Image Annotation with Concept Level Feature Using PLSA+CCA

Yu Zheng, Tetsuya Takiguchi, and Yasuo Ariki

Graduate School of Engineering, Kobe University
1-1, Rokkodai, Nada, Kobe, 657-8501 Japan
teiyiku@me.cs.scitec.kobe-u.ac.jp,
{takigu,ariki}@kobe-u.ac.jp
http://www.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+CCA,which generates the appropriate keywords to annotate the query image using large-scale image database.

Keywords: image annotation, PLSA, CCA, image recognition.

1 Introduction

With the production of large digital image collections favored by cheap digital recording and storage devices,these is a clear need for efficient indexing and retrieval systems. In QBE systems,various low-level visual features are preliminarily extracted from the data set and stored as image index. The query is an image example that is indexed by its features,and retrieved images are ranked with respect to their similarity to this query index. The natural query process is textual and images in a collection are indexed with words. Automatic image annotation has thus emerged as one of the key research areas in multimedia information retrieval.

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) and Canonical correlation analysis (CCA) are 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.

K.-T. Lee et al. (Eds.): MMM 2011, Part II, LNCS 6524, pp. 472–482, 2011.

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)[1],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 [2] 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.

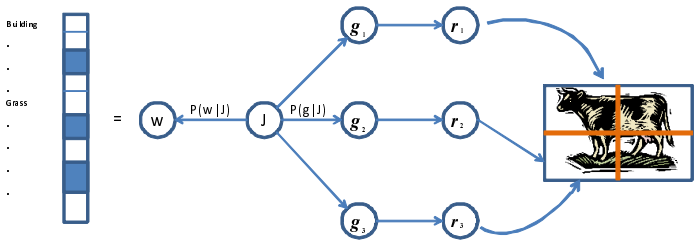

Fig. 1. MBRM.The annotation w is a binary vector. The image is produced by first sampling a set of feature vectors $g_1....g_n$, and then generating image regions $r_1....r_n$ from the feature vectors. Resulting regions are tiled to form the image.

Recent research efforts have focused on extensions of the translation paradigm that exploit additional structure in both visual and textual domains. For instance,[3]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.[4]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

and dimension problem and use the Internet database with concept groups for the training data to solve the large-scale dataset problem.By the experiment we can chose the best parameter, the number of topics K at PLSA model,and build the concept feature to learn correlation with label features.

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. 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. Automatic image annotation systems take advantage of existing annotated image data sets to link the visual and textual modalities by using machine learning techniques. The question is how to model the relation between captions and visual features to achieve the best textual indexing. This paper investigates this concept,proposing a new dependence between words and images based on latent aspects,we propose a probabilistic framework to analyze the contribution of the textual and the visual modalities separately. We assume that the two modalities share the same conditional probability distribution over a latent aspect variable that can be estimated from both or one of the two modalities for a given image.

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. 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

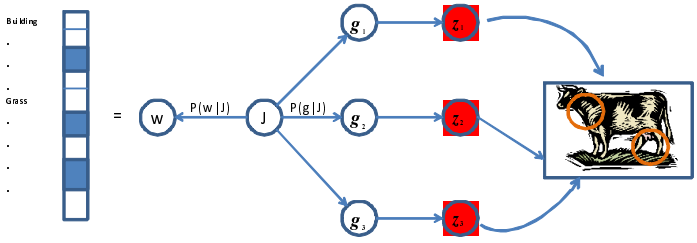

Fig. 2. Approch.The annotation w is a binary vector.The image is produced by first sampling a set of feature vectors $g_1....g_n$.The image regions $r_1.....r_n$ is replaced with concept representation $z_1.....z_n$.

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 search the relationship between the visual and texture modalities based on concept level by CCA.

The problem of generic object recognition algorithm is that the semantic gap between the image feature and the name of the object have not been solved. We proposed the annotation system that use two latent variable spaces. The computation of the conversion parameter in the CCA change with the number of dimension, the high dimension feature will not be suitable to CCA, so we need to reduce the number of dimension for CCA analysis.

CCA does not search the direct relationship between the two variables, but first convert to new variables P and Q can represent correlation of the two variables well. The concept feature vector z with p variables and the label feature vector w with q variables. The latent variable P_i derived from concept feature z. We obtain the conversion parameters matrix A and B by CCA use the training data D=$\{ (z_1, w_1)...(z_N, w_N) \}$. Learning the correlation between the image concept feature z and the label feature w, we build the model can convert the two variables to new variables with high correlation.We chose the matrix parameter A and B to maximize the correlation between P_i and Q_i shown in Eq.(1).and Eq.(2). $\bar{z}$ and $\bar{w}$ are the means of z and w.

