Image Annotation with Relevance Feedbacks

Cheng-Chieh Chiang (江政杰)
Department of Information Technology,
Takming University of Science and Technology, Taipei, Taiwan
e-mail: kevin@csie.ntnu.edu.tw

Abstract

This paper presents a novel approach for image annotation with relevance feedback to assist the user in annotating semantic labels for images. Our design for image annotation is based on a semi-supervised learning for building hierarchical classifiers associated with annotation labels. We construct individual hierarchical classifiers each corresponding to one semantic label that is used for describing the semantic contents of the images. Our semi-supervised approach for learning classifiers reduces the need of training images by use of both labeled and unlabeled images. We adopt hierarchical approach for classifiers to divide the whole semantic concept associated with a label into several parts such that the complex contents in images can be simplified. We also describe some experiments to show the performance of the proposed approach.

Keywords：Image Annotation, Relevance Feedback, Semi-supervised Learning

1. Introduction

Image understanding and retrieval [6][12] have become a very active research area since the 1990' s due to the rapid increase in the use of digital images. The most common approach to represent an image is to extract a set of visual features, either for the entire image or for its regions. However, humans recognize an image by use of their perceptions, not image features. The inconsistence between the low-level features extracted from images and the high-level concepts involved in human perception is called semantic gap problem.

Semantic gap is one of the most difficult problems in image understanding. In recent, many researchers have focused on the issue of semantic gap. Image annotation, which discovers the semantic contents from images, may be potential for bridging the semantic gap. The goal of image annotation is to annotate one or several labels to an image to describe the semantic contents of the image. In other words, image annotation extracts the semantic-based image features instead of the low-level visual features.

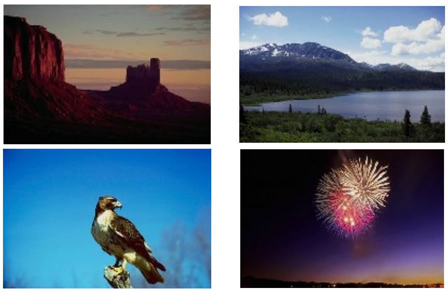


Figure 1: Different image contents with the same label ”sky”.

Image annotation is helpful to many applications, e.g., additional metadata within images for retrieval, or archiving personal photos. Unfortunately, it is a difficult task to build a model that can describe the contents of images with semantic labels. Regarding a simple case that images with a single label involve the same semantic meaning, the contents are not often homogeneous. For example, Figure 1 shows the four images that contain the same label ”sky”, but their semantic contents are very different – including sunset, cloud or cloudless, blue sky, and night. Obviously, it must be more complex if many kinds of labels are mixed. That is the main reason that most of the state-of-the-art approaches cannot annotate images well automatically.

Our opinion is to involve human feedbacks in image annotation because human should make the final decision for the semantic concept of images. Relevance feedback [13] is a query modification technique that attempts to capture the user's precise needs through iterative feedbacks and query refinement. In information retrieval, relevance feedback has been widely used in many researches to design powerful retrieval tools. This paper presents an annotation approach that is involved in an interactive scheme of feedback to assist the user in annotating images.

Image annotation is often considered a supervised learning problem in many state-of-the-art methods [4]. A main limit of the supervised learning approach for image annotation is that a large number of training images is necessary to avoid overfitting. However, it is often difficult to manually annotate a large set of images. Moreover, the number of labeled images must be also small at the beginning of annotating images. This limit motivates us to design a semi-supervised approach for image annotation by integrating labeled and unlabeled images to reduce the need of the training images. On the other hand, we build individual hierarchical classifiers each of them associated with a semantic label. This method can make the system more flexible because only the new classifier needs to be re-trained if a new label is added. Using an individual classifier with a label can reduce the complexity of the semantic contents for images, and the hierarchical approach can divide the whole concept within a label into several parts that could represent the different contents of images.

