# Using Statistics to Search and Annotate Pictures: an Evaluation of Semantic Image Annotation and Retrieval on Large Databases

Antoni B. Chan[1], Pedro J. Moreno[2], Nuno Vasconcelos[1]
Department of Electrical and Computer Engineering, University of California, San Diego[1]
Google, Inc.[2]

**Abstract**

We present the results of an extensive experimental evaluation of the supervised multi-class labeling (SML) model for semantic image annotation proposed by [16]. We test the robustness of this model to various parameters, and its scalability in both database and vocabulary size. The results of this study complement previous evaluations by [12], and [16], which were limited to smaller databases and vocabularies. We further compare the performance of SML to that of a model that explicitly trades off retrieval performance for scalability: the supervised category-based labeling (SCBL) model of [14]. This establishes a unifying view of the performance of two classes of labeling and retrieval systems that were previously only evaluated under different experimental protocols. This unification simplifies the evaluation of future proposals for semantic labeling and retrieval.

**Keywords**: semantic image annotation, semantic image retrieval, experimentation, evaluation.

## 1 Introduction

Content-based image retrieval, the problem of searching large image repositories according to their content, has been the subject of substantial research in the recent past [1]. While early retrieval architectures were based on the query-by-example paradigm, which formulates image retrieval as the search for the best database match to a user-provided query image, it was quickly realized that the design of fully functional retrieval systems would require support for semantic queries [2]. These are systems where the database of images are annotated with semantic keywords, enabling the user to specify the query through a natural language description of the visual concepts of interest. This realization, combined with the cost of manual image labeling, generated significant interest in the problem of automatically extracting semantic descriptors from images.

The earliest efforts in the area were directed to the reliable extraction of specific semantics, e.g. differentiating indoor from outdoor scenes [3], cities from landscapes [4], and detecting trees [5], horses [6], or buildings [7], among others. These efforts posed the problem of semantics extraction as one of *supervised* learning: a set of training images with and without the concept of interest was collected and a binary one-vs-all classifier was trained to detect the concept of interest. The classifier was then applied to all database images which were, in this way, annotated with respect to the presence or absence of the concept. This has two limitations. First, the complexity of learning each binary classifier depends on the size of the "absence" class, making the model very difficult to learn when there is a large vocabulary. Second, the binary formulation is not amenable to weakly labeled data, in the sense that training images that contain the concept, but that are mislabeled as not containing it, are incorrectly assigned to the "absence" class, and can compromise the classification accuracy.

In response to some of these limitations, namely the lack of scalability to problems involving large vocabularies (hundreds or even thousands of concepts), there has recently been a shift to the adoption of semantic labeling techniques based on *unsupervised* learning [8, 9, 10, 11, 12, 13]. The basic idea is to introduce a graphical model containing a set of latent variables that encode hidden states of the world, where each state defines a joint distribution on the space of semantic keywords and image appearance descriptors (in the form of local features computed over image neighborhoods). During training, the image is represented as a collection of regions (either block-based [10, 12, 14] or segmented regions [8, 9, 11, 15]), and an unsupervised learning algorithm is run over the entire database to estimate the parameters of the joint density of words and visual features. Given a new image to annotate, visual feature vectors are extracted, the joint probability model is instantiated with those feature vectors, state variables are marginalized, and a search for the set of labels that maximize the joint density of text and appearance is carried out.

While the unsupervised formulation leads to significantly more scalable (in database size and number of concepts) training procedures and places weaker demands on the quality of the manual annotations for training, it does not explicitly treat semantics as image classes and, therefore, provides little guarantees that the semantic annotations are optimal in a recognition or retrieval sense. That is, instead of annotations that achieve the smallest probability of retrieval error, it simply produces the ones that have largest joint likelihood under the assumed model. This has motivated the recent introduction of a model, *supervised multi-class labeling* (SML), that combines the advantages of both supervised and unsupervised methods of semantic annotation, through a slight reformulation of the supervised learning model (see [16, 17]). This reformulation consists of defining a multi-class classification problem where each of the semantic concepts of interest defines an image class, and has various important practical consequences. For example, it eliminates the need to compute a "non-class" model for each of the semantic concepts of interest, while producing a natural ordering of semantic keywords

(i.e., image classes) when annotating a test image. In result, the new model has learning complexity equivalent to that of the unsupervised formulation but, like the supervised one, it places weak requirements on the quality of manual labels. Finally, by explicitly treating semantics as image classes, it provides optimality guarantees in a recognition or retrieval sense.

Independently of algorithmic advances, one of the most significant recent contributions to the advance of semantic labeling and retrieval was the establishment, by [12], of an evaluation protocol that allows easy comparison of any new retrieval algorithm to most of the state-of-the-art techniques. In particular, [12] proposed a database (which we refer to as *Corel5K*), and a set of evaluation metrics (which we discuss in detail in Section 3), and evaluated a number of techniques [11, 12, 15] according to these metrics. The facts that 1) evaluation of SML under this protocol has outperformed all previously tested methods [16, 17], and 2) SML is relatively straightforward to implement, place SML as a natural benchmark against which to compare other algorithms. There are, nevertheless, a number of questions that remain open with respect to the evaluation of semantic labeling and retrieval systems.

