Bayesian Fusion of Camera Metadata Cues in Semantic Scene Classification

Matthew Boutell[1]
[1]*Department of Computer Science*
University of Rochester
boutell @cs.rochester.edu

Jiebo Luo[2]
[2]*Research & Development Laboratories*
Eastman Kodak Company
luo@image.kodak.com

Abstract

Semantic scene classification based only on low-level vision cues has had limited success on unconstrained image sets. On the other hand, camera metadata related to capture conditions provides cues independent of the captured scene content that can be used to improve classification performance. We consider two problems: indoor-outdoor classification and sunset detection. Analysis of camera metadata statistics for images of each class revealed that metadata fields, such as exposure time, flash fired, and subject distance, are most discriminative for both indoor-outdoor and sunset classification. A Bayesian network is employed to fuse content-based and metadata cues in the probability domain and degrades gracefully, even when specific metadata inputs are missing (a practical concern). Finally, we provide extensive experimental results on the two problems, using content-based and metadata cues to demonstrate the efficacy of the proposed integrated scene classification scheme.

1. Introduction

Determining the semantic classification (e.g., indoor, sunset, mountain, picnic) of an arbitrary image has been studied extensively in recent years. These classifiers use features derived from the image *content* only (e.g., colors, textures, edges) and achieve some success. With the advent and proliferation of digital cameras, an enormous number of digital images are created. Along with the need for automatic scene classification (e.g., for use in content-based enhancement and organization), digital cameras also bring with them a powerful source of information little-exploited for scene classification: camera metadata embedded in the digital image files. Metadata (or "data about data") for cameras records information related to the image capture conditions and includes values such as date/time stamps, presence or absence of flash, subject distance, exposure time, and aperture value.

Much research has been done on problems of scene classification [1,2,5,8,9,12,13,14,16,18]. The majority of these systems employed a learning-by-example approach based on low-level vision features derived exclusively from scene content.

Meanwhile, metadata has been used in the past for image analysis. For example, the use of key word annotations has been studied extensively in the context of image retrieval, e.g., [4,7]. Timestamps have been used successfully to *cluster* photographs by events [10]. However, none of the prior research exploited metadata related to image capture conditions (e.g., exposure time and flash), and none was used specifically for scene classification.

We present a probabilistic approach to fusing evidence from the camera metadata with that from a content-based image classifier. We start by discussing types of metadata cues appropriate for scene classification and using rigorous statistical discriminant analysis to identify valuable cues for a given problem. We apply our model successfully to the problems of indoor-outdoor scene classification and sunset scene detection. In addition, we demonstrate that our scheme functions gracefully when some or all of the cues are missing, leading to an indoor-outdoor scene classifier based solely on the metadata (without any content-based cues) that gives comparable results to existing scene classifiers using negligible computing resources.

Our main contribution is a Bayesian inference scheme capable of fusing *multi-modal* cues derived from both the image content and camera metadata for the purpose of scene classification, which also degrades gracefully with missing metadata. Rigorous statistical analysis and pertinent domain knowledge are used to facilitate such fusion.

2. Digital camera metadata

The Exif specification for camera metadata (used for JPEG images) includes hundreds of tags. Among these, 26 relate to picture taking conditions (e.g., FlashUsed, FocalLength, ExposureTime, Aperture, FNumber, ShutterSpeed, and Subject Distance). The Exif standard is open (http://www.exif.org).

It is clear that some of these cues can help distinguish various classes of scenes. For example, flash tends to be used more frequently with indoor images than with outdoor images. Some tags will be more useful than others for a given problem. We present intuitions about mutually independent tag categories, followed by a method for evaluating the discrimination power of various tags based on statistical analysis. Later in the paper, we use these analyses to identify tags most useful for the specific problems of indoor-outdoor and sunset classification.

2.1. Families of metadata tags

We have categorized these tags into four families that we believe to be useful for scene classification. These families are largely mutually independent from the physics of photography. Therefore, such categorization is likely to be valid beyond the two applications addressed in this study.

Scene Brightness. This category includes exposure time, aperture, f-number, and shutter speed. Natural lighting is stronger than artificial lighting. This causes outdoor scenes to be brighter than indoor scenes, even under overcast skies, and they, therefore, have a shorter exposure time, a smaller aperture, and a larger brightness value. The brightness value of sunset images tends to lie within a certain range, distinct from that under midday sky or of artificial lighting. The exception to this is night, outdoor scenes (which arguably should be treated as indoor scenes for many practical applications).

Flash. Because of the lighting differences described above, (automatic and manual) camera flash is used on a much higher percentage of images of indoor scenes than of outdoor scenes.

