Visual Topic Model for Web Image Annotation

Xianming Liu[1], Hongxun Yao[1], Rongrong Ji[1], Pengfei Xu[1], Xiaoshuai Sun[1], Qi Tian[2]

[1] School of Computer Science & Technology, Harbin Institute of Technology
No. 92 West Dazhi Street, Harbin, Heilongjiang, P.R. China
[2] Univeristy of Texas at San Antonio
One UTSA Circle, San Antonio
{liuxianming, yhx, rrji, pfxu, xssun}@vilab.hit.edu.cn, qitian@cs.utsa.edu

ABSTRACT

In this paper, we focus on image semantic understanding under large scale of image set, in which traditional approaches suffer from the limitations of scalability, tag correlation and noisy items. To solve these problems, a novel *Visual Topic Model* framework is proposed, via unsupervised clustering techniques. The framework aims at analyzing image semantics fusing both content and context, by considering tag correlations and ambiguities. In fact, the tags highly correlated in context may vary greatly in visual content and thus represent different semantics. Furthermore, a keyword selection and image annotation algorithm is also developed and applied to *Flickr* database with 175,770 images. Compared with the state-of-the-art methods, credible performance provides solid support for our framework.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing-indexing methods

General Terms

Algorithms, Theory, Experimentation

Keywords

Image Annotation, Visual Topics, Unsupervised Learning, Keyword Selection, Flickr, Image Retrieval

1. INTRODUCTION

The analysis of image semantics has become more and more important with the rapid development of online multimedia communities. However, the growing volume of multimedia data and the semantic gap pose a great challenge on further work. In image retrieval and annotation, the low precision is still a vital drawback for their practical applications. Except for the reason of incomplete understanding of images' content, it is partially because the noise existing in the original descriptions for images in training and annotation. One reasonable solution is keyword selection, with representative work in [1] by Lu et al. , which aimed at eliminating noisy semantic items and preserving the ones with smallest semantic gap, and annotating images with the selected keywords. It dramatically improved the performance of image annotation and retrieval as well as the training speed, especially when the database of images is huge and contains sufficient amount of keywords covering most user intentions.

ICIMCS'10, December 30-31, 2010, Harbin, China.

However, recently statistical research on keyword selection and distribution in [2] suggested that the correlation between keywords should also be considered into the selection, rather than regarding image annotation tasks as the mapping from visual features to high-level concepts merely. Thus, traditional strategies for image annotation and keyword selection are unilateral, which omit critical ontology existed between semantic items.

Within *NLP* and *IR*, topic models, aiming at finding a compact set of topics to describe the content of documents based on correlations between keywords, has been a credible solution for challenges in this domain recently. The classical work includes pLSA [3] by Hofmann, and LDA [4] by Blei. The topics are usually organized as set of keywords and their inside correlations. Topic Models improve understanding of documents' content at a higher level than traditional keyword level and make it available to achieve more precise information retrieval. Moreover, the generative processes in Topic Models reflected corresponding ontology to a certain extent.

The analysis for images' visual content and image retrieval followed the similar idea these years. The most representative work includes [5] by Barnard, which used a modified-LDA to generate a topic model coming from both image keywords and visual content, and adopted a hierarchical generative way for annotation tasks. In image annotation and keyword-based image retrieval, Liu proposed a theoretical framework in [6] to select topics from both image text descriptions and visual features, based on a mini-error-energy rule between content and context of images, named as *Complete Set of Keywords*. This framework also adopted generative approach to select keywords according to the complete set. Compared with the state-of-the-art methods, its performance is more stable especially under the cross-database conditions. To our knowledge, this is the first work that aims at selecting keywords based on content and context at the topic level for image descriptions.

Despite the positive, there still exist some critical problems. It can be concluded as the following two aspects:

1. *Scalability*: Current methods have limited ability as the image database is scaled up. This can be attributed to the strong prior assumptions on keywords and image distributions.
2. *Semantic Relationship*: Keywords are correlated, which is ignored in most current strategies, such as in the classical work [7] which followed a discriminative approach and required a pre-defined dictionary. To achieve more precise results, this correlation should be more emphasized.

