|z_k\right)}{\sum_{l=1}^{K}P\left(z_l\middle|d_i\right)P\left(x_j\middle|z_l\right)} \tag{4}$$

In the M-step, one has to maximize the expectation of the complete-data log-likelihood

$$\begin{aligned}E\left(\mathcal{L}^c\right)&=\sum_{i=1}^{N}\sum_{j=1}^{M}n\left(d_i,x_j\right)\\&\cdot\sum_{k=1}^{K}P\left(z_k\middle|d_i,x_j\right)\log\left[P\left(z_x\middle|d_i\right)P\left(x_j\middle|z_k\right)\right]\end{aligned} \tag{5}$$

Maximizing (5) with Lagrange multipliers to $P(z_k|d_i)$ and $P(x_j|z_k)$ respectively, under the following constraints

$$\sum_{k=1}^{K}P\left(z_k\middle|d_i\right)=1,\sum_{k=1}^{K}P\left(z_k\middle|d_i,x_j\right)=1 \tag{6}$$

For any d_i, z_k and x_j, the parameters are determined as

$$P\left(z_k\middle|d_i\right)=\frac{\sum_{j=1}^{M}n\left(d_i,x_j\right)P\left(z_k\middle|d_i,x_j\right)}{\sum_{j=1}^{M}n\left(d_i,x_j\right)} \tag{7}$$

$$\boldsymbol{\mu_k}=\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}n\left(d_i,\boldsymbol{x}_j\right)P\left(z_k\middle|d_i,\boldsymbol{x}_j\right)\boldsymbol{x}_j}{\sum_{i=1}^{N}\sum_{j=1}^{M}n\left(d_i,\boldsymbol{x}_j\right)P\left(z_k\middle|d_i,\boldsymbol{x}_j\right)}, \tag{8}$$

$$\boldsymbol{\Sigma_k}=\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}n\left(d_i,\boldsymbol{x}_j\right)P\left(z_k\middle|d_i,\boldsymbol{x}_j\right)\left(x_j-\boldsymbol{\mu_k}\right)\left(x_j-\boldsymbol{\mu_k}\right)^T}{\sum_{i=1}^{N}\sum_{j=1}^{M}n\left(d_i,\boldsymbol{x}_j\right)P\left(z_k\middle|d_i,\boldsymbol{x}_j\right)} \tag{9}$$

Alternating (4) with (7) - (9) defines a convergent procedure to a local maximum of (5). The EM algorithm terminates by either a convergence condition or *early stopping* technique.

As for the parameters, if parameter $P(x_j|z_k)$ is known, we could quickly infer the other parameters μ_k and Σ_k using folding-in method, and vice versa. Folding-in method is a partial version of the EM algorithm. It updates the unknown parameters with the known parameters kept fixed, so that it can maximize the likelihood with respect to the previously trained parameters.

4. Hybrid Generative/Discriminative Model

This section describes the construction and the related algorithms of Hybrid Generative/Discriminative Model (HGDM).

4.1. Hybrid Framework

On the basis of continuous PLSA, we propose a hybrid framework which combines generative and discriminative learning. The framework employs continuous PLSA to model visual features of images. As a result, each image can be represented as an aspect distribution. Then, this intermediate representation is used to build ensembles of classifier chains, which can learn semantic classes of images and consider the correlation between the labels at the same time. The framework is shown in **Figure 2**.

In training procedure, we firstly get the parameters μ_k and Σ_k given aspect z_k by modeling visual features of training images with continuous PLSA. At the same time, the aspect distribution $P(z_k|d_i)$ of each image is determined. This is the generative learning stage. The parameters μ_k and Σ_k are parameters of continuous PLSA. According to the independence assumption, these parameters remain valid for documents out of the training set. On the other hand, the aspect distribution $P(z_k|d_i)$ is only relative to the specific documents and cannot carry any prior information to an unseen image. This distribution, however, can represent each training image as a K-di-

