

# Natural scene classification using overcomplete ICA

Jiebo Luo<sup>a,\*</sup>, Matthew Boutell<sup>b</sup>

<sup>a</sup>Research and Development Laboratories, Eastman Kodak Company, Rochester, NY 14650, USA

<sup>b</sup>Department of Computer Science, University of Rochester, USA

Received 7 December 2004; accepted 11 February 2005

## Abstract

Principal component analysis (PCA) has been widely used to extract features for pattern recognition problems such as object recognition [Turk and Pentland, *J. Cognitive Neurosci.* 3(1) (1991)]. In natural scene classification, Oliva and Torralba presented such an algorithm in Oliva and Torralba [*Int. J. Comput. Vision* 42(3) (2001) 145–175] for representing images by their “spatial envelope” properties, including naturalness, openness, and roughness. Our implementation closely matched the original algorithm in accuracy for naturalness classification (or “manmade–natural” classification) on a similar (Corel) dataset. However, we found that consumer photos, which are far more unconstrained in content and imaging conditions, present a greater challenge for the algorithm (as they typically do for image understanding algorithms). In this paper, 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. Using both heuristic and support vector machine classifiers, we demonstrated that our ICA-based features are superior to the PCA-based features used in Oliva and Torralba [*Int. J. Comput. Vision* 42(3) (2001) 145–175]. In addition, we augment ICA-based features with camera metadata related to image capture conditions to further improve the performance of our algorithm.

© 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

**Keywords:** Semantic scene classification; Natural scenes; Manmade scenes; Sparse approximation; Independent components analysis (ICA); Principal component analysis (PCA)

## 1. Introduction

Finding the semantic classification, e.g., indoor vs. outdoor, manmade vs. natural, beach vs. desert, of arbitrary images has been studied extensively in the past decade [1]. Hierarchical methods [2] can be helpful, and often include classifying images into manmade vs. natural scenes [3–5] at some step (e.g., city vs. landscape [6]). We envision such system as first classifying images into indoor vs. outdoor [7], then outdoor images into man-made vs. natural,

and natural images further into beach, sunset, mountain, field, fall foliage, desert scenes, and so on [8].

We implemented [9] Oliva and Torralba’s original algorithm in Ref. [3] for representing images by their “spatial envelope” properties, including naturalness, openness, and roughness. One key novelty of their algorithm is to extract features from the power spectrum of an image using principal component analysis (PCA), leading to so-called discriminative structural templates (DST). This algorithm consists of the following steps:

- (1) Perform a Fourier transform on the image and consider only the power spectrum.
- (2) Sample the power spectrum across five frequencies and 12 orientations to obtain 60 features per image (obtained by convolving with Gabor filters (see Fig. 1(c))).

\* Corresponding author. Tel.: +1 585 722 7139;  
fax: +1 585 722 0160.

E-mail addresses: [jiebo.luo@kodak.com](mailto:jiebo.luo@kodak.com) (J. Luo),  
[boutell@cs.rochester.edu](mailto:boutell@cs.rochester.edu) (M. Boutell).

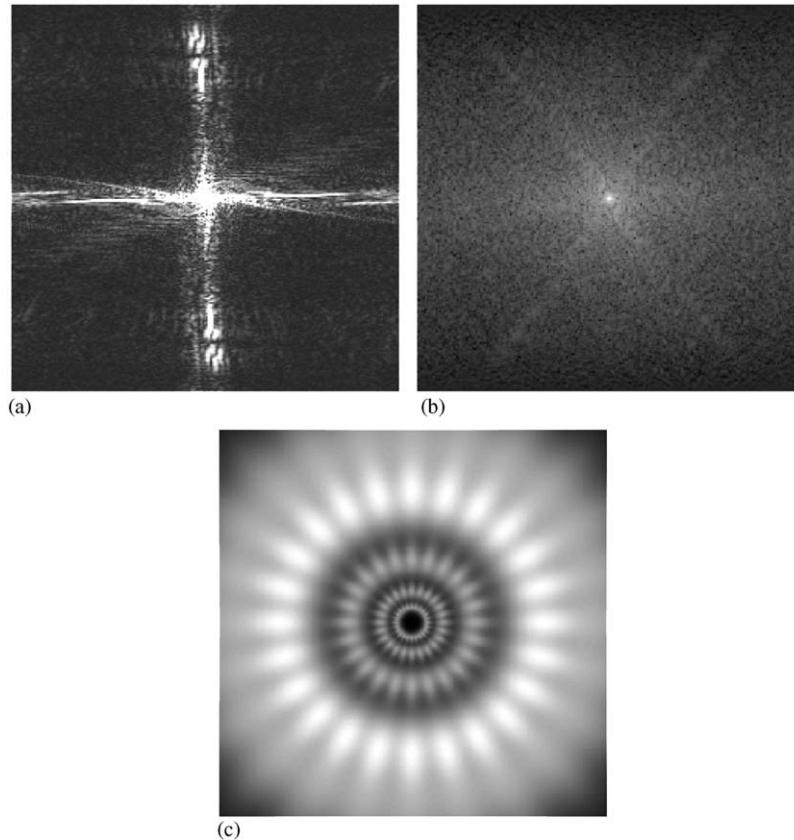


Fig. 1. Examples of the power spectrums and the Gabor filters used for feature extraction in Refs. [3,9].

- (3) Apply PCA to the training set and project data into first 16 principle components.
- (4) Compute the discriminant spectral templates (DST) from each feature.

Our implementation closely matched the original algorithm in accuracy for naturalness classification (or “manmade–natural”) on a similar dataset [3,9], consisting of Corel professional stock photographs. However, we found that consumer photos present a greater challenge for the algorithm (as they typically do for image understanding algorithms [10]). This is obvious from a plot of the output of the DST classifiers on Corel and consumer image sets (Fig. 2). The classifier we implemented was trained on both Corel and consumer photos so as to be representative. For binary problems (such as this one), the DST classifier is equivalent to linear discriminant analysis [3].