First, because Corel5K is a relatively small database, it is not clear that the conclusions based on the protocol of [12] will hold for larger databases, or databases annotated with larger vocabularies. Second, although relying on a simple image representation, SML requires the specification of a number of parameters (dimension of the feature space, number of mixture components, number of hierarchical stages for density estimation, etc.). It would be desirable to measure the sensitivity of retrieval performance to these parameters. Third, a number of competing semantic retrieval techniques were never considered in [12, 16, 17], including techniques specifically designed to achieve scalability in database size. Of particular interest is the method proposed in [14], that relies on a representation similar to that of SML, but utilizes an alternative image annotation philosophy, which we refer to as *supervised category-based labeling* (SCBL). Although this type of labeling is more error-prone than that adopted by the protocol of [12], it requires significantly less manual effort for the creation of training sets. A quantification of the loss, with respect to retrieval performance, of SCBL is currently not available.

This work presents the conclusions of a large-scale evaluation effort that aims to answer most of these questions. We adopt the SML retrieval model and start by testing its scalability, by evaluating it on two databases composed of 30,000 and 60,000 images. We then perform a fairly extensive evaluation of its robustness with respect to various parameters. This evaluation is followed with a detailed comparison with the model of [14]. The main conclusions of the study are as follows. First, the results obtained with the protocol of [12] on Corel5K seem to hold for larger databases. While precision-recall decreases with vocabulary size (larger vocabularies have more specialized words, and more words with the same semantic meaning), annotation performance is qualitatively similar to that observed on Corel5K. Second, the SML model is quite robust to the selection of its parameters. We found the performance to be nearly constant over a wide range of configurations. Third, the performance loss of SCBL can be quite significant. Finally, we compare the image representation of SML with the more sophisticated 2D-MHMM model of [14], and show that the former has significantly better performance. This illustrates the important point that more complex models are not always best, since the added complexity can compromise the generalization ability of retrieval algorithms.

Overall, we see as major contribution of this work, especially when considered as a complement to [16, 12], the unification of results of different annotation systems, previously only evaluated under different experimental protocols. As far as we are aware of, this is the first effort of this magnitude: some of the experiments reported in the following sections required more than 30 hours on a cluster of 2,000 state-of-the-art Linux processors. We hope that the unified view on the performance of existing annotation algorithms will 1) be of use for other researchers, and 2) help solidify the current trend, initiated by [12], towards the adoption of standard databases and evaluation protocols. The paper is organized as follows. In Section 2, we start by reviewing, and comparing, the SML and SCBL models. Next, we discuss image databases, experimental protocol, and details of experiment implementation in Section 3. Finally, we analyze results in Section 4.

## 2 Semantic Image Annotation for Large Databases

The goal of semantic image annotation is to, given an image $\mathcal{I}$, select a set of keywords $\mathbf{w}$ from the semantic vocabulary $\mathcal{L} = \{w_1, \ldots, w_L\}$ that best describes $\mathcal{I}$. Likewise, the goal of semantic retrieval is to, given a keyword $w$, select images from an image database that contain the associated visual concept. In both cases, learning is based on a training set of image-caption pairs, $\mathcal{D} = \{(\mathcal{I}_1, \mathbf{w}_1), \ldots, (\mathcal{I}_D, \mathbf{w}_D)\}$. We next review two annotations methods, SML and SCBL, that scale well with database and vocabulary size.

### 2.1 Supervised multi-class labeling

SML formulates both annotation and retrieval as classification problems. In particular, SML explicitly makes the elements of the semantic vocabulary the classes of an $M$-ary classification problem by introducing: 1) a random vector $\mathbf{X}$ of visual features; 2) a random variable $W$, which takes values in $\{1, \ldots, L\}$, so that $W = i$ if and only if $\mathbf{x}$ is a sample from the concept $w_i$; 3) a set of class-conditional distributions $P_{\mathbf{X}|W}(\mathbf{x}|i), i \in \{1, \ldots, L\}$ that model the distribution of visual features given the semantic class; and 4) a set of prior probabilities on the semantic class, $P_W(i)$.

