Integration of Visual and Text-Based Approaches for the Content Labeling and Classification of Photographs

Seungyup Paek[*] **Carl L. Sable**[†] **Vasileios Hatzivassiloglou**[†]
Alejandro Jaimes[*] **Barry H. Schiffman**[†] **Shih-Fu Chang**[*]
Kathleen R. McKeown[†]

[*]Department of Electrical Engineering
1312 S. W. Mudd Building
Columbia University
New York, N.Y. 10027
{syp, sfchang,
ajaimes}@ctr.columbia.edu

[†]Department of Computer Science
450 Computer Science Building
Columbia University
New York, N.Y. 10027
{sable, vh, bschiff,
kathy}@cs.columbia.edu

July 24, 1999

Topical categories: Content-based indexing/retrieval, text and image indexing/retrieval, text categorization, information access in digital libraries, architectures in Information Retrieval systems.

Abstract

Annotating photographs automatically with content descriptions facilitates organization, storage, and search over visual information. We present an integrated approach for scene classification that combines image-based and text-based approaches. On the text side, we use the text accompanying an image in a novel TF*IDF vector-based approach to classification. On the image side, we present a novel OF*IIF (object frequency) vector-based approach to classification. Objects are defined by clustering of segmented regions of training images. The image based OF*IIF approach is synergistic with the text based TF*IDF approach. By integrating the TF*IDF approach and the OF*IIF approach, we achieved a classification accuracy of 86%. This is an improvement of approximately 12% over existing image classifiers, an improvement of approximately 3% over the TF*IDF image classifier based on textual information, and an improvement of approximately 4% over the OF*IIF image classifier based on visual information.

1 Introduction

With the ease of creation and manipulation of multimedia data, an increasing amount of online multimedia information from various sources is now available. A recognized technical challenge is the development of accurate algorithms to support different functionalities that are useful across different multimedia content-focused applications [Chang *et al.* 1997].

For example, based on our experience in a previous Web image search engine, WebSEEk, we found that subject navigation and browsing is the most popular user operation in interactive image retrieval [Chang *et al.* 1997]. Users usually first browse through the subject hierarchy to get general ideas about the collection and then issue specific queries using keywords, visual features, or a combination of both.

Robust image classification is critical in successfully mapping images to specific classes in an image subject hierarchy. Automatic image classification systems are still relatively young. Recent work uses textual features such as keywords in URLs or manual annotations [Bach *et al.* 1996; Smith and Chang 1997a], image features

alone such as color histograms and transform-domain features [Szummer and Picard 1998; Vailaya *et al.* 1998], or classification specific to restricted tasks such as detection of pornography [Forsyth and Fleck 1997].

In this paper, we present a novel image classifier that uses the textual information accompanying an image, a novel image classifier that uses the visual information of an image, and an integration framework for combining the image classifiers which are based on information of different modalities.

On the text side, we use the text accompanying an image in a novel TF*IDF vector-based approach to classification, showing the effects on accuracy of using different amounts of text (e.g., full article, caption, or first sentence of caption), along with different types of information extracted from the text (e.g., all words, open class words only, words with their part of speech, etc.). To this we add single words which can discriminate between classes and a conversion from TF*IDF to estimated probability densities. For the set of approximately 1300 news images that we are working with, the TF*IDF approach gives a classification accuracy of 83.3% .

On the image side, we present a novel OF*IIF (object frequency) vector-based approach to classification. The OF*IIF approach can be based on objects that are defined through two different approaches: (1) Objects defined by clustering of automatically segmented regions of training images (2) Objects defined by knowledge-based models. In this paper, we focus on objects that are defined by clustering. In our current research, we are working on developing object recognition systems based on knowledge based models to be incorporated into the OF*IIF approach.

The OF*IIF approach based on objects defined through clustering of image regions gives a classification accuracy of 82.4%. The OF*IIF approach based on clustering of image regions achieves a classification performance that is approximately 8% better than an image classifier based on existing image-based approaches. Furthermore, we show that a core strength of the OF*IIF approach is that it is synergistic with the text based TF*IDF approach. Finally, the OF*IIF approach can incorporate object recognition systems in a modular and scalable way to increase classification accuracy.