Fig. 3 shows typical consumer images misclassified by the DST, along with the power spectrum of each image. These images are typical in that each manmade image contains a relatively large number of edges aligned in a single direction, causing at least one visible narrow “spike” in the

frequency domain. It does not appear that the DST classifier could learn the spike patterns from these features. We hypothesized that much of the discriminating information was lost in sampling the power spectrum (step 2 above). 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 principle 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 (e.g., Fig. 1(a)), and thus overwhelm the less frequent cases where edges are off the two main axes (e.g., Fig. 1(b)). Consequently, we found that the algorithm in Ref. [9] 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

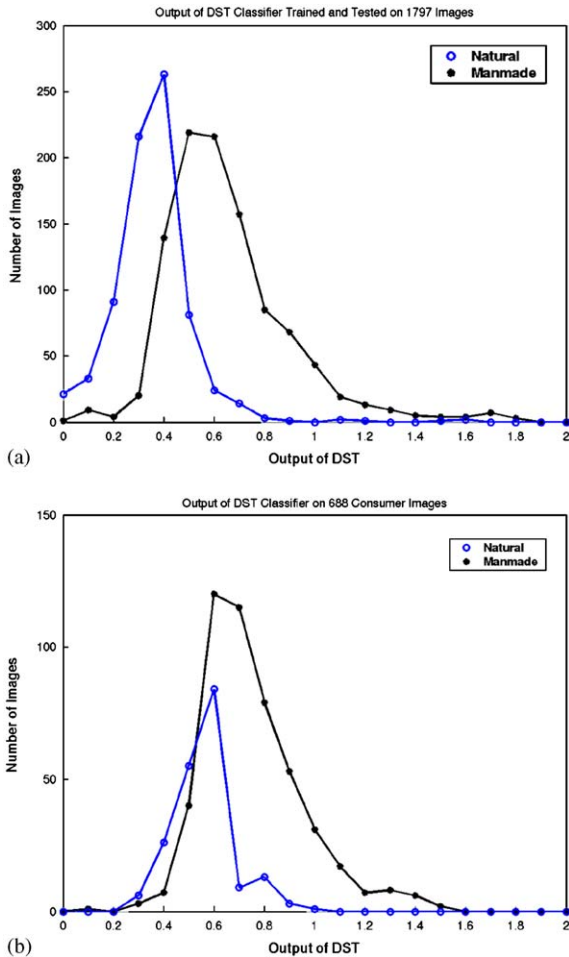


Fig. 2. Output of a DST classifier on the Corel set used in Ref. [9] and on a set of 688 consumer images. While the distributions are nicely separated for the Corel set, they overlap heavily for the consumer images. The classifier was trained on 1097 Corel images and 700 consumer images.

of edge directions present in an image (e.g., the first example in Fig. 3); and (3) when edges are not as discernable (e.g., distant shots of a city scene).

The remainder of the paper is organized as follows. In Section 3, we present an alternative approach to feature extraction, using independent components analysis (ICA) directly on the power spectrum. In Section 3, we design both a heuristic classifier and a support vector machine for naturalness classification using our ICA-based features. In addition, we describe in Section 4 how to add camera metadata related to image capture conditions to further improve the performance of our algorithm. In Section 5, we conduct experiments to compare the classifiers using our ICA-based features with the DST features used in Ref. [9]. Finally, we discuss our results and conclude in Sections 6 and 7.

## 2. Independent components analysis

Manmade images are typically characterized by many edges lying in a few well-defined directions. In the Fourier domain, we can characterize the content of each scene as a mixture of Gaussians (Fig. 3). Manmade components appear as sparse, thin Gaussian distributions, while a broad Gaussian (corresponding to low frequencies in many directions) appears in nearly every image and represents natural components (such as people, animals, rocks, trees, and plants) and noise. If we could estimate the parameters of the underlying mixture of Gaussians, then we could use these as features to discriminate manmade from natural scenes.

ICA seems to be a more suitable solution to this type of problem than PCA: recovering original independent sources  $S_i$ , in this case the “spikes”, from observed data  $x_j$ , in this case the power spectrum, as opposed to energy compaction. Specifically, it assumes that independent sources have been *linearly* mixed into a number of observations; ICA recovers the mixing matrix  $\{a_{ij}\}$

$$x_j = \sum_{i=1}^n a_{ij} s_i.$$

ICA has been studied extensively (see Ref. [11] for a good tutorial). ICA has often been applied to audio signals, and is closely related to blind source separation (BSS). The classic example of BSS is the cocktail party problem in which microphones placed around the room each record a “garbled” mixture of speakers and the algorithm attempts to recover each individual speaker.

While there are many methods of performing ICA, e.g., FastICA and JADE [11], the problem in this study poses a unique challenge because the number of salient edge directions in manmade scenes often exceeds the dimensionality of the Fourier space (two). Consequently, the so-called *overcomplete* ICA is required. Davies and Mitianoudis [12] 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 form of ICA [13], albeit a “poor man’s” version. We found it to work reasonably well as a proxy. 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. This has the effect

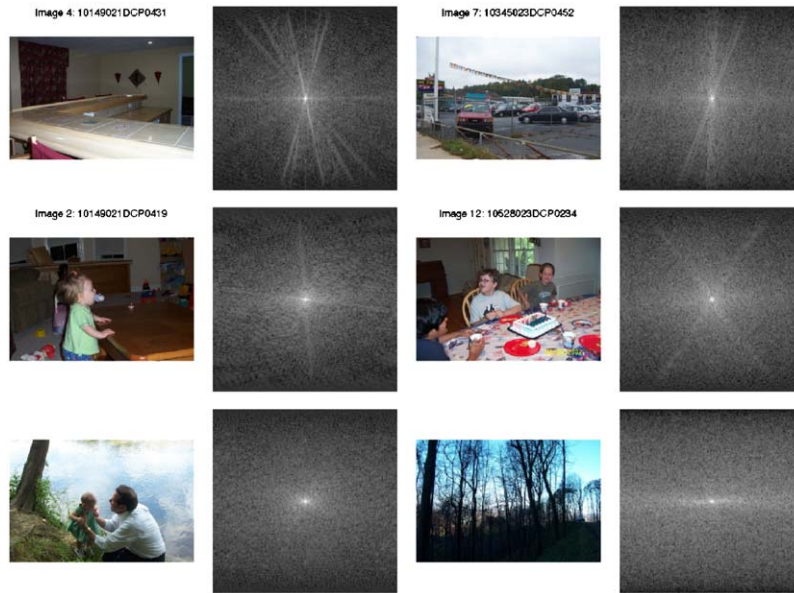


Fig. 3. Fourier transforms of manmade (rows 1 and 2) and natural (row 3) images. The manmade images have clearly identifiable spikes in the Fourier domain, while the natural images do not. Each of these images was misclassified by the DST algorithm [9], but classified correctly by the proposed ICA algorithm.

of preventing the high-energy areas from overpowering the distribution, while still giving them more weight than low-energy ones. This is a tradeoff between the extremes of (1) using a threshold to binarize the image and (2) doing nothing, i.e., using the original power spectrum.

