

What is a Complete Set of Keywords for Image Description & Annotation on the Web

Xianming Liu, Hongxun Yao, Rongrong Ji, Pengfei Xu, Xiaoshuai Sun

Department of Computer Science, Harbin Institute of Technology

No.92, West Dazhi Street, Harbin, P. R. China, 150001

Mail: {liuxianming, yhx, rrji, pfxu, xssun}@vilab.hit.edu.cn Tel: +86+45186416485

ABSTRACT

Does there exist a compact set of keywords that can completely and effectively cover the image annotation problem by expanding from it? In this paper, we answer this question by presenting a complete set framework for image annotation, which is motivated by the existence of semantic ontology. To generate this set, we propose a cross model optimization strategy from both textual and visual information for topic decomposition, based on a so-called Bipartite LSA model, which minimize multimodal error energy functions in a probabilistic Latent Semantic Analysis model. To achieve complete set based annotation, we present a Gaussian-Kernel-Generative process based keyword generation procedure, which analogizes keyword annotation in a probabilistic generative manner. A group of experiments is performed on Washington University image database and 80,000 Flickr images with comparisons to the state-of-the-arts. Finally, potential advantages and future improvements of our framework are discussed outside the scope of topic modeling.

Categories and Subject Descriptors

H.3.3 [Information storage and retrieval]: Information Search and Retrieval

General Terms: Algorithms, Design, Experimentation.

Keywords: Keyword Selection, ontology, semantic items, image annotation

1. INTRODUCTION

Keywords generating and selection is capable to improve the performance for image annotation and retrieval. In this paper, we focus on presenting a general framework for keywords selection in image annotation. From the intuitive point of view, it is reasonable since the relationship and co-occurrence existed between different semantic items. From the ontology point of view, it holds true as well. Ontology is a consensus in language processing, cognitive theory and AI. It is believed that all the keywords can be generated by the objects or categories on upper levels in a given set, in form of that the keywords lying on the leaves of the ontology tree can be accessed by travelling the tree from root. By applying the generative rule on the directory of vocabulary needed to label a set of objects, we can use a much smaller scale instead of the original one. This is fully discussed in [1]. It is beneficial for natural language processing, and the same holds for image semantic understanding.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'09, October 19–24, 2009, Beijing, China.

Copyright 2009 ACM 978-1-60558-608-3/09/10...\$10.00.

Based on this principle, we propose the following assumption: as the volume of image collection increase, the number of required keywords for annotating them converges to a limit. To validate our assumption, we perform the experiment on *Corel* image database by counting the keywords used for annotating images in which all these images are labeled manually, as the target image set increases. The statistical results are shown in Figure 1. It can be seen from this figure that the curve tend to boost more slowly along with the annotation procedure. As an ideal experiment, it must be close to a limit to an infinity small degree if the training images are sufficient. From the keyword-counting trend, it is obvious that our assumption holds true under the theory of ontology of semantic correlation.

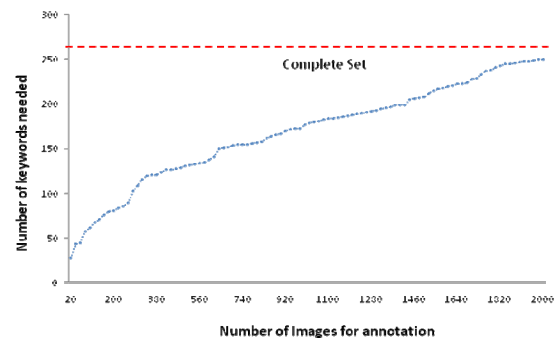


Figure 1 Number of Keywords needed for annotating images as the volume of image set increases.

The previous work on image annotation focus on topic models of images by treating them as documents to address this problem. These works can be mainly concluded into two groups: the probabilistic generative models and multi-model optimization methods. The former one supposes that the keywords or semantic items are generated from a single or mixture distribution under the guidance of latent variable – called latent semantic. It is a major approach used today for image annotation because that it sounds wonderful theoretically since the precise statistical inference. These models include Latent Semantic Analysis (LSA) [2], probabilistic LSA (pLSA) [3], LDA [4] and its variations [5] [6]. The multi-model approaches are mainly used in a web resource-oriented manner by optimizing multi-model for visual and text information. Outstanding performance in [7][9][10] has shown the effectiveness of the bipartite graph multi-model.