Using well known results in statistical decision theory [18], it is not difficult to show that both labeling and retrieval can be implemented with minimum probability of error if the posterior probabilities

$$P_{W|\mathbf{X}}(i|\mathbf{x}) = \frac{P_{\mathbf{X}|W}(\mathbf{x}|i)P_W(i)}{P_{\mathbf{X}}(\mathbf{x})} \quad (1)$$

are available. For annotation, the minimum probability of error rule is to, given a set of query feature vectors $\mathbf{x}$, pick con-

cept

$$i^*(\mathbf{x}) = \underset{i}{\operatorname{argmax}} P_{W|\mathbf{X}}(i|\mathbf{x}) = \underset{i}{\operatorname{argmax}} P_{\mathbf{X}|W}(\mathbf{x}|i)P_W(i). \quad (2)$$

For semantic retrieval, given concept $w_i$, the optimal rule is to select the database image of index

$$j^*(w_i) = \underset{j}{\operatorname{argmax}} P_{\mathbf{X}|W}(\mathbf{x}_j|i) \quad (3)$$

where $\mathbf{x}_j$ is the set of feature vectors extracted from the $j^{th}$ database image $\mathcal{I}_j$. In both cases, the ordering by decreasing posterior probability is a minimum probability of error ranking for the remaining keywords or images. Evaluation under the protocol of [12] has shown that SML outperforms a number of state-of-the-art unsupervised learning techniques.

### 2.2 Supervised category-based labeling

SCBL is specifically designed for labeling very large databases. Because it is time-intensive to hand-label individual images, it proposes the alternative of grouping images into disjoint categories and labeling all images in each category with a unique set of ground-truth words, that best describe the category as a whole. In [14], this idea is implemented in two steps. First, a classifier is applied to each test image to obtain the top five categories for that image, and the annotations from those categories are pooled into a list (with frequency counts for reoccurring annotations). The candidate annotations are then ordered using a statistical test for the hypothesis that an annotation has entered the list by chance. Specifically, the probability that the candidate word appears at least $j$ times in $k$ randomly selected categories is computed

$$P(j,k) = \sum_{i=j}^{k} I(i \leq m) \frac{\binom{m}{i}\binom{n-m}{k-i}}{\binom{n}{k}} \quad (4)$$

where $I(\cdot)$ is the indicator function, $n$ is the total number of image categories, and $m$ is the number of image categories containing the word. A small value of $P(j,k)$ indicates a low probability that the annotation occurred randomly (i.e. the word has high significance as an annotation), and hence lower values of $P(j,k)$ are better.

### 2.3 Comparison of annotation styles

In comparison with SML, SCBL has two advantages: 1) the amount of data required from each image category is small compared to the number of words learned (e.g. in [14], each image category contains 100 images and is labeled with an average of 4 words); and 2) expanding the database with a new image category does not require relearning the existing categories, i.e. only the new image category must be learned. These two properties allow easy addition of new words, or enhancement of previous words, by simply adding new image categories to the database. In contrast, SML annotation becomes less tractable as the database is expanded in the sense that: 1) common annotations, such as "landscape", are associated with many images, and hence learning these class densities becomes time intensive; and 2) when adding new images to the database, the class mixtures must be relearned for all the annotations associated with the new image[1]. In other words, not only is the learning problem more difficult because common classes may be associated with a very large number of images, but expanding an existing database requires relearning all the affected annotation classes.

One disadvantage of SCBL is that a representation of the actual annotation is never learned. Instead, the system learns a representation of groups of words, and uses a hypothesis test to determine the probability that a single word describes the test image. This is in contrast to SML, which learns a representation of a word directly from all its associated ground truth images, and is therefore optimal in the minimum probability of error sense. A second disadvantage of SCBL is that the hypothesis test tends to prefer unique over popular words. As a result, SCBL may not use generic annotations (e.g. landscape, water, grass, sky) that may be appropriate for many images. Also note that $P(j,k)$ tends to take a discrete number of values and that there is no natural ordering for words with the same value of $P(j,k)$. In contrast, ordering of annotations in SML is based on posterior probabilities that are very unlikely to take the same values, and there is almost always a strict ordering of words.

### 2.4 Density estimation

The estimation of class conditional densities raises an interesting complexity question: if the database is large, the direct estimation of $P_{\mathbf{X}|W}(\mathbf{x}|i)$ from the set of all feature vectors extracted from all images that contain the concept $w_i$ is usually infeasible. One solution is to discard part of the data, but this is suboptimal in the sense that important training cases may be lost. An efficient alternative is to adopt the hierarchical density estimation method of [19, 16, 17], based on a mixture hierarchy where children densities consist of different combinations of subsets of their parents' components. With this method, a class conditional mixture density is learned as follows. First, a mixture density is learned for the feature vectors extracted from each image, resulting in a mixture density for each image. Second, a class conditional density is learned by applying a hierarchical EM algorithm [19] to the densities of the images.

For classes with a large number of images, multiple hierarchical levels can be used by splitting the set of densities at the current-level into subsets, and then learning children-densities from these subsets, and so on. In this way, the class conditional distribution can be learned with logarithmic complexity on the number of training images. Note that the number of parameters in each image mixture is orders of magnitude smaller than the number of feature vectors in the image itself. Hence the complexity of estimating the class mixture parameters is negligible when compared to that of estimating the individual mixture parameters for all images in the class. It follows that the overall training complexity is dominated by the latter.