- (2) Threshold the images (conservatively) to reduce noise. An alternative, which both enhances the high values and reduces the low values, is to shape the image values using a sigmoid function; we found this to help in other domains (e.g., medical exam classification). It did not help as much here, because the noise, while reduced, can overpower the structure (salient peaks) in the histogram. We also remove the extremely low frequencies, because they are quantized in the conversion to polar coordinates more heavily than higher frequencies. Note that all images, whether natural or manmade, generally have significant low frequency energy. Some information is lost here, but would necessitate interpolation to overcome the quantization (see step 3).
- (3) Project the spectrum into a one-dimensional angular histogram. We convert each pixel to polar coordinates (using the arctangent function), and create a histogram in which each bin corresponds to an angle between  $1^\circ$  and  $180^\circ$ . A more sophisticated version is to perform a *backwards* mapping: we calculate, for each point with a given angle and radius, its coordinates in the original (rectangular) space and find its value using interpolation.
- (4) Find the spikes in the histogram (Fig. 3a). These correspond to the directions with the most well-defined en-

ergy. We search for local maxima using a 5-point window, relaxing the criterion such that a maximum cannot be less than any of its neighbors, and must be strictly greater than at least two neighbors.

- (5) Compute two features for each spike:
  - (a) *Sparsity* (“spikiness”) measures how well defined the edge directions are. As stated, manmade structure tends to give narrow spikes. This is the most discriminating feature. We define this as the ratio of height to width of the top 20% of the spike.
  - (b) *Energy* is given by the height of the histogram at the spike.
- (6) Compute the direction of the sharpest spike. This allows us to discriminate between *horizon* lines (natural), which yield sparse spikes if they are flat and of high contrast, and *vertical* edges, which usually signify manmade structure.
- (7) Sort spikes by sparsity, keep the two spikes with highest sparsity, since they are most salient. More than two could be used, but two is usually enough to separate the classes (“manmade” or “natural”). Further, using a greater number would require more training data, since the feature space would contain more dimensions.
- (8) Form a 5-dimension feature vector from spikes  $S_1$  and  $S_2$ :
 

(direction ( $S_1$ ), sparsity ( $S_1$ ), energy ( $S_1$ ), sparsity ( $S_2$ ), energy ( $S_2$ )).

Fig. 4 shows this process for a manmade and a natural image. Our algorithm was able to classify correctly these two

