Human Vision and Electronic Imaging XI (B. E. Rogowitz, T. N. Pappas, and S. J. Daly, eds.), Proc. SPIE Vol. 6057, (San Jose, CA), Jan. 2006.

Perceptually Based Techniques for Semantic Image Classification and Retrieval

Dejan Depalov,[a] Thrasyvoulos Pappas,[a] Dongge Li,[b] and Bhavan Gandhi[b]

[a]Electrical and Computer Engineering, Northwestern University
2145 Sheridan Road, Evanston,IL, 60208
[b]Motorola Labs, Motorola, 1303 E. Algonquin Road, Schaumburg, IL, 60196

ABSTRACT

The accumulation of large collections of digital images has created the need for efficient and intelligent schemes for content-based image retrieval. Our goal is to organize the contents semantically, according to meaningful categories. We present a new approach for semantic classification that utilizes a recently proposed color-texture segmentation algorithm (by Chen et al.), which combines knowledge of human perception and signal characteristics to segment natural scenes into perceptually uniform regions. The color and texture features of these regions are used as medium level descriptors, based on which we extract semantic labels, first at the segment and then at the scene level. The segment features consist of spatial texture orientation information and color composition in terms of a limited number of locally adapted dominant colors. The focus of this paper is on region classification. We use a hierarchical vocabulary of segment labels that is consistent with those used in the NIST TRECVID 2003 development set. We test the approach on a database of 9000 segments obtained from 2500 photographs of natural scenes. For training and classification we use the Linear Discriminant Analysis (LDA) technique. We examine the performance of the algorithm (precision and recall rates) when different sets of features (e.g., one or two most dominant colors versus four quantized dominant colors) are used. Our results indicate that the proposed approach offers significant performance improvements over existing approaches.

Keywords: Image analysis, retrieval, segmentation, semantic classification, perceptual models, adaptive clustering, steerable filters

1. INTRODUCTION

The accumulation of large collections of digital images has created the need for efficient and intelligent schemes for content-based image retrieval. Since humans are the ultimate users of most image retrieval systems, it is important to organize the contents semantically, according to meaningful categories. This requires an understanding of the important semantic categories that humans use for image classification, and the extraction of meaningful image features that can discriminate between these categories. Recent research efforts have addressed the first problem.[1–3] The goal of this paper is to address the second, namely, the extraction of perceptually-based image features that can be correlated with high level semantics and used to capture the semantic meaning of an image.

Current algorithms for low-level feature extraction, such as color, texture, and shape, are quite sophisticated and have met with e considerable success.[4,5] However, the extraction of low-level image features that can be correlated with high-level image semantics remains a challenging task. Several approaches have been proposed in the recent literature that attempt to bridge this semantic gap between low-level features and high-level semantics. Most of these approaches use an image segmentation scheme as an intermediate step, and then rely on the content of the segmented regions as well as their context within an image to obtain semantic information. Zhu *et al.*[6] partition the image into equal size blocks and index the regions using a codebook whose entries are obtained from the block features. Wang *et al.*[7] also propose a codebook based approach, whereby the codebook is used to segment the image based on the statistics of the region color and texture features. Their approach also attempts to take into account properties of the neighboring regions. Pan *et al.*[8] use a simple segmentation technique to segment an image into regions and extract their features. Each region is given a label called a blob-token. The authors attempt to find the association among the blob-tokens and the associated captions to index the

Further author information: Send correspondence to D. Depalov at email: depalov@ece.northwestern.edu, or T. N. Pappas at email: pappas@ece.northwestern.edu

(a) (b) (c) (d)

Figure 1. Color-texture image features and segmentation. (a) Original color image. (b) Adaptive dominant colors. (c) Texture classes (smooth regions shown in black, horizontal in gray, and complex in white). (d) Final segmentation

image. Li and Wang[9] use a statistical modeling approach in which images of a given concept are regarded as the instances of a random process characterizing this concept. Their method utilizes 2D hidden Markov models to calculate a measure of association between the image and the textual description of a concept. Finally, Mojsilovic and Rogowitz[3] attempt to link low-level image features directly to image semantics.

