Text-Based Approaches for the Categorization of Images

Carl L. Sable and Vasileios Hatzivassiloglou

Department of Computer Science, 450 Computer Science Building, Columbia University, 1214 Amsterdam Avenue, New York, N.Y. 10027, U.S.A.
{sable,vh}@cs.columbia.edu

Abstract. The rapid expansion of multimedia digital collections brings to the fore the need for classifying not only text documents but their embedded non-textual parts as well. We propose a model for basing classification of multimedia on broad, non-topical features, and show how information on targeted nearby pieces of text can be used to effectively classify photographs on a first such feature, distinguishing between indoor and outdoor images. We examine several variations to a TF*IDF-based approach for this task, empirically analyze their effects, and evaluate our system on a large collection of images from current news newsgroups. In addition, we investigate alternative classification and evaluation methods, and the effect that a secondary feature can have on indoor/outdoor classification. We obtain a classification accuracy of 82%, a number that clearly outperforms baseline estimates and competing image-based approaches and nears the accuracy of humans who perform the same task with access to comparable information.

1 Introduction

As digital collections on the World Wide Web, corporate intranets, and CD-ROMs increase vastly in size and availability, it is becomming increasingly important to find efficient methods of categorizing not only text documents but also images, video, sound files, and other multimedia embedded within a document. Work in information retrieval has focused primarily on text, and then on classifying an entire document as relevant to a particular query or as a member of a specific class. Yet, much is to be gained by independently categorizing and indexing pieces of a document from different media; multimedia information arguably follows a different classification hierarchy than text, and more factors than topical relevance come into play when an image or other non-text data is included within a document. For example, a news article on the recent events in Kosovo may include a picture of an airplane at a U.S. base, even though that particular aircraft never participated in the operations described in the article. The same image can frequently be found in multiple related documents, and, conversely, an independent classifier of images could help select an image from a broad, separate collection for illustrating a summary of a text-only source. Undesirable images (e.g., advertisements) could also be detected and pruned before a document is displayed to the user.

S. Abiteboul, A.-M. Vercoustre (Eds.): ECDL '99, LNCS 1696, pp. 19–38, 1999.

In the present work, we explore such an independent classification for images, using information from associated text sources such as captions and the surrounding text in the document in which the image is embedded. We are informed and motivated in this endeavor by the parallel development of a multimedia, multiple document summarizer (Aho *et al.* 1998), where appropriate images can enhance the text summary. Our approach centers on the development of a suitable class hierarchy of broadly applicable visual features that will facilitate the selection of appropriate images for such summaries, even when fine distinctions (such as the subject matter of the image) are not available. Such features include classifying the images as indoor or outdoor; as containing one or a few persons or a crowd or no people at all; and as depicting a natural landscape versus a city scene. If independent classifiers can be designed for these features, then we can infer the appropriateness of the image for a particular descriptive purpose with high likelihood given only a little domain guidance. For example, an outdoor image with no people from the terrorism news domain is likely to show the scene of an event or its aftermath, while an indoor photograph with a crowd of people probably refers to a related press conference. Additional techniques can refine these inferences, by using for example information extraction methods (Wacholder *et al.* 1997) to identify the location of an event or the names of specific participants in the images.

We report in this paper on our methods and results for classifying images as indoor versus outdoor. We chose this visual feature as a basis for a first division of the images because of its plausibility as an indicator of image content and because it is used as a high-level feature in image ontologies for image and digital signal processing (Vailaya *et al.* 1999). It is also a feature for which purely visual classifiers can be built (Szummer and Picard 1998); in fact, we are developing such classifiers in parallel with the text-based ones described here, and we plan to investigate ways to integrate them in the future. Although we have focused on this category, the methods described in the paper are independent of the specific feature and can be applied to any of the broad categories identified earlier.[1]

Our indoor/outdoor classifier for images is based on information retrieval measures of text similarity, such as term frequency and inverse document frequency (TF*IDF) (Salton and Buckley 1988; Salton 1989). Unlike information retrieval, however, we have to work with small pieces of text (a caption or a portion of a caption). Hence, we examined and evaluated several potential improvements to standard IR techniques, such as using targeted parts of the available text, limiting ourselves to particular word classes, and partially disambiguating words according to their part of speech. We collected a large sample of 1,675 images for training and evaluation, and had multiple human volunteers assign indoor/outdoor labels to them. We measured individual human performance on this task against the standard implied by their agreement, and compared our system's performance to the humans, a default baseline classifier, and image-based classifiers that operate on purely visual features (e.g., color, texture, and edge

[1] The main cost for moving on to new categories involves the necessary manual labeling of a large set of images for training and evaluation.

direction features). We optimized our classifier using three-fold cross-validation, varying several of the TF*IDF parameters and optional features and determining which of the features have a major effect on performance. Using probability density estimates for the output of the classifier, we are able to correct several potential misclassification errors. Our results show that the automatic system clearly outperforms the baseline and image-based classifiers, approaching the accuracy of the human volunteers. We extend these results by considering alternative evaluation methods and the effects of lenient versus strict definitions of the indoor and outdoor categories. We also explore another method for identifying words that discriminate between the two categories and measure the effect of additional high-level features (in this case, the number of people in each image) for indoor/outdoor classification.

2 Related Work

Our classification approach draws on a long line of work for measuring text similarity, mostly in an information retrieval context. Most of the information retrieval approaches rely on single words (e.g., (Salton and Buckley 1988; Salton 1989)), although sometimes compounds and collocations have been used (Smeaton 1992). Some of the features we explore (e.g., ignoring capitalization) are also used by default in most IR systems. Other, more natural language-informed features have found mixed success in information retrieval (e.g., (Salton and Smith 1989; Gay and Croft 1990; Smeaton 1992)), although the usefulness of each feature needs to be evaluated separately for each application (classifying image captions is different than classifying entire documents).