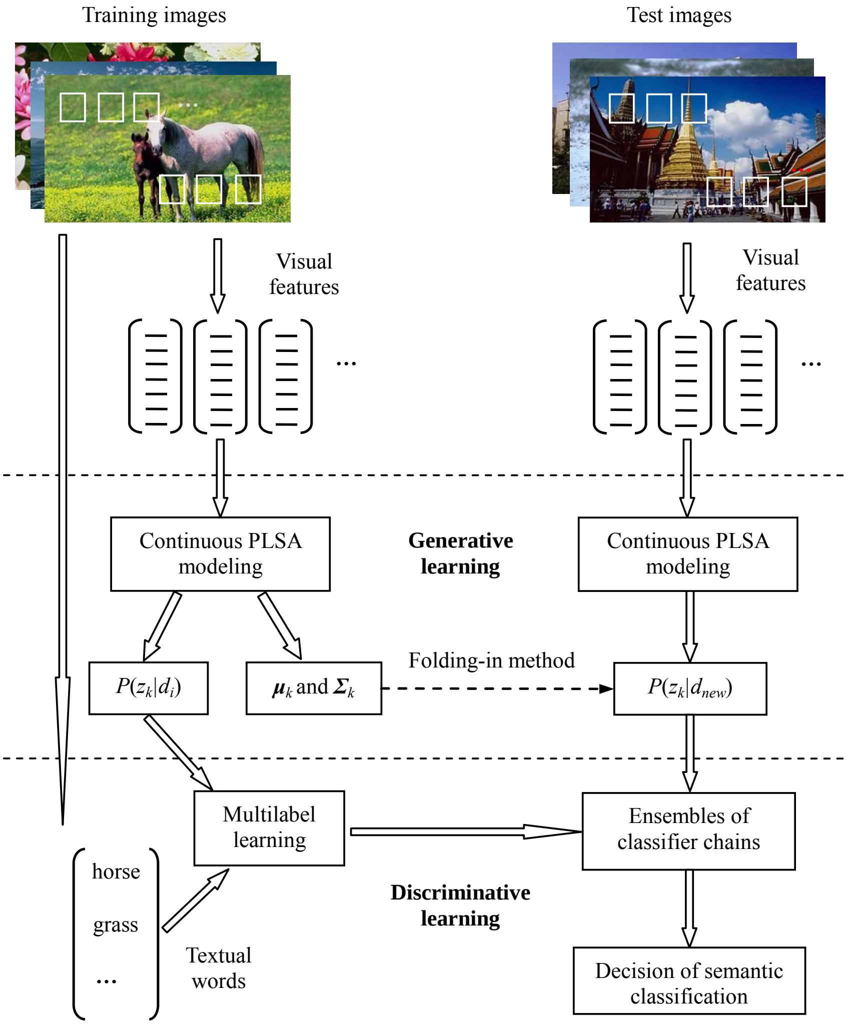


Figure 2. Learning procedure of hybrid framework.

mension vector. In addition, all the vectors can construct a simplex. Then, by making use of the aspect distribution and original annotation labels of each training image, we build a series of classifiers in which every word in the vocabulary is treated as an independent class. This is the discriminative learning stage. At this time, every image is represented as an aspect distribution, but has several semantic labels. This circumstance is in conformity with multi-label learning, which can construct multiclass classifiers and integrate correlative information of textual words at the same time.

Correspondingly, there are two steps in annotation procedure. Firstly, since model parameters μ_k and Σ_k are determined in training procedure, we can compute the aspect distribution of each test image using folding-in method. Secondly, we classify the aspect distribution of each test image with the trained ensembles of classifier chains. Furthermore, we choose 5 words with highest confidence as annotations of the test image. After each image in the database is annotated, the retrieval algorithm ranks the images labeled with the query word by decreasing confidence.

4.2. Ensembles of Classifier Chains

In discriminative learning stage, we employ ensembles of classifier chains [21] to accomplish the task of multi-label classification. Each binary classifier is implemented with SVM in classifier chains. Having taken the correlation between semantic labels into consideration, this approach can classify images into several semantic classes and it has higher confidence with acceptable computation complexity.

The classifier chain model involves $|L|$ binary classifyers, where L denotes the label set. Classifiers are linked along a chain where each classifier deals with the binary relevance problem associated with label $l_j \in L$. The feature space of each link in the chain is extended with the 0/1 label associations of all previous links. The training procedure is described in **Algorithm 1**. Note the notation for a training example ($\boldsymbol{x}$, S), where $S \subseteq L$ and $\boldsymbol{x}$ is an instance feature vector.