Subject Distance. With few exceptions, only outdoor scenes, and landscape images in particular, can have a large subject distance. Therefore, we expect distance measures to discriminate strongly between indoor and outdoor scenes and, to a lesser extent, between types of outdoor scenes.

Focal Length. Focal length is related to subject distance in less direct and intuitive ways through camera zoom. We expect a weak correlation between zoom level and scene type. The zoom-in function is more likely to be used for distant objects outdoors (but can also be used for close-ups in indoor pictures); zoom-out is used for long-distance, outdoor scenery images (and also for indoor occasions, such as group pictures) to expand the view. This effect is more pronounced for cameras equipped with a greater zoom ratio.

2.2. Cue selection using *Kullback-Leibler* divergence

Analysis of specific distributions can help decide which cues are most discriminative for a given problem. The Kullback-Leibler (KL) divergence [2,5] of two distributions, P and Q, is a measure of the disparity between the distributions, given by

$$D_{KL}(P,Q) = \sum_x Q(x)\log\left(\frac{Q(x)}{P(x)}\right)$$

Intuitively, cues that have a greater KL-divergence will be more discriminative. In the case of binary scene classification, P and Q are the cue histograms for the two scene classes, respectively, and x varies over each bin in the histogram (e.g., see Figure 2 and 3 later in the paper). Because the KL-divergence is asymmetric, a better measure is the average of *D(P,Q)* and *D(Q,P)* [5,15], which may be computed using:

$$D_{AvgKL}(P,Q) = \frac{1}{2}\sum_x \left(Q(x) - P(x)\right)\log\left(\frac{Q(x)}{P(x)}\right)$$

We calculate the average divergence for each individual cue's distributions; the maximum average corresponds to the most discriminative cue. Furthermore, KL divergence can be used for joint distributions of variables to find which cue *combinations* are most discriminative. Let P = {P_1, P_2} and Q = {Q_1, Q_2} be two joint distributions over two cues, and let x and y range over the bins of the joint histograms. Assuming cue independence, i.e., $P(x, y) = P_1(x)P_2(y)$ and $Q(x, y) = Q_1(x)Q_2(y)$, yields the following factored form:

$$D_{AvgKL}(P,Q) = \frac{1}{2}\sum_x\sum_y \left(Q_1(x)Q_2(y) - P_1(x)P_2(y)\right)\log\left(\frac{Q_1(x)Q_2(y)}{P_1(x)P_2(y))}\right)$$

Our approach draws from [5], in which KL divergence was used to determine discriminative features. The independence assumption is warranted by the category analyses described in the previous section.

3. Cue integration using a Bayesian network

We chose to use a Bayesian network as a robust method for integrating multiple sources of probabilistic information. First, it is a challenge to find a way to combine diverse evidence, measured by different means, and represented by different metrics. For example, color features are represented by histograms, and the presence of flash is Boolean. A probabilistic evidence-fusion framework would allow all the information to be integrated in common terms of probabilities. Second, domain knowledge is crucial for a visual inference process because it can bridge the “sensory gap” and “semantic gap” by including “human in the loop”, and serves as the agent for fusing low-level and metadata cues. Bayesian networks allow domain knowledge to be incorporated in the structure and parameters of the networks, which is more difficult – if not impossible – for other inference engines such as neural networks or support vector machines. Last but not the least, Bayesian networks are capable of handling incomplete information gracefully because the nodes corresponding to missing evidence are not instantiated, and no retraining of the network is needed. We use the topology shown in Figure 1.

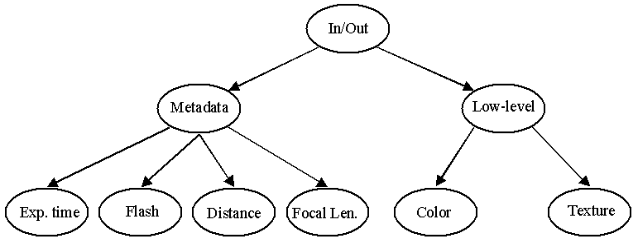

Figure 1. Bayesian network for evidence combination.

A Bayesian classifier according to the *maximum a posteriori* (MAP) criterion gives image classification *c'* by:

$$c' = \arg\max_i P(c_i \mid M, L) = \arg\max_i P(M \mid c_i) P(L \mid c_i) P(c_i)$$

$$= \arg\max_i \prod_{j=1}^{4} P(M_j \mid c_i) \prod_{k=1}^{2} P(L_k \mid c_i) P(c_i)$$

where M = metadata cues, L = low-level cues, P(c) = prior.

The low-level input is pseudo probabilistic (e.g., generated by treating the outcome of a neural network as probabilities (after suitable normalization), or by applying a sigmoid function to the output of a Support Vector Machine [17]). The metadata input is either binary (e.g., flash fired) or discrete (e.g., exposure time is quantized into discrete intervals). One advantage of Bayesian networks is that the conditional probability matrices (CPM) connecting the cues to the network can be set manually or learned from data.

