AUTOMATIC IMAGE ANNOTATION WITH CONTINUOUS PLSA

Zhixin Li$^{1,2}$, Zhiping Shi$^1$, Xi Liu$^{1,2}$ and Zhongzhi Shi$^1$

1 The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
2 Graduate University of Chinese Academy of Sciences, Beijing 100049, China
{lizx, shizp, liux, shizz}@ics.ict.ac.cn

ABSTRACT

Automatic image annotation has become an important and challenging problem due to the existence of semantic gap. In this paper, we firstly extend probabilistic latent semantic analysis (PLSA) to model continuous quantity. In addition, corresponding Expectation-Maximization (EM) algorithm is derived to determine the model parameters. Furthermore, in order to deal with the data of different modalities in terms of their characteristics, we present a semantic annotation model which employs continuous PLSA and standard PLSA to model visual features and textual words respectively. The model learns the correlation between these two modalities by an asymmetric learning approach and then it can predict semantic annotation for unseen images. We compare our approach with several state-of-the-art approaches on a standard Corel dataset. The experiment results show that our approach performs more effectively and accurately.

Index Terms— automatic image annotation, continuous PLSA, latent aspect model, image retrieval

1. INTRODUCTION

Content-based image retrieval (CBIR) has been studied and explored for decades. Its performance, however, is not ideal enough due to the notorious semantic gap [12]. CBIR retrieves images in terms of their visual features, while users often prefer intuitive text-based image searching. Since manual image annotation is expensive and difficult to be extended to large image database, automatic image annotation has emerged as a striking and crucial problem.

The state-of-the-art techniques of image auto-annotation can be categorized into two different schools of thought. The first one defines auto-annotation as a traditional supervised classification problem [3,10], 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 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, training 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 [6,8,9], while others associate them with image clusters [1,5] or latent aspects (topics) [2,11,13].

As a latent aspect model, PLSA [7] has been successfully applied to annotate and retrieve images. PLSA-WORDS [11] is a representative approach, which achieves the annotation task by constraining the latent space to ensure its consistency in words. However, since standard PLSA can only handle discrete quantity (such as textual words), this approach has to quantize 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 [2,9]. In order to model image data precisely, it is required to deal with continuous quantity using PLSA.

This paper proposes continuous PLSA, which assumes that feature vectors in an image are governed by a Gaussian distribution under a given latent aspect other than a multinomial one. In addition, corresponding EM algorithm is derived to determine the parameters. Then, each image can be treated as a mixture of Gaussians under this model. Furthermore, based on the continuous PLSA and the standard PLSA, we present a semantic annotation model that learns the correlation between the visual features and textual words. Once the parameters are learned by an asymmetric learning approach, this model can predict the semantic information for unseen images. We evaluate our approach on Corel dataset and the experiment results show that our approach outperforms several state-of-the-art approaches.

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 semantic annotation model and describes the asymmetric learning approach. Experiment results are reported and analyzed in section 4. Finally, the overall conclusions of this work are presented in section 5.
2. CONTINUOUS PLSA

Just like standard PLSA, continuous PLSA is also a statistical latent class model which introduces a hidden variable (latent aspect) \( z_k \) \((k \in \{1, \ldots, K\})\) in the generative process of each element \( x_j \) \((j \in \{1, \ldots, M\})\) in a document \( d_i \) \((i \in \{1, \ldots, N\})\). 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 standard PLSA. Using these definitions, continuous PLSA 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 the document \( d_i \);
3. Sample \( x_j \sim P(x_j|z_k) \) from a multivariate Gaussian distribution \( N(\mu_k, \Sigma_k) \) conditioned on the latent aspect \( z_k \).

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 \( z_k \),

\[
P(d, x) = P(d) \sum_{k=1}^{K} P(z_k | d) P(x | z_k).
\]

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

![Graphical model representation of Continuous PLSA](image)

**Fig. 1.** Graphical model representation of Continuous PLSA.

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

\[
P(x_j | z_k) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x_j - \mu_k)^T \Sigma_k^{-1}(x_j - \mu_k)\right),
\]

where \( D \) is the dimension, \( \mu_k \) and \( \Sigma_k \) are the \( D \)-dimensional mean vector and \( D \times D \) covariance matrix of elements belonging to \( z_k \) respectively.

