Scene Classification of Images and Video via Semantic Segmentation

Heather Dunlop
Digitalsmiths Corporation
5001 Hospitality Court, Suite 100
Morrisville, NC 27560
hdunlop@digitalsmiths.com

Abstract

Scene classification is used to categorize images into different classes, such as urban, mountain, beach, or indoor. This paper presents work on scene classification of television shows and feature films. These types of media bring unique challenges that are not present in photographs, as many shots are close-ups in which few characteristics of the scene are visible.

In our work, the video is first segmented into shots and scenes, and key frames from each shot are analyzed before aggregating the results. Each key frame is classified as indoor or outdoor. Outdoor frames are further broken down by a semantic segmentation which provides a label to each pixel. These labels are then used to classify the scene type by describing the arrangement of scene components with a spatial pyramid.

We present results from operating on a large database of videos and provide a comparison with selected work from the literature on photographs. Evidence of the success of the semantic segmentation is provided on a set of hand-labeled images. Our work improves the semantic segmentation and scene classification of images and, to the best of our knowledge, is the first paper that details a full working system on video.

1. Introduction

Scene classification applied to video identifies the time interval of occurrence of various scene types (*e.g.*, urban, indoor, desert, mountain, open water). Such a system provides context in further video analysis algorithms, either with the use of the final scene type or intermediate information such as the location of water and road within each frame. In the domain of content-based image or video retrieval, a database of media can be cataloged and easily searched for types of scenes.

Many challenges exist for a scene classification system on images. Variations in view-point and lighting can drastically change the appearance of a scene and objects in it. Even with successful recognition of the components of a scene, their spatial arrangement plays an important part in defining its class. The application of an image scene classification algorithm to television shows and feature films has even more challenges. Many shots in such videos are close-ups of characters, often alternating between characters exchanging dialogue. These close-ups are interspersed with wide-angle shots with a good view of the background. In very close-up shots and in action sequences, the background is often blurry, thereby confusing any part of the scene type that is visible. Although some of these issues may occur in photographs, they are less common due to the different objectives of a photographer in choosing the composition of the scene and in discarding photos that do not meet this standard. These types of effects must be accounted for in an application of scene classification to video.

The methods developed in our work take a divide-and-conquer approach by semantically segmenting the video and frames within it. We first break the video into shots and scenes. A shot is a sequence of frames from a single camera and a scene is a sequence of shots that present continuous action or are semantically correlated. Key frames are classified as indoor or outdoor using a two-stage process starting from raw features on regions of the image. On outdoor frames, further analysis is done to understand the scene. From an initial set of segmentations, adjacent segments containing similar materials are merged and the resulting regions are classified into material categories such as sky, water, or building. This assigns a category to each pixel in the image, known as a semantic segmentation. The arrangement of these scene components are characterized by dividing the image into a regular grid of different sizes and computing the distribution of materials in each region. This spatial pyramid is used to provide a final classification for the image. Results are aggregated across shots and across scenes, producing a result for each scene of the video.

In the domain of semantic segmentation, the novel ar-

978-1-4244-7030-3/10/$26.00 ©2010 IEEE

eas of our work include the use of multiple segmentations, merging segments containing similar materials before trying to classify them, and using strong color, texture, edge, line, and shape features.

We apply semantic segmentation to scene classification with the use of Spatial Pyramid Matching [20] rather than the standard computation of a histogram of scene concepts [5, 10, 17, 29, 32, 35]. Another unique aspect of our method is multi-label classification, whereby an image may fall into more than one scene category or none at all, a complication which most prior works do not even attempt (Boutell *et al*. [6] is one of the few that does).

The most unique aspect of our work is the application of scene classification to video. All known prior works in this area simply operate on some subset of frames from the video [17], but do not attempt to summarize the results across the video or deal with the unique challenges in this domain. Our methods handle wide-angle and close-up shots, and summarize results for each scene in the video.

2. Related Work

Published algorithms for scene classification of images often involve dividing the image into a regular grid and classifying the material components of each grid cell [5, 10, 17, 29, 32, 35]. The occurrence of each material over the image is then computed and the scene classified from this vector. Unsupervised materials can also be used by clustering the features extracted from each region into a pre-determined number of classes [15].

