Exploiting Ontologies for Automatic Image Annotation

Munirathnam Srikanth, Joshua Varner, Mitchell Bowden, Dan Moldovan
Language Computer Corporation
Richardson, TX, 75080
[srikanth,josh,mitchell,moldovan]@languagecomputer.com

ABSTRACT

Automatic image annotation is the task of automatically assigning words to an image that describe the content of the image. Machine learning approaches have been explored to model the association between words and images from an annotated set of images and generate annotations for a test image. The paper proposes methods to use a hierarchy defined on the annotation words derived from a text ontology to improve automatic image annotation and retrieval. Specifically, the hierarchy is used in the context of generating a visual vocabulary for representing images and as a framework for the proposed hierarchical classification approach for automatic image annotation. The effect of using the hierarchy in generating the visual vocabulary is demonstrated by improvements in the annotation performance of translation models. In addition to performance improvements, hierarchical classification approaches yield well to constructing multimedia ontologies.

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Information search and retrieval—*Retrieval models*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object Recognition*

General Terms

Algorithms, Measurement, Experimentation

Keywords

Automatic image annotation, image retrieval, ontologies, translation models, hierarchical classification models

1. INTRODUCTION

With significant improvements in text search and natural language question answering, there is growing interest in multimedia information retrieval and question answering. Early approaches to video and image retrieval were based on attributes like the color and texture of images or key frames of video. User queries, however, are better expressed in natural language than image attributes. It is easier to say `find images with aircraft in an hangar` than to generate the corresponding color, texture or sketch query. Automatic image annotation – the task of associating text to the semantic content of images – has been used as an intermediate step to image retrieval.

Different machine learning methods for image annotation model the association between words and images or image regions. These include translation models [7], classification approaches [14, 5] and relevance models [10, 13]. While most models have used the co-occurrence of image regions and words, few have explored the dependence of annotation words on image regions [3]. This paper proposes to exploit ontological relationships between annotation words and demonstrate their effect on automatic image annotation and retrieval.

The hierarchical aspect cluster model [2] for image annotation adapts a generative model that induces an hierarchical structure from co-occurrence data [9]. The topology of the hierarchy is externally defined and image regions and annotation words are attached to nodes of the hierarchy based on clustering of images from a training set. The hierarchy is intended to represent the various levels of generality of the concepts expressed in image regions and words. Based on their semantics, however, the annotation words themselves can be placed in an hierarchy of concepts. For example, the Corel data set provided by [7] includes images of `cougar`, `leopard`, `lion`, `lynx`, `tiger` and `cats`. Lexical and ontological resource like WordNet organizes different animals in an hierarchy and can place `cougar`, `leopard`, etc. under `cat`. In this paper, such hierarchical dependencies between annotation words are used to generate improved visual lexicons for the translation-based approaches. The intent is to use the hierarchies to capture the visual similarities among different cats.

The hierarchy induced by WordNet is the framework for the proposed *hierarchical classification approach to image annotation*. Given a visual vocabulary of blobs, each node in the hierarchy is associated with a statistical language model defined on blobs characterizing the concept represented by that node. Blob-likelihood probabilities for different concepts are estimated by combining the different language models of the nodes in an hierarchy using shrinkage. While hierarchical classification yields well to the image annotation problem, the hierarchical organization of concepts

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

and the associated language models defined on the blobs provides a representation for multimedia ontologies.

The next section motivates the relevance of the hierarchy induced on the annotation words for automatic image annotation and the need for such an organization to support multimedia information retrieval and natural language question answering.

2. MOTIVATION

Text ontologies have been shown to improve information retrieval and natural language question answering. Different TREC experiments [20] have explored the use of knowledge bases for document retrieval. In the context of imager retrieval based on image captions, Smeaton and Quigley [19] use Hierarchical Concept Graphs (HCG) derived from WordNet to estimate the semantic distance between caption words for similarity computation between query terms and image captions. In natural language question answering, Moldovan and Novichi [16] use the various relationships defined among WordNet concepts and their gloss to discover and weight lexical chains between concepts, and demonstrate the use of these lexical chains in question-answering.