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

\[
\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, x_j) \log P(d_i, x_j),
\]

\[
= \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, x_j) \left[ \log P(d_i) + \log \sum_{k=1}^{K} P(z_k | d_i) P(x_j | z_k) \right],
\]

where \( n(d_i, x_j) \) denotes the number of element \( x_j \) for document \( d_i \).

The standard procedure for maximum likelihood estimation in latent variable models is the EM algorithm [4]. In E-step, applying Bayes’ theorem to (1), one can obtain

\[
P(z_k | d_i, x_j) = \frac{P(z_k | d_i)P(x_j | z_k)}{\sum_{k=1}^{K} P(z_k | d_i)P(x_j | z_k)}.
\]

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

\[
E[\mathcal{L}] = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, x_j) \sum_{k=1}^{K} P(z_k | d_i, x_j) \log \left[ P(z_k | d_i)P(x_j | z_k) \right].
\]

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

\[
\sum_{k=1}^{K} P(z_k | d_i) = 1, \sum_{k=1}^{K} P(x_j | z_k) = 1
\]

for any \( d_i, z_k \) and \( x_j \), the parameters are determined as

\[
\mu_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, x_j) x_j}{\sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, x_j)},
\]

\[
\Sigma_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, x_j) (x_j - \mu_k)(x_j - \mu_k)^T}{\sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, x_j)},
\]

\[
P(z_k | d_i) = \frac{\sum_{j=1}^{M} n(d_i, x_j) P(z_k | d_i, x_j)}{\sum_{k=1}^{K} \sum_{j=1}^{M} n(d_i, x_j) P(z_k | d_i, x_j)}.
\]

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 the parameter \( P(z_k|d_i) \) is known, we could quickly infer the other parameters \( \mu_k \) and \( \Sigma_k \) 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.

3. SEMANTIC ANNOTATION MODEL

We discuss here a model to learn the semantic information from visual and textual modalities of annotated images. In order to deal with the data of different modalities in terms of their characteristics, we employ continuous PLSA and standard PLSA to model visual features and textual words respectively. These two models are linked by sharing the same distribution over latent aspects \( P(z_k|d) \). We refer to the semantic annotation model as Gaussian-multinomial PLSA (GM-PLSA), which is represented in Figure 2.

GM-PLSA 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 the document \( d_i \);
3. For each of the words, Sample \( w_m \) from a multinomial distribution \( \text{Mult}(x|\alpha_k) \) conditioned on the latent aspect \( z_k \).
4. For each of the feature vectors, Sample $f_i$ from a multivariate Gaussian distribution $N(x|\mu_i,\Sigma_i)$ conditioned on the latent aspect $z_i$.

$$P(d_i) \xrightarrow{|z_i} P(z_i|d_i) \xrightarrow{|w_i} P(w_i|z_i) \xrightarrow{|f_i} P(f_i|w_i)$$

Fig. 2. Representation of the annotation model GM-PLSA.

Under this modeling approach, each training image can be viewed as either a mixture of continuous Gaussian in visual modality or a mixture of discrete words in textual modality. Therefore, it can learn the correlation between features and words effectively and predict semantic annotation for an unseen image precisely.

We adopt asymmetric learning approach to estimate the model parameters because an asymmetric learning gives a better control of the respective influence of each modality in the latent space definition. As presented in [11], textual modality is firstly chosen to estimate the mixture of aspects in a given document, which constrains the definition of latent space to ensure its consistency in textual words, while retaining the ability to jointly model visual features.

In training stage, each image is processed and represented as bags of features and words. The aspect distributions $P(z|d)$ are firstly learned for all training documents from textual modality only. At the same time, the parameter $P(w_i|z_i)$ (i.e. $\theta_k$) is determined too. Then we use folding-in method described in section 2 to infer the parameters for the visual modality with the aspect distributions $P(z|d)$ kept fixed. Consequently, we can get the model parameters $\theta_k$, $\mu_k$ and $\Sigma_k$, which remain valid in images out of the training set.