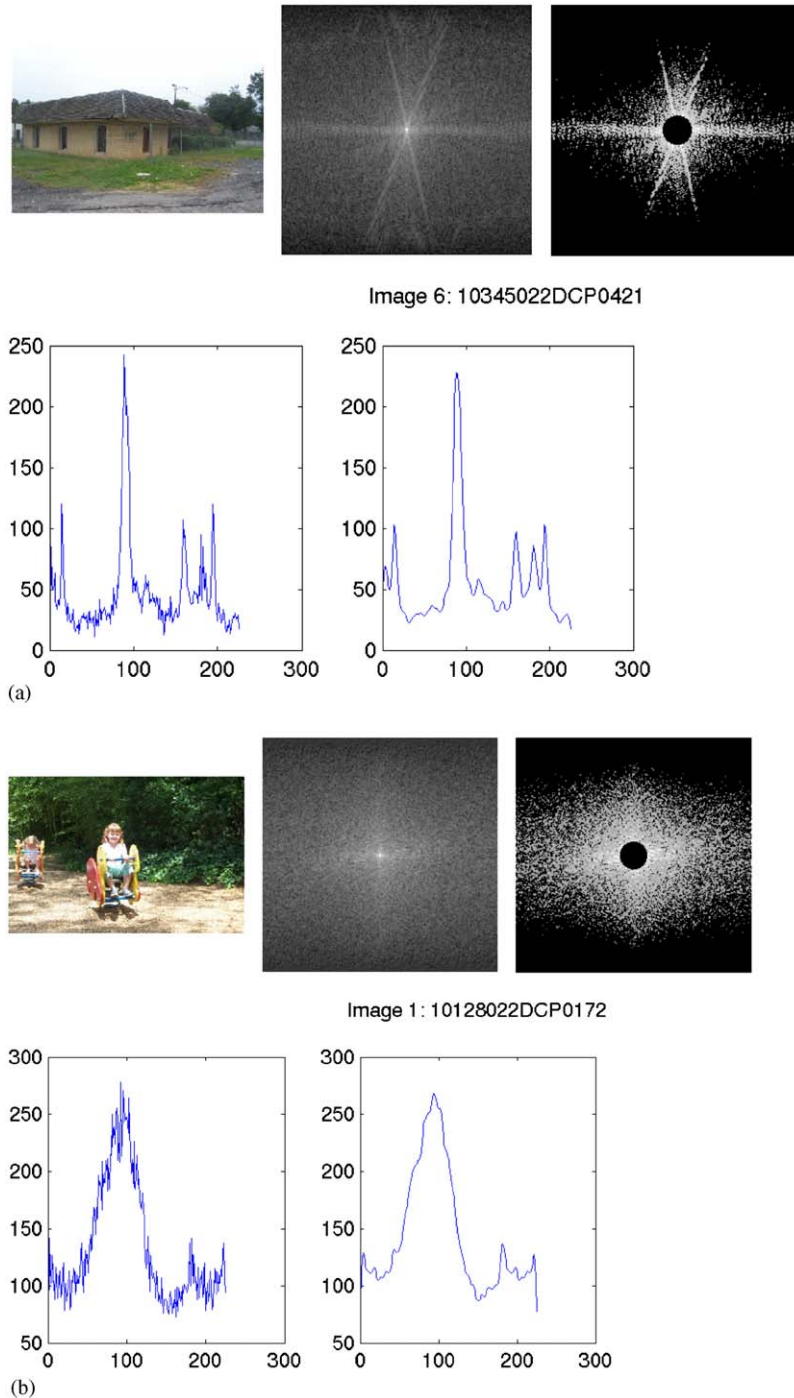


Fig. 4. (a) A manmade image misclassified in Ref. [9], but classified correctly by our method; (b) the three highly visible spikes in the frequency domain (the building's walls and two roof lines); (c) after thresholding and removing low-frequency data; (d) the 1D projection; and (e) the smoothed projection. Note that the spikes are highly visible at  $90^\circ$  and near  $180^\circ$  (the angles are rotated by  $90^\circ$  to make all arctangent output positive). The classifier in Ref. [9] did not learn to consider oblique angles as manmade, possibly because of the paucity of training data with those angles. (b) A natural image misclassified in Ref. [9], but classified correctly by our method. The only spike in the frequency domain corresponds to a horizon line, and is not spiky because the horizon is not clear.

images, which were misclassified by the baseline algorithm [9].

### 3. Image classification using ICA

Given these features, we can then classify an image using heuristics or using a classifier such as a support vector machine. We discuss each in the following sections.

#### 3.1. Heuristics for classifying images as manmade or natural

We observe that manmade images contain spikes with a large sparsity measure due to consistent edge directions. Natural images usually do not, because their angles are more spread out, with the possible exception of horizon lines. This is the basis of our heuristic classification algorithm:

```

if the image has no spikes
  Classify as NATURAL
else if the image has a single spike,  $S$ 
  if direction ( $S$ ) == horizontal
    Classify as NATURAL
  else if energy ( $S$ ) >  $Thr_{EN1}$  and sparsity ( $S$ ) >  $Thr_{SP1}$ 
    Classify as MANMADE
  else
    Classify as NATURAL
else if the image has two (or more) spikes  $S_1$  and  $S_2$ 
  if energy ( $S_1$ ) >  $Thr_{EN1}$  and sparsity ( $S_1$ ) >  $Thr_{SP1}$  and
  energy ( $S_2$ ) >  $Thr_{EN2}$  and sparsity ( $S_2$ ) >  $Thr_{SP2}$ 
    Classify as MANMADE
  else if energy ( $S_1$ ) >  $Thr_{EN1}$  and sparsity ( $S_1$ )
  >  $Thr_{SP1}$  and direction ( $S_1$ ) == vertical
    Classify as MANMADE
  else
    Classify as NATURAL

```

This heuristic classifier is fast, robust, and reasonably accurate on consumer images. However, we would like for it to output a confidence in its decision. This confidence could be used for image retrieval or in a semi-supervised classification algorithm, in which the user is asked to classify ambiguous images. We could also combine it with other evidence, e.g., color or camera metadata (see Section 3).

To be useful, such a confidence should be correlated with the degree to which the features match the model. Two non-horizontal spikes with extremely high energy and sparsity should be classified as *manmade* with *high* confidence, an image with no spikes should be classified as *natural* with *high* confidence, but an image with a single spike near the threshold could be classified as either, but with *low* confidence. Unfortunately, it is non-trivial to combine distances from multiple thresholds optimally into a single measure of confidence.

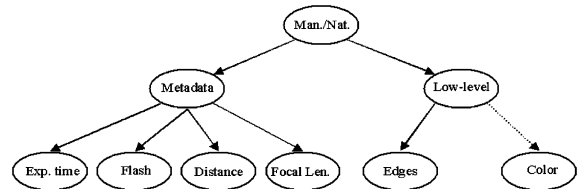


Fig. 5. Bayesian network for evidence combination.

#### 3.2. Support vector machine classifier

Turning our attention to the machine learning literature, support vector machines (SVMs) have been shown to have accurate performance and good generalization properties, even when the training data is scarce. Further, for a test example, they output a real number, in which the sign is the classification and the absolute value is the distance from the decision surface. This distance can be shaped into a *pseudo*-probability using a sigmoid function [14], which is exactly the confidence measure we desire.

#### 3.3. Orientation-awareness

Note that the DST [3] is orientation-aware, that is, all images were presented to the algorithm in their upright orientation. The power spectrum is invariant to rotation by  $180^\circ$  and to reflection across the vertical and horizontal directions, so at most, portrait images would have to be rotated  $90^\circ$ . Furthermore, the DST images in Ref. [9] appear to be at least approximately rotationally invariant. We confirmed this by training and testing two DSTs on the same sets of images. In the training set, 8% (141 of 1797) of images were in portrait orientation (since over half were correctly oriented Corel images), and in the test set, 17% (115 of 688). For the first, we rotated the portrait images by  $90^\circ$ ; for the second, we did not. The results on each were nearly identical.

The ICA-based feature extraction algorithm is similar; to process a portrait-oriented image, we can process it as if it were landscape, then rotate the spike direction by  $90^\circ$ .

## 4. Combining image content and camera metadata

Camera metadata recording image capture conditions can improve indoor–outdoor image classification and sunset detection [15]. The most salient cues were flash fired, exposure time, aperture, subject distance, and focal length. For indoor–outdoor classification, the metadata was so effective, it could be used *in place of* image content and still achieve 90% classification accuracy. However, this is not true in general (e.g., sunset detection), so the content and metadata cues would need to be combined. We combine the cues using the Bayesian network shown in Fig. 5 (color cues are not currently implemented). Section 5.2 contains discussion

on the cue distributions for naturalness classification calculated from our image set.

### 5. Experimental results

#### 5.1. Image databases

We started with a set of over 15,000 Kodak images collected with the intention of spanning “photospace” [16]. Fifty-six photographers from three US cities (Atlanta, Los Angeles, and Chicago) took pictures over the course of 12 months. Each image contains complete camera metadata. However, all were taken using single camera model, the Kodak DX3600. Dataset *D* is a subset of 3600 images randomly selected from the original set such that equal proportions of images were drawn from each of the three cities.

