1-2007

Mining multiple visual appearances of semantics for image annotation

Hung-Khoon TAN
City University of Hong Kong

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Mining Multiple Visual Appearances of Semantics for Image Annotation

Hung-Khoon Tan and Chong-Wah Ngo

Department of Computer Science,
City University of Hong Kong,
Kowloon, Hong Kong
{hktan,cwngo}@cityu.edu.hk

Abstract. This paper investigates the problem of learning the visual semantics of keyword categories for automatic image annotation. Supervised learning algorithms which learn only a single concept point of a category are limited in their effectiveness for image annotation. We propose to use data mining techniques to mine multiple concepts, where each concept may consist of one or more visual parts, to capture the diverse visual appearances of a single keyword category. For training, we use the *Apriori* principle to efficiently mine a set of frequent *blob-sets* to capture the semantics of a rich and diverse visual category. Each concept is ranked based on a discriminative or diverse density measure. For testing, we propose a level-sensitive matching to rank words given an unannotated image. Our approach is effective, scales better during training and testing, and is efficient in terms of learning and annotation.

Keywords: Image Annotation, Multiple-Instance Learning, Apriori.

1 Introduction

Content-based image indexing and retrieval is becoming a subject of significant importance. The earlier retrieval systems use only low-level features and the results are unsatisfactory because the semantic image contents are not well captured. An intuitive way is to manually annotate images with captions, and the users retrieve the relevant multimedia documents by typing in keywords at the system. Although low-tech, this approach is effective. However, as the size of the multimedia database explodes over years, such technique is no longer deemed feasible. Automatic annotation of images is becoming increasingly important and has since become an active area of research.

Image annotation systems could be broadly classified into unsupervised learning [1,2,3,4,5,6] and supervised learning [7,8,9,10] problems. The unsupervised learning approaches strive to learn the hidden states of concept, particularly the joint distribution between keywords and multiple visual features. Mori *et al.* [1] proposed the co-occurrence model to collect the co-occurrences between words and image features and used them to predict annotated words for images.

T.-J. Cham et al. (Eds.): MMM 2007, LNCS 4351, Part I, pp. 269–278, 2007.

Duygulu *et al.* [2] proposed a machine translation approach to learn a lexicon which maps a set of keywords to the set of regions of an image. In [3,4,5], relevance models were proposed to find the joint probability of observing a set of image regions together with another set of annotation words. As opposed to [2], the relevance models does not assume an underlying one-to-one alignment between the regions and words in an image and only assume that a set of keywords is related to a set of objects represented by regions. In [6], the Correlation LDA model is proposed to relate the keyword and the image.

The supervised learning approaches use generative or discriminative classifiers from a binary set of visual features with (positive) and without (negative) the semantic of interest. The classifier treats each annotated word as an independent class and a different image classification model is learnt for every semantic category. Recently, weakly supervised method, particularly multiple-instance learning (MIL) [12,13] is becoming a more attractive alternative for learning the semantics of images because of its less stringent requirement on manual labelling. In a MIL setting, we are aware of the presence of the object of interest in the image but which regions correspond to the object of interest is unknown. There are several drawbacks of supervised algorithms that have yet to be addressed before it could be effectively used for image annotation. Some algorithms, particularly MIL, learn a single concept (a point or region in a feature space). Learning multiple concepts for a single keyword category is crucial to the success of the adaptation of supervised approaches for large scale image annotation because (a) there are viewpoint, scale and lighting variations, and more seriously (b) some keyword categories are normally holistic or functional in nature, resulting in rich varieties of visual appearances. Second, the feature vectors generated by the segmentation algorithms are still far from desirable. Often, the object of interest is segmented into different parts. Therefore, it is interesting to investigate how useful modelling a concept with multiple visual parts is for image annotation.