For topical image classification, keywords extracted from a document have been used to index an associated image (Bach *et al.* 1996; Smith and Chang 1997), and image similarity has been measured on the basis of shared image features (Niblack *et al.* 1993; Pentland *et al.* 1994) and by a combination of textual and image feature matches (Ogle and Stonebraker 1995; Smith and Chang 1997). Rowe and Guglielmo (1993) and Smeaton and Quigley (1996) use information from captions for retrieving (rather than classifying) images given a query. Srihari (1995) uses face detection techniques along with name extraction from captions for linking images to specific people. Classification of images along broad, non-topical features such as those we are exploring has received less attention in the image processing literature, although this is beginning to change. Forsyth and Fleck (1996) present an image-based detector for naked people, while Szummer and Picard (1998) describe an approach for separating consumer photographs into indoor and outdoor classes. Both of these approaches utilize as their input only low-level visual features, such as color and edge direction.

3 Data Set

Our raw data set consists of 21,086 news postings from April 1997 to May 1998 from a variety of Usenet current news newsgroups. Of these, 1,490 contain, in

addition to a text article, one or more embedded images, each with an associated caption. Captions are generally two to four sentences long. The first sentence in the caption tends to describe the image, while the remainder usually gives background information and establishes the relevance of the image to the story. For example,

> BANGKOK, THAILAND, 9-NOV-1997: New Thai Prime Minister Chuan Leekpai gives a traditional "wai" to thank members of his party appplauding his entrance, November 9, during a ceremony appointing him as the country's 23rd prime minister in Bangkok, Thailand. Chuan was named prime minister for the second time, replacing Chavalit Yongchaiyudh at the helm of a country plagued by economic woes.

For training and testing, a web-based interface was set up allowing volunteers to label images according to two high-level features. The first feature corresponded to the indoor versus outdoor dichotomy, and the choices given were *Indoor*, *Outdoor*, *Likely Indoor*, *Likely Outdoor*, and *Ambiguous*. The second feature was number of people, and the available choices were *No People*, *One Person*, *Two People*, *Three or More People*, *Crowd*, and *Ambiguous*. In both cases, the authors went over a sample of images in advance, identified potential problems, and supplied the evaluators with detailed instructions which can be viewed at `http://www.cs.columbia.edu/~sable/research/readme.html`.

Using our interface, fourteen volunteers labeled the images under different access conditions: by viewing the image alone, the caption alone, both the image and the caption, or just the first sentence of the caption. Each image received two such labels under the full access condition (when volunteers viewed both the image and caption), which we consider representative of normal use of the images in multimedia documents.[2] We use the labels obtained for this condition as the basis for both our training and testing sets. A single label was obtained for each image under the other conditions; these are used to estimate human performance and to compare with our system (which uses only text information).

For the indoor versus outdoor distinction, analysis of the assigned labels reveals that in most cases (87.7%), a definite indoor or outdoor judgement was made, and only 3% of labels assigned were "Ambiguous". Agreement between humans was also high (90.4% of the images had compatible labels, although sometimes with different degrees of confidence). There was, however, some disagreement between human categorizers. 137 images had labels that differed by more than one step on the scale from "definite indoor" to "definite outdoor", and 39 of them had in fact one "definite indoor" and one "definite outdoor" label. Our analysis of the labels for the number of people feature indicated a somewhat lower but still significant level of agreement (80.4%). Inspection of the images that received conflicting labels reveals that several of the disagreements are due

[2] Four volunteers labeled the images under this specific access condition. One provided labels for all images, while each of the other three provided a second label for a third of the images. Thus, each of the 1,675 images received exactly two labels under this condition.

Fig. 1. An image that is hard to classify as indoor or outdoor.

to mistakes by the categorizers, but in some cases, even markedly different labels can be attributed to different opinions about how terms like "indoor" and "outdoor" should be defined. For example, close-ups of people within a vehicle such as a car or a plane, or pictures of people under a roof of a structure with no walls, were often labeled differently by different judges. Fig. 1 shows one of the images that reasonable people could disagree on; more can be inspected at `http://www.cs.columbia.edu/~sable/unusual.html`.

We have compiled four different sets of images according to these manual categorizations. First, we consider the images for which both evaluators provide a definite judgement in the same direction on the indoor versus outdoor question. This set contains 1,339 images (79.9% of the original 1,675) and is the primary focus of our experiments. 401 (29.9%) of these images were classified as indoor while 938 (70.1%) were classified as outdoor.

Our second experimental data set relaxes the requirement of strong beliefs from each evaluator. It consists of those images that received two judgements in the same direction on the indoor versus outdoor question, regardless of the reported degrees of confidence. This set includes 1,501 images (89.6% of the original 1,675). 475 (31.6%) of them are classified as indoor while 1,026 (68.4%) are classified as outdoor.

Turning to the number of people question, we define a third set, consisting of the images that received identical (non-ambiguous) judgements from both evaluators on that question. This set includes 1,346 images (80.4% of the total), further divided as 88 (6.5%) with no people, 304 (22.6%) with one person, 213 (15.8%) with two people, 609 (45.2%) with three or more people, and 132 (9.8%) with crowds. We also define a fourth experimental set for studying the interaction between the indoor/outdoor and number of people categories, as the intersection of the first and third sets described above. This last set contains 1,081 images (64.5% of the total).

4 Measuring Similarity for Indoor/Outdoor Classification

We base our classification of images into indoor or outdoor classes on a measure of similarity between each *document* we examine and the two *category prototypes* that correspond to the two classes. The term *document* is used above with a general sense, standing for any piece of text that is associated with the image under consideration; in many of our experimental runs, this is much smaller than the entire article that contains the image.