Figure 1 shows only a few of the potential input cues that could be used for metadata. For indoor-outdoor scene classification, they are the best cues from each of the categories discussed previously. When used, nodes for other metadata, such as brightness value, are *substitutes* for the ones shown and are never used simultaneously because they are correlated.

Bayesian networks are very reliable in the presence of (either partially or completely) missing evidence. This is ideal when dealing with metadata because some of the metadata tags, e.g., subject distance, are sometimes not given a value by many camera manufacturers.

4.Problem 1: Indoor-outdoor classification

Our baseline low-level classifier is similar to [13,16], extracting color histograms and wavelet (texture) features in a 4x4 block configuration, classifying each using an SVM, summing the outputs over all blocks, and shaping the sum using a sigmoid into a pseudo-probability. We trained it on an independent set of film and digital images not used elsewhere in this study.

Our image database consists of 24,000 pictures broadly spanning “photo-space” in terms of image content [11]: 56 photographers from 3 U.S. cities (Atlanta, Los Angeles, and Chicago) took over 24,000 pictures over the course of 12 months, all using a single camera model that provides complete metadata. To learn the scene-class distributions, we randomly selected a subset, D1, of 3071 images (1564 indoor and 1507 outdoor) such that equal proportions of the subset were taken from each of the 3 locations.

Our test set, D2, consists of 2049 (1205 indoor, 844 outdoor) manually-labeled images and contains no semantically ambiguous images (e.g., images where the photographer standing indoors took a picture through a doorway to the outdoors, or vice versa).

4.1. KL-divergence analysis

From our discussion in Section 2.1, we computed the KL divergence for following cues: subject distance (SD), focal length (FL), exposure time (ET), aperture value (AP), f-number (FN), and flash fired (FF). We dropped shutter speed, since it is closely related to (and recorded less reliably than) exposure time. Results are given in Table 1.

Table 1. Statistical evidence for cues and cue combinations.

Cue	D(P,Q)	D(Q,P)	Average	(Rank)
SD	0.341	0.420	0.380	7
FL	0.022	0.021	0.021	10
ET	3.450	1.167	2.308	4
AP	0.225	0.409	0.317	8
FN	0.180	0.299	0.239	9
FF	1.193	1.411	1.302	6
ET + FF	4.643	2.578	3.611	2
ET + SD	3.790	1.587	2.689	3
SD + FF	1.534	1.831	1.683	5
ET + FF + SD	4.984	2.998	3.991	1

For individual cues, exposure time has the greater KL divergence and is thus most salient, followed by flash and subject distance. Other cues, such as focal length, have a low average divergence, and thus do not discriminate indoor from outdoor scenes as well.

We chose a greedy approach to cue combination, combining only the most salient cues from each family. To compute the KL divergence of joint distributions, we assumed cue independence based on the analysis discussed earlier. As expected, the cue combinations tend to have higher KL-divergence values. Note that the four largest KL-divergence values are for the combinations including exposure time, which is so salient that, even alone, it has greater divergence than the combination of subject distance and flash. The highest KL divergence is for the combination of a single cue from each of the first three main families.

4.2. Cue distributions for indoor-outdoor images

Figure 2, Figure 3, and Table 2 lend insight into the saliency of the top cues for each family. Figure 2 shows the distributions of exposure times. Those over 1/60 (0.017) second are more likely to be indoor scenes because of lower lighting. Figure 3 shows the distribution of subject distance. Most indoor scenes have a distance of between 1–3 meters, while outdoor scenes have a relatively flat distribution of distances. The graph shows some effects of quantization. However, while subject distance may not be perfect, even estimates can help classification performance. Table 2 shows camera flash statistics on the data set.

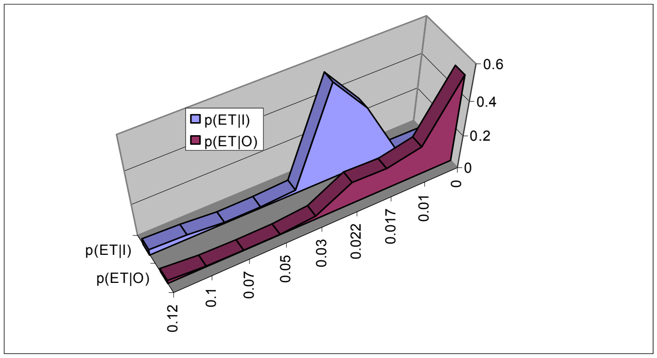

Figure 2. Distribution of exposure times (ET) of indoor and outdoor scenes.