In this paper, we address the fundamental issues of utilizing multi-facet visual concept points to characterize keyword categories. For clarity, we term each keyword as a category and each category is basically formed by multiple concepts. Every concept point can further contain one or several visual parts. The highlights of this work are as follows. First, data mining technique is proposed to learn multiple concepts to effectively handle multi-facet keyword categories. Each concept is composed of several visual parts to characterize its appearance. The learnt concepts are further ranked either by a discriminative measure or a diverse density measure. Second, region independence is not assumed in our approach. Most approaches [3,4,5] assume the process that generates the regions b_i are independent where $P(b_1...b_n) = \prod_{i=1}^{n} P(b_i)$ and neglect the correlation among visual parts. Our approach avoids this drawback by processing groups of visual parts. Third, the proposed technique is computationally efficient and scales well with data size compared to methods such cross-media relevance model (CMRM) [4] that does not scale well with the training set size.

2 Multi-facet Visual Appearance Model

We model the appearance model of a keyword category as a lattice structure shown in Figure 1. The lattice captures the multi-facet concepts while presenting them at different levels of visual granularities. In this structure, each node represents a concept which captures one or several visual parts. Basically the nodes at a higher level carry more specific, and thus more discriminant, categorical information for image annotation. In this section, we first propose techniques to mine, while simplifying, the lattice structure. Two novel measures, from the perspectives of discriminativeness and diverse density, are presented to encode the usefulness of each concept in a probabilistic manner. Image annotation is then performed by capitalizing on the level-sensitive information provided by the hierarchical structure of lattice representation.

2.1 Apriori Based Concept Mining

The different visual appearances of a keyword category can be effectively modelled by a lattice of visual part groups. To generate the structure, all the visual parts in the images of the same category are extracted and then used to create all permutations of visual part groups hierarchically. Modelling each category with a full lattice structure is inefficient because clearly a portion of the nodes in the lattice is uninteresting and does not correspond to the semantics of the keyword category. The extraction of the interesting subset of the lattice structure is posed as a data mining problem [11] where the *Apriori* algorithm can be used to mine for the significant sub-structure of lattice which contains frequent, and thus likely to be more interesting, concepts.

We use a discrete image representation as in [1,2,3]. Regions are extracted from the images using a general purpose segmentation algorithm. Features such as color, texture, position and shape information are computed for these regions and K-means is applied on the collection of all features to form clusters of features known as *blobs*. A training image is represented by a set of blobs $B_I = \{b_1...b_m\}$ and a word list $W_I = \{w_1...w_n\}$. To be consistent with data mining terminologies, we term a candidate concept (a collection of one or more blobs) as a *blobset* and a positive training image B_I as a *transaction* (of blobs). A n-blobset is a set of n blobs. The *level* of a blobset is the number of items in the blobset, which is n. One property of a blobset is its *support count*, which refers to the cardinality of a blobset in the set of all transactions, $T = \{B_1...B_M\}$. A *frequent* blobset is a blobset which satisfies a minimum support count *min_sup_count*.

We are interested to mine all frequent blobsets from the transactions of positive training set. Initially, all frequent 1-blobsets are extracted from T. Then, all the subsequent n-blobsets are recursively generated from the initial list. A data set that contains k 1-blobsets can potentially generate up to $2^k - 1$ frequent blobsets, resulting in a lattice model as shown in Figure 1. We use the well-known Apriori principle to generate only frequent blobsets.

Theorem 1 (*Apriori* Principle). *If an itemset is frequent, then all of its subsets must also be frequent.*

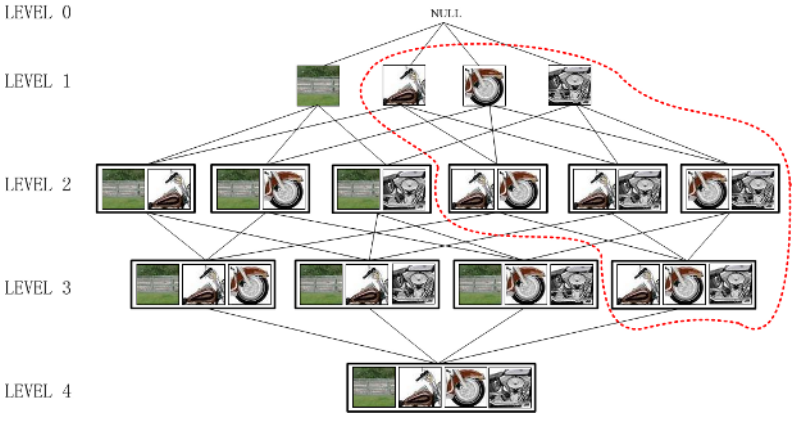

Fig. 1. A simplified lattice structure for representing the multiple visual appearance of a keyword category 'motorcycle' with a four 1-blobset b_i. Apriori algorithm extracts the visually significant blobsets (dotted red line) from the full lattice-set.