Grid methods produce a rough semantic segmentation; however, other works that target this area (but not scene classification) achieve more successful results. Shotton *et al*. model texture, layout, and context with a conditional random field [30]. Starting from a segmentation, various probabilistic models can be applied to classify each region of an image [1, 18, 34]. Yang *et al*. use appearance and a bag of key points model with mean-shift to assign labels to regions of the image [36], while Corso *et al*. apply a graph-shifts method to the problem [11].

In lieu of a semantic segmentation, additional approaches typically use a bag of key points [24, 26] in combination with Latent Dirichlet Allocation [13], probabilistic Latent Semantic Analysis [2], Spatial Pyramid Matching [20], or a combination of such techniques [3, 8]. A 2D Hidden Markov Model has also been attempted [21] as well as a wavelet coefficients representation of features with a hierarchical Dirichlet process hidden Markov trees [19].

Examining the power spectrum of an image is another approach taken. By using the power spectrum and Principal Components Analysis, the type of scene can be determined [31]. However, this approach is specific to photographs where care is taken in framing the scene from a distance with the horizon visible across the image.

Some methods only tackle simple tasks such as indoor vs. outdoor, city vs. landscape, or detecting a sunset [4, 22, 28, 32]. The two-stage indoor/outdoor classification approach by Serrano *et al*. [28] is also used in our work.

Few prior works on scene classification of video exist. Israel *et al*. [17] use scene classification results on images to produce results on video by selecting key frames to classify, but do not further describe how the scene results from each frame are used. Bosch *et al*. [3] show results on a selection of frames from the film Pretty Woman, while Shotton *et al*. [30] produce semantic segmentation results on a selection of frames from television shows. However, these results are purely anecdotal as this is not the main focus of their work.

3. Approach

Our method starts by detecting shot and scene boundaries in the video. Key frames are extracted from each shot for processing. Key frames that are too dark are discarded. Remaining frames are classified as indoor, outdoor, or undetermined. Undetermined frames include close-up shots in which little background is visible. In the case of an outdoor result, the frame is segmented and the materials within are classified to produce a semantic segmentation. The spatial arrangement of these scene components are then characterized and used to classify the outdoor scene type. The results on individual frames are combined across shots and scenes to produce a final set of results for the video.

3.1. Semantic Segmentation

A semantic segmentation consists of breaking an image into regions and categorizing each into a set of pre-defined classes. We refer to the semantic categories as materials and have selected the following groups: building, grass, person, road, rock, sand/gravel, sky/clouds, snow/ice, trees/bushes, vehicle, water, and miscellaneous. All segments in the image are classified as one of these materials.

Figure 1 provides an overview of the algorithm used. Multiple segmentations of the image are generated. Segments are merged together when they are likely to belong to the same material class. Each segment is classified according to its local features and the final result is generated by averaging the results across the set of segmentations.

We first provide the background details on features before proceeding with their usage in semantic segmentation.

3.1.1 Feature Extraction

Given a region or segment of any shape in an image, a number of features can be extracted to characterize its appearance. We represent color in CIELAB space. K-means is used to compute a dictionary of colors, then a histogram represents the distribution of colors in a region. We compute edge strength and edge direction histograms by using

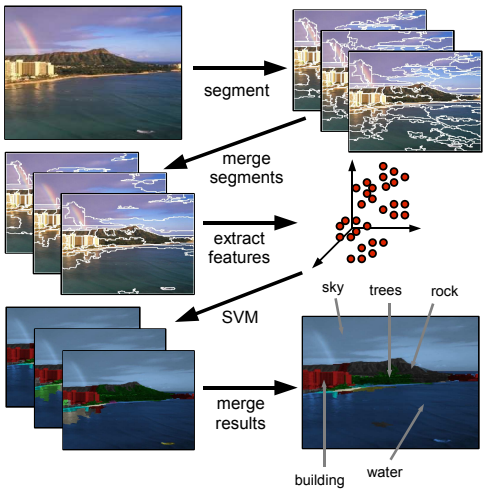

Figure 1. Overview of semantic segmentation. Multiple segmentations are produced and similar segments are merged. From extracted features, segments are classified and the labels combined across segmentations.

the Sobel transform. A line length histogram is formed with Hough transform to detect lines. Texture takes the form of a texton histogram [33]. Shape is represented by circularity, convexity, distance to the best fit polygon, and angularity, all computed from the boundary of the region [12]. The results for each of these features are concatenated together, forming a feature vector to characterize the appearance of the region.