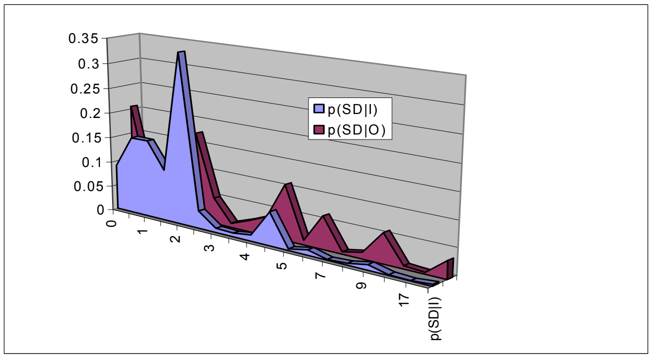

Figure 3. Distribution of subject distance (SD) of indoor and outdoor scenes.

Table 2. Distribution of flash in indoor and outdoor scenes.

Class	P(on | scene class)	P(off | scene class)
Indoor	0.902	0.098
Outdoor	0.191	0.809

4.3. Experimental results

To evaluate the performance of metadata cues and combinations, the conditional probability matrices on the links to the metadata nodes are taken directly from the distributions such as those in Figure 2, Figure 3, and Table 2. We set the prior probabilities at the root node to be equal per the “photo space”. Once evidence is propagated to the root node, images with belief values above a threshold, T, are classified as outdoor images. While T = 0.5 is natural, it may be varied to obtain other operating points.

Bayesian networks are robust in the face of missing data. Figure 4 presents recall rates of indoor and outdoor images (varying T) using the same cues and cue combinations as Table 1. Table 3 compares select individual cues with cue combinations. It is important to note that the accuracy in ranking is similar to the cue ranking given by the KL-divergence measure (Table 1), with exposure time and flash being the strongest individual cues and cue combinations giving higher performance than individual cues. One exception is that flash is a stronger cue empirically. In general, the empirical performance agrees with the statistical measure. Slight discrepancies may be attributed to the fact that the features may not be completely independent (e.g., flash and exposure time), or that the KL-divergence is not necessarily a perfect measure of the empirical classifier performance even if the features are perfectly independent (there is no guarantee of *exact* correlation between class separability and a particular classifier).

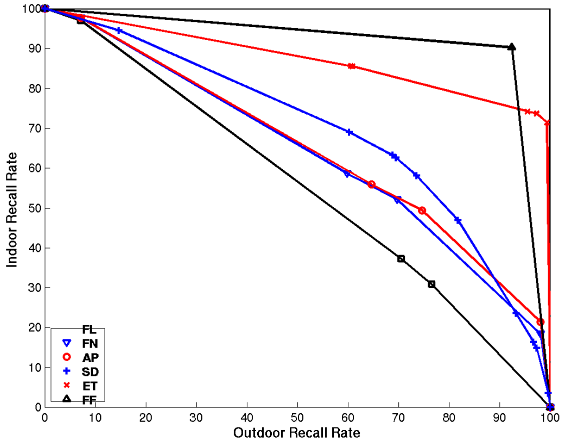

Figure 4. Comparison of individual metadata cues.

What is the effect of adding evidence obtained from our low-level classifier? Figure 5 shows two results: (1) metadata cues alone can outperform low-level cues alone, and (2) the combination of both (complementary) types of cues is most successful. Table 4 shows accuracies of selected combinations at T = 0.5.

Table 3. Accuracy using metadata cues and combinations.

Cue	Indoor Accuracy	Outdoor Accuracy	Total Accuracy
SD	73.53%	58.18%	67.20%
ET	99.34%	71.45	87.85%
FF	92.36%	90.28%	91.51%
ET + FF	93.11%	90.88%	92.19%
ET + SD	95.60%	77.01%	87.95%
SD + FF	91.95%	90.28%	91.26%
ET + FF + SD	93.94%	88.03%	91.51%

Table 4. Accuracy when low-level cues are added.

Cue	Indoor Accuracy	Outdoor Accuracy	Total Accuracy
Low-level (LL) only	83.15%	77.96%	81.02%
LL + ET + FF	96.10%	90.40%	93.75%
LL + ET + FF + SD	96.02%	91.47%	94.14%

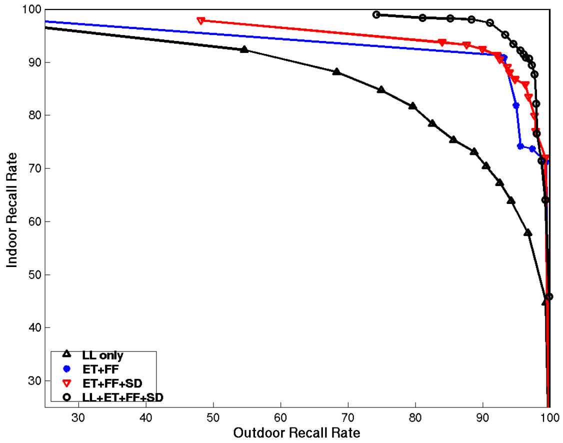

Figure 5. Comparison of performance using low-level, metadata and combined cues.

