A Novel Approach to Auto Image Annotation Based on Pair-wise Constrained Clustering and Semi-naïve Bayesian Model

Shi Rui[1], Wanjun Jin[1,2] and Tat-Seng Chua[1]

[1] *School of Computing National University of Singapore, Singapore*

[2] *Dept of Computer Science and Engineering, Fudan University, China*

ABSTRACT

Automatic image annotation has been intensively studied for content-based image retrieval recently. In this paper, we propose a novel approach for this task. Our approach first performs the segmentation of images into regions, followed by the clustering of regions, before learning the associations between concepts and region clusters using the set of training images with pre-assigned concepts. The main focus of this paper and our main contributions are as follows. First, in the learning stage, we perform clustering of regions into region clusters by incorporating pair-wise constraints derived by considering the language model underlying the annotations assigned to training images. Second, in the annotation stage, to alleviate the restriction of the independence assumption between region clusters, we develop a greedy selection and joining algorithm to find the independent sub-sets of region clusters and employ a semi-naïve Bayesian (SNB) model to compute the posterior probability of concepts given those independent sub-sets. Experimental results show that our proposed system utilizing these two strategies outperforms the state-of-the-art techniques in large image collection.

Keywords

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

1. INTRODUCTION

Image annotation refers to the process of automatically labeling the image contents with a predefined set of concepts representing image semantics, which can be used primarily for image database management. Recent studies [15] suggest that users are likely to find it more useful and convenient to search for images based on text annotations rather than using visual-based features. Thus automatic image annotation (AIA) aims to invest a large amount of preprocessing efforts to annotate the images as accurately as possible to support keyword-based image search.

Most current systems share a general three-step pipeline to tackle AIA problem: (a) image component decomposition: by decomposing an image into a collection of sub-units, which could be segmented regions, equal-size blocks or entire image; (b) image content representation: by modeling each content unit based on a feature representation scheme; and (c) content classification: by computing the associations between unit representations and textual concepts. The general AIA framework is shown in Figure 1.

For step a, three kinds of image components are often used as image analysis units in most CBIR (content-based image retrieval) and AIA systems. In [10, 18], the entire image was used as a unit. Some recent systems use segmented regions as sub-units in images [2, 4, 5, 6, 8, 20]. However, the accuracy of segmentation is still an open problem. As a compromise, several systems adopt fixed-size sub-image blocks as sub-units for an image [12, 19] since block-based methods can be implemented easily.

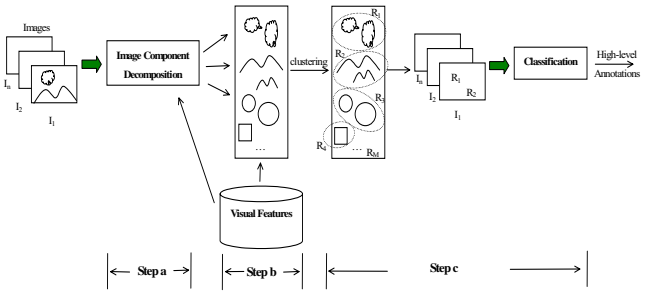

Figure 1: Framework for auto-image annotation system

For step b, conventional CBIR approaches employ color, texture, and statistical shape features to model image contents [10, 16, 18]. In addition, some special purpose systems tend to rely on domain-specific features, such as the use of face detectors and face recognizers to look for images of people [1]. Recent work examined the use of adaptive visual features to model a wide range of classification tasks [20].

For step c, current approaches focus on employing statistical learning models to associate visual representations of image analysis units with concepts, by using a training set of images with pre-assigned concept annotations. The learning techniques employed include: co-occurrence model [12], 2D HMM [23], translation model [2], LDA model [4], cross-media relevance model [6] and continuous-space relevance model [8]. The key idea behind these models is in deriving the probability of associating concepts with image regions. Thus as Figure 1 shows, given segmented regions, these approaches first perform clustering of segmented regions to region clusters, then finding joint probability of region clusters and concepts, or the posterior probability of concepts given region clusters. The use of region clusters, rather than regions themselves, aims to convert the problem of region-concept associations into a discrete model so as to alleviate the data sparseness problem of needing to deal with many diverse regions. 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 automatic image annotation, the performance of annotation is strongly influenced by the quality of clustering. Currently, most approaches perform region clustering merely based on visual features. Thus regions with different semantic concepts but share similar appearance may be grouped, leading to a poor clustering performance. Second, to estimate the joint probability of region clusters and concepts, most current techniques assume that the events of observing region clusters