$$P_i = A^T(z_i - \bar{z}) \tag{1}$$

$$Q_i = B^T(w_i - \bar{w}) \tag{2}$$

Two types of features that images feature and label feature cannot be compared directly before, but by this model they can be compared. We first take out training sample image data from the image database. And extract the image feature and label feature from the training data. Conduct the CCA analysis for the two vectors expressed in Eq.(3). p(z,w) is the co-occurrence probability of z and w. N_l is the number of latent variables.

$$p(z, w) = p(P_i) \sum_{i=1}^{N_l} p(z|P_i) * p(w|P_i) \tag{3}$$

$$p(z|P_i) = \frac{exp(-\frac{1}{2}(P - P_i)^T \Sigma^{-1}(P - P_i))}{\sqrt{(2\pi)^d \mid \Sigma \mid}} \tag{4}$$

$$p(w|P_i) = \Pi\mu\delta_{w,P_i} + (1 - \mu)\frac{N_w}{N_W} \tag{5}$$

The probability of z when given latent variable P_i $p(z|P_i)$ expressed in Eq.(4),it is the Gauss distribution with the mean of P_i in the latent variable space. z is the concept feature vector. P_i is the latent variable derived from the concept feature vector z. The probability of the label feature derived from the latent variable P_i $p(w|P_i)$ expressed in Eq.(5). It is designed to top-down by language model. N_w is the occurrence of labels w in the training data. δ_{w,P_i} will be 1 when label w labeling to P_i, otherwise will be 0. μ is fixed to 0.99. The PLSA-CCA construction has been shown in Fig.3.

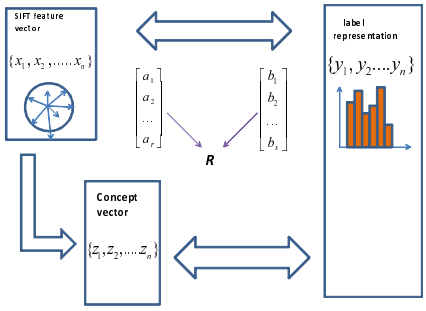

Fig. 3. We first get the SIFT feature x_1 , ..., x_n of the image and then convert it to concept feature z_1 , ..., z_n by PLSA and to do the CCA analysis with label feature

3.2 Training Data Search

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. 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 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 PLSA-Mixed Concept Feature

A document is a mixture of latent aspects. These latent aspects are defined by multinomial distributions over words that are learned for each text corpus

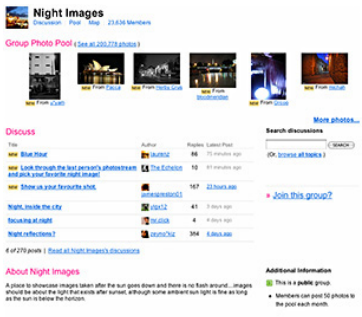

Fig. 4. Flickr is almost certainly the best online photo management and sharing application in the world. With millions of users, and hundreds of millions of photos and videos, Flickr is an amazing photographic community, with sharing at its heart.

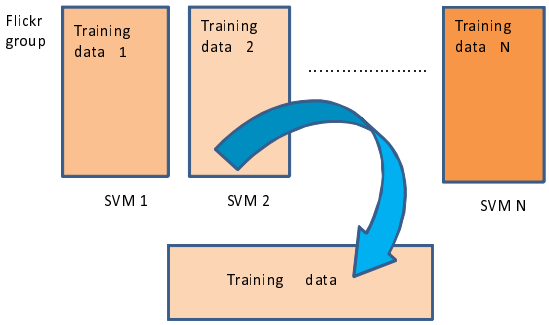

Fig. 5. Search training data with Flickr group.We download thousands of images from many Flickr groups.Groups that we use are organized by objects. For each group,we train a SVM classifier. For a test image,we use the trained group classifiers to predict likely group memberships. We train classifiers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

considered. These distributions characterize the aspects and show that a correspondence between topics identified by humans and latent aspects can exist. The concept of latent aspects is not restricted to text documents. Images are intuitively seen as mixtures of several content types, which make them good candidates for a latent aspect approach. Different latent aspect models, adapted from the LDA model for text, have been proposed to model annotated images.

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. 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. PLSA is a statistical model as

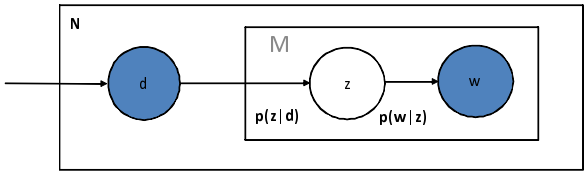

Fig. 6. Joint probability model. Plate notation representing the PLSA model. d is the document variable,z is a topic drawn from the topic distribution for this document,$p(z|d)$. w is a word drawn from the word distribution for this topic,$p(w|z)$. The d and w are observable variables,the topic z is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

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.

$$P(v_j, d_i) = P(d_i) \sum_{l=1}^{N_A} P(z_l|d_i) P(v_j|z_l) \tag{6}$$

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.(6) 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.(7). Eq.(7)is the concept level image representation.

$$(P(z_l|d))_{l=1,2.....N_A} \tag{7}$$

Comparing to feature representation,concept representation shown in Eq.(7) can find out accurate images because it searches data based on objects in the images.