To support questions addressed to multimedia data where the answer is represented in multiple modalities, multimedia ontologies must be able to express concepts in multiple media formats and provide cross-modal relationships to support reasoning. Consider the question, `What is the jersey color of Brazil's national soccer team?`. Given that the picture and caption in Figure 1 is in the image collection, cross-modal reasoning is required to identify that Ronaldo is a Brazilian player and the color of his jersey is yellow.

Figure 1: Image with the answer to the question on the color of the Jersey of Brazil. The associated text is *Ronaldo seals Brazil's place in the last eight with a shot through Geert de Vlieger's legs late on to eliminate Belgium.* The color of Ronaldo's jersey will not show up in black and white versions of this picture.

An ontology should include an organization of concepts of interest and the relationships among them. In WordNet, each concept node is associated with (1) the different surface forms in which the concept occurs in text, (2) its sense and (3) a gloss of sentences that show the use of the concept in that particular sense. While identifying the occurrence of a concept in a document is easily accomplished by searching for the occurrence of one of its surface forms, detecting concept occurrence in image or video (or any multimedia data) is not trivial. For example, it is easier to perform string match and detect the occurrence of `tiger` in text than recognizing an orange patch with black stripes as a tiger in an image. The models used for image annotation define the mechanism for detecting concept occurrence in images or video key frames. Associating a model and its parameters with a concept node extends a text ontology to a multimedia ontology. This can be done by associating the visual tokens for a concept with its node in the ontology. An image of tiger on grass contains image regions that identify tiger and grass. The image regions that characterize tiger can be placed under the concept of tiger in the multimedia ontology. The image can be made part of the gloss for that concept.

The use of text ontologies as a basis for defining visual vocabulary or as a framework for automatic image annotation increases the number of concepts an image annotation system can recognize for a given image. Based on the hypernymy (ISA hierarchy) of annotated words, it is possible to annotate images with words that do not have explicit examples in the training set. Image annotation system using the Corel data set can annotate an image as an `animal` even when no image in the training set that is so annotated. Regions corresponding to images of `tiger`, `cougar`, `cat`, `horse`, `cow`, etc., can be placed under the concept node of `animal` and be used to learn the characteristics of animal allowing its prediction as an annotation.

We restrict ourselves to using lexical resources like WordNet to induce hierarchies in the annotation words. Additional aspects of WordNet, like the gloss, that can be used in disambiguating the sense of the annotation words are left for future work.

The rest of the paper is organized as follows: the next section presents related work in automatic image annotation. In the context of translation models, Section 4 discusses the use of an hierarchy induced on the annotation words to generate visual vocabulary used to represent image regions. Section 5 presents the hierarchical classification approach to image annotation. The experiments reported here are based on the annotated images from the Corel data set by Dugulu et al. [7]. Each image is associated with up to 5 keywords. They provide limited linguistic context to use natural language processing techniques to identify the sense of an annotation word. Section 6 discusses an approach adopted to identify the sense of a word for the purposes of generating the annotation word hierarchy. Section 7 reports on the experiments comparing the translation and classification methods for image annotation under different settings of the visual vocabulary used for representing images. Section 8 concludes the paper and describes future work.

3. RELATED WORK

A number of machine learning approaches have been explored for the automatic image annotation problem. With an annotated training set of images, models proposed for image annotation learn from the co-occurrence of words and images or image regions. Image regions can be generated using image segmentation techniques like N-cuts [18] or from grids. In their co-occurrence model, Mori et al. [17] used a

grid-based segmentation method to identify image regions and used the co-occurrence of words and image regions to predict image annotations.

Duygulu et al. [7] modeled image annotation as translating visual representation of concepts in an image to their textual representation. The visual vocabulary is generated by clustering the image regions identified using N-cuts segmentation algorithm. IBM Model 2 [4] was used in their image annotation experiments. Observing the skewness in the estimation of translation probabilities of uncommon words, Jin et al. [12] proposed regularization of the translation model based on Brown's Model 1. While the translation models capture the correlation between blobs and words, the dependencies between blobs or words are not captured. Dependence between annotation words in the training set, based on an hierarchy derived from WordNet, is used in this paper for selecting the visual vocabulary.