For a single piece of text, a word's TF, or *term frequency*, is the number of times that this word occurs in that text. For a category (such as all indoor images), the TF assigned to a word is the number of times that word occurs in all documents of that category. A word's IDF, or *inverse document frequency*, is the logarithm of the ratio of the total number of documents to the number of documents that contain that word; this measure remains constant independently of the particular document or category examined. The product TF*IDF,

$$TFIDF(word) = TF(word) \times \log \frac{\text{Total number of documents}}{DF(word)} \tag{1}$$

is therefore highest when a word contains a balance of high frequency within a document or category (signifying high importance to the document or category) and low overall dispersion within the collection (signifying high specificity).

Every document and category is represented by a vector of TF*IDF values, with each dimension corresponding to a word. By abstracting content in this manner, word vectors of documents and categories can be compared to determine how well a document fits in each category. We use the inner product between document and category vectors, i.e.,

$$Score(document, category) = \sum_i TFIDF_{document}[i] \times TFIDF_{category}[i] \tag{2}$$

as our measure of similarity. Each document is then assigned to the category for which the fit is best, i.e., for which (2) is maximized.

We varied this measure of similarity in different experimental runs by using different restrictions on what enters the TF*IDF formula (i.e., what a "word" is) and by modifying (2) with the introduction of normalizing factors. Our first set of parameters, corresponding to the definition of words, involves four choices:

- **Text span considered**. What is the text that should be associated with each image, becoming the "document" in the TF*IDF calculations above? We have experimented by using the entire article, the article without the image caption, just the caption, or only the first sentence of the caption. While the articles are longer and provide more information about the related story than the caption, they are less related to the specific image, and therefore may contain too much noise to be helpful for the type of categorization we are performing. Hence, we can trade some information of questionable quality for increased specificity by limiting ourselves to the caption only. Similarly, the first sentence of the caption tends to be more descriptive of the image than the rest, which often provides background information.

- **Resctriction to specific grammatical categories**. Should all the words in the selected text span be included in the TF*IDF computations? Open-class words (adjectives, nouns, verbs, and adverbs) carry in general most of the content information, while words such as numbers and pronouns do not usually affect an image's classification. We used a statistical part-of-speech (POS) tagger (Church 1988) to automatically assign a grammatical category tag to each word, and then experimented with using all words, only open-class words and prepositions (because of the nature of the indoor/outdoor distinction), and open-class words and prepositions with proper nouns excluded.
- **Disambiguation of words**. A word's sense is frequently ambiguous, and sometimes knowing its grammatical part-of-speech can help disambiguate it. For example, *can* is most often an auxiliary verb, but sometimes a noun with a different meaning. We experimented with keeping the POS tag as part of the word (thus distinguishing between the two senses of *can/verb* and *can/noun* above), versus ignoring this information.
- **Case sensitivity**. Should capitalization matter for treating words as different? Capitalization may indicate a proper noun, but may also be the result of sentence-initial placement. We experimented with collapsing words that differ only in capitalization to the same token versus treating words as different if they differ in case.

Each combination of the above parameters results in a different set of TF*IDF vectors for each document. Three more parameters were varied when calculating the similarity between a document and a category:

- **Ignoring words with low TF*IDF during similarity computations**. We have experimented with optionally ignoring words whose TF*IDF values within a document fall below a given constant, for several alternative values of that constant. This eliminates relatively insignificant words, which have minimal impact on the classification, while potentially speeding up the necessary calculations and avoiding some rare words whose TF and IDF is hard to estimate accurately.
- **Normalization of category vectors**. The size of each of the two classes does not enter (2) or the TF*IDF calculations. Yet, it is natural to expect that the a priori most frequent category will have higher TF values, simply because it contains more documents. This is a concern for our experiments, since the "outdoor" category contains more than two thirds of the images in our collection. We therefore experiment with a modification to (2), where the TF*IDF value of each word in a category vector is divided by the total number of documents that fall into that category. This modification, which replaces total frequency with average per document frequency, makes the TF*IDF values directly comparable across categories.
- **Density estimation**. The standard approach for assigning documents to categories is to select the category for which similarity is largest. This, however, implicity assumes that the similarity scores are on the same scale for both categories, and makes it hard to tell when a difference between the

similarity scores for the two categories is large enough for the system to be confident in its decision. We experimented with a modification of the category decision rule by transforming the difference of the raw similarity scores between the two categories into the corresponding probability that a document with the given score difference belongs in the indoor category. In other words, we empirically estimate the probability density of the composite random variable $Score(document, indoor) - Score(document, outdoor)$. We calculate the histogram of this difference function from the training part of the data (see the next section), and then use a rectangular smoothing window on top of the histogram to estimate the probability density (Scott 1992). For a new image in the test set, we again compute the difference and apply the conversion procedure that was fixed during training. The resulting probability is more directly interpretable than the difference of the raw similarity scores, automatically adjusts the cut-off point between the two categories (from the arbitrary 0 on the unrestricted difference scale to the now well-justified 0.5 on the 0 to 1 probability scale), and provides a measure of confidence in the system's decision (values near 0 or 1 indicate higher confidence) that can be easily combined with information from other independent categorizers.

5 Results and Evaluation

We randomly selected 894 (approximately two thirds) of the 1,339 images that had definite human agreement on the indoor versus outdoor classification question for training, and the remaining 445 images for testing. 276 (30.9%) of the training images were indoor while 618 (69.1%) were outdoor. 125 (28.1%) of the testing images were indoor while 320 (71.9%) were outdoor. So, on that particular breakdown of our main experimental image set, a default classifier would achieve 71.9% accuracy on the test set by labeling every image with the more frequent category in the training set.