4.4. Simulating the availability of metadata

Data set D2 is representative of the scene content – but not the metadata availability – of consumer images. While all of our images contained full metadata, a more accurate measure of performance of our system should take missing metadata into account. To this end, we obtained from a major on-line photo service provider the availability of various metadata tags by various camera models.

Simulating the statistical availability of metadata, we obtained the results shown in Table 5. In particular, we used the same images of D2 but, for each image, we first sampled the statistics to determine if the image contained metadata (71%). If so, we further sampled the statistics to determine which tags were present, restricting ourselves to flash, exposure time, and subject distance, giving only "none", FL, FL+ET, FL+SD, and FL+ET+SD as possible combinations of metadata (flash was present in 100% of the images with metadata). All simulated metadata evidence was presented to the Bayesian network. Note that we ran the simulation 20 times to generate reliable results.

Table 5. Simulated performance with missing metadata tags. Note that the first 3 rows simply indicate the likelihood of having each of the 3 metadata cues available.

Statistic	Mean	Standard Deviation
Has Flash	71.03%	1.00%
Has Exposure Time	69.59%	1.17%
Has Subject Distance	21.15%	0.72%
Indoor Accuracy	94.00%	0.49%
Outdoor Accuracy	85.53%	1.12%
Total Accuracy	90.51%	0.61%

The overall accuracy is just over 90%, closer to the best-case scenario (94% with all tags available) than the worst-case (81% with no tags) scenario. This is a more realistic estimate of how the algorithm might perform with general consumer images.

The complexity of the classifier depends dramatically on which cues are used for inference. The average runtime of the full system (using low-level and metadata cues) is 1.402 sec./image on a 502 MHz SunBlade, with all but 0.032 sec used by the low-level feature extraction and classifier.

Based on the complexity of low-level cues and the availability of these statistics, our system offers a "lite" option, which only invokes the low-level cues when no camera metadata is present. On average, this system ran in 0.45 sec./image and obtained accuracy of 89%. However, the performance of this "lite" system, while impressive, should be taken with a grain of salt: it is unlikely that metadata alone would perform as admirably for other problems, such as differentiating among outdoor scenes.

Table 6. Number of images in each category from D1.

Category	Indoor	Outdoor
Correct by both	982	606
Gained by metadata	175	166
Lost by metadata	20	52
Incorrect by both	28	20
Total	1205	844

4.5. Discussions of indoor-outdoor classification

Metadata and content-based cues are complementary, capturing different information useful for deciding the semantic scene category of an image. Which types of images are suited to analysis using combined (metadata + low-level) cues, and which can be classified correctly using low-level cues alone? We compared accuracy by the low-level detector to accuracy using all cues (LL + FL + ET +

SD) on data set D2. We have broken down the indoor and outdoor categories further into the following subcategories (Table 6): correct (by both); correct by combined cues, but not by low-level cues ("gained" by metadata), correct by low-level cues, but not by metadata ("lost" by metadata), and incorrect (by both). Figures 6 and 7 include example images from each of the above categories.

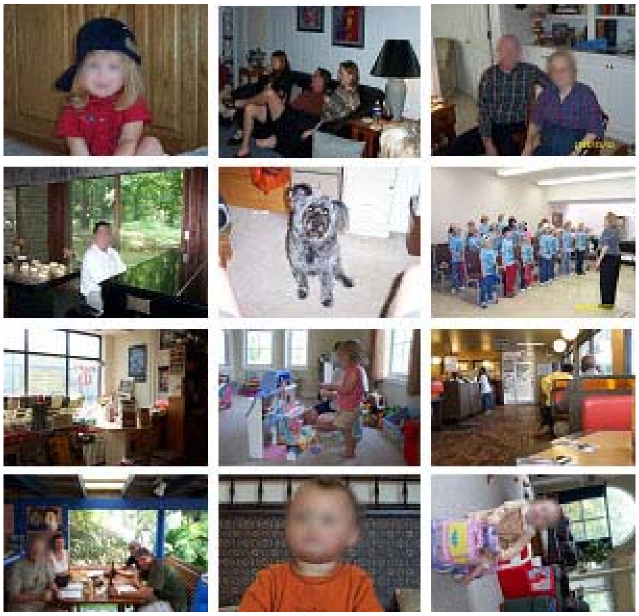

Figure 6. Indoor image samples, classified correctly by both (Row 1), gained by metadata (Row 2), lost by metadata (Row 3), and incorrectly, regardless of cues (Row 4).