Proceedings of the 11th International Multimedia Modelling Conference (MMM'05)
1550-5502/05 $20.00 © 2005 **IEEE**

within an image are mutually independent once an image is picked. But this assumption does not often hold. For example, some images with image-level concepts like "Cityscapes" or "lndoor" can only be represented by a group of co-occurring region clusters, as a single region cluster is hard to induce these high level concepts. In another example, tigers are more likely to co-occur with grass, water and trees, but less often with objects like "Furniture, "Tables" or "Chairs". These facts indicate that some region clusters often co-occur in an image. That is, we could not simply assume that the events of observing region clusters are mutually independent.

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 [22, 24] 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 (SNB) model [7, 13] to perform the AIA task. The basic idea behind our SNB is to strike a good balance between the simplicity of Naïve Bayes (NB) model and the need to incorporate co-occurrence information of region clusters. Experimental results demonstrate that the combined approach, including pair-wise constrained clustering and SNB model, 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 derived from language model. Second, we formulate a SNB model for concept prediction and inference and develop a greedy selection and joining algorithm (GSJ) to find independent region cluster subsets for SNB model. This paper discusses the design and implementation of our system.

The rest of this paper is organized as follows. In Section 2 we introduce our proposed approach of region clustering with pair-wise constraints. In Section 3 we present how to predict concepts for an image based on SNB model. Section 4 shows experimental results and compares our performance with other AIA systems. Finally, the last section concludes with a discussion of future work in this area.

2. REGION CLUSTERING WITH PAIR-WISE CONSTRAINTS

Most AIA systems first segment the images into regions, and represent each region by a set of low-level visual features. However, the features, naturally associated with image regions, are often very sparse in the whole feature space. So most approaches employ clustering techniques to vector quantize the image region representation. As far as we know, 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 poor clustering performance. A natural solution to overcome this problem is to impose some constraints to the process of clustering, which have been shown to perform considerably better than the unconstrained ones [22]. We consider a similar framework as in [22] that uses the cannot-link constraints between pairs of regions, with an associated cost P when violating each constraint. Here we use R_i to denote region cluster i, and r_j to denote individual region j.

2.1 Formulation of pair-wise constraints

In our approach, each image is composed of some segmented regions, and every segmented region inherits all the concepts from its image. Thus, image concepts (or annotations) reflect the semantics of this image as well as its regions, and we would like to induce from the image concepts the cannot-link and must-link relationships between different regions. In general, the cannot-link relationship can be easily deduced from shared image concepts but the must-link relationship is harder to deduce. For example, it is obvious that regions in image of "sky water grass" cannot link to regions in image of "furniture indoor"; however, it is harder to say that certain regions in image "sky water trees" must exactly correspond to certain regions in another image "sky water grass". Thus, we deduce only the cannot-link relations from semantic annotations, leaving others as "possible-links" to be further evaluated by visual features.

We further 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 about. Under this assumption, we assert that for every image pair x_p and x_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 $\forall\ r_i \in x_p$, $\forall\ r_j \in x_q$; $s(r_i, r_j) = 1$, 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 $C_p \cap C_q = \Phi$. This strategy, referred to as "simple constraint" in our experiment, is quite straightforward and easy to realize. However, terms might be correlated statistically or lexically through a language model. Thus, we need to further consider statistical and language model in deducing whether "C_p and C_q are irrelevant" to each other. The resulting model is called "languge model constraint" in our experiment. The model considers two kinds of semantic correlations [9]:

a) Co-occurrence based correlation

In general, high co-occurrence concepts are likely to be used together to describe (or annotate) 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:

$$R_C(c_1, c_2) = \begin{cases} df(c_1 \cup c_2) / df(c_1 \cap c_2), & if\ df(c_1) > s \\ 0, & otherwise \end{cases} \quad (1)$$

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

In practice, the co-occurrence correlation is meaningful only if there is sufficient statistical support which can be approximated by requiring: $df(c_1) > \sigma$, where σ is a pre-defined threshold.