Despite the positive, training speed and accuracy are still bottom-neck for both approaches on real-world scenario, especially for large-scale Web search engines. The reason can be attributed to the problem we validated at the beginning, expressed specifically as redundancy and noise in keywords distribution. An original work covering this point is proposed in [8]. It indicates a framework to choose a set of most frequent keywords within a set of images with small semantic gap using clustering on “content-context K-nearest neighbor” matrix (KNN-C2). However, this

method bases on a simple assumption of more frequent keywords should be more important and lacks theoretical supporting and the measurable analysis.

This paper concentrates on to what extent the keywords converges to the limit. We also try to discover the content of these keywords in order to a more accurate and efficient image annotation strategy. To solve these problems, we firstly give the definition of *Complete Set* in form of the scaling of semantic gap as the error energy: A complete set here is a set of latent components with least count to span all the semantic items. Thus, it is essential for analyzing the content of multimedia resources and understanding the relationship between multiple models in different levels of semantic. Secondly, we perform an optimization on this model to analyze the methodology to obtain complete set. As a simple generalization, a Bipartite LSA optimization is proposed under the assumption of Gaussian generating procedure existing in image annotation. Similarly, other optimization rules are also available under our complete set theory and framework.

2. CRITERION FOR COMPLETE SET AND FRAMEWORK

2.1 The Definition

The complete set motivates on scaling topic models and keywords selection based on the ontology existed in semantic correlation.

Definition: A *complete set of semantics* is the Minimal complete set of semantic items (*visual topics*) that a set of images and documents covering in the semantic space as entities.

To discover the complete set of semantics, we firstly present a *minimizing error energy rule*. Its basic principle is to optimize the energy of distribution errors among different forms of representation of ontology semantics since it is believed that the existence of different forms of semantics should follow the same distribution. Thus, the *Complete Set* is a set of middle-level *latent components* with least error energy between both models of semantic existences. In other words, it represents the semantic in both visual and textual models. By adopting the representation of energy as the quadratic error, we get the following rule:

$$\text{CompleteSet } C_s = \arg \min_{C \subset O} \left(\frac{1}{2} \sum (M_1 - C)^T (M_2 - C) + \frac{\lambda}{2} |C|^2 \right)$$

Where O is the corpora of semantic ontology, and M_1, M_2 are the models of semantic existence. The target of this optimization is to find a compact subset C of corpora O which connect both visual and textual models together with least error. To avoid over-fitting and control the volume of complete set, a penalty $\lambda \cdot |C|^2 / 2$ is introduced into the optimization. The parameter λ is usually a small constant which has little effect on the optimization but would avoid over-fitting especially when the corpora are large.

Complete set can be used for automatic image annotation. Keywords needed for annotating a certain image collections can be generated from such set of semantic components. Therefore, image annotation is a generative procedure given the context of visual contents inferred from our assumption.

2.2 The Generation – Bipartite LSA Rule

Based on above framework, we propose a Bipartite LSA Optimization rule for image auto-annotation, aiming at validating the definition and theory of complete set in this section. LSA is usually used to discover topics of documents. Suppose matrix M_v is the visual feature – image matrix which represents the visual content for each image, while M_w is the keywords – image matrix

which reflects the text descriptions for images. Each image is organized in form of a column vector. It is believed that these two distinct types of organization of semantic items should follow the same underlying distribution since the co-occurrence relation between visual content and text description [6]. Under the basic assumption of LSA, the latent semantic items are available via Singular-Value-Decomposition as shown below:

$$M_v = T_v \cdot S_v \cdot D_v^T, M_w = T_w \cdot S_w \cdot D_w^T$$

Where matrix T_v, T_w are all dimensional left singular vectors and D_v, D_w are $n \times k$ dimensional right singular vectors while n is the volume of image collection and k is the pre-defined number of latent semantic items. Diagonal matrices S_v and S_w are composed by singular values of original matrix with decreasing orders.

