A PROBABILITY MODEL FOR IMAGE ANNOTATION

Yong Ge, Richang Hong, Zhiwei Gu, Rong Zhang, Xiuqing Wu

Dept of EEIS, University of Science and Technology of China
Hefei, 230027 China
{ygesi@mail, richong@mail, guzhiwei@mail, zrong@, xqwu@}.ustc.edu.cn

ABSTRACT

Automatic image annotation is a promising solution to enable more effective image retrieval by keywords. Traditionally, statistical models for image auto-annotation predicate each annotated keyword independently without considering the correlation of words. In this paper, we propose a novel probability model, in which the correspondence between keywords and image visual tokens/regions and the word-to-word correlation are well combined. We employ the conditional probability to express two kinds of correlation uniformly and obtain the correspondence between keyword and visual feature with the cross-media relevance model (CMRM). Experiments conducted on standard Corel dataset demonstrate the effectiveness of the proposed method for image automatic annotation.

1. INTRODUCTION

Nowadays, the digital images have increased tremendously with the rapid development of digital photography. In order to organize and search these images efficiently, content-based image retrieval (CBIR) was proposed in early 1990's. It takes example image as query and computes relevance based on the similarity of low-level features such as texture, color, and shape etc. However, there is a gap between low-level visual feature and semantic meaning, which is a major problem that leads to low retrieval accuracy for most CBIR approaches.

To capture the semantic of image, automatic image annotation has received extensive attention recently. Furthermore indexing by key-words facilitates organizing and searching an image database. Previously, many statistical models have been proposed [2, 3, 4, 5]. Based on abundant training samples, these models determine the correspondence between keywords and image visual tokens/regions, and then use this association to annotate images that do not have captions. However, the common problem shared by most models is that each keyword for an image is predicated independently from other keywords, eventually the current annotation accuracy with these models is quite low for the captions too noisy (Fig1). For an image, a set of true keywords constitutes the whole semantic environment, in which keyword is commonly correlated each other. Keyword that is irrelevant to others on semantic concept is considered as noisy keyword. To remove the noisy keyword, we propose a novel probability model, in which the correlation of keywords and the correspondence between keyword and image visual tokens/regions are combined well. We consider the word-to-word correlation from two aspects. One is the semantic similarity, such as "plane and jet" and "house and building". The other is semantic concomitance, such as "sky and cloud" and "coral and ocean". Even though, the co-occurring words, such as "coral and ocean", are different on semantic meaning, they have close correlation for annotating some images. [1] attempts to improve the annotation accuracy by using the semantic similarity between keywords. However, they only consider the semantic similarity. We employ condition probability to express the two kinds of correlation uniformly. For example, *p(plane/jet)* and *p(ocean/coral)* denote the probability of the plane (ocean) when the image is annotated by jet (coral). We employ the cross-media relevant model (CMRM) to obtain the correspondence between keyword and image visual tokens.

17018 water sky cloud dune snow

131019 garden branch bird leaf plants

Fig 1 Automatic annotation example with noisy terms

2. RELATED WORK

In recent years, a variety of methods have been applied to automatic image annotation. Mori et al. [6] developed a co-occurrence model, in which they looked at the co-occurrence of keywords with image regions. In [3], a machine translation model was proposed to translate from a vocabulary of blobs to a vocabulary of words and achieved improvement over the co-occurrence model. Jeon et al. [2] introduced a cross-media relevance model (CMRM) that learns the joint distribution of a set of regions and a set of keywords rather than the correspondence between a single

1-4244-1017-7/07/$25.00 ©2007 IEEE

region and a single keyword. The CMRM was subsequently improved through the continuous-space relevance mode (CRM)[4] and the multiple Bernoulli relevance model (MBRM) [5]. In the scenario that each word is treated as a distinct class, image annotation can be viewed as multi-class classification problem. The representative works are content-based annotation method with SVM[7], Bayes Point Machine[8],and asymmetrical support vector machine-based MIL algorithm[9]. All above methods require a well organized training set to predict the correlation between the keyword and the visual feature, which leads to that only a very limited number of concepts can be modeled on the small-scale training set. So these models lack generalization capability. To overcome this problem, some researchers proposed novel method based by search [10].

