massachusetts institute of technology — artificial intelligence laboratory

Multiple Resolution Image Classification

Jake V. Bouvrie

AI Memo 2002-022 December 2002

Abstract

Binary image classification is a problem that has received much attention in recent years. In this paper we evaluate a selection of popular techniques in an effort to find a feature set/classifier combination which generalizes well to full resolution image data. We then apply that system to images at one-half through one-sixteenth resolution, and consider the corresponding error rates. In addition, we further observe generalization performance as it depends on the number of training images, and lastly, compare the system's best error rates to that of a human performing an identical classification task given the same set of test images.

1 Introduction

The classification of images into one of several categories is a problem which arises naturally under a wide range of circumstances. Tumor diagnosis [1], photographic developing, and visual image querying have each received considerable attention. Binary classification of images based on scene information is one particular case where there is no definitive language with which to describe scenes depicted in an image. Various attempts have been made to extract global [3] and local [2] cues from an image, in an effort to construct features that can accurately capture the essence of a scene type. Classification performance thus hinges mainly on the accuracy and reliability of the feature set in representing the various scene classes.

In the vast majority of prior work, it is assumed that images are available at a suitably high resolution. Low resolution representations on the other hand reveal significant additional distortions. Harsh negative artifacts might be unavoidable in some applications, where storage or capture constraints enforce strict scaling/compression requirements. Thus, features which might have done well in the presence of ample scene information, may prove to be no longer suitable at very low resolutions.

It is this question that is the focus of the paper. We attempt to classify indoor-outdoor scenes, given grayscale image data only. Using features which work well on full resolution images, we observe classification performance as images are scaled downwards linearly to 1/16th resolution in the worst case. We additionally make comparisons to human classification accuracy over a similar set of resolutions, using an identical collection of test images. And lastly we evaluate system performance as it depends on training set size.

2 Classifier

For purposes of comparison, two classifiers were chosen to discriminate between indoor and outdoor images on the basis of a vector of features. The Support Vector Machine [4] was utilized in order to achieve high generalization performance. SVMs with both polynomial and Gaussian kernels were investigated. This allowed for exploration of higher order interactions amongst the existing features without having to explicitly compute additional feature vector elements. Gaussian kernels give radial basis function classifiers, where the centers, weights, and threshold are provided implicitly [5]. SVMs with Gaussian RBF kernels have been shown to be superior compared to RBF networks trained using classical methods (e.g. K-means clustering plus error backpropagation) [6]. In addition, new examples are compared to only those training points which double as support vectors, rendering a speedy verdict. To locate optimal kernel parameters, the SVM was separately trained in the polynomial case using orders $n = 1, \ldots, 20$, while for Gaussian kernels the variance was located using a heuristic search method not unlike bisection.

A K-nearest neighbors classifier was also examined, as a basis of comparison for the SVMs and also because it is popular in the literature. At each resolution, KNN performance was evaluated using two or three of the following distance metrics:

$$\begin{aligned} d_1(\mathbf{x},\mathbf{y}) &= \|\mathbf{x}-\mathbf{y}\|_1 \\ d_2(\mathbf{x},\mathbf{y}) &= \|\mathbf{x}-\mathbf{y}\|_2^2 \\ d_h(\mathbf{x},\mathbf{y}) &= \sum_{i=1}^{N}(x_i - \min(x_i, y_i)) \end{aligned}$$

Where we have respectively the sum of the absolute differences, the squared Euclidean length, and the histogram intersection norm proposed in [7] and adopted by [2]. A distance resembling Pearson's χ^2 statistic is used in [5], however that metric gave exceptionally poor performance during initial trials and was subsequently omitted from further comparisons.

3 Features

3.1 Image Database

The images used in this experiment were drawn primarily from the Corel stock photo collection. Three hundred indoor and outdoor images each were hand picked from 200 possible pre-labeled categories and further subdivided into training and testing sets of size 460 and 140 images respectively (with equal proportions of each scene category). The full resolution images were 80-by-120 pixels. Sample images are shown in Figure 1.

