A Computationally Efficient Approach to Indoor/Outdoor Scene Classification

Navid Serrano[1], Andreas Savakis[2] and Jiebo Luo[1]

[1]*Electronic Imaging Products, R & D*
Eastman Kodak Company
{serrano,luo}@image.kodak.com

[2]*Department of Computer Engineering*
Rochester Institute of Technology
savakis@mail.rit.edu

Abstract

Prior research in scene classification has shown that high-level information can be inferred from low-level image features. Classification rates of roughly 90% have been reported using low-level features to predict indoor scenes vs. outdoor scenes. However, the high classification rates are often achieved by using computationally expensive, high-dimensional feature sets, thus limiting the practical implementation of such systems. We show that a more computationally efficient approach to indoor/outdoor classification can yield classification rates comparable to the best methods reported in the literature. A low complexity, low-dimensional feature set is used in conjunction with a two-stage Support Vector Machine classification scheme to achieve a classification rate of 90.2% on a large database of consumer photographs.

1. Introduction

Scene classification is important in a number of applications that deal with consumer photographs. Knowledge of the scene type is useful in event classification, which constitutes a fundamental component of automatic albuming systems [1]. Scene classification is also valuable in image retrieval from databases because it provides understanding of scene content that can be used along with color, texture, and shape for database browsing. The general problem of automatic scene categorization is difficult to solve and is best approached by a divide-and-conquer strategy. A good first step is to consider two classes such as indoor/outdoor [2-4] and then further subdivide into city/landscape [5], etc.

Scene classification is often approached by computing low-level features (e.g. color and texture) that are processed with a classifier engine for inferring high-level information about the image [2,5]. Another approach is to combine the low-level features with semantic scene content in order to improve the classification performance [3,4]. For example, in [4], a Bayesian network is used to integrate low-level color and texture features with semantic sky and grass features to improve classification performance. In fact, the low-level and semantic features have the greatest impact when combined.

The design of a practical scene classification system must address the issue of computational efficiency. Ideally, the computational efficiency should not compromise the accuracy. In [6], it was shown that a computationally efficient object detection method could match the classification rates of the best previous face detection systems. In the area of indoor/outdoor classification, however, high classification rates are typically achieved with a notable computational cost. For instance, in [2,4], the Multiresolution Simultaneous Autoregressive (MSAR) model is used to predict texture despite the computational cost. A k-nearest neighbor (k-nn) classifier is used to train and classify low-level features in [2,4], again representing a significant computational burden. Another common drawback is high feature dimensionality, as is the case in [5], where the feature set is on the order of 600 dimensions. Such issues obviously limit the practical use of these systems.

In this paper, we introduce an improved approach to indoor/outdoor classification using a low complexity, low-dimensional feature set while still achieving classification rates comparable to existing methods. The gains in efficiency are achieved by first, using wavelet [7] texture features, rather than the MSAR model, and second, by significantly reducing the feature dimensionality. High classification rates are achieved, despite the dimensional reduction, by using Support Vector Machines (SVMs) [8] in a two-stage classification scheme. The choice of SVMs is motivated by the fact that they have been shown to achieve equivalent or significantly lower error rates than comparative methods [8] and can, in theory, result in more efficient classification than the often-used k-nn classifier, for example. The proposed method was trained and tested on a database of 1200 consumer photographs.

2. Image Database

A database of 1200 consumer photographs collected by Kodak was used to train and test the indoor/outdoor classification performance. It is the same image database as the one used in [2], where we reduced the number of images from 1343 to 1200 by eliminating images with near duplicate scene content and/or ambiguous indoor/outdoor labeling. The removal of near duplicates can, in general,

1051-4651/02 $17.00 © 2002 IEEE

result in higher error rates, as will be shown later. The indoor and outdoor images are equally distributed in the set.

The images in the database are 36-bit color, 512 x 768 resolution scanned photographs. The preprocessing stage included quantization to 24-bit color, and a simple color balance that clipped the top and bottom 0.5% of each color channel, centered, and equalized the histogram. In addition, the images were subsampled to 256 x 384 pixels for increased processing speed.

3. Features

We employ low-level color and texture features. The features are extracted from image subblocks. We evaluated several subblock configurations and found that a 4 x 4 tessellation, as used in [2], yielded better results. The color and texture features used in our approach are analogous to those used in [2,4], yet possess roughly half the number of dimensions. Further reduction in computational complexity is achieved by using wavelet texture features instead of the computationally intensive MSAR features used in [2,4].