This paper is organized as the follows. Section 2 introduces related works for this research. Section 3 formulates our problem and Section 4 presents the overview of our approach. Then, the details of the classifier training and the computation of confidence values are described in Section 5 and 6, respectively. Section 7 presents our experiments to show the effectiveness of our approach, and Section 8, in final, draws the conclusion and the future work.

2. Related Works

Image annotation is a mapping from digital images to annotation keywords. Some of state-of-the-art works for image annotation and concept detection were provided in [6]. Mori et al. proposed a co-occurrence model to count the frequency of keywords and image-regions [14]. Duygulu et al. designed a translation model to translate images to labels [7]. Besides, Relevance Model [11], Cross-Media Relevance Model (CMRM) [9], Multiple Bernoulli Relevance Model (MBRM) [8], and Image-Keyword Document Model (IKDM) [15] are other famous models for image annotation. Chang et al. designed an approach called soft annotation to give images a confidence level for each trained semantic label [5]. Carneiro and Vasconcelos formulated image annotation as a supervised learning problem [4]. Jin et al. designed a K-means clustering with pair-wise constraints for image annotation [10]. Srikanth et al. proposed methods for image annotation by use of a hierarchy defined on the annotation labels derived from a textual ontology [14].

This paper designs a semi-supervised learning with relevance feedback for image annotation. Hence, we briefly review the related work of semi-supervised learning and relevance feedback. Semi-supervised learning in general is defined by using both labeled and unlabeled data for learning, and there are good reviews in [2], [1], and [16]. In this paper, we design the learning model based on the unsupervised K-means clustering and apply labeled images to evaluate the clustering. Relevance feedback is a query modification technique that attempts to capture the user's precise needs through iterative feedbacks and query refinement [13]. Relevance feedback has been widely used for image retrieval recently [6][12], and we apply relevance feedback to assist the user in image annotation.

3. Formulation

Let the entire dataset, denoted as D, contain M images. Suppose that K annotation labels $\{L_1, ..., L_K\}$ are predefined to describe the semantic contents of the images. Because the number M is usually huge, it is hard to annotate all images in D manually. However, it is also difficult to build a model to automatically annotate images within many labels. Hence, our approach allows that the user annotates a single label at a time.

Assume that the user annotates images for label L_k, $1 \le k \le K$. We denote all labeled images associated with L_k as D_k, images labeled without L_k as $D_k^{'}$, and other unlabeled images as D_U. Note that $M = D_U \cup D_k \cup D_k^{'}$ for each $1 \le k \le K$. Our goal is to build an annotation classifier C_k associated with label L_k. Table 1 lists these notations used in this paper.

Table 1. The notations used in this paper.

M: the number of images in the entire dataset

$\{L_1, ..., L_K\}$: the predefined set of labels for describing the semantic contents of images

D_k: the labeled images associated with label L_k.

$D_k^{'}$: the labeled images that are not associated with label L_k.

C_k: the annotation classifier associated with label L_k.

N: the number of returned images at each iteration of relevance feedbacks.

4. Overview of Our Approach

In this paper, our basic idea is like a retrieval task – (i) the user submits which label she/he wants to annotate, (ii) the system returns images to the user with the most confident for the label, and (iii) the user assigns which images are relevant. This method focuses on a single label for image annotation at the same time because the user could annotate images more consistent in semantic contents.