3. PROPOSED PROBABILITY MODEL

3.1. Problem Formulation

In this section, we show how to combine visual and textual information through our method. For an un-annotated image, the essential of annotation is estimate the conditional probability of keyword $P(w_i/I_q)$, where $w_i\,(i=1\ldots k)$ is the i^{th} keyword in the vocabulary and I_q is an uncaptioned image. We formulate the calculation of conditional probability as follows:

$$P(w_i/I_q)\cong \partial p(w_i/I_q)+(1-\partial)\sum_{j=1,j\neq i}^{k} p(w_i/w_j)p(w_j/I_q) \qquad (1)$$

In equation (1), $p(w_i/I_q)$ and $p(w_j/I_q)$ represent the conditional probability that is obtained through the correspondence between the keyword and image visual feature. Although here we use the CMRM model, however, even though we employ TM and CRM, the equation (1) can be applied. The $p(w_i/w_j)$ denotes the conditional probability that is the possibility of w_i when w_j is annotated for I_q. In our work, we estimate the word-to-word correlation $p(w_i/w_j)$ with keyword co-occurrence matrix. The parameter ∂ determines the degree of the correspondence between the keyword and image visual feature and the word-word correlation. In this paper, the parameter ∂ is empirically set as 0.6. Finally we could select top M keywords with the largest conditional probability as the annotation for image I_q. In our work, we set M to be 5.Our algorithm is composed of two steps:

1. **Link between keyword and blob-token.** For an uncaptioned image I_q, we first calculate the probability of all words in the vocabulary with CMRM and get a vector of probabilities for all words in the vocabulary. It is represented by $p_{I_q}^{CMRM}$

2. **Combing the correlation of keywords.** We use the keyword co-occurrence matrix to estimate the $p(w_i/w_j)$, $i,j \in \{1,\ldots\ ,k\}$ for every keyword-pair. With the vector of probabilities $p_{I_q}^{CMRM}$ and $p(w_i/w_j)$, $i,j \in \{1,\ldots\ ,k\}$, we calculate $P(w_i/I_q)$ through equation (1).

Considering the efficiency, for every vector $p_{I_q}^{CMRM}$ we only use the top 10 keywords with the largest probability when calculating the $P(w_i/I_q)$ for one uncaptioned image. Therefore, the equation (1) becomes:

$$P(w_i/I_q)=\partial p(w_i/I_q)+(1-\partial)\sum_{j=1,j\neq i}^{10} p(w_i/w_j)p(w_j/I_q) \qquad (2)$$

3.2. The Cross-Media Relevance Model(CMRM)

Following the derivation of Jeon et al.[2], the CMRM model can be described as follows. Suppose we have a training set T, of labelled images, and a test set Q, of unlabelled images.

Firstly, each training image is segmented into regions. Secondly, all regions of the training images are clustered based on the visual features, such as color, shape or texture. We call these clusters 'blobs'. Thus, each image of the training set can be represented as a set of blobs and words, $J=\{b_1,\ldots,b_n;\ w_1,\ldots,w_l\}$. For the test images I, they are also partitioned into regions, each of which is assigned to the blob that is closest to it. Then each test image can be represent as a set of blobs $I=\{\ b_1,\ldots,b_m\ \}$. So the annotation process for the test image is to estimate the conditional probability $P(w/I)\approx P\ (w/\ b_1,\ldots,b_m)$. They use the training set T of annotated images to estimate the joint probability $P(w,\ b_1,\ldots,b_m)$ as follow:

$$P(w,b_1,\ldots,b_m)=\sum_{J\in T}P(J)P(w,b_1,\ldots,b_m\mid J) \qquad (3)$$

They assume that the events of observing w and $b_1,\ldots,b_m$ are mutually independent once one image J is chosen. Therefore, the equation (3) becomes:

$$P(w,b_1,\ldots b_m)=\sum_{J\in T}P(J)P(w\mid J)\prod_{i=1}^{m}P(b_i\mid J) \qquad (4)$$

$P(J)$ is treated uniformly over all images in T. $P(w/J)$ and $P(b/J)$ are estimated by smoothed maximum likelihood as follow:

$$P(w\mid J)=(1-\partial_J)\frac{\#(w,J)}{|J|}+\partial_J\frac{\#(w,T)}{|T|} \qquad (5)$$

$$P(b\mid J)=(1-\beta_J)\frac{\#(b,J)}{|J|}+\beta_J\frac{\#(b,T)}{|T|} \qquad (6)$$

where, *#(w,J)* denotes the number of times word w occurs in the caption of J, and *#(w,T)* denotes the number of times words w occurs in all captions of image in T. *#(b,J)* is the number of times blob b occurs in J, and *#(b,T)* is that of the whole training set. $|J|$ is the aggregate count of all keywords and visual terms in J, and $|T|$ is that of the whole training set. ∂ and β are smoothing parameters obtained by optimizing system performance on a held-out portion of the train set.

