Translating Topics to Words for Image Annotation

Yong Wang
Department of Computer Science
Queen Mary, University of London
Mile End Road, London, UK, E1 4NS
ywang@dcs.qmul.ac.uk

Shaogang Gong
Department of Computer Science
Queen Mary, University of London
Mile End Road, London, UK, E1 4NS
sgg@dcs.qmul.ac.uk

ABSTRACT

One of the classic techniques for image annotation is the language translation model. It views an image as a document, i.e., a set of visual words which are obtained by vector quatitizing the image regions generated by unsupervised image segmentation. Annotating images are achieved by translating visual words to textual words, just like translating a document in English to a document in French. In this paper, we also view an image as a document, but we view the annotation processes as two consecutive processes, i.e., document summarization and translation. In the document summarization process, an image document is firstly summarized into its own visual language, which we called *visual topics*. The translation process translates these visual topics to textual words. Compared to the original translation model, our visual topics learned by the probabilistic latent semantic analysis (PLSA) approach provide an intermediate abstract level of visual description. We show improved annotation performance on the Corel image dataset.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Algorithms, Measurement, Experimentation

Keywords

Automatic image annotation, translation model, probabilistic latent semantic analysis

1. INTRODUCTION

Image annotation is an important and promising research topic in image retrieval. Most of the existing working systems of web image retrieval such as Google Images and Yahoo! Images are implemented by annotating images firstly and then searching image based on the textual keywords. In such kind of systems, the annotations are obtained from the text information such as the surrounding texts, captions and URLs. However, there exist a lot of non-web images such as personal photos without any associated texts, and even for web images, the annotations by Google and Yahoo! are very noisy. Thus, automatic image annotation can be very helpful for computer users to organize and retrieve images. One of the classic techniques for image annotation is the language translation model [5]. It regards an image as a document in its visual language. The visual words are obtained by vector quatitizing image regions, which is obtained by unsupervised image segmentation. Annotating images with a set of words is viewed as translating a document in one language (visual language) to a document in another (text language). However, textual words are very concise and abstract in their meaning, and in many cases, the image regions obtained by image segmentation do not make any good sense in semantics. This makes the correspondence between an image region and a textual word not very meaningful. Motivated by these observations, we propose a method to annotate images based on the summarization of an image document. In our model, the annotation is viewed as *document summarization* instead of a *document* in another language. Image annota-

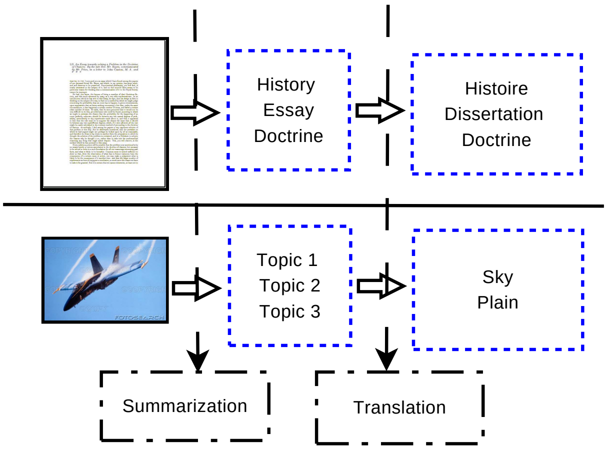

Figure 1: Annotation by summarization and translation. The upper row illustrates the process of summarizing an English document in French. The lower row illustrates the analogous process of annotating an image. Both of these two tasks have two consecutive processes: summarization and translation.

CIKM'07, November 6–8, 2007, Lisboa, Portugal.

tion is achieved by firstly summarizing the image document in its own visual language, which we call *visual topics*, and then translating these visual topics to textual words. The extracted visual topics from the low level visual features provide an equivalent abstract space which can be corresponded to the textual words. Our experimental results on the Corel image dataset show the improved annotation performance compared to the original translation model [5]. Fig. 1 illustrates the motivation of our approach. The upper row of the figure shows the process of summarizing an English document in French. The lower row shows the analogous process of annotating an image. Both of these two tasks have two consecutive processes, i.e, summarization and translation. The rest of this paper is organized as follows. Section 2 briefly review the existing work on image annotation. Section 3 introduce the details of our approach. In Section 4 shows some experimental results and we conclude this paper in Section 5.