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:

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

The relevance of two sets of annotations C_p and C_q is defined as

$$Rel(C_p, C_q) = \underset{c_i \in C_p, c_j \in C_q}{argmax} (R_C(c_i, c_j)) \mid R_L(C_p, C_q)) \quad (3)$$

In our experiments, we just use the co-occurrence based correlation as the language model constraints. If the relevance of two annotations *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 from different images, 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 Constrained K-Means (PCK-Means) [24] to perform the clustering. We formulate the goal of pair-wise constrained 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 N regions, and $\{R_h\}_{h=1}^{M}$ be the set of M region clusters (with $M << N$) with corresponding cluster centroids of $\{\mu_h\}_{h=1}^{M}$. Let $l(i)$ be the cluster assignment of a region r_i, where $l(i) \in \{1,2,...,M\}$, and P be the cost incurred when the "cannot-link" pair-wise constraints are violated. The value of P is chosen empirically and it is dependent to selected region features. Our aim is to minimize the target function:

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

Traditional K-Means method is an EM-like algorithm. Compared with K-Means, PCK-Means alternates between cluster assignment in the E-step, and centroids estimation in the M-step. In the E-step, every point is assigned to the cluster that minimizes its distance to the cluster centroids according to the local metric and the cost of any constraint violations incurred by this cluster assignment. Since earlier studies have shown that the order of assignment does not result in statistically significant differences in clustering quality; we therefore employ random ordering. Here, each point moves to a new cluster only if the component of target function J contributed by this point decreases.

In the M-Step, every cluster centroids is first re-estimated using the points in corresponding R_h. As a result, the contribution of each cluster to J is minimized. The pair-wise constraints do not take part in this centroids re-estimation step because the constraint violations only depend on cluster assignments, which do not change in this step.

We repeat the E-step and M-step until the clustering result gets converges. Further details of PCK-means clustering steps can be found in [24]. Here the selection of appropriate K and good initial centroids is critical to the success of greedy clustering algorithms such as K-means. In our experiments, the number of cluster K is set to 300 empirically.

After clustering, we obtain a set of M region clusters: $\Gamma = \{R_1, R_2 \ldots, R_M\}$. 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:

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

where $R_i \in \Gamma$, x_j is an observed image and r is one of segmented regions of x_j.

3. A SEMI-NAÏVE BAYESIAN APPROACH TO ANNOTATION

Given the set of region clusters, the next problem is to find the associations or probabilities between region clusters and concepts. To estimate these probabilities, most existing approaches assume that the events of observing region clusters within an image are mutually independent once an image is picked. But this assumption does not often hold. To alleviate the restriction of this "naive" assumption, we explore the use of semi-naïve Bayesian principle to incorporate some forms of dependencies among region clusters and/or concepts in an efficient and effective manner. A Semi-Naive Bayes (SNB) classifier decomposes the input variables into subsets and represents the statistical dependence within each subset, while treating the subsets as statistically independent [7, 13]. Thus, during the annotation stage, given an image and its representation as a set of region clusters, we first need to find the independent region cluster subsets by a greedy selection and joining (GSJ) algorithm. Then we estimate the joint probabilities or posterior probabilities between concepts and these independent region cluster subsets instead of region clusters.

3.1 Formulate AIA task as a probability problem

The Bayesian framework requires that all the entities involved in decision making be adequately formalized:

- Each observed image x belongs to a collection T of un-annotated image.
- The set Ω of concepts used for image annotation, $\Omega = \{c_1, c_2, \ldots, c_N\}$, and if we regard each concept $c \in \Omega$ as a class, any image x from T belongs to one or more classes.

- Each observed image x is modeled as a sample of a random variable X, whose class-conditional probability density function for a concept $c \in \Omega$ is written as $f_X(x \mid c)$.
- A priori knowledge concerning the concepts is expressed via a probability function defined on the set of concepts, $\{p(c_1), p(c_2), \ldots, p(c_N)\}$.
- After pair-wise constraint clustering, we can obtain a set Γ of region clusters to represent the whole low-level feature space, $\Gamma = \{R_1, R_2 \ldots, R_M\}$.

Firstly, each image $x \in \mathcal{T}$ is segmented into regions $\{r_1, r_2, \ldots, r_m\}$. Note that the number of regions m is not necessarily fixed. Next, we need to find an appropriate region cluster $R \in \Gamma$ for each region $r \in \{r_1, r_2, \ldots, r_m\}$. Here, we assume that each region corresponds to only one region cluster. Such correspondence can be achieved by calculating the cosine similarity between R ($R \in \Gamma$) and r. Thus the un-annotated image x is represented by a set of region clusters $\{R_{k_1}, R_{k_2}, ..., R_{k_m}\} \subseteq \Gamma$, which can be described as $x = \{R_{k_1}, R_{k_2}, ..., R_{k_m}\}$. In terms of Bayesian framework, we can derive the posterior probability of any concept $c \in \Omega$ for image x given its region clusters $\{R_{k_1}, R_{k_2}, ..., R_{k_m}\}$ as follows.