3.3. The word-to-word correlation

The word-to-word correlation contains semantic similarity and semantic concomitance. [1] strives to find the semantic similarity of keyword with WordNet. However, the semantic similarity of WordNet only reflects the hyponymy hierarchy, i.e. it can easily find the similarity of "plane" and "jet" but omits the correlation of "plane" and "sky". In the captions of training images, two keywords with semantic similarity or semantic concomitance will frequently occur simultaneously. So statistical co-occurrence for annotated keywords is an effective way to estimate the word-to-word correlation.

In our work, we count the frequency of every keyword-pair simultaneously occurring in the annotation for one image, then we gain the keyword co-occurrence matrix M_c $(k \times k)$, where k is the total number of keywords in training set. $M_c(w_i,w_j)$ is the frequency of co-occurrence for keyword w_i and w_j. $M_c(w_i)$ is occurring time of keyword w_i.

$$M_c(w_i) = \sum_{j \in \{1,2..k\}} M_c(w_i, w_j) \tag{7}$$

We estimate the $p(w_i/w_j)$ as follow:

$$p\ (w_j/w_i)=M_c(w_i,w_j)\ /M_c(w_i) \tag{8}$$

$$p\ (w_i/w_j)=M_c(w_i,w_j)\ /M_c(w_j) \tag{9}$$

Note that $p(w_j/w_i)$ is usually not equal to $p(w_i/w_j)$. With sufficient training set, this kind of associations is effectual and convenient. As shown in Table 1, we report some keyword-pairs conditional probability calculated by equation (7).

4. EXPERIMENT

We conduct our experiments with the data set downloaded from [11], which is the same as the data set used in the [1]. The data package contains 5,000 images from 50 Corel Photo CDs. Each Cd contains 100 images on the same topic. Images are segmented using Normalized cut [12]. Only regions larger than a threshold are used, each image is typically represented by 5-10 regions. Each region is represented as a 30 dimensional vector, including region color, region average orientation energy, region size and location and so on. Each image is labeled by 1-5 keywords. The vocabulary contains 371 different words. 4,500 images are used as the training set and the remaining 500 images constitute the testing set. *Precision* and *recall* are used to evaluate the performance of our method:

$$precision = \frac{N_c}{N} \tag{10}$$

Table 1 Examples of word-to-word Correlation

Keywords-pair	Conditional Probability
p(cat/tiger)	0.9780
p(sky/plane)	0.4422
p(tracks/car)	0.6567
p(beach/palm)	0.5357
p(snow/arctic)	0.7222
p(plane/jet)	1.0000
p(ocean/coral)	0.9663

Table 2 Performance of some Frequent Keywords

Keywords	CMRM		Our Method	
	Precisipn	Recall	Precision	Recall
petals	0.3750	0.7500	0.7500	0.7500
pool	0.4167	0.4545	0.7143	0.4545
sand	0.2857	0.2105	0.4000	0.2105
wall	0.2500	0.0769	0.5000	0.0769
statue	0.1666	0.0909	0.5000	0.0909
nest	0.3636	0.1333	0.4444	0.1333
stone	03870	0.5714	0.4400	0.5238
field	0.2307	0.3529	0.4000	0.5882
sky	0.2758	0.8380	0.2538	0.9524
tree	0.2105	0.7234	0.2005	0.9043
buildings	0.2643	0.4259	0.2647	0.5000
train	0.2500	0.0909	0.3333	0.1819

$$recall = \frac{N_c}{N_r} \tag{11}$$

where, for a single word, N_c is the number of correctly predicted test images, N is the number of all test image predicted by the word, N_r is the number of test images actually annotated by the word. To illuminate the efficiency of removing noisy keywords, we propose *Noisy-Coefficient (N-Coe)*:

$$N - Coe = \frac{N_w}{N_r} \tag{12}$$

where N_r is the number of return words for an image, N_w is the number of wrong keywords. The average *N-Coe* over all test images is the final measure.

4.1. Model Comparison

We give some examples of annotation result with our method and CMRM in Fig 2. As Table2 shows, we report results for some frequently used keywords for CMRM and Our Method. For the keywords, petals, pool, sand, wall, statue, nest, and stone, the *precision* of Our Method is substantially higher than that of CMRM; at the same time, except for the keyword stone, *recall* is the same in both cases. This happens due to the removal of only noisy keywords. For keywords, field, sky, tree, buildings, and train, our method gains substantial improvement on *recall*; on the other hand, *precision* is almost as good as that of CMRM. So our method can catch more relevant keywords.

