Image Annotation by Hierarchical Mapping of Features

Qiankun Zhao   Prasenjit Mitra   C Lee Giles
College of Information Sciences and Technology
Pennsylvania State University, University Park, PA
{qzhao, pmitra, giles}@ist.psu.edu

ABSTRACT

In this paper, we propose a novel approach of image annotation by constructing a hierarchical mapping between low-level visual features and text features utilizing the relations within and across both visual features and text features. Moreover, we propose a novel annotation strategy that maximizes both the accuracy and the diversity of the generated annotation by generalizing or specifying the annotation in the corresponding annotation hierarchy. Experiments with 4500 scientific images from Royal Society of Chemistry journals show that the proposed annotation approach produces satisfactory results at different levels of annotations.

Categories and Subject Descriptors

H.4.8 [**Image Processing and Computer Vision**]: Scene Analysis object recognition; H.3.3 [**Information Storage and Retrieval**]: Information Retrieval search process

General Terms

Algorithm, Experiment, Performance

Keywords

Image Annotation, Hierarchical Relation, Feature Mapping

1. INTRODUCTION

Automatic image annotation is an important problem given the fact that annotation based image retrieval outperforms the content-based image retrieval [1] and only few of the images on the web are annotated. Most existing annotation approaches proposed to learn the mapping between low-level visual features and keywords using co-occurrence, correlation, and probabilistic models [2, 4].

However, most of the existing image annotation approaches ignored the relations between features within the visual features and textual annotations. That is, the visual or textual context, which is reflected by relations within visual/textual features, plays an important role in determining the mapping model between visual and textual features. In this paper, we propose to construct a hierarchical mapping model between visual and textual features of images by exploring relations between features within and across the visual and textual dimensions. More importantly, we propose a novel

annotation strategy to maximize the diversity and accuracy of the predicted annotations based on the hierarchical mapping model.

2. HIERARCHICAL IMAGE CLUSTERING

First, we propose to take into account the relations among features within the visual dimension and textual annotation to build two cluster hierarchies. An image usually contains multiple objects and the correlations among objects is expected to improve the annotation. We use hierarchical clustering because each cluster in the cluster hierarchy is expected to be characterized by a subset of distinguishing features of these images in the cluster. The characteristics of an image is the sum of the distinguishing features of all image clusters, which the image belongs to. Images are clustered into two hierarchies based on the following visual features and textual annotation features, respectively.

For each image, the color, texture, and shape features are extracted as the visual features. The color features consist of 32 color histogram and cumulative histogram features, 36 gray-level co-occurrence features extracted using the co-occurrence image matrix. Texture features are extracted by calculating the means and variations of the filtered image regions on 8 orientations at 6 scales. Shape features include edge-map-based features and line features.

Figure caption, text references, and surrounding text in the scientific papers are extracted as image textual features. The text segments are takenized, and part-of-speech tags are added, stop words are removed, stemming of words is also applied. As a result, for each type of text annotation, a term vector is constructed for the corresponding image. Then, the *term frequency* and *inverted image frequency* is used as the weight of each term in the vector.

To explore the hierarchical relations between images, we propose to represent the set of images as a graph, $G = (V, E)$, where each vertex represents an image, each edge denotes the similarity between the pair of images it connects. Then, the graph partition algorithm proposed by Shi and Malik [3] is applied to cluster images into small groups, whereas the hierarchical relations are constructed. For both the visual feature-based clustering and the annotation-based clustering, each image is represented as a vector of features and the cosine similarity between the vector representations is taken as the similarity between two images.

3. CONSTRUCTION OF MAPPING MODEL

Given an image cluster in the hierarchy, there are two types of features: discriminative features and non-discriminative

features. We propose the conditional Kullback-Leibler divergence metrics to measure the discriminative power of feature subsets with respect to the child image cluster and the parent image cluster. By maximizing the conditional KL-divergence, for each image cluster in the image hierarchy, the corresponding discriminative features can be extracted. The conditional Kullback–Leibler divergence is defined as:

$$D_{\mathrm{KL}}(P\|Q|f_i) = \int_{-\infty}^{\infty} p(x|f_i) \log \frac{p(x|f_i)}{q(x|f_i)} \, dx$$

Then, to calculate the strength of the links between image clusters in the two image hierarchies, initially they are connected based on the common images. Basically, the weights of the links are measured by the mutual information of two clusters X and Y. Note that here both X and Y are represented by the discriminative features rather than the entire set of visual or textual features. The idea is: the larger the mutual information is, the stronger the correlation is between the corresponding discriminative features. For a given visual/textual cluster, we rank the corresponding textual/visual clusters based on the values of the mutual information.