2. RELATED WORK

Existing techniques for automatic image annotation can be summarized into two categories [1, 6, 2]. In the first category of approaches, image annotation is formulated as an image classification problem. Specifically, each concept or textual word is viewed as a unique class label. For each class, a binary classifier is trained from the training data. To annotate a new image, the trained classifiers of each classes are applied to the image and produce the most likely class labels. The advantage of the classification approach is that we have various well-studied machine learning techniques available such as Naive Bayes, Support Vector Machine (SVM), Bayes Point Machines (BPM) and the 2-D multi-resolution Hidden Markov Model (HMM) [6] etc. In the second category of approaches [5, 4, 8, 9]. Textual words associated with images are viewed as another type of features besides the visual features provided by the images. These approaches usually make an assumption of the joint distribution of the visual features and the textual features. The model parameters are learned from the training data. Image annotation is achieved by predicting the missing textual features given the observed visual features.

3. OUR APPROACH

As mentioned in Section 1, our model of image annotation has two consecutive processes, i.e., summarizing images into visual topics and translating visual topics to textual keywords. The image summarization process is achieved by making an analogy between an image and a text document and learning a number of visual topics by the probabilistic latent semantic analysis (PLSA) [3]. The translation process translates these visual topics to textual words.

3.1 PLSA

PLSA is proposed to automatically learn topics from text documents, so we describe it in the scenario of text analysis and then extend it to images. Suppose we are given a set of text documents $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, each of which is represented by a term frequency vector, i.e.,

$$d_i = [n(d_i, w_1), n(d_i, w_2), \ldots, n(d_i, w_m)] \tag{1}$$

where $n(d_i, w_j)$ is the number of occurrence of word w_j in document d_i, and m is the vocabulary size. PLSA assumes that each word in a document is generated by a specific *hidden topic* z_k, where $z_k \in \mathcal{Z}$ and $\mathcal{Z}$ is the vocabulary of hidden topics. Since z_k is a hidden variable, the conditional probability of a word w_j given document d_i is a marginalization over the topics, i.e.,

$$P(w_j|d_i) = \sum_{k}^{K} P(w_j|z_k, d_i)P(z_k|d_i) \tag{2}$$

where K is the number of hidden topics, $P(w_j|z_k, d_i)$ is the conditional probability of a word w_j given topic z_k and the document d_i, $P(z_k|d_i)$ is the conditional probability of topic z_k given d_i. Furthermore, PLSA assumes that the conditional probability of generating a word by a specific topic is independent from the document, i.e.,

$$P(w_j|z_k, d_i) = P(w_j|z_k) \tag{3}$$

So, Eq. (2) can be simplified as

$$P(w_j|d_i) = \sum_{k}^{K} P(w_j|z_k)P(z_k|d_i) \tag{4}$$

The model parameters $P(w_j|z_k)$ and $P(z_k|d_i)$ can be learned by an Expectation-Maximization (EM) algorithm [3]. Given the learned model parameters and a new document d, its topic distribution $\{P(z_k|d)\}_{k=1}^{K}$ can be estimated by an EM algorithm similar to the training process [3].

3.2 Learning Visual Topics From Images

To learn visual topics from images, we make an analogy between an image and a text document and represent an image as a bag of visual words. More specifically. each image is partitioned into a number of small patches using a regular grid. For each patch, we extract a 128-D SIFT [7] descriptor and 6-D colour descriptor separately. The colour descriptor is a concatenation of the mean and the variance values of the R, G and B channels in a patch. We consider two types of visual words, i.e., texture words and colour words. The texture words are obtained by clustering a subset of SIFT descriptors and the colour words are obtained by clustering a subset of colour descriptors. The whole visual vocabulary is obtained by the combination of texture words and colour words. It is worth noting that we do not concatenate the texture descriptor and colour descriptor into one descriptor since we intend to extract visual topics on the two types of visual words separately. We consider that the roles of texture and colour are different in discriminating different concepts. This is also to avoid the heteroscadastic problem (i.e. different variances exist in texture and colour subspace distributions).