3.1.2 Segmentation

Image segmentation is a difficult task and no single result is likely to be correct. However, each result performs accurately on some portion of the image and together they can be used to obtain better performance [16, 23]. The image is resized to 720×480 pixels, or one with roughly the same number of pixels that maintains the original aspect ratio. Segmentations are created using Efficient Graph-Based Segmentation [14] with three different sets of parameters.

Some of the segments produced consist of only part of an object or larger region of material. We achieve a better result in classification if some of these smaller segments are merged together. Adjacent segments are used to train a Random Forest [7] as an affinity classifier. We use positive examples of adjacent segments that belong to the same material class and negative examples of adjacent segments that belong to different material classes. The feature vector for each segment is computed as the color, edge, line, texture, and shape features described in Section 3.1.1. The absolute value of the element-wise difference between the feature vectors is used as the feature set for the classifier. We can now compute the difference between the feature vectors of two adjacent segments and decide whether to merge them based on the affinity score produced by the classifier.

An affinity score is calculated for each pair of adjacent segments. If the highest affinity score is above a predetermined merging threshold, then the associated pair of segments are merged into a single segment and the feature vector of all affected pairs of segments is recalculated. The merging threshold is computed during training, such that for all data points producing a score above this threshold, a precision of 95% is attained. The merging and recalculation process is repeated until the highest affinity score is no longer above the threshold. The procedure is followed independently for each segmentation.

3.1.3 Material Classification

The next step is to use the feature vectors to develop a classifier that predicts the material class given an image segment. The color, edge, line, texture, and shape features are computed on the segment. Each element in the feature vector is normalized to fall between zero and one. Each segment has its material labeled by hand as ground truth. The labeled feature vectors are used in training a multi-class Support Vector Machine (SVM) with a radial basis function kernel [9]. Thus, any novel image segment can be classified by extracting features and predicting its class label with the SVM.

The resulting material scores are averaged across the three different segmentations at each pixel. Now to assess the material results over a specific region (not necessarily a segment), the material scores from each pixel in the region can be averaged to produce a material occurrence vector. To produce a semantic segmentation, the material result with the greatest score is selected for each pixel.

3.2. Scene Classification

Now that we have detected the presence of different materials as components of a scene, we turn to the problem of classifying the category of the whole image. Images are first classified as indoor, outdoor, or undetermined. Outdoor images are then further classified with a semantic segmentation and the arrangement of their components characterized with a spatial pyramid. Figure 2 provides an overview of this process. The outdoor scene categories chosen in our work are coast/beach, desert, forest, grassland, highway, lake/river, mountainous, open water, sky, snow, and urban. We treat this as a multi-label problem by allowing an image to belong to any number of these categories or none at all.

3.2.1 Indoor/Outdoor Classification

We wish to classify each image as indoor, outdoor, or undetermined. An undetermined image is when an observer cannot clearly tell whether the camera is viewing an indoor

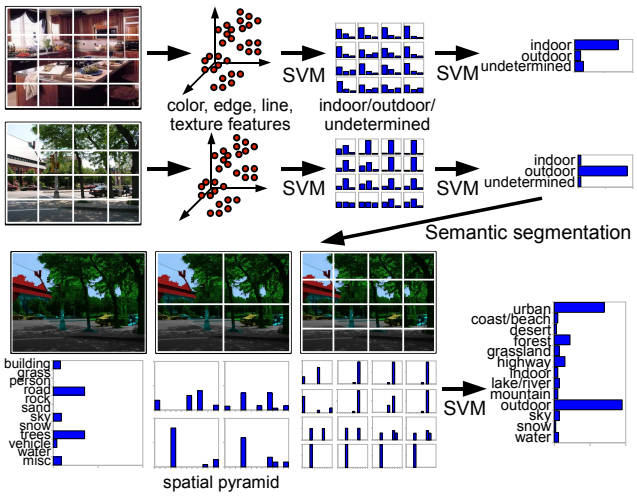

Figure 2. Overview of scene classification on images. A two-stage SVM classification is used to classify the image as indoor, outdoor, or undetermined. On outdoor images, a semantic segmentation is performed and used in a spatial pyramid to compute scores for outdoor categories.

or outdoor scene. An example of this case is a close-up shot where little of the background is visible, and people or objects are taking up most of the foreground.