One final interesting property of the hierarchical method is that it enforces a data-driven form of regularization which im-

[1]Note, however, that this computation is greatly reduced if the hierarchical density estimates described later are used, and the image mixtures are saved.

proves generalization [19]. We have observed that, due to this property, hierarchical class density estimates are much more reliable than those obtained by direct learning based on standard EM.

## 3 Experimental Evaluation

Recent works [11, 12, 15] have adopted a "de-facto" experimental protocol, which we refer to as Corel5K, for evaluating semantic annotation systems. This protocol has two significant limitations. First, because the database on which it is based is relatively small, it relies on a very small number of examples for many of the semantic labels. This makes it difficult to guarantee that the resulting annotation systems have good generalization. Second, because the concept vocabulary is also relatively small, it does not necessarily test the scalability of annotation/retrieval algorithms. Some of these limitations are corrected by the Corel30K protocol, an extension of Corel5K based on a substantially larger database. None of these protocols has, however, been used to evaluate some of the most scalable techniques available in the literature. One alternative protocol, originally proposed by [14], is more suitable to test such systems. We refer to it as PSU, and review the three protocols in this section.

### 3.1 Corel5K and Corel30K

The evaluation of a semantic annotation/labeling and retrieval system requires three components: an image database with manually produced annotations, a strategy to train and test the system, and a set of measures of retrieval and annotation performance. The Corel5K benchmark is based on the Corel image database [11, 12, 15]: 5,000 images from 50 Corel Stock Photo CDs, were divided into a training set of 4000 images, a validation set of 500 images, and a test set of 500 images. An initial set of model parameters is learned on the training set. Parameters that require cross-validation are then optimized on the validation set, after which this set is merged with the training set to build a new training set of images. Non-cross-validated parameters are then tuned with this training set. Each image has a caption of 1-5 keywords, and there are 371 keywords in the entire data set (the test set only contains 280 of these).

Image annotation performance is evaluated by comparing the captions automatically generated for the test set, with human-produced ground-truth. Similarly to [12, 15] we define the automatic annotation as the five semantic classes of largest posterior probability, and compute the recall and precision of every word in the test set. For a given semantic descriptor, assuming that there are $|w_H|$ human annotated images in the test set, and the system annotates $|w_a|$, of which $|w_c|$ are correct, recall and precision are given by:

$$recall = \frac{|w_c|}{|w_H|}, \quad precision = \frac{|w_c|}{|w_a|} \tag{5}$$

As suggested by [12, 15], the values of recall and precision are averaged over the set of words that appear in the test set. Finally, we also consider the number of words with non-zero recall (NZR), i.e. words with $|w_c| > 0$. This can be seen as the number of words the system has effectively learned.

The performance of semantic retrieval is also evaluated by measuring precision and recall. Given a query term and the top $n$ image matches retrieved from the database, recall is the percentage of all relevant images contained in the retrieved set, and precision the percentage of the $n$ which are relevant (where relevant means that the ground-truth annotation of the image contains the query term). Under the protocol of [12], retrieval performance is evaluated by the mean average precision (MAP). The retrieved images are ordered according to their scores, and the average precision is computed by averaging over only the precision values where the recall changes (i.e. where relevant items occurred), while increasing the number of retrieved images. The MAP is the mean of the average precision values for all the query words.

The Corel30K protocol is similar to Corel5K but substantially larger, containing 31,695 images and 5,587 words. Of the 31,695 images, 90% were used for training (28,525 images) and 10% for testing (3,170). Only the words (950 in total) that were used as annotations for at least 10 images were learned. Corel30K is much richer than Corel5K, in terms of number of examples per label and database size, therefore posing a much stronger scalability challenge.

### 3.2 PSU protocol

PSU is also based on the Corel set, and contains 60,000 images [14] with 442 annotations. Unlike Corel5K and Corel30K, the PSU ground truth was not obtained by annotating each image separately. Instead, images were associated with an image category (the image set was split into 600 image categories with 100 images each), which were then annotated with a general description that reflects the image category as a whole, but may not accurately characterize each individual image. This strategy saved human annotation time but produced somewhat noisy annotations. For example, some images from the "tiger" category, which is labeled with "tiger", "grass", and "trees", do not contain "trees". For performance evaluation, 40% of the PSU images were reserved for training (23,878 images), and the remainder (35,817 images) used for testing. Note that [14] only used 4,630 of the 35,817 possible test images, whereas the experiments reported here are based on the entire test set.

Following [14], the performance of the image category classifier is evaluated by its accuracy, where a test image is considered to be categorized correctly if any of the top $r$ categories is the true category. Also following [14], the performance of SCBL annotation is measured with the "mean coverage", which is the percentage of ground-truth annotations that match the computer annotations. In addition, we also evaluate the annotation and retrieval performance with the standard measures used in Corel5K and Corel30K. A summary of the experimental protocols is given in Table 1.

| DB name | images | labels | training set | testing set |
|---|---|---|---|---|
| Corel5K | 5,000 | 260 | 4,500 (90%) | 500 (10%) |
| Corel30K | 31,695 | 950 | 28,525 (90%) | 3,170 (10%) |
| PSU | 59,695 | 442 | 23,878 (40%) | 35,817 (60%) |

Table 1: Image databases and experimental setup

| Class Representation | $r=1$ | $r=2$ | $r=3$ | $r=4$ | $r=5$ |
|---|---|---|---|---|---|
| GMM-DCT | 0.209 | 0.270 | 0.309 | 0.338 | 0.362 |
| 2D-MHMM [14] | 0.119 | 0.171 | 0.208 | 0.232 | 0.261 |

Table 2: Accuracy of image categorization on PSU database. A image is correctly categorized if any of the top $r$ categories is the true category.

