doi:10.3772/j.issn.1006-6748.2015.01.011

Exploiting PLSA model and conditional random field for refining image annotation[①]

Tian Dongping(田东平)[②][*][**]
([*] Institute of Computer Software, Baoji University of Arts and Sciences, Baoji 721007, P. R. China)
([**] Institute of Computational Information Science, Baoji University of Arts and Sciences, Baoji 721007, P. R. China)

Abstract

This paper presents a new method for refining image annotation by integrating probabilistic latent semantic analysis (PLSA) with conditional random field (CRF). First a PLSA model with asymmetric modalities is constructed to predict a candidate set of annotations with confidence scores, and then model semantic relationship among the candidate annotations by leveraging conditional random field. In CRF, the confidence scores generated by the PLSA model and the Flickr distance between pairwise candidate annotations are considered as local evidences and contextual potentials respectively. The novelty of our method mainly lies in two aspects: exploiting PLSA to predict a candidate set of annotations with confidence scores as well as CRF to further explore the semantic context among candidate annotations for precise image annotation. To demonstrate the effectiveness of the method proposed in this paper, an experiment is conducted on the standard Corel dataset and its results are compared favorably with several state-of-the-art approaches.

Key words: automatic image annotation, probabilistic latent semantic analysis (PLSA), expectation-maximization, conditional random field (CRF), Flickr distance, image retrieval

0 Introduction

With the prevalence of digital imaging devices such as webcams, phone cameras and digital cameras, the number of accessible images is growing at an exponential speed. Thus how to make the best use of these resources becomes an emerging problem. An ideal image retrieval system can establish exact correspondence between image visual content and semantic description so that many well-developed text retrieval approaches can be easily used to retrieve images by ranking the connection between image annotations and text quires. Therefore, how to efficiently annotate the images serves as a key problem for image retrieval. The traditional method for image annotation is to let people manually annotate images by some keywords. However, this method is labor-intensive and time-consuming. Furthermore, the annotating result is subjective to different people and it is difficult to be extended to large image dataset. To address these limitations, automatic image annotation (AIA) has emerged as an important topic and becomes an active research area in recent years. Its goal is to automatically assign some keywords to an image that can well describe the content comprising in it.

In recent years, many methods have been developed for AIA, and most of them can be roughly classified into two categories, viz., classification-based method and probabilistic modeling method. The representative work of the former involves automatic linguistic index for pictures[1], content-based annotation method with SVM[2] and asymmetrical support vector machine-based MIL algorithm[3]. The probabilistic modeling methods include the translation model (TM)[4], cross-media relevance model (CMRM)[5], continuous relevance model (CRM)[6], multiple- Bernoulli relevance model (MBRM)[7] and latent aspect model PLSA[8,9]. However, all the annotation methods aforementioned, to some extent, can achieve certain success compared to the manual annotation, but they are still far from satisfaction due to the little effort on exploiting the semantic context and correlations among annotation keywords. Recently some researchers pro-

[①] Supported by the National Basic Research Priorities Programme (No. 2013CB329502), the National High Technology Research and Development Programme of China (No. 2012AA011003), the Natural Science Basic Research Plan in Shanxi Province of China (No. 2014JQ2-6036) and the Science and Technology R&D Program of Baoji City (No. 203020013, 2013R2-2).

[②] To whom correspondence should be addressed. E-mail: tdp211@163.com, tiandp@ics.ict.ac.cn

Received on Nov. 4, 2013

pose to refine image annotation by taking the word correlation into account. Jin, et al.[10] have implemented pioneer work on annotation refinement based on the knowledge of WordNet. This method, however, can only achieve limited success as it totally ignores the visual content of images. In Ref. [11], Wang, et al. apply random walk with restarts model to refine candidate annotations by integrating word correlations with the original candidate annotation confidence together. Followed by they propose a content based approach by formulating the annotation refinement as a Markov process[12]. In addition, Wang, et al.[13] employ conditional random field to refine image annotation by incorporating semantic relations between annotation words. More recently Liu, et al.[14] rank the image tags according to their relevance with respect to the associated images by tag similarity and image similarity in a random walk model. Xu, et al.[15] come up with a new graphical model termed as regularized latent Dirichlet allocation (rLDA) for tag refinement. Zhu, et al.[16] put forward an efficient iterative approach for image tag refinement by pursuing the low-rank, content consistency, tag correlation and error sparsity by solving a constrained yet convex optimization problem. Besides, several nearest-neighbor-based methods have also been proposed for refining image annotation in the most recent years[17,18].