Given the built visual vocabulary, we can transform an image into a text document by assigning a visual word label to each image patch. Then we can use PLSA to learn a number of visual topics, each of which is characterised by a multinomial distribution of visual words. After a PLSA model is learned from the training images, we can obtain the topic labeling o of a visual word v in a specific document d by the following equation

$$P(o|v, d) = \frac{P(v|o)P(o|d)}{P(v|d)} \tag{5}$$

The ending results of PLSA is that each image patch has a topic label. To differing the visual words and visual topics from the textual words and textual topics, we have used v to denote the visual words and o the visual topics in the above discussion. It is worth noting that we have learned two type of visual topics, i.e., texture topic and colour topic respectively from the corresponding visual words.

3.3 Machine Translation Model

After the learning of visual topics of each image, we train a language translation model from the training data which can map the visual topics to textual words. Although there exist several statistical machine translation models, the IBM model by Brown [5] has been proved to be more effective [5] than the others and the computation cost of this model is not very expensive, so we have only considered this model.

Suppose the number of the visual topics is m and the size of textual vocabulary is n. The set of annotated images is represented by $\mathcal{J}$. The total number of images is denoted as $|\mathcal{J}|$. The i^{th} image in $\mathcal{J}$ is denoted as J_i. J_i can be represented as a combination of visual topics and textual words, i.e.,

$$J_i = \{\vec{O}_i; \vec{W}_i\} = \{b_{i,1}, b_{i,2}, \ldots, b_{i,m}; a_{i,1}, a_{i,2}, \ldots, a_{i,n}\} \quad (6)$$

where $b_{i,j}$ is the number of times that the j^{th} visual topic o_j appearing in J_i and $a_{i,j}$ is a binary variable indicating whether the j^{th} word w_j appearing in the caption of J_i.

The essence of the translation model for image annotation is the translation probability table $\Theta = \{t_{j,k}\}$, where $t_{j,k}$ is the probability of translating the k^{th} visual topic to the j^{th} textual word. The probability of annotating J_i with $\vec{W}_i$ given $\vec{O}_i$ is:

$$p(\vec{W}_i|\vec{O}_i) = \prod_{l=1}^{n} \Big(p(w_l|\vec{O}_i) \Big)^{a_{i,l}} = \prod_{l=1}^{n} \Big(\sum_{k=1}^{m} t_{l,k} b_{i,k} \Big)^{a_{i,l}} \quad (7)$$

In the training stage, the probability table $\{t_{i,j}\}$ is obtained by maximizing the likelihood of the training set, i.e.,

$$L(\mathcal{J}) = \prod_{i=1}^{|\mathcal{J}|} L(J_i) = \prod_{i=1}^{|\mathcal{J}|} L(\vec{W}_i, \vec{O}_i) = \prod_{i=1}^{|\mathcal{J}|} p(\vec{W}_i|\vec{O}_i) p(\vec{O}_i) \quad (8)$$

this is equivalent to maximize

$$\prod_{i=1}^{|\mathcal{J}|} p(\vec{W}_i|\vec{O}_i) = \prod_{i=1}^{|\mathcal{J}|} \prod_{l=1}^{n} \Big(\sum_{k=1}^{m} t_{l,k} b_{i,k} \Big)^{a_{i,l}} \quad (9)$$

The Expectation-Maximization (EM) algorithm [5] is used to find the optimal translation probability table $\Theta^* = \{t_{i,j}\}^*$. The E-M algorithm updates iteratively the translation probabilities using the following equation:

$$t_{j,k}^{new} = \frac{1}{Z_k} \sum_i \frac{a_{i,j} b_{i,k} t_{j,k}^{old}}{\sum_l b_{i,l} t_{j,1}^{old}} \quad (10)$$

where Z_k is a normalization factor to ensure $\sum_{j=1}^{n} t_{j,k}^{new} = 1$ After the optimal model parameters $\Theta^* = \{t_{j,k}^*\}$ are obtained, the probability of annotating a test image J with a word w_j given $\vec{O}$ is,

