Image Annotation by Concept Level Search Using PLSA

Yu ZHENG†, Tetsuya TAKIGUCHI††, and Yasuo ARIKI††

† Graduate School of Engineering, Kobe University 1-1, Rokkodai, Nada, Kobe, 657-8501 Japan
†† Organization of Advanced Science and Technology, Kobe University 1-1, Rokkodai, Nada, Kobe, 657-8501 Japan
E-mail: †teiyiku@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract Digital cameras have made it much easier to take photos, but organizing those photos is difficult. As a result, many people have thousands of photos in some miscellaneous folder on their hard disk . If computer can understand and manage these photos for us, we can save time. Also it will be useful for indexing and searching the web images. In this paper we propose an image annotation system with concept level search using PLSA,which generates the appropriate keywords to annotate the query image using large-scale image database.
Key words image annotation , PLSA , image recognition

1. Introduction

Image annotation has been an active research topic in recent years due to its potentially large impact on both image understanding and web image search. We target at solving the automatic image annotation in a novel search framework. Given an uncaptioned image, first in the search stage a set of visually similar images are found from a large-scale image database. The database consists of images from the World Wide Web(Flickr Group)with rich annotations and surrounding text made by user. In the mining stage, a search result clustering technique (PLSA) is utilized to find most representative keywords from the annotations of the retrieved image subset. These keywords, after ranking, are finally used to annotate the uncaptioned image.

2. Prior Work

A large number of techniques have been proposed in the last decade. Most of these deal with annotation as translation from image instances to keywords. The translation paradigm is typically based on some model of image and text co-occurrences. One of this translation model is the Correspondence Latent Dirichlet Allocation(CorrLDA),a model that finds conditional relationships between latent variable representations of sets of image regions and sets of words. Although it considers associations through a latent topic space in a generatively learned model, this class of models remains sensitive to the choice of topic model,initial parameters and prior image segmentation. MBRM shown in Fig.1 proposed approaches to automatically annotating and retrieving images by learning a statistical generative model called a relevance model using a set of annotated training images.The images are partitioned into rectangles and features are computed over these rectangles.A joint probability model for image features and words called a relevance model will be learned and is used to annotate test images which have not been seen.Words are modeled using a multiple Bernoulli process and images modeled using a kernel density estimate.However,the complexity of the kernel density representations may hinder MBRM's applicability to large data set.

Recent research efforts have focused on extensions of the translation paradigm that exploit additional structure in both visual and textual domains. For instance,[1]utilizes a coherent language model, eliminating independence between keywords. The added complexity, however, makes the models applicable only to limited settings with small-size dictionaries.[2]developed a real-time ALIPR image search engine which uses multiresolution 2D Hidden Markov Model to model concepts determined by a training set. While this method successfully infers higher level semantic concepts based on global features, identification of more specific categories and objects remains a challenge.

In this paper,we propose a method to solve the problem of the trade off between the computational efficiency with the large-scale dataset and the precision performance on complex annotation tasks.We use the concept level representation to solve the precision problem and use the Internet database with concept groups for the training data to solve the large-scale dataset

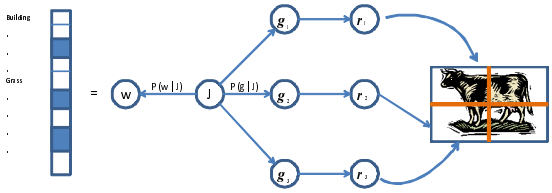

Fig. 1 MBRM.The annotation w is a binary vector.The image is produced by first sampling a set of feature vectors $g_1....g_n$, and then generating image regions $r_1....r_n$ from the feature vectors. Resulting regions are tiled to form the image.

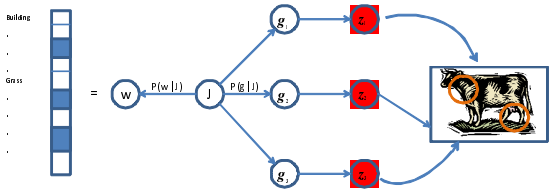

Fig. 2 Approch.The annotation w is a binary vector.The image is produced by first sampling a set of feature vectors $g_1....g_n$.The image regions $r_1.....r_n$ is replaced with concept representation $z_1.....z_n$

problem.By the experiment we can chose the best parameter, the number of topics K at PLSA model.