Using this training/testing partition, we calculated the TF*IDF vectors and similarity scores described in the previous section for each of the 768 possible combinations of parameters, performing a complete designed experiment (Hicks 1982). The training set was randomly divided into three equal parts, and for each such experiment, we repeatedly trained on two parts and measured system performance on the third. This three-fold cross validation on the training set gives us the ability to compare the relative performance of the various settings for the experimental parameters. It also allows us to select the best combination of parameters, which is fixed for subsequent experiments, and in particular for scoring against the completely unseen test set.

We found a wide variety in the obtained average accuracy score (percentage of correct categorizations) depending on the parameter settings. The parameters which had the most major effect were:

- **Text Span**. Restricting analysis to the first sentences of captions accounted for the 37 top scoring experiments. First sentences clearly outperformed captions, while text spans that included the entire article (with or without the

caption) were far behind. This provides support to our thesis that specifically selected and narrowly targeted pieces of text can be more useful for classifying embedded multimedia information than the document as a whole.

- **Restriction to specific grammatical categories**. Using only open-class words plus prepositions accounted for 4 of the top 5 experiments. The average accuracy over all experiments for this setting was also higher that that for using all parts of speech, which, in turn, was higher than that using open-class words plus prepositions but excluding proper nouns. So it appears that proper nouns help in this classification task, a somewhat counter-intuitive result, especially since we generally have a high number of low-frequency proper nouns.
- **Normalization of category vectors**. Normalizing category vectors accounted for 12 of the top 15 experiments, and had a higher average accuracy among all experiments, even more so for cases where density estimates were used.
- **Density estimation**. Using probability densities instead of raw similarity scores improved performance in almost every case, including all combinations of parameters ranked near the top. This optional component had one of the most pronounced effects in overall system performance.

On the other hand, ignoring words with low TF*IDF, keeping the part of speech information for disambiguation, and ignoring case differences played much smaller roles. High thresholds for including words in the TF*IDF vector were clearly bad, but other than that, all setting of these parameters were used in some of the best experiments, and the average accuracy for each were similar. Table 1 summarizes the effect of each value of each parameter over all experiments, while Table 2 shows the top fifteen combinations of parameters (those which achieved over 82.5% accuracy) in terms of performance during the three-fold cross validation on the training set. The average cross-validated accuracy of all 384 experiments that directly use the TF*IDF scores was 71.74%, and of the 384 experiments that include the probability conversions, 74.26%. Note that these overall accuracies are close to the baseline of the default classifier (71.9%), while 31 of the 768 combinations of parameters performed better than 82% during cross validation. This indicates that an informed choice of the parameters is important for this classification task.

On the basis of these cross-validation experiments, we selected the following combination of parameters for our system: using the first sentences of captions only; restricting words to those of an open class plus prepositions; treating words that differ only in part of speech as identical; keeping capitalization information; not applying any thresholds for including words in the TF*IDF vector; normalizing according to category size; and applying the density transformation. These are the parameters that were used in the experiment represented by the first line of Table 2, which was one of two that tied for the best results during cross-validation. With these parameters fixed, we retrained on the full training set and tested on the unseen test set. The corresponding categorizer achieved

Table 1. Average overall accuracy during cross-validation of all experiments with the given value of each parameter.

Parameter	Value	Average Accuracy
Text Span	first sentences of captions	79.45%
	captions	76.06%
	articles (including captions)	69.22%
	articles (excluding captions)	67.26%
Part of speech restriction	open-class and prepositions	73.54%
	all words	73.09%
	open-class and prepositions, exluding proper nouns	72.36%
Keeping tags for disambiguation	yes	73.08%
	no	72.91%
Case sensitivity	yes	73.01%
	no	72.99%
Threshold on TF*IDF	medium	73.63%
	low	73.57%
	none	73.21%
	high	71.57%
Normalization according to category size	yes	73.36%
	no	72.64%
Using probability density estimates	yes	74.26%
	no	71.74%

on the test set 82.02% accuracy, and 90.72% on the training set.[3] If the density estimate transformation were not employed, the accuracy on the tests set falls dramatically to 72.36%. Tables 3 and 4 are contingency tables further breaking down these accuracy scores on a per category basis, separately for the cases where the density adjustments are used or not. Note that the use of probability densities tends to shift the system's categorizations from the smaller category to the larger category. Therefore, the smaller category winds up having a higher precision and lower recall, while the larger category ends with a lower precision and higher recall. Detailed results on our 445 individual test images can be observed at `http://www.cs.columbia.edu/~sable/research/demo_results/demo_results.cgi`.

Naturally, we want to compare these results with alternative classifiers, including humans. Our accuracy on the test set (82.02%) clearly surpasses that of the default classifier which always selects the "outdoor" label for every image (71.9%). We estimate human performance on this task by measuring the percentage of correct classifications achieved by a human volunteer who looked only at the captions of the images (i.e., who had access to the same kind of information that our system does). Of the 1,339 images in our main set, 1,172 (87.52%)

[3] An indoor output probability of more than 50% is translated to a decision in favor of the indoor category during this evaluation.

Table 2. Top fifteen combinations of TF*IDF experiment parameters after three-fold cross validation on the training set. The "tags" column indicates whether tags were kept for disambiguating words; the "case" column indicates whether word comparisons were case sensitive; and the "norm." column indicates whether the normalization for category size was applied during the similarity calculations.