As illustrated in Figure 1, if the blobset b_1b_2 is infrequent, then all its supersets such as $b_1b_2b_3$, $b_1b_2b_4$ and $b_1b_2b_3b_4$ are also infrequent. Thus we can utilize the anti-monotonic property of the support count of blobs to prune the generation of uninteresting blobset, resulting in a set of frequent concepts as bounded by the dotted lines in Figure 1. Although *min_sup_count* is determined empirically, it is not critical and is more for the purpose of speed optimization.

The frequent blobset generation has its significance in several aspects. First, the Apriori algorithm mines for a compact and succinct set of concepts to fully model the different visual appearances of a keyword category. Similar to [1], we are using the co-occurrence between words and blobs. However, the co-occurrence model does not explicitly learn a class model for each word and only the co-occurrence between a blob and a word is considered. The model is incomplete in the sense that it does not model all the variations of visual appearances of a keyword category. Second, blobsets of different levels describe a concept with multiple visual parts. In this aspect, we also avoid the assumption that blobs are mutually independent of each other. The underlying assumptions by the relevance models [3,4,5] that the process that generates each blob is independent become invalid when some blobs which are correlated with some other blobs. Third, we can interpret the subset of a blobset as the incomplete or occluded version of the blobset. For example, the blobset $b1b2$ could be interpreted as the occluded version of its superset blobset $b1b2b3$ where the blob $b3$ is occluded. Thus, the Apriori algorithm creates a succinct model to represent a complete description of a keyword category.

2.2 Characterizing Concept Uniqueness

The lattice structure provides a platform to effectively model the multi-facet appearances of a keyword category. However, how do we determine the ownership

of a blobset? Apparently, the frequency of a blobset in the positive examples of a keyword is not a reliable cue if it is common across the whole training set. This scenario is analogous to the use of *idf* in text document retrieval. A reliable measure for characterizing the uniqueness of concept is "discriminativeness" where a blobset is unique if it is only frequent in positive images but rare, as a whole, in the training samples. However, a low discriminative measure does not necessarily rule out the usefulness of blobsets in describing a keyword. This is true for keyword categories that always co-exist together such as the keyword category "plane" is semantically related to "sky" and they share some similar visual content which could still be rare in other keyword categories. For such cases, additional information such as the ratio of the blobset in the positive examples, negative examples and total examples is useful. Therefore, we propose another measure "diverse density", inspired by the fundamental of multiple instance learning (MIL) [12], to handle such cases.

Discriminative Measure. The first measure $conf$ is a measure of how discriminative the concept is in describing a keyword category w_i. It assigns higher confidence values to concepts which appear only frequently in the positive images with respect to the whole training set. It is an asymmetric measure where only the presence of a blob is regarded as important. It ignores the blob's relative size with respect to the positive, negative and the whole training set. The $conf$ of c, a candidate concept for the keyword w_i, could be formulated as

$$conf(c, w_i) = \frac{|c|_+}{|c|_J} \tag{1}$$

where $|\bullet|_S$ denotes the cardinality of $\bullet$ in a set S. $+$ and $-$ denote the positive and negative training set for the keyword category w_i and J denotes the whole training set.