3.2 Image Features

The extraction of meaningful features forms the key difficulty in classifying images according to scene. In addition to reducing the dimensionality of the feature vectors, it was our goal to reduce redundant information and noise present in the images while imposing translational, rotational and scale invariance.

3.2.1 Wavelet Decomposition

In an effort to form a lower dimension representation, approximation coefficients were taken from the 2D wavelet decomposition [8]. Experiments were conducted using the Haar "mother wavelet" and the 9/7 biorthogonal wavelet (also used in the JPEG2000 specification).

3.2.2 Histograms

Global graylevel features were captured by way of grayscale histograms using 32,16,8,6,4,3 and 2 bins. The histograms are naturally invariant to translations and rotations in the image.

3.2.3 DCT/DFT Corner

Following work done in [2], frequency information was collected by taking the 2D DCT of the 2D DFT magnitude. The DFT is shift invariant and reveals periodic textures in an image (e.g. grass, ocean, walls), while the DCT combines related frequencies into one value and conveniently focuses energy into the top left corner of the resultant image. Triangles with side lengths $n = 2, 4, 8, 12$ were taken from the top left and "flattened" into feature vectors with length $\frac{n(n+1)}{2}$.

3.2.4 Edge Direction Histograms

To represent shape information present in an image, histograms of the edge directions were formed [9]. Edges were detected using either the Prewitt or the Sobel operator, with no prior Gaussian smoothing applied to the image. Sobel filters performed slightly better and were thus chosen for the final system. The fact that Sobel's approximation rendered an improvement over Prewitt's is not completely alarming in light of the fact that Sobel kernels are generally more isotropic in their response.

For either choice, if h_y and $h_x = h_y^T$ represent the horizontal and vertical convolution kernels, and we apply them to an image to get g_y and g_x respectively, then the edge directions are defined as $\phi \equiv \arctan(-g_y/g_x)$. Quantization into 72 bins of 5° each is recommended by [9], however we observe results while employing 16,8, and 4 bins as well. Eight bins corresponding to 45° levels proved to be by far the most effective, as we shall see below.

As pointed out by [9], edge direction histograms are translation invariant with respect to objects in an image. Scale invariance is achieved by normalizing the histogram by the number of edge points in the image of interest. To reduce rotational effects in an image, [9] further advises a smoothing scheme which amounts to a non-causal moving average. Smoothing was indeed carried out, however it proved to hamper rather than enhance system performance in some cases. This might be explained by the fact that indoor and outdoor images simply do not have many objects common to a class which might be subject to rotation across images. The final configuration does however attempt to impose this additional rotational invariance anyhow.

3.3 Scaling Methods

At each resolution, training and testing images alike were scaled down and then subsequently enlarged to the original resolution using one of nearest-neighbor, bilinear, or bicubic interpolation. This was done so as to extract features through a process identical to that of the full resolution case. The results presented below show that performance at some resolutions was

Figure 1: Sample indoor/outdoor images.

significantly better given one particular interpolation method over the others; a simple nearest-neighbor approach did not always yield the best performance. An illustration of the various scaling methods in action is shown in Figure 2.

Wavelet	*SVM*	*KNN:* d_1	*KNN:* d_2
9/7	0.48	0.32	**0.31**
Haar	0.42	0.37	0.37

Table 1: Wavlet Classification Error

Bins	*SVM*	*KNN:* d_h
32	0.31	0.31
16	0.31	0.30
8	0.31	0.28
4	0.26	0.26
2	**0.26**	0.34

Table 2: Histogram Classification Error

4 Full Resolution Feature Performance

Summarized in Table 1, performance given Wavelet approximation coefficients is not particularly encouraging. Graylevel histograms (without equalization) also perform rather poorly on their own, as shown in Table 2. KNN with distance metrics d_1, d_2 performed poorly and is not included. It is interesting to note, however, that even with two histogram bins the SVM is able to do substantially better than chance.

Classification with DFT/DCT features worked rather well, achieving less than 20% error in most cases. Results are given in Table 3. Edge direction histograms showed slight improvement over the DFT/DCT corners, and performance is noted in Table 4. Edge direction results were generated using the Sobel operator which generally performed better than Prewitt's approximation.