The indoor images gained include primarily "non-typical" indoor scenes (e.g., aerobics classes, close-ups of animals, or other objects, like vases) with flash (the only two images that did not use flash had borderline low-level belief already: 0.50–0.51) and a longer exposure time. Many also had a short subject distance, being close-ups. The indoor images lost primarily included longer-distance indoor scenes and those with ample external lighting (none used flash). Many of the indoor images that were misclassified, regardless of any combination of cues, were those of rooms with large windows, giving outdoor-like lighting and color/texture cues typical of outdoor images (e.g., green trees visible through the window). Furthermore, only one of these images (one under a picnic pavilion) used flash.

The outdoor images gained through the use of metadata were primarily those with bright (yet overcast) skies. In these cases, the color distribution of gray sky may be mistaken for indoor structures (e.g., ceilings), but the short exposure time and lack of flash (all except one image) are strong outdoor cues. Outdoor images that were lost included primarily scenes with greenery but with a longer exposure time. Flash was also used in every scene in this category. Outdoor images incorrectly classified varied greatly, but usually included man-made structures (e.g., walls), which usually occur indoors. None of these images includes sky and all have longer exposure times; flash use was varied.

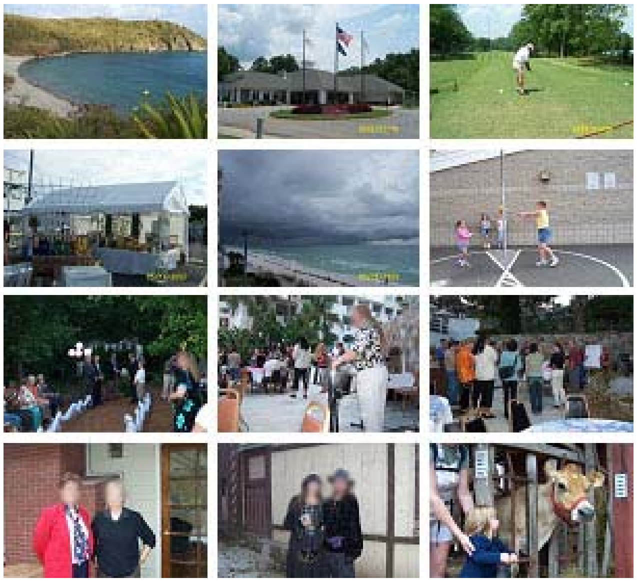

Figure 7. Outdoor image samples (same notation).

Figure 8 shows that, when all cues are used, the final belief value is a good measure of confidence because there is little overlap between the two classes. This makes our system superbly amenable to a reject option by thresholding the belief values. Figure 9 shows accuracy vs rejection rate and verifies this claim: the accuracy is 97.5%, if 10% of the images are left unclassified ("rejected").

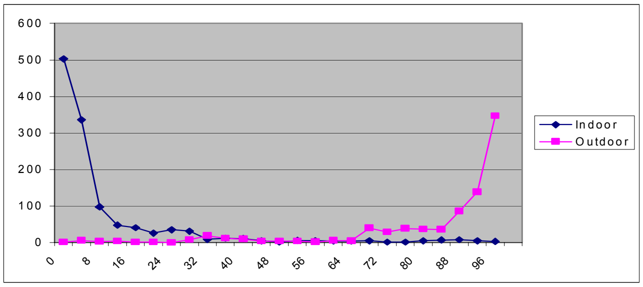

Figure 8. Distributions of beliefs for indoor and outdoor scenes shows that belief is a good measure of confidence.

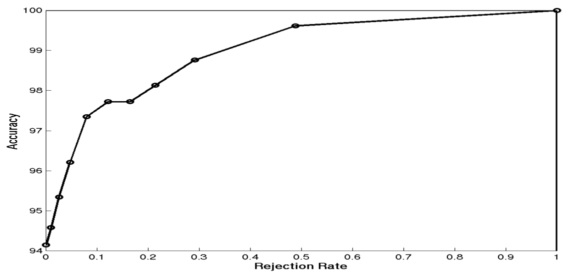

Figure 9. Accuracy vs rejection rate obtained by thresholding the final beliefs.