By integrating the TF*IDF approach and the OF*IIF approach, the classification accuracy is increased to a combined accuracy of 86.2%. This is an improvement of approximately 12% over existing image classifiers, and an improvement of approximately 3% over the TF*IDF image classifier alone, and an improvement of approximately 4% over the OF*IIF image classifier alone.

We have developed a general integration framework and classification techniques that can be used for any choice of output categories, but in the present work we demonstrate how they work on classifying images as indoor/outdoor. In the domain of terrorist news in which we are working, outdoor images are often photographs taken on the scene, illustrating damage. Indoor images include, among others, press conferences and speeches related to terrorist events. In future work, we will investigate how the indoor/outdoor categories combine with other categories to create other meaningful categories for the domain.

After presenting related work, we first describe our experimental setup, including our collection of images and how indoor/outdoor labels were assigned to them. We then describe the individual text and image classifiers in turn. Finally, we turn to the integration of the classifiers, showing the impact on results.

2 Related research

Several approaches have been developed recently to improve image retrieval mechanisms. Keywords are used for image indexing by extracting significant words from associated documents or manual annotation [Bach *et al.* 1996; Smith and Chang 1997a]. Feature-level similarity search has been explored in several image search engines using image features only [Niblack *et al.* 1993; Pentland *et al.* 1994; Smith and Chang 1996]. A combination of textual and visual features has been used in integrated image queries [Ogle and Stonebraker 1995; Smith and Chang 1997a]. In [Srihari 1995], integration of image features (e.g., face detection) and textual features (e.g., corresponding names and locations) has been used to achieve cross-modality indexing and provide efficient access methods.

There have also been several encouraging research results in scene classification. [Szummer and Picard 1998] present a system that classifies indoor and outdoor images on the basis of color histograms and discrete cosine transform coefficients. For a set of 1,300 consumer photographs, the system achieves 90% classification accuracy. While the system achieves relatively high classification accuracy for the Kodak consumer photographs reported in [Szummer and Picard 1998], we found that the accuracy is significantly lower (74.7%)

on the set of news images we are working with.

[Smith and Chang 1997b] use a multi-stage system to classify images into several classes, sequentially assigning images to *type* (e.g., color graphics, black and white), *domain* (e.g., center surround, silhouette) and *semantic* classes (e.g., beach, buildings, nature, sunsets). Image semantics were determined by a novel system which matches the arrangements of regions in the images to composite region templates, and an overall classification accuracy of 78% was achieved.

For restricted classification, [Forsyth and Fleck 1996] detect naked people in an image, using a representation (*body plan*) for people and animals, which is adapted to segmentation and to recognition in complex environments. The representation is an organized collection of grouping hints obtained from a combination of constraints on color and texture and constraints on geometric properties such as the structure of individual parts and the relationships between parts.

3 Experimental Setup

For both classifier training and evaluation, we need access to data that has been carefully labeled according to the classes of interest, in this case as indoor or outdoor images. Our raw data consists of news articles containing images and corresponding captions that were extracted from a variety of Clarinet current news newsgroups, spanning the period from April 1997 to April 1998. We extracted 1,675 images with corresponding captions and articles from such documents.

Each of these images was labeled as indoors or outdoors on the basis of independent human judgements. Fourteen volunteers accessed the images together with their captions sequentially through a web-based interface and assigned indoor or outdoor labels. Volunteers were supplied with guidelines that helped them deal with some common ambiguities, and had the option of labeling an image as *Indoor*, *Outdoor*, *Likely Indoor*, *Likely Outdoor*, or *Ambiguous*. Each image was labeled by at least two volunteers. The analysis of the assigned labels indicates that, in most cases (87.7%), a definite indoor or outdoor judgement was made, and only 3% of the labels used were "ambiguous". Agreement between humans was also high (90.4% of the images had compatible labels, although sometimes with different degrees of confidence). However, there are still some hard cases where different internal definitions of the task led to different classification decisions; for example, close-ups of people inside a vehicle such as a car or plane and pictures of people under the roof of a structure with no walls were often labelled differently by different judges. 39 of the 1,675 images received one "definite indoor" and one "definite outdoor" label, and while some of these are in retrospect attributable to human error, most correspond to differences in the definition of what makes a scene indoors or outdoors.

