Semantic Scene Classification for Image Annotation and Retrieval*

Özge Çavuş and Selim Aksoy

Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey
{cavus,saksoy}@cs.bilkent.edu.tr

Abstract. We describe an annotation and retrieval framework that uses a semantic image representation by contextual modeling of images using occurrence probabilities of concepts and objects. First, images are segmented into regions using clustering of color features and line structures. Next, each image is modeled using the histogram of the types of its regions, and Bayesian classifiers are used to obtain the occurrence probabilities of concepts and objects using these histograms. Given the observation that a single class with the highest probability is not sufficient to model image content in an unconstrained data set with a large number of semantically overlapping classes, we use the concept/object probabilities as a new representation, and perform retrieval in the semantic space for further improvement of the categorization accuracy. Experiments on the TRECVID and Corel data sets show good performance.

1 Introduction

Image annotation and content-based retrieval have been very active research areas with open problems due to the constant increase in the richness of the available image content. Contextual information plays a very important role for characterizing such content. A promising method for modeling context in images is scene classification because associating scenes with semantic labels has a high potential for providing a natural grouping of images instead of relying only on low-level features.

Scene classification and the related retrieval problems have two critical components: representing scenes and learning models for associating labels to these scenes. Given the difficulty of image segmentation in unconstrained data sets, most of the recent work use histograms of local features [1] or fixed grid layouts [2,3,4] for image representation but region-based models [5] can also be found. Intermediate semantic models that make use of the occurrence of common concepts, such as sky, water, grass, snow, have also been shown to improve the classification of natural scenes [4]. Although, recent work was mostly limited in terms of grid-based representations [2,3,4] or the small number of classes used [4], contextual modeling of image scenes using combinations of concepts and objects looks promising for decreasing the semantic gap.

* This work was supported in part by the TUBITAK Grant 104E077.

N. da Vitora Lobo et al. (Eds.): SSPR&SPR 2008, LNCS 5342, pp. 402–410, 2008.

In this paper, we describe an image annotation and retrieval framework that uses scene classification to detect high-level concepts for image indexing. First, images are segmented using clustering of color features and grouping of line structures (Sections 2 and 3). Next, each image is modeled using the histogram of quantized region types, and Bayesian classifiers are used to obtain the occurrence probabilities of high-level concepts and objects in images using these histograms (Section 4). An important observation is that a single class with the highest probability is not always sufficient to model image content in an unconstrained data set with semantically overlapping classes [2]. Therefore, we use the concept/object probabilities as a new representation, and perform retrieval in the semantic space for further improvement of the categorization accuracy (Section 5). Performance of the proposed models is evaluated using the TRECVID and Corel data sets.

2 Segmentation Using Color Information

Image segmentation is still an unsolved problem in computer vision. Although numerous algorithms have been shown to work well for images with only a few objects and a simple background, it seems impossible to find a fixed set of parameters that produces reasonable results in a large unconstrained data set. In this paper, we assume that a very precise segmentation of an image is not required for the scene classification and retrieval problem. Therefore, our goal is to obtain a rough estimate of important regions using unsupervised classification of color features and grouping of line segments.

For segmentation using color, we use the combined classifier approach in [6] because it fuses color and spatial information, and does not require the number of regions as an input parameter. First, an initial labeling of an image is done using k-means clustering of only the HSV values of pixels. Next, these pixel labels are used to train a multi-class nearest mean classifier on the HSV color features and a Parzen window classifier with a Gaussian kernel using the pixel positions as spatial features. The nearest mean classifier is selected for its simplicity and the Parzen window classifier is selected for its nonparametric nature for modeling a distribution with an indefinite number of modes (each mode corresponds to a segment). Then, the posterior probability outputs of each classifier for each pixel are combined using the product rule, and the pixels are assigned to the class with the largest probability. A new pair of classifiers are trained using these new pixel labels and the iterations continue until the pixel labels stabilize. Note that the number of clusters in the initial k-means clustering does not directly correspond to the number of segments, and can be empirically estimated using the number of dominant colors that can be found in the images in the selected data set.