4. ANNOTATION STRATEGY

Given the mapping model between visual features and textual annotation features, the goal of image annotation is to provide as complete and diverse as possible and as accurate as possible annotations.

Based on the mapping model, there will be a ranked list of textual annotation clusters that correspond to a given image. One goal of the image annotation is to produce as diverse as possible annotations, We define the diversity of annotation as:

$$A_d = Max(\sum_{a_i \neq a_j \in A} Dist(a_i, a_j))$$

where A_d is the diversity of the annotation results, $Dist(a_i, a_j)$ is the distance between two annotation clusters and is defined as $Dist(a_i, a_j) = min(|a_i| - |a_c| + |a_j| - |a_c|)$, where $|a_i|$ is the depth of the cluster in the hierarchy and a_c is the common ancestor of a_i and a_j.

To maximize the accuracy of the predicted annotation, both the strength of the relation between the predicted annotation cluster and the image cluster and the depth of the corresponding annotation clusters are taken into account. The accuracy of annotation is defined as:

$$A_a = Max(\sum_{a_i \in A} I(image, a_i) \times |a_i|)$$

where A_a is the accuracy of the annotation results, $I(image, a_i)$ is the strength between the annotation cluster and the $image$, which is represented as the visual image cluster. The final annotation will be based on the combination of the diversity and the accuracy of the annotation results.

$$A = Max(\alpha \cdot A_d + \beta \cdot A_a)$$

where α and β are the weights of the diversity and accuracy, and $\alpha + \beta = 1$.

5. PERFORMANCE EVALUATION

To evaluate the performance our proposed image annotation approach, training and testing images are from the Royal Society of Chemistry. We extracted 4500 images, 4000 of them are used as training data and 500 are used for testing.
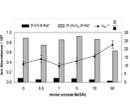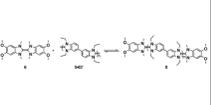

Figure 1: Examples of Image Annotations

Figure 1[1] presents three image examples with the original and predicted cluster IDs and keyword annotations($\alpha = 0.3, \beta = 0.7$). Three types of images are used as representatives. The results show that in the image cluster ID level, our prediction is more accurate than in the keyword level. The predicted cluster ID and the annotation cluster ID are the clusters with the maximum similarity. The list of keywords are the top-5 keywords with largest sums of weights.

The following experiments have been conducted: (1) using visual features to predict the annotation clusters, (2) using visual features to predict the detail annotation keywords; (3) annotation of visual images with partial knowledge such as cluster ID, top-1 keyword, and top-2 keywords. The results are shown in Table 1. It can be observed that our annotation approach can produce stratificatory results at both the cluster ID and keywords levels. Partial annotation can improve the quality of full annotation generated by our algorithm. The cluster ID improved the quality a bit, whereas the first keyword improved the annotation substantially and most significantly.

Partial Annotation	Precision	Recall
No Annotation	0.81	0.79
Cluster ID	0.84	0.81
Cluster ID, Top-1 Keyword	0.88	0.84
Cluster ID, Top-2 Keywords	0.91	0.89

Table 1: Performance of Partial Annotation

6. CONCLUSION

In this paper, we propose the first approach of image annotation by utilizing not only the correlations between features but also the correlations between features within the same modality. We propose a novel annotation prediction method that maximizes the diversity and accuracy. Experiments with real data show that the proposed image annotation approach produces satisfactory results.

7. REFERENCES

[1] T. A. S. Coelho, *et al.* Image retrieval using multiple evidence ranking. *IEEE T KDE*, 16(4):408–417, 2004.
[2] V. Lavrenko, *et al.* A model for learning the semantics of pictures. In *NIPS*, 2004.
[3] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE T PAMI*, 22(8):888–905, 2000.
[4] R. Zhang, *et al.* A probabilistic semantic model for image annotation and multi-modal image retrieva. In *ICCV*, 846–851, 2005.

---

[1]Images are from the Royal Society of Chemistry