For the experiments reported in the rest of the paper, we restricted our image collection to those that received definite judgements with absolute confidence and with the same label by at least two evaluators. This set covers the vast majority of images, containing 1,339 out of the original 1,675 images. 401 of these images (29.9%) are classified as indoor and 938 (70.1%) as outdoor.

We also varied the amount of information that was available to the human volunteers. In order to measure how well a system that sees only the text or only the image would perform, we asked some of the volunteers to label the images by looking only at the images or only at the captions. We compared their performance to our reference set described in the previous paragraph. Their average accuracy was 95.5% when only the image was available, and 87.5% when only the caption was available. This provides upper bounds on the performance of classifiers working with a single medium, and also indicates that full resolution of the problem requires access to both types of information. It also indicates that humans can make the indoors/outdoors distinction more easily from the image than from the corresponding text, something that, as we shall see, is reversed for the automatic system.

4 Text-Based Approaches For Indoor/Outdoor Classification

We examine two alternative classification approaches that draw on textual information: TF*IDF scores and machine learning of words that discriminate between the classes. Both techniques are general and can be immediately applied to other classification problems, e.g., for the computation of other types of content labels.

4.1 TF*IDF scoring of image captions

Methodology Our first text-based classifier relies on TF*IDF scores [Salton and Buckley 1988; Salton 1989] to categorize images via their corresponding captions or articles. For a single document, a word's TF, or "term frequency", is the number of times the word occurs in that document; for categories such as all indoor or all outdoor images, the TF is the number of times the word occurs over all documents in that class. A word's IDF, or "inverse document frequency", is the log of the ratio of the total number of documents to the number of documents that contain the word. The product TF*IDF term is therefore higher when a word combines a balance of high frequency within a document (signifying high importance for this document) and low overall dispersion within the collection (signifying high specificity).

We apply TF*IDF in a novel way to image categorization by examining different definitions of a word, experimenting by restricting our analysis to different spans of text associated with an image and to different sets of words extracted from a span of text, sometimes including, for example, all words within the span, sometimes only open class words, and sometimes collating together different orthographic strings. Overall, we varied four experimental parameters in a complete designed experiment [Hicks 1982]:

1. The extent of each input data item over which TF*IDF values are computed. We have articles and captions available, but it is clear that captions are much more closely related to images, and omitting the article may actually improve performance by avoiding "background" noise. Similarly, we observed that the first sentence of a caption usually describes the image, while subsequent sentences provide historical information, 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.

   Thus we experiment with word vectors drawn from the full article plus caption, the article only, the caption only, or just the first sentence of the caption.

2,3. Part-of-speech information. Syntactic parts of speech allow for the restriction of the input words to members of "interesting" classes only such as open-class words (nouns, verbs, adjectives, and adverbs). Limiting input to such words and some closed-class words of particular relevance to this task (i.e., prepositions) is similar to using a flexible stop list. Keeping the part of speech as part of the word during comparison also serves as another experimental parameter, since it offers partial semantic disambiguation (e.g., "cross" as a noun and "cross" as a verb).

4. Case sensitivity. We chose to experimentally test the effect of word capitalization, sometimes collapsing words that only differ in capitalization to the same token.

For each combination of the above parameters, we compute TF*IDF scores for each document and also class TF*IDF scores for all documents associated with indoor or outdoor images (given a specific training set). We then aggregate the various TF*IDF scores by computing the dot product between the TF*IDF vector corresponding to the document and each of the two class TF*IDF vectors, i.e.,

$$Score(image, class) = \sum_i TFIDF_{image}[i] \times TFIDF_{class}[i] \quad (1)$$

Two more experimental parameters enter the calculations of equation (3). First, we optionally ignore elements of the document vector that are lower than a prespecified constant, eliminating relatively insignificant words. Second, we optionally normalize for the a priori probability that a new case will fall in one of the two classes by dividing all elements of the class vectors by the total number of images in the training set that fall within the corresponding class.

