IMAGE ANNOTATION WITH RELEVANCE FEEDBACK USING A SEMI-SUPERVISED AND HIERARCHICAL APPROACH

Cheng-Chieh Chiang [1], Ming-Wei Hung [2], Yi-Ping Hung [2] and Wee Kheng Leow [3]

[1] *Department of Information Technology, Takming University of Science and Technology, Taipei, Taiwan*

[2] *Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan*

[3] *Department of Computer Science, School of Computing, National University of Singapore, Singapore*

Keywords: Image Annotation, Relevance Feedback, Semi-supervised Learning, Hierarchical Classifier.

Abstract: This paper presents an approach for image annotation with relevance feedback that interactively employs a semi-supervised learning to build 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. 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 design a semi-supervised approach for learning classifiers reduces the need of training images by use of both labeled and unlabeled images. This proposed semi-supervised and hierarchical approach is involved in an interactive scheme of relevance feedbacks to assist the user in annotating images. Finally, we describe some experiments to show the performance of the proposed approach.

1 INTRODUCTION

Image understanding and retrieval (Datta et al., 2005) have become a very active research area since the 1990's due to the rapid increase in the use of digital images. However, the semantic gap between the low-level features extracted from images and the high-level concepts involved in human perception is still a challenging problem. 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 several labels to an image to describe the semantic contents of the image. 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 – 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. Our opinion is to involve human feedbacks in image annotation because human should make the final decision for the semantic concept. Hence, we design a method with interactive human feedbacks to assist the user in annotating images in this paper.

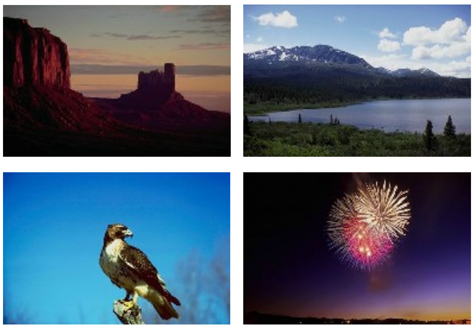


Figure 1: Different image contents with label "sky".

Image annotation is considered a supervised learning problem in many state-of-the-art methods (Carneiro and Vasconcelos, 2005). 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

Chiang C., Hung M., Hung Y. and Kheng Leow W. (2008).
IMAGE ANNOTATION WITH RELEVANCE FEEDBACK USING A SEMI-SUPERVISED AND HIERARCHICAL APPROACH.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 173-178
DOI: 10.5220/0001082001730178

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, and Section 3 formulates our problem and presents the overview of our approach. Then, the details of classifier training and confidence values are described in Section 4 and 5, respectively. Section 6 presents our experiments to show the effectiveness of our approach, and Section 7, in final, draws the conclusion and the future work.

2 RELATED WORK

Some of state-of-the-art works for image annotation and concept detection were provided (Datta et al., 2005). Many previous researches related to image annotation were based on the probabilistic model between features and labels, for example, a co-occurrence model (Mori et al., 1999), a translation model (Duygulu et al., 2002), a relevance model (Lavrenko and Croft, 2001), Cross-Media Relevance Model (CMRM) (Jeon et al., 2003), and Multiple Bernoulli Relevance Model (MBRM) (Feng et al., 2004), etc. Soft annotation is designed to give images a confidence level for each trained semantic label (Chang et al., 2003). Image annotation was formulated as a supervised learning problem (Carneiro and Vasconcelos, 2005). Jin et al. designed a K-means clustering with pair-wise constraints for image annotation (Jin et al., 2004). Srikanth et al. proposed methods for image annotation by use of a hierarchy defined on the annotation labels derived from a textual ontology (Srikanth et al., 2005).

In additional, 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 (Bilenko, 2004) (Zhu, 2005). 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 (Rui et al., 1998). Relevance feedback has been widely used for image retrieval recently, and we apply relevance feedback to assist the user in image annotation.

3 OVERVIEW AND FORMULATION

Let the entire dataset, denoted as *D*, contain *M* images. Suppose that *K* annotation labels $\{L_1, \ldots, L_K\}$ 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. This paper proposes a method with relevance feedback to assist the user in image annotation: the user can easily annotate images with labels as metadata, or summarize a set of images with semantic concepts.

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.

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 retrieve images in D_U with the most confidence values associated with label L_k. Figure 2 shows the flow that describes our interactive process for image annotation. Considering a 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.