3. Approach

3.1 Outline

Automatically assigning keywords to images is of great interest as it allows one to index, retrieve, and understand large collections of image data. We can treat annotation as a retrieval problem. First to find nearest neighbor of a given image, the keywords are then assigned using a label transfer mechanism. Given an input image, the goal of automatic image annotation is to assign a few relevant text keywords to the image that reflect its visual content. Utilizing image content to assign a richer, more relevant set of keywords would allow one to further exploit the fast indexing and retrieval architecture of web image search engines for improved image search. This makes the problem of annotating images with relevant text keywords of immense practical interest.

Image annotation is a difficult task for two main reasons: First is the semantic gap problem, which points to the fact that it is hard to extract semantically meaningful entities using just low level image features. Doing explicit recognition of thousands of objects or classes reliably is currently an unsolved problem. The second is to find the training image set with the keywords. One of the simplest annotation schemes is to treat the problem of annotation as that of image-retrieval. For instance, given a test image, one can find its nearest neighbor from the training set, and assign all the keywords of the nearest image to the input test image. One obvious modification of this scheme would be to use K-nearest neighbors to assign the keywords instead of relying on just the nearest one. We do not only use the low level image features,but also the concept level of the images.

In our system shown in Fig.2 for image automatic annotation, a user gives query image to the system in the beginning, and obtains keywords associated with given query image finally. In this paper, we propose a new method to select relevant keywords to the given query image from images gathered from the Web. Our method is based on generative probabilistic latent topic models such as Probabilistic Latent Semantic Analysis (PLSA), Firstly, we gather images related to the given query image from the Web based on image features extracted from images themselves. Secondly, we use the gathered images for the training data in the PLSA model, and train a probabilistic latent topic model with them. Finally, we select strong relevant images from the training data images based on concept level with the learned model and extract the keywords,following the below described equations.

$$w^* = p(w|I_i) * p(I_i|I_d) \tag{1}$$

$$p(w|I_i) = \sum_{z \epsilon Z} p(w|z) * p(z|I_i) \tag{2}$$

$$p(w,d) = p(d) \sum_{z \epsilon Z} p(z|d) p(w|z) \tag{3}$$

We use the Eq.(1) to extract the keywords w^* . I_d is the query images and I_i is the extracted strong relevant images. $p(w|I_i)$ which gets the strong relevant keywords ω in I_i ,can be obtained by using the concept variables z. $p(z|d)$ in Eq.(3) can be replaced with the $p(z|I_i)$ in Eq.(2). The relationship between the above equations has been shown in Fig.3. D is the images in WWW,and d is the images related to the query image.

3.2 Training Data Search

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. A support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification,

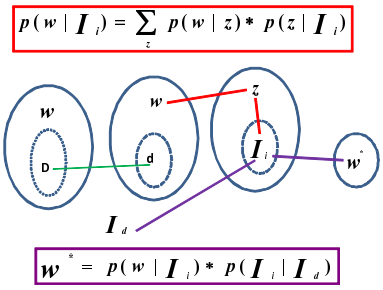

Fig. 3 The relation between the images. We can get a lot of images D with relevant keywords from the photo album site. In the different topic groups,we choose the group of images d related to the query image in the feature level. The group will be used for the training data in the next step. The keyword ω include many noisy keywords that have to be removed. We search the images strongly related to the query image based on the concept z and get the keywords w^* without noisy keywords.

regression or other tasks. Multiclass SVM aims to assign labels to instances by using support vector machines, where the labels are drawn from a finite set of several elements. The dominating approach for doing so is to reduce the single multiclass problem into multiple binary classification problems. Each of the problems yields a binary classifier, which is assumed to produce an output function that gives relatively large values for examples from the positive class and relatively small values for examples belonging to the negative class.

Fig. 4 Flickr is almost certainly the best online photo management and sharing application in the world. With millions of users, and hundreds of millions of photos and videos, Flickr is an amazing photographic community, with sharing at its heart.