Text span	Part of speech restriction	Tags	Case	Threshold on TF*IDF	Norm.	Accuracy without densities	Accuracy with densities
first sentences of captions	open-class plus prepositions	no	yes	none	yes	75.06%	83.22%
first sentences of captions	open-class plus prepositions	no	yes	low	yes	75.06%	83.22%
first sentences of captions	all words	yes	no	medium	yes	78.08%	82.89%
first sentences of captions	open-class plus prepositions	no	no	low	yes	74.83%	82.89%
first sentences of captions	open-class plus prepositions	no	no	none	yes	74.61%	82.89%
first sentences of captions	all words	no	no	medium	yes	79.08%	82.77%
first sentences of captions	open-class plus prepositions	no	yes	none	no	78.75%	82.77%
first sentences of captions	all words	yes	no	medium	no	78.97%	82.66%
first sentences of captions	all words	no	yes	low	yes	77.29%	82.66%
first sentences of captions	all words	no	no	low	yes	76.73%	82.66%
first sentences of captions	open-class plus prepositions	yes	no	low	yes	75.17%	82.66%
first sentences of captions	open-class plus prepositions	no	yes	medium	no	81.99%	82.55%
first sentences of captions	all words	no	yes	none	yes	77.40%	82.55%
first sentences of captions	all words	no	no	none	yes	77.07%	82.55%
first sentences of captions	all words	yes	no	none	yes	76.96%	82.55%

were correctly categorized under this access condition.[4] This figure can serve as a reasonable, approximate upper bound for how well we might hope our system to perform given only text information.

Recently, an image-based approach for classifying photographs as indoor or outdoor has been proposed (Szummer and Picard 1998). This approach is based on a decomposition of the image by applying a 4×4 grid on it and taking

[4] For this purpose, any categorization in the right direction (i.e., indoor or outdoor), regardless of the degree of confidence, was considered correct while assignments of the "Ambiguous" label received half credit.

Table 3. Contingency table showing the breakdown of the system's categorizations on the test set with conversions to probability densities.

	Actual Indoor	Actual Outdoor	Precision
System Indoor	75	30	71.43%
System Outdoor	50	290	85.29%
Recall	60.00%	90.63%	

Table 4. Contingency table showing the breakdown of the system's categorizations on the test set using the raw similarity scores.

	Actual Indoor	Actual Outdoor	Precision
System Indoor	106	104	40.48%
System Outdoor	19	216	91.91%
Recall	84.80%	67.50%	

measures of low-level image features such as color and texture on each of the 16 image regions. Then, similarities between blocks in a given image and blocks in known indoor and outdoor images are calculated, and the image is assigned to one of the two categories. In cooperation with image processing researchers at Columbia,[5] we reimplemented this technique and measured its performance on our collection of photographs. We found that its accuracy on our test set was 74%, significantly less than what we obtain with our text-based methods. We also added supplemental low-level features, such as edge direction histograms, to those used by Szummer and Picard, and a machine learning component for estimating classification thresholds. The resulting classifier (Paek *et al.* 1999) achieves 76% performance, still less than the method described in this paper.

For each of the above comparisons, we calculated a level of significance by applying Pearson's chi-square test (Fleiss 1981) on the contingency table that represents the cross-classification of the answers of the two compared methods.[6] We observe that the difference between the performance of our system and either the default baseline, Szummer's and Picard's image-based classifier, or (regrettably) the human judges, is strongly significant at the 1% level or less; the probability that similar or more pronounced differences in the observed accuracy rates between the compared methods would be observed by chance is 0.046%, 0.464%, and 0.460%, respectively. When comparing our system to our enhanced image-based model (Paek *et al.* 1999), the difference is still significant at the 5% level (P-value of 3.24%).

[5] Seungyup Paek, Alejandro Jaimes, and Shih-Fu Chang, of the Department of Electrical Engineering, Columbia University.

[6] The large-sample assumption of the chi-square test is satisfied for these contingency tables. Because we test on several hundreds of images, the exact Fisher test (Fisher 1934) is computationally impractical.

Table 5. System accuracy stratified according to high, medium, or low confidence.

Confidence Level	Number Correct	Number Incorrect	Accuracy
$p \geq 0.9$ or $p \leq 0.1$	234	21	91.76%
$0.7 \leq p < 0.9$ or $0.1 < p \leq 0.3$	89	32	73.55%
$0.3 < p < 0.7$	42	27	60.87%
Total	365	80	82.02%

A final evaluation question is how reliable the confidence estimates provided by our system's output probabilities are. Preferably, decisions with a high degree of confidence should be more likely to be accurate than decisions given a low degree of confidence. We have therefore broken down the test set into three subsets according to the probability assigned by our system, p, that a given image is indoor. These three ranges of p were defined as high confidence ($p \geq 0.9$ or $p \leq 0.1$), medium confidence ($0.7 \leq p < 0.9$ or $0.1 < p \leq 0.3$), and low confidence ($0.3 < p < 0.7$). Note that the indoor probability equals 1 minus the outdoor probability, with the classifier selecting the indoor category when $p > 0.5$ and the outdoor category otherwise; hence, probabilities of p and $1-p$ are equivalent in terms of the expressed confidence. Table 5 shows the accuracy of our system within each confidence category, and verifies that decisions given a higher level of confidence are more likely to be correct, thus validating our confidence estimates. In particular, 255 (57.3%) of the 445 test images were labeled with over 90% confidence, and 91.76% of these categorizations were correct.

6 Identifying Words with High Discriminating Power

Methodology. A second approach to the classification problem is to automatically locate words (or multi-word phrases) whose presence strongly indicates one of the competing classes. We explore this technique by first extracting all open-class words plus prepositions from the first sentences of captions. We exclude proper nouns from this analysis since they are unlikely to be general indicators of one of the categories, and only consider words occurring five times or more in our training set. This last step is done to ensure that the words we keep will be frequent enough to be general discriminators, and to avoid cases where a particular word occurs in a few captions of images from a particular class simply by chance.