Figure 1 shows example segmentations. The regions that are smaller than an area threshold are removed from the final segmentation where the results contain only contiguous sets of pixels that have a relatively uniform color distribution and are large enough.

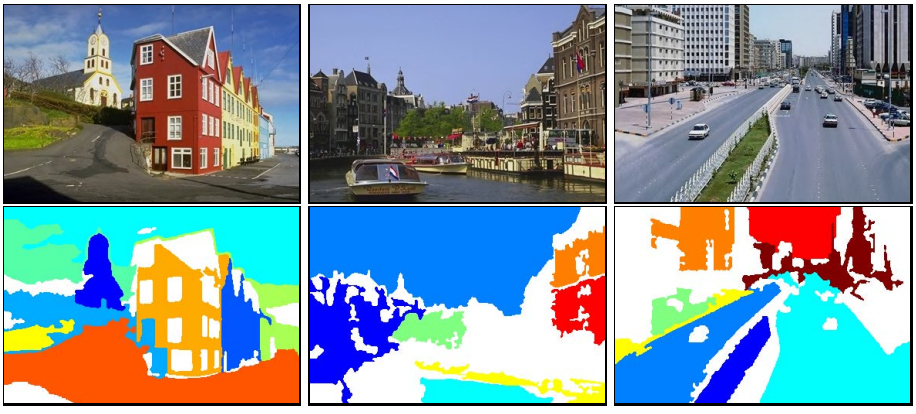


Fig. 1. Segmentation using color. Top row: color images; bottom row: segmentation results in pseudocolor (pixels marked as white are not segmented).

3 Segmentation Using Line Structure

Not all objects/regions of interest can be characterized using uniformity of colors. Li and Shapiro [7] showed that features of line segments can be exploited for building recognition. We also use properties of nearby line segments to extract regions.

A line segment is found either at the boundary of two different regions or within a highly textured region. We expect that nearby line segments that have a similar color distribution around them may belong to the same object. Given the line segments in an image, they are grouped using average linkage hierarchical clustering according to the average color values on opposite sides of each segment. The average link criterion is used because we want all line segments that are selected as belonging to the same object to have similar color values.

The resulting clusters can contain line segments that have similar color pairs but belong to different objects. Therefore, a second level of clustering is performed to select the neighboring ones. This is done separately for each cluster using single linkage hierarchical clustering according to the distances between the end points of line segments. The single link criterion is used because we want to merge two groupings of line segments only by considering their nearest pair of end points. In both clustering steps, the number of clusters is automatically determined from the dendrogram [8].

Li and Shapiro [7] also required the line segments to have similar orientations in their building recognition algorithm. We do not consider orientation here because we observed that clustering of line segments can also detect highly textured regions such as trees where the tree branches have an almost random orientation. However, we perform a final post-processing step to eliminate the clusters with a very small number or a very sparse spatial distribution of line segments where the spatial coverage of a line cluster is computed as the ratio

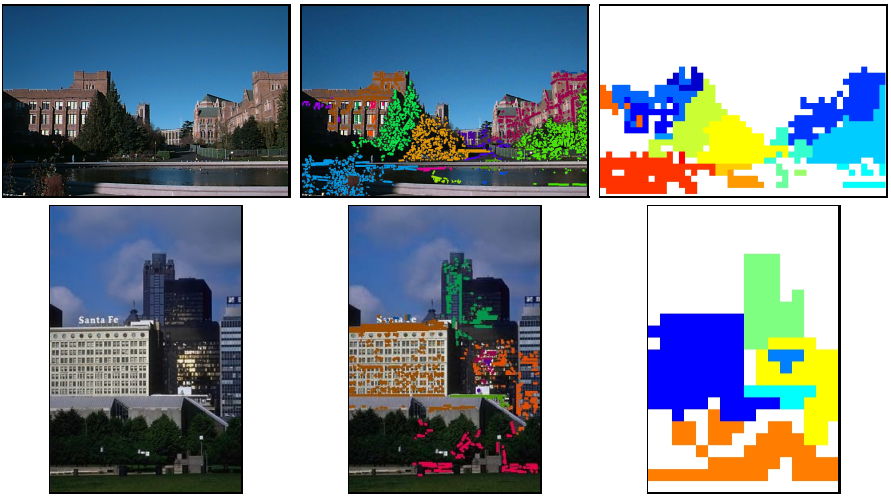


Fig. 2. Segmentation using line structure. First column: color images; second column: line clusters in pseudocolor; third column: final regions (pixels marked as white are not segmented).

of the area of the convex hull formed by the corresponding line segments to the number of line segments.