Motivated by the statistical clustering models proposed by Hoffmann and Puzicha [9] for co-occurrence data, Barnard and Forsyth [2] adapted the hierarchical aspect cluster model for image annotation. The hierarchy of models for generating word and image elements is derived by clustering images in the training set. The clusters capture contextual similarities while the nodes capture generality of concepts. Words and blobs are then represented as a distribution over the nodes of the hierarchy, thus words and blobs with similar distributions can be considered correlated. While the hierarchies induced by image clusters provide some semantic interpretation for the models, the structure of the hierarchy is manually specified. The proposed hierarchical classification approach for image annotation uses an hierarchy derived from a knowledge source that is based on concept organization in natural language. It is expected that removal of this arbitrariness in the structure of the hierarchy will provide improved results.

By viewing each annotated word as an independence class, text classification approaches have been adapted for the image annotation task. These models are defined on image attributes that can be generated to predict the likelihood of an annotation word for a given image. Classification approaches for image annotation and retrieval include linguistic indexing of images [14] and Support Vector Machines (SVM) [5]. Motivated by the hierarchical classification approach using shrinkage proposed by McCallum et al. [15], the proposed hierarchical classification approach exploits the hierarchy in the annotation words derived from WordNet for image annotation.

Different probabilistic models have been proposed for automatic image annotation. Blei and Jordon [3] have proposed graphical models of increasing sophistication of capturing the dependence between words and image regions: Gaussian mixture models, Guassian-Multinomial LDA and correspondence Latent Dirichlet Allocation (LDA) for the image annotation problem. Recently, Relevance models [10, 13, 8] for image annotation have shown significant performance improvements. The model learns the joint probability of associating words to image features from training set and uses it to generate the probability of associating a word to a given query image. While the above models predicted the probability, $P(w|I)$ of an annotation word w given an image, I, one is interested in generating a the set of annotation words,$\{w\}$. While annotation of certain length can be selected by ranking the probabilities $P(w|I)$, Jin et al. [11] proposed the Coherent Language Model that predicts the annotation words, $\{w\}$, by relaxing the estimation of $P(\{w\}|I)$ to estimating the probability, $P(\theta_w|I)$ of a language model θ_w to generate the annotation words for the image I.

Except for the correspondence LDA model [3] that captures the dependence between image regions and word annotations, and Coherent Language Model [11] that captures the correlation between annotated words **of an image**, all the above models assume independence of word and image element events in their generative models for image annotation. The image annotation and retrieval approaches proposed here use the dependencies between annotation words represented by the hierarchy derived from an text ontology.

The ARDA VACE program studies the use of ontologies in the visual domain extensively, but the ontologies used there are mainly related to event detection and the process of representing complex events as combinations of simple events. Using ontologies in the context of annotations and object recognition would aid in the identification of the entities involved in an event.

4. ANNOTATION WORD HIERARCHY AND TRANSLATION MODELS

Annotation words of an image describe the semantics of the image in text. Concepts expressed in text can be organized in hierarchies derived from a knowledge base, in our case, a lexical resource (WordNet). This section proposes to use the hierarchy induced in the annotation words by WordNet in the translation model for image annotation.

In the translation model, images are segmented to images regions. Feature vectors capturing the image attributes of a region are clustered to generate the visual vocabulary or the lexicon for image representation. Each image region is mapped to a element of this lexicon referred to as a blob. We use the IBM translation model 2 [4] that includes assignment probabilities of translating a blob in an image to a word. Given a training set, T, of N images, with each image J_i represented by $\mathbf{w_i}, \mathbf{b_i}$ with blobs $\mathbf{b} = \{b_{i1}, b_{i2}, \ldots, b_{in_i}\}$ and annotated words $\mathbf{w} = \{w_{i1}, w_{i2}, \ldots, w_{im_i}\}$, the translation probability $P(w|b)$ is estimated from the likelihood

$$L(T) = \prod_{i=1}^{N} \prod_{j=1}^{m_i} \sum_{k=1}^{n_k} P(a_{ijk}) P(w_{ij}|b_{ik}). \quad (1)$$

where $P(a_{ijk})$ corresponds to the probability of assigning the blob b_k to annotation word w_j in the image J_i. The Expectation-Maximization (EM) [6] is used to estimate the translation probabilities.