3.4 Annotation Using PLSA-CCA

PLSA has been recently shown to perform well on image classification tasks, using the aspect mixture proportions to learn the classifiers. The conditional

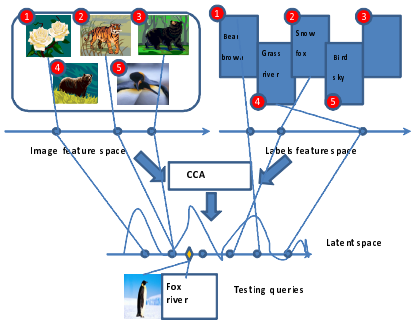

Fig. 7. The algorithm of CCA

probability distribution over aspects $P(z|d_{new})$ can be inferred for an unseen document d_{new} . The folding-in method maximizes the likelihood of the document d_{new} with a partial version of the EM algorithm, where $P(x \mid z)$ is obtained from training and kept fixed.In doing so,$P(z|d_{new})$ maximizes the likelihood of the document d_{new} with respect to the previously learned $P(x \mid z)$ parameters. The PLSA-MIXED model learns a standard PLSA model on a concatenated representation of the textual and the visual features x=(w,v).Using a training set of captioned images, $P(x \mid z)$ is learned for both textual and visual co-occurrences to capture simultaneous occurrence of visual features and words. Once $P(x \mid z)$ has been learned,it can be used to infer a distribution over words for a new image as follows: The new image d_{new} is represented in the concatenated vector-space, where all word elements are zero(no annotation): x_{new} =(0,v_{new}).The multinomial distribution over aspects given the new image $P(z|d_{new})$ is then computed with the partial PLSA steps and allows the computation of $P(x|d_{new})$ .The conditional probability distribution over words $P(w|d_{new})$ is extracted from $P(x|d_{new})$ and allows the annotation of the new image d_{new}.

Given two column vectors $X = (x_1, ...x_n)^T$ and $Y = (y_1, ..., y_m)^T$ of random variables,canonical correlation analysis seeks vectors a and b such that the random variables $a^T X$ and $b^T Y$ maximize the correlation ρ =cor($a^T X, b^T Y$) expressed in Eq.(8). The random variables $U = a^T X and V = b^T Y$ are the first pair of canonical variables.

$$\rho = \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} a}\sqrt{b^T \Sigma_{YY} b}} \tag{8}$$

$$h = \sum_{i=1}^{N} p(z_{new}|P_i)p(w|P_i) \tag{9}$$

The unkown images is inputed and the concept feature z_{new} will be computed and then the posterior probability of the label $p(w|z_{new})$ will be computed. $p(z_{new})$ and $p(P_i)$ have the same value for all the label.The h value Eq.(9) of all the label w will be ordered and assigned to the unknown image based on the h.

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.

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

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

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

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

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 proposed method performs best. Precision-recall have been shown in fig.8.

As shown in Fig.9,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.

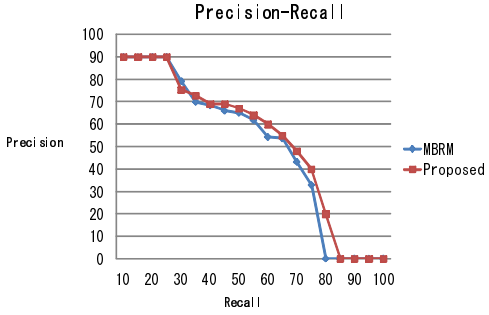

Fig. 8. Precision-recall when K=8 with the 1000 input images

Fig. 9. 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. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: Proc. ACM SIGIR, pp. 127–134 (2003)
2. Feng, S.L., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. In: IEEE Conf. Computer Vision and Pattern Recognition (2004)
3. Jin, R., Chai, J.Y., Si, L.: Effective automatic image annotation via a coherent language model and active learning. In: ACM Multimedia Conference, pp. 892–889 (2004)
4. Li, J., Wang, J.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (2003)
5. Barnard, K., Duygulu, P., Forsyth, D., Freitas, N., Blei, D., Jordan, M.: Matching words and pictures. JMLR (2003)
6. Garneiro, G., Vasconcelos, N.: A Database Centric View of Semantic Image Annotation and Retrieval. In: SIGIR (2005)
7. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
8. Barnard, K., Forsyth, D.A.: Learning the semantics of words and pictures. In: ICCV, pp. 408–415 (2001)
9. Zhang, H., Berg, A., Maire, M., Malik, J.: SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. In: Proc. IEEE CS Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 2126–2136 (June 2006)
10. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (2008)