Of these, 990 (28%) were clearly manmade and 398 (11%) were clearly natural. By clearly, we mean it was unambiguous to a human observer to which class each image belonged because the majority of the scene content belonged to that class. These numbers are consistent with internal studies. For film images, 11% were nature/landscape, 10% were building structure, and 24% were in or around the house [17]. Because of low costs, digital cameras

Table 1  
Image database descriptions

Database	Size	Manmade	Natural	Comments
<i>D</i>	3600	990	398	Full data set
DH	1388	990	398	<i>D</i> with ambiguous, close-ups removed
DHTr	700	500	200	First half of DH
DHTE		490	198	Second half of DH
DC	1097	529	568	Corel stock photos, some web images

allow consumers to take many more photos of daily life (37% in/around house, usually cluttered), leaving a smaller percentage of images in “clear” categories (e.g., landscapes taken while on vacation) [16] and thus making the problem significantly more difficult unless we constrain the problem.

Another 591 (16%) of the images were close-ups. We separated out close-ups, whether of a person or an object, because they do not contain enough of the environment for an algorithm (nor a human in many cases) to determine the class of the image as manmade or natural. Digital photos contain many close-ups, more so than film [17]. For example, in “children-only” digital photos (ignoring zoom), 20% of images have subject distance less than 2 ft and 50% have

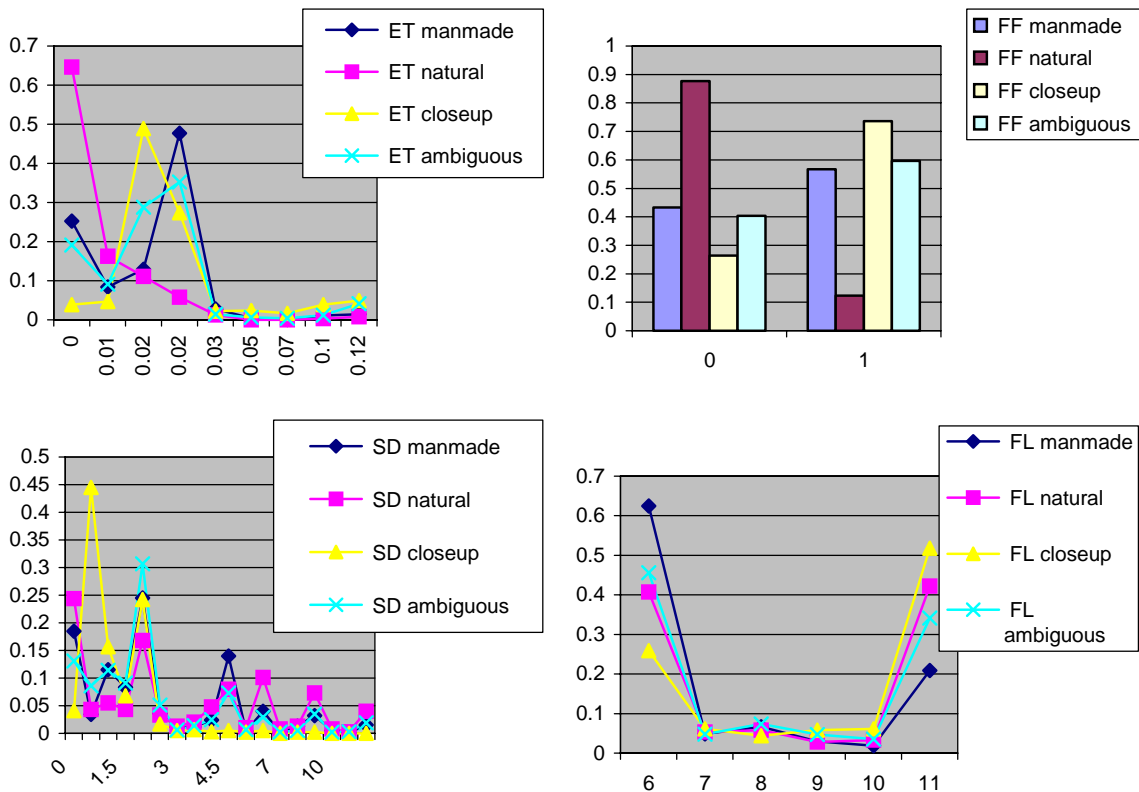


Fig. 6. Distribution of exposure times (ET), flash fired (FF), subject distance (SD), and focal length (FL) on dataset *D*.

distance less than 4 ft. For film, the 20% and 50% marks move out to 4 and 8 ft. The percentages of images with average distance (6–12 ft) are similar for both. Separating close-ups has been done in the literature [18]. Accurate focal length and subject distance could be used to automatically filtering out close-ups, but we leave that to future work.

The remaining 1621 (45%) images were ambiguous or of low quality. This may seem like a large percentage, until one considers that most of a person’s daily life is spent in an environment with both manmade and natural components; it is unclear how to classify images containing a large area of each (unless we allow “both” as a category [19]).

In our study, we consider only the 1388 images that are clearly manmade or natural, and denote this data set as DH (for home). We further break it down into independent training and test sets, DTr and DTe, which were taken by different photographers. We made no effort to ensure that each set was identically distributed, but merely took the first 500 manmade and first 200 natural images for training and used the rest for testing. We call the 1097, primarily Corel images used in Ref. [9] as dataset DC. Table 1 summarizes the datasets.

### 5.2. Metadata distributions for manmade and natural images

We considered the same eight cues as we did for indoor–outdoor and sunset detection. Fig. 6 shows the distributions of the strongest cues (exposure time, flash fired, subject distance, and focal length) on dataset *D* (close-up and ambiguous classes included). Photos of natural scenes tend to have shorter exposure times and no flash, due to being captured primarily outdoors. Subject distances and focal lengths, which are somewhat weaker cues, appear correlated to indoor and outdoor classes as well. Subject distance is heavily quantized, but even an approximation can help (e.g., if distance is greater than 7 m, it is much more likely to be a natural image). Close-up images mostly have distances of less than 3 m. Focal length is bimodal. Because the zoom ratio on the camera is small, it is difficult to use an intermediate zoom setting. However, for the extremes, photographers tend to use higher zoom on natural images, perhaps to capture distant scenes.

Of the other cues, shutter speed is correlated with exposure time (but recorded less often). Scene energy, aperture, and *f*-number, were weak cues, showing little discriminatory power.