Ontology-induced Visual Vocabulary

The hierarchy induced on the annotation words is used to define the visual vocabulary for images. In their translation model experiments, Duygulu et al. [7] use K-means clustering to cluster the images regions to generate the visual vocabulary. The effectiveness of the clusters generated by K-means clustering depends on the selection of the exemplars assigned as cluster centers. Instead of a random selection of initial cluster centers for the K-means clustering, the hierarchy in the annotation words are used to perform the selection.

Given the hierarchy of annotation words that includes other concepts related to the annotation words, images re-

gions are grouped under each node in the hierarchy. An image region r in image I is placed under a concept w if either (1) w is an annotation word for the image I, or, (2) one of its hyponyms (descendants in the hierarchy) annotates the image, I. While image regions placed under a concept may characterize aspects of the image not necessarily related to the concept, averaging the feature vectors of the images regions in this set produces a semantically-motivated initial cluster center for the K-means clustering. This initial clustering of image regions is used in the K-means clustering algorithm to generate the blobs.

Weighted K-means clustering
In each iteration of the K-means clustering algorithm, cluster centers are computed by averaging the feature vectors of the image regions placed in the cluster. Based on the association between between image regions and annotation words, weights can be assigned to image regions quantifying their contribution to the particular cluster. Let W_r be the set of words from the set of annotation words, W, that is associated with a region r. The weight $wt(r, c)$ is computed as

$$wt(r,c) = \prod_{w \in W_r} \frac{n(w,c)+1}{n(c)+|W|} \tag{2}$$

where $n(w, c)$ is the number of image regions in cluster c associated with the word w and $n(c)$ is the number of regions in the cluster c. The cluster center for a cluster c with regions R_c assigned to it is then given by

$$c^* = \sum_{r \in R_c} wt(r,c) f(r) \tag{3}$$

where $f(r)$ is the feature vector for region r. The weight corresponds to the relevance of a region to the cluster based on the words associated with the cluster.

The results of translation models for image annotation based on the above selection criteria for visual vocabulary are presented in Section 7.

5. IMAGE ANNOTATION BY HIERARCHICAL CLASSIFICATION

Motivated by the statistical language modeling approach based on shrinkage for hierarchical classification [15], we propose a hierarchical classification approach for image annotation based on the hierarchy induced on annotation words derived from WordNet. Viewing each annotation word as a class label, models defined on blobs are generated for prediction the the likelihood probability, $P(w|I)$, of assigning the class label w given an image, I.

Adapting the notation of McCallum et al. [15] we describe to approach to image annotation as a classification problem. Assuming that the image data is generated using a mixture model, θ, with one-to-one correspondence between the model components and annotation words, $w_k \in W$. The generative model for an image J_i then corresponds to selecting an annotation word, w_k based on the priors $P(w_k|\theta)$ and using the corresponding mixture component to generate the image J_i according to $P(J_i|w_k, \theta)$. The marginal probability of generating an image J_i is given by

$$P(J_i|\theta) = \sum_{k=1}^{|W|} P(w_k|\theta) P(J_i|w_k,\theta) \tag{4}$$

Each image can be viewed as a sequence of blobs or visual terms from a visual vocabulary, V. An image can be modeled to have been drawn from a multinomial distribution of blobs by making the naive Bayes assumption that each selection of a blob for the image is independ of its context given the annotation word, and it is also independent of the position of the blob in the image. The image likelihood given the annotation word, w_k, is given by

$$P(J_i|w_k,\theta) = P(N_{J_i}) \prod_{l=1}^{N_{J_i}} P(b_{J_{il}}|w_k,\theta) \tag{5}$$

The parameters for this model includes the mixture components θ_k for each annotation word w_k and the blob probabilities, $\theta_{kl} = P(b_l|w_k, \theta)$ such that $\sum_k \theta_{kl} = 1$.