Inspite of all this effort, the effectiveness of CBIR systems has not been satisfactory and they are still a long way from matching the performance of the human visual system (HVS). The basis of this paper is a new methodology for image segmentation, semantic classification, and retrieval, that combines an understanding of image characteristics with perceptual models and principles about the processing of texture and color information. Like the schemes described above, the proposed approach relies on image segmentation as an intermediate step for extracting high-level information. However, it is the segmentation and feature selection that sets apart the proposed approach. It utilizes a recently proposed perceptual segmentation technique that uses spatially adaptive color and spatial texture features to segment natural scenes into perceptually, semantically uniform regions.[10–12] It is aimed at segmentation of natural scenes, in which the color and texture of perceptually distinct regions do not typically exhibit uniform statistical characteristics. The focus of this paper is on assigning semantic labels to the resulting segments. This requires the derivation of region-wide color and texture features as medium level descriptors. These descriptors are the key to bridging the gap between low-level image primitives and high-level image semantics. However, as we pointed out above, it is the use of perceptually uniform segments and the selection of perceptually-motivated region-wide features that is critical to the success of the proposed approach. We present linear discriminant analysis techniques for assigning labels to image segments based on color composition and spatial texture descriptors. Further improvements can be achieved by incorporating the segment location, size, and boundary shape, as well as the properties of the neighboring segments. However, our focus is on what can be achieved using only color and texture.

We demonstrate the effectiveness of the proposed approach using a database of approximately 2500 images of natural scenes, which were segmented using the algorithm in Ref. 10. Our results indicate that the proposed approach offers significant performance improvements over existing approaches. While the focus of this paper is on still images, the techniques we discuss can also form the basis for content-based analysis of video sequences.

2. COLOR-TEXTURE FEATURE SELECTION

The selection of appropriate color-texture features is critical for both image segmentation and segment classification. The segmentation approach proposed in Ref. 10 is based on two types of spatially adaptive features. The first provides a localized description of the color composition of the texture and the second models the spatial characteristics of its grayscale component. Both incorporate models of human perception and signal characteristics.

The color composition feature exploits the fact that the HVS cannot simultaneously perceive a large number of colors. In addition, it accounts for the spatially varying image characteristics and the adaptive nature of the HVS. It thus consists of a small number of spatially adaptive dominant colors and the corresponding percent occurrence of each color in the vicinity of a pixel:

$$f_c(x,y,N_{x,y}) = \{(c_i,p_i), i = 1,\ldots,M, p_i \in [0,1]\} \quad (1)$$

where c_i is a 3-D color vector and p_i is the corresponding percentage. $N_{x,y}$ denotes the neighborhood of the pixel at (x,y) and M is the number of dominant colors in $N_{x,y}$; a typical value is $M = 4$. The spatially adaptive dominant colors are obtained using the adaptive clustering algorithm (ACA).[13] An example is shown in Fig. 1(b). The perceptual similarity

(a) (b) (c) (d)

Figure 2. Segment-wide feature extraction. (a) Segmented image. (b) Selected segment. (c) Its color composition. (d) Its texture composition.

between two color composition feature vectors is based on the "Optimal Color Composition Distance (OCCD)," which finds the optimal mapping between the color composition features of two segments and computes the average distance between them in the *CIE L*a*b** color space.

The spatial texture feature extraction is based on a multiscale frequency decomposition with four orientation subbands (horizontal, vertical, +45°, -45°). Here, we use a one-level steerable filter decomposition with four orientation subbands. The local energy of the subband coefficients is used as a simple but effective characterization of spatial texture. At each pixel location, the maximum of the four subband coefficients determines the texture orientation. A median filtering operation boosts the response to texture within uniform regions and suppresses the response resulting from to transitions between regions. Pixels are then classified into smooth and non-smooth classes, and non-smooth pixels are further classified on the basis of dominant orientation, as horizontal, vertical, +45°, -45°, and complex (i.e., no dominant orientation). An example is shown in Fig. 1(c).

The segmentation algorithm combines the color composition and spatial texture features to obtain segments of uniform texture. It is a fairly elaborate algorithm that relies on spatial texture to determine the major structural composition of the image and combines it with color, first to obtain a crude estimate of the major segments, and then to refine it in order to obtain accurate and precise localization of the border between regions. Several critical parameters of the texture features and segmentation algorithm can be determined by subjective tests.[12] These include thresholds for the smooth/non-smooth classification, for determining the dominant orientation, and for the color-composition feature similarity. The goal of the tests is to relate human perception of isolated (context-free) texture patches to the statistics of natural textures.

