Overcomplete ICA-based Manmade Scene Classification

Matthew Boutell
Department of Computer Science
University of Rochester
boutell@cs.rochester.edu

Jiebo Luo
Research and Development Laboratories
Eastman Kodak Company
jiebo.luo@kodak.com

Abstract

Principal Component Analysis (PCA) has been widely used to extract features for pattern recognition problems such as object recognition. Oliva and Torralba used "spatial envelope" properties derived from PCA to classify images as manmade or natural. While our implementation closely matched theirs in accuracy on a similar (Corel) dataset, we found that consumer photos, which are far less constrained in content and imaging conditions, present a greater challenge for the algorithm (as is typical in image understanding). We present an alternative approach to more robust naturalness classification, using overcomplete Independent Components Analysis (ICA) directly on the Fourier-transformed image to derive sparse representations as more effective features for classification. We demonstrated that our ICA-based features are superior to the PCA-based features on a large set of consumer photographs.

1. Introduction

Semantic classification of photographs, e.g., indoor vs. outdoor, manmade vs. natural, beach vs. desert, is a difficult problem studied extensively in the past decade [2]. The ability to classify scenes as manmade vs. natural [3,7,8] is useful as part of a hierarchical system of binary classifiers [13], which can be easier to design and more accurate than multi-class ones. Such a system can aid in content-based image retrieval, organization, and enhancement.

We implemented Oliva and Torralba's algorithm [8] for describing images in terms of their "spatial envelope" properties, including *openness* and *naturalness*. This algorithm extracts features from the image's power spectrum by convolving it with Gabor-like filters at 12 orientations and 5 scales. Because many filters are needed to cover the spectrum, they extract only the first 16 principal components of the images as determined by performing PCA on the training set.

Out implementation closely matched the original algorithm in the reported manmade-natural classification accuracy on a similar data set, composed of Corel professional stock photos [8]. However, the accuracy on a set of home photos dropped greatly (over 15%), because these photographs vary more in content and viewpoint, and some are of lower quality. We encountered this drop regardless of whether the training set was composed of home photos, stock photos, or a combination of both.

Figure 1 shows two consumer images misclassified by the PCA-based algorithm, along with the power spectrum of each image. The manmade image is typical in that it contains groups of edges aligned in the same direction, causing visible narrow "spikes" in the frequency domain, while the edges in the natural image vary in direction, and thus produce no such spike.

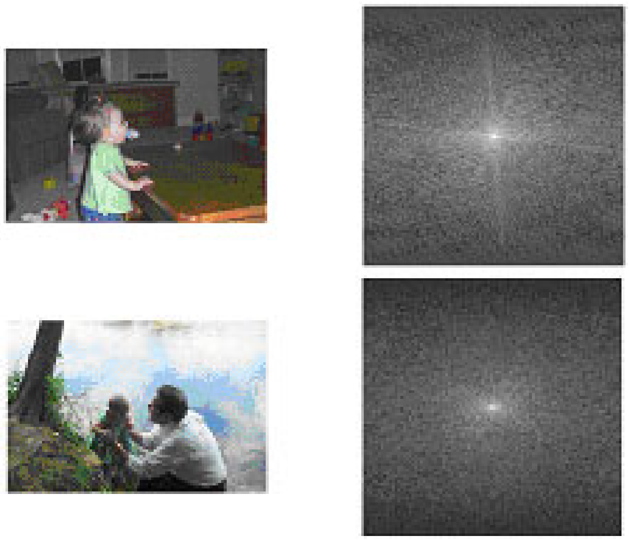

Figure 1: Typical manmade and natural consumer snapshots. Note that spikes occur in the power spectrum of the manmade image only.