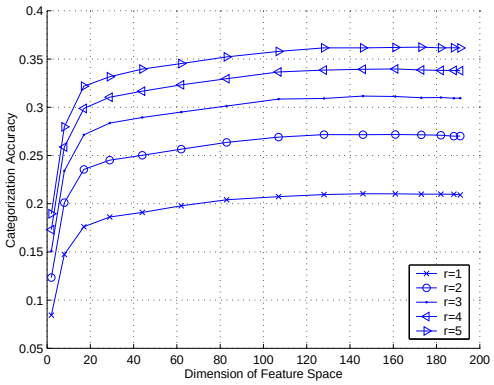



Figure 1: Accuracy of image categorization on PSU using GMM-DCT versus the dimension of the DCT feature space.

### 3.3 Image representation

All experiments were based on the image representation of [16, 17]. In particular, we use the YBR color space, and each image was decomposed into a set of overlapping 8 x 8 windows, which were extracted with a sliding window that moved by one pixel between consecutive samples. A feature vector was obtained, from each location of the three color channels, by the application of the discrete cosine transform (DCT). The image was then represented as a bag $X = \{x_1, \cdots, x_N\}$ of independently sampled feature vectors.

### 3.4 Implementation details

For the implementation of SML, a Gaussian mixture was fit to the set of images associated with a given annotation using the hierarchical method discussed in Section 2.4. A 2-level hierarchy was constructed by first learning mixtures for the feature vectors of the images, and then learning the class conditional from the image mixtures. For the 3-level hierarchy, an additional level split the image mixtures into groups of 250, and mixtures were learned for those groups. The class conditional density was finally learned from the group mixtures. We refer to this representation as GMM-DCT. Unless otherwise noted, the image densities used 8 mixture components and the class conditionals used 64 components. The Gaussian components had diagonal covariance matrices and used the full 192-dimensional space.

Using SML, test images were annotated with the five labels of largest posterior probability. For single query image retrieval (i.e. ranked retrieval), each image was first annotated with five labels. When retrieving images for a query word, all images that received the word as an annotation were selected, and ordered based on the posterior probability of word given image.

The implementation of SCBL for the PSU database used the GMM-DCT representation for each image category. Class conditional distributions also used 64 mixture components and the full feature space. Experiments were conducted on a cluster of 2,000 state-of-the-art Linux machines, and ran between 1 hour for Corel5K and 34 hours for PSU. These times are overestimates, since the experiment was sometimes preempted by other jobs on the cluster.

## 4 Experimental Results

In this section we present the results of annotation experiments. We first compare the performance of SML and SCBL, and then present robustness results of SML on the three databases.

### 4.1 Comparison of SML and SCBL

We started by comparing the image categorization performance between the GMM-DCT class representation and the representation of [14]. In [14], an image category is represented by a two-dimensional multi-resolution hidden Markov model (2D-MHMM) defined on a feature space of localized color and wavelet texture features at multiple scales. Table 2 shows the accuracy of image categorization using the two class representations. GMM-DCT outperformed the 2D-MHMM in all cases, with an improvement of about 0.10 (from 0.26 to 0.36). Figure 1 shows the categorization accuracy of GMM-DCT versus the dimension of the DCT feature space. It can be seen that the categorization accuracy increases with the dimension of the feature space, but remains fairly stable over a significant range of dimensions.

We next compared the annotation performance of SCBL, using GMM-DCT and the 2D-MHMM (we denote the combinations by SCBL-GMM-DCT and SCBL-2D-MHMM). The images were annotated by applying a threshold of 0.0649 to $P(j,k)$, as described in [14]. Table 3 shows the mean coverage of SCBL-GMM-DCT and SCBL-2D-MHMM with and without thresholding. GMM-DCT annotations outperformed those of 2D-MHMM by about 0.12 (from 0.22 to 0.34 with threshold, and 0.47 to 0.61 without). Figure 2 shows the mean coverage versus the dimension of the DCT feature space. Again, performance increases with feature space dimension, but remains fairly stable over a large range of dimensions.

Finally, we compared SCBL and SML when both methods used the GMM-DCT representation. SCBL annotation was performed by thresholding the hypothesis test (SCBL-GMM-DCT threshold), or by selecting a fixed number annotations

| Method | threshold=0.0649 | no threshold |
|---|---|---|
| SCBL-GMM-DCT | 0.3420 | 0.6124 |
| SCBL-2D-MHMM [14] | 0.2163 | 0.4748 |

Table 3: Mean coverage of annotation on PSU database using SCBL.