$$P(w_j|\vec{O}; \Theta^*) \propto \sum_{k=1}^{m} t_{j,k}^* b_k \quad (11)$$

4. EXPERIMENT

The dataset in our experiment is the same subset of Corel images as used in [5]. There are 5000 colour images. Each image has $1 \sim 5$ caption words as its annotation. There are totally 374 words in the captions. We use 4500 images with annotation as the training data and the other 500 images as the test. We used a regular grid with 13×13 pixels for each patch. The vocabularies of texture words and colour words are built by running the k-means clustering on the descriptors from 200 images randomly chosen from the training data. The size of the vocabulary of texture words and colour words are both chosen as 500. Then each image can be represented as a bag of texture words and a bag of colour words. We run the learning algorithm of PLSA on the whole training data. After obtaining the parameters of PLSA, we then estimate the topic representation of each test image. The numbers of learned texture topic and colour topic are both chosen as 50 initially, so the total number of visual topics is 100. Given the visual topic representation of the training images and the ground truth annotations, the next stage is training a translation model on the whole training data to obtain the best translation probability table. The annotation of test images is achieved by the translation from the visual topics to annotation words. The top five words are selected as the final annotation. We measure the annotation performance by the averaged single query precision and recall of all the words with at least one image correctly annotated. Table 1 shows the comparison between our approach and the original translation model. The number of words with non-zero annotation precision has increased around 30%. The averaged precision and recall values over these words are also improved from 0.1988 to 0.2410 and from 0.1688 to 0.2017 respectively.

Table 1: Comparison of the performance between the original translation model and our new topic-based translation model.

	[5]	Our Approach
# of Textual Words	62	81
Ave. Recall	0.1988	0.2410
Ave. Precision	0.1688	0.2017

The number of visual topics is also important to the final annotation performance. To evaluate the effect of this quantity, we tried experiments with different number of visual topics on the same training data and the same test data. In all the settings, we have kept the number of texture topics and color topics as the same. This is not necessary but it enables us to focus on the total number of visual words. The precision and recall with different number of visual topics are shown in Table 2. From the results, we can find that when the number of visual topics is 120 both the original translation model and our approach have the best performance. However, even the best performance does not have big different from the worst, so both of our approach and the original translation model are relatively stable wrt. the number of visual topics.

Table 2: Comparison of the performance between the original translation model and our new topic-based translation model with different number of visual topics.

# of Visual Words=80		
Approaches	[5]	Our Approach
# of Textual Words	58	79
Ave. Recall	0.1859	0.2221
Ave. Precision	0.1601	0.1939
# of Visual Words=120		
Approaches	[5]	Our Approach
# of Textual Words	63	83
Ave. Recall	0.2023	0.2631
Ave. Precision	0.1769	0.2180
# of Visual Words=140		
Approaches	[5]	Our Approach
# of Textual Words	61	80
Ave. Recall	0.1961	0.2517
Ave. Precision	0.1701	0.2177
# of Visual Words=200		
Approaches	[5]	Our Approach
# of Textual Words	60	80
Ave. Recall	0.1931	0.2458
Ave. Precision	0.1672	0.1978

5. DISCUSSION

We have presented an approach for image annotation based on the language translation model. Instead of viewing the annotation words as a document in another language, we view the annotation words as a document summarization in another language. To annotate an image we firstly summarize the image document in its own visual language, i.e., visual topics, and then translate these visual topics into textual words. A visual topic is a kind of intermediate feature capturing the co-occurrence relationship between different visual words. Our experiments on the Corel dataset has demonstrated that our approach outperforms the original translation model. More broadly, the idea of summarizing image documents as visual topics can be easily exploited further to other machine learning models for image annotation such as the cross media relevance model (CMRM) [4].

6. REFERENCES

[1] E. Chang, K. Goh, G. Sychay, and G. Wu. CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines. *CSVT, IEEE Transactions on*, pages 26–38, Janary 2003.

[2] G. Carneiro, and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *Proceedings of IEEE CVPR*, June 2005.

[3] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001.

[4] J. Jeon et al. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of ACM SIGIR*, pages 119–126, 2003.

[5] Kobus Barnard et al. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003.

[6] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1075–1088, 2003.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[8] S. L. Feng et al. Multiple bernoulli relevance models for image and video annotation. In *Proceedings of IEEE CVPR*, June 2004.

[9] V. Lavrenko, R. Manmatha and J. Jeon. A model for learning the semantics of pictures. In *Proceedings of NIPS*, 2003.