Each image is split into a 4×4 grid. The color, edge, line, and texture features are computed separately on each grid cell to produce a set of feature vectors. These feature vectors, along with the label of indoor, outdoor, or undetermined for the image, are used to train an SVM with a radial basis function kernel. Given a feature vector for a rectangular portion of an image, the classifier will predict whether the region belongs to an indoor, outdoor, or undetermined image. The results from this classifier on each grid cell in the image, consisting of a score between zero and one for each of the three categories, are concatenated and then used as a feature vector for another SVM classifier with a linear kernel which predicts the class of the whole frame.

This two-stage classification is used on each frame of the video being processed. If the frame is indoor or undetermined, it is the final result for the frame. If it is outdoor, further processing is done to narrow down the class.

3.2.2 Outdoor Scene Classification

In classification of an outdoor scene, a semantic segmentation is first produced following Section 3.1. From this, each pixel will have a score for each material category with the largest score representing the most likely one.

A material occurrence vector is computed for a region in an image by averaging the score vectors for each pixel as a measure of the prevalence of each material. For example, this would describe the proportion of water, sand, sky, etc.

To characterize the placement of materials in the image, a spatial pyramid of material occurrence vectors is computed as was shown in Figure 2. We follow the method of Lazebnik *et al*. [20], placing a sequence of increasingly finer grids over the image and computing the material occurrence vector for each grid cell. Three levels of this pyramid are used, with grid sizes of 1×1, 2×2, and 4×4. The vectors are weighted such that those at a finer resolution receive a larger weight; we use weights of 1/4, 1/4, and 1/2, respectively. Thus, we have one material occurrence vector for the 1x1 grid, 4 for the 2×2, and 16 for the 4×4, each weighted accordingly. After concatenating these, we use a histogram intersection kernel with an SVM to classify the image. Note once again that an image may belong to more than one class or none at all. Thus, a classifier is trained independently for each category using positive and negative examples for the class. The set of trained classifiers can now produce a score from zero to one for each scene category on a novel image.

3.3. Video

Now turning to the application of video, we apply the concepts developed on images in combination with techniques unique to video. The first step we take in processing a video is to segment it into shots and scenes. Key frames are extracted from each shot, dark frames are discarded, and the scene is classified in each remaining key frame. The results are then combined across shots and then across scenes to produce a result for each scene.

3.3.1 Segmenting Video

We first need to segment the video into shots and scenes. Shot and scene boundary detection are performed by the algorithm outlined by Rasheed and Shah [25]. Key frames can be extracted from each shot also by the method of Rasheed and Shah [25], or simply by selecting a set of equally spaced frames from each shot. We chose the later for simplicity, using five equally spaced frames per shot.

3.3.2 From Frames to Shots to Scenes

If a shot is too dark, this algorithm cannot accurately classify the scene. Thus, we perform a quick first step to discard such frames.

Taking the remaining frames, scene classification is performed on each following the method in Section 3.2. In a typical shot, a single scene is within view. Characters and objects may move and the camera may also change its field of view. Our assumption is that a simple method of averaging results from the key frames within the shot will produce a reasonable result. Thus, for each class, the scores are averaged across the key frames in the shot.

Many shots within a video tend to be close-up, where little background is visible and the results of the scene clas-

sifier are not meaningful. This problem tends to be consistent across a single shot, but varies from shot to shot in a scene. Often 80 or 90% of the shots in a scene are close-ups. Thus, averaging the results across a scene the way we did across a shot does not produce a useful result. Taking the maximum result of the scores for each class is one possible solution. We have achieved better results by taking the 95th percentile result, interpolating if necessary. For class c, let the scores for the N shots in a scene be represented by $s_1, s_2, ..., s_k, ..., s_N$, sorted in an ascending manner. The percentile of each score can be computed as $p_k = \frac{100}{N}k$, where k is the index in the ascending list of scores. If there exists a p_k that equals 95, we take the result s_k. Otherwise, we interpolate by finding values p_k and p_{k+1} such that $p_k \leq 95 \leq p_{k+1}$ and take the result $s_k + \frac{N}{100}(95 - p_k)(s_{k+1} - s_k)$.

Resulting scores for each scene class can be compared with a pre-determined threshold, selecting only those greater than this value. The threshold for each class is learned by assessing the performance on a set of test videos and selecting the threshold that achieves the desired performance for the application.

4. Experiments

This section discusses the results from a set of experiments used for assessing the performance of the algorithm. We first look at the accuracy of semantic segmentation. Next, we provide results on scene classification of photographs, including a comparison of our method with that of Lazebnik *et al*. [20]. Finally, we produce results on a large video data set, establishing that the methods described herein have successfully been applied to production video.