By the definition of complete set, the target of optimization is to find a representation of semantics to minimize the error energy. Suppose it is in form of a matrix C , then the error energy becomes:

$$\frac{1}{2} (M_v - C)^T (M_w - C) + \frac{\lambda}{2} |C|^2$$

It is equivalent to minimize the target function as:

$$\frac{1}{2} (S_v - S)^T (S_w - S) + \frac{\lambda}{2} |S|^2$$

Where S is the diagonal matrix composed with the singular values of target matrix C ordered in a decreasing magnitude as well, which stands for on what components of a given sets the visual contents and text descriptions sharing the most significant latent semantic items. These are just the components of complete set. Therefore, the target of finding complete set for a given image-text collection can be formulated as optimizing the target function:

$$\arg \min_{k, \text{RankList}} \left\{ \sum_{d=1}^k \frac{(S_v(d) - S(d))(S_w(d) - S(d))}{d} + \frac{\lambda}{2} d^2 \right\} \quad (1)$$

Where $S(d)$ is the d^{th} element on the diagonal of S . And the value of d is the number of entities in complete set needed for describing the semantics items for current image collection, while *RankList* is a specific ranked list of items for all semantic entities. In other words the target of optimization is to find top d of all entities in what a ranked list. Here we adopt descending order for all elements in diagonal matrix S since larger values indicate components that are more effective. The solution of this optimization problem can be accessed by taking derivative of S from Eq. 1 shown below:

$$\frac{\partial \text{error}}{\partial S(d)} = \frac{1}{d} \sum ((S(d) - S_w(d)) + (S(d) - S_v(d))) = \frac{1}{d} \sum (2 \cdot S(d) - (S_w(d) + S_v(d)))$$

Therefore, solution of each dimension for this optimization is:

$$S(d) = \frac{1}{2} (S_w(d) + S_v(d))$$

In this condition, the explicit value of d can be discovered by choosing an optimal sub-matrix of S in an incremental manner.

3. ANNOTATION ON COMPLETE SET

3.1 Kernel-Based Image Annotation Rule

Image annotation can be viewed as the procedure of generating keywords set W from complete set C based on the context of visual representation V for each image in current collection, which can be interpreted in probabilistic point of view as:

$$\arg \max_W \{ p(W, C | V) \}$$

$$p(W, C | V) = \sum_{c \in C} p(c | V) \cdot p(W | c, \theta)$$

$$p(W|c, \theta) = \frac{p(c|W, \theta)p(W|\theta)}{p(c)} \propto p(c|W, \theta)p(W|\theta)$$

Without loss of generality, we adopt the Gaussian distribution for these two equations, denoted as function g . Then the above probabilistic generative expression can be written as the following matrix form, if we treat all the likelihood probability to be equal:

$$\begin{aligned} p(W, C|v) &= \frac{1}{\omega} g(C, v) \cdot g((C, W), C^{-1}) \\ &= \frac{1}{\omega} g(C, T_{v,d}|v) \cdot g(C^{-1}, T_{w,d}^{-1}) \end{aligned} \quad (2)$$

Matrices $T_{v,d}$ and $T_{w,d}$ are d -column sub-matrices of left singular matrices of M_v and M_w after optimization in Eq. 1 respectively, while v is the visual feature vector of input image to be annotated. Here the inverse on $T_{w,d}$ stands for Pseudo-inverse matrix.

By introducing Gaussian Kernel Function K_g instead of the probabilistic distribution g above, we present a kernel-based image annotation rule using complete set as shown below:

$$w = v^t \otimes (T_{v,d} \otimes C) \otimes (C^{-1} \otimes T_{w,d}^{-1}) \quad (3)$$

w is the set of keywords within the annotation results and operation \otimes is defined as:

$$\alpha \otimes [\beta_1, \beta_2, \dots, \beta_n] = [K_g(\alpha, \beta_1), K_g(\alpha, \beta_2), \dots, K_g(\alpha, \beta_n)]$$

α is a vector and $[\beta_1, \beta_2, \dots, \beta_n]$ is a matrix.

3.2 Gaussian Process for Keywords Selection

Keywords are selected automatically by using annotation strategy based on complete set. The selection procedure can be shown in a probabilistic way based on Gaussian assumption as follows:

Rule: the keywords which have high joint probabilities with items in complete set will be more valuable during annotation. It means that these keywords will response most significantly with respect to current complete set items, and be more meaningful in ontology.

Generation: Under the Gaussian assumption adopted above, this rule can be formulated as:

$$W_c = \{w | p(w|C) > \text{threshold} \wedge w \in W\}$$

Where the probability $p(w|C)$ can be calculated by statistical method from current training set, which relies on the Right-Singular Vectors of SVDs for both visual-image matrix and text-image matrix. Under the Gaussian assumption, this selection is a trim procedure for a multivariate Gaussian distribution upon complete set items denoted as c_i . The statistical variables are:

- 1) μ_v, μ_w - The mean vector for complete set items via Right-Singular Vectors of SVD for matrix M_v, M_w respectively.
- 2) $\Sigma_c = \varepsilon[(D_v - \mu_v)(D_w - \mu_w)^t]$ - Covariance Matrix for Complete Set C .