3. SEGMENT WIDE FEATURE EXTRACTION

As we pointed out in the intorduction, the key to bridging the gap between low-level image primitives and high-level semantics is the extraction of medium-level segment descriptors. In this section, we discuss the development of segment-wide color composition and spatial texture features. These are derived from the features we described in the previous section, but there are important differences. Image segmentation required a careful combination of global and local information.[10,13] However, once the segments are obtained, region interpretation should rely on region-wide features. Thus, the features must be calculated on a segment by segment basis, using only information from within the segment. To accomplish this, the local averages and medians for both the color composition and spatial texture features must be computed across and strictly within each segment. An example is shown in Fig.2, where Fig.2(a) shows a segmented image, Fig.2(b) shows a selected segment, and Fig.2(c) shows the color composition of the segment (dominant colors and percentages). The texture features of the segment can be similarly described by the percentage of smooth, horizontal, vertical, +45°, -45°, and complex pixels as shown in Fig.2(d)

The spatial texture features can be represented as a six-dimensional vector that consists of the percentages for each texture category. The color composition features, on the other hand, require further simplification, because in their original form (trichromatic colors and associated percentages) they are not well suited for the statistical analysis techniques we consider in Section 5. We consider two possibilities.

The first[14] uses a perceptually quantized color space, whereby a relatively small number of perceptually distinct representative colors (color names) are selected (e.g., Boynton's "eleven colors which are almost never confused"[15]), and the

feature vector consists of the percentage of each color for the segment. Thus, the feature vector of each segment will have up to $M = 4$ nonzero components. The advantage of this representation is that, in addition to reducing the dimensionality of the color composition feature vector, it provides a symmetry between the two types of features, as noted in Ref. 14. Moreover, it is a natural extension our perceptual approach: Humans use a few color names to describe the color composition of a scene. For example, Boynton[15] found that when people are asked to categorize colors, the number of perceptually distinguishable color categories is small. The procedure we use to assign color names to the dominant colors of a region is described in Ref. 16. The color names (labels) are consistent with a National Bureau of Standards recommendation for color names. The syntax contains names for 267 regions in color space, and employs English terms to describe colors along the three dimensions of the color space: hue, lightness and saturation. There are seven discrete values for lightness, five discrete values for saturation, and a basic set of eleven prototypical hues, as shown in Table 1. Thus, if we assign labels based on the primary hues and achromatic colors, we end up with 14 labels, for a 14-dimensional texture vector.

Hue-primary	Hue-secondary	Saturation	Lightness	Achromatic
Red	Reddish	Grayish	Blackish	Black
Orange	Brownish	Moderate	Very-dark	Gray
Brown	Yellowish	Medium	Dark	White
Yellow	Greenish	Strong	Medium	
Green	Bluish	Vivid	Light	
Blue	Purplish		Very-light	
Purple	Pinkish		Whitish	
Pink				
Beige				
Magenta				
Olive				