Given the good generalization of SVMs with edge direction and DFT/DCT features, the final system was based on a combination of the two with parameter settings equal to the best performing value for the individual case (DCT/DFT corner edge length 12, 8 edge direction bins). System performance based on selections of the edge finding filter and histogram smoothing function is shown in Table 5. The addition of additional features did not improve performance, and in some cases made the error rate substantially worse. The system thus utilized only DFT/DCT features and edge direction histograms. It was additionally determined that the number of misclassifications was the same or very nearly so for both polynomial and Gaussian SVM kernels, given optimal kernel parameter settings. Given the parameter search methods described above, the search for an optimal polynomial kernel order is normally faster than locating a good Gaussian kernel variance. The final implementation therefore used a polynomial kernel, and searched for an optimal order $1 \leq n \leq 20$.

5 Low Resolution Performance

The system detailed above was trained given images at each resolution $1, \ldots, 1/16$, and for each scal-

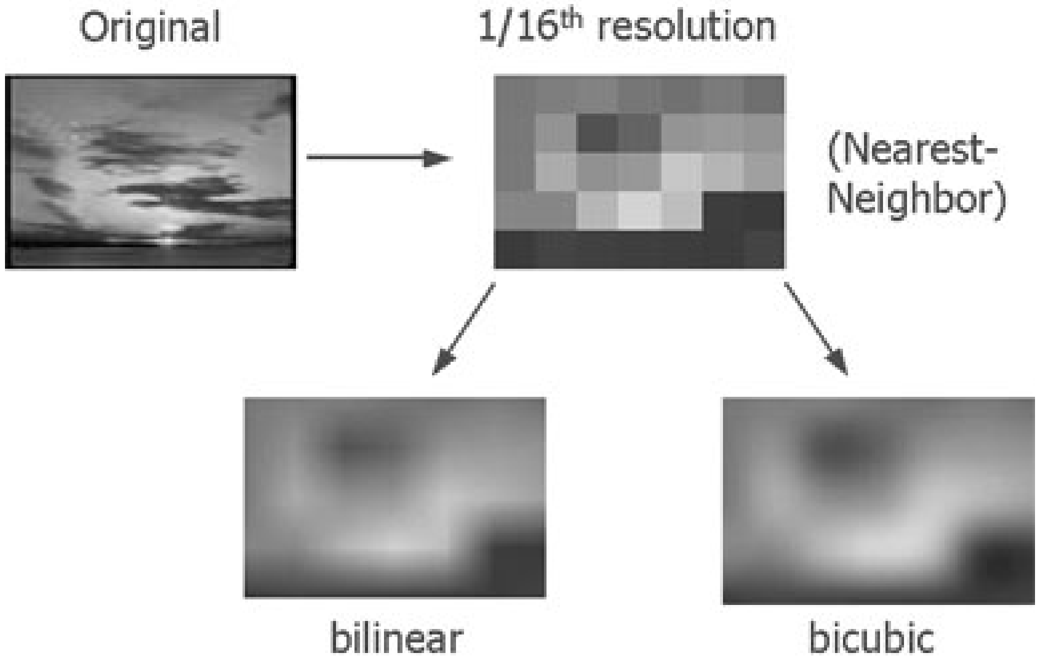

Figure 2: Scaling Methods

ing method. At each scaling/resolution combination, KNN performance was also computed using the two best performing distance metrics d_2, and d_h for comparison. The combination yielding minimal error was then selected as the final classifier for the corresponding resolution.

Results for each of the three interpolation methods are shown in Figure 3.