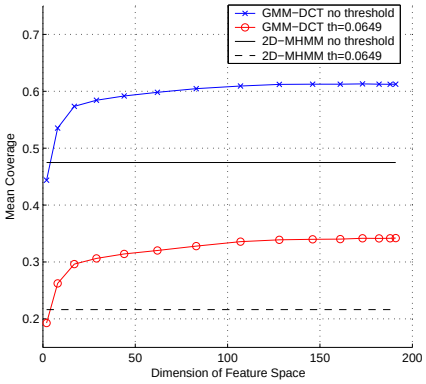



Figure 2: Mean coverage of annotation of PSU using SCBL-GMM-DCT versus the dimension of the DCT feature space.

(SCBL-GMM-DCT fixed). SML classifiers were learned using both 2-level and 3-level hierarchies. Figure 3 presents the precision-recall (PR) curves produced by the two methods. Note that SML trained with the 3-level hierarchy outperforms the 2-level hierarchy. This is evidence that the hierarchical EM algorithm provides some regularization of the density estimates, which improves the performance.

The SML curve has the best overall precision at 0.236, and its precision is clearly superior to that of SCBL at most levels of recall. There are, however, some levels where SCBL-GMM-DCT leads to a better precision. This is due to the coupling of words within the same image category, and to the noise in the ground truth annotations of PSU. Note that if the correct category is in the top 5 classified categories, then the list of candidate words will contain all of the ground truth words for that image. Eventually, as the image is annotated with more words from the candidate list, these ground truth words will be included, regardless of whether the ground truth actually applies to the image (i.e. when the ground truth is noisy). In result, recall and precision are artificially inflated as the number of annotations increases. On the other hand, for SML, each word class is learned separately from the other words. Hence, images will not be annotated with the noisy word if the concept is not present, and the precision and recall can suffer. Finally, for SCBL-threshold, the PR curve has an unusual shape. This is an artifact that arises from thresholding a hypothesis test that has discrete levels.

In summary, the experimental results show that the GMM-DCT representation substantially outperforms the 2D-MHMM of [14] in both image categorization and annotation using SCBL. When comparing SML and SCBL based on the GMM-DCT representation, SML achieves the best overall precision, but for some recall levels SCBL can achieve a better precision due to coupling of annotation words and noise in the

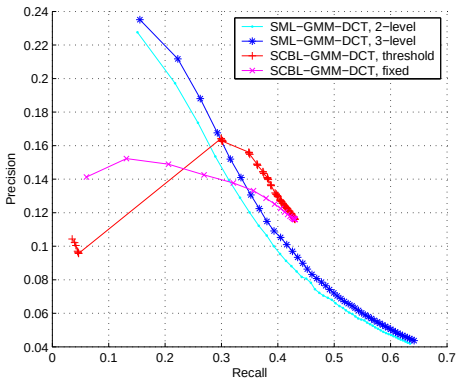



Figure 3: Precision-Recall for SCBL and SML using GMM-DCT on the PSU database.

annotation ground truth.

### 4.2 Robustness and scalability of SML

We have already seen that, under the SCBL model, both the categorization and annotation performance of the GMM-DCT representation are fairly stable with respect to the number of mixture components and feature space dimension. We now report on experiments performed to evaluate the robustness of SML-GMM-DCT to these parameters. Figure 4 (left) shows the PR curves obtained for annotation on Corel5K, as a function of the number of mixture components used to model class conditional densities. Note that the PR curve remains fairly stable above 64 components. Figure 4 (right) shows the PR curve for annotation with 64 components while varying the feature space dimension. In this case, stability is achieved above 63 dimensions.

To test scalability, the SML annotation experiment was repeated on the larger Corel30K. Figure 5 shows the performance obtained with 64 and 128 mixture components, learned with either the 2-level or 3-level hierarchy. Annotation performance on the larger database is qualitatively similar to that obtained on the smaller Corel5K database (e.g. compare the shape of the PR curves with those of Figure 4 (left)), albeit with overall lower precision and recall levels. This is due to the difficulty of learning specialized annotations, and to the presence of different annotations with the same semantic meaning, which are both more frequent on Corel30K. In addition, a 3-level hierarchy outperforms the standard 2-level hierarchy for both 64 and 128 components. This indicates that the regularization of the 3-level structure is superior to that of the standard hierarchical organization. The differences are nevertheless not staggering, suggesting some robustness with respect to this parameter.

Overall, these experiments indicate that 1) SML is fairly stable with respect to its parameter settings, and 2) results on Corel5K are a good indication of the relative performance of different techniques (albeit the absolute PR values are likely to be overestimated). Finally, Table 4 shows the average precision and recall for semantic annotation with five labels using SML with 64 mixture components and the full DCT space.