3.1. Color Features

A color space transformation is used to de-correlate the color channels in the original *RGB* image. We use the *LST* color space defined by:

$$L = \frac{k}{\sqrt{3}}(R+G+B)\,, S = \frac{k}{\sqrt{2}}(R-B)\,, T = \frac{k}{\sqrt{6}}(R-2G+B) \quad (1)$$

Where, L is the luminance channel, S and T are the chrominance channels, and k = 255/max$\{R, G, B\}$. This color space is akin to the Ohta color space with the exception of the scale factors. Once the image is transformed to *LST* space, a histogram is computed for each channel. The approach described in [2] used 32 bin histograms per channel. Instead, we use 16 bin histograms, thus reducing the dimensionality by one half. The three concatenated histograms compose a color feature vector of 48 dimensions. When the color features are computed for an entire image and trained using SVM, a classification rate of 74.5% is achieved. As a matter of comparison, the 96-dimensional color feature used in [2] achieved a classification rate of 74.2% using a k-nn classifier. In our final system, however, the color features (as well as the texture features) are computed on image subblocks and classified independently.

3.2. Texture Features

The texture features are obtained from a two-level wavelet decomposition. The decomposition is performed on the L-channel using Daubechies' 4-tap filters [7], where $h(n) = [-0.129 \;\; 0.224 \;\; 0.837 \;\; 0.483]$ is the low-pass filter and $g(n) = [-0.483 \;\; 0.837 \;\; -0.224 \;\; -0.129]$ is the high-pass filter. In a two-dimensional separable wavelet implementation, the filters are applied to the rows and columns of the image separately, where $LL(i,j) = h(i)h(j)$ are the approximation coefficients, $LH(i,j) = h(i)g(j)$, $HL(i,j) = g(i)h(j)$, and $HH(i,j) = g(i)g(j)$ are the detail coefficients, and i and j, denote the row and column image coordinates.

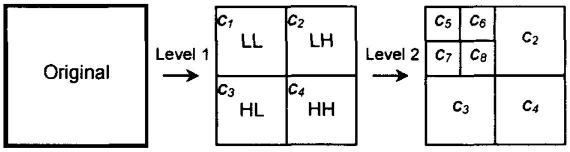

Figure 1. Two-level wavelet pyramid structure.

Let c_2, c_3, c_4, c_5, c_6, c_7, and c_8 represent the subband coefficients of the two-level wavelet decomposition as shown in Figure 1. The texture features are obtained by first filtering the low-frequency coefficients c_5 using the Laplacian filter and second, obtaining a measure of the subband energy for all wavelet coefficients according to:

$$e_k = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left|c_k(i,j)\right|^2 \,, \quad k = 2,3,\ldots 4K \qquad (2)$$

Where, M and N are the image dimensions of coefficient c_k, and K is the number of decomposition levels (in this case 2). The use of wavelets results in increased computational efficiency (see section 5), as well as reduced dimensionality, as only 7 texture features are used compared to the 15 MSAR features used in [2]—again, roughly half the dimensionality. When the wavelet texture features are computed on the entire image and trained with SVM, a classification rate of 83.0% is achieved, compared to 82.2% using MSAR texture features and a k-nn classifier [2].

4. Classification

We employ a two-stage classification approach using Support Vector Machines (SVMs). An SVM can be used to learn various representations such as neural networks, polynomial estimators, etc. while achieving excellent generalization performance [8]. The SVMs described here were trained using a radial basis function (RBF) representation. The first stage involves training color and texture SVMs based on image subblocks. The block-based classification rates will be lower than for the entire image. However, in the second stage, another SVM is used to interpret the color and texture classification results and yield a high accuracy, final indoor/outdoor classification. The two-stage approach proposed in this paper is shown graphically in Figure 2.