5. Problem 2: Sunset scene detection

To further demonstrate the efficacy of the proposed Bayesian cue-fusion scheme, we also applied it to the problem of sunset detection. Our low-level classifier was a SVM using spatial color-moment features trained on an independent set of images not used elsewhere in this study (see [1] for details). For testing data, we used a separate set of 4678 personal images containing 191 (4%) sunsets. Analysis of KL divergence showed that flash fire, focal length, aperture, and subject distance were the most salient cues for sunset detection. A Bayesian network similar to that in Figure 1 was constructed and trained. Figure 10 shows the performance of a content-only classifier compared to one augmented with metadata. The benefit is clear: we see that using metadata can increase the true positive rate by as much as 15% for a given false positive rate.

As we did for indoor-and outdoor images in the previous section, we now discuss typical sunset and non-sunset images. Figure 11 shows sample sunset images classified correctly by both low-level and combined cues, incorrectly by both sets of cues, and those gained and lost by adding metadata. The images correct by both methods are typical salient sunsets. Those gained by adding metadata cues include ones with little warm color and those that contained occluding regions. Only one sunset image was lost by adding metadata because of the nonstandard aperture used in the low light. The sunsets that were always classified incorrectly include one with people as the main subject (giving a small subject distance) and those with predominantly bluish or dark colors.

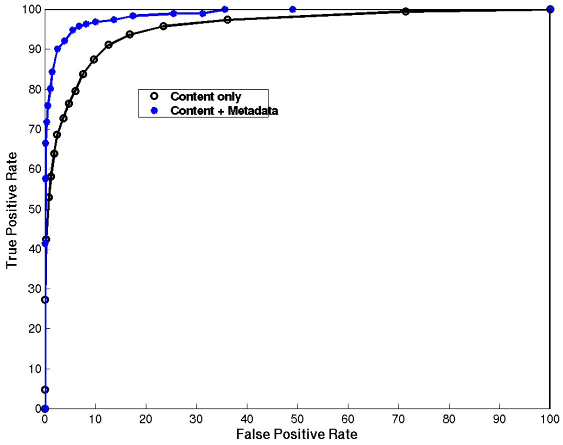

Figure 10. Performance of content-only vs metadata-enhanced sunset detection. As an example, at the 0.5 threshold, sunset recall rose from 79.6% to 94.8%, while the false positive rate dropped slightly from 6.0% to 5.5%.

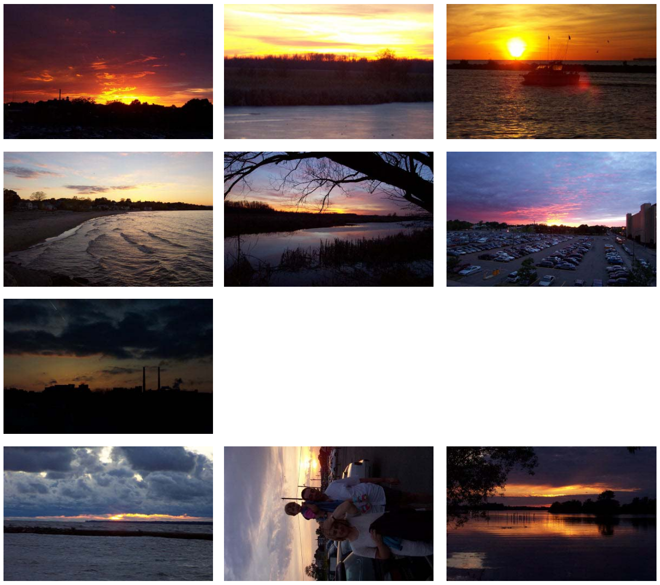

Figure 11. Sunset image samples, classified correctly by both (Row 1), gained by metadata (Row 2), lost by metadata (Row 3), and incorrectly, regardless of cues (Row 4). Only a single sunset image was lost by metadata.

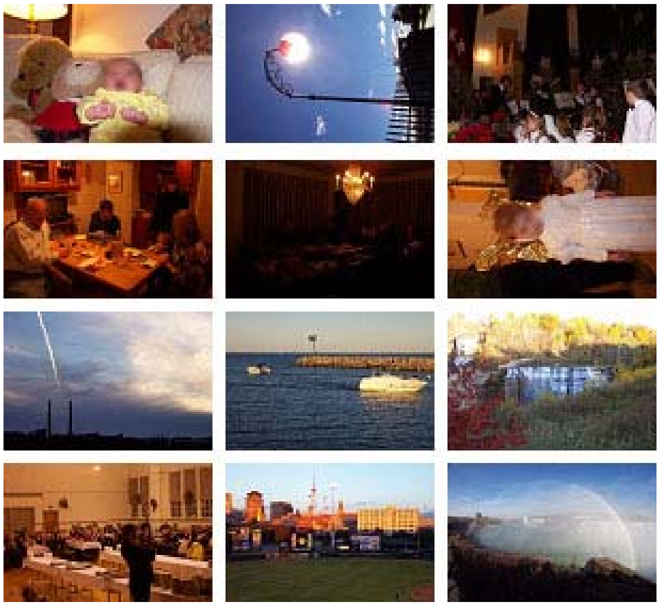

Figure 12. Non-sunset image samples, classified correctly by both (Row 1), gained by metadata (Row 2), lost by metadata (Row 3), and incorrectly, regardless of cues (Row 4).