As briefly reviewed above, most of these approaches can achieve encouraging performance and motivate us to explore better image annotation methods with the help of their excellent experiences and knowledge. Hence, in this paper a new method for refining image annotation is presented based on a fusion of probabilistic latent semantic analysis and conditional random field (PLSA-CRF). For a given image, a PLSA model with asymmetric modalities is first constructed to predict a candidate set of annotations with confidence scores, and then model semantic relationship between these keywords using conditional random field (CRF) where each vertex indicates the final decision (true/false) on a candidate annotation and the refined annotation is given by inferring the most likely states of these vertices. The method is evaluated on the standard Corel dataset and the experimental results are superior or highly competitive to several state-of-the-art approaches. To the best of our knowledge, this study is the first attempt to integrate PLSA with conditional random field in the task of refining image auto-annotation.

The rest of the paper is organized as follows. Section 1 presents how to apply PLSA to predict a candidate set of annotations with confidence scores. Section 2 elaborates the PLSA-CRF model, in which the confidence scores generated by the PLSA model and the concept similarity between pairwise candidate annotations are considered as local evidences and contextual potentials respectively. Experimental results on the Standard Corel dataset are reported and analyzed in Section 3. Finally, this paper is ended with some important conclusions and future work in Section 4.

1 PLSA model

PLSA[19] is a statistical latent class model which introduces a hidden variable (latent aspect) z_k in the generative process of each element x_j in a document d_i. Given this unobservable variable z_k, each occurrence x_j is independent of the document it belongs to, which corresponds to the following joint probability:

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

The model parameters of PLSA are the two conditional distributions: $P(x_j|z_k)$ and $P(z_k|d_i)$. $P(x_j|z_k)$ characterizes each aspect and remains valid for documents out of the training set. On the other hand, $P(z_k|d_i)$ is only relative to the specific documents and cannot carry any prior information to an unseen document. An EM algorithm is used to estimate the parameters through maximizing the log-likelihood of the observed data.

$$L = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, x_j) \log P(d_i, x_j) \quad (2)$$

where $n(d_i, x_j)$ is the count of element x_j in document d_i. The steps of the EM algorithm can be succinctly described as follows.

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

$$P(z_k \mid d_i, x_j) = \frac{P(z_k \mid d_i) P(x_j \mid z_k)}{\sum_{l=1}^{K} P(z_l \mid d_i) P(x_j \mid z_l)} \quad (3)$$

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

$$P(x_j \mid z_k) = \frac{\sum_{i=1}^{N} n(d_i, x_j) P(z_k \mid d_i, x_j)}{\sum_{m=1}^{M} \sum_{i=1}^{N} n(d_i, x_m) P(z_k \mid d_i, x_m)} \quad (4)$$

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{M} n(d_i, x_j) P(z_k \mid d_i, x_j)}{\sum_{j=1}^{M} n(d_i, x_j)} \quad (5)$$

If one of the parameters ($P(x_j|z_k)$ or $P(z_k|d_i)$) is known, the other one can be inferred by using the folding-in method, which updates the unknown parameters with the known parameters kept fixed, so that it can maximize the likelihood with respect to the previously trained parameters. Given a new image visual features $v(d_{new})$, the conditional probability distribution $P(z_k | d_{new})$ can be inferred with the previously estimated model parameters $P(v | z_k)$, then the posterior probability of words can be computed by

$$P(w \mid d_{new}) = \sum_{k=1}^{K} P(w \mid z_k) P(z_k \mid d_{new}) \qquad (6)$$

From Eq. (6), a candidate set of annotations with confidence scores (i. e., the posterior probabilities of words) can be easily obtained.