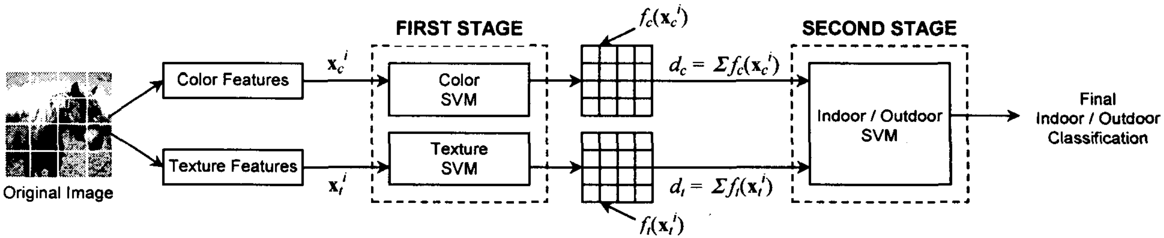

Figure 2. Graphic illustration of the two-stage indoor/outdoor classification approach

4.1. First Classification Stage

The first stage involves the classification of the low-level color and texture features described in section 3. Let $\mathbf{x}_c^i$ and $\mathbf{x}_t^i \in R^d$ be the color and texture feature vectors, respectively, corresponding to image subblock i. Two SVMs are trained for the color and texture features independently. Each SVM is trained on subblocks of the tessellated image. The database of 1200 images was divided into independent training and testing sets (each consisting of 600 images) with an equal number of indoor and outdoor scenes. The 16 subblocks for each of the 600 images were trained and tested separately. The resulting classification rates for the block-based color and texture SVMs are reported in Table 1.

Table 1. Subblock classification results after first stage

Feature	Training Set	Test Set
Color	73.7%	67.6%
Texture	75.8%	73.0%

As stated in sections 3.1 and 3.2, computing the color and texture features over the entire image results in classification rates of 74.5% and 83.0%, respectively. As shown in Table 1, the block-based results are quite lower, which is to be expected because there are fewer and weaker signatures in image subsections. In comparing the results of Table 1 with those obtained in [2], we find that they obtained block-based classification rates of 70.3% for color features and 74.7% for MSAR texture features. These figures are slightly higher than those obtained on the test image set, as shown in Table 1. However, as stated in section 3, our features have half the dimensionality of those used in [2].

4.2. Second Classification Stage

In the first stage, given a color feature vector $\mathbf{x}_c^i$ the color SVM produces a value $f_c(\mathbf{x}_c^i)$, which is a measure of the distance from $\mathbf{x}_c^i$ to the separating hyperplane—the trained color SVM decision boundary in feature space. A point $\mathbf{x}_c^i$ that lies on the decision boundary has a value of zero, while positive distances correspond to indoor scenes, and negative distances correspond to outdoor scenes. A large positive distance indicates that a point (subblock) $\mathbf{x}_c^i$ has strong indoor cues, whereas a large negative distance indicates that $f_c(\mathbf{x}_c^i)$ has strong outdoor cues. An analogous value $f_t(\mathbf{x}_t^i)$ is also produced by the texture SVM. Therefore, $f_c(\mathbf{x}_c^i)$ and $f_t(\mathbf{x}_t^i)$ can be used to describe the likelihood that a particular image region and, in turn, the image itself, is indoor vs. outdoor.

Because $f_c(\mathbf{x}_c^i)$ and $f_t(\mathbf{x}_t^i)$ represent a distance measure, the values can be summed to obtain a distance feature (for color and texture each) corresponding to the entire image. Hence, we define two new features:

$$d_c = \sum_{i=1}^{16} f_c(\mathbf{x}_c^i) \tag{3}$$

$$d_t = \sum_{i=1}^{16} f_t(\mathbf{x}_t^i) \tag{4}$$

In the second classification stage, we train a new SVM using d_c and d_t as global color and texture features, respectively, for the entire image. The training procedure can be summarized as follows. After training the color and texture SVM in the first stage, the resulting distances $f_c(\mathbf{x}_c^i)$ and $f_t(\mathbf{x}_t^i)$ are collected for the training set of images and combined to obtain the d_c and d_t values for each image. These values are used to train a new SVM, which will produce the final indoor/outdoor classification. The resulting indoor/outdoor classification rates for the 600 images in the independent test set are shown in Table 2. As a point of comparison, we tested the majority classifier suggested in [2] on the same image set. In this approach a binary label (1 = Outdoor, 0 = Indoor) is assigned to each block and the final indoor/outdoor classification is decided by majority vote.

As can be seen from Table 2, the second stage SVM yields a 3% accuracy increase on the independent test set compared to the majority classification suggested in [2]. One important advantage our method clearly has is that ambiguous blocks with a borderline indoor/outdoor belief would not affect the final decision as opposed to the forced binary labels.

