Fusing semantic aspects for image annotation and retrieval

Zhixin Li\(^{a,b,*}\), Zhiping Shi\(^{a,c}\), Xi Liu\(^{a}\), Zhiqing Li\(^{a}\), Zhongzhi Shi\(^{a}\)

\(^{a}\) Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
\(^{b}\) College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, China
\(^{c}\) Information Engineering College, Capital Normal University, Beijing 100048, China

**Abstract**

In this paper, we present an approach based on probabilistic latent semantic analysis (PLSA) to achieve the task of automatic image annotation and retrieval. In order to model training data precisely, each image is represented as a bag of visual words. Then a probabilistic framework is designed to capture semantic aspects from visual and textual modalities, respectively. Furthermore, an adaptive asymmetric learning algorithm is proposed to fuse these aspects. For each image document, the aspect distributions of different modalities are fused by multiplying different weights, which are determined by the visual representations of images. Consequently, the probabilistic framework can predict semantic annotation precisely for unseen images because it associates visual and textual modalities properly. We compare our approach with several state-of-the-art approaches on a standard Corel dataset. The experimental results show that our approach performs more effectively and accurately.

**1. Introduction**

With the development of digital imaging and data storage, searching and indexing large image databases efficiently and effectively has become a challenging problem. In order to solve the problem, there exist two distinct approaches in the literature. One solution is to annotate each image manually with keywords or captions and then search images using a conventional text search engine. This technique uses text to capture semantic content of images and allows query by text. However, expensive labor makes this solution difficult to be extended to large image databases. The other solution is to query by visual example. 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 retrieval results are usually semantically irrelevant to the query example due to the notorious semantic gap [1].

As a result, automatic image annotation has emerged as a striking and crucial problem for semantic image retrieval [2].

As a latent aspect model, PLSA has been applied in many research areas of computer vision, such as object recognition and scene classification. Furthermore, many approaches based on PLSA successfully achieve the task of automatic image annotation [3–5]. However, most of these approaches learn the latent space from either two modalities (visual features or textual words) equivalently or one modality only. In this paper, we present an extended model to fuse aspects learned from both visual and textual modalities asymmetrically. In addition, an adaptive learning approach is proposed to fit the model. From the theoretical perspective, the proposed probabilistic framework is similar to PLSA-WORDS [3]. The proposed learning approach, however, is quite different from theirs. First, when constructing latent space, PLSA-WORDS uses aspects of one PLSA model to learn the semantic information from textual modality, while our approach employs two sets of aspects to learn the semantic information from visual and textual modalities respectively. More important, the learning process of PLSA-WORDS is relatively static. In contrast, our approach learns semantic information in an adaptive mode. That is, it fuses two sets of aspects with different weights which are determined by the visual representations of images in training data set.

The main contributions of this work are the following. Firstly, we present a probabilistic framework based on PLSA to achieve the task of automatic image annotation and retrieval. The framework constructs the latent space by using two PLSA models to capture semantic aspects from visual and textual modalities respectively. Secondly, an adaptive asymmetric learning algorithm is proposed to fuse the
aspects of these two models. For each image document, the aspect distributions of different modalities are fused by multiplying different weights. Finally, we test the performance of our approach using a standard Corel dataset which consists of 5000 images. In comparison with several state-of-the-art approaches, our approach achieves higher annotation accuracy and superior retrieval effect.

The rest of the paper is organized as follows: Section 2 discusses the related work. Section 3 presents PLSA model and its principles. Section 4 describes image representation and proposes the adaptive asymmetric approach to learn the correlation between visual features and textual words. Furthermore, this section gives the training, annotating and retrieving algorithms. Experimental results are reported and analyzed in Section 5. Finally, the overall conclusions of this work are presented in Section 6.

2. Related work

Various approaches have been proposed for semantic image annotation and retrieval. The state-of-the-art techniques can be roughly categorized into two different schools of thought.

The first one defines auto-annotation as a traditional supervised classification problem, which treats each word (or semantic category) as an independent class and creates a different class model for every word (or semantic category). This approach separates the textual components from the visual components, computing similarity at the visual level. Then it annotates a new image by propagating the corresponding class weights. A representative work is automatic linguistic indexing of pictures (ALIP) proposed by Li and Wang [6]. ALIP uses two-dimensional multiresolution 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. [7] is based on binary classifiers (RPMs and SVMs) trained for each word and it indexes a new image with the output of each classifier. Caneiro et al. [8] 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.