2 PLSA-CRF model for refining image annotation

2.1 Concept similarity measure

To measure the similarity between pairwise concepts related to an image from the viewpoint of computer vision, in actual fact, is still a tough problem in multimedia information processing. The commonly used methods include WordNet[20] and normalized Google distance (NGD)[21]. From the comparison of their definitions, it can be seen easily that NGD puts emphasis on the measure of the contextual relation while WordNet focuses on the semantic meaning of concept itself. What's more, both of them build word correlations only based on textual descriptions of images and do not fully take visual information of the corresponding images into account, which also plays a crucial role in precise image auto-annotation. So in this paper, the simple yet very efficient Flickr distance (FD)[22] is adopted to measure the similarity between two concepts C_1 and C_2, which can be calculated as the average square root of Jensen-Shannon (JS) divergence between the corresponding visual language models as follows (K denotes the total number of latent topics).

$$D_{Flickr}(C_1, C_2) = \sqrt{\frac{\sum_{i=1}^{K} \sum_{j=1}^{K} D_{JS}(Pz_i^{c_1} \mid Pz_j^{c_2})}{K^2}} \qquad (7)$$

where $D_{JS}(Pz_i^{c_1} \mid Pz_j^{c_2}) = (D_{KL}(Pz_i^{c_1} \mid M) + D_{KL}(Pz_j^{c_2} \mid M))/2$ denotes the JS divergence and is defined based on Kullback-Leibler (KL) divergence to describe the distance metric between these visual language models, $M = (Pz_i^{c_1} + Pz_j^{c_2})/2$ is the average of $Pz_i^{c_1}$ and $Pz_j^{c_2}$. Here, given that $Pz_i^{c_1}$ and $Pz_j^{c_2}$ be the trigram distributions under latent topic $z_i^{c_1}$ and $z_j^{c_2}$ respectively, and $z_i^{c_1}$ denotes the ith latent topic of concept C_1. Besides, the KL divergence can be defined as

$$\begin{aligned} D_{KL}(Pz_i^{c_1} \mid Pz_j^{c_2}) &= \sum\nolimits_l Pz_i^{c_1}(l) \log \frac{Pz_i^{c_1}(l)}{Pz_j^{c_2}(l)} \\ &= -\sum\nolimits_l Pz_i^{c_1}(l) \log Pz_j^{c_2}(l) \\ &\quad + \sum\nolimits_l Pz_i^{c_1}(l) \log Pz_i^{c_1}(l) \\ &= H(Pz_i^{c_1}, Pz_j^{c_2}) - H(Pz_i^{c_1}) \end{aligned} \qquad (8)$$

where $Pz_i^{c_1}(l)$ and $Pz_j^{c_2}(l)$ correspond to the probability densities of the lth trigram in these two distributions respectively. From the perspective of information theory, the KL divergence, in fact, is a measurement of the mutual entropy between two visual language models. Thus $H(Pz_i^{c_1}, Pz_j^{c_2})$ is the cross entropy of the two distributions and $H(Pz_i^{c_1})$ is the entropy of $Pz_i^{c_1}$.

2.2 Conditional random field

Conditional random field (CRF) has been widely used in computer vision community[23-25]. A CRF can be viewed as an undirected graphical model in which each vertex represents a random variable whose distribution is to be inferred, and each edge represents a dependency between two random variables. In CRF, the distribution of each discrete random variable y_i in the graph is conditioned on an input sequence x. Mathematically, the conditional probability of $y = (y_1, y_2, \cdots, y_n)$ given x is formulated as

$$P(y \mid x) = \frac{e^{\phi(y, x; \Phi)}}{\sum_{y'} e^{\phi(y', x; \Phi)}} \qquad (9)$$

where

$$\begin{aligned} \phi(y, x; \Phi) &= \sum_i \sum_k \theta_k^1 f_k^1(y_i, i, x) \\ &\quad + \sum_{i,j} \sum_l \theta_l^2 f_l^2(y_i, y_j, i, j, x) \end{aligned} \qquad (10)$$