Hence a chain $C_1, C_2, \cdots, C_j$ of binary classifiers is formed. Each classifier C_j in the chain is responsible for learning and predicting the binary association of label l_j given the feature space, augmented by all prior binary relevance predictions in the chain $l_1, l_2, \cdots, l_{j-1}$. The classification process begins at C_1 and propagates along the chain: C_1 determines $Pr(l_1|\boldsymbol{x})$ and every following classifier $C_2, \cdots, C_j$ predicts $Pr(l_j|\boldsymbol{x}_i, l_1, l_2, \cdots, l_{j-1})$. This classification procedure is described in **Algorithm 2**.

Algorithm 1. Training procedure of classifier chain.

```
Input: Example set $D= \{(\boldsymbol{x}_1, S_1), (\boldsymbol{x}_2, S_2), \cdots, (\boldsymbol{x}_n, S_n)\}$.
Output: Classifier chains $\{C_1, C_2, \cdots, C_{|L|}\}$.
Process:
for j $\in$ 1, 2, $\cdots$, $|L|$
do single-label transformation and training
  $D' \leftarrow \{\}$
  for $(\boldsymbol{x}, S) \in D$
    do $D' \leftarrow D' \cup ((\boldsymbol{x}, l_1, l_2, \cdots, l_{j-1}), l_j)$
  Train $C_j$ to predict binary relevance of $l_j$
  $C_j: D' \rightarrow l_j \in \{0, 1\}$
```

Algorithm 2. Classifying procedure of classifier chain.

```
Input: Test example $\boldsymbol{x}$.
Output: Results of all classifiers in the chain $Y = \{l_1, l_2, \cdots, l_{|L|}\}$.
Process:
$Y \leftarrow \{\}$
for j $\in$ 1, 2, $\cdots$, $|L|$
  do $Y \leftarrow Y \cup (l_j \leftarrow C_j: (\boldsymbol{x}, l_1, l_2, \cdots, l_{j-1}))$
return $(\boldsymbol{x}, Y)$
```

This training method passes label information between classifiers, allowing classifier chain take into account label correlations and thus overcoming the label independence problem of binary relevance method. However, classifier chain still remains advantages of binary relevance method including low memory and runtime complexity.

The order of the chain itself clearly has an effect on accuracy. This problem can be solved by using an ensemble framework with a different random train ordering for each iteration. Ensemble of classifier chains trains m classifiers $C_1, C_2, \cdots, C_m$. Each C_k is trained with a random chain ordering of L and a random subset of D. Hence each C_k model is likely to be unique and able to give different multi-label predictions. These predictions are sumed by label so that each label receives a number of votes. A threshold is used to select the most popular labels which form the final predicted multi-label set.

5. Experimental Results

In our prototype system, we have implemented PLSA-WORDS, PLSA-FUSION, GM-PLSA and the hybrid generative/discriminative model (HGDM) proposed in this paper. The process of PLSA-based models fitting and classifiers training are executed offline; the task of image annotation and retrieval is performed online.

In order to test the effectiveness and accuracy of HGDM, we conduct our experiments on an annotated image data set which was originally used in [12]. The dataset consists of 5000 images from 50 Corel Stock Photo cds. Each cd includes 100 images on the same topic. We divided this dataset into 3 parts: a training set of 4000 images, a validation set of 500 images and a test set of 500 images. The validation set is used to determine system parameters. After fixing the parameters, we merged the 4000 training set and 500 validation set to form a new training set. This corresponding to the training set of 4500 images and the test set of 500 images used by [12].

5.1. Parameters Setting

An important parameter of the experiment is the number of latent aspects for the PLSA-based models. Since the number of latent aspects defines the capacity of the model—the number of model parameters, it can determine the training time and system efficiency to a large extent. We choose three values (90, 120 and 150) of aspect number to do experiments. Through a series of experiments, we found that the system performs better when aspect number is 150. Therefore, we use 150 as aspect number, without ruling out the possibility that another aspect number would make the system performs much better. Furthermore, our approach constructs an ensemble including 90 classifier chains. Each classifier chain randomly chooses a subset of 500 images for training.