Table 2. Final classification results after second stage

Classifier	Training Set	Test set
Majority classifier	92.8%	87.2%
Second stage SVM	95.0%	90.2%

5. Discussion

In the previous sections, we described a computationally efficient approach to indoor/outdoor classification that yields a classification rate of 90.2% when training and testing on a database of 1200 consumer photographs. We now would like to put these results in context by comparing them with those of existing methods, and noting the gains in computational efficiency.

A slightly different two-stage approach was introduced in [2] and a final indoor/outdoor accuracy of 90.3% was reported on a superset of the image database used here. Because the image database used in [2] was composed of consumer photographs shot on film rolls, some of the images actually contain near duplicate scene content. When a *k*-nn classifier is used, as in [2], duplicate scenes actually facilitate classification. As expected, after retraining and testing the method of [2] without duplicate scenes, the classification rate dropped to about 85%. Hence, higher classification rates are achieved using our method than the method of [2] on essentially the same image database. Furthermore, the MSAR texture features used in [2] are more computationally elaborate than the wavelet texture features we propose. To highlight this point, computation (on a Sun Ultra 5) of the MSAR features on an image of comparable size to those in our database required 194 seconds compared to 0.3 seconds for the wavelet features.

The use of an SVM as opposed to a *k*-nn classifier [2] also provides added computational efficiency. Whereas a *k*-nn classifier must scan the entire training space (equal to the number of training samples) to classify a given image, the number of points in the SVM training space is equal to the number of support vectors [8] (typically less than the number of training samples). In fact, the color and texture SVMs described in section 4.1 represent a combined 33% decrease in training vectors compared to a *k*-nn classifier.

In [5], a distinct approach to indoor/outdoor scene classification is proposed. Low-level features (color moments) are calculated on image subsections, concatenated into a single feature vector, and trained with a Bayesian classifier. Indoor/outdoor classification rates of 88.2% and 88.7% are reported on two different test sets containing 2540 and 1850 images, respectively. Both rates are close, though lower than the 90.2% that we achieve. The image sets used in [5] are not related to ours and contain mostly professional stock photos such as those in the Corel data set. In general, there is less variation in professional photos because greater care is given to color and composition, and as a result, simple color features alone, as in [5], may provide sufficient generalization. In addition, a major drawback of the method described in [5] is the feature dimensionality of 600, which is challenging to manage, especially in classification.

6. Conclusions

We have shown that a low complexity, low dimensional feature set can, in fact, be used to achieve a high indoor/outdoor classification rate when applied in a two-stage SVM classification scheme. Our proposed method achieves a success rate of 90.2%, which compares favorably with the best existing indoor/outdoor scene classification methods. Furthermore, we envision this low-level indoor/outdoor classification system being used in conjunction with semantic scene features, such as the presence of sky and grass. It is our hope to test our low-level feature approach in conjunction with semantic features. The combination of low-level and semantic features using, for instance, a Bayesian network as in [4] may conceivably increase classification accuracy beyond our reported 90.2%.

7. References

[1] A. C. Loui and A. E. Savakis, "Automatic Image Event Segmentation and Quality Screening for Albuming Applications," *Proc. Int. Conf. Multimedia and Expo*, New York, NY, 2000.

[2] M. Szummer and R. W. Picard, "Indoor-Outdoor Image Classification", *IEEE International Workshop on Content-Based Access of Image and Video Databases, ICCV '98*, 1998.

[3] S. Paek, et al, "Integration Of Visual and Text Based Approaches for the Content Labeling and Classification of Photographs," ACM SIGIR '99 Workshop on Multimedia Indexing and Retrieval, Berkeley, CA, August 19, 1999.

[4] J. Luo and A. Savakis, "Indoor vs. Outdoor Classification of Consumer Photographs Using Low-Level and Semantic Features," *Proc. Int. Conf. Image Process.*, Thessaloniki, Greece, 2001.

[5] A. Vailaya, M. A. T. Figuereido, A. K. Jain and H. J. Zhang, "Image Classification for Content-Based Indexing," *IEEE Trans. Image Process.*, vol. 10, pp. 117-130, January 2001.

[6] P. Viola and M. Jones, "Robust Real-Time Object Detection," Second Int. Workshop on Statistical and Computational Theories of Vision-Modeling, Computing, and Sampling, Vancouver, Canada, July 2001.

[7] I. Daubechies, *Ten Lectures on Wavelets*, SIAM Publications, Philadelphia, 1992.

[8] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining And Knowledge Discovery*, vol. 2, pp. 1-43, 1998.