The second perspective takes a different stand and treats images and texts 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 and posing annotation as statistical inference in a graphical model. Mori et al. [9] propose co-occurrence model which collects the co-occurrence counts between words and features and uses them to predict annotated words for unseen images. Duygulu et al. [10] improve the co-occurrence model by utilizing 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. [11] 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. [12] employ correspondence latent Dirichlet allocation (LDA) model [13] 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. Afterwards, Monay et al. [3] 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. In addition to this, Jeon et al. [14] propose cross-media relevance models (CMRM) to annotate image, assuming that the blobs and words are mutually independent given a specific image. Lavrenko et al. [15] 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. [16] propose multiple Bernouli relevance model (MBRM), in which a multiple Bernouli distribution is used to generate words instead of the multinomial one as in CRM.

3. PLSA model

Although the LDA model [13] has been shown to improve over PLSA [17] in terms of perplexity in text collections, we still choose PLSA to construct our model for two main reasons. First, PLSA allows for an exact EM algorithm. This makes the intended modifications of learning procedure easier. Second, PLSA has been shown to perform well on image classification tasks [18,19], using the aspect mixture proportions to learn the classifiers.

PLSA [17] is a statistical latent aspect model for co-occurrence data which associates an unobserved class variable with each observation. The model can be fitted to a training set through an Expectation–Maximization (EM) based iterative algorithm.

3.1. The aspect model

PLSA model introduces a hidden variable $z_k (k \in 1,\ldots,K)$ in the generative process of each element $x_j (j \in 1,\ldots,M)$ in a document $d_i$ $(i \in 1,\ldots,N)$. Given this unobservable variable (latent aspect) $z_k$, each occurrence $x_j$ is independent of the document it belongs to, which corresponds to the following joint probability: $P(d_i, z_k, x_j) = P(d_i)P(z_k|d_i)P(x_j|z_k)$. The joint probability of the observed variables is obtained by marginalizing over the latent aspect $z_k$:

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

A representation of the aspect model in terms of a graphical model is depicted in Fig. 1(a). Since the cardinality of the latent aspects is typically smaller than the number of documents (and elements) in the collection, $K \ll \min(N,M)$, it acts as a bottleneck variable in predicting words. The model (1) expresses each document as a convex combination of $K$ aspect vectors. This amounts to matrix decomposition as shown in Fig. 1(b). Essentially, each document is modeled as a mixture of aspects — the histogram for a particular document.

![Graphical model representation of PLSA. (b) Matrix decomposition of conditional distribution.](image-url)
being composed of a mixture of the histograms corresponding to each aspect.

The model parameters of PLSA are the two conditional distributions: \( P(x_j|z_k) \) and \( P(z_k|d) \), which are estimated by an EM algorithm on a set of training documents. \( P(x_j|z_k) \) characterizes each aspect and remains valid for documents out of the training set. On the other hand, \( P(z_k|d) \) is only relative to the specific documents and cannot carry any prior information to an unseen document.

3.2. Model fitting with the EM algorithm

An EM algorithm is used to compute the parameters \( P(x_j|z_k) \) and \( P(z_k|d) \) through maximizing the log-likelihood of the observed data

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

(2)

where \( n(d_i, x_j) \) is the count of element \( x_j \) in document \( d_i \). The steps of the EM algorithm are described as follows [17]:

1. **E-step.** The conditional distribution \( P(z_k|d_i, x_j) \) is computed from the previous estimate of the parameters:

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

   (3)

2. **M-step.** The parameters \( P(x_j|z_k) \) and \( P(z_k|d) \) are updated with the new expected values \( P(z_k|d_i, x_j) \):

   \[
P(x_j|z_k) = \frac{\sum_{i=1}^{N}n(d_i, x_j)P(z_k|d_i, x_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N}n(d_i, x_m)P(z_k|d_i, x_m)}
\]

   (4)

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

   (5)

As for the two parameters, if one parameter \( P(x_j|z_k) \) (or \( P(z_k|d) \)) is known, we could quickly infer the other one \( P(z_k|d) \) (or \( P(x_j|z_k) \)) using folding-in method — a partial version of the EM algorithm. The method 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. Modeling annotated images by adaptive asymmetric learning

We discuss here an approach to annotate and retrieve images by learning correlations between visual features and textual words with PLSA model.