The focus of this paper is not on image feature selection and our approach is independent of visual features. So our prototype system uses similar features to [11] for easy comparison. We simply decompose images into a set of blocks (the size of each block is empirically determined as 16 × 16 through a series of experiments on the validation set), then compute a 36 dimensional feature vector for each block, consisting of 24 color features (auto correlogram computed over 8 quantized colors and 3 Manhattan Distances) and 12 texture features (Gabor energy computed over 3 scales and 4 orientations). As a result, each block is represented as a 36 dimension feature vector. Then each image is represented as a bag of features, that is, a set of 36 dimension vectors. All the feature vectors of training images compose the inputs of continuous PLSA. Therefore, this preprocessing procedure provides a uniform interface for continuous PLSA modeling.

5.2. Results of Automatic Image Annotation

In this section, the performance of HGDM is compared with some state-of-the-art approaches—the Translation Model [12], CMRM [9], CRM [10], MBRM [11], PLSA-WORDS [15], SML [7], TGLM [21] and MSC [8]. We evaluate the performance of image annotation by comparing the captions automatically generated with the original manual annotations. Similarly to [10], we compute the recall and precision of every word in the test set and use the mean of these values to summarize the system performance.

We report the results on two sets of words: the subset of 49 best words and the complete set of all 260 words that occur in the training set. The systematic evaluation results are shown in **Table 1**. From the table, we can see that HGDM performs significantly better than all other models. We believe that using hybrid framework to learn semantic classes is the reason for this result.

Several examples of annotation obtained by our prototype system are shown in **Table 2**. Here top five words are taken as annotation of the image. We can see that even the system annotates an image with a word not contained in the ground truth, this annotation is frequently plausible.

Table 1. Performance comparison on the task of automatic image annotation.

Models	Translation	CMRM	CRM	MBRM	PLSA-WORDS	SML	TGLM	MSC	HGDM
#words with recall > 0	49	66	107	122	105	137	131	136	137
Results on 49 best words									
Mean Recall	0.34	0.48	0.70	0.78	0.71	—	—	0.82	0.83
Mean Precision	0.20	0.40	0.59	0.74	0.56	—	—	0.76	0.78
Results on all 260 words									
Mean Recall	0.04	0.09	0.19	0.25	0.20	0.29	0.29	0.32	0.32
Mean Precision	0.06	0.10	0.16	0.24	0.14	0.23	0.25	0.25	0.28

Table 2. Comparison of annotations made by PLSA-WORDS and HGDM.

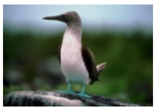
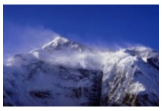
Image				
Ground truth	blue-footed, booby, rock, bird	elk, bugle, antlers, grass	peak, mountains, snow, sky	building, courtyard, sky, trees
Annotations of PLSA-WORDS	bird, nest, close-up, booby, branch	forest, sun, elk, antlers, sunset	mountains, sky, peak, snow, landscape	sky, building, wall, sand, temple
Annotations of HGDM	booby, bird, trees, rock, sky	antlers, trees, grass, elk, sand	sky, mountain, snow, peak, ice	building, sky, wall, trees, courtyard

Table 3. Comparison of mAPs in ranked image retrieval.

Models	CMRM	CRM	MBRM	PLSA-WORDS	SML	TGLM	MSC	HGDM
All 260 words	0.17	0.24	0.30	0.22	0.31	0.29	0.42	0.35
Words with recall ≥0	0.20	0.27	0.35	0.26	—	—	—	0.41
Words with recall >0	—	—	—	0.55	0.49	0.52	0.79	0.67