4.1. Semantic Segmentation

The semantic segmentation algorithm requires a ground-truth labeling of segments in images. 1019 images were segmented and each segment hand-labeled as one of the material categories. These images were obtained from the LabelMe database [27], Google Image Search, and frames from a selection of movies. Five-fold cross-validation is used to train on 80% of the data and evaluate results on the remaining 20%. This is done five times, averaging the results over each try. A confusion matrix is used to analyze the results, as shown in Table 1. For each material, it shows the percentage of the time that a segment labeled as that material is correctly classified and the percentage that it is incorrectly classified as each of the other materials.

The most accurately classified material is sky/clouds which was performed correctly 91% of the time. The most common misclassification is snow/ice, incorrectly classified as water 28% of the time. The results are partially dependent on the quantity of training data per class. For example,

True Class \ Predicted Class	building	grass	person	road/sidewalk	rock	sand/gravel	sky/clouds	snow/ice	trees/bushes	vehicle	water	miscellaneous
building	71	1	6	1	2	2	5	0	2	7	1	3
grass	1	71	1	0	3	7	0	0	11	0	3	2
person	5	1	75	2	1	0	2	0	4	3	1	6
road/sidewalk	7	1	3	52	2	12	0	3	0	4	15	1
rock	4	4	7	3	47	16	0	0	13	1	1	4
sand/gravel	4	9	3	3	14	48	2	6	3	1	2	6
sky/clouds	2	0	0	2	1	1	91	1	0	0	3	0
snow/ice	4	0	2	6	2	2	13	41	0	2	28	1
trees/bushes	4	3	2	0	3	1	2	1	77	1	3	4
vehicle	22	0	15	2	4	0	0	1	1	51	0	3
water	1	3	1	3	1	11	8	4	1	1	66	1
miscellaneous	13	6	26	1	7	2	3	0	4	13	1	23

Table 1. Material confusion matrix. Average classification rates for individual classes are listed along the diagonal. The off-diagonal entries represent the mis-classification rates.

a smaller set of data was used to train rocks and snow/ice, while a larger amount was available for sky and trees.

Classification errors are common in images with explosions, fire, smoke, fog, or smog. These translucent obstructions result in the background being partially visible, but not clear. Regions containing smoke or fog are commonly misclassified as sky. Fall leaves and dead grass are examples in which the materials were not trained in the altered color state, resulting in errors. Blurred backgrounds due to high motion or an unfocused camera also produce incorrect results.

4.2. Scene Classification on Photographs

For computing results on scenes, all of the labeled images described in the previous section are used to train the material classifier. 9855 images from the LabelMe database [27], Google Image Search, and frames from movies were labeled with the appropriate scene class. There is no overlap between this image set and that used for materials. Recall that an image may belong to any number of classes. To evaluate the results of this method, we use five-fold cross-validation and the measures precision and recall. Precision represents the percentage of detected scenes that truly belong to the category, while recall measures the portion that are detected rather than missed. A precision-recall curve shows the results as the detector threshold is varied. To provide a simple summarization, the area under the precision-recall curve is computed, known as average precision.

We compare two methods: that detailed in this work and a bag of key points model with spatial pyramid matching by Lazebnik *et al*. [20], both run on our data set. We use code

	Our Method	Lazebnik *et al*. [20]
Coast/Beach	**.60**	.44
Desert	**.76**	.48
Forest	.71	**.84**
Grassland	**.79**	.56
Highway	.67	**.79**
Lake/River	**.44**	.42
Mountainous	.73	**.81**
Open Water	**.70**	.67
Sky	.82	**.83**
Snow	**.75**	.69
Urban	**.90**	.87
Outdoor	.94	**.99**
Indoor	.73	**.87**
Average	**.73**	.71

Table 2. Performance comparison on image data set, measured as the average precision.

provided by Lazebnik, with a dictionary of size 400 and a 3-level pyramid. Results from both are shown in Table 2.

Examining the results with our method, the outdoor and urban categories achieve the greatest accuracy. This is partly due to the fact that there are more training images for these categories. Lake/river achieves quite poor results. This category is very difficult as the reflection of surrounding terrain on to a small body of water produces a confusing appearance, much different than the blue water of an ocean.