4.1. Image representation and preprocessing

A large amount of visual features have been employed in existing image auto-annotation approaches. However, most approaches use either global features [6,7] or regional features [10,14] exclusively. Since each of these two feature representations provides different kinds of information, they have their own advantages in classifying some certain categories. On the other hand, there are many situations where the annotation of images should be judged based on the combination of global and regional features. Thus, we believe that the combination of these features should be beneficial in annotating images.

In order to describe image content accurately, each image is processed and represented as a bag of words (BOW), which is originally proposed in natural language processing and has been successfully applied in many research areas of computer vision [20]. BOW is used to combine the features into one uniform representation so that each image could be considered as a concatenating histogram and each feature could make its contributions to the description of image content. Assuming that a set of features are quantized into \( N_w \) visual words, then the image document \( d \) can be represented as a histogram \( v(d) \) of size \( N_w \):

\[
v(d) = \{n(d, w_1), \ldots, n(d, w_i), \ldots, n(d, w_{N_w})\},
\]

(6)

where \( n(d, w_i) \) denotes the number of the visual word \( w_i \) in the image \( d \). Moreover, the representation of the caption of an image \( d \) is relatively straightforward for caption itself is discrete data. If the size of the vocabulary is \( N_w \), the caption can be represented as a histogram \( w(d) \) of size \( N_w \):

\[
w(d) = \{n(d, w_1), \ldots, n(d, w_i), \ldots, n(d, w_{N_w})\},
\]

(7)

where \( n(d, w_i) \) denotes the count of the word \( w_i \) in the caption of the image \( d \).

In our work, several visual features are employed and integrated into BOW model. First, SIFT keypoints [21] are used to identify objects in images. Each image is sampled with the difference-of-Gaussians (DOG) point detector at different scales and locations. Then each detected region is represented by SIFT descriptors, which are based on histograms of local orientation. These SIFT descriptors are quantized by the K-means clustering algorithm to obtain a discrete set. Second, we impose a fixed-size rectangular grid on each image to divide it into blocks. HSV histogram [22] and Local Binary Patterns (LBP) [23] are used to describe the features of each block. Again, K-means algorithm is used to quantize these blocks into clusters so as to get a corresponding discrete representation. Third, just like [10], each image is segmented into several regions and clustered into blobs. Although automatic segmentation algorithm cannot always segment image into semantically-coherent regions, the blobs are still expected to represent corresponding semantic objects to some extent.

After the preprocessing procedure, all kinds of quantized features can be combined by simply concatenating them. More important, this representation can accept or reject specific visual features flexibly, which provides great convenience for our experiment. As a result, every image can be seen as a bag of visual words which consist of many representative visual features. Particularly, every annotated image can be seen as a mixture of a bag of visual words and textual words.

4.2. Fusing semantic aspects by adaptive asymmetric learning

The basic principle of annotating images with PLSA can be described as follows. In learning stage, aspect distribution \( P(\omega|d) \) of each training image is learned by various approach. Given each semantic aspect, distribution over visual words \( P(v|\omega) \) and textual words \( P(w|\omega) \) can be learned in terms of the aspect distribution. Because of the independent assumption of PLSA, these two parameters are independent of training images and remain valid for images out of the training set. In annotation stage, given an unseen image \( d_{new} \), its visual representation \( v(d_{new}) \) can be gotten automatically and the aspect distribution \( P(\omega|d_{new}) \) can be inferred by folding-in method. Then, posterior probability \( P(\omega|d_{new}) \) is computed by multiplication rule. As a result, the unseen image can be annotated with several words of largest posterior probability.

In order to annotate images automatically, it is important to learn the correlation between visual and textual modalities. However, most existing approaches treat visual and textual modalities equivalently. In particular, PLSA-WORDS and PLSA-FEATURES claim that they use asymmetric learning approaches[3]. In fact, these two approaches learn the aspect distribution \( P(\omega|d) \) from one modality only (corresponding to words or features respectively). In other words, they construct the latent space with statistical data obtained from one modality and then link the space to
another modality. Consequently, the correlation between the modalities can be learned through the latent space.