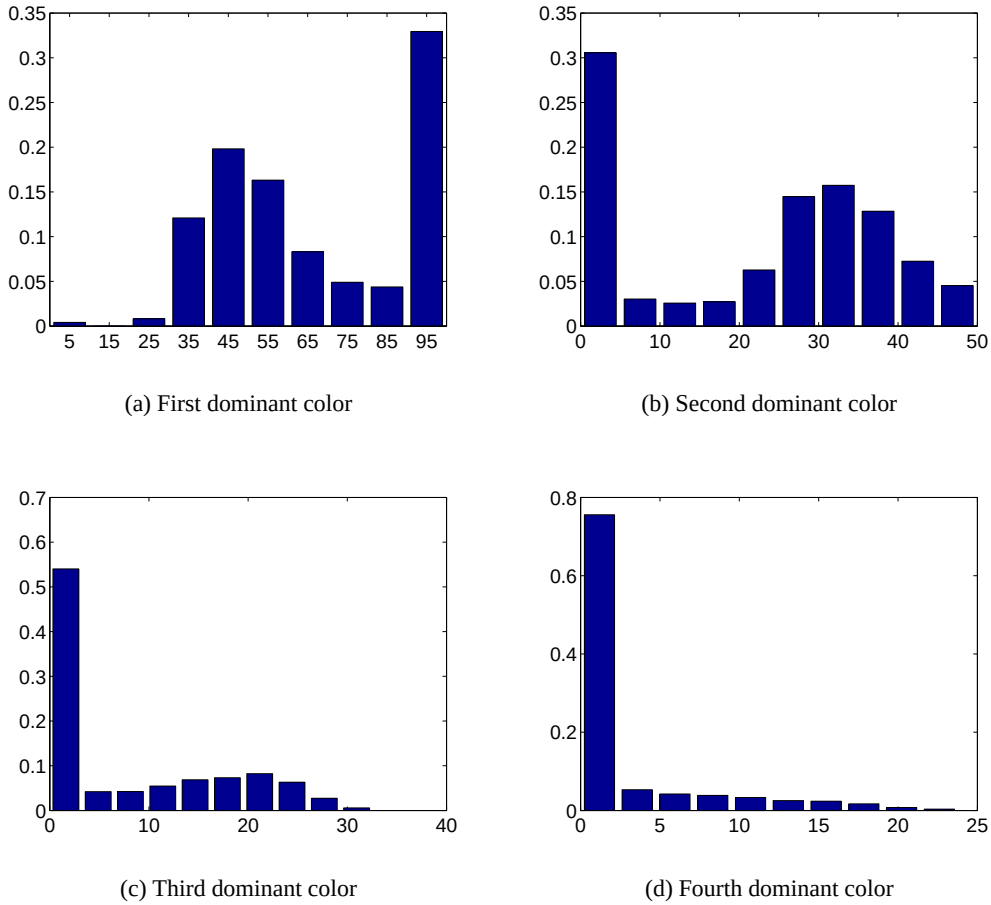
Table 1. Color Naming Syntax

The second approach is to use only one or two dominant colors, the ones with highest percentages, and to represent them by their (unquantized) $L^*a^*b^*$ coordinates. In the remainder of the paper, we will refer to these as the first and second dominant colors. In our implementation, we used the first dominant color and the difference between the first and second dominant color. As we will see in Section 6, this representation results in better classification performance than the color naming approach. In order to explain this somewhat surprising result, we did a statistical analysis of the dominant colors associated with segments in our database. The relevant statistics are shown in Fig. 3, which shows histograms for the first, second, third, and fourth dominant color across all segments in the database. The horizontal axis represents the percentage of the area that the corresponding dominant color occupies in a segment, while vertical axis represents the probability of occurence for each bin. The analysis reveals that the great majority of segments could be described by first two dominant colors with very little loss of information. Further analysis of the $L^*a^*b^*$ distance among the dominant colors indicates that, in the majority of cases, the second dominant color is less than twenty units away from the first. This means that for a great majority of segments the second dominant color is similar to the first. The histogram of the distance between the first and second dominant colors is shown in Fig. 4(a), and the distance between the first and third is shown in Fig. 4(b).

4. SEMANTIC LABELING

In Refs. 1–3, subjective experiments were conducted in order to identify important semantic categories that humans use for image organization and retrieval. For example, they have discovered two important dimensions in human similarity perception: "natural" vs. "man-made," and "human" vs. "non-human." In addition, they found that certain cues, such as "sky," "water," "mountains," etc., play an important role in semantic categorization by humans.[17] Thus, rather than trying to obtain a complete and detailed description of every object in the scene, this suggests that it may be sufficient to isolate segments of such perceptual significance, and use them to correctly classify an image into a given category. Based on this observation, the first step towards semantic classification is assigning semantically important labels to image segments.