Eq. (10) is the potential function. i, j are used to index the vertexes, $f_k^1(y_i, i, x)$ and $f_l^2(y_i, y_j, i, j, x)$ denote the node feature function and edge feature function respectively, $\Phi = \{\theta^1, \theta^2\}$ indicates the model parameters to be learned. To simplify the computation, we adopt a variant of the potential function employed in work of Ref. [13] as follows.

$$\begin{aligned} \phi(y, x; \Phi) &= \alpha_1 \times \sum_i \omega^1(y_i, i, x) + \alpha_2 \\ &\quad \times \sum_{i,j} \omega^2(y_i, y_j, i, j) \end{aligned} \qquad (11)$$

where ω^1 indicates the local evidence of the state of y_i. It depends on the image observation x. ω^2 is a prior parameter that indicates the contextual potential between the states of two variables y_i and y_j. Here, we take the local evidence as the logarithm of the confidence scores provided by the PLSA model.

$$\omega^1(y_i,i,x) = \begin{cases} \log P(c(y_i)=1 \mid x), & y_i = 1 \\ \log[1 - P(c(y_i)=1 \mid x)], & y_i = 0 \end{cases} \tag{12}$$

where $c(y_i)$ is used to indicate the concept and P is the posterior probability of annotation keywords according to Eq. (6). Alternatively, the contextual potential is assumed to be independent of x and can be obtained as general knowledge provided by the converted FD as

$$\omega^2(y_i,y_j,i,j) = \begin{cases} -\log FD(c(y_i),c(y_j)), & y_i = y_j = 1 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

In this simplified CRF model, the weight parameters $\Phi = \{\alpha_1, \alpha_2\}$ to be learned are utilized to control the balance between local evidence and contextual potential.

2.3 Parameter estimation and refining image annotation

As we know, the task of CRF for image annotation is to infer the most probable labels given an input image and the model parameters which are learned from the training set. As can be seen from the above description, it is difficult to choose the weight parameters manually since the local evidence and contextual potential come from different sources. Following the work in Ref. [13], a similar learning algorithm is adopted to estimate the parameters. And the whole process for refining image annotation by fusing PLSA with conditional random field as well as the weight parameter estimation is described in Algorithm 1.

Algorithm 1 PLSA-CRF for refining image annotation

```
Training
Input: the training image set T, validation image set V
Train PLSA model on T
Select candidate annotations with some top confidence
  scores generated by the trained PLSA on V
Construct indicator vector $y$ for all images in V
Compute local evidence by Eq. (12) as well as the con-
  textual potentials of CRF by Eq. (13)
Learn $\Phi$ by maximizing the log posterior of the following
  equation by the deepest gradient descent algorithm
  $L(\Phi) = \sum_k \log(y^k \mid x^k) - \alpha_1^2/2\sigma^2 - \alpha_2^2/2\sigma^2$
Testing
Input: a test image I
Generate candidate annotations of I by the trained PLSA
Construct the corresponding indicator vector
Infer the indicator variable $y_i^* = \arg\max_{y_i} P(y_i \mid x; \Phi^*), y_i \in \{0,1\}$
Output: refined annotation results
```

Note that the indicator vector in the pseudo-code described above is constructed in such a way that variable y_i is true if the corresponding concept appears among the keywords with top 10 confidence scores and also in the ground truth labels, otherwise it is false.

3 Experimental results and analysis

In this section, experimental results and some analysis for the proposed PLSA-CRF will be reported. The experiment is conducted on the Corel5k dataset comprising 5000 images, in which 4500 images are used as training set and the remaining 500 images as testing set. Here, features similar to Ref. [6] are used since the focus of this paper is not on image feature selection. For fair comparison, each image is divided into a set of 32 × 32 sized blocks and a 36 dimensional feature vector for each block is extracted, which includes 24 dimensional color features computed over 8 quantized colors and 3 Manhattan distances, 12 dimensional texture features computed over 3 scales and 4 orientations. As a result, each image is represented as a bag of features, i. e., a set of 36 dimensional vectors. The following features of images are clustered by k-means algorithm and discretized into clusters, which are considered as visual words. Thus, the clustering process generates a visual-word vocabulary describing different local patches in images. The number of clusters determines the size of the vocabulary. By mapping all the blocks to visual words, we can represent each image as a bag-of-visual-words (or bag-of-visterms). Similar to Ref. [8], the dimension of the bag-of-visual-words for images is set to 1000 dimension in our experiment.