We construct a *log-linear regression model* (Santner and Duffy 1989) using binary variables corresponding to the occurrence of each of these words as predictors and the output feature (e.g., indoor or outdoor image) as the response. The model is fitted with *iterative reweighted least squares* (Bates and Watts 1988), and the fit assigns a weight to each of the candidate discriminators. Words with higher weights are those that actually help discriminate between the two classes.

As an alternative machine learning technique, we also consider *decision trees* (Quinlan 1986). The prediction model remains the same, but now the tree is constructed with *recursive partitioning*, with the most discriminating variable being selected first. The resulting tree is *shrunk* (Hastie and Pregibon 1990) (node probabilities are optimally regressed to their parents) to reduce the possibility of overfitting; we select the shrinking parameter α through cross-validation within the training set.

Results. Using the same training/test set division as with the TF*IDF experiments reported in the previous section, our list of candidate discriminators contains 665 words. Both the log-linear regression model and the tree select a subset of these words as classification features; in the case of the selected tree, 80 words are used during classification.

It is interesting to note which these words are, especially since the results of this procedure are likely to generalize to other sets of images. The five words most favoring an indoor classification are **conference**, **meeting**, **meets**, **hands** (plural noun), and **L**, while the five words most strongly indicating an outdoor image are **of**, **from**, **soldiers**, **police**, and **demonstration**. Some of them are expected (e.g., *demonstration* or *police* for an outdoor image, or *conference* for an indoor one), but some come as a surprise, for example, the "words" *C*, *L*, and *R* (indicating an indoor image) used in parentheses to identify people in images by position (i.e., center, left, or right).

Overall performance of the word discriminant method was 93.62% over the training set and 78.65% over the test set.

Integrating the two classifiers. The two classifiers discussed in the present and the previous section utilize different approaches to arrive at similar classification performance. Hence, it is natural to investigate how correlated their answers are, and whether a combined classifier might improve overall performance on the indoor/outdoor classification task.

We have built such combined classifiers using both general machine learning techniques discussed above (log-linear models and decision trees). However, the overall performance of the composite classifiers was in both cases slightly less than that obtained by the best individual classifier (82.02%). We attribute this to overtraining during the construction of the composite classifiers, especially since the same training set was used for building each of them and for combining them.[7] Nevertheless, our implementation of two classification methods provides us with two different general tools that can be easily ported to other high-level classification tasks; and the ability to identify key discriminating words may prove helpful in future exploration of what makes images in distinct categories different.

[7] A further subdivision of our image data in two separate training sets and a test set would leave us with too few images in each set.

7 An Alternative Evaluation Metric

So far, all reported accuracies considered the system to be completely right if the category with the higher probability was correct and completely wrong otherwise. An alternative evaluation method is to take the probability assigned by the system to the correct category and consider that to be the system's score for that document, in a manner similar to the partial credits proposed in (Hatzivassiloglou and McKeown 1993). For example, let's say that the system analyzes an image and says the probability that the image is indoor is 65% (meaning that the probability that the image is outdoor is 35%). If the image is actually indoor, the system is given a score of 0.65 for this image, while if the image is actually outdoor, the system is given a score of 0.35. The overall accuracy of the system is then the sum of the system's scores for all images divided by the total number of images. In the ideal case, the system would assign all indoor images a probability of 1 of being indoor, and all outdoor images a probability of 0 of being indoor. Thus its total overall accuracy would be 1, or 100%. Indeed, if the system always has complete confidence in its decisions, the revised evaluation method becomes equivalent to the standard one.

In this way, the system receives partial credit for each answer, more if the system leans in the correct direction and directly increasing as the system's confidence in a correct decision increases. In general, when a system already classifies most images correctly under the original 0/1 scoring method, it will tend to be penalized for its uncertainty on correct decisions more than it is credited for uncertain wrong answers. This is the case in our task when our classifier is evaluated on our main set of images (those with definite agreement by the human volunteers); the system achieves 82.02% accuracy under the original evaluation method, and 76.71% under the revised one. However, we consider this modified method as more revealing, as it offers a way to evaluate the system's confidence in its decisions.

To further illustrate this alternative evaluation technique, and also the generality of our parameter selection mechanism, we repeated our training of the indoor/outdoor classifier on our second set of images, those that had any kind of agreement from the human judges (not necessarily with strong beliefs; see Sect. 3). We randomly selected 1,000 (approximately two thirds) of the 1,501 images in that set for training and the remaining 501 images for testing, and retrained the classifier using the optimal combination of parameters determined in Sect. 5. 308 (30.8%) of the training images were indoor while 692 (69.2%) were outdoor; within the test set, 167 (33.3%) images were indoor while 334 (66.7%) were outdoor. Our system achieved on the test set 77.05% accuracy using the raw TF*IDF similarities and 80.04% after converting those to probability estimates. The latter of these results is the most important, and it is 1.98% lower than the result from the main set with definite agreement. This makes sense, since manual categorizations with a lower degree of confidence are less likely to be accurate, and also may indicate images that are inherently harder to classify. This is in fact reflected in the system's confidence measure, which tends to be lower on these problematic cases; applying the alternative evaluation method to

Table 6. Breakdown of the set of images with definite agreement on indoor/outdoor and number of people features into indoor and outdoor images for each value of the number of people feature.

Number of people	Indoor images	Outdoor images	Percentage of indoor images
No People	2	75	2.6%
One Person	122	108	53.0%
Two People	75	85	46.9%
Three or More People	155	332	31.8%
Crowd	8	119	6.3%
Total	362	719	33.5%

this second test set, we obtain overall accuracy of 76.56%, almost as high as that measure is for the first test set.