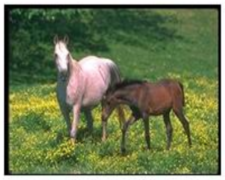
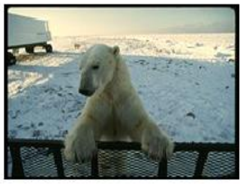
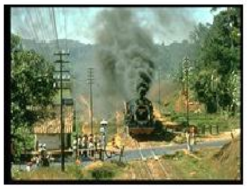
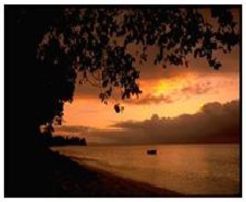
Image				
Ground Truth	field foals mare horses	bear polar snow vehicle	locomotive railroad train tree	beach boats water sunset
CMRM	buildings house mare window door	snow sky tree water people	tree plants garden ice frost	sky hill sand rainbow stone
Our Method	field mare foals tree house	bear snow polar tree sky	train locomotive tree water sky	water tree hill sky sand

Fig 2 Annotation Examples

Table 3 Performance Comparison of Automatic Annotation on the Corel dataset

Models	CMRM	Our Method	TM	TMHD
Result on 49 best words				
Mean per-word Recall	0.47	0.47	0.35	0.21
Mean per-word Precision	0.40	0.47	0.20	0.30
Result on keywords with recall>0				
Mean per-word Recall	0.37	0.36	0.35	0.21
Mean per-word Precision	0.35	0.40	0.20	0.30
Results on all test 500 images				
Average N-Coe	0.7756	0.6923	—	—

The average of *precision*, *recall* and *N-Coe* are shown in Table 3. Including the improvement of *precision* and *recall*, the drop of *N-Coe* explicitly illustrates efficiency of removing noisy keywords. Another related work is Yohan Jin's TMHD. We compare the annotation performance of our method and TMHD in Table 3. Compared with CMRM, our method gains improvement on *precision* and equal on *recall*. While the TMHD gains increased *precision* but losses on *recall*.

5. CONCLUSION

In this paper, we develop an efficient probability method for image annotation. This method combines the text and visual information well. We employ the CMRM to estimate the correspondence between the keyword and visual feature. We propose two kinds of word-to-word correlation and express the correlation with conditional probability uniformly. Experiment on Coral dataset show engaging performance of the proposed method.

6. REFERENCES

[1] Yohan Jin, Khan and etc. Image Annotation By Combining Multiple Evidence & WordNet. Proceedings of the 13th Annual ACM MM, pp. 706-715, Singapore, 2005.

[2] J. Jeon, V. Lavrenko and R. Manmatha. Automatic Image Annotation and Retrieval Using Cross-media Relevance Models. In Proceeding of ACM SIGIR, pp. 119-126, July 2003.

[3] Pinar, D. and Kobus, B. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In Seventh ECCV, 4:97-112, 2002.

[4] R. Manmatha, V. Lavrenko, and J. Jeon, A Model for Learning the Semantics of Pictures. In Proceeding of the 17th Annual Conference on Neural Information Proceeding Systems, 2003

[5] S. L. Feng, R. Manmatha and C. Lavrenko. Mulitple Bernouli Relevance Models for Image and Video Annotation. In Proc. Of CVPR, Washington, DC, June, 2004.

[6] Y. Mori, H. Takahashi, and R. Oka, Image-to-word transformation based on dividing and vector quantizing image with words, In MISRM'99 First Intl. Workshop on Multimedia Intelligent Storage and Retrieval Management, 1999.

[7] Claudio, C., Gianluigi, C., Raimondo, S. Image annotation using SVM. In proceeding of Internet image IV, Vol. SPIE, 2004

[8] Edward Chang, Kingshy Goh. Gerard Sychay, Gang Wu. CBSA: content-base soft annotation for multimodal image retrieval using bayes point machines. CirSysVideo, pp. 26-38,13(1), 2003.

[9] C. B. Yang, M. Dong, J. Hua. Region-based Image Annotation using Asymmetrical Support Vector Machine-based Multiple-Instance Learning. In proceeding of IEEE CVPR, 2006

[10] X. Wang. Lei Zhang and etc. AnnoSearch: Image Auto-Annotation by Search. In CVPR, 2006

[11] "http://www.cs.arizona.edu/people/kobus/research/data/eccv_2002,"

[12] Jiabo Shi and Jitendra Malik, Normalized cuts and image segmentation. In Proc of IEEE CVPR 97, Puerto Rico, 1997.