In order to assign labels to segments, we need to relate the segment features to semantic labels, but first, we must decide what the labels will be. To this end, we have assembled a vocabulary of labels consistent with the above findings, as well



(a) First dominant color

(b) Second dominant color

(c) Third dominant color

(d) Fourth dominant color

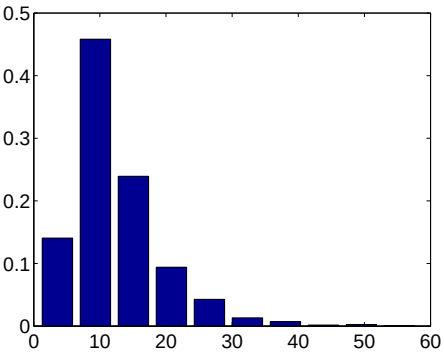
Figure 3. Statistics of dominant colors. The horizontal axis represents the percentage of the area that the dominant color occupies in a segment and the vertical axis represents the probability of occurence for each bin.

as those used in annotation of the NIST TRECVID 2003 development set.[18] The set of labels we selected is a subset of NIST lexicon. To describe the content of an image we use two types of labels, segment and scene labels. The segment labels describe the semantics of a particular segment (e.g., building, sky), while the scene labels describe the (higher-level) semantic content of the image (e.g., beach scene). The segment labels we have chosen are shown in Table 2. They are arranged in a hierarchical manner, with natural, man-made, human, and animal at the top. Note that in our classification algorithms, we use only leaf nodes. The higher order categories are then derived from the leaves.

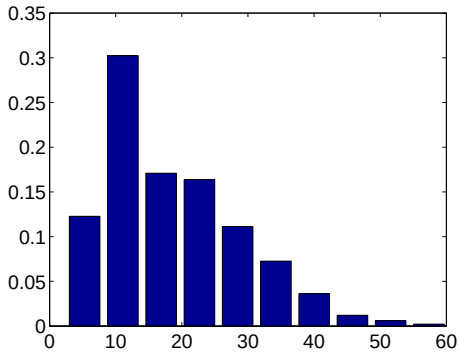
The initial focus of our experiments was on natural vs. man-made classification. Thus, we did not include any scenes with humans or animals. The human detection problem, and especially face detection, is well-studied in the literature,[19] and the existing techniques can easily be combined with the proposed approach. The problem of animal detection is more complicated because of natural camouflage. However, we have found that the proposed approach is capable of segmenting and detecting animals.

5. LEARNING AND CLASSIFICATION

We performed several sets of experiments using approximately 2500 photographs. The majority of the images were obtained from the Corel Stock Photo Library. Additional images were obtained from a Key Photos Library and the investigators' personal repository. The images in the database cover a variety of outdoor scenes, with a wide range of themes. The



(a) Histogram of distances between first and second dominant color



(b) Histogram of distances between first and third dominant color

Figure 4. Distances between dominant colors in *L*a*b** color space

Natural				Man-made	Human	Animal
Vegetation	Sky	Landform	Water			
Grass	Day-sky	Snow		Building/House	Face	
Trees/bushes	Night-sky	Mountain/Hill		Bridge	Person	
Forest	Sun	Ground		Car	People	
Flowers	Clouds	Pavement/Road*		Boat		
	Sunrise/Sunset			Airplane		
				Other		

Table 2. Segment Labels

images were segmented using the algorithm in Ref. 10 described above. The resulting segments were manually labeled to be used as the ground truth in supervised learning. Each segment was assigned exactly one label. Segments whose area was less than three percent of total image area were not considered. This resulted in approximatelly 9000 labeled segments, 80% of which were used for training and the rest for testing.

For the training and classification we used the Linear Discriminant Analysis (LDA) method.[20] LDA belongs to the class linear classifiers, which try to find a subspace projection such that samples from the different classes are well separated, i.e., to find directions in the data space that facilitate data classification. This is done by finding a direction that maximizes the variance between the class means and at the same time minimizes the variance within each class.

The measure of separation between the class means is defined as:

$$S_B = \sum_c N_c(\mu_c - \bar{\mathbf{x}})(\mu_c - \bar{\mathbf{x}})^T \tag{2}$$

where,

$$\mu_c = \frac{1}{N_c}\sum_{i\in c} x_i \tag{3}$$

and

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_c N_c\mu_c \tag{4}$$

*While "pavement/road" is man-made, its features are almost identical to those for "ground."

The measure of within class variance is defined as:

$$S_W = \sum_c \sum_{i \in c} (\mathbf{x}_i - \mu_c)(\mathbf{x}_i - \mu_c)^T \tag{5}$$

S_B is usually referred to as the "between the classes scatter matrix," while S_W is known as the "within the classes scatter matrix."