Corner Size	*SVM*	*KNN: d_2*
16	0.18	0.24
12	**0.17**	0.27
8	0.18	0.31
4	0.19	0.29
2	0.29	0.26

Table 3: DFT/DCT Classification Error

Bins	*SVM*	*KNN (metric)*
100	0.15	0.165 (d_h)
72	0.15	0.136 (d_2)
32	0.143	0.18 (d_2)
16	0.143	0.21 (d_h)
8	**0.136**	0.21 (d_h)
4	0.17	0.17 (d_2)

Table 4: Edge Direction Histogram Classification Error

SVM Error	*Edge Filter/Smoothing*
0.129	(Prewitt/None)
0.114	(Sobel/None)
0.107	(Prewitt/Smoothing)
0.100	(Sobel/Smoothing)

Table 5: Combined Feature Classification Error

5.1 Man vs. Machine

The set of test images used to gauge classification performance was chosen to be identical to that used by [10] in their study of human classification accuracy at resolutions following an exponential curve: $1, 1/2, 1/4, 1/8, 1/16$. Five images from each category which were possibly ambiguous to a human viewer were thrown out, and the remaining 140 indoor/outdoor images retained. But it was not expected that this would have a significant effect on performance. A comparison of the machine classification system at its best versus human performance observed in [10] is shown in Figure 4. The point at which both perform similarly is roughly one-quarter resolution, while the system's performance is far superior for resolutions below this threshold. Human vision at full resolution is still significantly better than this machine implementation.

5.2 Training Set Size

Test error was also observed as a function of the training set size. Selecting four resolutions, we have noted best possible system performance as the training set is reduced to as low as 5 images. As the number of images is decreased, the error does not correspond monotonically. This indicates that the error is highly sensitive to which particular 5 images the system is trained upon. If we are training the system on n images, then a better generalization estimate might be based on the mean performance over all $\binom{460}{n}$ possible image selections. For the purposes of this experiment however, averaging was not pursued.