It does not appear that the PCA-based classifier could learn these spike patterns. We hypothesized that much of the discriminating information was lost in sampling the power spectrum. First, Gabor filter-based sampling of the power spectrum does not explicitly capture the correlation between edges in a single direction; in particular, if the spike lies between the directions in which the image was sampled, the spike will be missed entirely. Second, even if the features did capture this correlation, it would require a large amount of training data to learn it simply because edges could line up in any arbitrary direction. Third, even if the correlations could be learned, it is possible that the principal components would not preserve them because PCA is computed over an ensemble of images, in which the majority of manmade images contain edges aligned along vertical and horizontal directions, and thus overwhelm the less frequent cases where edges are off the two main axes, e.g., Figure 1(a). Consequently, we found that the algorithm in [8] tends to fail mostly (1) when the linear structures in the images deviate from the horizontal and vertical directions (e.g., due to perspective distortion), (2) when there are a large number of edge directions present in an image (e.g., the first example in Figure 4), and (3) when edges are not as discernable (e.g., distant shots of a city scene).

We next present an alternative approach to feature extraction, using Independent Components Analysis (ICA) directly on the power spectrum. We then compare the classifiers using our ICA-based features with the PCA features-based classifier in [8]. We conclude with directions for future work.

2. Independent Components Analysis

In the Fourier domain, one can characterize scene content as a mixture of Gaussians (moG). Manmade structures, typically characterized by many edges clustered in a few well-defined directions, appear as sparse, thin Gaussians, while a wide Gaussian (corresponding to all low frequencies and higher frequencies in many directions) appears in nearly every image. Low frequencies occur in all natural images and can be ignored, while the high frequencies represent natural components (such as people, animals, rocks, trees, and plants) and noise. The parameters of the underlying moG, if they can be estimated, can be used as features to discriminate manmade from natural scenes.

Independent Components Analysis (ICA) seems well suited to this problem: recovering original independent sources S_i (the "spikes"), from observed data x_j (the power spectrum). Specifically, ICA assumes that independent sources have been linearly mixed into a number of observations, recovering the mixing matrix $\{a_{ij}\}$ in $x_j = \sum_{i=1}^{n} a_{ij} s_i$.

While there are many methods of performing ICA (see [6] for a good tutorial), in our problem, the number of salient edge directions in manmade scenes often exceeds the dimensionality of the Fourier space (two), requiring the use of overcomplete ICA. In speech recognition, Davies and Mitianoudis [4] modeled the source distributions as mixtures of Gaussians, and were able to estimate the parameters of the model by the expectation-maximization (EM) algorithm. They overcame the exponential complexity by assuming that the source distributions are sparse. In their application, audio data was transformed using discrete cosine transform (DCT), making them sparse in that space. We believe that their algorithm would apply to our Fourier-transformed images as well.

Simply projecting the spectrum into a single dimension (i.e., the orientation) and obtaining the maxima of the histogram is a simplified form of ICA [15]. Once the orientations of the sparse distributions, which correspond to the mixing matrix, are uncovered, we can then estimate the parameters of the Gaussians from the histogram. Our full feature extraction algorithm is as follows:

1. Compute the FFT of the image, take its power spectrum, and convert it to log-space. We use log to prevent the high-energy frequency componenets from overpowering the distribution, while still giving them more weight than low-energy ones.
2. Threshold the images (conservatively) to reduce noise, which can overpower the structure (salient peaks) in the histogram. We also remove the extremely low frequencies (found in all images), because they are quantized in the conversion to polar coordinates more heavily than higher frequencies (thus undesirably over-weighted).
3. Project the spectrum into a one-dimensional angular histogram. We convert each pixel to polar coordinates and create a histogram; we use a bin for each angle between 1 and 180 degrees.
4. Find the spikes (local maxima) in the histogram (Fig. 2a). These correspond to the directions with the most well-defined energy. We smooth the distribution using a sigma filter to eliminate noise.
5. Compute two features for each spike:
    a. *Sparsity* ("spikiness") measures how well defined the edge directions are, defined as the ratio of height to width of the top 20% of the spike. This is the most discriminating feature, as manmade structures tend to yield narrow spikes.
    b. *Energy* is given by the height of the histogram at the spike.
6. Retain only the two spikes with highest sparsity, since they are most salient. The number of spikes was chosen empirically; two was usually enough to distinguish manmade from natural images, without needing an abundance of training data to populate a higher-dimensional space.
7. Compute the direction of the sharpest spike, S1. This allows us to discriminate between horizon lines (in natural scenes), which yield sparse spikes if they are flat and of high contrast, and vertical edges, which usually signify manmade structure. The PCA-based algorithm [8] also required properly oriented images; we can reorient them automatically with good accuracy if need be [1].
8. Extract 5 features from the two spikes, S1 and S2: direction, sparsity, and energy of S1, and sparsity and energy of S2.