Using these measures, the objective function can be represented as:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \tag{6}$$

where $\mathbf{w}$ is a transformation matrix. The objective function $J(\mathbf{w})$ is maximized by solving the generalized eigenvalue problem:

$$S_B \mathbf{w} = \lambda S_W \mathbf{w} \tag{7}$$

so that columns of an optimal $\mathbf{w}$ are the eigenvectors asoociated with largest eigenvalues. One of the underlying assumptions that are necessary for LDA to work is that the data for each class form a single cluster.

6. RESULTS

We evaluated the performance of the proposed techniques using the standard measures that are used for evaluating search strategies in the literature. The **recall** is the ratio of the correctly labeled segments to the total number of relevant segments in the database (i.e., those with the particular label). The **precision** is the ratio of the correctly labeled segments to the total number of segments that the algorithm assigned to the particular label (both correct and incorrect). Both performance measures are expressed as percentages. Overall performance can be expressed as the accuracy over the whole database.

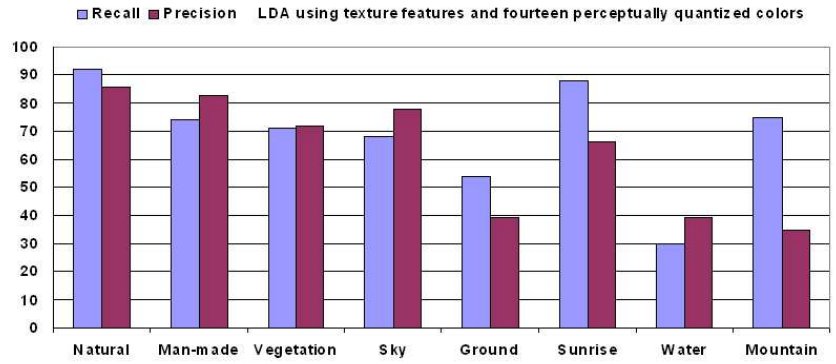
We compared the classification performance of the proposed approach using the spatial texture and color composition feature representations we described in Section 3, and applying LDA to the labeled set of segmented images. In all experiments we used the same set of texture features, but different sets of color composition features. As we saw in Section 5, 80% of the labeled segments were used for training and 20% for testing. The results are shown in Fig. 5 for the perceptually quantized colors and Fig. 6 for the most dominant colors.

In the first experiment, we used the perceptually quantized color features with 15 colors. Thus, each segment was represented by a 21 dimensional feature vector (six spatial textures and 15 colors). The results are shown in Fig. 5(a), which shows the recall and precision rates for the most important semantic categories.

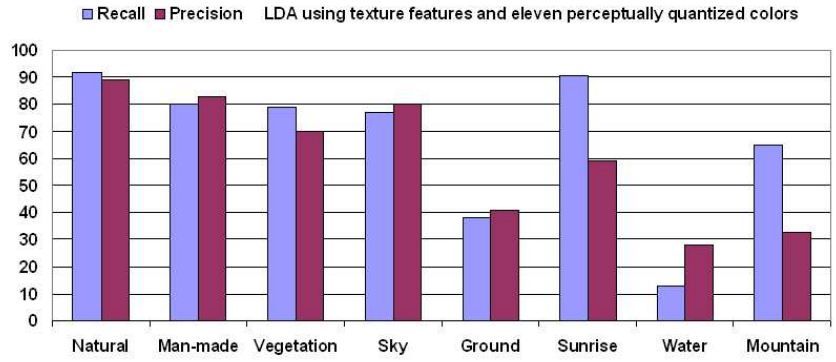
We then evaluated the classification performance using the spatial texture features and the most dominant color expressed in CIE *L*a*b** coordinates. In this case, the dimension of the feature vector is nine (three dimensions for color compontent and six for texture). Figure 6(a) presents the results of this classification. Comparing the recall and precision rates of the two expreiments it is clear that using only the most dominant color expressed as an exact coordinate in the *L*a*b** color space outperforms the perceptual quantization approach. At first, this appears to be surprizing, which is what led us to the statistical analysis presented in Section 3. Recall that the key conclusion of our analysis was that a great majority of segments either have only one dominant color or have the second dominant color similar to the first. To make sure that the result is not an artifact of the number of quantization levels, we experimented with different numbers of colors. Figure 5(b) shows the classification results when the number of quantization levels is reduced to eleven. Observe that there is no significant difference between the two quantization schemes. Note also that increasing the number of color quantization levels will not increase the classification performance. This is because increasing the number of quantization levels increases the feature dimensionality in such a way that it creates an artificial distance between perceptually similar colors. This statement has been confirmed by our experiments.