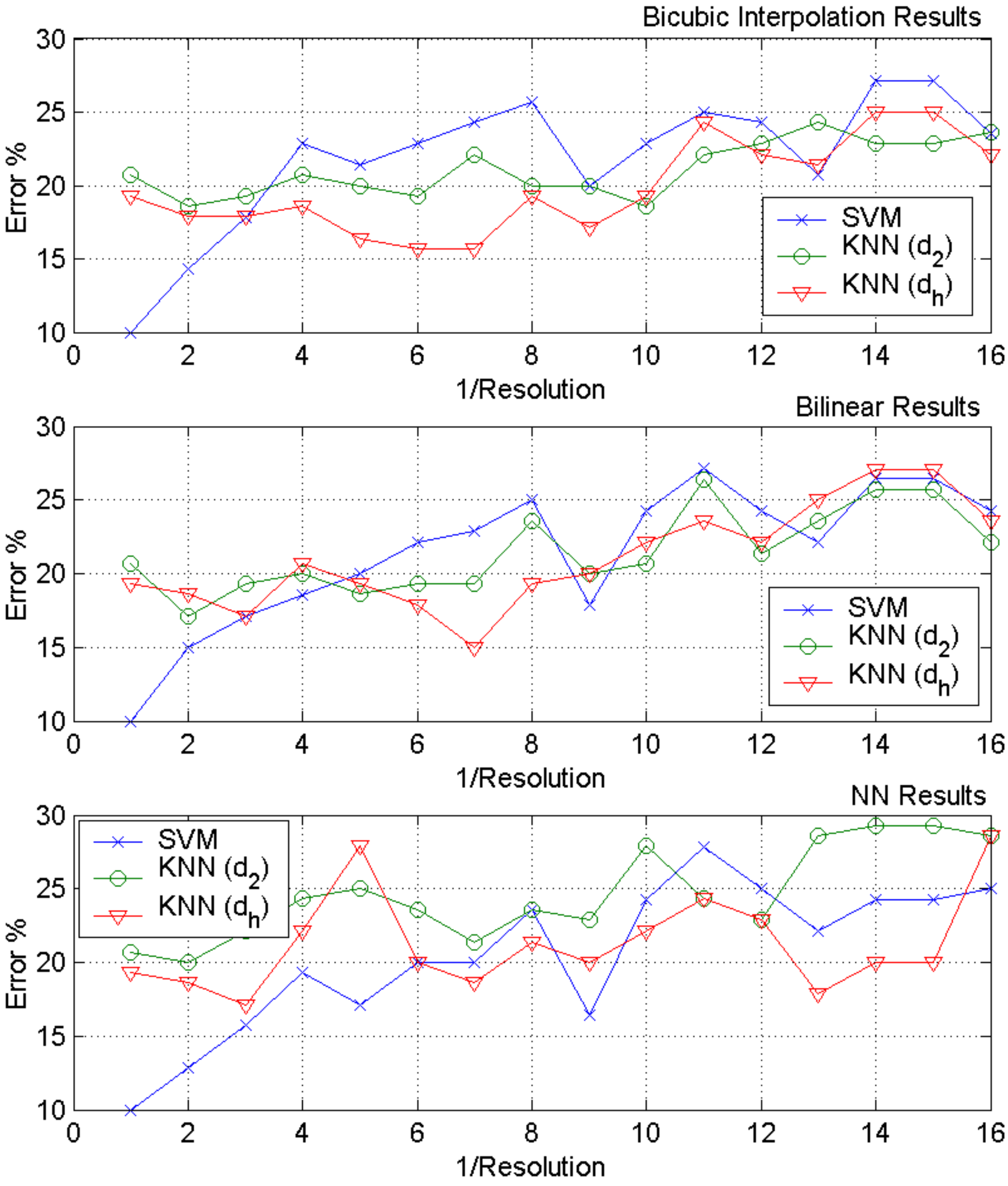

Figure 3: Scaling Results.