Figure 12 shows example non-sunset images. Those classified correctly, regardless of cues, span a wide variety of images; therefore only examples of those containing potentially confusing content are shown. Without metadata cues, indoor images under low incandescent lighting can be mistaken for sunsets because of their similar color distributions; however, their (short) subject distance can often be used to disambiguate them. Likewise, outdoor non-sunsets weakly classified as such can be pushed over the threshold because of their long subject distance and lack of flash. Those non-sunsets classified incorrectly as sunsets regardless of cues include indoor images with strong low-level evidence for sunset, outdoor images that occur under sunset lighting (but facing the opposite direction), and a landscape including a rainbow.

6. Conclusions and future work

We have introduced a probabilistic scheme that can be used to fuse multi-modal low-level and camera metadata cues for scene classification. We used KL-divergence as a measure of cue discrimination power and found tags to accurately discriminate indoor from outdoor and sunset from non-sunset scenes. It is advantageous to use as many *independent* tags available. However, the proposed scheme is robust even if one or more tags were missing. This is helpful for on-line image storage for which metadata is often missing. Because metadata has the distinct advantages of being computationally cheap and relatively accurate, it also allows a "lite" indoor-outdoor classifier: ignoring low-level cues or only computing them when metadata is missing.

We applied our model to the problems of indoor-outdoor and sunset scene classification, in both cases increasing accuracy while holding processing time relatively constant.

Interesting directions for future work include generalizing the model to handle multiple scene classes beyond indoor and outdoor and heterogenous camera models. Because some camera manufacturers' metadata is more accurate than others', this may necessitate including confidence values on metadata, which is nontrivial.

References

[1]M. Boutell, J. Luo, and Robert T. Gray. Sunset scene classification using simulated image recomposition. *Proceedings of International Conference on Multimedia and Expo*, Baltimore, MD, 2003.

[2] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 24(8):1026–1040, 2002.

[3] R. Duda, R. Hart, and D. Stork. *Pattern Classification*, 2nd Edition. John Wiley and Sons, Inc., New York, 2001.

[4] A. Hauptmann and M. Smith. Text, speech, and vision for video segmentation: The infomedia project. *Proceedings of AAAI Symposium on Computational Models for Integrating Language and Vision*, 1995.

[5] P. Lipson, E. Grimson, and P. Sinha, "Configuration based scene classification and image indexing," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1997.*

[6] C. Liu and H.-Y. Shum. Kullback-Leibler Boosting. *Proceedings of International Conference on Computer Vision and Pattern Recognition*, 2003.

[7] Y. Lu, C. Hu, X. Zhu, H. J. Zhang, and Q. Yang. A unified framework for semantics and feature based relevance feedback in image retrieval systems. *ACM Multimedia Conference*, Los Angeles, CA, 2000.

[8] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision,* 42(3):145-175, 2001.

[9] S. Paek and S.-F. Chang. A knowledge engineering approach for image classification based on probabilistic reasoning systems. *Proceedings of IEEE International Conference on Multimedia and Expo, 2000.*

[10]J. Platt. AutoAlbum: Clustering digital photographs using probabilistic model merging. *IEEE Workshop on Content-based Access of Image and Video Libraries*, 2000.

[11]R. Segur, "Using photographic space to improve the evaluation of consumer cameras," *Proceedings of IS&T Image Processing, Image Quality, Image Capture and Systems (PICS) Conference*, 2000.

[12]N. Serrano, A. Savakis and J. Luo. A Computationally Efficient Approach to Indoor/Outdoor Scene Classification. *Proceedings of International Conference on Pattern Recognition*, 2002.

[13]J. R. Smith and C.-S. Li. Image classification and querying using composite region templates. *Journal of Computer Vision and Image Understanding*, 75(1/2):165–174, 1999.

[14]Y. Song and A. Zhang. Analyzing scenery images by monotonic tree. *ACM Multimedia Systems Journal*, 2002.

[15]Z. Sun, "On Multiple Cue Integration", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, 2003.

[16]M. Szummer and R. W. Picard. Indoor-outdoor image classification. *Proceedings of IEEE Workshop on Content-based Access of Image and Video Databases*, 1998.

[17]D. Tax and R. Duin. Using two-class classifiers for multi-class classification. *International Conference on Pattern Recognition*, Quebec City, QC, Canada, August 2002.

[18]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, 1999.