Figure 2 shows this process for a manmade image. Our algorithm was able to classify correctly this image, which was misclassified by the baseline algorithm [8].

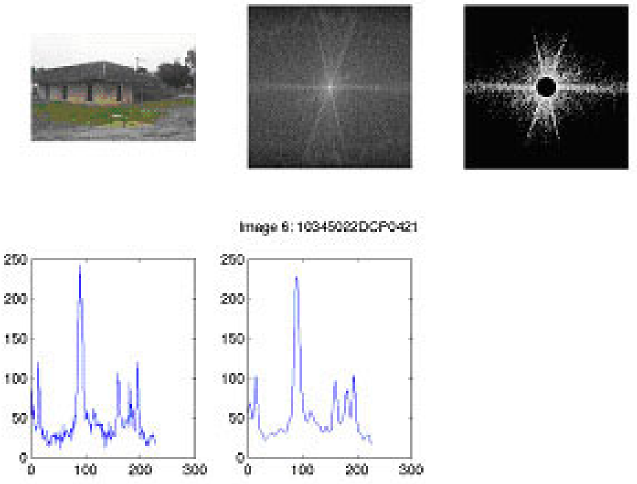

Figure 2: (a) A manmade image misclassifed by [8], but classified correctly by our method. (b) The three highly visible spikes in the frequency domain (the building's walls and two roof lines). (c) After thresholding and removing low-frequency data. (d) The 1D projection. (e) The smoothed projection. Note that the spikes are highly visible at 90 and near 180 degrees (the angles are rotated by 90 degrees to make all arctangent output positive; the bins near 0/180 degrees are copied so spikes near this periodic boundary can be detected). The classifier in [8] did not learn to consider oblique angles as manmade, possibly because of the paucity of training data with those angles.

3. Experimental Results

We started with a set of over 24000 Kodak images collected with the intention of spanning "photospace": 56 photographers from three U.S. cities took pictures during a 12 month period. We chose a subset randomly selected from the original set such that equal proportions of images were drawn from each of the three cities. We then removed close-ups (those containing not enough of the environment to determine the class of the image, as in [9]) and images with ambiguous classification (containing a large area of both manmade and natural components).

In our study, we considered only the 1388 images that are clearly manmade or natural, and denote this data set as DH (for home photos). We further broke it down into independent training and test sets, DHTr and DHTe, which were taken by different photographers. We also used 1069 Corel images as dataset DC, leaving out the 28 images misclassified by the PCA-based classifier. We trained our classifier on DC + DHTr and tested on DHTe and the 28 Corel images.

As a classifier, we chose to use a Support Vector Machine (SVM). They have been shown to have accurate performance and good generalization properties, even when the training data is scarce. Further, the magnitude of the output (the distance from the decision surface) can be used as a measure of confidence in the algorithm's output, which can in turn be used in image retrieval or in a setting where the user is asked to classify ambiguous images.

We obtain high accuracy as shown in Figure 3, with equal recall for each class of 79%, as compared with the baseline performance of 75%. Furthermore, each photograph shown in Figures 1 and 2 was classified incorrectly by the baseline algorithm, but now correctly by the proposed ICA-based algorithm.