In addition, the visual language model is constructed to calculate Flickr distance so as to measure the semantic correlation between the annotation keywords. Without loss of generality, precision and recall metrics are utilized to evaluate the image annotation results. Furthermore, the top _ N precision and coverage rate[26] are adopted to measure the performance of annotation, in which top _ N precision measures the precision of top _ N ranked annotations for one image whereas the top _ N coverage rate is defined as the percentage of images that are correctly annotated by at least one word among the first N ranked annotations. Both of them can be defined as

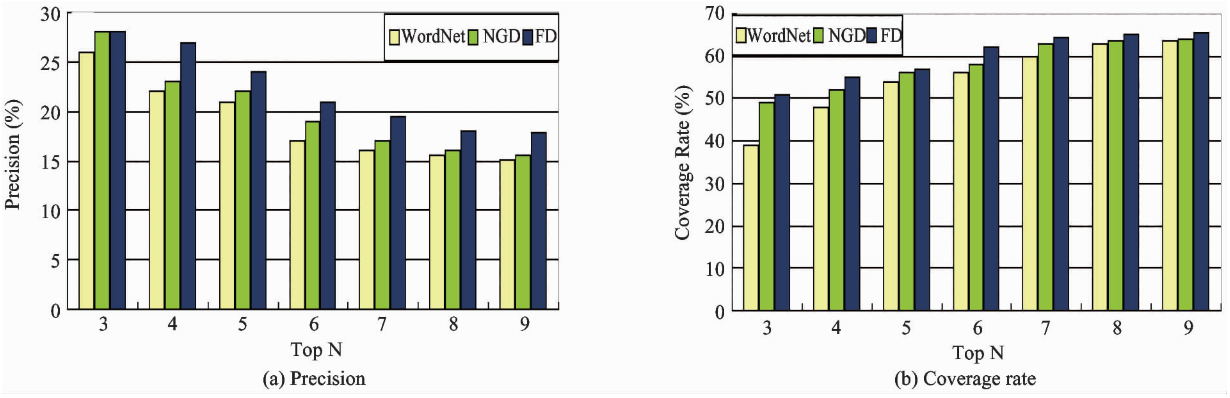
$$Top\_N\_P = \frac{1}{|T|}\sum_{i \in T}\frac{precision(i,N)}{N} \tag{14}$$

$$Top\_N\_C = \frac{1}{|T|}\sum_{i \in T} coverage(i,N) \tag{15}$$

where $precision(i,N)$ denotes the number of correct annotations in top _ N ranked annotations for image i, T is the test image set and $|T|$ denotes the size of T. *coverage*(i,N) judges whether image i contains correct annotations in the top _ N ranked ones. If at least one correct annotation of image i belongs to the top _ N annotations, then *coverage*(i,N) is set to 1, otherwise by 0. To evaluate the performance of final annotations, the precision and coverage rate are adopted together in our experiment.

3.1 Comparison of different measurements

To demonstrate the advantage of Flickr distance applied in the CRF model over the WordNet and normalized Google distance (NGD), we make use of WordNet, NGD and FD to define the contextual potentials for the conditional random field proposed in this paper respectively, and the corresponding results based on the complete set of all 260 words are illustrated in Fig. 1. It's easy to see that the top _ N precision descends gradually with the increase of N in Fig. 1(a) for three different measures. To be specific, PLSA-CRF based on the Flickr distance can get 8%, 23%, 14%, 24%, 22%, 16% and 18% as well as 1%, 17%, 9%, 11%, 15%, 13% and 15% precision improvements over that based on the WordNet and normalized Google distance respectively. On the contrary, it is also worth noting that the top _ N coverage rate displayed in Fig. 1(b) increases gradually when N is varied from 3 to 9 for the three different approaches. This fact suggests that the conditional random field based on the Flickr distance is apparently superior to the other two methods. The reason lies in that the FD is more precise for visual domain concepts and it can capture the visual relationship between the concepts instead of their co-occurrence in text search results.