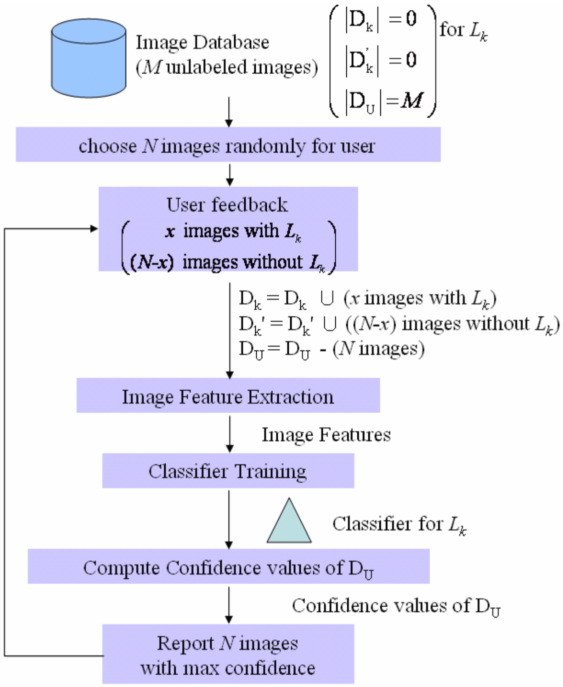


Image Database
(M unlabeled images)
$\begin{pmatrix} |D_k| = 0 \\ |D_k^{'}| = 0 \\ |D_U| = M \end{pmatrix}$ for L_k
choose N images randomly for user
User feedback
$\begin{pmatrix} x \text{ images with } L_k \\ (N\text{-}x) \text{ images without } L_k \end{pmatrix}$
$D_k = D_k \cup$ (x images with L_k)
$D_k' = D_k' \cup$ ((N-x) images without L_k)
$D_U = D_U$ - (N images)
Image Feature Extraction
Image Features
Classifier Training
Classifier for L_k
Compute Confidence values of D_U
Confidence values of D_U
Report N images
with max confidence

Figure 2. The flowchart for the proposed image annotation method with relevance feedback.

Figure 2 shows the flowchart that describes our interactive process for image annotation. The user annotates images with one label at a time within interactive feedbacks. Considering a single label L_k, we do not have any labeled images at the beginning of annotation, i.e., $|D_U| = M$ and $|D_k| = |D_k^{'}| = 0$. The user specifies all positive images, D_k, for label L_k displayed by the system, and then the other non-specified images are negative, $D_k^{'}$. Next, we mix D_U, D_k, and $D_k^{'}$ to train a hierarchical classifier, denoted as C_k, for label L_k using a semi-supervised clustering. Then, all unlabeled images are tested by the classifier

C_k to compute the confidence values of the images associated with label L_k. Finally, the system returns N (N=100 in our experiments) unlabeled images with the highest confidence values to the user to make the decision of the annotation.

This work designs an interactive method to assist the user in image annotation. In general, we often have only few positive examples at the beginning iterations in the relevance feedbacks. That will make the learning difficult for overfitting. Hence, we integrate unlabeled images into the training images for the classifier training to avoid this problem. Also, we adopt the hierarchical approach to build a classifier associated with each of labels. The main reason is that we divide the whole semantic concept of images with a label into several sub-concepts by use of the hierarchical classifier such that the complex contents, illustrated as Figure 1, of images can be simplified. Moreover, our proposed method by use of individual classifiers for image annotation makes the system flexible because it is independent of the number of labels.

5. Training

Table 2 shows the algorithm that constructs the hierarchical classifier C_k for label L_k. In this algorithm, the root node N_{root}^k of the tree C_k initially contains the mixture of images D_k, $D_k^{'}$, and D_U. In most cases, $|D_k^{'}| >> |D_k|$; hence we randomly ignore some of negative-labeled images such that $|D_k^{'}| = |D_k|$ to avoid the imbalance problem in the training. In the algorithm, we first decide which node needs to be split. If a node needs to be split, we go on to decide how many branches are appropriate to split the node. Here, K-means clustering is applied to divide a node into several child nodes. We try a range of branch number and calculate a score for each branch number to select the best one.

Table 2: The algorithm of constructing the hierarchical classifier C_k for label L_k.

```
Input: unlabeled images $D_U$, positive images $D_k$, and negative images $D_k^{'}$
Output: a hierarchical classifier $C_k$ for label $L_k$
Initialization: root node $N_{root}^k$ contains $D_U \cup D_k \cup D_k^{'}$
// $N_i^k$: node i for the hierarchical classifier $C_k$.
// construct the tree by splitting each node $N_i^k$.

1. for each leaf node $N_i^k$ not fitting the **stopping condition** {
2. for z = 2 to b {
// b is the max number of the trying ranges
3.     **node splitting method** to divide $N_i^k$ into z classes.
4.     compute **score**$(N_i^k, z)$
// evaluate how many branches are appropriate for node $N_i^k$.
    }
5.     $z_i^k = \arg\min_z score(N_i^k, z)$
6.     $N_i^k$ is divided into $z_i^k$ classes by use of k-means clustering
       // $N_i^k$ is divided into, w.l.o.g., $Nc_{i,1}^k, ..., Nc_{i,z_i^k}^k$.
}
```