We verified that this improvement was not due to the classifier; even a heuristic classifier using the ICA features yielded accuracy of almost 79%. Unfortunately, these heuristics did not include a parameter that could be changed to obtain a ROC-like curve, nor an indicator of confidence.

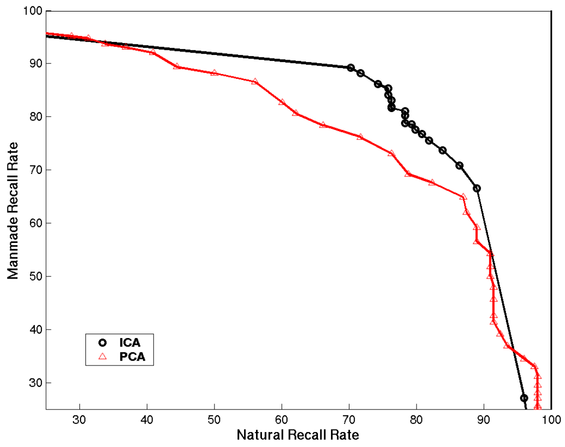

Figure 3: Comparison of performance between the ICA-based and PCA-based classifiers. The ICA classifier shows approximately 4% improvement over the PCA-based classifier at the point of equal class recall.

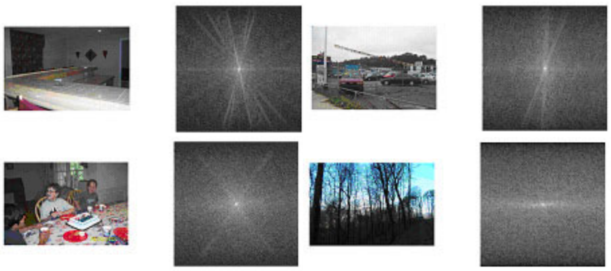

Figure 4. Examples of images classified incorrectly by the PCA-based method in [8], but correctly by the proposed ICA-based algorithm.

As noted above, because our method is to recover spikes from the power spectrum explicitly, we expect to achieve more accurate results than the PCA-based method of [8]. Figure 4 shows example images for which this is the case. Note that the direction of many of the edges in these images is not vertical or horizontal. Our method can handle these oblique angles, which can be due to perspective distortion or camera rotation. These images are typical of the 4% of the images with improved classification. For distant scenes, ICA and PCA both fail, as the edge features become less salient.

Figure 5 shows all 28 of the images misclassified previously by the PCA-based algorithm. The 19 outlined in green were classified correctly by the proposed ICA-based algorithm, while the images outlined in pink remain misclassified. The top 11 images are of natural scenes, while the remainder are of man-made scenes.

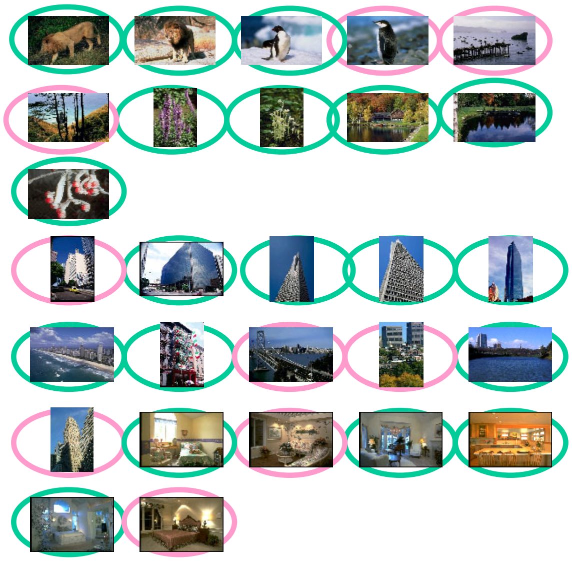

Figure 5. Examples of Corel images classified incorrectly by the PCA-based method in [8].