Our data set contains a lot of variability for each scene category, including close-up images and ones that belong to multiple categories (or none at all). For close-up views of characters in a car, the correct classification is ambiguous. The classifier sometimes produces an indoor result and sometimes outdoor, depending on the visible background.

In comparing our results with Lazebnik *et al*. [20], we found that our method achieves significantly higher performance on some categories (up to a 28% increase). Their greater success on the indoor category is likely due to the SIFT descriptors and bag of words model better capturing the variability of indoor scenes. Our method's success on coast/beach, desert, and grassland is likely due to the classification of water, sand, and grass materials, so that the location of these specific scene components are used in classifying the scene. In addition, Lazebnik *et al*. do not use color in their model. Overall, we have shown that our method using semantic segmentation performs comparable to a state-of-the-art method on our challenging data set, with exceptionally large gains in some categories.

4.3. Scene Classification on Video

The scene classifiers were trained using the entire data set described in the previous section. Now to evaluate it on video, we use a set of 281 videos from 49 different television shows consisting of 110 hours of content, as well as six feature films. Each video was segmented into shots and scenes, and the class of each scene was labeled by hand.

We present a comparison of performance on individual key frames versus the result after aggregating over shots and scenes. Table 3 shows the average precision for each class. By aggregating results using our method, the performance increases by 50% - clearly a necessary step when operating on video.

With the exception of the indoor category, the precision and recall scores achieved on video were lower than on our image data set. This emphasizes that the image data set is not truly representative of the variability seen in video.

When taking a photo, much care is taken in selecting the view point and lighting. Landscape photos typically have little foreground to occlude the background. Any people present are placed such that other important parts of the scene are still readily visible. In addition, many photos are discarded, leaving the few remaining that meet the expectations of the photographer. Different objectives are considered in videos for television shows and movies. Typically, the character speaking takes up most of the view and the background is of lesser importance. Often very few of the frames in a scene have enough background visible to enable classification. In our data set, 70% of the frames analyzed were close-up views. All of these factors led to the lower performance with video.

We also provide a sampling of results by running the algorithm on video, as shown in Figure 3. For each class, sample frames from a scene are shown. Each was classified as indoor, outdoor, or undetermined. In the case of an outdoor result, a semantic segmentation was performed and the frame classified. Through combining the results from these frames and others in the scene, a correct classification was achieved.

5. Conclusions and Future Work

We have developed a system that integrates segmentation, recognition of scene components, and classification of whole images and video sequences. Many of the challenges in scene classification of video were explored. Techniques addressing the unique properties of video content are a necessity. Further improvements could still be made to address the numerous close-up shots present in video. Face and body detection and tracking could be used to identify when the background is obstructed. Background segmentation with the use of motion cues might also provide assistance. With the current system, only a few shots from each scene end up providing results for the final classification. Other techniques can be developed to extract useful information from a larger portion of the shots in a scene.

As other works have achieved success with bag of key points techniques [3, 20], we would like to investigate their

	Key Frames	Scenes
Coast/Beach	.13	**.34**
Desert	.04	**.09**
Forest	.29	**.45**
Grassland	.32	**.47**
Highway	.16	**.33**
Lake/River	.02	**.07**
Mountainous	.05	**.11**
Open Water	.33	**.52**
Sky	.24	**.34**
Snow	.04	**.08**
Urban	.33	**.62**
Outdoor	.67	**.86**
Indoor	.72	**.82**
Average	.26	**.39**

Table 3. Comparison of performance on video data set, measured as the average precision. The performance when classifying individual key frames is compared with using our method of combining results across shots and scenes.

usage in classifying material concepts. Thus, a pixel-wise segmentation would not be necessary, but the power of annotating such concepts can be employed. Concepts applicable to indoor scenes can also be incorporated and the separate indoor/outdoor classifier will no longer be necessary.

Although a large database of videos has been used in the analysis of this system, examples are still lacking for some classes, such as desert, mountain, and snow. A more varied set of videos with a larger variety of terrain would be useful. Exploring the performance difference when training the material or scene classifiers on video frames instead of photographs would help to further understand the differences in processing video versus photographs.

Acknowledgments

Thanks to Matt Berry and Liang Zhao for helpful discussions, Aleksander Ivanovic and Arindam Mitra for providing feedback on drafts of this work, and Lindsay Richardson for the time spent labeling scenes in video.