Our proposed semi-supervised approach learns the classifier in the two ways: (i) evaluate the stopping criteria for node splitting according to the positive and negative images in the node and (ii) split a node by use of the mixture of labeled and unlabeled images in order to cover more information in learning. Finally, each leaf node in a hierarchical classifier represents a sub-concept for the label.

In the algorithm, we need to design three tasks: (i) the node splitting method (line 3), (ii) the stopping condition (line 1) which checks whether a node needs to be split, and (iii) the score function (line 4) which evaluates how many branches for the node to split are appropriate. Note that we use the two notations given a node N_i^k: d_i is the number of positive images in the node and $d_i^{'}$ is the number of negative images in the node.

5.1. Node Splitting

An unsupervised clustering is used to divide node N_i^k into several classes. Here we used *K*-means clustering. To employ it in our work, an image should first be converted into be a vector of image features. We adopt the model of visual words [3] to build the region-based representation for an image, which is briefly described as follows. All images are first segmented into a set of regions, and then feature vectors are extracted from these regions. The region features can be divided into *v* clusters (using another *K*-means clustering) in the feature space. The *v* clusters are viewed as visual words for representing images. An image can be then represented by a *v*-D vector that is accumulated by the appearance of visual words in the image. Note that either features or unsupervised clustering method are independent of the proposed algorithm.

5.2. Stopping Condition

A node containing consistent or unified information means that this node is high confident to classify data. Hence, a node shouldn’t be split if it only contains either positive or negative data. We define the stopping condition of splitting a node as:

$$Stop(N_i^k) = \begin{cases} true, \ \text{if } \dfrac{d_i}{d_i + d_i^{'}} > H_S \text{ or } \dfrac{d_i^{'}}{d_i + d_i^{'}} > H_S \ , \\ \text{or } (d_i + d_i^{'}) < H_d \\ \\ false, \qquad\qquad\qquad \text{if otherwise} \end{cases} \tag{1}$$

where H_s and H_d are two thresholds, set 0.8 and 5, respectively, in our experiments.

5.2. Score Function

When deciding how many branches to split a node, we use the score function to calculate a score for each number of a range of branch number, and compare the scores to choose the most appropriate number to split the node. Denote the score of branch number z for splitting the node N_i^k as *score*(N_i^k, z). Here, we hope the child nodes either can contain much more positive images than negative images that means this node can present a cluster of images associated with this label, or can contain much more negative images than positive images means this node can present a cluster of images not associated with this label. We adopt entropy to measure the score of the branch number. For subnodes $Nc_{i,j}^k$ split from node N_i^k, we define:

$$\begin{aligned} &entropy(Nc_{i,j}^k) = (-1)\times(\tau_{ij}\log\tau_{ij} + (1-\tau_{ij})\log(1-\tau_{ij})), \\ &\text{where } \tau_{ij} = \frac{d_{ij}}{d_{ij} + d_{ij}^{'}} \text{ is the ratio of positive images} \\ &\text{with } L_k \text{ in } Nc_{i,j}^k, \end{aligned} \tag{2}$$

and

$$\begin{aligned} &score(N_i^k, \text{z}) = \min\{entropy(Nc_{i,j}^k) \ \ , 1 \le j \le \text{z}\}, \\ &\text{where } Nc_{ij}^k \text{ is the } j\text{-th child of } N_i^k. \end{aligned} \tag{3}$$

In Equation (3), we use the minimal function because we expect that there exists at least one node with the best criteria in the next level. Other nodes with worse scores can be divided again. Thus, the best branch number for splitting node N_i^k is

$$z_i^k = \arg\min_z score(N_i^k, z) \tag{4}$$

While dividing a node N_i^k into z_i^k nodes using *K*-means clustering, the semantic label L_k can be grouped z_i^k subclasses according to the positive and negative images in the node.