(a) Precision

(b) Coverage rate

Fig. 1 Performance comparison of top _ N precision and coverage rate

3.2 Comparison with state-of-the-art results

We apply MATLAB 7.0 to implement the proposed PLSA-CRF model. The experiments are carried out on a 1.80GHz Intel Core Duo CPU personal computer (PC) with 2.0G memory running Microsoft windows XP professional. To validate the effectiveness of PLSA-CRF, we make a direct comparison with several previous approaches[4-9] except for RVM-CRF[13] because its experimental results cannot be accessed directly from the literature. Similarly, we compute the recall and precision of every word in the test set and use the mean of these values to summarize the model performance. The experimental results listed in Table 1 are based on two sets of words: the subset of 49 best words and the complete set of all 260 words that occur in the training set. From Table 1, it is easy to see that our model PLSA-CRF outperforms all the others, especially the first three approaches.

Table 1 Performance comparison of AIA on Corel5k dataset

Models	Translation	CMRM	CRM	PLSA-WORDS	PLSA-FUSION	MBRM	PLSA-CRF
#words with recall > 0	49	66	107	108	112	122	128
Results on 49 best words							
Mean per-word recall	0.34	0.48	0.70	0.76	0.76	0.75	0.78
Mean per-word Precision	0.20	0.40	0.59	0.58	0.65	0.73	0.75
Results on all 260 words							
Mean per-word recall	0.04	0.09	0.19	0.22	0.22	0.25	0.26
Mean per-word Precision	0.06	0.10	0.16	0.16	0.19	0.23	0.25

Alternatively, Table 2 presents some examples of the annotations (only four cases are listed here due to the limited space) generated by RVM-CRF and PLSA-CRF respectively. As can be seen from Table 2, the performance of PLSA-CRF is superior or highly competitive to that of RVM-CRF, which further demonstrates the effectiveness of PLSA-CRF proposed in this paper.

Table 2 Illustration of some annotation results obtained by RVM-CRF and PLSA-CRF

Image				
Ground Truth Annotation	mountain, lake, water, grass, ocean	flower, plant, leaves, garden	building, city, sky, ocean	grass, tiger, cat, forest
RVM-CRF Annotation	mountain, lake, water, building, ocean	grass, plant, flower, garden, people	building, landscape, sky, city, ocean	grass, plant, tiger, people, landscape
PLSA-CRF Annotation	mountain, lake, water, grass, ocean	flower, leaves, garden, plant, pergola	building, city, sky, landscape, ocean	tiger, grass, landscape, cat, plant

To further illustrate the effect of PLSA-CRF, mean average precision (*m*AP) is also applied as a metric to evaluate the performance of single word retrieval. Here, we only compare our model with CMRM, CRM, MBRM and PLSA-FUSION due to the *m*AP of other methods cannot be accessed directly. As shown in Table 3, our model is obviously superior to CMRM, CRM and PLSA-FUSION. Compared with MBRM, it can also get 7% and 3% improvements on 260 words and words with positive recall respectively.

Table 3 Ranked retrieval results based on one word queries

Mean Average Precision for Corel5k Dataset		
Models	All 260 words	Words with recall >0
CMRM	0.17	0.20
CRM	0.24	0.27
MBRM	0.30	0.35
PLSA-FUSION	0.26	0.30
PLSA-CRF	0.32	0.36

4 Conclusion

This paper presents a novel method for refining image annotation by integrating probabilistic latent semantic analysis with conditional random field. In particular, the confidence scores generated by the PLSA model and the Flickr distance rather than WordNet or NGD between two candidate annotations are applied to define the unary and binary potential functions for CRF, respectively. The experimental results on the Corel5k dataset show that our model is superior or highly competitive to several state-of-the-art approaches. In the future, we plan to introduce semi-supervised learning into our approach to utilize the labeled and unlabeled data simultaneously. In addition, we intend to employ other more complicated real-world image datasets, such as NUS-WIDE and MIRFLICKR to further evaluate the scalability and robustness of PLSA-CRF comprehensively.