In this paper, we focus on mining topics of given image collections considering both keywords' semantic correlations and diversities for images visual content, named as *VISUAL TOPIC MODEL*, in order to address the above two problems. The *Visual Topic* is defined as the semantic correlated set of keywords and

their corresponding visual content – the virtual *Averaged Images*. To facilitate its scalability, unsupervised clustering is adopted, which makes weak prior hypothesis on data distribution but work under large scale of data set. The experiments on large scale web image show that accurately mined visual topics are more helpful and suitable for the representation of images semantics. The visual topics are available for keyword selection, image annotation and on-line image label suggestion. In this paper, we address the keyword selection problem by choosing the items with the largest likelihood considering the visual diversities within visual topics, which is supposed to hold most information across the whole distribution.

2. FRAMEWORK OF VISUAL TOPIC MODEL

In this section, the general idea and framework for the proposed Visual Topic Model will be introduced. The visual topics are organized as clusters of context and visual content with similar semantics. Unfortunately, the optimization on these two sources of media under large scale is usually hard to deal with. Instead, we adopt a sequential way for context and content and it is more convenient in practice.

In order to mine the visual topics, the procedure can be mainly partitioned into two stages. It firstly divides keywords into groups according to their co-occurrence. It is believed that those keywords with high co-occurrence probabilities share more similar or related semantics. However, though similar in context, they may vary greatly in content. Therefore, further division is performed based on their visual content. Beyond basic visual similarities, the basic principles used in this framework include *Maxmin Principle* of visual diversities, which we will introduce in Section 4. Finally, the obtained sets of keywords and their corresponding visual content (*Averaged Images*) are organized as visual topics, which will present image content at a higher level beyond traditional topic models or merely content analysis.

In this paper, we focus on keyword selection and image tag refinement based on the selected keywords. To achieve this target, keywords with top correspondence probabilities within each *Visual Topic* are selected, expecting representation of these topics more significantly. Alternatively, a candidate list of selective keywords for each *Visual Topic* is also taken into consideration compared with the most straightforward approach above to improve performance. Furthermore, a generative probabilistic model at the visual topic level is developed.

3. KEYWORDS SEMANTIC CORRELATION

3.1 Preliminary & Denotation

Our basic assumption is that: keywords with top large correlation share similar semantics with large probabilities. Thus, it is a straightforward but effective idea to cluster these keywords together, representing the correlated semantics in terms of topics, and the topics are organized as keywords with their correlations as well. Similar with those in information retrieval, we treat image tags in a near-identical way. In this paper, tag correlation is evaluated according to their co-occurrence probabilities. The co-occurrence probability of tag *i* and *j* is defined as:

$$Co\_P(t_i,t_j)=\frac{Co\_Freq(t_i,t_j)^2}{Freq(t_i)\cdot Freq(t_j)} \quad (1)$$

where $Freq(t_i)$, $Freq(t_j)$ are occurrence frequencies for tag *i* and *j*, $Co\_Freq(t_i,t_j)$ is their co-occurrence frequency. It is bounded with [0,1] obviously and the larger co-occurrence probability indicates higher correlation between keywords.

Consequently, keyword correlation matrix is calculated following the same way over the whole image collections, and we denote it as *W*, with each element $w_{i,j}$ being the co-occurrence probability of keyword *i* and *j*. Furthermore, the co-occurrence matrix is an adjacent matrix for a certain weighted graph *G*. Each keyword (tag) is a node of the graph, while each element in matrix represents the graph weight for a corresponding pair of graph nodes. Any partition of graph *G*: $\{P_1, P_2, \ldots, P_k\}$, $P_1 \bigcup P_2 \ldots \bigcup P_k = G$ and $\forall i,j, P_i \bigcap P_j = \phi$, corresponds with a certain pattern of clustering on current dictionary. Therefore, the task of mining groups of semantically similar tags can be formulated as a partition on graph *G*.

3.2 Spectral Clustering on Keywords Semantic Correlation

To cope with the keyword correlations, we explore a clustering based algorithm following the properties suggested in [2]. Though hard in matrix operation, spectral clustering is more direct in the analysis of the performance and effectiveness than other strategies. Therefore, we adopt *Normalized Spectral Clustering* by Ng [8]. It is proved that normalized spectral clustering is equivalent to Normalized Graph Cut, which will avoid clustering most points into a large group and producing unbalance. The Laplacians Matrix L_{sym} is constructed from *Keyword Correlation Matrix W* as follows:

$$L_{sym} = I - D^{-1/2} W D^{-1/2}$$

where *D* is the diagonal matrix, with elements $d_i = \sum_{j=1}^{n} w_{i,j}$

Algorithm 1: Spectral Clustering on Keyword Correlation Matrix

Input: Keyword Distribution S, threshold t
Output: Tag cluster label array Idx, Cluster Spectrum $Spec$

```
Construct W from S:
w_{i,j} = Co_P(t_i, t_j), Spec = Null, clusterNum = 0;
Construct graph G from W: KNN / ε-Ball;
D = diag(d_i), d_i = Σ_{j=1}^{n} w_{i,j};
Construct L_sym = I − D^{-1/2} W D^{-1/2};
SVD on L_sym;
foreach eigenvalue of L_sym e_i do
    if e_i < t then
        Add corresponding eigenvector to Spec
        clusterNum++;
    end
end
Spectral Clustering on Spec, into clusterNum clusters;
return Idx, Spec
```