Finally, the resulting line segment clusters are converted to region representations by partitioning an image into non-overlapping grid cells and labeling each grid cell with the label of the cluster whose lines most frequently intersect with the cell. Figure 2 shows example segmentations.

4 Scene Classification

In this paper, a scene's content is represented as a collection of its regions. The regions that are extracted using pixel-based color information (Section 2) are modeled using their HSV histograms with 8 bins used for the H channel and 3 bins for each of S and V channels. Then, the k-means algorithm is used to create a codebook of k_1 region types. The regions that are extracted using line structure information (Section 3) are modeled using a 10-bin histogram of the orientation of their line segments. Similarly, the k-means algorithm is used to create a codebook of k_2 region types for the whole data set. After each region is assigned a type label, as the final representation, each scene is modeled using the histogram of $k_1 + k_2$ region types.

We use a Bayesian framework for scene classification and investigate two different settings for probability estimation. In both settings, the goal is to estimate the posterior probabilities $p(w_j|\mathbf{x})$, $j = 1, \ldots, c$, where w_j represents the j'th class, c is the number of classes, and $\mathbf{x}$ is the histogram of the quantized region types.

The first setting treats the region types independently and estimates the class-conditional probabilities using the multinomial model. The parameters of the model are computed using maximum likelihood estimates that involve counting the number of times each region type is observed for each class in the training data [9].

The second setting treats each class separately and trains a one-class Gaussian classifier using the training examples for each class independently. The one-class setting is particularly suitable here because the classes are not mutually exclusive and the commonly used two-class (target vs. others) approach is not applicable because sampling a sufficient number of training data from the "others" class is not always possible. One-class classifiers model only the "target" class and assume a low uniform distribution for the "others" class [10]. After a probability density (in this case, a Gaussian) is estimated using the training examples of the target class, a threshold is set on the tails of this distribution and a specified amount of the target data is rejected.

Experiments: We used the TRECVID 2005 and Corel data sets and their ground truth to evaluate the algorithms proposed in this paper. The TRECVID data set contains 24517 video shots in 18 classes and the Corel data set contains 4999 images in 21 classes. Two thirds of the images were used for training and the remaining one third for testing. Both k_1 and k_2 were set as 1000.

The Bayesian classifier assigns each image to the class with the highest posterior. Table 1 shows the confusion matrices when the multinomial model was used. The matrices show that the error rates were rather high (12.83% accuracy for the TRECVID data set and 34.76% accuracy for the COREL data set) but most of the misclassifications occurred among the classes that have a semantic overlap (e.g., boat–water, bus–road, sky–sunset). The confusions for the one-class Gaussian model were worse as most of the images were assigned to the outdoor class for the TRECVID data set and to the vegetation class for the Corel data set when the highest posterior was used. We also ran the bag-of-words model with probabilistic latent semantic analysis [1] on the Corel data set and obtained 19.58% accuracy (compared to 34.76% by the multinomial model in Table 1(b)).

An important observation was that, when all probabilities were considered, the image content can be modeled by multiple classes instead of a single class that has the highest probability. Figure 3 shows the posterior probability values for example images. These examples also confirm that the class models and the resulting probabilities are not necessarily wrong but a single class assignment is not appropriate when an unconstrained data set with a diverse set of classes is considered [2].