At this point, each image receives two scores, one measuring similarity with the prototypical indoor images and the other with the outdoor ones. Comparing the two scores directly assumes implicitly that the two scores are on the same scale, and that equal values indicate no preference for either class. We also explore

the alternative of empirically estimating the probability density of the difference of the two scores, using a rectangular smoothing window on top of the histogram of this difference function [Scott 1992]. In this manner, still using our training set, we estimate the probability of each class given a particular value of the difference of our two similarity scores (indoor minus outdoor). This alternative not only automatically adjusts the cut-off point between the two categories away from zero differences, but also scales the classifier's answer between 0 and 1, allowing easy combination with other independent classifiers.

Results We randomly selected approximately one half of our 1,339 images with definite agreement on indoor/outdoor label as the training set, and the remaining half of the images as the test set. We applied the TF*IDF computations of the previous subsection for each of the 384 combinations of experimental parameters. In each such run, we randomly divided the training set into three equal parts, and repeatedly trained on two of these and tested 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.

We found wide variety in the obtained average accuracy score (percentage of correct answers for both classes) depending on these parameter settings. Some of the parameters had a major effect, namely,

- Restricting analysis to just the first sentences of captions accounts for the top 37 best-scoring experiments. First sentences clearly outperformed full captions, which in turn outperformed the full article.
- Class frequency normalization accounts for 12 of the top 15 experiments, and clearly outperforms the alternative of unmodified class vectors.
- Density estimation improves performance in almost all cases, and is included in all combinations of parameters ranked near the top.
- Using only words from open syntactic classes (plus prepositions) was better than using all the words, and better than open class words with proper nouns removed.

On the other hand, using thresholds on individual TF*IDF components, transforming words to lower case, and keeping or discarding the part of speech from a word had each little effect on the accuracy score. Table 1 summarizes the effect of each value for each experimental parameter, while Table 2 shows the top ten combinations of parameters in terms of overall accuracy during cross-validation.

On the basis of the cross-validation of the different parameters over the training set, we selected the following combination for our system: first sentences only, open-class words plus prepositions, no transformation of capitalization, no distinction between polysemous words according to part of speech, normalization of class vectors, and no thresholds during TF*IDF computations. Then, we retrained on the full training set and tested on the unseen test set. The corresponding classifier achieves 90.72% accuracy on the training set and 83.3% accuracy on the test set.

4.2 Words As Class Discriminators

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.

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.

Parameter	Value	Average Accuracy
Text Fields	first sentences	81.02%
	captions	78.27%
	articles + captions	69.35%
	articles	68.39%
Parts of Speech	open POS	78.84%
	all POS	74.48%
	open POS except proper nouns	73.45%
Threshholds	low	74.83%
	medium	74.71%
	none	74.51%
	high	72.99%
Normalization	yes	75.81%
	no	72.71%
Case Sensitivity	no	74.27%
	yes	74.25%
Keep Tag	yes	74.30%
	no	74.22%

Table 1: Average overall accuracy of all experiments with a given value of each parameter.

Parts of Speech	*Case Sensitive*	*Keep Tag*	*Normalization*	*Thresholds*	*Text Span*	*Accuracy*	
						without density	*with density*
open POS	yes	no	yes	none	first sentences	75.06	83.22
open POS	yes	no	yes	low	first sentences	75.06	83.22
all POS	no	yes	yes	medium	first sentences	78.08	82.89
open POS	no	no	yes	low	first sentences	74.83	82.89
open POS	no	no	yes	none	first sentences	74.61	82.89
all POS	no	no	yes	medium	first sentences	79.08	82.77
open POS	yes	no	no	none	first sentences	78.75	82.77
all POS	no	yes	no	medium	first sentences	78.97	82.66
all POS	yes	no	yes	low	first sentences	77.29	82.66
all POS	no	no	yes	low	first sentences	76.73	82.66

Table 2: Top ten combinations of TF*IDF experiment parameters after three-fold cross validation on the training set.

Results Using the same training/test set division as with the TF*IDF experiments reported earlier, our list of candidate discriminators contains 665 words. Both the log-linear regression model and the tree select a subset of these words; 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.

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

5 Image-Based Approaches for Indoor-Outdoor Classification

5.1 OF*IIF scoring of images

We present a novel image classification approach that draws on visual information: OF*IIF scoring of images. The approach is general and can be immediately applied to other classification problems.