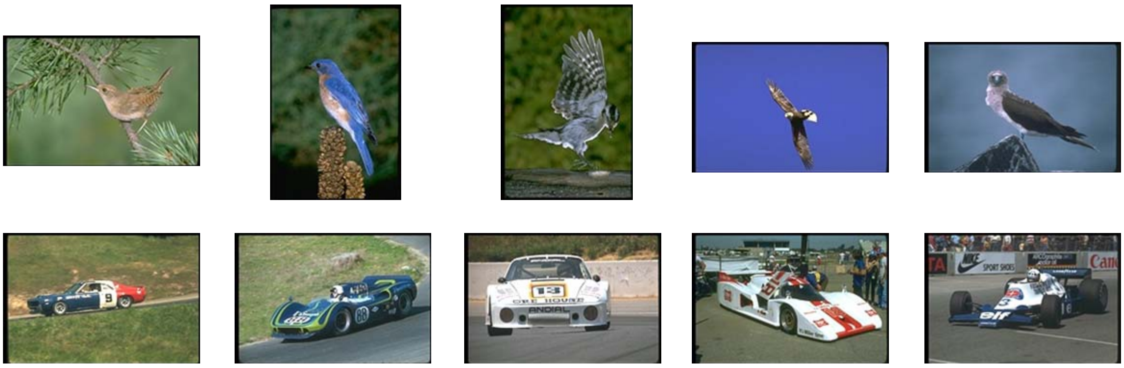


Figure 3. Semantic retrieval results on Corel 5 k dataset.

5.3. Results of Ranked Image Retrieval

In this section, mean average precision (mAP) is employed as a metric to evaluate the performance of single word retrieval. We compare our approach with all models in previous section except the translation model, because mAP of the model cannot be accessed directly from the literatures.

The annotation results ignore rank order. However, users always like to rank retrieval images and hope that the top ranked ones are relative images. In fact, most users do not want to see more than even 10 or 20 images in a query. Therefore, rank order is very important for image retrieval. Given a query word, our system will return all the images which are automatically annotated with the query word and rank the images according to the posterior probabilities of that word. **Table 3** shows that HGDM performs better than other models.

Figure 3 presents two ranked retrieval results obtained with single word queries for challenging concepts. Each row shows the top 5 matches to a semantic query. From top to bottom, the retrieval keywords are "*bird*" and "*car*". The diversity of visual appearance of the returned images indicates that HGDM has good generalization ability.

In summary, the experiment results show that HGDM outperforms some state-of-the-art approaches in many respects, which proves that the continuous PLSA and the hybrid framework is effective in modeling visual features and learning semantic classes of images.

6. Conclusion

In this paper, we have proposed continuous PLSA to model continuous quantity and develop an EM-based iterative procedure to estimate the parameters. Furthermore, we present a hybrid generative/discriminative model, which employs continuous PLSA to deal with the visual features and uses ensembles of classifier chains to learn semantic classes of images. Experiments on the Corel dataset prove that our approach is promising for automatic image annotation and retrieval. In comparison to some state-of-the-art approaches, higher accuracy and superior effectiveness of our approach are reported.

7. Acknowledgements

This work is supported by the National Natural Science Foundation of China (Nos. 61165009, 60903141, 61105052), the Guangxi Natural Science Foundation (Nos. 2012 GXNSFAA 053219, 2012GXNSFBA053166) and the "Bagui Scholar" Project Special Funds.

REFERENCES

[1] A. W. M. Smeulders, M. Worring, S. Santini, *et al.*, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 3, 2000, pp. 1349-1380. doi:10.1109/34.895972

[2] R. Datta, D. Joshi, J. Li, *et al.*, "Image Retrieval: Ideas, Influences, and Trends of the New Age," *ACM Computing Surveys*, Vol. 42, No. 2, 2008, pp. 1-60.

[3] Z. X. Li, Z. P. Shi, Z. Q. Li, *et al.*, "A Survey of Semantic Mapping in Image Retrieval," *Journal of Computer-Aided Design and Computer Graphics*, Vol. 20, No. 8, 2008, pp. 1085-1096. (in Chinese)

[4] J. Li and J. Z. Wang, "Automatic Linguistic Indexing of

Pictures by a Statistical Modeling Approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 9, 2003, pp. 1075-1088. doi:10.1109/TPAMI.2003.1227984

[5] E. Chang, K. Goh, G. Sychay, *et al.*, "CBSA: Content-Based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 13, No.1, 2003, pp. 26-38. doi:10.1109/TCSVT.2002.808079

[6] C. Cusano, G. Ciocca and R. Schettini, "Image Annotation Using SVM," *Proceedings of SPIE Conference on Internet Imaging V*, San Jose, Vol. 5304, 2004, pp. 330-338.