Furthermore, our ICA-based method uses fewer features than the PCA-based method (5 vs. 16), making it very efficient.

4. Conclusions and Future Work

We have shown that sparse features derived from performing ICA on the power spectrum of images are more effective and more efficient (with fewer features) for classifying photos into natural and manmade classes than PCA-based features. The main difference from the earliest related work to our knowledge in [16], where ICA resulted in two different clusters of basis functions (local, spatial edge "patches") to help classify newspaper text from natural images, is that we applied ICA to the Fourier domain in order to generate optimal features global to the image.

Interesting directions for future work include combining these cues with color cues (e.g., green and brown are more correlated with natural scenes). We also plan to investigate using expectation maximization (EM) to learn the parameters of the mixture of Gaussians (constrained to be zero mean), similar to what Davies did in speech recognition [4] but in the presence of a higher magnitude of outliers.

Acknowledgments

We would like to thank Anlei Dong for implementing the original MIT algorithm, and Professor Michael Davies for valuable discussions on EM-based ICA.

References

1. J. Luo and M. Boutell. Automatic image orientation detection via confidence-based integration of low-level and semantic cues. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(5):715-716, 2005.
2. M. Boutell, J. Luo, A. Singhal, and C. Brown. Survey on the state of the art in semantic scene classification. Technical Report 799, University of Rochester, Computer Science Department, June 19, 2002.
3. B. Bradshaw. Semantic-based image retrieval: A probabilistic approach. *Proceedings of ACM Multimedia*, 167-176, 2000.
4. M. Davies and N. Mitianoudis. A simple sparse mixture model for overcomplete ICA. *IEE Proceedings-Vision Image and Signal Processing*, 151(1):35-43, 2004.
5. R. Duda, R. Hart, and D. Stork. *Pattern Classification*, John Wiley and Sons, Inc., New York, 2001.
6. A. Hyvarinen and E. Oja. Independent components analysis: A tutorial. Technical Report, Helsinki University of Technology, Espoo, Finland, 1999.
7. S. Kumar and M. Hebert. Man-made structure detection in natural images using causal multiscale random field. *Proceedings of International Conference on Computer Vision and Pattern Recognition*, 1:119-126, 2003.
8. A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. of Computer Vision*, 42(3):145-175, 2001.
9. R. Schettini, C. Brambilla, A. Valasna, and M. De Ponti. An Indoor/Outdoor/Close-up Photo Classifier. *Proceedings of SPIE Human Vision and Electronic Imaging*, 2002.
10. N. Serrano, A. Savakis, and J. Luo. Improved indoor-outdoor scene classification. *Pattern Recognition*, 37(9), pp. 1757-1771, September 2004.
11. D. Tax and R. Duin. Using two-class classifiers for multi-class classification. *Proceedings of International Conference on Pattern Recognition*, Quebec City, QC, Canada, August 2002.
12. M. Turk and A. Pentland. Eigen faces for recognition. *J. of Cognitive Neuroscience*, 3(1), 1991.
13. A. Vailaya, M. Figueiredo, A. Jain, and H.-J. Zhang. Content-based hierarchical classification of vacation images. *Proceedings of IEEE Multimedia Systems* '99 (International Conference on Multimedia Computing and Systems), Florence, Italy, June 1999.
14. A.Vailaya, A.K. Jain, and H.-J. Zhang. On image classification: City images vs. landscapes. *Pattern Recognition*, 31:1921--1936, December 1998.
15. L. Vielva, I. Santamaria, D. Erdogmus, and J. Principe. On the estimation of the mixing matrix for underdetermined blind source separation in an arbitrary number of dimensions, *Proceedings of ICA'04*, Granada, Spain, May 2004.
16. T. Lee, M.S. Lewicki, and T.J. Sejnowski. ICA mixture models for unsupervised classification of non-Gaussian classes and automatic context switching in blind signal separation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10):1078-1089, 2000.