A Semi-Naïve Bayesian Method Incorporating Clustering with Pair-wise Constraints for Auto Image Annotation

ABSTRACT

We propose a novel approach for auto image annotation. In our approach, we first perform the segmentation of images into regions, followed by clustering of regions, before learning the relationship between concepts and region clusters using the set of training images with pre-assigned concepts. Our main contribution is two-fold. First, in the learning stage, we perform clustering of regions into region clusters by incorporating pair-wise constraints which are derived by considering the language model underlying the annotations assigned to training images. Second, in the annotation stage, we employ a semi-naïve Bayes model to compute the posterior probability of concepts given the region clusters. Experiment results show that these two strategies result in considerable improvements in the performance of auto image annotation, and our proposed approach outperforms the state-of-the-art techniques in large image collection.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval Models; **I.4.8 [Image Processing and Computer Vision]**: Scene Analysis – Object Recognition

General Terms

Algorithms, Experimentation

Keywords

Image annotation, pair-wise constraint, semi-supervised clustering, semi-naïve Bayes

1. INTRODUCTION

With the proliferation of multimedia contents, especially images and videos, there is a strong need to automatically annotate the contents of images using a pre-defined set of concepts frequently used by the users. The annotated concepts can then be used as the basis to retrieval images based on concepts, perform question answering, and various semantic image processing tasks. Starting from a training set of images with pre-assigned concept annotations, current approaches employ statistical learning models to associate visual features of sub-units within images with concepts. The learning techniques employed includes: co-occurrence model [10], 2D HMM [8], translation model [1], LDA model [3], cross-media relevance model [5] and continuous-space relevance model [7]. The sub-unit employed could be a fixed size block [10, 8] or a region [1, 3, 5, 7]. Almost all approaches use traditional color, texture, position and statistical shape as visual features.

The key idea behind auto image annotation is in deriving the probability of associating concepts with image regions. Most existing approaches share the same two-step pipeline in tackling this problem: (a) clustering image regions to region clusters; and (b) finding joint probability of region clusters and concepts, or the posterior probability of concepts given region clusters. This paper aims to tackle two of the key limitations of the current approaches. First, since most approaches rely on clustering as the basis for auto image annotation, the performance of annotation is strongly influenced by the quality of clustering. Currently, most approaches perform regions clustering merely based on visual features. Thus regions with different semantic concepts but share similar appearance may be easily grouped, leading to a poor clustering performance. Second, most current techniques do not take co-occurrence of region clusters (and/or co-occurrence of concepts) into consideration. This co-occurrence information is useful in providing the context that identifies impossible configuration of region clusters (and/or concepts).

To address the above problems, we first consider the use of a language model underlying the annotations assigned to training images to impose additional semantic pair-wise constraints when clustering the regions. Recently research on clustering [12, 2] shows that clustering with pair-wise constraints, a kind of realistic semi-supervised clustering method, performs considerably better than the unconstrained methods. Next, we formulate a semi-naïve Bayesian model to perform auto annotation. It aims to strike a good balance between the simplicity of naïve Bayesian model and the need to incorporate co-occurrence information of region clusters. Experimental results demonstrate that both clustering with pair-wise constraints and semi-naïve Bayesian model are effective, and the combined approach outperforms the state-of-the-art systems.

Our main contribution is two-fold. First, we develop a semi-supervised region clustering method incorporating pair-wise constraints which are derived from language model. Second, we formulate a semi-naïve Bayesian model for concept prediction and inference. This paper discusses the design and implementation of our system.

2. REGION CLUSTERING WITH PAIRWISE CONSTRAINTS

In most cases, it is computationally intractable to construct a continuous model for auto annotation, so most approaches employ discrete models that rely on clustering as the basis for auto-annotation. Currently, most approaches perform image regions clustering merely based on visual features. Thus regions with different semantic concepts but similar appearance may be easily grouped, which will lead to a poor clustering performance. A natural solution to overcome this problem is to impose constraints to the process of clustering, which have been shown to perform considerably better than the unconstrained ones. We consider the

ACM Multimedia '04, New York, USA.

framework that uses the cannot-link constraints between pairs of regions [12], with an associated cost **P** when violating each constraint. Here we use $\mathbf{R_i}$ to denote region cluster i, and $\mathbf{r_j}$ to denote region **j**.