[7] G. Carneiro, A. B. Chan, P. J. Moreno, *et al*., "Supervised Learning of Semantic Classes for Image Annotation and Retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 3, 2007, pp. 394-410. doi:10.1109/TPAMI.2007.61

[8] C. Wang, S. Yan, L. Zhang, *et al.*, "Multi-Label Sparse Coding for Automatic Image Annotation," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Hefei, 20-25 June 2009, pp. 1643-1650.

[9] J. Jeon, V. Lavrenko and R. Manmatha, "Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models," *Proceedings of the* 26*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, 2003, pp. 119-126.

[10] V. Lavrenko, R. Manmatha and J. Jeon, "A Model for Learning the Semantics of Pictures," *Advances in Neural Information Processing Systems* 16, Vol. 16, 2004, pp. 553-560.

[11] S. L. Feng, R. Manmatha and V. Lavrenko, "Multiple Bernoulli Relevance Models for Image and Video Annotation," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, 2004, pp. 1002-1009.

[12] P. Duygulu, K. Barnard, J. F. G. de Freitas, *et al.*, "Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary," *Lecture Notes in Computer Science*, Vol. 2353, 2002, pp. 97-112. doi:10.1007/3-540-47979-1_7

[13] K. Barnard, P. Duygulu, D. Forsyth, *et al.*, "Matching Words and Pictures," *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 1107-1135.

[14] D. M. Blei and M. I. Jordan, "Modeling Annotated Data," *Proceedings of the* 26*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, 28 July-1 August 2003, pp. 127-134.

[15] F. Monay and D. Gatica-Perez, "Modeling Semantic Aspects for Cross-Media Image Indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 10, 2007, pp. 1802-1817. doi:10.1109/TPAMI.2007.1097

[16] R. Zhang, Z. Zhang, M. Li, *et al.*, "A Probabilistic Semantic Model for Image Annotation and Multi-Model Image Retrieval," *Proceedings of the* 10*th IEEE International Conference on Computer Vision*, 15-21 October 2005, pp. 846- 851.

[17] A. Bosch, A. Zisserman and X. Munoz, "Scene Classification Using a Hybrid Generative/Discriminative Approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 4, 2008, pp. 712-727. doi:10.1109/TPAMI.2007.70716

[18] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, Vol. 42, No. 1-2, 2001, pp. 177-196. doi:10.1023/A:1007617005950

[19] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 993-1022.

[20] J. Read, B. Pfahringer, G. Holmes, *et al.*, "Classifier Chains for Multi-Label Classification," *Lecture Notes in Artificial Intelligence*, Vol. 5782, 2009, pp. 254-269.

[21] J. Liu, M. Li, Q. Liu, *et al.*, "Image Annotation via Graph Learning," *Pattern Recognition*, Vol. 42, No. 2, 2009, pp. 218-228. doi:10.1016/j.patcog.2008.04.012

[22] Z. X. Li, Z. P. Shi, X. Liu, Z. Q. Li and Z. Z. Shi, "Fusing Semantic Aspects for Image Annotation and Retrieval," *Journal of Visual Communication and Image Representation*, Vol. 21, No. 8, 2010, pp. 798-805. doi:10.1016/j.jvcir.2010.06.004

[23] Z. X. Li, Z. P. Shi, X. Liu and Z. Z. Shi, "Automatic Image Annotation with Continuous PLSA," *Proceedings of the* 35*th IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, 14-19 March 2010, pp. 806-809.

[24] Z. X. Li, Z. P. Shi, X. Liu and Z. Z. Shi, "Modeling Continuous Visual Features for Semantic Image Annotation and Retrieval," *Pattern Recognition Letters*, Vol. 32, No. 3, 2011, pp. 516-523. doi:10.1016/j.patrec.2010.11.015

[25] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, New York, 2006.

[26] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Vol. 39, No. 1, 1977, pp. 1-38.