Methodology Our novel image-based classifier is analogous to the TF*IDF approach outlined in the previous section for text based classification of images. The image based approach is referred to as OF*IIF scoring of images. For a single image, the OF, or object frequency, is the number of times an object occurs in that image; for categories such as all indoor or all outdoor images, the OF is the number of times the object occurs over all images in that class. An object's IIF, or inverse image frequency, is the log of the ratio of the total number of images to the number of images that contain the object. The product OF*IIF is therefore higher when an object combines a balance of high frequency within an image (signifying high importance for this image) and low overall dispersion within the entire collection (signifying high specificity).

The OF*IIF scoring of images depends on two assumptions. First, that a set of objects can be defined for a given set of training images. Second, that once a set of objects is defined, the occurrence of an object can be detected in an image. In the experiments reported in this paper, we have restricted our analysis to a cluster based approach to defining and detecting image objects. The approach that we developed has three main components: visual feature extraction of image regions, clustering image regions, and cluster matching of image regions. We describe the main features of each of these components below.

Visual feature extraction for image regions An image is first divided into a set of regions, which can be formed by segmenting the image into blocks, or by segmenting the image based on color and texture coherence [Zhong and Chang 1997]. Currently, in our experiments, we are dividing each image into 64 sub-images or blocks of equal dimensions i.e. the image is divided into an 8×8 grid. For each sub-image (block) we generate a set of color and texture related features. The color related visual feature we generate is the *HSV* color histogram of each block. The texture related visual feature we generate is the edge direction histogram of each block.

The *HSV* color representation is attractive because it represents with equal emphasis the three color attributes that correspond to the human perception of colors: hue (H), saturation (S), and value (V). Value corresponds to the brightness of a color. For each block of an image, a histogram is computed by recording the number of occurrences of each quantized HSV color in the pixels of the block. Currently, we are using a total of 166 quantized HSV colors in our experiments. Details of the HSV color reqresentation and quantization are given in [Smith and Chang 1996a].

Edge direction histogram texture features have previously been used in image classification with good performance [Vailaya *et al.* 1998]. We follow the method first proposed in [Vailaya *et al.* 1998]. The Sobel edge detection algorithm [Kasturi and Jain 1991] is used to extract the edges in a block. An extra bin in the histogram is used to measure the frequency of non-edge pixels in each block of an image. A total of 73 bins are used to represent the edge direction histogram of an image. The first 72 bins are used to represent edge directions quantized at 5 degree intervals and the last bin represents a count of the number of pixels that did not contribute to an edge.

Defining and detecting image objects Once the visual features have been extracted for the blocks of all the training images, the feature vectors associated with the blocks are clustered. We use a clustering technique that incorporates the concept of similarity based on the sharing of near neighbors. The clustering technique is an essentially parallel approach and the scheme is applicable to problems involving large sample size and high dimensionality [Jarvis and Patrick 1973]. For each point to be clustered, *k1* nearest neighbors are computed. Once the *k1* nearest neighbors have been computed for all the points, clusters are defined by points which have at least *k2* of the *k1* nearest neighbors that are the same i.e. points in a cluster share at least *k2* nearest neighbors. In our current experiments, we are clustering feature vectors based on single features and not on composite feature vectors based on multiple features. The cluster centroid of each cluster that is generated defines an object, as required for the OF*IIF approach. This stage defines a set of objects *(o1, o2, o3,...)*, each with a corresponding cluster centroid.

Once a set of objects and corresponding cluster centroids have been defined based on all the blocks of the training images, we have to detect the occurrence of these objects in both the training images and testing images. In our approach, objects are detected by matching the blocks of an image to the cluster centroids of the objects that have been defined. To match a block to a cluster centroid, we find the cluster centroid for which a distance measure between the block and the cluster centroid is a minimum. In this stage, every block of an image is matched to one of the objects from the set of all objects *(o1, o2, o3,...)*. In this way, we compute the OF*IIF scores for each image and also the class OF*IIF scores for all images associated with indoor or outdoor images (given a specific training set). For a single image, the OF, or object frequency, is the number of times an object occurs in that image; for categories such as all indoor or all outdoor images, the OF is the number of times the object occurs over all images in that class. An object's IIF, or inverse image frequency, is the log of the ratio of the total number of images to the number of images that contain the object.