$$p(c \mid x) \approx p(c \mid R_{k_1}, R_{k_2}, \ldots, R_{k_m}) = \frac{f_\Gamma(R_{k_1}, R_{k_2}, ..., R_{k_m} \mid c)\, p(c)}{f_\Gamma(R_{k_1}, R_{k_2}, ..., R_{k_m})} \quad (6)$$

where the denominator, $f_\Gamma(R_{k_1}, R_{k_2}, ..., R_{k_m})$, in Eq. (6) gives the unconditional (or marginal) probability density function of the observed region clusters, which serves simply as a normalizing constant. Thus, the image annotation problem can be stated as: given a set of observed region clusters, $\{R_{k_1}, R_{k_2}, ..., R_{k_m}\}$, from an image x, classify x into one or some of the concept classes in Ω. So according to the criterion of the *maximum a posteriori* (MAP), the final decision is given by

$$\hat{c} = \arg\max_{c \in \Omega} \{p(c \mid (R_{k_1}, R_{k_2}, ..., R_{k_m}))\} = \arg\max_{c \in \Omega} (f_\Gamma((R_{k_1}, R_{k_2}, ..., R_{k_m}) \mid c)\, p(c)) \quad (7)$$

Note that the reason that MAP can be used is because regions and region clusters within an image inherit all the concepts from their corresponding image.

3.2 Find the independent region cluster subsets

For estimating the class-conditional density function, $f_\Gamma(R_{k_1}, R_{k_2}, ..., R_{k_m} \mid c)$, most existing approaches simply assumes that the region clusters are conditionally independent, which is the basic assumption of Naïve Bayes (NB) Classifier [11] . As a result, the class-conditional density function can be written as:

$$f_\Gamma(R_{k_1}, R_{k_2}, ..., R_{k_m} \mid c) = \prod_{j=k_1}^{k_m} f_{\Gamma^{(j)}}(R_j \mid c) \quad (8)$$

However, from our earlier discussions, we know that this assumption does not often hold. Thus, a straightforward solution to this problem is to find the dependency between region clusters, $\{R_{k_1}, R_{k_2}, ..., R_{k_m}\}$. In our system, we apply semi-naïve Bayesian principle to achieve this goal. Semi-naïve Bayes method is proposed by [7, 13] and has been proven to be efficient. Inspired by the method in [13], we propose a greedy selection and joining (GSJ) algorithm to find the independent region cluster sub-sets for $\{R_{k_1}, R_{k_2}, ..., R_{k_m}\}$ and in each subset the region clusters are dependent.

$$R_{co}(R_i, R_j) = \frac{|\Pi(R_i) \cap \Pi(R_j)|}{|\Pi(R_i) \cup \Pi(R_j)|} \quad (9)$$

The dependency between any two of region clusters can be measured by Eq. (9), which shows evaluate how much these two region clusters are likely to show up in the same image. In our approach, the dependency can be easily calculated using the inversion list of region cluster, as shown in Eq. (5). In order to limit the level of dependency between region cluters so as to save computational costs, we use the parameter t to control the number of region clusters to be included in each independent region cluster subset. t=1 gives the naïve Bayesian model, while higher t values result in more complex models that use more detailed co-occurrence information from similar region clusters. The greedy selection and joining (GSJ) algorithm for finding the independent subsets of region clusters for the SNB model is given in Fig. 2.

1) Initialization:
$B = \varnothing$; $S = 1$; Choose $R_i \in H$ randomly,
$k_1 \le i \le k_m$, $B_S = \{R_i\}$; $H = H / \{R_i\}$;

2) Selection step:
Select $R_j = \arg\max_{R_h \in H} \sum_{R_g \in B_S} |R_{co}(R_g, R_h)|$,
and for any $R_g \in B_S$, $|R_{co}(R_g, R_j)| > \varepsilon$,
ε is a pre-defined threshold;

3) Joining step:
If R_j exists and $|B_S| < t$
$B_S = B_S \cup \{R_j\}$; $H = H / \{R_j\}$;
Go to 2);
Else If $H \neq \varnothing$
$S = S + 1$; $B = B \cup \{B_S\}$;
Go to 1);
Else
Exit
End
End

Figure 2: The GSJ (greedy selection and joining) algorithm

Based on our proposed GSJ algorithm, the set of region clusters H can be decomposed into the independent subsets $B = \{B_1, B_2, \ldots, B_l\}$, where $\bigcup_{i=1}^{l} B_i = H$, and for any $B_i, B_j \in B$, $B_i \cap B_j = \varnothing$.