In annotation procedure, given feature vectors of each test image and the previously estimated parameters $\mu_k$ and $\Sigma_k$, the aspect distribution $P(z_i|d_{new})$ can be inferred using folding-in method. The posterior probability of each word in the vocabulary is then computed by

$$P(w|d_{new}) = \sum_{z_i} P(z_i|d_{new})P(w|z_i).$$

As usual, we choose five words with the largest posterior probabilities as annotations of an unseen image.

Having annotated all images in the database, the procedure of semantic image retrieval can be put into practice directly.

4. EXPERIMENT RESULTS

In order to test the effectiveness and accuracy of the proposed approach, we conduct our experiments on an annotated image data set which was originally used in [5]. The dataset consists of 5000 images from 50 Corel Stock Photo cds. Each cd includes 100 images on the same topic. We select 90 images from each topic as training images and other 10 images as test images, which corresponding to the training set of 4500 images and the test set of 500 images used by [5].

This paper is not focus on image feature selection and our approach is independent of visual features. We simply decompose images into a set of 32×32 blocks, 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 Distance) and 12 texture features (Gabor energy computed over 3scales and 4 orientations).

4.1. Results of Automatic Image Annotation

In this section, the performance of our model (GM-PLSA) is compared with several previous models — the Translation Model [5], CMRM [8], CRM [9], MBRM [6] and PLSA-WORDS [11]. Similarly to [9], 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 our model performs much better than most models. Besides, the model performs slightly better than MBRM. We believe that using continuous PLSA and standard PLSA to model visual and textual data respectively is the reason for this result.

Several examples of annotation obtained by our prototype system are shown in Figure 3. 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>
<thead>
<tr>
<th>Models</th>
<th>Translation</th>
<th>CMRM</th>
<th>CRM</th>
<th>MBRM</th>
<th>PLSA-WORDS</th>
<th>GM-PLSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>#words with recall &gt; 0</td>
<td>49</td>
<td>66</td>
<td>107</td>
<td>122</td>
<td>105</td>
<td>125</td>
</tr>
<tr>
<td>Mean Per-word Recall</td>
<td>0.34</td>
<td>0.48</td>
<td>0.70</td>
<td>0.78</td>
<td>0.71</td>
<td>0.79</td>
</tr>
<tr>
<td>Mean Per-word Precision</td>
<td>0.20</td>
<td>0.40</td>
<td>0.59</td>
<td>0.74</td>
<td>0.56</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Results on all 260 words

| Mean Per-word Recall | 0.04 | 0.09 | 0.19 | 0.25 | 0.20 | 0.25 |
| Mean Per-word Precision | 0.06 | 0.10 | 0.16 | 0.24 | 0.14 | 0.26 |
4.2. 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.

The annotation results ignore rank order. However, 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 2 shows that the retrieval performance of GM-PLSA is better than that of other models.

Table 2. Comparison of ranked retrieval results

<table>
<thead>
<tr>
<th>Models</th>
<th>All 260 words</th>
<th>Words with recall ≥ 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMRM</td>
<td>0.17</td>
<td>0.20</td>
</tr>
<tr>
<td>CRM</td>
<td>0.24</td>
<td>0.27</td>
</tr>
<tr>
<td>MBRM</td>
<td>0.30</td>
<td>0.35</td>
</tr>
<tr>
<td>PLSA-WORDS</td>
<td>0.22</td>
<td>0.26</td>
</tr>
<tr>
<td>GM-PLSA</td>
<td>0.32</td>
<td>0.37</td>
</tr>
</tbody>
</table>

In summary, the experiment results show that GM-PLSA outperforms several previous models in many respects, which proves that the continuous PLSA is effective in modeling visual features.

5. CONCLUSIONS

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 semantic annotation model, which employ continuous PLSA and standard PLSA to deal with the visual and textual data respectively. Experiments on the Corel dataset prove that our approach is promising for automatic image annotation. In comparison to previous proposed annotation methods, higher accuracy and superior effectiveness of our approach are reported.

6. ACKNOWLEDGEMENTS

This work is supported by the National Basic Research Priorities Programme (No.2007CB311004), National Science Foundation of China (No. 60933004, No. 60903141, No. 60805041).

7. REFERENCES