Diverse Density Measure. Motivated by the classical MIL diverse density algorithm [12], the second measure dd assigns the similarity based on how frequent a concept is in positive images and how infrequent it is in negative images. It takes into consideration how discriminative the concept is ($conf$ measure) with respect to the sizes of the positive, negative and the whole training sets. For instance, a keyword category with a higher number of training examples is assigned a higher dd value because there are more examples to support the presence of the category. The diverse density dd of a concept c of the keyword w_i is defined as

$$dd(c, w_i) = P(c|w_i, +)P(c|w_i, -)P(w_i)conf(c, w_i) \tag{2}$$

where $P(c|w_i, +)$ is the ratio of the number of concept c in all positive images, $P(c|w_i, -)$ is the ratio of images not containing the concept c in all negative images and $P(w_i)$ is the the ratio of the training images for keyword w_i over all images.

Both the $conf$ and dd measures have their own strength. The $conf$ measure looks into the exclusiveness of a blobset where higher values are assigned to blobsets which are mainly found in only one particular keyword category regardless

of the cardinality of training set. In other words, it ignores the global statistics of the blobset in the training set. In this aspect, it is similar to the CMRM platform [3] and is well-suited for keyword categories with diverse visual features, such as "boat" and "house". Since visual correlation is expected to be weak for such keyword categories, $conf$ is a better measure by highlighting the visual parts which are unique to the keyword only. The dd measure emphasizes on the global statistics of the candidate concept. In this aspect, it is similar to the MIL platform [12]. It is more useful for keyword categories with prominent visual features, such as "tiger" and "forest", where strong visual correlation exists among the positive examples. Besides, the measure is better positioned to handle semantically related keywords with overlapping visual parts such as "plane" and "sky". The localized overlapping of visual parts have a negative effect on $conf$ but less profound impact on the dd since the visual parts would still be rare statistically in the set of all negative examples as defined in $P(c|w_i, -)$.

2.3 Level-Sensitive Annotation

As we move down in the lattice level as shown in Figure 1, concepts become rarer, more discriminative and have less chance of happening by chance. Basically concepts at higher-level are capable of eliciting more evidence in terms of the number of visual parts to support their keyword category. We thus tap into this implicit feature of the lattice and propose a novel *level-sensitive* annotation. The approach strives to prioritize the scores of a concept according to its level, or more specifically the number of visual parts residing in a concept.

To determine the conditional probability of a keyword category w_i given an unannotated image I, $P(w_i|I)$, we select the best concept from the pool of candidate concepts of the category that matches the unannotated image. A concept matches the unannotated image when all the blobs in a concept are present in the unannotated image. Depending on which measure being used, we embed the notion of level-sensitivity into $P(w_i|I)$ using the following formulations

$$P(w_i|I) = \max_{c \in \mathbf{C}_{w_i}} \{dd(c, w_i) + L(c)\} \tag{3}$$

or

$$P(w_i|I) = \max_{c \in \mathbf{C}_{w_i}} \{conf(c, w_i) + L(c)\} \tag{4}$$

where $\mathbf{C}_{w_i}$ is the set of all frequent visual concepts of the word w_i learnt during the Apriori step. $L(c)$, representing the level of concept c, is aimed for assigning higher score to c at higher level. Then, annotation is performed based on a maximum a posteriori (MAP) criterion as follows

$$\hat{w} = \arg \max_{w_i \in V} P(w_i|I) \tag{5}$$

where the keyword category with the highest conditional probability is assigned to the unannotated image. For multiple annotations, the top-N keywords are selected.

3 Experiment and Results

3.1 Data Set and Evaluation

To evaluate the effectiveness of our approach, we use the data set provided by Duygulu *et al.* [2]. A total of 4,500 images is used as training set and the remaining 500 images as testing set. Each image is annotated with 1-5 keywords with a total vocabulary of 371 keywords. Images are segmented into 5-10 regions using normalized cut [14]. A 36-dimensional feature vector, which is composed of color, texture, mean oriented energy and other features, is extracted for each region. The set of all feature vectors is then quantized by K-means into 500 blobs. Details of the feature extraction process can be found in [2]. We follow the experimental methodology used by [2,3]. Given an unannotated image I from the test set, we use Equation 3 or 4 to arrive at the conditional probability $P(w_i|I)$. We perform a ranking and select the top 5 words as an annotation of image I using the recall and precision measure. Recall is the number correctly annotated images divided by the number of relevant images in the ground truth. Precision is the number of correctly annotated images divided by the total number of images annotated with that particular word. Recall and precision are then averaged over the word set. As in [3], we report the results on two sets of words, the subset of 49 best words and the complete set of all 260 words in the testing set.