For each row-vector in Left-Singular Vector T_w , calculate the posterior probability and reserve top ones significantly consistent with current complete set items. The keywords selection strategy above is denoted as Complete Set Keyword Trim (CS-K-Trim).

4. EXPERIMENTAL RESULT

4.1 Preliminary

To validate and evaluate the performance of our complete set framework and Bipartite LSA rule for image annotation, a group of experiments are performed. We use a crawler to collect 80,000 images and tags as keywords from www.flickr.com, and randomly select 60,000 of them as training set while the other 20,000 as test set for performance evaluation. Except to compare with state-of-the-art approaches, the image annotation data set of Washington University is used as well.

Parameters for Gaussian Kernel Function can be estimated by sampling training set of image collection via calculating mean and variance on visual features and text distribution respectively. The parameter λ is set as 0.05 exponential.

The measurement of accuracy for image annotation is defined as the precision shown following:

$$\text{precision} = \frac{\text{number of keywords matched for ground truth of each image}}{\text{total number of keywords annotated for each image}}$$

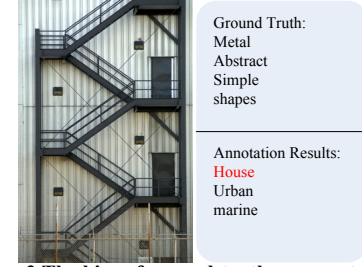


Figure 2 The bias of ground-truth vs. annotation.

In an approximate way, we choose the ground-truth as the labels downloaded with images from web pages, such as users' labels for images. Since the subjective bias for different users, these labels are not so accurate for evaluating annotation performance. It can be inferred that the actual precision is higher than the one we state here. An example is shown in Figure 2. The upper ones are ground truth while the below are annotation results. The annotated keyword "house" is correct although it doesn't appear in the ground-truth.

4.2 Validation of Complete Set Framework

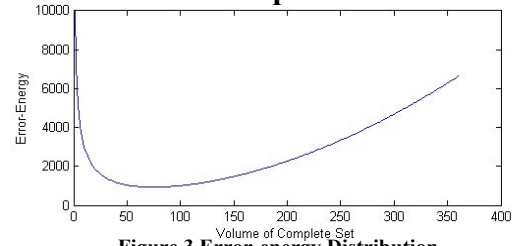


Figure 3 Error-energy Distribution.

To validate the correctness of our complete set framework, the following measurements are calculated during training procedure, as shown in Figure 3 and Figure 4 respectively. Figure 3 is the error energy for each volume of complete set from 2 to 200 defined in Eq.1, while Figure 4 shows the average precision of annotation using different size of complete set. They both convergent at the same point - $d = 75$, which means 75 visual topic components are consisted within the given image set. This accordance shows that the optimization rule is correct.

It is obvious from Fig 4 that the performance improvement slows down to nearly zero after the optimized point at $d = 75$, which indicates that the complete set is able to cover nearly all semantics within current images and text descriptions.

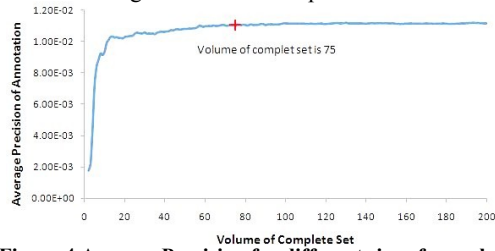


Figure 4 Average Precision for different size of complete set.

4.3 Performance Comparison

In this group of experiments, we compare our performance on complete set framework based on Bipartite LSA with the state-of-the-arts including pLSA [3] and approaches in [8] on Washington University Image Database. Figure 5 shows the experiment results.