5 Image Annotation and Retrieval

The classification results show that an image cannot always be assigned to a single class if the categories are not mutually exclusive. Furthermore, different users may have different requirements where one may be interested in urban scenes but another may specifically look for vehicles. In image annotation, a

Table 1. Confusion matrices for Bayesian classification using the multinomial model

(a) TRECVID

	Airplane	Boat/ship	Building	Bus	Car	Desert	Fire	Mountain	Disaster	Outdoor	Road	Sky	Snow	Sports	Truck	Urban	Vegetation	Waterscape
Airplane	0.19	0.07	0.07	0.04	0.00	0.04	0.07	0.07	0.00	0.00	0.04	0.04	0.00	0.04	0.04	0.00	0.00	0.30
Boat/ship	0.00	0.07	0.07	0.00	0.04	0.00	0.04	0.00	0.00	0.00	0.07	0.07	0.07	0.07	0.00	0.00	0.04	0.44
Building	0.01	0.01	0.14	0.02	0.09	0.05	0.04	0.02	0.02	0.08	0.07	0.08	0.02	0.04	0.03	0.12	0.09	0.06
Bus	0.00	0.00	0.27	0.00	0.07	0.00	0.00	0.07	0.00	0.00	0.27	0.07	0.00	0.00	0.07	0.00	0.20	0.00
Car	0.01	0.02	0.06	0.00	0.20	0.02	0.03	0.00	0.03	0.06	0.12	0.04	0.01	0.05	0.04	0.12	0.11	0.08
Desert	0.01	0.01	0.04	0.00	0.03	0.31	0.12	0.01	0.03	0.03	0.09	0.06	0.03	0.07	0.03	0.04	0.03	0.04
Fire	0.01	0.01	0.10	0.01	0.06	0.16	0.22	0.00	0.01	0.04	0.07	0.07	0.00	0.00	0.01	0.03	0.03	0.14
Mountain	0.02	0.07	0.04	0.02	0.00	0.04	0.02	0.22	0.02	0.07	0.17	0.02	0.02	0.00	0.00	0.02	0.04	0.20
Disaster	0.00	0.00	0.17	0.00	0.13	0.04	0.04	0.04	0.00	0.00	0.13	0.04	0.00	0.04	0.00	0.17	0.13	0.08
Outdoor	0.02	0.02	0.09	0.02	0.07	0.05	0.04	0.02	0.02	0.09	0.07	0.05	0.02	0.08	0.03	0.12	0.12	0.08
Road	0.01	0.01	0.10	0.03	0.14	0.05	0.05	0.01	0.03	0.05	0.09	0.04	0.04	0.06	0.03	0.08	0.09	0.08
Sky	0.03	0.03	0.08	0.01	0.06	0.08	0.05	0.04	0.02	0.05	0.08	0.10	0.03	0.04	0.03	0.07	0.10	0.10
Snow	0.04	0.00	0.07	0.00	0.04	0.07	0.07	0.11	0.00	0.04	0.04	0.00	0.29	0.00	0.07	0.00	0.04	0.14
Sports	0.01	0.02	0.04	0.03	0.06	0.02	0.01	0.02	0.00	0.06	0.06	0.01	0.03	0.38	0.03	0.06	0.14	0.06
Truck	0.00	0.00	0.08	0.02	0.06	0.10	0.06	0.04	0.00	0.04	0.18	0.04	0.04	0.00	0.00	0.10	0.08	0.14
Urban	0.01	0.01	0.12	0.02	0.09	0.04	0.04	0.02	0.01	0.10	0.09	0.06	0.01	0.06	0.03	0.13	0.09	0.07
Vegetation	0.01	0.02	0.09	0.02	0.07	0.04	0.03	0.03	0.01	0.07	0.07	0.04	0.02	0.11	0.03	0.08	0.20	0.06
Waterscape	0.03	0.07	0.05	0.01	0.03	0.04	0.02	0.02	0.01	0.03	0.05	0.05	0.05	0.02	0.04	0.05	0.05	0.38