3.2 Performance Comparison

We compare our proposed approach with the Co-occurrence (CO) [1], Machine Translation (MT) [2] and Cross-Media Relevance Model (CMRM) [3]. Our approach is named separately as APR_CF and APR_DD which uses the $conf$ and dd measure, respectively. During learning, we learn an average of 100 concepts up to a maximum level of 5 for each keyword category. The performance of the five tested approaches are summarized in Table 1 and illustrated in Figure 2. Our approach, although using co-occurrence between blobs and words, has significantly better performance compared to the CO and MT models. Both APR_CF and APR_DD are comparable to CMRM. They perform better in terms of average recall, and with a higher number of images with at least one correct annotation (i.e., recall > 0). Our approach, however, has a lower precision, partly because the Corel training set of different keyword categories is not well-balanced in terms of number of training examples. We notice that the keyword categories with too few training examples (some as few as 1) end up with a trivial lattice structure, impacting the precision performance. It is also observed that there are no notable differences in performance between the $conf$ and dd measures. We investigate the results and find that this is attributed to the level-sensitive matching scheme which filters out the noisier lower-level matchings.

Sensitivity of lattice height. The height (number of levels) of a lattice formed by the frequent concepts indeed impacts the performance of annotation. Here, we define "height" as the maximum level in a lattice that the frequent concepts of its category reach. The higher the lattice of a category, the more discriminant

Table 1. Performance of our approach (APR_DD and APR_CF) with Co-occurrence (CO), Machine Translation(MT) and Cross-Media Relevance Model(CMRM)

	All 260 Words		Best 49 Words		Recall>0
	Avg. Re.	Avg. Pr.	Avg. Re.	Avg.Pr.	
CO	0.02	0.03	-	-	19
MT	0.04	0.06	0.34	0.20	49
CMRM	0.09	0.10	0.48	0.40	66
APR_DD	0.11	0.07	0.50	0.26	75
APR_CF	0.11	0.07	0.50	0.27	76

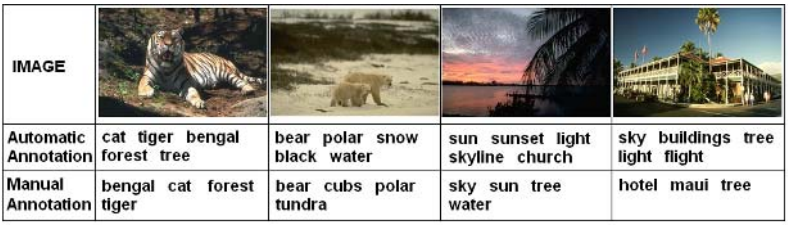

Fig. 2. Some annotation results of our approach compared to manual annotations

the concepts being learnt, and thus leads to a more reliable annotation. The lattice height of different keywords varies depending on the visual appearances of the keyword images, and also partly the number of training examples. In this experiment, we group the keyword categories according to the height of their lattice models. For each group, we compute their average recall, precision and percentage of words > 0. The results are shown in Table 2. Apparently, all the performance measures improve with the increase of the height of the lattice model. This shows that the height of lattice, which translates to the number of visual parts in a concept of a keyword category, is useful in describing the semantics of an image.

Table 2. Performance of APR_DD and APR_CF for keyword groups based on the height of their lattice model

Height of lattice model	APR_DD					APR_CF				
	1	2	3	4	5	1	2	3	4	5
#keywords in group	34	50	99	54	23	34	50	99	54	23
%words with recall>0	0	0.22	0.25	0.46	0.60	0	0.22	0.26	0.46	0.61
Average recall	0	0.05	0.08	0.16	0.40	0	0.05	0.08	0.16	0.41
Average precision	0	0.07	0.05	0.09	0.17	0	0.09	0.05	0.09	0.16