Through a number of experiments, we find that if the distribution over visual words is demonstrated to be sparse and with high kurtosis, the aspect distribution learned from the visual modality is more reliable and can bring higher performance. Correspondingly, more weight should be assigned to the visual modality, and vice versa. Thus, we adopt the entropy to measure the sparseness of the distribution over visual words of each image and determine the weight of the visual modality.

In order to take full advantage of the potential of both modalities, we employ a probabilistic framework with two linked PLSA models sharing the same latent space, which is learned adaptively and asymmetrically from both modalities. To be specific, we fuse semantic aspects learned from both visual features and textual words in an adaptive mode. For a training image document, the aspect distribution \( P(z_k|d) \) is learned from the two modalities with different weights. Further, the weight of each modality is determined by its contribution to the content of image, which is measured by the entropy of its distribution over visual words. In summary, our approach constructs the latent space by fusing the two sets of aspects adaptively. The process of adaptive asymmetric learning is shown in Fig. 2.

Assuming that the two PLSA models corresponding to visual and textual modalities have \( m \) and \( n \) aspects respectively, the fusing model has \( m + n \) aspects. We use \( P_s(s|d) \) and \( P_t(t|d) \) to denote the distribution over aspects, with symbols \( s \) and \( t \) to denote aspects of the two models. After obtaining the parameters \( P_s(s|d) \) and \( P_t(t|d) \) of each image through fitting these two PLSA models, the distribution \( P(z_k|d) \) is fused through

\[
P(z_k|d) = \begin{cases} 
\alpha_v P_v(s_k|d), & k = 1, \ldots, m, \\
\alpha_w P_w(t_k|d), & k = m + 1, \ldots, m + n,
\end{cases}
\]

where \( \alpha_v \) and \( \alpha_w \) are weights of visual and textual modalities. They are computed by empirical formula

\[
\alpha_v = \begin{cases} 
1, & H(v(d_i)) \leq 3, \\
\exp(3 - H(v(d_i))), & H(v(d_i)) > 3,
\end{cases}
\alpha_w = 1 - \alpha_v,
\]

where \( H(v(d_i)) \) is the entropy of the visual representation \( v(d_i) \) of a specific image document \( d_i \).

The adaptive asymmetric learning gives a better control of the respective influence of each modality in the latent space definition. Once the distribution \( P(z_k|d) \) is determined properly, \( P(z_k|d) \) corresponding to two modalities can be inferred by folding-in method. Since \( P(z_k|d) \) of each image document is learned from two modalities in an adaptive mode, our approach, referred to as PLSA-FUSION, is believed to possess better learning and generalization ability than PLSA-WORDS and PLSA-FEATURES.

4.3. Algorithm description

In this section, we describe three algorithms used in our work, namely, training, annotation, and retrieval. In the following description, the element \( x \) in some cases may be substituted by \( v \) or \( w \), which correspond to the visual and textual elements, respectively.

For the training algorithm, we assume a training set \( D = \{ (d_1, c_1), \ldots, (d_N, c_N) \} \) of image-caption pairs, where image \( d_i \in \mathcal{D} \) with \( \mathcal{D} = \{ d_1, \ldots, d_N \} \), and caption \( c_i \in \mathcal{L} \) with \( \mathcal{L} = \{ w_1, \ldots, w_T \} \), which is usually referred to as semantic vocabulary. The following gives a detailed illustration of the training algorithm:

1. For each image \( d_i \in \mathcal{D} \), visual features are extracted and quantized into visual representation \( v(d_i) \); The caption linked to the image are processed and represented as \( w(d_i) \).
2. Both visual and textual representations are used to estimate the conditional distribution \( P(v(z)|d) \) and \( P(z|d) \) by fitting PLSA model. Consequently, we can get two sets of parameters: \( P_s(s|d) \), \( P_v(v|d) \) and \( P_w(w|d) \).
3. Fusion parameters \( \alpha_v \) and \( \alpha_w \) are introduced to evaluate the importance of the visual and textual modality. These parameters are computed by Eq. (9). For each image document \( d_i \), aspect distribution \( P(z_k|d_i) \) is computed by Eq. (8).
4. Using folding-in method, we can get the final training results \( P(v(z_i)) \) and \( P(w(z_i)) \), which remain valid in images out of the training set.