We also tried including a second dominant color (in addition to first and texture). This resulted in a modest gain in classification performance, as is evident from Fig. 6(b), which shows the results of using spatial texture features and two most dominant colors expressed in CIE *L*a*b** coordinates. Finally, we should also note that additional experiments verified that using a third dominant color does not increase performance, and that using a fourth dominant color actually reduces the classification ability of LDA.

Overall, our segment classification results compare favorably to the methods described in the literature (e.g., Refs. 21–26).



(a) LDA using texture and 14 perceptually quantized colors



(b) LDA using texture and 11 perceptually quantized colors

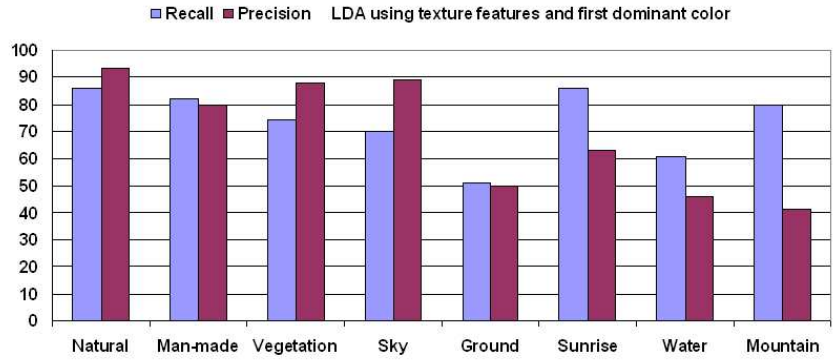
Figure 5. Classification results with perceptually quantized colors

7. CONCLUSIONS

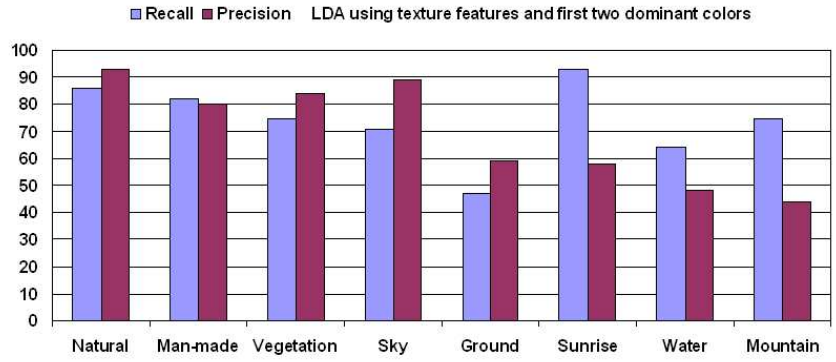
We presented a new approach for semantic classification that utilizes perceptual models for image segmentation and classification. The main innovations of the proposed apprach are the use of an algorithm that produces perceptually uniform segments and the selection of perceptually-motivated region-wide color and texture features. The features of these regions are used as medium level descriptors and are the key to bridging the gap between low-level image primitives and high-level image semantics. Our results indicate that the proposed approach offers significant performance improvements over the existing literature.

8. ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation (NSF) under Grant No.CCR-0209006. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF. This work was also supported by the Motorola Center for Telecommunications at Northwestern University.