(b) Corel

	Airplane	Boat/ship	Building	Bus	Car	Castle	Coast	Desert	Harbor	Mountain	Night	Road	Rock	Ruin	Sky	Snow	Sunset	Surfing	Train	Vegetation	Waterfall
Airplane	0.21	0.08	0.02	0.00	0.06	0.02	0.08	0.03	0.00	0.06	0.00	0.00	0.02	0.06	0.02	0.08	0.02	0.05	0.06	0.11	0.03
Boat/ship	0.02	0.38	0.04	0.01	0.01	0.02	0.12	0.01	0.02	0.00	0.01	0.00	0.00	0.02	0.04	0.08	0.01	0.11	0.08	0.00	0.01
Building	0.00	0.06	0.48	0.04	0.05	0.06	0.01	0.00	0.01	0.00	0.01	0.00	0.04	0.07	0.01	0.02	0.00	0.03	0.05	0.04	0.01
Bus	0.00	0.00	0.15	0.56	0.03	0.06	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.09	0.00	0.00
Car	0.00	0.03	0.05	0.01	0.38	0.07	0.03	0.02	0.00	0.01	0.01	0.00	0.02	0.02	0.04	0.00	0.14	0.02	0.04	0.10	0.00
Castle	0.00	0.03	0.08	0.00	0.02	0.20	0.03	0.02	0.03	0.07	0.01	0.00	0.08	0.16	0.00	0.04	0.01	0.01	0.06	0.14	0.00
Coast	0.01	0.15	0.02	0.00	0.00	0.03	0.29	0.02	0.01	0.08	0.00	0.01	0.02	0.05	0.02	0.05	0.03	0.11	0.04	0.04	0.01
Desert	0.02	0.02	0.05	0.02	0.03	0.00	0.03	0.17	0.02	0.03	0.02	0.03	0.25	0.05	0.05	0.03	0.03	0.03	0.02	0.11	0.02
Harbor	0.00	0.21	0.18	0.06	0.03	0.03	0.06	0.00	0.15	0.00	0.00	0.00	0.00	0.12	0.00	0.03	0.03	0.03	0.03	0.03	0.00
Mountain	0.01	0.02	0.01	0.00	0.00	0.03	0.13	0.00	0.00	0.32	0.00	0.02	0.06	0.03	0.01	0.09	0.00	0.12	0.03	0.07	0.04
Night	0.00	0.00	0.00	0.00	0.04	0.04	0.00	0.00	0.09	0.00	0.52	0.00	0.04	0.00	0.00	0.00	0.17	0.04	0.00	0.04	0.00
Road	0.00	0.00	0.00	0.00	0.00	0.09	0.12	0.15	0.00	0.00	0.00	0.03	0.09	0.00	0.00	0.09	0.03	0.00	0.12	0.26	0.03
Rock	0.00	0.03	0.00	0.00	0.01	0.03	0.01	0.10	0.00	0.05	0.03	0.00	0.47	0.09	0.01	0.02	0.01	0.02	0.00	0.10	0.01
Ruin	0.01	0.02	0.07	0.01	0.01	0.15	0.09	0.02	0.01	0.04	0.00	0.01	0.09	0.17	0.00	0.02	0.01	0.02	0.03	0.17	0.03
Sky	0.02	0.04	0.00	0.00	0.02	0.02	0.07	0.04	0.01	0.04	0.01	0.01	0.02	0.02	0.19	0.11	0.24	0.07	0.02	0.00	0.02
Snow	0.04	0.04	0.02	0.00	0.00	0.03	0.08	0.03	0.01	0.10	0.00	0.01	0.04	0.04	0.04	0.29	0.02	0.13	0.00	0.05	0.02
Sunset	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.05	0.03	0.02	0.03	0.00	0.05	0.00	0.13	0.00	0.62	0.02	0.02	0.00	0.00
Surfing	0.03	0.07	0.00	0.02	0.01	0.04	0.04	0.00	0.00	0.04	0.03	0.00	0.01	0.02	0.01	0.07	0.00	0.58	0.00	0.02	0.01
Train	0.00	0.11	0.09	0.05	0.01	0.12	0.03	0.02	0.01	0.00	0.01	0.02	0.01	0.07	0.00	0.02	0.00	0.02	0.30	0.09	0.01
Vegetation	0.01	0.00	0.00	0.00	0.01	0.08	0.02	0.01	0.00	0.01	0.00	0.00	0.08	0.08	0.00	0.04	0.00	0.02	0.05	0.58	0.01
Waterfall	0.00	0.03	0.00	0.00	0.00	0.03	0.03	0.00	0.00	0.09	0.00	0.00	0.06	0.03	0.00	0.21	0.00	0.03	0.00	0.21	0.27