Thus, Eq. (8) can be rewritten as

$$f_\Gamma(R_{k_1}, R_{k_2}, ..., R_{k_m} \mid c) = f_B(B_1, B_2, ..., B_l \mid c) = \prod_{i=1}^{l} f_{B^{(i)}}(B_i \mid c) \quad (10)$$

Compared with the methods in [7, 13], the advantage of GSJ algorithm lies in its simplicity and efficiency in computation,

since more parameters are needed to be estimated in [7] and more computational costs in [13].

3.3 Estimate the class-conditional density functions

Consider that there is a total of n training images from a concept c in the database. So we can simply approximate the class-conditional density, i.e. $f_{B^{(i)}}(B_i \mid c)$, by Eq. (11),

$$f_{B^{(i)}}(B_i \mid c) \approx \frac{vol(B_i, c)}{vol(c)} \tag{11}$$

where $|B_i| \leq t$, $B_i \subseteq H$, $vol(c) = n$, and $vol(B_i, c)$ is the volume of images containing region clusters B_i with the concept c.

3.4 Annotate new image

After we have derived the posterior probability of every concept c, we annotate the new image by choosing some concepts with top posterior probabilities. In our experiments, we use a fixed number of concepts for annotation. We also compare the performance of different number concepts for annotation in section 4.2.

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 list of concepts used is given in Figure 3. The concepts include image-level concepts like "Cityscapes", "Indoor", "Sunrises&sunsets"; region-level concepts like "Fruits", "Bears"; together with a concept for "none". These concepts are chosen based on the hierarchical concepts described in TGM I (Thesaurus for Graphics Materials) [25].

Animals, Apes, Bears, Birds, Dogs, Elephants, Lions, Penguins, Tigers, Wolves, Zebras, Painting, Sculpture, Clothing, Cityscapes, Beaches, Deserts, Forests, Glaciers, Mountains, Meadows, Valleys, Buildings, Historic buildings, Skyscrapers, Towers, Bridges, Roads, Railroads, Food, Fruits, Dessert, Furniture, Tables, Chairs, Indoor, Clouds, Frost, Night, Reflections, Sky, Snow, Sunrises&sunsets, Plants, Grasses, Trees, Flowers, People, Athletes, Rocks, Churches, Temples, Vehicles, Trains, Water, Waterfall, Waves, Unknown, None

Figure 3: The list of concepts used in annotation

We have labeled all concepts at the image level, with 1~5 concepts for each image. We perform segmentation using Blobworld [5] and ensure that there are 1-12 regions for each image. Each region is represented by a 69-D feature vector including Luv color histogram and our Matching-Pursuit texture feature [20], 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

As discussed in Section 3.2, the parameter t is very important for the SNB classifier. 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-annotating the test images are 0.241, 0.297 and 0.302 respectively. From the results, we can see that semi-naïve Bayes is more effective than naïve Bayesian especially when t is large. This is because a larger t value captures the co-occurrence information of region clusters well. However, a larger t value requires considerably more computational costs and larger training set in evaluating the dependencies between larger set of region clusters. As the increase in F_1 measure from t=2 to t=3 is quite small, we therefore consider t=2 for the rest of experiments.

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 4).

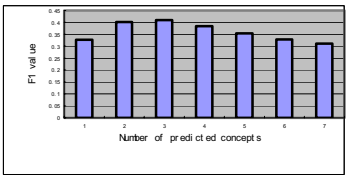

Figure 4. The influence of number of predicted concepts

4.3 Two models for comparison

As baseline for comparison with our methods, we choose two popular classifiers to perform the association between region clusters and concepts. They are: the cross-media relevance model (CMRM) and the probabilistic SVM (PSVM).

CMRM is the stat-of-the art AIA system. Different from our method, CMRM attempts to discover the statistical links between region clusters and concepts by estimating the joint distribution of concepts and region clusters [6]. However, CMRM assumes that the events of observing the region clusters and concepts are mutually independent, which is similar to the principle of NB.