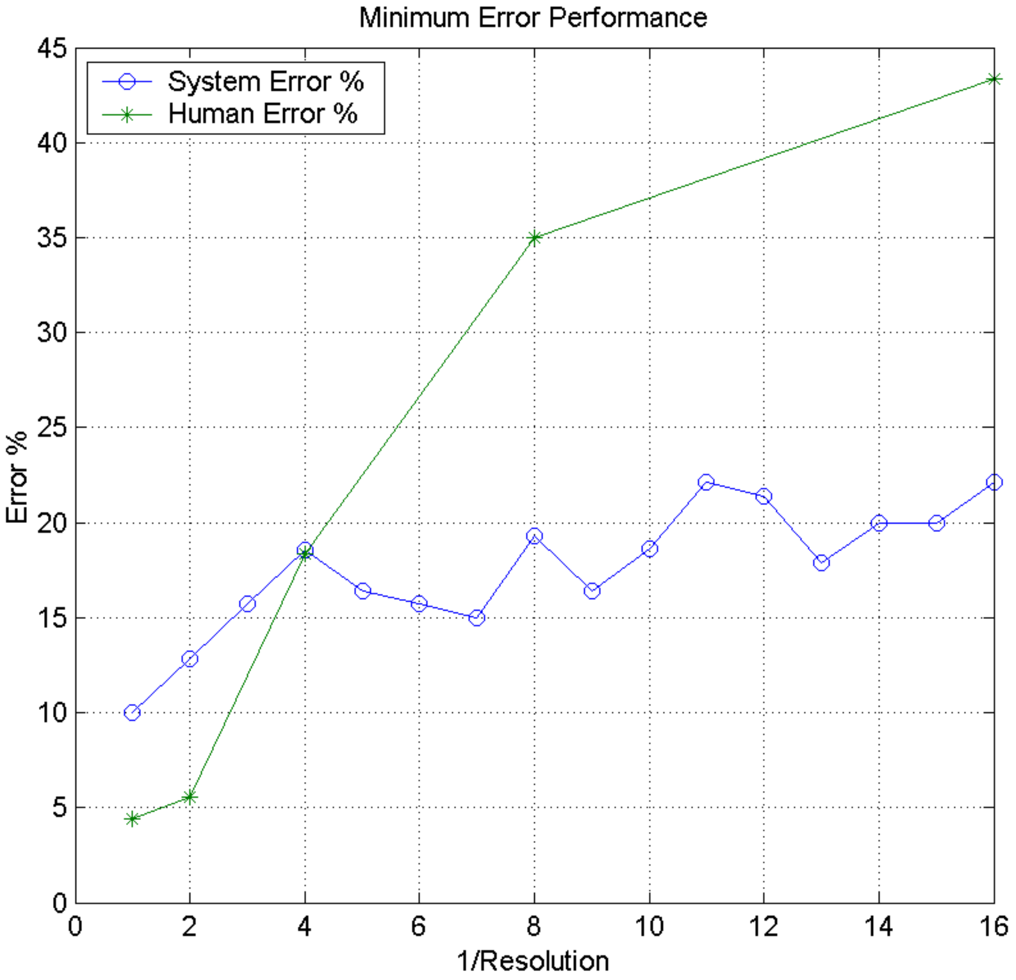

Figure 4: Human Error Vs. System Error.

6 Conclusions

6.1 Findings

We have evaluated the performance of a binary image classifier given low resolution training and testing data. At best, the system is able to achieve a 10% misclassification rate given full resolution images, to roughly 22% error with one-sixteenth resolution images. Compared to human performance given a nearly identical test set, the system vastly outperforms human discrimination ability at resolutions 1/4 and lower. We additionally showed that the system's dependence on training set size does not progress monotonically as the set is decreased, and that even with as low as 5 or 10 training images the system is able to capture meaningful structure.

6.2 Problems

Compared to similar studies, this experiment made use of relatively small training and test sets. Images numbering in the thousands would have given a more accurate estimate of test performance, and could relieve overfitting. The process by which features are computed is also computationally taxing. Extracting edges, computing DCT's and generating histograms is by no means a "fast" process. Preparing features and training the SVM comprises the bulk of the computational burden.

Previous work achieved slightly better error rates: [2] produced 9.7% error at best on a nearly identical classification task. [9] achieved 6% error given city/landscape images (with color information), and 5.5% error a for mountain/forest versus sunrise/sunset classification task (also with color).

6.3 Possible Improvements

It is likely that performance can be increased further given the incorporation of additional features. Pixels which have high mutual information with the labels [11] is one selection method with promise. Other methods with potential include Non-Negative Matrix Factorization [12] and oriented bandpass filtering [13].

References

[1] G. Yeo and T. Poggio. *Multiclass Classification of SRBCTs.* MIT AI Memo No. 2001-018, CBCL Memo No. 206, 2001.

[2] M. Szummer and R.W. Picard. *Indoor-Outdoor Image Classification.* MIT Media Laboratory Perceptual Computing Section Technical Report No. 445, 1998.

[3] A. Torralba and P. Sinha. *Recognizing Indoor Scenes.* MIT AI Memo No. 2001-015, CBCL Memo No. 202, 2001.

[4] V. Vapnik, *Statistical Learning Theory.* John Wiley, New York, 1998.

[5] O. Chapelle, P. Haffner, V. Vapnik. *SVMs for Histogram-Based Image Classification.* Submitted to IEEE Transactions on Neural Networks, 1999.

[6] B. Scholkopf, K. Sung, et. al. *Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers.* MIT AI Memo No. 1599, CBCL Paper No. 142, 1996.

[7] M. Swain and D. Ballard. *Indexing via Color Histograms.* International Journal of Computer Vision, vol. 7:11-32, 1991.

[8] G. Strang and T. Nguyen. *Wavelets and Filterbanks.* Wellesley-Cambridge Press, Wellesley, MA, 1996.

[9] A.K. Jain and A. Vailaya. *Shape-Based Retrieval: A Case Study With Trademark Image Databases.* Pattern Recognition, Vol. 31, No. 9, pp.1369-1390, 1998.

[10] J. Bouvrie and E. Conwell. *Human Classification of Indoor-Outdoor Images.* MIT 9.63 Report No.3, 2001.

[11] D. Koller and M. Sahami. *Toward Optimal Feature Selection.* Proc. ICML, 1996.

[12] D.D. Lee and H.S. Seung. *Learning the Parts of Objects by Non-Negative Matrix Factorization.* Nature 401, 788-791, 1999.

[13] M.M. Gorkani and R.W. Picard. *Texture Orientation For Sorting Photos At A Glance* Proceedings, International Conference on Pattern Recognition, Jerusalem, Vol I, 459-464, 1994.