### 5.3. Performance

#### 5.3.1. Gabor features and DST classifier

Fig. 7 shows the performance of the DST classifier trained on DC [9] on two datasets, DC, and DHTe. We compute the recall of each class while varying the threshold between the two classes to obtain the graphs (similar to ROC curves).

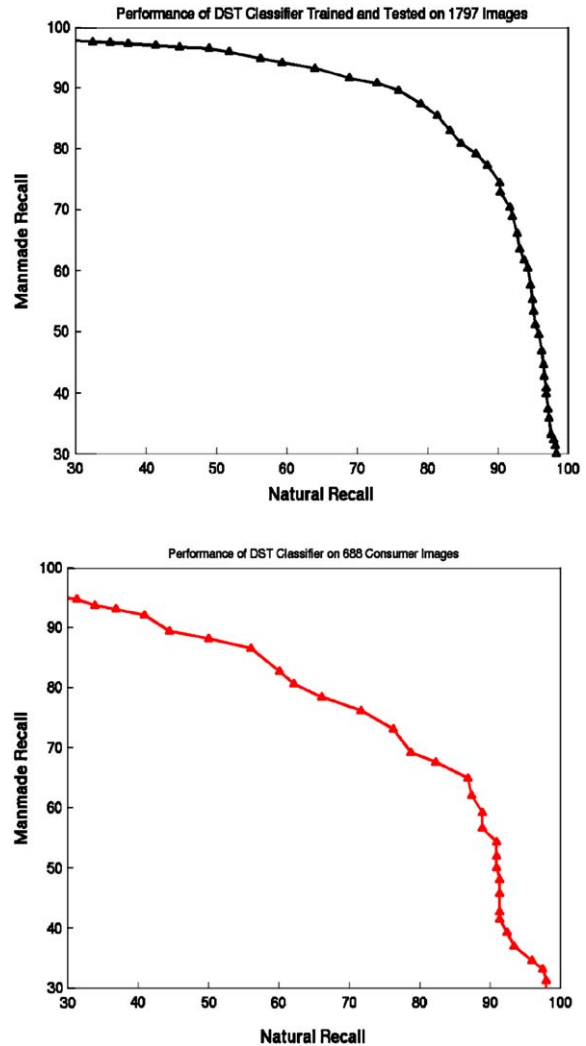


Fig. 7. Performance of the DST classifier on the Corel/consumer training set (left) and the consumer test set. On the training set, the breakeven point is about 83% (at threshold  $t = 0.50$ ), while on the test set, it is about 74% (at threshold  $t = 0.67$ ). The lack of a stable threshold makes the DST method difficult to use in practice.

Classification on the training set is accurate, as expected, and is similar to that reported in Ref. [9]. However, we see that the classifier does not generalize well to consumer images, even when including some consumer images in the training set.

We also implemented linear discriminant analysis (LDA) [20] directly, and found that it gave similar results as DST, as stated in Ref. [3].

#### 5.3.2. ICA features and heuristic classifier

Our heuristic classifier using ICA features (Section 2.1) performed as well as the DST classifier on Gabor features on the same training set used above (DC+DHtr), obtaining

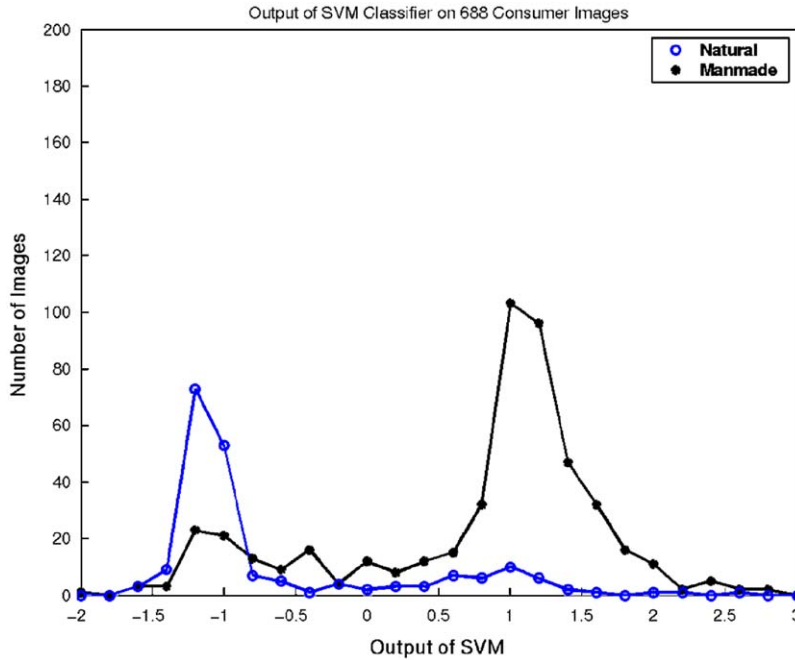


Fig. 8. Distribution of SVM output on data set DHtr for each class. Note the large separability (compared to similar plots in Fig. 2).

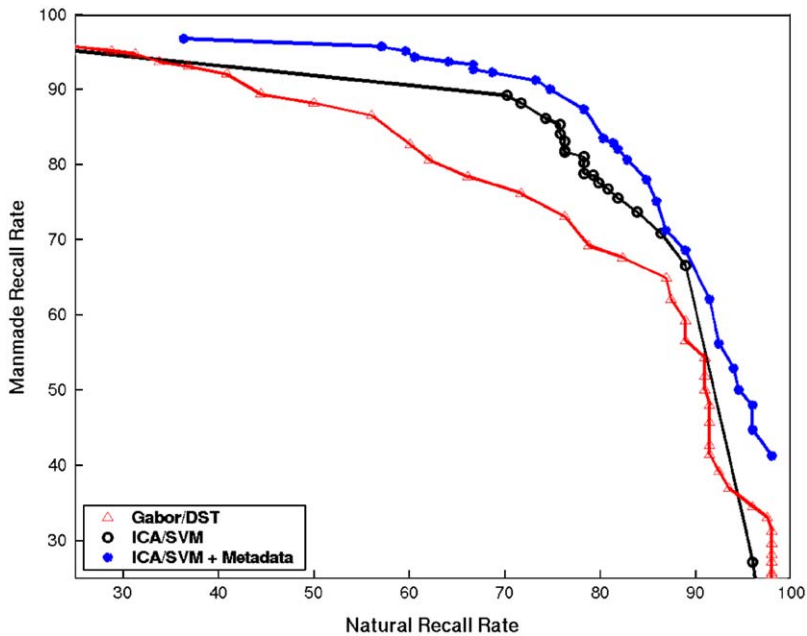


Fig. 9. Comparison of performance between the content-only classifiers (Gabor/DST and ICA/SVM) with the metadata-enhanced classifier. The ICA/SVM classifier shows approximately 4% improvement over the Gabor/DST classifier at the point of equal class recall. Further, using metadata increases accuracy over the best content-based classifier regardless of the operating point used, e.g. about 3% (from 79% to 82%) at the point of equal class recall.