To search the most similar image group for the training data,measuring image similarity became an effective way. Two images are similar if they are likely to belong to the same Flickr groups. We use SIFT as the image feature and quantize them. Using online photo sharing sites, such as Flickr be shown in Fig.4. People have organized many millions of photos into hundreds of thousands of semantically themed groups. How can we learn

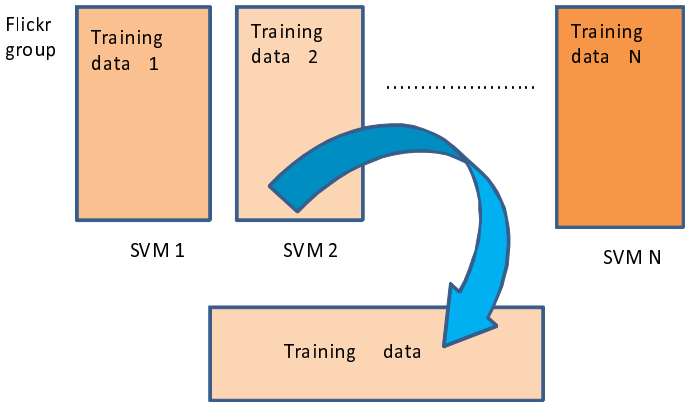

Fig. 5 Search training data with Flickr group.We download thousands of images from many Flickr groups.Groups that we use are organized by objects. For each group,we train a SVM classifier. For a test image,we use the trained group classifiers to predict likely group memberships. We train classifiers to predict whether an test image is likely to belong to a Flickr group. The group will be took out for the training data.

whether a photo is likely to belong to a particular Flickr group? we can easily download thousands of images belonging to the group and many more that do not ,and then we calculate the SIFT value of the images from the Flickr groups,finally quantize them to form the feature,suggesting that we train a classifier SVM as shown in Fig.5. For each group, we train a SVM. For a test image, we also calculate the SIFT feature of the test image and use the trained group classifiers to predict likely group memberships. We use these predictions to measure similarity,and decide which group is the test image belongs to.

3.3 Strong Relevant Images Search

Existing approaches are mainly based on global features extracted from the whole image or on fixed spatial layouts, while images of a given object are usually characterized by the presence of a limited set of specific visual parts tightly organized into different view-dependent geometrical configurations. An image is generally composed of several entities (car, house, door, tree, rocks...) organized in often unpredictable layouts. Hence, the content of images from a specific scene type exhibits a large variability. We expect that the specificity of a particular scene type greatly rests on particular co-occurrences of a large number of visual components. PLSA, an unsupervised probabilistic model for collections of discrete data, integrates the recently proposed scale-invariant feature and probabilistic latent space model frameworks,has dual ability to

generate a robust, low-dimensional representation. The bag-of-visual representation is simple to build. Recently probabilistic latent space models have been proposed to capture co-occurrence information between elements in a collection of discrete data. PLSA is a statistical model as shown in Fig.6 that associates a latent variable $z_l \in Z = \{z_1, ..., z_{N_A}\}$ with each observation (the occurrence of a word in a document). These variables, usually called aspects, are then used to build a joint probability model over images and visterms, defined as the mixture.

$$P(v_j, d_i) = P(d_i) \sum_{l=1}^{N_A} P(z_l|d_i) P(v_j|z_l) \qquad (4)$$

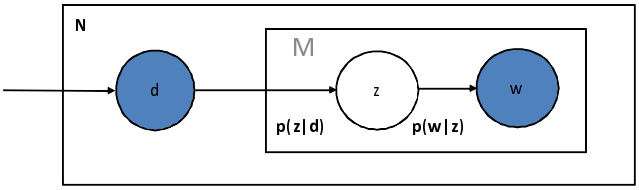

Fig. 6 Joint probability model. Plate notation representing the PLSA model. d is the document variable,z is a topic drawn from the topic distribution for this document,$p(z|d)$. w is a word drawn from the word distribution for this topic,$p(w|z)$. The d and w are observable variables,the topic z is a latent variable.

We use the PLSA to find the more stronger relevant images and the keywords included in them. PLSA introduces a conditional independence assumption: it assumes the occurrence of a visual word v_j to be independent of the image d_i it belongs to, given an aspect z_l. The model in Eq.(4) is defined by the conditional probabilities $P(v_j|z_l)$ which represent the probability of observing the visual word v_j given the aspect z_l, and by the image-specific conditional multinomial probabilities $P(z_l|d_i)$. The model expresses the conditional probabilities $P(v_j|d_i)$ as a convex combination of the aspect specific distributions $P(v_j|z_l)$.