References

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

[2] Cusano C, Ciocca G, Schettini R. Image annotation using SVM. In: Proceedings of the International Society for Optical Engineering, California, USA, 2003. 330-338

[3] Yang C, Dong M, Hua J. Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, New York, USA, 2006. 2057-2063

[4] Duygulu P, Barnard K, De Freitas J, et al. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark, 2002. 97-112

[5] Jeon L, Lavrenko V, Manmantha R. Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, 2003. 119-126

[6] Lavrenko V, Manmatha R, Jeon J. A model for learning the semantics of pictures. In: Proceedings of the 17th International Conference on the Advances in Neural Information Processing Systems, Vancouver, Canada, 2003. 553-560

[7] Feng S, Manmatha R, Lavrenko V. Multiple Bernoulli relevance models for image and video annotation. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, Washington, USA, 2004.

1002-1009

[8] Monay F, Gatica-Perez D. Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(10): 1802-1817

[9] Li Z, Shi Z, Liu X, et al. Fusing semantic aspects for image annotation and retrieval. *Journal of Visual Communication and Image Representation*, 2010, 21(8): 798-805

[10] Jin Y, Khan L, Wang L, et al. Image annotations by combining multiple evidence and wordnet. In: Proceedings of the 13th International Conference on Multimedia, Singapore, 2005. 706-715

[11] Wang C, Jing F, Zhang L, et al. Image annotation refinement using random walk with restarts. In: Proceedings of the 14th International Conference on Multimedia, California, USA, 2006. 647-650

[12] Wang C, Jing F, Zhang L, et al. Content-based image annotation refinement. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, Minnesota, USA, 2007. 1-8

[13] Wang Y, Gong S. Refining image annotation using contextual relations between words. In: Proceedings of the 6th International Conference on Image and Video Retrieval, Amsterdam, Netherlands, 2007. 425-432

[14] Liu D, Hua X, Yang L, et al. Tag ranking. In: Proceedings of the 18th International Conference on World Wide Web, Madrid, Spain, 2009. 351-360

[15] Xu H, Wang J, Hua X, et al. Tag refinement by regularized LDA. In: Proceedings of the 17th International Conference on Multimedia, Beijing, China, 2009. 573-576

[16] Zhu G, Yan S, Ma Y. Image tag refinement towards low-rank, content-tag prior and error sparsity. In: Proceedings of the 18th International Conference on Multimedia, Firenze, Italy, 2010. 461-470

[17] Makadia A, Pavlovic V, Kumar S. A new baseline for image annotation. In: Proceedings of the 10th European Conference on Computer Vision, Marseille, France, 2008. 316-329

[18] Guillaumin M, Mensink T, Verbeek J, et al. TagProp: discriminative metric learning in nearest neighbor models for image auto-annotation. In: Proceedings of the 12th International Conference on Computer Vision, Kyoto, Japan, 2009. 309-316

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

[20] Miller G, Fellbaum C. WordNet: An electronic lexical database. Cambridge: MIT press, 1998

[21] Cilibrasi R, Vitanyi P. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(3): 370-383

[22] Wu L, Hua X, Yu N, et al. Flickr distance. In: Proceedings of the 16th International Conference on Multimedia, Vancouver, Canada, 2008. 31-40

[23] Li W, Sun M. Semi-supervised learning for image annotation based on conditional random fields. In: Proceedings of the 5th International Conference on Image and Video Retrieval, Arizona, USA, 2006. 463-472

[24] Xu X, Jiang Y, Peng L, et al. Ensemble approach based on conditional random field for multi-label image and video annotation. In: Proceedings of the 19th International Conference on Multimedia, Arizona, USA, 2011. 1377-1380

[25] Huang Q, Han M, Wu B, et al. A hierarchical conditional random field model for labeling and segmenting images of street scenes. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, Colorado, USA, 2011. 1953-1960

[26] Li J, Wang J. Real-time computerized annotation of pictures. In: Proceedings of the 14th International Conference on Multimedia, California, USA, 2006. 911-920

Tian Dongping, born in 1981. He received his M. S. and Ph. D. degrees from Shanghai Normal University and Institute of Computing Technology, Chinese Academy of Sciences in 2007 and 2013, respectively. His main research interests include computer vision and machine learning.