88.9% accuracy on natural images and 82.5% accuracy on manmade images. On DC, it was even better: 90.8% accuracy on natural images and 90.1% accuracy on manmade

images (similar to Ref. [9]). Note that there is no bias gained in testing on this data set, because the heuristic classifier is not “trained”.



Fig. 10. Examples of Corel images classified incorrectly by the PCA-based method in Refs. [3,9]. The 19 outlined in green are now 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 remainders are of manmade scenes.

On data set DHte, it obtained 83.8% accuracy on natural images, and 76.9% accuracy on manmade images, far better than the Gabor features and DST classifier. It is noteworthy that this improvement is not a reflection of a better classifier (our heuristics are certainly suboptimal), but of *more expressive* features.

### 5.3.3. ICA features and support vector machine classifier

We trained the SVM on a combined set composed of DC+DHtr. The SVM classifier obtained slightly higher accuracy on DHte than the heuristics, as expected. However, as noted, the primary advantage of using the SVM for this application is its confidence measure for each image, which will be later integrated with the camera metadata cues.

Note that SVMs can memorize (and thus obtain arbitrarily high accuracy on) training sets, so we do not report results

on DC or DHtr. Fig. 8 shows a histogram of the SVM output for each class of images in data set DHte. Note the high degree of separability using this feature set (as compared with Fig. 2). We obtain high accuracy as shown in Fig. 9, with equal recall for each class of 79%.

In Fig. 10, the advantage of the ICA-based features is demonstrated in the corrections made on 19 (indicated by green circles) of the 28 Corel images the original PCA-based algorithm misclassified. In addition, all the consumer images shown in Fig. 3 are now correctly classified by the ICA-based method.

### 5.3.4. Metadata-enhanced classifier

We extracted four metadata cues (exposure time, flash, subject distance, and focal length) for each image and combined them with the image content evidence (the output



Fig. 11. Manmade image samples, classified correctly by both content-based and combined cues (row 1), gained by metadata (row 2), lost by metadata (row 3), and incorrectly regardless of cues (row 4).

of the SVM just described) using the Bayesian network in Fig. 5. Fig. 9 shows the accuracy of the metadata-enhanced classifier, compared with the accuracy of both content-only baselines (Gabor and ICA features). We see that metadata improves the accuracy of the classifier across every point on the operating curve. At equal class recall, accuracy is approximately 82%.

## 6. Discussions

Fig. 11 shows examples of manmade scenes. Those classified correctly regardless of the method (row 1) typically have strong edge content, as expected. Those gained by metadata (row 2) have somewhat less salient edge content, but were captured in a typical indoor environment (with flash fired and long exposure time). The images lost by metadata (row 3) are all outdoor scenes in which the flash usually was not used and most of them do not have very salient edge content to begin with. Metadata is unable to help classify correctly images with very weak edge content (row 4), since the belief that they are natural is too strong to overcome.

Fig. 12 shows typical natural images (with same notational layout). Of particular interest are those lost by metadata, because they include the rare cases in which flash was used outdoors (see row 3). Natural images with some strong edges (e.g., flat horizon, tree trunks) were often misclassified by the baseline classifier (but usually just beyond the threshold); metadata helped correct these mistakes by pushing back over the threshold those images captured under typical outdoor conditions (see row 2). Metadata cannot help classify images with strong vertical edge content, which are mistaken for manmade structure in the feature space (see row 4, and it is interesting to note that some of these images do have manmade structures, albeit in natural settings).

## 7. Conclusions and future work

In this study, we have shown that sparse features derived from performing ICA on the power spectrum of images are more effective and more efficient (with fewer number of features) in classifying images into natural and manmade classes, compared to PCA type of features. We