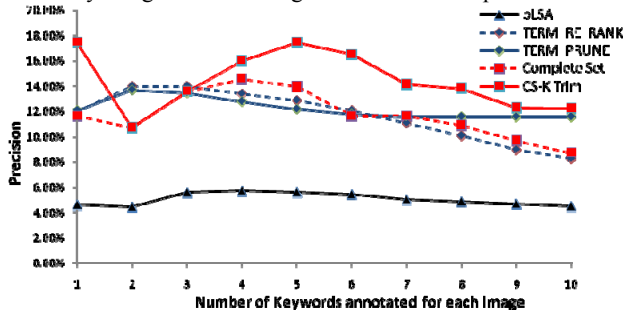






Figure 5 Comparison with State-of-the-arts

Where pLSA is the baseline, and Term_Re_Rank, Term_Prune are from [8]. We perform two different strategies of complete set on image annotation: without CS-K-Trim and the one with. It is obvious that our methods perform much better than the baseline, and are very close to the performance of Term_Prune. Since the latter are trained on a much huger image dataset containing 2.4 million images, our performance is much more satisfying trained on an 80,000 one and tested on a totally different Washington University image data set. This can be mainly attributed into two reasons: Firstly, our probabilistic generative assumption is correct based on pre-knowledge provided by all training samples, especially the Gaussian Kernel. Compared with approaches in [8], it is more reasonable to address image annotation problem into generative model instead of frequency-trimming dictionary constructing. Secondly, the cross-database annotation result shows that constructing a unified corpus is available to for different collections. For the images in a specified domain, specialized sub-corpus could be considered. The fluctuation of precision can be seen from Figure 5. This is because the noise existed in keywords list which is produced automatically via word frequent count. The highest precision appears at 5 keywords for each image. That is because most images for WU database contain 5 manual labels. Besides, the annotation of CS-K-Trim is more accurate and stable than the algorithm without. It may be because that some certain pre-distribution existing within keywords correlation. This fact also validates the existence of Complete Set which is proposed just based on this pre-distribution.

Table 1 shows some annotation results from WU image database.

Table 1 Annotation Results by CS-K-Trim algorithm

	Cannonbeach\Image34.jpg	Beach, Winter, Water, Thanks giving
	Greenlake\Image41.jpg	Spring, art, people, cat, trees
	Football\Image40.jpg	Football, Alabama, school
	Yellowstone\Image23.jpg	Waterfall, rural, decay, HDR

Finally, as an implementation of this theory, Bipartite LSA model is just one of the simplest one with limited ability. This can be attributed as the generalization capability of Complete Set. Other approaches may also be candidates, as long as the model satisfies the min Error-Energy rule under the Complete Set framework.

5. CONCLUSION

This paper analyzes the principle of keywords selection, which is essence for more quick and accurate image annotation, especially for web image search engines. A Complete Set framework is proposed to address this problem by defining an error energy function and an optimization is performed. Then a Bipartite LSA rule is presented as an implementation and the criterions for keywords selection and keywords generating are also presented. In further discussion, the generalization capability is mentioned and future work can be concluded as more practical and strict strategies under the Complete-Set framework.

6. ACKNOWLEDGMENTS

This work is supported by National Basic Research Program of China (2009CB320906), National Natural Science Foundation of China (60775024) and Specialized Research Fund for the Doctoral Program of Higher Education of China (20060213052).

7. REFERENCES

- [1] S. Zinger, C. Millet, B. Mathieu, G. Grefenstette, P. H  de, and P.-A. Mo  llic. 2005. Extracting an Ontology of Portrayable Objects from WordNet. In *Proceedings of the MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*, 2005.
- [2] T.K. Landauer, P.W. Foltz and D. Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 1998.
- [3] T. Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd annual international ACM SIGIR*, 1999.
- [4] D. Blei, A. Ng and M. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003, 3:993-1022.
- [5] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei and M. Jordan. 2003. Matching words and pictures. *Journal of Machine Learning Research*, 2003.
- [6] D. Blei and M. I. Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th Intl. ACM SIGIR*, 2003.
- [7] X. Rui, M. Li, Z. Li, W.Y. Ma and N. Yu. 2007. Bipartite graph reinforcement model for web image annotation. In *Proceedings of the ACM International Conference on Multimedia*, 2007.
- [8] Y. Lu, L. Zhang, Q. Tian and W.Y. Ma. 2008. What are the High-Level Concepts with Small Semantic Gaps?. *CVPR*, 2008.
- [9] Xianming Liu, Rongrong Ji, Hongxun Yao, Pengfei Xu, Xiaoshuai Sun, Tianqiang Liu. "Cross-Media Manifold Learning for Image Retrieval & Annotation". *ACM MIR* 2008, pp: 141-148, 2008.
- [10] Rongrong Ji, Hongxun Yao, "Visual & Textual Fusion for Region Retrieval: From Both Fuzzy Matching and Bayesian Reasoning Aspects," *ACM Conference on Multimedia Information Retrieval (MIR)*, pp.159-168, 2007.