The annotation algorithm processes test images \( d_{\text{new}} \notin \mathcal{D} \) according to the steps as follows:

1. For each test image \( d_{\text{new}} \notin \mathcal{D} \), the algorithm executes step (1) of the training algorithm.
2. Given visual representation \( v(d_{\text{new}}) \) and the previously estimated parameters \( P(v(z_i)) \), the conditional probability distribution \( P(z|d_{\text{new}}) \) can be inferred using folding-in method.
3. The posterior probability of each word in the vocabulary \( \mathcal{L} \) is then computed by

\[
P(w(d_{\text{new}})) = \sum_{z_{\text{new}}} P(w|z_i) P(z_i|d_{\text{new}}).
\]

4. Annotate the test image with several words of largest posterior probability \( P(w(d_{\text{new}})) \).

Finally, The retrieval algorithm takes as inputs a semantic word \( w \) and a database of test images \( \mathcal{T} \), such that \( \mathcal{T} \cap \mathcal{D} = \emptyset \). It consists of the following steps:

1. For each image \( d_i \in \mathcal{T} \), perform steps (1)–(4) of the annotation algorithm.
2. Rank the images labeled with the query word by decreasing posterior probability \( P(w(d_i)) \).

By use of these three algorithms, representative visual information, contextual information and correlation between images and words are integrated with PLSA. Therefore, the task of image annotation and retrieval can be achieved effectively and accurately.

5. Experimental results

We have implemented our approach in a prototype system. The process of PLSA model fitting is executed offline; the task of semantic image annotation and retrieval is performed online.

5.1. Dataset and evaluation measures

In order to test the effectiveness of the prototype system, we conduct our experiments on the annotated image data set which was originally used in [10]. The dataset, referred to as Corel5K, consists of 5000 images from 50 Corel Stock Photo cds. Each cd includes 100 images on the same topic. We divided this dataset
into three 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 [10]. Furthermore, only the words (260 in total) used as annotations for at least eight images were selected into vocabulary.

Image annotation performance is evaluated by comparing the captions automatically generated for the test set with the human-produced ground truth. Similar to [16], we define the automatic annotation as the five semantic words of largest posterior probability, and compute the recall and precision of every word in the test set. For a given semantic word, recall = B/C and precision = B/A, where A is the number of images automatically annotated with a given word; B is the number of images correctly annotated with that word; C is the number of images having that word in ground truth annotation. The average word precision and word recall values summarize the system performance. Besides, the performance of semantic retrieval is also evaluated by measuring precision and recall.

These measures, however, are not good enough to evaluate the retrieval performance comprehensively. So we use another important metric called mean average precision (mAP), which has been a standard measure for the retrieval of text document for years. mAP has the ability to summarize the retrieval performance in a meaningful way. To compute it, the AP of a query q is first defined as the sum of the precisions of the correctly retrieved images at rank i divided by the total number of relevant images rel(q) for this query,

\[
AP(q) = \frac{\sum_{i=1}^{N_{rel}} \text{precision}(i)}{N_{rel}}.
\]

The AP measure of a query is thus sensitive to the entire ranking of documents. The mean of the AP of \(N_q\) queries summarizes the performance of a retrieval system in one mAP value,

\[
mAP = \frac{\sum_{q=1}^{N_q} AP(q)}{N_q}.
\]

### 5.2. Parameters setting

There are two types of parameters needed to be set. The first is the number of K-means clusters that defines the quantization of the visual features. This number affects the performance for it determines the granularity of BOW. According to experiment experience, we set 500, 700 and 800 as the cluster number of SIFT descriptor, block-base global features and blobs, respectively. Then each image can be represented as a vector of 2000 dimensions, that is, a histogram with 2000 bins. Each bin counts the number of visual element \(v_i\) in image document \(d_i\). The second type of parameter 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. Through series of experiments on the validation set, we choose 200 as aspect number for PLSA-WORDS. As for PLSA-FUSION, we set 120 aspects to learn textual modality and 80 aspects to learn visual modality.