The parameters of the model are estimated using the maximum likelihood principle, using a set of training images D. The training images have been got in the first step. The optimization is conducted using the Expectation-Maximization (EM) algorithm. This estimation procedure allows to learn the aspect distributions $P(v_j|z_l)$. These image independent parameters can then be used to infer the aspect mixture parameters $P(z_l|d)$ of any image d given its bag-of-visterms (BOV) representation. Consequently, the second image representation we will use is defined by Eq.(5). Eq.(5)is the concept level image representation.

$$(P(z_l|d))_{l=1,2.....N_A} \qquad (5)$$

Eq.(5) has the same meaning with $p(z|d)$ in (3). We can use the concept representation to search for images strongly related to the query image. Comparing to feature representation,concept representation shown in Eq.(5) can find out accurate images because it searches data based on objects in the images.

3.4 Label Transfer

We propose here a simple method to transfer n keywords to a query image $\hat{I}$ from the query 's K nearest neighbors in the training set. Let I_i …… be these K nearest neighbors, ordered by increasing distance. The number of keywords associated with I_i is denoted by $| I_i |$. following are the steps of our label transfer algorithm.

（1） Rank the keywords of I_1 according to their frequency in the training set

（2） Of the $| I_1 |$ keyword of I_1, transfer the n highest ranking keywords to query $\hat{I}$, If $| I_1 | < n$ proceed to step 3

（3） Rank the keywords of neighbors I_2 through I_k according to two factors:1)co-occurrence in the training set with the keywords transferred in step 2, and 2)local frequency(how often they appear as keywords of images I_2 through I_k. Select the highest ranking $n- | I_1 |$ keywords to transfer to $\hat{I}$.

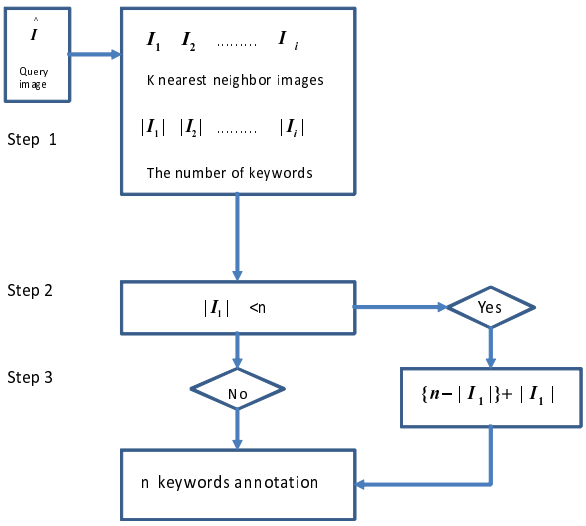

Fig. 7 Graph of the label transfer algorithm

This transfer algorithm shown in Fig.7 is somewhat different from other obvious choices. One can imagine simpler algorithm where keywords are selected simultaneously from the entire neighborhood (all the neighbors

are treated equally),or there the neighbors are weighed according to their distance from the test image. However, an initial evaluation showed that these simple approaches underperform in comparison to two-stage transfer algorithm.

4. Experiment

Predicting annotations with an unlimited vocabulary, which is a significant advantage of this annotation system benefited from Web-scale data, to get a better similarity measure to obtain a more semantically relevant image set

To obtain a well-annotated image database, we gathered 1K images from photo forum site, images in photo forums have rich and accurate descriptions provided by photographers. We used the random 1K images for the test images.

The number of topic: In the Table.1 we compared the performance of precision at concept level with different number of topics. It can be seen that the precision will change with different number of topics. Meanwhile noisy or irrelevant words resulting in some drop in precision can be improved by concept search.We chose the best parameter K which gets the highest precision.

The number of image: In the Table.1 we also change the number of test images.The performances improved when the number of images increased. This implies that more images may bring more noises, and at the concept level the noises can be reduced effectively. High precision can be achieved benefited from the large-scale data.

Table 1 Average precision with different number of topics and different number of images

topic	K=1	K=2	K=3	K=4	K=5
10	56.7	67.8	42.7	79.8	86.6
20	52.0	47.8	69.0	54.9	67.9
50	43.7	57.9	48.2	64.7	59.9
100	41.5	49.2	45.3	51.5	49.8
200	53.6	57.1	49.6	65.9	58.8
500	57.9	67.5	59.2	69.9	72.3
1000	56.8	58.3	52.4	63.1	59.0
topic	K=6	K=7	K=8	K=9	K=10
10	70.1	73.8	69.7	87.9	82.5
20	57.8	71.5	72.7	70.4	68.9
50	79.8	69.0	80.3	76.5	80.1
100	63.9	55.1	57.7	64.2	65.4
200	67.8	73.4	70.8	75.5	67.9
500	71.4	68.0	70.4	77.3	77.0
1000	72.3	74.4	79.2	76.0	68.6

As be shown in Table.2 we compared the performance on the task of automatic image annotation with different models. CRM and CRM-Rectangles are essentially the same model but the former uses regions produced by a segmentation algorithm while the latter uses a grid. We can see that when inputed 100 images,MBRM performs best. When inputed 500 or 1000 images,the concept performs best.