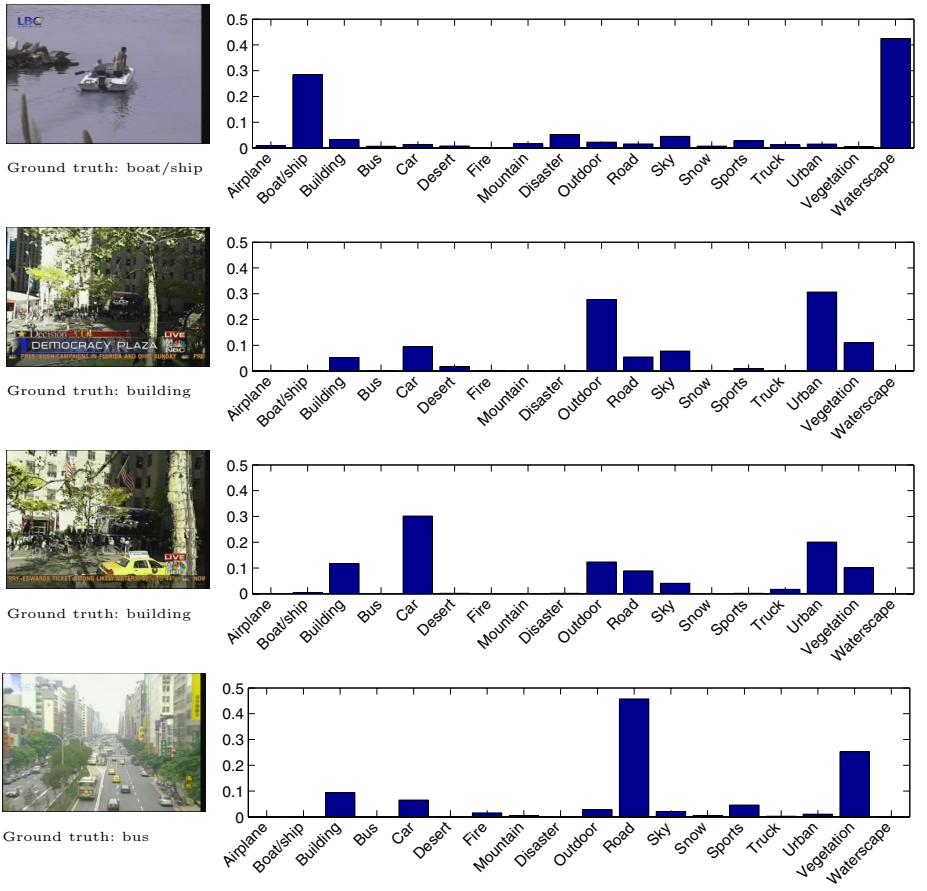
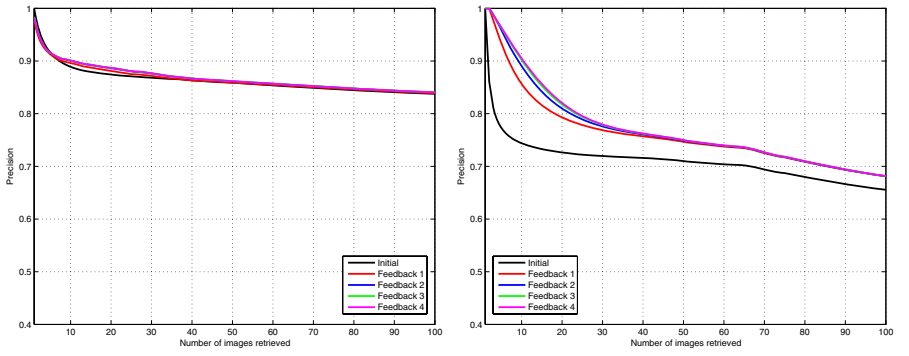


Fig. 3. Posterior probability examples. Although the ground truth class is not the one that received the highest posterior (i.e., counted as error in Table 1), the probability values are semantically correct

threshold is often applied to the posterior probabilities, and the concepts/classes with probabilities higher than this threshold are assigned to the image. To further improve the classification accuracy in an adaptive way regarding different user interests, we use the probabilities obtained from the Bayesian classifier as new features for each scene. This corresponds to a new representation in terms of a feature vector of length c for c classes.