PLSA is known for overfitting data. We control the overfitting of our model by early stopping, in which we do not necessarily achieve the optimization until convergence, but instead stop updating the parameters once the performance on hold-out data is not improving. Our algorithm stops the iterative process based on the likelihood of a validation set. We consider the folding-in-likelihood, which allows good performance prediction and overfitting control without the need for a tempered version of the EM algorithm [24]. The folding-in likelihood of the validation set is defined as

\[
L_{\text{valid}}(\theta) = \prod_{i=1}^{N_{\text{valid}}} \prod_{j=1}^{M} \sum_{k=1}^{K} P(z_k|d_i) P(x_j|z_k).
\]

For PLSA, the EM algorithm is initially randomly and typically converges in 40–100 iterations, while the folding-in method often converges in less than 25 iterations.

### 5.3. Results of automatic image annotation

In this section, we compare the automatic image annotation performance of several models — the Co-occurrence Model [9], the Translation Model [10], CMRM [14], CRM [15], PLSA-WORDS [3] and the model proposed in this paper (PLSA-FUSION). We evaluate the performance of image annotation by comparing the captions automatically generated with the original manual annotations. Similar to [15], we compute the recall and precision of every

![Fig. 3. Precision-recall curves of PLSA-WORDS and PLSA-FUSION for automatic image annotation.](image)

<table>
<thead>
<tr>
<th>Models</th>
<th>Co-occurrence</th>
<th>Translation</th>
<th>CMRM</th>
<th>CRM</th>
<th>PLSA-WORDS</th>
<th>PLSA-FUSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Words with recall &gt; 0</td>
<td>19</td>
<td>49</td>
<td>66</td>
<td>107</td>
<td>105</td>
<td>112</td>
</tr>
<tr>
<td>Results on 49 best words, as in [9,10,14,15]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean per-word recall</td>
<td>–</td>
<td>0.34</td>
<td>0.48</td>
<td>0.70</td>
<td>0.71</td>
<td>0.76</td>
</tr>
<tr>
<td>Mean per-word precision</td>
<td>–</td>
<td>0.20</td>
<td>0.40</td>
<td>0.59</td>
<td>0.56</td>
<td>0.65</td>
</tr>
<tr>
<td>Results on all 260 words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean per-word recall</td>
<td>0.02</td>
<td>0.04</td>
<td>0.09</td>
<td>0.19</td>
<td>0.20</td>
<td>0.22</td>
</tr>
<tr>
<td>Mean per-word precision</td>
<td>0.03</td>
<td>0.06</td>
<td>0.10</td>
<td>0.16</td>
<td>0.14</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Table 1: Performance comparison on the task of automatic image annotation.
word in the test set and use the mean of these values to summarize
the system performance.

In addition, we report the results on two sets of words: the sub-
set of 49 best words which was used by [10,14,15], and the com-
plete set of all 260 words that occur in the training set. Table 1
shows the performance on both word sets. From the table, we
can see that our model performs much better than the first three
approaches and slightly better than CRM and PLSA-WORDS. We
believe that the enhanced performance is attributed to the integra-
tion of various visual features and the incorporation of adaptive
asymmetric learning.

Fig. 3 presents the precision–recall curves of PLSA-WORDS and
PLSA-FUSION on the Corel5k dataset, with the number of annota-
tions from 2 to 10. The precision and recall values are the mean
values calculated over all words. From the figure we can see that
PLSA-FUSION consistently outperforms PLSA-WORDS.

Fig. 4 shows several examples of annotation obtained by our
prototype system. Top five words are taken as annotation of the
image here. The figure shows that, for some images with sparse vi-
sual representation (such as the second image), annotation results

<table>
<thead>
<tr>
<th>Image</th>
<th>Ground Truth</th>
<th>PLSA-WORDS</th>
<th>PLSA-FUSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>polar, bear, snow, tundra</td>
<td>waved, albatross, flight, sky</td>
<td>water, boats, village, harbor</td>
<td>zebra, grass, planes, profile</td>
</tr>
<tr>
<td>city, flight, ceremony, pond, swallow-tailed</td>
<td>water, beach, boats, harbor, skyline</td>
<td>grass, zebra, planes, herd, cat</td>
<td></td>
</tr>
<tr>
<td>flight, bird, sky, waved, albatross</td>
<td>water, boats, harbor, sky, beach</td>
<td>grass, zebra, planes, herd, trees</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 4. Comparison of annotations made by PLSA-WORDS and PLSA-FUSION.