8 Using Information about Number of People

Earlier on, we noted that our goal in this line of research is to develop multiple classifiers for a number of broadly applicable classification features. It is natural to consider interactions between such classifiers, as information about one feature may well help the categorization according to another feature. In this section, we report on investigations regarding the effect knowledge about the number of people in a photograph has on our ability to classify the image as indoor or outdoor.

We have not yet built a text-based classifier for this second feature,[8] so we use instead ideal knowledge, provided by the human categorization of images according to this feature. We analyze the set of images that has both definite agreement between the human judges in the indoor/outdoor question and agreement in the number of people question (excluding ambiguous labels). This set contains 1,081 images, 362 (33.5%) of which are indoor and 719 (66.5%) are outdoor, a similar distribution as in the larger set which we used for our main experiments. However, if we take the number of people as given, the distribution of indoor versus outdoor images within each category of the secondary feature changes, sometimes dramatically, as Table 6 shows.

To utilize this information, we need a formula that connects $f(I|c,d)$, the probability density of an image being indoor given that it belongs to category c according to the number of people feature and that it receives a similarity difference of d, to our old probability density estimates, $f(I|d)$. Unfortunately, a Bayesian expansion of $f(I|c,d)$ involves the joint density $f(c,d)$, which we cannot estimate without a classifier that predicts the number of people c from the difference d (or vice versa). Therefore, we intuitively derive an approximate

[8] Although work is under way for building one based on face detection combined with name extraction from captions.

formula for $f(I|c,d)$ as follows: Given N images with similarity difference in a small neighborhood Δd around d, approximately $P(I|\Delta d) \cdot N$ of them will be indoor. Now, for any image that has a specific number of people c, its odds for being indoor will change (for better or worse) from the global proportion of indoor images $P(I)$ by the ratio $P(I|c)/P(I)$. If $P(c)$ is the global proportion of images with c people in them, the overall number of indoor images with c people among the initial N images with similarity difference close to d can be estimated as

$$N(I|c,\Delta d) \approx P(c) \cdot \frac{P(I|c)}{P(I)} \cdot P(I|\Delta d) \cdot N \tag{3}$$

Similarly, the overall number of outdoor images with c people among the same N images can be estimated as

$$N(O|c,\Delta d) \approx P(c) \cdot \frac{1-P(I|c)}{1-P(I)} \cdot (1-P(I|\Delta d)) \cdot N \tag{4}$$

By combining (3) and (4), we get our formula for updating $P(I|\Delta d)$:

$$\begin{aligned} P(I|c,\Delta d) &\approx \frac{N(I|c,\Delta d)}{N(I|c,\Delta d)+N(O|c,\Delta d)} \\ &\approx \frac{\frac{P(I|c)}{P(I)} \cdot P(I|\Delta d)}{\frac{P(I|c)}{P(I)} \cdot P(I|\Delta d) + \frac{1-P(I|c)}{1-P(I)} \cdot (1-P(I|\Delta d))} \end{aligned} \tag{5}$$

We applied this update formula to the images in the set with definite agreement on both the indoor/outdoor and number of people questions. Since that set is a subset of our main experimental image set, we took those images that were in the training set for the main set (see Sect. 5) as our training images, and the remaining as test images. The resulting training set had 732 images, of which 249 (34.0%) were indoor, and the testing set contained 349 images, of which 113 (32.4%) were indoor. If the methods of Sect. 5 are applied to this training/test set partition while ignoring the number of people information, we obtain 79.94% accuracy on the test set. If instead we assume perfect knowledge of the number of people variable and update the probability estimates by applying (5) (estimating quantitities such as $P(I)$ and $P(I|c)$ from the training set), we obtain 80.23% accuracy on the test set. This is only a minor improvement, not statistically significant. However, if the alternative evaluation metric of the previous section is employed, accuracy improves from 74.96% to 77.19%. So while few categorizations actually changed from wrong to right or vice versa, the system's confidence values in its decisions were more appropriate when the number of people was taken into account. In other words, on average, correct decisions were given higher confidence while the reverse happened to incorrect decisions.

9 Conclusions and Future Work

We have shown that our methods for categorization of images as indoor or outdoor strongly beat baseline performance and competing, image-based techniques, and even begin to approach human performance. In fact, our system provides 93.72% of the correct answers that a human judge with access to the same kind of information does (82.02% versus 87.52% overall accuracy). By staying within the TF*IDF paradigm but experimenting with several parameters and adding the use of probability density estimates, we have created a system that achieves 82% accuracy on unseen images. The output of our system is in terms of a probability, which is readily interpretable and provides a level of confidence in the system's decision. We have explored additional techniques both for image classification and for evaluating the constructed classifiers. In addition, we investigated the possibility of using additional information about images that might change a priori probabilities of an image being indoor or outdoor, and there is some promise that the system's results may be improved. Our methods are general, and could be applied to other high-level visual features, although currently our model of probability densities assumes dichotomous classifications.

We have examined a classification approach that relates to the Rocchio paradigm (Rocchio 1971) and combines TF*IDF estimates with a probabilistic normalization. A future alternative is to compare our results with pure probabilistic approaches such as naive Bayes (Lewis 1998) and connectionist models (Lewis *et al.* 1996). Certainly, we have not exhausted the space of possible features and transformations of the input data; we plan to examine additional such options, including morphological transformations/stemming, semantic information linking related words, and different weighing of identified named entities.

Our immediate next step is to integrate this text-based classifier with image-based ones that are being developed at Columbia, and expand the range of classification questions considered. We will explore high-level classifications such as indoor/outdoor, number of people, and city versus landscape, and complement the general classifiers with specific image feature detectors (e.g., detectors of skies, handshakes, or faces). Our goal is to provide a hierarchy of such classifiers and analyze their interactions so that we can build a model that relates a combination of the high-level visual features to specific conditions under which an image is appropriate for inclusion in a multimedia document.