Effectiveness of Level-Sensitive Annotation. To assess the performance improvement due to level-sensitive annotation, we compare the cases with and without the level-sensitive matching. When the level-sensitive matching is disabled, the annotation is performed purely on the dd or $conf$ measure. The result

shown in Table 3 clearly indicates that the performance decreases without level sensitivity matching. Compared to the lower level blobsets, higher level blobsets are more discriminant and thus provides more visual evidence to support the presence of a keyword category. In addition, the *conf* measure is found to be less sensitive to the concept level compared to *dd*. We believe it is because most of the keyword categories in the Corel data set have diverse range of visual appearances. As discussed in Section 2.2, the *conf* measure is more robust to such data set than the *dd* measure. Level-sensitive matching is able to reduce this gap and improve the performance of both measures through selective matching.

Table 3. Performance of APR_DD and APR_CF *without* level-sensitive matching

	All 260 Words		Best 49 Words		Re.>0
	Avg. Re.	Avg. Pr.	Avg. Re.	Avg.Pr.	
APR_DD	0.04	0.03	0.32	0.25	38
APR_CF	0.06	0.04	0.47	0.23	60

Speed Efficiency. The complexity of our approach is $O(W \times N \times C)$ per image, where W is number of words in the vocabulary, N is the number of visual parts in a concept and C is the average number of concept points per word category. As a comparison, the CMRM has a time complexity of $O(W \times R \times J)$, where J is the average training sample of keywords and R is the number of regions in the data set. Obviously, $R > N$. For the Corel data set, $R = 9$ and $N = 3$. Besides, notice that $J > C$ in general since J is required to be large for reliable learning. In the case of Corel data set, the training data of keyword categories varies a lot, $1 \leq J \leq 1004$. In our current implementation, $C = 100$ on average. Our approach requires only 32.48 seconds for training and 1.61 seconds for annotating all the 500 testing images on a Pentium 4 3GHz and 512MB of memory.

4 Conclusions

In this paper, we propose a new approach for image annotation by learning the multiple concept points of keyword categories. Each concept is supported by one or more visual parts and mined using the Apriori principle. Under the guidance of lattice structure, the level-sensitive selection of concepts based on the discriminative and diverse density measure is exploited for effective image annotation. Experiment results show that learning multi visual parts in a model like lattice structure is useful in capturing the semantics of keyword categories.

Acknowledgments. The work described in this paper was fully supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 118905 and CityU 118906).

References

1. Mori, Y., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. MIRSM. (1999)
2. Duygulu, D., Barnard, K., Freitas, N. de, Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. ECCV. (2002) 97–112
3. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. SIGIR. (2003) 119–126
4. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. NIPS. (2003)
5. Feng, S. L., Lavrenko, V., Manmatha, R.: Multiple Bernoulli relevance models for image and video annotation. CVPR. (2004) 1002–1009
6. Blei, D., Jordan, M. I.: Modeling annotated data. SIGIR. (2003) 127–134
7. Carneiro, G., Vasconcelos, N.: Formulating semantic image annotation as a supervised learning problem. CVPR. **2**(2005) 163–168
8. Ghoshal, A., Ircing, P., Khudanpur, S.: Hidden Markov Models for Automatic Annotation and Content-Based Retrieval of Images and Video. SIGIR. (2005) 544–551
9. Szummer, M., Picard, R.: Indoor-Outdoor Image Classification. Workshop in Content-based Access to Image and Video Databases. (1998)
10. Shi, R., Chua, T. S., Lee, C. H., Gao, S.: Bayesian Learning of Hierarchical Multinomial Mixture Models of Concepts for Automatic Image Annotation. CIVR. (2006) 102-112.
11. Tan, P., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison Wesley. (2006)
12. Maron, O., Raton, A. L.: Multiple-Instance Learning for Natural Scene Classification. ICML. (1998) 341–349
13. Zhang, Q., Yu, W., Goldman, S. A., Fritts, J. E.: Content-Based Retrieval Using Multiple-Instance Learning. IMCL. (2002) 682–689
14. Shi, Y. Malik, J.: Normalized cuts and image segmentation. CVPR (1997) 731–737