Given a annotated training set of images, T, smoothed estimates can be obtained for the parameters of the model. Let $N(b_l, J_i)$ be the number of times blob b_l occurs in image J_i and define $P(w_k|J_i) \in \{0, 1\}$ (Bernoulli model for annotations as they do not repeat for an image) as given by the class label. Then,

$$\theta_{kl} = P(b_l|w_k,\theta) = \frac{1 + \sum_{i=1}^{|T|} N(b_l, J_i) P(w_k|J_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|T|} N(b_s, J_i) P(w_k|J_i)} \tag{6}$$

The Laplace smoothing of the maximum likelihood estimate is used in the above estimate to ensure non-zero probabilities for all blobs; The prior probabilities for annotation words is given by

$$P(w_k|\theta) = \sum_{i=1}^{|T|} P(w_k|J_i)/|T| \tag{7}$$

For a given test image, I, represented by its sequence of blobs, $b_1, b_2, \ldots, b_m$, the probability of a word w_k being its annotation is given by

$$P(w_k|I,\theta) = \frac{P(w_k|\theta) P(I|w_k,\theta)}{P(I|\theta)} \tag{8}$$

$$= \frac{P(w_k|\theta) \prod_{l=1}^{m} P(b_l|w_k,\theta)}{\sum_{r=1}^{|W|} P(w_r|\theta) \prod_{l=1}^{m} P(b_l|w_k,\theta)} \tag{9}$$

In the above no assumption is made on the dependencies, if any, between the annotation words; Each word is associated with a mixture component in the mixture model. The dependencies between the annotation words, as identified by the ISA hierarchy in WordNet is used in the following to generate an hierarchy model for estimating $P(b_l|w_k, \theta)$. Let H be the hierarchy induced on the annotation concepts. Using shrinkage results in the following equation

$$P(b|w,\theta) = \sum_{v \uparrow w} \alpha(v,w) P_{mle}(b|v) \tag{10}$$

As an example, assume the hierarchy of 'tiger' ISA 'cat' ISA 'animal' ISA 'ROOT'. The probability $P(b|'tiger')$ is given by

$$\begin{aligned} P(b|'tiger') = {} & \alpha('tiger','tiger') P_{mle}(b|'tiger') + \\ & \alpha('cat','tiger') P_{mle}(b|'cat') + \\ & \alpha('animal','tiger') P_{mle}(b|'animal') + \\ & \alpha('tiger'|'ROOT') P(b) \end{aligned}$$

Here 'ROOT' is the highest node in all hierarchies that corresponds to the uniform distribution for $P(b)$.

To learn the α's we use held-out images, H. From the training set, T, maximum likelihood estimates are obtained for the concepts in the hierarchy. For an annotation word, w_j, let $\{v_1 = w_j, v_2, \ldots, v_m =' ROOT'\}$ be the set of concepts from the leaf to the root. Note that some of the intermediate nodes themselves can occur as annotations for some images. The corresponding ML estimates be given by $\{\theta_j^1, \theta_j^2, \ldots, \theta_j^m\}$. The objective is to estimate

$$P(b_l|w_k, \theta_j) = \sum_{i=1}^{m} \alpha(v_i, w_k)\theta_{kl}^i \quad (11)$$

The weight $\alpha(v, w)$ is estimated using a simple form of EM [15]. We compare the results of using flat organization of annotation words (baseline classification model) with the hierarchy induced from WordNet (hierarchical classification model).

6. INDUCING HIERARCHIES IN WORD ANNOTATIONS

The text annotations of images in the Corel Data set represent the semantics of the objects in the images. Each image has at most 5 words describing its content. There is limited linguistic context to use a syntactic parser to identify the sense in which the particular words are used in the annotation. Deriving from WordNet, Barnard et al. [1] use the sense of a word that has the largest hypernym in common with the neighboring words in an image annotation. This based on the assumption that senses are shared, in this case with shared parentage, for text associated with pictures.

Since the words used as annotations in the Corel Data set are nouns, we made a coarser assumption that a particular word is used in only one sense in the the whole corpus. The sense for a word is selected based on the number of times its (1) hyponyms (in that sense) and (2) hypernyms (in that sense) appear as annotations of an image in the collection. For example, 'tiger' has two senses in WordNet: Sense 1 - tiger: a fierce or audacious person, and Sense 2 - tiger: large feline of forests; a big cat. Due to the corpus having a number of animal pictures, the second sense is assigned for 'tiger' and accordingly placed in the hierarchy under animals. The context provided by other annotation words is important to determine the sense of a word. In the settings of a multimedia ontology, different senses of a concept will have nodes of their own right and accordingly, image samples included in their gloss representing the correct sense of the words.