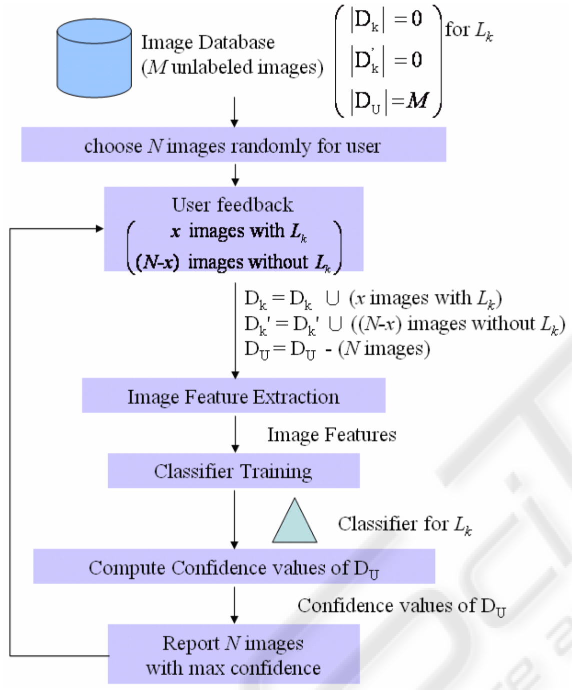


Figure 2: The flowchart of our approach.

4 CLASSIFIER TRAINING

Table 1 shows the algorithm that constructs the hierarchical classifier C_k for label L_k. In this algorithm, the root node N^k_{root} 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. 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.