2.1 Formulation of pair-wise constraints

Annotation of an image reflects the semantics of the image as well as its regions, and we would like to induce from the annotations the cannot-link and must-link relations between different regions. In general, the cannot-link relationship can be easily deduced from shared concepts but the must-link relationship is harder to deduce. For example, it is obvious that regions in image "sky water grass" cannot link to regions in image "furniture indoor"; but it is harder to say that certain regions in image "sky water grass" must exactly correspond to certain regions in another image "sky water grass". Thus, we deduce only the cannot-link relations from semantic concepts, leaving others as "possible-link" to be further evaluated by visual features.

We assume that the semantic irrelevance of two regions can be deduced by the irrelevance of all concepts (or annotations) between two images. If two images show little correlation in their annotations, we can say with high confidence that regions in these images are semantically irrelevant to each other. This assumption is reasonable as although annotation of an image is likely to be incomplete, it is always complete for those concepts that we care most. Under this assumption, we assert that for every image pair I_p and I_q, if their annotations C_p and C_q are irrelevant, then all relationships across their regions are marked as cannot-link. We denote that $\forall\ r_i \in I_p$, $\forall\ r_j \in I_q$; $s(r_i, r_j) = 1$, where **s** is a relationship function between regions r_i and r_j, and it is set to 1 for cannot-link and 0 for no restriction (possible-link).

"C_p and C_q are irrelevant" can be simply interpreted as C_p and C_q do not have any terms in common, or $C_p \cap C_q = \Phi$. However, terms might be correlated statistically or lexically through a language model. Here, we adopt two relatively simple and yet effective approaches to derive semantic correlations [9]:

a) Co-occurrence based correlation

In general, high co-occurrence concepts are likely to be used together to describe (or annotation) the same image. In other words, two concepts are likely to belong to the same conceptual group if they have high co-occurrence and vice visa. The co-occurrence-based correlation of two concepts c_1 and c_2 is computed as:

$$R_{co}(c_1, c_2) = df(c_1 \wedge c_2) / df(c_1 V c_2) \tag{1}$$

where $df(c_1 \wedge c_2)$ $(df(c_1 V c_2))$ is the fraction of images with annotations containing c_1 and (or) c_2.

b) Thesaurus based correlation

WordNet is an electronic thesaurus popularly used in research on lexical semantic acquisition. In WordNet, the meaning of a word is represented by a network of synonym (synset) and hypernym etc between words. The thesaurus-based correlation between the two concepts c_1 and c_2 is computed as:[1]

$$R_L(c_1, c_2) = \begin{cases} 1 & (c_1 \text{ and } c_2 \text{ in the same synset, or } c_1 = c_2) \\ 0.8 & (c_1 \text{ and } c_2 \text{ have "antonym" relation}) \\ 0.5 & (c_1 \text{ and } c_2 \text{ have relations of "is\_a",} \\ & \quad \text{"part\_of", or "member\_of")} \\ 0 & (\text{others}) \end{cases} \tag{2}$$

The relevance of two annotations C_p and C_q is defined as

$$\mathrm{Rel}(C_p, C_q) = \underset{c_i \in C_p, c_j \in C_q}{\arg\max} (R(c_i, c_j)) \tag{3}$$

where the correlation definition R could be either R_{co} or R_L. If the relevance of two annotations $\mathrm{Rel}(C_p, C_q)$ is smaller than a predefined threshold, then C_p and C_q and their corresponding image regions are regarded as "irrelevant" to each other.

2.2 Clustering with pair-wise constraints

After the construction of pair-wise constraints between regions, we perform clustering to generate region clusters. K-Means is a popular clustering method. Since K-Means cannot directly handle pair-wise constraints, we adapt a variant of K-Means called Pair-wise Constrains K-Means (PCK-Means) [2] to perform the clustering. We formulate the goal of pair-wise constraint clustering as the minimization of a combined objective function, defined as the sum of the total squared distances between the regions and their region cluster centroids, and the cost incurred by violating any of the pair-wise constraints. Let $\{r_i\}_{i=1}^{N}$ be the whole set of regions, $\{\mu_h\}_{h=1}^{K}$ represent the centroids of K region clusters $\{R_h\}_{h=1}^{K}$, l(i) be the cluster assignment of a region r_i, where $l(i) \in \{1,2,\ldots,K\}$, and P be the cost incurred when the "cannot-link" pair-wise constraints are violated. Our aim is to minimize the target function:

$$J_{pckmeans} = \sum_{i=1}^{N} \| r_i - \mu_{l_i} \|^2 + \sum_{\{(r_i, r_j) | l(i) = l(j)\}} s(r_i, r_j) * P \tag{4}$$

Traditional K-Means method is sorts of EM-like algorithm. Compared with K-Means, PCK-Means alternates between cluster assignment in the E-step, and centroids estimation in the M-step. We repeat the E-step and M-step until the clustering result converges (see [2]). In addition, the selection of appropriate K and good initial centroids are critical to the success of greedy clustering algorithms such as K-means. In our experiments, the number of clusters K is set to be 300 empirically and initial centroids are selected using some heuristics like farthest-first traversal.