6. Confidence Value

Given an unlabeled image I_{new} and the classifiers C_k that is trained by the procedures in Section 4, we compute the confidence value of image I_{new} associated with label L_k, which confidence is denoted by $\gamma(L_k, N_{\text{root}}^k | I_{new})$ that is computed according to the hierarchical classifier C_k with root node N_{root}^k for label L_k. We therefore design a recursive computation for the confidence values and describe it as the follows.

Given a node N_i^k in the hierarchical classifier C_k for label L_k, the confidence value $\gamma(L_k, N_{\text{i}}^k | I_{new})$ can be regarded as the confidence of image I_{new} involving the sub-concepts in node N_i^k, and it can be recursively computed by

$$\gamma(L_k, N_i^k | I_{new}) = \begin{cases} \sum_{j=1}^{z_i^k} \gamma(L_k, Nc_{ij}^k | I_{new}) p(Nc_{ij}^k | I_{new}), \text{if } N_i^k \text{ is not a leaf} \\ \dfrac{d_i}{d_{\text{root}}} \qquad\qquad , \text{if } N_i^k \text{ is a leaf} \end{cases} \tag{5}$$

If N_i^k is a leaf node, we define d_i / d_{root}, where d_i and d_{root} are the number of positive images in node N_i^k and root N_{root}^k, respectively, to judge how confident node N_i^k involves sub-concepts associated with label L_k. Note that we adopt d_i / d_{root} instead of $d_i / (d_i + d_i^{'})$ for the judgement; the main reason is that overfitting will be obvious for the latter in most nodes which contain a small number of images. If N_i^k is not a leaf node, it can be propagated by its children, Nc_{ij}^k, as well as the weight $p(Nc_{ij}^k | I_{new})$ that means the possibility of image I_{new} belonging to node Nc_{ij}^k. The weight can be defined as the normalized inverse of distances from I_{new} to the mean of the cluster, denoted as N_i^k in general, by

$$p(N_i^k | I_{new}) = \frac{dist^{-1}(I_{new}, N_i^k)}{\sum_{j=1}^{J} dist^{-1}(I_{new}, Nc_{parent,j}^k)} \tag{6}$$

where J is the number of sibling nodes of N_i^k.

Figure 3 illustrates the computation, in equation (5), of the confidence values. Assume that the classifier is trained for label L_k, and an unlabeled image I_{new} is annotated now. In initial, $|D_k| = |D_k^{'}| = 100$, and the numbers of positive and negative images in nodes are shown. The red digits means $p(N_i^k | I_{new})$ of all nodes N_i^k. Then, the final confidence value of I_{new} associated with label L_k is the sum of all values computed in the leaves, and it is 0.15745.

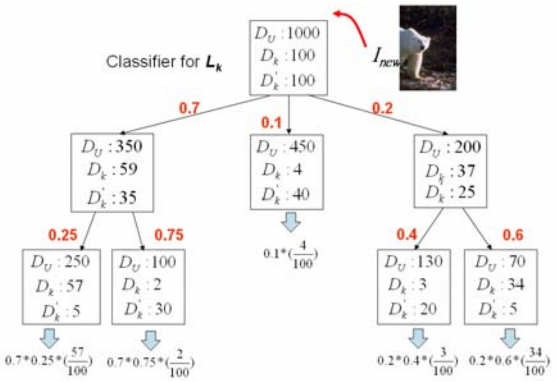
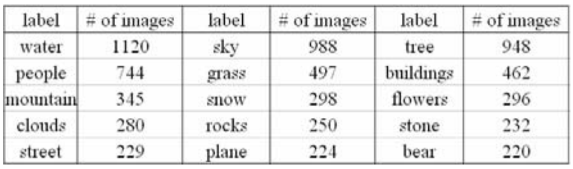


Figure 3. Illustration of computing the confidence of an image I_{new} associated with label L_k. The total confidence value is the sum of all values computed in the leaves.