This representation maps images to a semantic space where each component of the new feature vector corresponds to the probability of observing a particular concept/object in an image. Furthermore, the features complement each other and provide contextual information. For example, a car detector that uses low-level features may not be very reliable when used alone but the occurrence of a region that exhibits feature characteristics of a car becomes more relevant if that image also has a high probability of containing a road and having an



(a) TRECVID (b) Corel

Fig. 4. Precision vs. the number of images retrieved for the initial query and four feedback iterations

urban context. Therefore, this new representation compensates the limitations of individual object and scene detectors that use low-level features, and brings us one more step closer to bridging the semantic gap.

We use this representation in a retrieval framework. First, the user selects a category and the images that have a high probability of belonging to that category are retrieved. Next, the user selects one of these images as the query and an initial retrieval is performed.

The result of this query is a ranked list of the images in the database. If the user marks some of the images as relevant or irrelevant, this relevance feedback information can be used for further improvements. In this paper, we use the support vector data description (SVDD) model [11] as a one-class classifier. The one-class framework is intuitively applicable to the feedback problem because the images that are labeled as relevant provide a good sample for the target class of interest whereas the images that may be labeled as irrelevant often do not provide sufficient training data to learn a separate class. The SVDD model uses the relevant examples to learn a hypersphere that encloses the target class, and uses the irrelevant examples to minimize the volume of this sphere. Once the classifier is trained for a given iteration, the images can be re-ranked according to their distances to the resulting hypersphere. These results can also be used for annotating the query image using the most common concept/object labels among the images that fall into the hypersphere.

Experiments: The ground truth was used to automatically generate queries and provide feedback using the top 30 images by automatically labeling them as relevant or irrelevant at each iteration. Figure 4 shows the average precision results when the one-class Gaussian model was used for the Bayesian classifier. The first iteration gave the largest increase. Following iterations provided minor improvements. Although the classification accuracy was low when only the class with the highest posterior was considered, feedback iterations successfully

converged to the true set of images by making use of the concept/object occurrence probabilities as semantic features.

6 Summary

We described an image annotation and retrieval framework that used scene classification for extracting semantic features for image representation. Scene classification was performed using Bayesian modeling of histograms of regions segmented using clustering of color features and line structures. Given the observation that multiple classes can describe the image content, a new semantic representation was constructed by contextual modeling of images using the occurrence probabilities of concepts and objects. The experiments showed that this new representation provided an increased precision in an unconstrained data set with a large number of semantically overlapping classes.

References

1. Quelhas, P., Monay, F., Odobez, J.M., Gatica-Perez, D., Tuytelaars, T.: A thousand words in a scene. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(9), 1575–1589 (2007)
2. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition 37(9), 1757–1771 (2004)
3. van Gemert, J.C., Geusebroek, J., Veenman, C.J., Snoek, C.G.M., Smeulders, A.W.M.: Robust scene categorization by learning image statistics in context. In: CVPR (2006)
4. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. International Journal of Computer Vision 72(2), 133–157 (2007)
5. Li, Y., Shapiro, L.G., Bilmes, J.A.: A generative/discriminative learning algorithm for image classification. In: ICCV (2005)
6. Paclik, P., Duin, R.P.W., van Kempen, G.M.P., Kohlus, R.: Segmentation of multispectral images using the combined classifier approach. Image and Vision Computing 21(6), 473–482 (2003)
7. Li, Y., Shapiro, L.G.: Consistent line clusters for building recognition in CBIR. In: ICPR (2002)
8. Mojena, R.: Hierarchical grouping methods and stopping rules: An evaluation. The Computer Journal 20(4), 359–363 (1977)
9. Gokalp, D., Aksoy, S.: Scene classification using bag-of-regions representations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Beyond Patches Workshop, Minneapolis, Minnesota, June 23 (2007)
10. Tax, D.M.J.: One-Class Classification. PhD thesis, Delft University of Technology, Delft, The Netherlands (2001)
11. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Machine Learning 54(1), 45–66 (2004)