Table 1: 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 $N^k_{root}$ contains $D_U \cup D_k \cup D^{'}_k$
// $N^k_i$: node i for the hierarchical classifier $C_k$.
// construct the tree by splitting each node $N^k_i$.
1. for each leaf node $N^k_i$ not fitting the **stopping
   condition**
{
2. for $z = 2$ to $b$
   { // $b$ is the max number of the trying range
3.   **node splitting method** to divide $N^k_i$ into $z$
     classes.
4.   compute **score**$(N^k_i, z)$
     // evaluate how many branches are appropriate for
        node $N^k_i$.
   }
5. $z^k_i = \arg\min_z score(N^k_i, z)$
6. $N^k_i$ is divided into $z^k_i$ classes
// $N^k_i$ is divided into, w.l.o.g., $Nc^k_{i,1}, \ldots, Nc^k_{i,z^k_i}$.
}
```

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^k_i: d_i is the number of positive images in the node and $d^{'}_i$ is the number of negative images in the node.

4.1 Node Splitting

An unsupervised clustering is used to divide node N^k_i 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 (Fei-Fei and Perona, 2005) 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.

4.2 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:

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

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

4.3 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

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

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.

5 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 $\gamma(L_k, N_{\mathrm{root}}^k \mid I_{new})$ 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 $\gamma(L_k, N_{\mathrm{i}}^k \mid I_{new})$ 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

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

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 $p(Nc_{ij}^k \mid I_{new})$

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

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

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 $p(N_i^k \mid I_{new})$ 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.

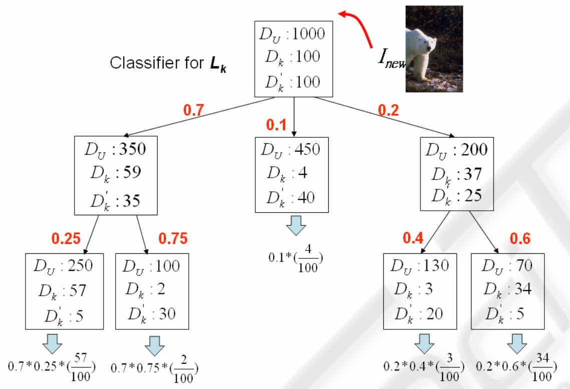


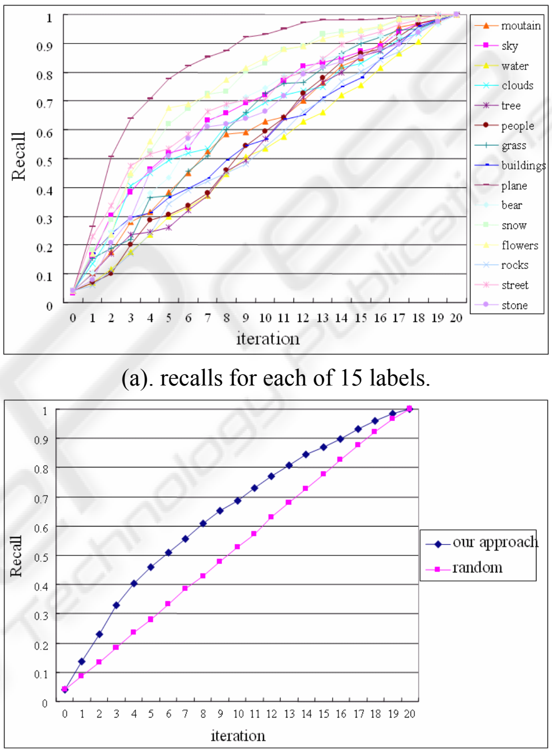
Figure 3: Illustration of computing the confidence of an image I_{new} associated with label L_k. The total confidence value is the sum of all values computed in the leaves.

6 EXPERIMENTAL RESULTS

In our experiments, we adopted the public dataset (Duygulu et al., 2002) 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 Table 2, that are associated with images many enough.

Table 2: The labels used in the experiments and their original numbers of associated images.

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



(a). recalls for each of 15 labels.

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

Figure 4: The recalls of our proposed method with different iterations.

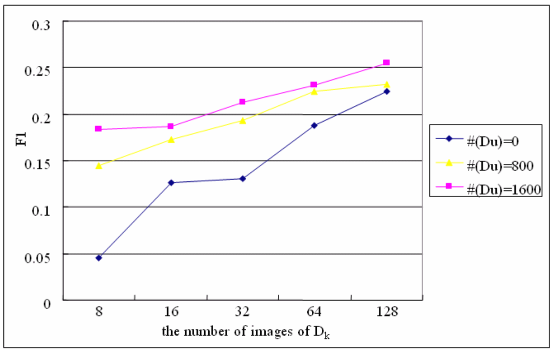


Figure 5: The performances of our method with different number of labeled images and with different number of unlabeled images.

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 4(a) shows each of the recalls for 15 labels with different iterations, and Figure 4(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×precision×recall)/(precision+recall). 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 5, 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.

7 CONCLUSIONS 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 to reach a better performance even though fewer training images are included. We construct hierarchical classifiers each corresponds to an individual label that can make the annotation system more flexible. In the future, we will use another unsupervised clustering instead of *K*-means clustering in our method. We also plan to embed prior knowledge, e.g., ontology, in the annotation task. Moreover, we plan to apply the annotation results to image retrieval.

ACKNOWLEDGEMENTS

This work was supported in part by the grants of 96-2752-E-002-007-PAE and 96R0062-03.

REFERENCES

Bilenko, M., Basu , S., and Mooney, R. J. (2004). Integrating Constraints and Metric Learning in Semi-Supervised Clustering. *Proceedings of ICM.*

Fei-Fei, L. and Perona, P. (2005). A Bayesian Hierarchical Model for Learning Natural Scene Categories. *Proceedings of CVPR*, pp. 524-531.

Carneiro, G. and Vasconcelos, N. (2005). Formulating Semantic Image Annotation as a Supervised Learning Problem. *Proceedings of CVPR*.

Chang, E. Y., Goh, K., Sychay, G., and Wu, G. (2003). 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.

Datta, R., Li, J., and Wang, J. Z. (2005). Content-Based Image Retrieval - Approaches and Trends of the New Age. *Proceedings of the ACM SIGMM international workshop on MIR*.

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

Feng, S. L., Manmatha, R., and Lavrenko, V. (2004). Multiple Bernoulli Relevance Models for Image and Video Annotation. *Proceedings of CVPR*.

Jeon, J., Lavrenko, V., and Manmatha, R. (2003). Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. *Proceedings of ACM SIGIR*.

Jin, W., Shi, R., and Chua, T. –S. (2004). A Semi-Naïve Bayesian Method Incorporating Clustering with Pair-Wise Constraints for Auto Image Annotation. *Proceedings of ACMMM*.

Lavrenko, V. and Croft, W. (2001). Relevance-Based Language Models. *Proceedings of ACM SIGIR*.

Mori, Y., Takahashi, H., and Oka R., (1999). Image-to-word transformation based on dividing and vector quantizing images with words. *Proceedings of First International Workshop on Multimedia Intelligent Storage and Retrieval Management.*

Rui, Y., Huang, T. S., Ortega, M., and Mehrotra, S. (1998). 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.

Srikanth, M., Varner, J., Bowden, M., and Moldovan, D. (2005). Exploiting Ontologies for Automatic Image Annotation. *Proceedings of ACM SIGIR*.

Zhu, X. (2005). Semi-Supervised Learning with Graphs. *Ph.D. Thesis, CMU.*