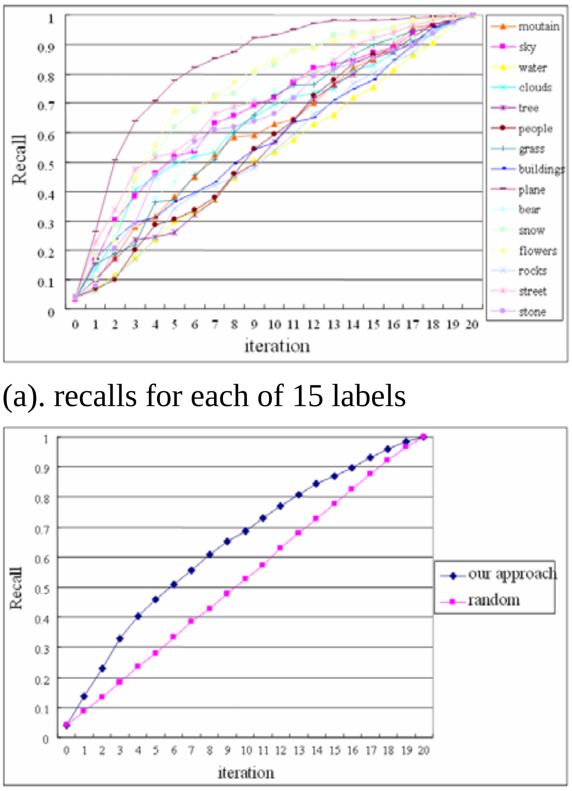
7. Experimental Results

In our experiments, we adopted the public dataset [7] that is widely used for the evaluation in image annotation. This dataset includes a total of 5,000 images of Corel Photo, the ground truth of labeling (1-5 labels for each image), a set of region features (36D), and visual words generated by *K*-means clustering (*K*=500). In the dataset, some labels are associated with a huge number of images, but some labels are not. For example, there are 1,120 images labeled by “water” but only one image labeled by “glacier”. Because our method is independent of the number of labels, we select 15 labels, shown in Figure 4, that are associated with images many enough.



label	# of images	label	# of images	label	# of images
water	1120	sky	988	tree	948
people	744	grass	497	buildings	462
mountain	345	snow	298	flowers	296
clouds	280	rocks	250	stone	232
street	229	plane	224	bear	220

Figure 4. The labels used in the experiments and their original numbers of associated images.



(a). recalls for each of 15 labels

(b). average recalls of our methods and random choice

Figure 5: The recalls of the proposed method with different iterations.

For the quantitative evaluation, we randomly and roughly selected 200 images for each of the 15 labels and computed the average recalls of image annotation for each label. Note that we adopted the region features and the visual words that are provide within the dataset. Figure 5(a) shows each of the recalls for 15 labels with different iterations, and Figure 5(b) draws the average recalls of all. For the comparison, we depict the average recalls using random choice.

Moreover, we perform another experiment, without relevance feedback, to show the effect of using unlabeled images in classifier training. We adopt F1 value, which considers both precision and recall, as the evaluation measure, where F1= (2×precisionrecall)/(precision+recall). In this experiment, we change the numbers of the labeled images with $|D_k|$=8, 16, 32, 64, and 128, and we also change the numbers of the unlabeled images with $|D_U|$=0, 800, and 1,600. The result, in Figure 6, shows that using unlabeled images can significantly improve the performance, especially the cases with few labeled images (e.g., $|D_k|$=8 or 16). That will be very helpful for relevance feedback because we cannot get many labeled images at the beginning of the iterations for image annotation. Using unlabeled images to help the clustering can reach to a better performance at first iterations for image annotation.