After clustering, we obtain a set of m region clusters: $R_1, R_2, \ldots, R_m$. Each region is assigned to one region cluster. For each region cluster, we keep an inversion list in order to facilitate subsequent processing by semi-naïve Bayes model. The inversion list of region cluster records the list of all images containing at least one region which has been assigned to this region cluster as:

[1] $R_{co}(c_i, c_j) > \sigma$ is also required to ensure that c_i and c_j co-occur sufficiently to ensure they are in the same context or have same linguistic sense. Here we set $\sigma = 0$.

$$II(R_i) = \{I_j \mid \exists r \in I_j, l(r) = i\} \tag{5}$$

3. A SEMI-NAÏVE BAYESIAN APPROACH TO ANNOTATION

Instead of directly building intractable relationship between regions and concepts, we try to build the relationship between region clusters and concepts. A semi-naive Bayes classifier decomposes the input variables into subsets and represents the statistical dependency within each subset, while treating the subsets as statistically independent. During the annotation stage, we use the probabilities to model the relationships between annotations and region cluster, and solve it using semi-naïve Bayesian model in 4 steps as follows.

Step 1: Formulate it as a probability problem

Given a new un-annotated image, we first segment it into regions r_1, r_2, …, r_m using a region segmentation method. Note that the number of regions m is not necessarily fixed. We derive the posterior probability of concept c_i given the regions r_1, r_2, …, r_m as follows.

$$\begin{aligned}
&\Pr(c_i \mid r_1, r_2, \ldots, r_m) \\
&= \sum_{(k_1, k_2, \ldots, k_m)} \{\Pr(c_i \mid R_{k_1}, R_{k_2}, \ldots, R_{k_m}) \\
&\quad * \Pr(R_{k_1}, R_{k_2}, \ldots, R_{k_m} \mid r_1, r_2, \ldots, r_m)\}
\end{aligned} \tag{6}$$

$$\begin{aligned}
&\approx \arg\max_{(k_1, k_2, \ldots, k_m)} \{\Pr(c_i \mid R_{k_1}, R_{k_2}, \ldots, R_{k_m}) \\
&\quad * \Pr(R_{k_1}, R_{k_2}, \ldots, R_{k_m} \mid r_1, r_2, \ldots, r_m)\}
\end{aligned} \tag{7}$$

Thus, one needs to sum up the probabilities over all possible correspondence between regions and region clusters, as shown in Eq. (6). Since this is in general intractable, it is common to perform a (saddle-point) approximation of the sum around the optimal point estimate which is the maximum a posterior (MAP) estimate. We approximate it using the one with the largest probability (see Eq. (7)). Through experimentation, we found that this approximation is reasonable and can greatly improve computation efficiency.

Step 2: Calculate the posterior probability of concepts given region clusters.

Semi-naïve Bayes method is proposed by [6] and has been proven to be efficient. We apply semi-Bayes principle to dissemble joint possibility of all elements into multiplication of joint possibility of each element and its nearest neighbors. From Eq. (7), we have:

$$\Pr(c_i \mid R_{k_1}, R_{k_2}, \ldots, R_{k_m})$$

$$= \frac{\Pr(c_i, R_{k_1}, R_{k_2}, \ldots, R_{k_m})}{\Pr(R_{k_1}, R_{k_2}, \ldots, R_{k_m})} \tag{8}$$

$$= \frac{\prod_{j=1}^{m} \Pr(c_i, R_{k_j}, near^t_{\{R_{k_1}, R_{k_2}, \ldots, R_{k_m}\}}(R_{k_j}))}{\prod_{j=1}^{m} \Pr(R_{k_j}, near^t_{\{R_{k_1}, R_{k_2}, \ldots, R_{k_m}\}}(R_{k_j}))} \tag{9}$$

$$= \prod_{j=1}^{m} \frac{\Pr(c_i, R_{k_j}, near^t_{\{R_{k_1}, R_{k_2}, \ldots, R_{k_m}\}}(R_{k_j}))}{\Pr(R_{k_j}, near^t_{\{R_{k_1}, R_{k_2}, \ldots, R_{k_m}\}}(R_{k_j}))} \tag{10}$$