Classification Once we have computed the OF*IIF scores for each image and also the class OF*IIF scores, we compute the dot product between the OF*IIF vector corresponding to the image and each of the two class OF*IIF vectors. Each image receives two scores, one measuring similarity with the prototypical indoor images and the other with the outdoor ones. The OF*IIF scores are found by computing the dot product between the OF*IIF vector corresponding to the image and each of the two class OF*IIF vectors, i.e.,

$$Score(image, class) = \sum_i OFIIF_{image}[i] \times OFIIF_{class}[i] \tag{2}$$

We normalize for the a priori probability that a new case will fall in one of the two classes by dividing all elements of the class vectors by the total number of images in the training set that fall within the corresponding class. At this point, each image receives two scores, one measuring similarity with the prototypical indoor images and the other with the outdoor ones. As in the TF*IDF approach, we empirically estimate the probability density of the difference of the two scores, using a rectangular smoothing window on top of the histogram of the difference function.

Results We randomly selected approximately one half of our 1,339 images with definite agreement on indoor/outdoor label as the training set, and the remaining half of the images as the test set. Based on the HSV histogram feature alone, the image classifier achieves 79.5% accuracy on the test set. Based on the edge direction histogram feature alone, the image classifier achieves 77.7% accuracy on the test set. Based on the HSV histogram and the edge direction features, the image classifier achieves 82.4% accuracy on the test set. Following the general approach described in [Szummer and Picard 1998], we achieved a classification accuracy of 74.7%.

6 Integrating image and text based approaches

In the previous sections, we presented two approaches for image classification. One approach draws on textual information (TF*IDF scoring of image captions). Another approach draws on visual information (OF*IIF scoring of images). These techniques can be integrated naturally by combining the scores from each approach.

We compute OF*IIF scores for each image and also the class OF*IIF scores for indoor and outdoor classes. We also compute TF*IDF scores for each image caption and also the class TF*IDF scores for indoor and outdoor classes. We then aggregate the various OF*IIF scores for each image together with the various TF*IDF scores for each image caption as follows:

$$Score(image, class) = \sum_i w_T(TFIDF_{image}[i] \times TFIDF_{class}[i]) + \sum_j w_O(OFIIF_{image}[j] \times OFIIF_{class}[j]) \tag{3}$$

Firstly, the equation shows that we compute the dot product between the OF*IIF vector corresponding to the image with each of the class OF*IIF vectors. This sum is weighted by a normalization factor which is equal to $1/NO$, where NO is equal to the average number of objects in all the images in the training set.

Secondly, the equation shows that we compute the dot product between the TF*IDF vector corresponding to the image with each of the class TF*IDF vectors. This sum is weighted by a normalization factor which is equal to $1/NT$, where NT is equal to the average number of terms in all the images in the training set. Finally, the equation shows that the normalized dot product of the OF*IIF vectors and the normalized dot product of the TF*IDF vectors are added together. This gives us two scores for each image, one score measures the similarity with the prototypical indoor images and the other with the outdoor ones. As we did in both the image based and text based approaches, we empirically estimate the probability density of the difference of these two scores.

Results The integration experiments demonstrated that combining information of the two modalities can achieve a classification accuracy of 86.2%. This is an improvement of approximately 12% over existing image classifiers, an improvement of approximately 3% over the text-based classifier alone, and an improvement of approximately 4% over the image-based classifier alone.

7 Conclusion and Future Research

We have described classifiers that, operating on textual or visual information, are able to classify a photograph as indoor or outdoor with high accuracy. The obtained performance is in many situations close to the performance of humans; for example, our text-based classifier achieves 83.3% accuracy while humans correctly perform the same task 87.5% of the time when looking at the textual information only. We have also developed an framework for integrating classifiers, and the integration experiments demonstrated that the combined approach can result in a classification accuracy of 86.2%. This is an improvement of approximately 12% over existing image classifiers, an improvement of approximately 3% over an image classifier based on text information alone, and an improvement of approximately 4% over an image classifier based on visual information alone.