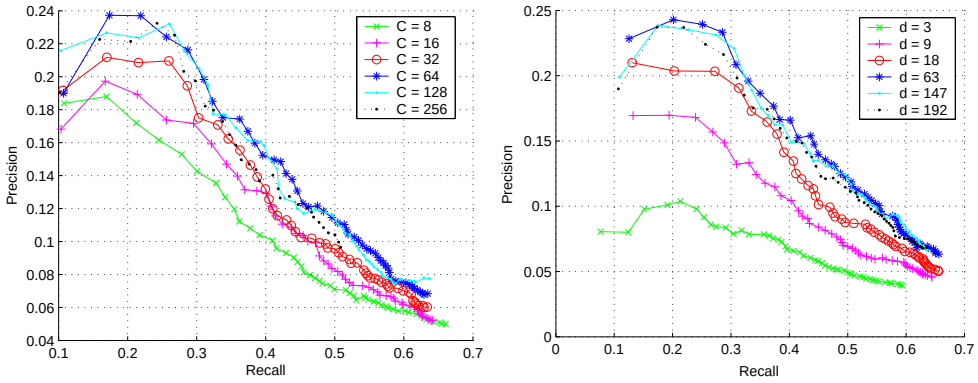



Figure 4: Precision-recall curves for annotation on Corel5K using SML while varying: (left) the number of mixture components ($C$); (right) the dimension of the DCT feature space ($d$) for 64 mixture components.

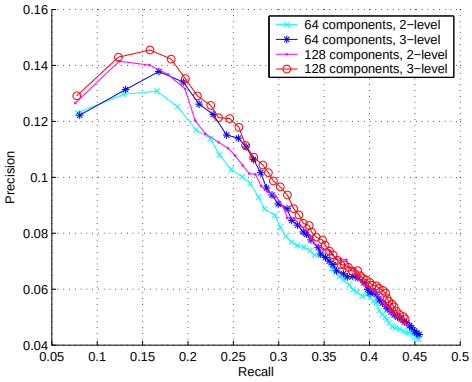



Figure 5: Precision-recall curves for annotation on Corel30K using SML-GMM-DCT.

Figure 6 shows several examples of SML on Corel30K. The automatic annotations are plausible, even though not perfectly matching the human ground-truth.

### 4.3 Ranked retrieval results

Figure 7 presents results of ranked retrieval on Corel5K for different numbers of mixture components and DCT dimensions. The left plot depicts the MAP for all 260 words, while the one in the center shows the same curves for words with non-zero recall. In both cases, the MAP increases with the number of mixture components, stabilizing above 128 components. The plot on the right shows the number of words with non-zero recall, which decreases with the number of mixture components, once again stabilizing above 128 components. Table 4 presents retrieval results obtained with SML on the the three databases.

## 5 Conclusions

In this work we have presented a comparison of the SML and SCBL models for semantic annotation and retrieval on very large databases. Various conclusions can be taken from our experiments. First, the GMM-DCT classifier outperforms the 2D-MHMM classifier of [14] in both image categorization and SCBL image annotation. When using GMM-DCT, SML has better overall precision than SCBL, although the latter may have some values of precision and recall that are better when the test image is annotated with a substantial number of words. This is due to the noisy ground-truth labels, and label coupling within image categories. Second, the performance of SML is robust to changes in the dimension of the feature space, and number of mixture components. Finally, SML seems to be scalable to large databases, where annotation performance displays similar characteristics to that of smaller databases. For large databases, using a 3-level hierarchical method for learning the class mixtures can improve the generalization of the system, as compared to the standard 2-level hierarchy. The differences are however not staggering, confirming the robustness of SML to its parameter settings.

## Acknowledgments

The authors would like to thank Kobus Barnard for providing the Corel dataset used in [11], David Forsyth for providing the Corel30K dataset, James Wang for the PSU dataset used in [14], Google Inc. for providing the computer resources for the experiments, and Gustavo Carneiro for helpful conversations. This research was partially supported by NSF CAREER award IIS-0448609 and a grant from Google Inc.

## References

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain. Content-based image retrieval: the end of the early years. *PAMI*, 22(12):1349-80, 2000.

[2] R. Picard. Digital libraries: meeting place for high-level and low-level vision. In *Proc. Asian Conf. on Computer Vision*, Dec. 1995.

[3] M. Szummer and R. Picard. Indoor-outdoor image classification. In *Workshop in Content-based access to image and video databases*, 1998.

| | Semantic Annotation | | | | | Ranked Retrieval | |
|---|---|---|---|---|---|---|---|
| Database | Precision | Recall | Precision NZR | Recall NZR | # of NZR words | MAP | MAP NZR |
| Corel5K | 0.2163 | 0.2871 | 0.4499 | 0.5972 | 125 | 0.3023 | 0.6288 |
| Corel30K | 0.1261 | 0.2122 | 0.2826 | 0.4756 | 424 | 0.2113 | 0.4734 |
| PSU | 0.1519 | 0.3158 | 0.1626 | 0.3379 | 413 | 0.2552 | 0.2732 |

Table 4: Results for semantic annotation using SML-GMM-DCT: average precision and recall for all words, and words with non-zero recall (NZR). Results for ranked retrieval: mean average precision (MAP) for all words, and for words with non-zero recall.