4. VISUAL DIVERSITY

4.1 Why Visual Diversity?

Review the clustering results shown in Figure 1(d), the problem comes that there exist some clusters being notably larger in tag amount than others. It is not because of the parameter chosen or

connexity of Correlation Graph, but the insufficiency of keywords' representation, or ambiguity of context.

A reasonable solution to this phenomenon is to further divide these clusters into sub-groups, which is supported by the truth that contextually similar items may vary greatly in their content, hence may be different in semantics. These sub-groups are organized as *Visual Topics*.

Traditional approaches merely focused on the similarities of visual contents, which lead the loss of topic diversification and the lack of richness in image retrieval results. Content diversification provides a discriminative and informative way to some extent. Song [9] acknowledged the need for visual diversities in image retrieval, and the similar idea is also used in [10] to re-rank the image retrieval results based on visual diversities of the return list. Zeigler [11] also studied topic diversification in order to reflect users' spectrum of interests. Thus, it is reasonable to take diversification of visual content into consideration.

4.2 Evaluation of Visual Diversification

In this part, we focus on evaluating the diversification of images' visual content. Suppose matrix V_i represents visual feature matrix for keyword i, with each row vector of V_i being visual feature vector for each image labeled by current keyword. We define the diversity of this keyword on visual representation as:

$$Diveristy(i) = tr(\Sigma(V_i)) \tag{2}$$

where Σ is the covariance matrix of V_i and $tr()$ the trace of the given matrix. Following this rule, a diversity vector is composed, denoted as *Diversity*. A larger value indicates more diverse in visual representation.

Furthermore, distance between two keywords' visual content matrices V_i and V_j are defined as:

$$dist(i,j) = \sqrt{(\mu(i)-\mu(j))\Sigma'(\mu(i)-\mu(j))^t} \tag{3}$$
$$\Sigma' = (\Sigma(i)\Sigma(j))^{-1/2}$$

$\tilde{\mu}(\cdot)$ is the human-constructed virtual *Averaged Image*, which is represented by the mean of all row vectors within visual feature matrix V. In fact, it is just for convenience in algorithms.

4.3 Diversity based Mining Algorithm

To perform visual topics mining, we proposed the following clustering algorithm modified from Maxmin clustering in [10]:

Algorithm 2: Diversity Based Clustering - Modified Maxmin

Input: $Diversity, V$
Output: Cluster Label array Idx_{div}, Number of Cluster $TopicNum$

```
Idx_div = NULL, TopicNum = 0 ;
centerList = Null;
foreach Cluster obtained in Algorithm 1 K do
    Construct Diversity(K) for each tag in K;
    clear centerList;
    while not all of Diversity(K) < threshold do
        t = argmax(Diversity(K));
        Diversity(K)_t = 0;
        Add K_t into centerList;
    end
    Assign each K to nearest center in centerList;
    clusterNum+ = #(centerList)
end
```

The *threshold* in the Algorithm 2 is given by the mean of all diversities across entire clusters. Generally speaking, this is a lightweight clustering algorithm capable for online application, and its basic idea is to maximize the intra-cluster diversities and minimize the inter-cluster distance. We called it the Maxmin principle for visual content. The obtained further divided clusters are one type of organization of the so-called *Visual Topics*.

5. EXPERIMENT

In this section, we firstly give two strategies for keyword selection and image annotation, and then experiments are shown.

5.1 Keyword Selection & Image Annotation

It is feasible to select informative keywords from *Visual Topics*. Our basic idea is to extract the most representative items within each visual topic. To achieve the evaluation of representativeness, the following criterion is proposed:

$$Pres(i) = \sqrt{(\mu(i)-\tilde{\mu})\Sigma(i)^{-1}(\mu(i)-\tilde{\mu})^t} \tag{4}$$

where $\tilde{\mu}$ is the mean of all *Averaged Images* within current visual topic. And the keyword with the smallest value is chosen to present the topic it lies in. As an alternative, we also provide strategy selecting K keywords for each visual topic, which increases the generalization of visual topic models.