The function $near^t_A(R)$ returns the **t** nearest neighbors (region clusters) to region cluster R among all region clusters in set A. Two region clusters are said to be "near to each other" if they have high correlation, i.e. they are likely to show up in the same image. Correlation of two region clusters can be easily calculated using the inversion list of region cluster.

$$R_{co}(R_i, R_j) = \frac{|II(R_i) \cap II(R_j)|}{|II(R_i) \cup II(R_j)|} \tag{11}$$

To calculate $\Pr(R_{k_j}, near^t_{\{R_{k_1}, R_{k_2}, \ldots, R_{k_m}\}}(R_{k_j}))$, we examine all images and count the images containing the region clusters $\{R_{k_j}, near^t_{\{R_{k_1}, R_{k_2}, \ldots, R_{k_m}\}}(R_{k_j})\}$. This can be easily accomplished by using the intersection of inversion list of region clusters. The probability is then the ratio between the total count and the total region cluster pair throughout all images. To calculate $\Pr(c_i, R_{k_j}, near^t_{\{R_{k_1}, R_{k_2}, \ldots, R_{k_m}\}}(R_{k_j}))$, we examine all images that have concept c_i in their annotations, and perform the same computation as above. The ratio of the first count and second count is equal to the inner part of Eq. (10).

Step 3: Calculate the posterior probability of region clusters given regions

We use the independent assumption to estimate the second part of Eq. (7) as:

$$P(R_{k_1}, R_{k_2}, \ldots, R_{k_m} \mid r_1, r_2, \ldots, r_m) = \prod_{i=1}^{m} P(R_{k_i} \mid r_i) \tag{12}$$

Under the assumption that each region corresponds to one and only one region cluster, the calculation of the posterior $P(R_{k_1}, R_{k_2}, \ldots, R_{k_m} \mid r_1, r_2, \ldots, r_m)$ can be regarded as the problem of finding an appropriate region cluster for each region. Such correspondence can be achieved by calculating the cosine similarity between R_i and r_j.

Step 4: Annotate new images

After we have derived the posterior probability of every concept c_i, we annotate the new image by choosing concepts with high probabilities. This can be further refined by using the co-occurrence statistics of concepts. But how many concepts are enough? We employ the following heuristics to determine the list of concepts.

a) We choose a fixed number of concepts for annotation.

b) We adaptively select the number of concepts for the annotation based on a threshold to be determined empirically.

4. EXPERIMENTS & DISCUSSION

4.1 Database

We collect 4,850 images from Corel image CD, and select 59 concepts to be used for the annotation experiments. The concepts include scene-level concepts like "Beaches" "Sky" "Sunset"; object-level concepts like "Fruits" "Bears"; together with a concept for "none". The concepts are chosen based on the hierarchical concepts described in TGM I (Thesaurus for Graphics Materials) [14]. We have labeled all images at the image level, with 1~5 concepts for each image. We perform segmentation using Blobworld [4] and ensure that there are 1-12 regions for each image. Each region is represented by a 69-D feature vector using our Matching-Pursuit feature [11], which has been found to be effective in our earlier experiments. For each test, we randomly select 10% of images for testing and the rest for training, and measure the performance in terms of F_1 measure.

4.2 Selection of model parameters

The core idea behind semi-naïve Bayesian model is reflected by the t value used in $near_A^t(R)$ function (see Eqn. (9-10)). t=1 gives the naïve Bayesian model, while higher t values result in more complex models that uses more detailed co-occurrence information from the similar region clusters. In order to evaluate the effect of changing t values, we carry out experiments by using t=1, 2 and 3. The corresponding F_1-values of auto annotation are 0.241, 0.297 and 0.282 respectively. From the results, we can see that semi-naïve Bayes is more effective than naïve Bayesian since it captures the co-occurrence information of region clusters well. Also, t=2 (i.e. semi-naïve Bayes with degree=2) seems to give the best performance.