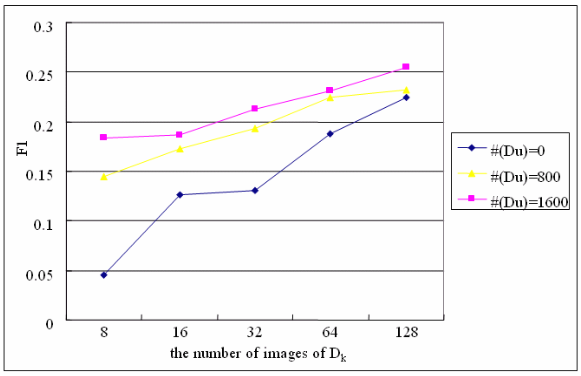


Figure 6: The performances of our method with different number of labeled images and with different number of unlabeled images.

8. Conclusion and Future Work

This paper presents an interactive method for image annotation using a semi-supervised and hierarchical approach. We apply unlabeled images to assist classifiers in training in order to reach a better performance even though with fewer training images. We construct hierarchical classifiers each corresponds to an individual label that can make the system more scalable when a new label is added and re-train the classifier easily in relevance feedback. These approaches proposed in this paper can make the annotation system more flexible.

For the future work, we are trying to embed ontological structure in the annotation process in order to reduce the complexity of the semantic contents in images. Ontology may link the relationship of labels, and then the task of image annotation could be more accurate. Besides, another unsupervised clustering method instead of *K*-means clustering can be used. Moreover, we plan to apply the annotation results to image retrieval.

References

[1] S. Basu, “Semi-supervised Clustering: Probabilistic Models, Algorithms and Experiments”, Ph.D. Thesis, Department of Computer Sciences, University of Texas at Austin, 2005.

[2] M. Bilenko, S. Basu, and R. J. Mooney, “Integrating Constraints and Metric Learning in Semi-Supervised Clustering”, in Proceedings of ICML, 2004.

[3] Fei-Fei, L. and P. Perona, “A Bayesian Hierarchical Model for Learning Natural Scene Categories”, in Proceedings of CVPR, pp. 524-531, 2005.

[4] G. Carneiro and N. Vasconcelos, “Formulating Semantic Image Annotation as a Supervised Learning Problem”, in Proceedings of CVPR, 2005.

[5] E. Y. Chang, K. Goh, G. Sychay, and G. Wu, “CBSA: Content-based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines”, IEEE Transaction on Circuits and Systems for Video Technology, 13(1):26–38, 2003.

[6] R. Datta, J. Li, and J. Z. Wang, “Content-Based Image Retrieval - Approaches and Trends of the New Age”, in Proceedings of the ACM SIGMM international workshop on MIR, 2005.

[7] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, “Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary”, in Proceedings of ECCV, pp. 97-112, 2002.

[8] S. L. Feng, R. Manmatha, and V. Lavrenko, “Multiple Bernoulli Relevance Models for Image and Video Annotation”, in Proceedings of CVPR, 2004.

[9] J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic Image Annotation and Retrieval using Cross-Media Relevance Models”, in Proceedings of ACM SIGIR, 2003.

[10] W. Jin, R. Shi, and T. S. Chua, “A Semi-Naïve Bayesian Method Incorporating Clustering with Pair-Wise Constraints for Auto Image Annotation”, in Proceedings of ACMMM, 2004.

[11] V. Lavrenko and W. Croft, “Relevance-Based Language Models”, in proceedings of ACM SIGIR, 2001.

[12] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, “Content-Based Multimedia Information Retrieval: State of the Art and Challenges”, ACM Transactions on Multimedia Computing, Communications and Applications, vol.

2(1): 1–19, 2006.

[13] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, “Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 8(5): 644-655, 1998.

[14] M. Srikanth, J. Varner, M. Bowden, and D. Moldovan, “Exploiting Ontologies for Automatic Image Annotation”, in Proceedings of ACM SIGIR, 2005.

[15] X. D. Zhou, L. Chen, J. Ye, Q. Zhang, and B. Shi, “Automatic Image Semantic Annotation Based on Image-Keyword Document Model”, in Proceedings of CIVR, 2005.

[16] X. Zhu, “Semi-Supervised Learning with Graphs” Ph.D. Thesis, CMU, 2005.