A probabilistic algorithm for image annotation based on the selected dictionary is also proposed in this paper. We choose a candidate list of visual topics for target image *img* based on the visual content similarity between the *Averaged Image* for each visual topic, and put the top ones into the candidate list $T=\{t_1,t_2,...t_n\}$. And within the entire set of representative keywords $K=\{k_1,k_2,...k_N\}$ from visual topics t in T, we preserve those with largest correspondence posterior probabilities defined as follows:

$$p(img \mid k_i) = \frac{p(k_i \mid img)p(img)}{p(k_i)} = \frac{\sum_{j=1}^{n} p(k_i \mid t_j)p(t_j \mid img)p(img)}{p(k_i)} \tag{5}$$

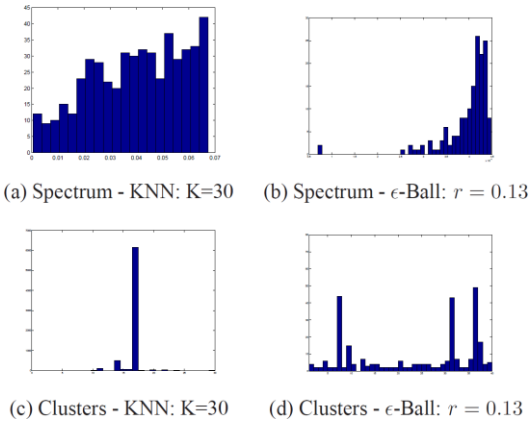
All these probabilities can be calculated from density estimation, *KNN* or inverse of distance, and we will not discuss the detail here due to the space limit.

5.2 Database & Experimental Result

The experiments are performed on crawled *Flickr* image database with 175,770 images. Without loss of generality, we use global features by extracting 360 dimensional color histogram and 8 dimensional texture co-occurrence features as visual features for each image.

5.2.1 Graph Construction: KNN vs. ε - Ball

Graph construction is a key factor affecting the performance and stability of clustering algorithm. Common methods include *K*-Nearest Neighbors (*K*-NN) and ε-Ball. The basic idea of both is to preserve the local topology for data points embedded in high-dimensional manifold. However, different strategies may cause different results during *Visual Topic* mining. Figure 1 shows experimental results of spectral clustering based on *K*-NN and ε-Ball respectively. (a) and (b) shows the distributions of spectrum, from which it is obvious that spectrum for KNN is dispersive while the one for ε-Ball more accumulational. (c) and (d) shows the distributions of clustering results for these two strategies, and the clusters for *K*-NN focus on a few connected components but those for ε-Ball are more averaged.



(a) Spectrum - KNN: K=30 (b) Spectrum - ϵ-Ball: $r = 0.13$

(c) Clusters - KNN: K=30 (d) Clusters - ϵ-Ball: $r = 0.13$

Fig. 1. Affection of graph construction on spectral clustering. In (a) and (b) the x-axes represent the interval of eigenvalues while y-axes the number of eigenvalues who lies in corresponding interval. In (c) and (d) the x-axes represent the ID of clusters while y-axes the number of points belongs to these clusters.

The reason comes from the nature of these two approaches. In fact, there are always tags which are isolated points either by wrong-written or stop-words from being visually connected, the Sparsity. Yet *K*-NN preserves *K* relationships for each node and connects unwanted components into adjacency, which lead to the result above - the majority nodes are clustered together and lose their discriminations. On the contrary, ε-Ball solves this problem reasonably by preserving the feasible link only. In the rest experiments, the graph construction adopts ε-Ball with *r>0.13*.

5.2.2 Experiment Results & Discussion

After the first step, spectral clustering on *Keyword Correlation Matrix*, totally 40 groups of tags are obtained. We preserve the clusters with fewer than 5 keywords and further divide the others based on Algorithm 2. Figure 2 shows the distribution of diversities for tags defined by Equation 2 over all keywords in the collection. It is obvious that our evaluation of tags' diversities is significant and discriminative for informative tags.