In addition to **t**, the number of concepts assigned to each image influences the final performance. Annotation with too many concepts may boost recall but sacrifice precision, and vice visa. We experiment with different number of assigned concepts, and found that annotating each image with three concepts gives the best performance (see Figure 1).

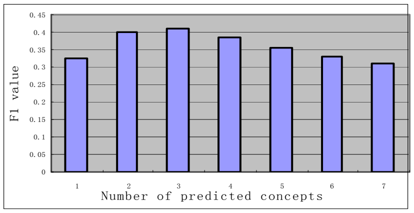

Figure 1. The influence of number of predicted concepts

4.3 Performance

Table 1 shows the performance of different approaches. SNB_{NC} represents the semi-naïve Bayesian approach with no constraints and is also used as our Baseline run. SNB_{SC} represents semi-naïve Bayesian approach with simple constraints, and SNB_{LC} represents the semi-naïve Bayesian approach with language model constraints. For comparison, we also include the results of the state-of-the-art cross-media relevance model (CMRM) [5]. The results in Table 1 show that SNB_{LC} performs significantly better than CMRM and SNB_{NC} with an F_1 measure of about 0.41. Table 2 shows the per-concept recall and precision averaged over all 59 concepts. It again shows that our model with constrains based on language model (SNB_{LC}) performs the best.

Table 1. The performance of different test configurations

Different Models	**F_1 value**	**Comparison With SNB_{NC}**	**Comparison with CMRM**
CMRM	0.326	-	0
SNB_{NC}	0.297	0	-9.0%
SNB_{SC}	0.386	+29.8%	+18.3%
SNB_{LC}	0.410	+37.9%	+25.7%

Table 2. Comparison per-concept performance of the models

Models	**CMRM**	**SNB_{SC}**	**SNB_{LC}**
# Concepts with recall>0	22	26	27
Average per-concept Recall	0.107	0.141	0.157
Average per-concept Precision	0.277	0.308	0.342

5. CONCLUSIONS & FURTHER WORK

We have presented a novel semi-naïve Bayesian approach incorporating clustering with pair-wise constraints for auto image annotation, which has been shown by experiments to give considerable improvements to the annotation performance. Future work includes incorporating different types of pair-wise constraints and auto annotation of other media types.

6. REFERENCES

[1] Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D. & Jordan, M. I.: Matching words and pictures. *Journal of Machine Learning Research*, 3,:1107-1135, 2003.

[2] Bilenko, M., Basu, S. & Mooney, R.J.: Integrating constraints and metric learning in semi-supervised clustering, *to appear in the Proc. of the 21st Int. Conf. on Machine Learning (ICML-2004), Banff, Canada, July 2004.*

[3] Blei, D. & Jordan, M.I.: Modeling annotated data. *Proc. of ACM SIGIR,* 127-134. ACM Press, 2003.

[4] Carson, C., Belongie, S., Greenspan, H., & Malik, J.: Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying, *IEEE Trans. on Patter Analysis and Machine Intelligence, vol. 24, no. 8*, Aug, 2002

[5] Jeon, J., Lavrenko, V. & Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. *Proc. of ACM SIGIR Conf.* 119-126, 2003.

[6] Kononenko, I.: Semi-naïve Bayesian classifier. *Sixth European Working Session on Learning.* 206-219. 1991.

[7] Lavrenko, V., Manmatha, R. & Jeon, J.: A model for learning the semantics of pictures. *Neural Information Processing System (NIPS),* 2003.

[8] Li, J. & Wang, J. Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(10):14, 2003.

[9] Liu, J.M. & Chua, T.-S., Building semantic perceptron net for topic spotting. *39th Annual Meeting of Association for Computational Linguistic (ACL 2001)*, 370-377, 2001.

[10] Mori, Y., Takahashi, H. & Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. *First Int'l Workshop on multimedia Intelligent Storage & Retrieval Management*, 1999.

[11] Shi, R., Feng, H.M., Chua, T.-S. & Lee, C.-H., An adaptive image content representation and segmentation approach to automatic image annotation. *Int'l Conf. on Image and Video Retrieval*, July 21-23, 2004

[12] Wagstaff, K., Cardie, C., Rogers, S. & Schroedl, S.: Constrained K-means clustering with background knowledge. *Proc. of Int'l Conference on Machine Learning (ICML-2001)*.

[13] Yan, R. & Hauptman, A.: A discriminative learning framework with pair-wise constraints for video object classification. *CVPR 2004.*

[14] http://www.loc.gov/rr/print/tgm1/