7. EXPERIMENTS

The experiments are performed on the Corel Data set provided by [7]. The data set consists of 5000 annotated images split into a training set of 4500 images with the remaining 500 used for testing. Each image is segmented into multiple regions represented by 36 image features, such as color, position, texture and shape [7]. A total of 371 words annotate the data set with up to 5 annotation words for each image. The reference translation model results are based on the blobs generated by clustering the image regions from the training set into 500 clusters using K-means clustering algorithm. This set of blobs is referred to as *KM-500*.

Based on the method described in Section 6, an hierarchy with 714 nodes is generated for the 371 unique annotation words. This hierarchy is used in the experiments described below. Assuming one blob per concept node in the hierarchy, 714 clusters are generated by initializing the K-means clustering algorithm with the images regions associated with the concept. This set of clusters is referred to as *ONT-714*. To compare the results obtained using KM-500 with 500 clusters, we reduced the set of 714 blobs to 500 by iteratively combining closest clusters. In each iteration, the two closest clusters are replaced by averaging their cluster centers. This set is referred to as *ONT-500*. While assigning one blob for a concept node ensures that each concept should have at least one 'visual surface form' in the vocabulary, more sophisticated methods can be explored for the number of blobs per concept and the selection of image regions for generating the visual vocabulary. This is left for future work.

Weighted K-means clusters are generated based on the method presented in Section 4. This set of referred to as *WKM-500*.

As with previous studies on automatic image annotation, the quality of annotation is evaluated by comparing the generated annotations with actual image annotations in the test set. Each image is annotated with the top 5 words based on their likelihood probability. Some images do not have five annotation words, so this corresponds to the unnormalized score in [7]. For a given word, w, let $N(w)$ be the number of images annotated by w of which let $r(w)$ be the number of images that were *correctly* annotated. Let $R(w)$ be the total number of with w as its true annotation. Precision and recall are defined by

$$Precision(w) = r(w)/N(w) \quad (12)$$

$$Recall(w) = r(w)/R(w) \quad (13)$$

Overall performance is measured by the precision and recall values averaged over all annotation words.

In addition to these measures, the number of predicted words and words with positive recall are included in the results. The number of words predicted as an annotation at least once for any image in the test set differs for different annotation methods. To make fair comparison of the different methods, the average precision and recall values are evaluated for the union of all predicted words. This includes all words that are predicted by at least one of the compared methods.

7.1 Use of Ontology in Translation Model

The results of using the different visual vocabularies in translation-based approach to image annotation is given in Table 1. The precision and recall values along with the number of words predicted and the number of such words with positive recall is presented. The reference K-Means vocabulary gives higher precision and recall over weighted K-Means and ontology-induced vocabularies. However, the number of words predicted is less compared to the ontology-induced vocabularies.

The improvement obtained by using the ontology induced visual vocabulary is evident from the precision and recall values in the last two rows. These are averaged over the words that were predicted by at least one of the annotation methods. The ONT-714 vocabulary gives 19.5% improvement in average precision and 13% increase in average precision over the KM-500 vocabulary. Reducing the ontology to 500 clusters (ONT-500) results in 12.6% and 3.6%

Measures	KM-500	WKM-500	ONT-714	ONT-500
Precision	0.3306	0.3177	0.3074	0.3159
Recall	0.3618	0.3926	0.3178	0.3180
Predicted	28	27	36	33
Positive Recall	27	26	35	32
All 42 words				
Precision	0.2204	0.2042	0.2634	0.2482
Recall	0.2412	0.2524	0.2724	0.2499

Table 1: Comparing the performance of Translation Models using different visual vocabularies