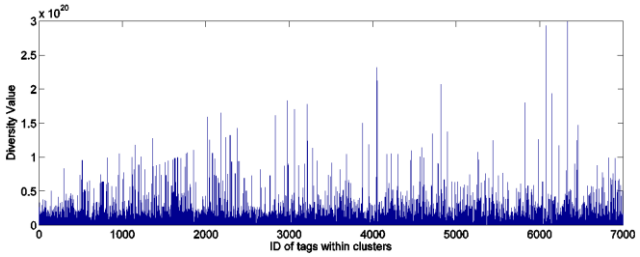


Fig. 2. Distribution of Diversities defined by Equation 2.

We apply the mined visual topics from Flickr image collection to annotate *Washington University Database*, which is manually labeled and capable of more credible ground-truth. Mean average precision (MAP) is adopted with annotating 1-10 tags for each image respectively. The reason to use MAP instead of Precision and Recall are: 1) Precision need to manually label all the test data which is not practical; 2) We only download part of web images, and the recall on such a down-sampled set can not reflect the true evaluation of the whole web.

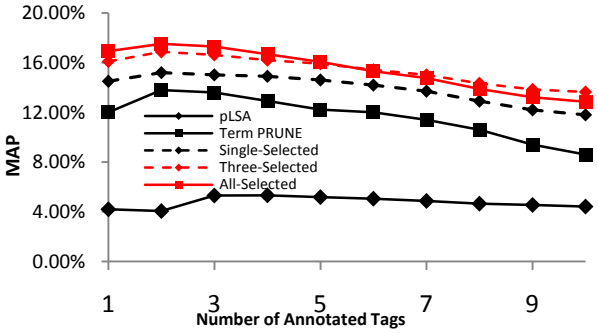


Fig. 3. Performance of experiments compared with state-of the-arts and different strategies proposed in this paper.

Detailed experimental results are shown in Figure 3, where pLSA is the baseline; **Term_Prune** is from [1]. We also select 1, 3 and all tags from each visual topic and compare their performances, denoted as Single-Selected, Three-Selected and All-Selected, respectively. Considering the cross-database situation, our performance is more stable and satisfactory. It is also essential to point out that, selecting appropriate number of keywords will benefit for improving performance, but reducing the computational cost instead of using all tags. From Figure 3, selecting three tags improves performance significantly, and is more stable because of the elimination of noisy tags. On the other hand, using all keywords does not improve the performance much but reduces the accuracy of annotation as the number of annotated tags increases.

For the future work, two aspects are suggested: 1) the dynamic process and life cycle of web communities should be considered, thus a dynamic visual topic model will be proposed. 2) the proposed work uses only global features, and some local descriptors may be more helpful for object level applications.

6. CONCLUSION

We firstly analyze the current situation of image retrieval and propose two challenges for new techniques. To cope with these problems, a visual topic model is developed via unsupervised approach to facilitate its scalability, by organizing images' semantics considering the correlations between context and diversification of visual content. Algorithms for keyword selection and image annotation are further developed. A group of experiments on Flickr image collection and *Washington University* image set validate its effectiveness. For future work, dynamic visual topic model and usage of local descriptor at object level will be investigated.

7. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No.61071180 and No. 60775024), National Basic Research Program of China (No. 2009CB320906) and Space Science Foundation (No. 20105577015).

8. REFERENCES

[1] Y. Lu, L. Zhang, J. Liu, and Q. Tian. Constructing lexica of high-level concepts with small semantic gap, in *IEEE Transactions on Multimedia*, 2010.

[2] X. Liu, H. Yao, and R. Ji. Exploring statistical properties for semantic annotation: Sparse distributed and convergent assumptions for keywords. in *IEEE ICASSP,* 2010.

[3] T. Hofmann. Probabilistic latent semantic indexing. in *ACM SIGIR*, 1999.
[4] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent dirichlet allocation. *JMLR*, vol. 3, pp. 993–1022, 2003.
[5] K. Barnard, P. Duygulu, D. Forsyth, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, vol. 3, 2003.
[6] X. Liu, H. Yao, R. Ji, P. Xu, and X. Sun. What is a complete set of keywords for image description & annotation on the web. in *ACM Multimedia*, 2009, pp. 613 – 616.
[7] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE Trans. PAMI*, 2008.
[8] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. in *NIPS* 2001.
[9] K. Song, Y. Tian, W. Gao, and T. Huang. Diversifying the image retrieval results. in *ACM Multimedia*, 2006.
[10] R. H. van Leuken, L. G. Pueyo, X. Olivares, and R. van Zwol. Visual diversification of image search results. in *WWW*, 2009.
[11] C. N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. in *ACM WWW*, 2005, pp. 22–32.