Table 2 Performance comparison on the task of automatic image annotation with different model.

models	Translation	CRM	CRM-Rectangles
100	34	70	75
500	20	59	72
1000	18	47	63
models	MBRM	Concept(best)	
100	78	65.4	
500	74	77.3	
1000	69	79.2	

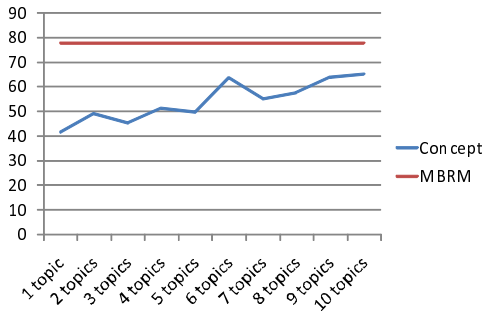

Fig. 8 100 images precision

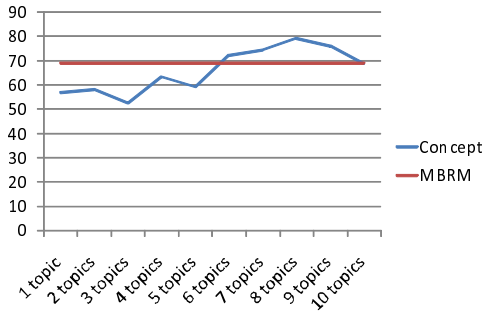

Fig. 9 1000 images precision

When 100 test images were input,the MBRM performs better than concept as shown in Fig.8. But when 1000 images were input,the best precision of the concept performs better than MBRM as shown in Fig.9.

As shown in Fig.10,the images with much prior knowledge such as building and mountain can achieve high precision. But the images with less prior knowledge such as Ferris wheel dose not perform well.

Ferris wheel Mountain Building	Lighthouse Sea cloud	Building Mountain	Tree Road lawn	Flower Snow Mountain	Trees Bicycle person

Fig. 10 Example

5. Conclusion

In this paper, we have presented a practical and effective image annotation system. We formulate the image annotation as searching for similar images and mining key phrases from the descriptions of the resultant images, based on two key techniques: image search -index and the search result clustering technique. We use these techniques to bridge the gap between the pixel representations of images and the semantic meanings. However identifying objects ,events, and activities in a scene is still a topic of intense research with limited success. In the future we will investigate how to improve the annotation quality without any prior knowledge.

References

[1] Jin,R.,Chai,J.Y.,Si,L.:Effective automatic image annotation via a coherent language model and active learning.In:ACM Multimedia Conference.(2004)892-889

[2] Li,J.,Wang,J.:Automatic linguistic indexing of pictures by a statistical modeling approach.IEEE Transactions on Pattern Analysis and Machine Intelligence 25(2003)

[3] K.Barnard,P.Duygulu,D.Forsyth,N.Freitas,D.Blei,and M.Jordan.Matching words and pictures.JMLR,2003.

[4] G.Garneiro and N.Vasconcelos.A Database Centric View of Semantic Image Annotation and Retrieval.SIGIR,2005.

[5] P.Duygulu,K.Barnard,N.Freitas and D.Forsyth.Object Recognition as Machine Translation:Learning a Lexicon for a Fixed Image Vocabulary.ECCV,2002.

[6] K.Barnard and D.A.Forsyth.Learning the semantics of words and pictures.In ICCV.pages 408-15,2001.

[7] D.Blei,Y.Andrew,and M.Jordan.Latent Dirichlet allocation.Journal of Machine Learning Research,3:993-1020,2003

[8] G.Dorko and C.Schmid.Selection of scale invariant parts for object class recognition.In Proc.ICCV,Oct.2003

[9] H.Zhang,A.berg,M.Maire,and J.Malik,"SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition,"Proc.IEEE CS Conf.Computer Vision and Pattern Recognition,vol.2,pp.2126-2136,June 2006.