(a) LDA using texture and first dominant color



(b) LDA using texture and two dominant colors

Figure 6. Classification results with most dominant colors

REFERENCES

1. B. E. Rogowitz, T. Frese, J. R. Smith, C. A. Bouman, and E. Kalin, "Perceptual image similarity experiments," in *Human Vision and Electronic Imaging III*, B. E. Rogowitz and T. N. Pappas, eds., **Proc. SPIE, Vol. 3299**, pp. 576–590, (San Jose, CA), Jan. 1998.
2. A. Mojsilovic and B. Rogowitz, "A psychophysical approach to modeling image semantics," in *Human Vision and Electronic Imaging VI*, B. E. Rogowitz and T. N. Pappas, eds., **Proc. SPIE Vol. 4299**, pp. 470–477, (San Jose, CA), Jan. 2001.
3. A. Mojsilović and B. E. Rogowitz, "Semantic metric for image library exploration," *IEEE Trans. Multimedia* **6**, pp. 828–838, Dec. 2004.
4. Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions and open issues," *J. Visual Communication and Image Representation* **10**, pp. 39–62, Mar. 1999.
5. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Machine Intell.* **22**, pp. 1349–1379, Dec. 2000.
6. L. Zhu, A. Zhang, A. Rao, and R. Srihari, "Keyblock: An approach for content-based image retrieval," in *ACM Multimedia 2000*, pp. 157–166, (Los Angeles, CA), Oct. 2000.
7. W. Wang, Y. Song, and A. Zhang, "Semantics retrieval by content and context of image regions," in *Proc. of the 15th International Conference on Vision Interface*, pp. 17–24, (Calgary, Canada), May 2002.

8. J. Y. Pan, H. J. Yang, P. Duygulu, and C. Faloutsos, "Automatic image captioning," in *ICME*, 2004.
9. J. Li and J. Wang, "Automatic linguistic indexing of pictures by a atatistical modeling approach," *IEEE Trans. Pattern Anal. Machine Intell.* **25**, Sept. 2003.
10. J. Chen, T. N. Pappas, A. Mojsilovic, and B. E. Rogowitz, "Adaptive perceptual color-texture image segmentation," *IEEE Trans. Image Processing* **14**, pp. 1524–1536, Oct. 2005.
11. J. Chen, T. N. Pappas, A. Mojsilovic, and B. E. Rogowitz, "Perceptually-tuned multiscale color-texture segmentation," in *Proc. Int. Conf. Image Processing (ICIP-04)*, pp. 921–924, (Singapore), Oct. 2004.
12. J. Chen and T. N. Pappas, "Experimental determination of visual color and texture statistics for image segmentation," in *Human Vision and Electronic Imaging X*, B. E. Rogowitz, T. N. Pappas, and S. J. Daly, eds., **Proc. SPIE Vol. 5666**, pp. 227–236, (San Jose, CA), Jan. 2005.
13. T. N. Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Trans. Signal Processing* **SP-40**, pp. 901–914, Apr. 1992.
14. D. Depalov, T. N. Pappas, D. Li, and B. Gandhi, "A perceptual approach for semantic image retrieval," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-06)*, (Toulouse, France), May 2006. To appear.
15. R. M. Boynton, "Eleven colors that are almost never confused," in *Human Vision, Visual Proc., and Digital Display*, B. E. Rogowitz, ed., **Proc. SPIE Vol. 1077**, pp. 322–332, (Los Angeles, CA), Jan. 18–20 1989.
16. A. Mojsilović, "A computational model for color naming and describing color composition of images," *IEEE Trans. Image Processing* **14**, pp. 690–699, May 2005.
17. A. Mojsilovic and B. Rogowitz, "Capturing image semantics with low-level descriptors," in *Proc. Int. Conf. Image Processing (ICIP-01)*, pp. 18–21, (Thessaloniki, Greece), Oct. 2001.
18. C. Y. Lin, B. L. Tseng, and J. R. Smith, "Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets." NIST TRECVID, 2003.
19. P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision* **57**(2), pp. 137–154, 2004.
20. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2nd ed., Oct. 2000.
21. J. Pan, H. J. Yang, P. Duygulu, and C. Faloutsos, "Automatic image captioning," in *ICME*, June 2004.
22. K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan, "Matching words and pictures," *Journal of Machine Learning Research* **3**, pp. 1107–1135, 2003.
23. J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. Pattern Anal. Machine Intell.* **25**(9), pp. 1075–1088, 2003.
24. J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *SIGIR*, pp. 119–126, ACM, 2003.
25. J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries," *IEEE Trans. Pattern Anal. Machine Intell.* **23**, pp. 947–963, Sept. 2001.
26. C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, "Blobworld: A system for region-based image indexing and retrieval," in *Proc. Third International Conference on Visual Information Systems*, pp. 509–516, Springer, June 1999.