Support vector machines (SVMs) are an effective classification technique with strong background in statistical learning theory [21]. But SVMs only output a positive or negative prediction without any associated measure of confidence. So in our experiment, we consider an extended version of SVM, called PSVM [14], which can output a *poseriori* class probability. Thus, an associated probability of class membership can be obtained. The training examples for each class are the set of the training images, each represented by a set of region clusters. Thus, the inputs to PSVM are the binary vectors.

Both CMRM and PSVM are based on the results of K-means clustering without constraints. In our experiment, we use the giniSVM toolkit as our probabilistic SVM [26].

4.4 Performance

Table 1 shows the performance of different approaches. SNB_{NC} represents the semi-naïve Bayesian approach with no constraints. SNB_{SC} represents semi-naïve Bayesian approach with simple

constraints, and SNB_{LC} represents the semi-naïve Bayesian approach with language model constraints (see Section 2 for details of these models). For comparison, we also include the results of the state-of-the-art cross-media relevance model (CMRM) [6] and PSVM.

In this experiment, we evaluate the F_1 value of each model in annotating all the testing images. The results are presented in Table 1. From the Table 1, we can see that the performance of SNB_{NC} is comparable or slightly better than the baseline models, namely, the CMRM and PSVM models. Moreover, the performance of SNB_{SC} and SNB_{LC} are significantly better than that of SNB_{NC}. In particular, SNB_{LC} performs the best with an F_1 measure of about 0.41. It again shows that our model with constraints 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	0.263	-11.45%	0
PSVM	0.284	-4.38%	+7.98%
SNB_{NC}	0.297	0	+12.93%
SNB_{SC}	0.386	+29.97%	+46.77%
SNB_{LC}	0.410	+38.05%	+55.89%

5. CONCLUSIONS & FURTHER WORK

We have presented a novel semi-naïve Bayesian approach incorporating clustering with pair-wise constraints for automatic image annotation. The model has been found experimentally to be considerably better than the naïve Bayesian model, and CMRM and PSVM based models in auto-image annotation task.

Our current research is focused in three areas. First, we will experiment with the use of different types of pair-wise constraints and clustering models. Second, we will extend our model to automatically annotate data of other media types. Third, we plan to apply our model to large-scale TRECVID dataset.

6. REFERENCES

[1] Y. A. Aslandogan and C. T. Yu, "Multiple evidence combination in image retrieval: Diohenese searches for people on the web," *ACM SIGIR'2000*, Athens, Greece, 2000.

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

[3] M. Bilenko, S. Basu & R.J. Mooney, "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.*

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

[5] C. Carson, M. Thomas, S. Belongie, J.M. Hellerstein and J. Malik, "Blobworld: A system for region-based image indexing and retrieval," *Proc. Int'l Conf. Visual Information System*, 1999.

[6] J. Jeon, V. Lavrenko and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," *ACM SIGIR'03*, July 28-Aug 1, 2003.

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

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

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

[10] B. Manjunath and W. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837-842, Aug 1996.

[11] T.M. Mitchell, "Machine Learning," McGraw Hill, 1997.

[12] Y. Mori, H. Takahashi and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," In *Proc. of First International Journal of Computer Vision*, 40(2): 99-121, 2000.

[13] M.J. Pazzani, "Searching dependency in Bayesian classifiers," *Learning from data: Artificial intelligence and statistics V*, New York, NY: Springer-Verlag, editor D. Fisher and H.-J. Lenz, pp. 239-248, 1996.

[14] J.C. Platt, "Probabilities for SV machines," In A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61-74, Cambridge, MA, 1999, MIT Press.

[15] K. Rodden, "How do people organize their photographs?" In *BCS IRSG 21st Ann. Colloq. on Info. Retrieval Research*, 1999.

[16] J.R. Smith and S.F. Chang, "VisualSeek: A fully automated content-based query system," *ACM Multimedia*, 1996.

[17] J.R. Smith and C.S. Li, "Image classification and querying using composite region templates," *Journal of Computer Vision and Image Understanding*, 2000.

[18] M. Swain and D. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11-32, 1991.

[19] M. Szummer and R.W. Picard, "Indoor-outdoor image classification," *IEEE Intl Workshop on Content-based Access of Image and Video Databases*, Jan 1998.

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

[21] V. Vapnik, "The Nature of Statistical Learning Theory," Springer, New York, 1995.

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

[23] J.Z. Wang and J. Li, "Learning-based linguistic indexing of pictures with 2-D MHMMs," *Proc. ACM Multimedia*, pp. 436-445, Juan Les Pins, France, Dec 2002.

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

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

[26] http://bach.ece.jhu.edu/svm/ginisvm/