<table>
<thead>
<tr>
<th>Image</th>
<th>Ground Truth</th>
<th>PLSA-WORDS</th>
<th>PLSA-FUSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>garden, flowers, landscape, trees</td>
<td>pyramids, stone, people, camels</td>
<td>mountain, water, sky, clouds</td>
<td>beach, people, water, sky</td>
</tr>
<tr>
<td>flowers, garden, farm, trees, bench</td>
<td>stone, pyramids, mountain, columns, range</td>
<td>mountain, clouds, boat, coast, hut</td>
<td>sky, beach, snow, sand, mountain</td>
</tr>
<tr>
<td>flowers, garden, grass, trees, farm</td>
<td>stone, pyramids, sand, sky, antelope</td>
<td>sky, mountain, water, clouds, trees</td>
<td>water, sky, beach, wave, snow</td>
</tr>
</tbody>
</table>

Table 2
Comparison of ranked retrieval results.

<table>
<thead>
<tr>
<th>Models</th>
<th>Mean average precision for corel dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All 260 words</td>
</tr>
<tr>
<td>CMRM</td>
<td>0.1697</td>
</tr>
<tr>
<td>CRM</td>
<td>0.2353</td>
</tr>
<tr>
<td>PLSA-WORDS</td>
<td>0.2213</td>
</tr>
<tr>
<td>PLSA-FUSION</td>
<td>0.2578</td>
</tr>
</tbody>
</table>

Fig. 5. Average precision of selected 10 words when annotating with PLSA-WORDS and PLSA-FUSION.
of PLSA-FUSION are much better than that of PLSA-WORDS. In addition, we can see that even if the system annotates an image with a word not contained in the ground truth, this annotation is frequently plausible (such as the annotation “trees” in the fourth image and the annotation “sand” in the sixth image).

5.4. Results of ranked image retrieval

We use mAP as a metric to evaluate the performance of ranked retrieval with single word queries. This section only compares our model (PLSA-FUSION) with CMRM, CRM and PLSA-WORDS, because mAPs of other models 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 ranked according to the posterior probabilities of that word. Table 2 compares the mAP of PLSA-FUSION with several models and we can see that PLSA-FUSION is better than other models.

Fig. 5 compares the average precision of selected 10 words when annotating with PLSA-WORDS and PLSA-FUSION. As shown in the figure, no matter a word is learned well or bad, its average precision of PLSA-FUSION is mostly higher than that of PLSA-WORDS.

Fig. 6 presents three ranked retrieval results obtained with single word queries for challenging concepts. The diversity of visual appearance of the returned images indicates that our model has good generalization ability.

In summary, the experimental results show that our model performs better than several state-of-the-art approaches in many respects, which proves that the adaptive asymmetric learning approach is feasible and effective. However, there exists two limitations in our approach. First, since PLSA-FUSION has to model data of two modalities and fuse the aspects, it requires more time to learn the model parameters. Second, PLSA-FUSION use discrete visual words to represent a image. Therefore, its annotation quality is sensitive to clustering granularity.

6. Conclusions

In this paper, we have developed a PLSA-based automatic image annotation system, which uses two linked PLSA models to learn the mixture of aspects from both the visual and the textual modalities. Furthermore, in order to handle these two modalities, an adaptive asymmetric learning approach is proposed and put into practice. Our system is promising for the task of semantic image annotation and retrieval, which is experimented on a standard Corel dataset. In comparison with previous proposed annotation approaches, higher accuracy and superior effectiveness of our approach are reported.

Acknowledgments

This work is supported by the National Basic Research Priorities Programme (No. 2007CB311004), the National Science and Technology Support Plan (No. 2006BA088006) and the National Natural Science Foundation of China (Nos. 60775035, 60933004, 60903141, 60970088, 60903079).

References