Measures	KM-500	WKM-500	ONT-714	ONT-500
Baseline Classification Method				
Precision	0.1627	0.1867	0.1647	0.1643
Recall	0.2766	0.2831	0.2724	0.2697
Predicted	152	153	150	141
Positive Recall	86	90	84	80
Hierarchical Classification Method				
Precision	0.1805	0.1882	0.1723	0.1754
Recall	0.3174	0.3135	0.2926	0.2903
Predicted	146	140	150	137
Positive Recall	93	91	91	81

Table 2: Comparing the performance of classification methods using different visual vocabularies

improvement in average precision and recall, respectively. With the simple assumption of one blob per concept, the ontology-induced clusters improves image annotation using translation-model. The hierarchy-based initial clusters seem to attach better semantics to the clusters generated by the K-means algorithm.

7.2 Using Ontology in Classification Approaches

Different visual vocabularies were used to evaluate the classification approaches for image annotation. The baseline classification model views each annotation word as an independent class and learns the blob-likelihood model. The hierarchical classification model uses the hierarchy induced in annotation words to learn an interpolated model for likelihood probabilities. Table 2 summarizes the results of comparing of the two classification models.

The hierarchical classification approach performs better than the baseline classification method on different visual vocabulary settings. Using the KM-500 vocabulary the average precision increase of about 10% and 14% increase observed in average recall. While the number of predicted words reduces, there is increase in the number of words with positive recall. Across different vocabularies, WKM-500 provides the best precision and recall values.

To make fair comparison of the performance of the classification methods, Table 3 presents the precision and recall values for predicting the pooled set of predicted annotation words. The comparison is performed by first fixing the visual vocabulary. Each column corresponds to the selected visual vocabulary. The second row identifies the total number of predicted words for the two classification methods using the same visual vocabulary.

Measures	KM-500	WKM-500	ONT-714	ONT-500
# Words	168	169	171	163
Baseline Classification Method				
Precision	0.1481	0.1701	0.1455	0.1431
Recall	0.2519	0.2580	0.2405	0.2350
Hierarchical Classification Method				
Precision	0.1580	0.1570	0.1521	0.1485
Recall	0.2777	0.2616	0.2584	0.2458

Table 3: Comparing the performance of classification models predicting pooled set of annotation words

Measures	KM-500	WKM-500	ONT-714	ONT-500
# Words	168	169	171	163
Translation Method				
Precision	0.0551	0.0508	0.0647	0.0640
Recall	0.0603	0.0627	0.0669	0.0644
Baseline Classification Method				
Precision	0.1481	0.1701	0.1455	0.1431
Recall	0.2519	0.2580	0.2405	0.2350
Hierarchical Classification Method				
Precision	0.1579	0.1570	0.1521	0.1485
Recall	0.2777	0.2616	0.2584	0.2458

Table 4: Comparing the performance of translation model and classification models

While the two classification methods have comparable number of predicated words (Ref. Predicted word count in Table 2), the pooled results indicate that significant improvements using the KM-500 and ONT-714 vocabulary. The best precision and recall values observed under WKM-500 in Table 2 are not observed in the pooled set. The precision decreases for the hierarchical classification method.

Based on the above results, the annotation word hierarchy seems is effective as a framework for the hierarchical classification model as well as a source for selecting the initial clusters for the visual vocabulary. While the hierarchical classification approach improved in precision and recall under all visual vocabulary settings except WKM-500, the ONT-714 is the only one that does not lose predictive ability in the hierarchical classification. Same number of words are predicted with 5 more words with non-zero recall value.

Table 4 compares the performance of the translation and classification models on the pooled set of predicted words. The classification algorithms show a three to four times improvement over the translation model, mainly due to the ability to predict three or four times as many words as the translation model. The KM-500 set performs the best over the union of predicted words.

The hierarchy induced in annotation words provides overall improvements in automatic image annotation both when used for generating the visual vocabulary for image representation and as a framework for the hierarchical classification method.

8. CONCLUSIONS

This paper proposed methods to use hierarchies induced on annotation words to improve automatic image annotation. While hierarchical clustering models have been explored for the image annotation problem, the hierarchies were statistically derived from image clusters. This paper presents a method for generating visual vocabularies based on the semantics of the annotation words and their hierarchical organization in an ontology like WordNet. The semantically-motivated K-means clustering for generating the blobs improved the performance of the translation models for image annotation.