We are currently looking into several additional features and classification approaches. On the text side, we are experimenting with a knowledge-based approach that relies on more extensive linguistic analysis of the image captions, parsing their relatively stylized structure and recovering salient information about the subject and main action described in the image. We will also condition the results of text-based classifiers on secondary features that might be easier to compute independently, for example an indoor/outdoor decision on the number of people present in the image.

On the image side, we are currently developing object detectors (e.g., fire, pavement, soil, vegetation, crowds, etc.) to be used in the OF*IIF approach. The OF*IIF technique currently only uses objects that are defined automatically for a set of training images using a clustering scheme. However, the OF*IIF approach can be based on objects that are defined through objects defined by knowledge-based models. A core strength of the OF*IIF approach is that it can incorporate objects in a modular and scalable way to increase classification accuracy. In the *Visual Apprentice* system [Jaimes and Chang 1999], a graphical user interface is utilized to define the object hierarchy (consisting of regions, object parts, and objects) and to label example regions in images for the desired class. The system then uses those examples to automatically learn detectors at each level in the object hierarchy. By providing the Visual Apprentice with 40 labeled training images, we achieved 55% recall and 87% precision in recognizing general (not just blue) skies. Currently the *Visual Apprentice* system is being used to build detectors for handshakes, number of people, and vegetation regions.

Finally, we are currently working on integrating various machine learning techniques to optimize the components in our image classifier that perform object definition and detection. For a variety of image collections, we are also performing extensive experiments to compare our approach with other image classification systems.

References

[Bach *et al.* 1996] 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.

[Bates and Watts 1988] Douglas M. Bates and Donald G. Watts. *Nonlinear Regression Analysis and its Applications*. Wiley, New York, 1988.

[Chang *et al.* 1997] S.-F. Chang, J. R. Smith, M. Beigi, and A. Benitez. Visual Information Retrieval from Large Distributed On-Line Repositories. *Communications of the ACM*, **40**(12):63–71, December 1997. Special issue on Visual Information Management.

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

[Forsyth and Fleck 1997] D. A. Forsyth and M. M. Fleck. Body Plans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Puerto Rico, June 1997.

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

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

[Jaimes and Chang 1999] Alejandro Jaimes and Shih-Fu Chang. Model Based Classification In *Proceedings of the International Society for Optical Engineering (SPIE)*, 1999 (to appear, January 1999).

[Kasturi and Jain 1991] R. Kasturi and R. C. Jain *Computer Vision: Principles*. IEEE Computer Society Press.

[Jarvis and Patrick 1973] R.A. Jarvis and Edward. A. Patrick Clustering Using similarity measure based on Shared Near Neighbors *IEEE Transactions on Computers* **C-22**(11):1027–1034, November 1973.

[Niblack *et al.* 1993] 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.

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

[Pentland *et al.* 1994] 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.

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

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

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

[Santner and Duffy 1989] Thomas J. Santner and Diane E. Duffy. *The Statistical Analysis of Discrete Data*. Springer-Verlag, New York, 1989.

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

[Smith and Chang 1996a] J. R. Smith and S.-F. Chang. Tools and Techniques for Color Image Retrieval. 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.

[Smith and Chang 1996] J. R. Smith and S.-F. Chang. VisualSEEk: A Fully Automated Content-Based Image Query System. In *Proceedings of the ACM Multimedia Conference*, Boston, Massachusetts, November 1996.

[Smith and Chang 1997b] J. R. Smith and S.-F. Chang. Multi-Stage Classification of Images from Features and Related Text. In *Proceedings of the Fourth DELOS Workshop*, Pisa, Italy, August 1997.

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

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

[Szummer and Picard 1998] Martin Szummer and Rosalind 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.

[Vailaya *et al.* 1998] A. Vailaya, A. Jain, and H. J. Zhang. On Image Classification: City vs. Landscape. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, Santa Barbara, California, June 1998.

[Zhong and Chang 1997] D. Zhong and S.-F. Chang. Video Object Model Segmentation for Content-Based Video Indexing. In *IEEE International Conference on Circuits and Systems*, Hong Kong, June 1997. Special Session on Networked Multimedia Technology and Applications.