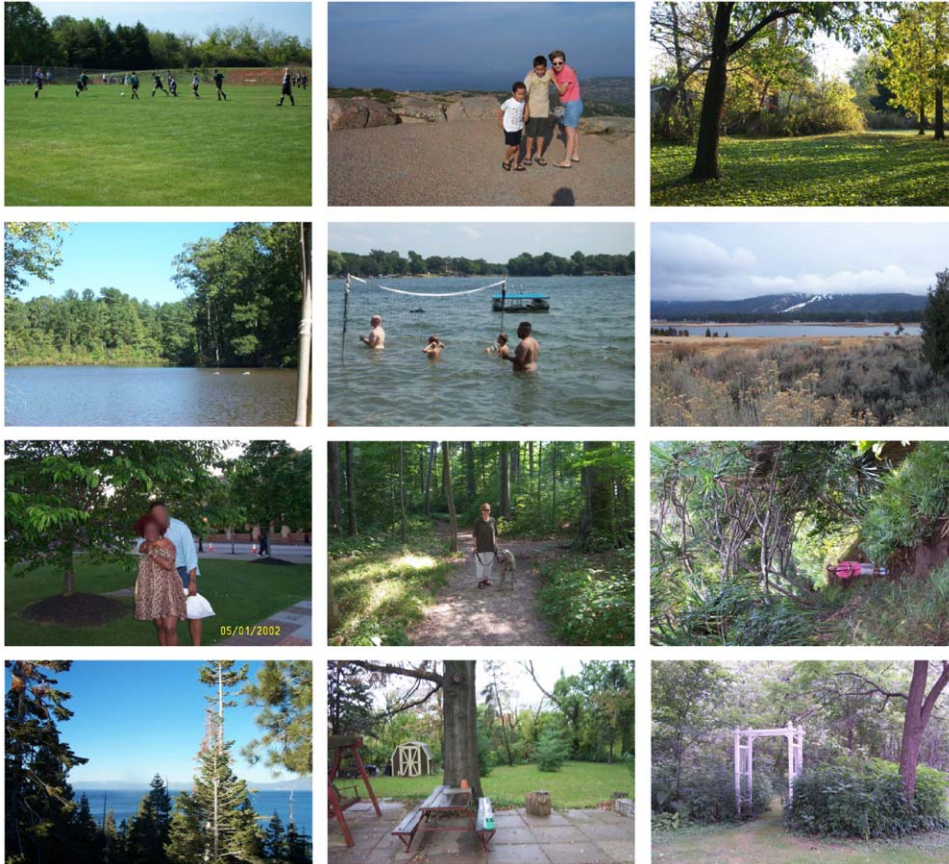


Fig. 12. Natural image samples, classified correctly by both content-based and combined cues (row 1), gained by metadata (row 2), lost by metadata (row 3), and incorrectly regardless of cues (row 4).

also showed that metadata cues, which add negligible computational overhead, improve the classification accuracy by an average of 2–3%, regardless of the operating point chosen.

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 Davies and Mitianoudis [12] but in the presence of many outliers.

### Acknowledgements

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] M. Boutell, J. Luo, A. Singhal, 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.

- [2] A. Vailaya, M. Figueiredo, A. Jain, H.-J. Zhang, Content-based hierarchical classification of vacation images, in: *Proceedings of the IEEE Multimedia Systems '99, International Conference on Multimedia Computing and Systems*, Florence, Italy, June 1999.
- [3] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, *Int. J. Comput. Vision* 42 (3) (2001) 145–175.
- [4] S. Kumar, M. Hebert, Man-made structure detection in natural images using a causal multiscale random field, *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 119–126.
- [5] B. Bradshaw, Semantic-based image retrieval: a probabilistic approach, *Proceedings of the ACM Multimedia*, 2000, pp. 167–176.
- [6] A. Vailaya, A.K. Jain, H.-J. Zhang, On image classification: city images vs. landscapes, *Pattern Recognition* 31 (1998) 1921–1936.
- [7] N. Serrano, A. Savakis, J. Luo, Improved indoor–outdoor scene classification, *Pattern Recognition* 37 (9) (2004) 1757–1771.

- [8] M. Boutell, J. Luo, C. Brown, Learning spatial configuration models using modified Dirichlet priors, in: Proceedings of the 2004 Workshop on Statistical Relational Learning (in conjunction with ICML2004), Banff, Alberta, July 2004.
- [9] A. Dong, J. Luo. Towards holistic scene descriptors for semantic scene classification, Eastman Kodak Company Technical Report, October 1, 2003.
- [10] M. Turk, A. Pentland, Eigen faces for recognition, *J. Cognitive Neurosci.* 3 (1) (1991).
- [11] A. Hyvarinen, E. Oja, Independent components analysis: a tutorial, Technical Report, Helsinki University of Technology, Espoo, Finland, 1999.
- [12] M.E. Davies, N. Mitianoudis, A simple sparse mixture model for overcomplete ICA, *IEE Proc.-Vision Image Signal Process.* 151 (1) (2004) 35–43.
- [13] L. Vielva, I. Santamaria, D. Erdogmus, J.C. Principe, On the estimation of the mixing matrix for underdetermined blind source separation in an arbitrary number of dimensions, in: Proceedings of the ICA'04, Granada, Spain, May 2004.
- [14] D. Tax, R. Duin, Using two-class classifiers for multi-class classification, in: Proceedings of the International Conference on Pattern Recognition, Quebec City, QC, Canada, August 2002.
- [15] M. Boutell, J. Luo, Bayesian fusion of camera metadata cues in semantic scene classification, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Washington, DC, 2004.
- [16] R.K. Segur, Using photographic space to improve the evaluation of consumer cameras, Eastman Kodak Company Technical Report, November 1999.
- [17] K. Donnelly, D. Howell, The optical photographic usage study, Eastman Kodak Company Technical Report, November 2000.
- [18] R. Schettini, C. Brambilla, A. Valasna, M. De Ponti. An indoor/outdoor/close-up photo classifier, Proceedings of the IS&T Conference on Human Vision and Electronic Imaging, 2002.
- [19] M. Boutell, J. Luo, Multi-label semantic scene classification, *Pattern Recognition* 37 (9) (2004) 1757–1771.
- [20] R. Duda, R. Hart, D. Stork, *Pattern Classification*, second ed., Wiley, New York, 2001.

**About the Author**—MATTHEW BOUTELL received the B.S. degree in Mathematical Science from Worcester Polytechnic Institute in 1993 and the M.Ed. degree from the University of Massachusetts in 1994. He served for several years as a mathematics and computer science instructor at Norton High School and at Stonehill College. Currently, he is a Ph.D. student in Computer Science at the University of Rochester. His research interests include computer vision, pattern recognition, probabilistic modeling, and image understanding. He is a student member of the IEEE.

**About the Author**—JIEBO LUO received his Ph.D. degree in Electrical Engineering from the University of Rochester in 1995. He is currently a Senior Principal Research Scientist in the Eastman Kodak Research Laboratories. His research interests include image processing, pattern recognition, and computer vision. He has authored over 100 technical papers and holds over 30 granted US patents. Dr. Luo was the Chair of the Rochester Section of the IEEE Signal Processing Society in 2001, and the General Co-Chair of the IEEE Western New York Workshop on Image Processing in 2000 and 2001. He was also a member of the Organizing Committee of the 2002 IEEE International Conference on Image Processing, a Guest Co-Editor for the *Journal of Wireless Communications and Mobile Computing* Special Issue on Multimedia Over Mobile IP, lead Guest Editor for the *Pattern Recognition* journal Special Issue on Image Understanding for Digital Photos, and a Member of the Kodak Research Scientific Council. Currently, he is serving as an Associate Editor of the journals *IEEE Transactions on Multimedia*, *Pattern Recognition*, and *Journal of Electronic Imaging*. Dr. Luo is a Kodak Distinguished Inventor, a winner of the 2004 Eastman Innovation Award (Kodak's most prestigious technology prize), and a Senior Member of the IEEE.