In the context of classification approaches to image annotation, the hierarchy derived from WordNet is used as a framework to generate classification models defined on blobs. While both classification methods performed better than translation model in our experiments, the hierarchical classification provides significant improvements under different settings of the visual vocabulary. The representation used in the hierarchical classification approach also yields well to defining multimedia ontologies by extending a text ontology like WordNet. Models for detecting concept occurrences in images enable the definition 'visual surface forms' for a multimedia ontology, and visual glosses provide training data for creating these models.

Our experiments and observations are based on the Corel Data set. Captions and automatic speech recognition transcripts provide more linguistic clues than the keyword annotation for images in the Corel Data set. We intend to verify the improvements obtained using the proposed methods on a larger and more challenging data set like TREC Video dataset [21].

Image regions were identified using N-cuts algorithm in our experiments. Grid based image regions have been explored as image surrogates for image annotation [8]. Either methods for segmenting images capture different aspects of an image. Methods that combine both of these representations can be explored for image annotation. While we have explored only the hypernymy relations in WordNet other relations between concepts can be explored to capture such correlations between blobs.

We have proposed methods to use ontologies in the preprocessing stage to define the visual vocabularies for images as well as generating hierarchical models for automatic annotation. Ontologies can also be used to define a contextual model at the end of automatic annotation to disambiguate the annotations words assigned for a given image.

9. REFERENCES

[1] K. Barnard, P. Duygulu, and D. A. Forsyth. Modeling the Statistics of Image Features and Associated Text. In *Document Recognition and Retrieval IX - Electronic Imaging*, 2002.

[2] K. Barnard and D. A. Forsyth. Learning the Semantics of Words and Pictures. In *Proceedings of International Conference on Computer Vision*, pages 408–415, 2001.

[3] D. Blei and M. Jordan. Modeling Annotated Data. In *Proceedings of SIGIR'03*, 2003.

[4] P. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 32(2):263–311, 1993.

[5] C. Cusano, G. Ciocca, and R. Schettini. Image Annotation using SVM. In *Proceedings of Internet Imaging IV*, 2004.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[7] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In *Proceedings of European Conference on Computer Vision*, pages 97–112, 2002.

[8] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli Relevance Models for Image and Video Annotation. In *Proceedings of CVPR'04*, 2004.

[9] T. Hoffmann and J. Puzicha. Statistical Models for Co-occurrence Data. A. I. Memo 1635, 1998.

[10] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. In *Proceedings of SIGIR'03*, 2003.

[11] R. Jin, J. Y. Chai, and L. Si. Effective Automatic Image Annotation via A Coherent Language Model and Active Learning. In *Proceedings of MM'04*, 2004.

[12] F. Kang, R. Jin, and J. Y. Chai. Regularizing Translation Models for Better Automatic Image Annotation. In *Proceedings of CIKM'04*, 2004.

[13] V. Lavrenko, R. Manmatha, and J. Jeon. A Model for Learning the Semantics of Pictures. In *Proceedings of NIPS'03*, 2004.

[14] J. Li and J. Z. Wang. Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(19):1075–1088, 2003.

[15] A. K. McCallum, R. Rosenfeld, T. M. Mitchell, and A. Y. Ng. Improving Text Classification by Shrinkage in a Hierarchy of Classes. In *Proceedings of ICML-98, 15th International Conference on Machine Learning*, pages 359–367, Madison, US, 1998.

[16] D. Moldovan and A. Novischi. Lexical Chains for Question Answering. In *Proceedings of COLING'02*, pages 674–680, 2002.

[17] Y. Mori, H. Takahashi, and R. Oka. Image-to-Word Transformation based on Dividing and Vector Quantizing Images with Words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[18] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

[19] A. F. Smeaton and I. Quigley. Experiments on Using Semantic Distances Between Words in Image Caption Retrieval. In *Proceedings of SIGIR'96*, pages 174–180, 1996.

[20] Text retrieval conference. `http://trec.nist.gov`.

[21] Trec video data set. `http://www-nlpir.nist.gov/projects/trecvid`.