Acknowledgements

We would like to thank Seungyup Paek, Alejandro Jaimes, Shih-Fu Chang, Luis Gravano, Barry Schiffman, Hongyan Jing, and especially Kathy McKeown for numerous suggestions and discussions during the development of this work. We are also grateful to the volunteers who categorized the images in our collection, and to the anonymous reviewers for additional suggestions that helped to improve this paper. The work reported here has been funded in part by a National Science Foundation STIMULATE grant, IRI-96-19124. Any opinions, findings,

or recommendations are those of the authors, and do not necessarily reflect the views of the NSF.

Additional information about this work, including our collection of images, human response collection protocol, and detailed results, can be found at `http://www.cs.columbia.edu/~sable/inout.html`.

References

A. V. Aho, S.-F. Chang, K. R. McKeown, D. Radev, J. R. Smith, and K. Zaman. Columbia Digital News Project. *International Journal of Digital Libraries*, **1**(4):377–385, 1998.

J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. C. Jain, and C. Shu. The VIRAGE Image Search Engine: An Open Framework for Image Management. In *Proceedings of the Symposium on Electronic Imagic: Science and Technology—Storage and Retrieval for Image and Video Databases IV*. IS&T/SPIE, February 1996.

D. M. Bates and D. G. Watts. *Nonlinear Regression Analysis and its Applications*. Wiley, New York, 1988.

K. W. Church. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the Second Conference on Applied Natural Language Processing (ANLP-88)*, pages 136–143, Austin, Texas, February 1988.

R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, United Kingdom, 5th edition, 1934.

J. L. Fleiss. *Statistical Methods for Rates and Proportions*. Wiley, New York, 2nd edition, 1981.

D. A. Forsyth and M. M. Fleck. Finding Naked People. In *Proceedings of the European Conference on Computer Vision*, Berlin, Germany, 1996.

L. S. Gay and W. B. Croft. Interpreting Nominal Compounds for Information Retrieval. *Information Processing and Management*, **26**(1):21–38, 1990.

T. Hastie and D. Pregibon. Shrinking Trees. Technical report, AT&T Bell Laboratories, 1990.

V. Hatzivassiloglou and K. R. McKeown. Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 172–182, Columbus, Ohio, June 1993.

C. R. Hicks. *Fundamental Concepts in the Design of Experiments*. Holt, Rinehart, and Wilson, New York, 3rd edition, 1982.

D. Lewis, R. Schapire, J. Callan, and R. Papka. Training Algorithms for Linear Text Classifiers. In *Proceedings of the 19th International ACM SIGIR Conference on Researce and Development in Information Retrieval (SIGIR-96)*, 1996.

D. Lewis. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In *Proceedings of the European Conference on Machine Learning*, 1998.

W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin. The QBIC Project: Quering Images by Content Using Color, Texture, and Shape. In *Proceedings of Symposium on Electronic Imaging: Science and Technology—Storage and Retrieval for Image and Video Databases*. SPIE, February 1993.

V. E. Ogle and M. Stonebraker. Chabot: Retrieval from a Relational Database of Images. *IEEE Computer Magazine*, **28**(9):40–48, September 1995.

S. Paek, C. L. Sable, V. Hatzivassiloglou, A. Jaimes, B. H. Schiffman, S.-F. Chang, and K. R. McKeown. Integration of Visual and Text-Based Approaches for the Content Labeling and Classification of Photographs, 1999. In preparation.

A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Tools for Content-Based Manipulation of Image Databases. In *Proceedings of the Symposium on Electronic Imagic: Science and Technology—Storage and Retrieval for Image and Video Databases II*, pages 34–47, Bellingham, Washington, 1994. SPIE.

J. R. Quinlan. Induction of Decision Trees. *Machine Learning*, **1**(1):81–106, 1986.

J. Rocchio. Relevance Feedback in Information Retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter 14, pages 974–979. Prentice-Hall, 1971.

N. C. Rowe and E. J. Guglielmo. Exploiting Captions in the Retrieval of Multimedia Data. *Information Processing and Management*, **29**(4):453–561, 1993.

G. Salton and C. Buckley. Term Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, **25**(5):513–523, 1988.

G. Salton and M. Smith. On the Application of Syntactic Methodologies in Automatic Text Analysis. In *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1989.

G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, Massachusetts, 1989.

T. J. Santner and D. E. Duffy. *The Statistical Analysis of Discrete Data*. Springer-Verlag, New York, 1989.

D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley and Sons, New York, 1992.

A. F. Smeaton and I. Quigley. Experiments on Using Semantic Distances Between Words in Image Caption Retrieval. In *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

A. F. Smeaton. Progress in the Application of Natural Language Processing to Information Retrieval Tasks. *The Computer Journal*, **35**(3):268–278, 1992.

J. R. Smith and S.-F. Chang. Visually Searching the Web for Content. *IEEE Multimedia*, **4**(3):12–20, July–September 1997.

R. K. Srihari. Automatic Indexing and Content-Based Retrieval of Captioned Images. *IEEE Computer Magazine*, **28**(9):49–58, September 1995.

M. Szummer and R. W. Picard. Indoor-Outdoor Image Classification. In *IEEE Workshop on Content Based Access of Image and Video Databases (CAIVD-98)*, pages 42–51, Bombay, India, January 1998.

A. Vailaya, M. Figueiredo, A. K. Jain, and H. Zhang. Bayesian Framework for Semantic Classification of Outdoor Vacation Images. In *Proceedings of SPIE—Storage and Retrieval for Image and Video Databases VII*, San Jose, California, 1999.

N. Wacholder, Y. Ravin, and M. Choi. Disambiguation of Proper Names in Text. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing (ANLP-97)*, pages 202–208, Washington, D.C., April 1997.