References

[1] A. Bosch, X. Muñoz, and J. Freixenet. Segmentation and description of natural outdoor scenes. *Image and Vision Computing*, 25(5):727–740, May 2007. 2

[2] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification via plsa. In *Proc. ECCV*, pages 517–530. 2006. 2

[3] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Trans. PAMI*, 30(4):712–727, 2008. 2, 6

[4] M. Boutell, J. Luo, and R. T. Gray. Sunset scene classification using simulated image recomposition. In *Proc. International Conference on Multimedia and Expo*, 2003. 2

[5] M. R. Boutell, J. Luo, and C. M. Brown. Factor graphs for region-based whole-scene classification. In *CVPR Workshop on Semantic Learning*, 2006. 2

[6] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, May 2004. 2

[7] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001. 3

[8] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Proc. ICCV*, 2007. 2

[9] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001. 3

[10] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *The Journal of Machine Learning Research*, 5:913–939, 2004. 2

[11] J. Corso, A. Yuille, and Z. Tu. Graph-shifts: Natural image labeling by dynamic hierarchical computing. In *Proc. CVPR*, 2008. 2

[12] H. Dunlop. Automatic rock detection and classification in natural scenes. Master's thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, August 2006. CMU-RI-TR-06-40. 3

[13] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR*, volume 2, pages 524–531, 2005. 2

[14] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004. 3

[15] G. Heidemann. Unsupervised image categorization. *Image and Vision Computing*, 23(10):861–876, 2005. 2

[16] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *Proc. ICCV*, 2005. 3

[17] M. Israel, E. L. van den Broek, P. van der Putten, and M. D. Uyl. Automating the construction of scene classifiers for content-based video retrieval. In *Proc. of the Fifth International Workshop on Multimedia Data Mining*, pages 38–47, Seattle, WA, 2004. 2

[18] J. Kaufhold, R. Collins, A. Hoogs, and P. Rondot. Recognition and segmentation of scene content using region-based classification. In *Proc. ICPR*, 2006. 2

[19] J. J. Kivinen, E. B. Sudderth, and M. I. Jordan. Learning multiscale representations of natural scenes using dirichlet processes. In *Proc. ICCV*, 2007. 2

[20] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, 2006. 2, 4, 5, 6

[21] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. PAMI*, 25(9):1075–1088, 2003. 2

[22] J. Luo and A. Savakis. Indoor vs outdoor classification of consumer photographs using low-level and semantic features. In *Proc. ICCV*, 2001. 2

[23] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: combining segmentation and recognition. In *Proc. CVPR*, volume 2, pages 326–333, 2004. 3

Figure 3. Sample results on video. For each class, four sample frames from a scene are shown. Each was classified as indoor, outdoor, or undetermined. In the case of an outdoor result, a semantic segmentation is shown using the legend at the bottom. Through combining the results from these frames and others in the scene, a correct classification was achieved.

[24] P. Quelhas and J.-M. Odobez. Natural scene image modeling using color and texture visterms. In *Proc. International Conference on Image and Video Retrieval*, pages 411–421, 2006. 2

[25] Z. Rasheed and M. Shah. Scene detection in hollywood movies and tv shows. In *Proc. CVPR*, 2003. 4

[26] N. Rasiwasia and N. Vasconcelos. Scene classification with low-dimensional semantic spaces and weak supervision. In *Proc. CVPR*, 2008. 2

[27] B. Russell, A. Torralba, K. Murphy, and W. Freeman. LabelMe: A database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, May 2008. 5

[28] N. Serrano, A. Savakis, and A. Luo. A computationally efficient approach to indoor/outdoor scene classification. In *Proc. ICPR*, volume 4, pages 146–149, 2002. 2

[29] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Proc. CVPR*, 2008. 2

[30] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81(1):2–23, January 2009. 2

[31] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412, 2003. 2

[32] A. J. Vailaya and A. H. J. Zhang. On image classification: city vs. landscape. *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 3–8, 1998. 2

[33] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *IJCV*, 62(1-2):61–81, 2005. 3

[34] J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *Proc. CVPR*, 2007. 2

[35] J. Vogel and B. Schiele. Semantic scene modeling and retrieval for content-based image retrieval. *IJCV*, 72(2):133–157, April 2007. 2

[36] L. Yang, P. Meer, and D. Foran. Multiple class segmentation using a unified framework over mean-shift patches. In *Proc. CVPR*, 2007. 2