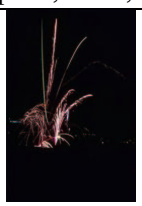
| | | | | |
|---|---|---|---|---|
| |  |  |  |  |
| Human | sky, track, rail | cowboys, horses, water, people | boats, water, people, carnival | castle, building, water, reflection |
| SML | rail, train, tracks, harbor, boats | cowboys, horses, fields, landscape, fence | coast, waves, harbor, shore, glacier | palace, mosque, sculpture, castle, arch |
| |  |  |  |  |
| Human | sailboards, people, water, waves | musician, people | man, donkey, people, building | fireworks, night, sky, burst |
| SML | sailboard, sea, waves, shore, canoe | plants, market, model, costumes, costume | castle, tower, sculpture, landmark, statue | burst, fireworks, quartz, gem, lights |
| |  |  |  |  |
| Human | valley, grass, sky | monkey, leaf, zoo | food, meal, cuisine | butterfly, side, leaves |
| SML | desert, canyon, pyramid, valley, formation | model, zoo, branch monkey, rainbow | cuisine, dessert, meal, snack, sunset | butterfly, wings, fungus, mushroom, branch |

Figure 6: Examples of semantic annotation using SML-GMM-DCT on the Corel30K database.

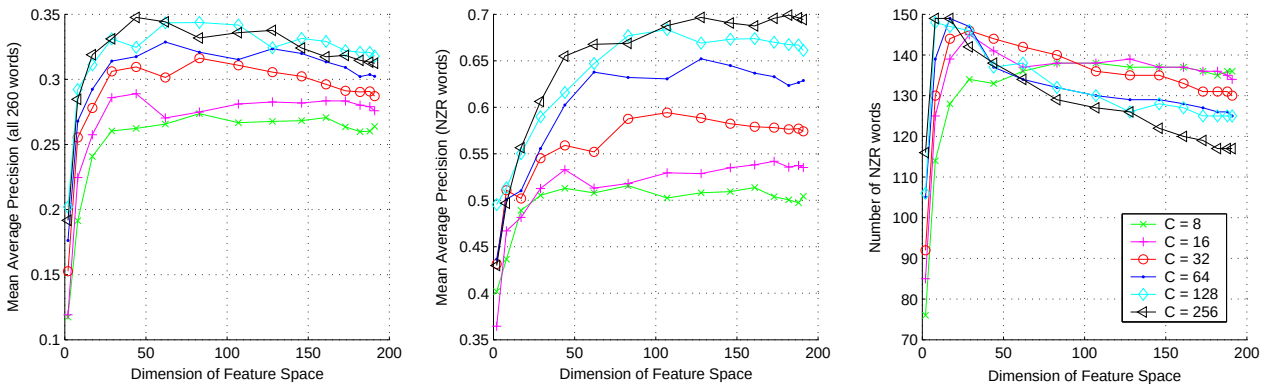



Figure 7: Ranked retrieval on Corel 5k using SML with different mixture components ($C$): (left) MAP for all the words; (middle) MAP for words with non-zero recall; (right) number of words with non-zero recall.

[4] A. Vailaya, A. Jain, and H. J. Zhang. On image classification: city vs. landscape. *Pattern Recognition*, vol. 31, pp. 1921-36, 1998.

[5] N. Haering, Z. Myles, and N. Lobo. Locating deciduous trees. In *Workshop in Content-based Access to Image and Video Libraries*, pp. 18-25, 1997.

[6] D. Forsyth and M. Fleck. Body plans. In *Proc. CVPR*, 1997.

[7] Y. Li and L. Shaprio. Consistent line clusters for building recognition in CBIR. In *Proc. ICPR*, vol. 3, pp. 952-956, 2002.

[8] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 3:1107-1135, 2003.

[9] D. Blei and M. Jordan. Modeling annotated data. In *Proc. ACM SIGIR*, 2003.

[10] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *Proc ECCV*, 2004.

[11] P. Duygulu, K. Barnard, N. Freitas, and D. Forsyth. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *Proc. ECCV*, 2002.

[12] S. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli relevance models for image and video annotation. In *Proc. CVPR*, 2004.

[13] P. Carbonetto, H. Kueck, and N. Freitas. A constrained semi-supervised learning approach to data association. In *Proc. ECCV*, 2004.

[14] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical approach. *PAMI*, 2003.

[15] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.

[16] G. Carneiro and N. Vasconcelos. A database centric view of semantic image annotation and retrieval. In *Proc. ACM SIGIR*, 2005.

[17] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *Proc. CVPR*, 2005.

[18] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, 2001.

[19] N. Vasconcelos. Image indexing with mixture hierarchies. In *Proc. CVPR*, 2001.