October 2005

oe magazine

The SPIE Magazine of Photonics Technologies and Applications

Computers Play 'Tag' with Image Archives

>> ASICs kiss red eye goodbye

>> Building a better SAN

>> IR optics leap forward

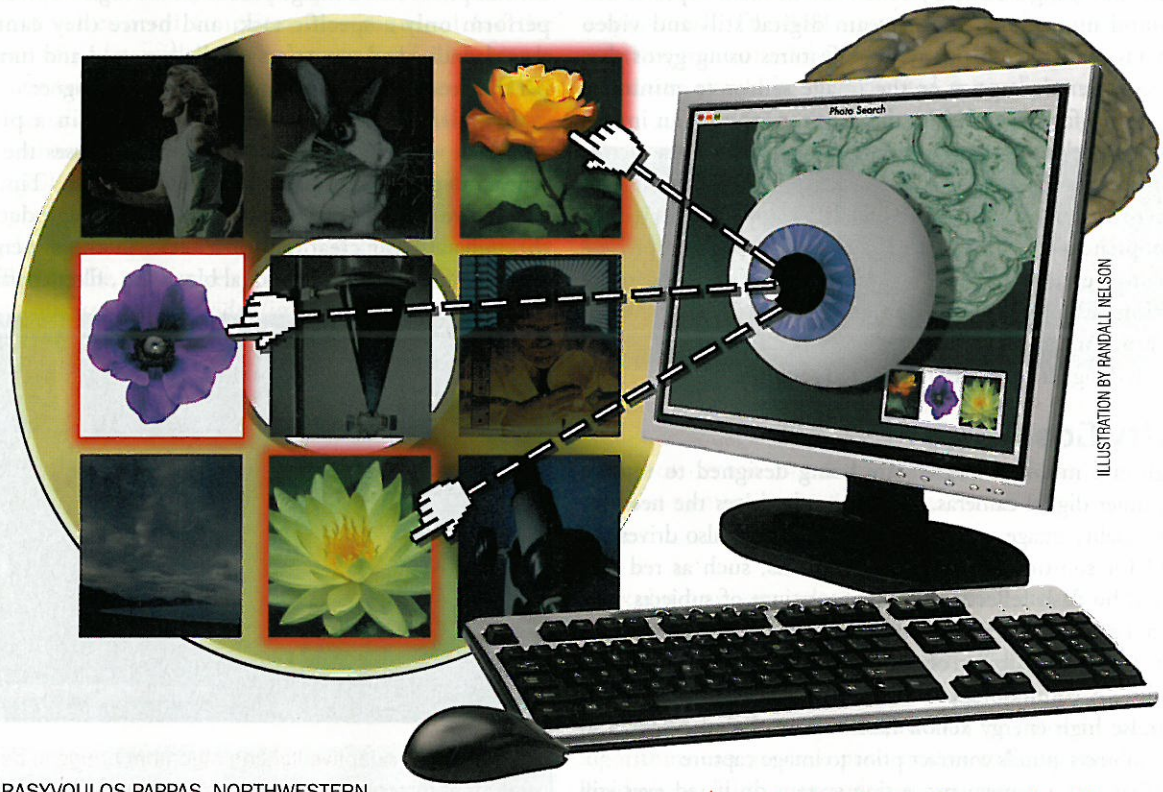

ILLUSTRATION BY RANDALL NELSON

BY THRASYVOULOS PAPPAS, NORTHWESTERN UNIVERSITY; JUNQING CHEN, UNILEVER RESEARCH; AND DEJAN DEPALOV, NORTHWESTERN UNIVERSITY

Image segmentation, classification, and retrieval algorithms incorporate models of human vision and signal attributes.

Learning Perception

Signal processing systems are still a long way from matching the performance of the human visual system (HVS). This could be intimidating or discouraging, but human performance also can provide inspiration and a performance goal for scientists and engineers working in the field. Many image- and video-processing algorithms rely, at least implicitly, on the properties of human vision. Image halftoning (displaying continuous-tone images using a limited number of colors), for example, would not be possible without the ability of the eye to act as a spatial lowpass filter. Movies would be perceived as a sequence of still-frame images were it not for the ability of the eye to act as a temporal lowpass filter.

There has been a lot of progress recently in incorporating explicit HVS models in image processing algorithms to maximize system performance. We can achieve "perceptually lossless" compression of images and video, for example, using perceptual models that determine the amount of distortion that can be introduced in the signal without being noticed by the human observer. Similar ideas are used in audio compression algorithms for MP3 players and in multimedia watermarking.

An exciting topic of current research is the development of perceptual models for image analysis and understanding; for example, one can apply perceptual principles to the segmentation of complex natural scenes and extract semantic information that can be used for intelligent and efficient image organization and retrieval. Our group has developed an adaptive perceptual color-texture segmentation algorithm that combines knowledge of human perception with an understanding of signal characteristics to segment natural scenes into perceptually/semantically uniform regions.[1] The method can be used for image labeling and classification.[2] Still images are the focus of this article, but the techniques

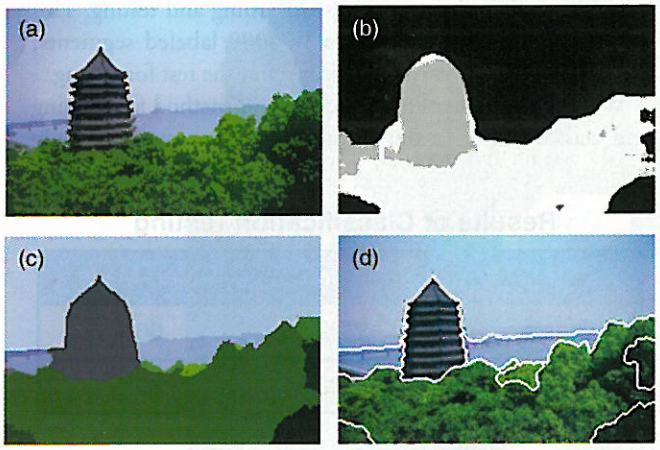

Figure 1 To produce a segmented image, the algorithm divides the image into adaptive dominant colors (a) and texture classes (b; smooth regions shown in black, horizontal in gray, and complex in white). It then performs crude segmentation (c). A border-refinement step produces the final segmentation (d), shown here overlaying the original image.

we discuss also form the basis for content-based analysis of video sequences.

Feature Selection

Developing models for the segmentation of images of natural scenes is difficult because unlike the detection of faces or specific objects, natural textures do not have a specific structure. In addition, in natural scenes, the texture characteristics of perceptually distinct regions are not statistically uniform due to effects of lighting, perspective, scale changes, and so on. In spite of such difficulties, the HVS can effortlessly segment natural scenes into perceptually/semantically uniform regions. A successful segmentation algorithm needs to incorporate both signal characteristics and models of the HVS.

In contrast to texture analysis/synthesis that requires a large number of parameters,[3] the spatially varying characteristics of natural texture dictate simple models, the parameters of which can be robustly estimated from relatively few sample points. The proposed approach 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 gray-scale component.

The color composition feature exploits the fact that the HVS cannot perceive a large number of colors simultaneously. 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 $M = 4$. We obtain the spatially adaptive dominant colors using the adaptive clustering algorithm (see figure 1).[4] Finally, we use a perceptually based similarity metric to compare color-composition feature vectors.

The spatial-texture feature extraction is based on a multiscale frequency decomposition with four orientation subbands (horizontal, vertical, +45°, −45°). Such decompositions offer an efficient and flexible approximation of early processing in the HVS. We use the local energy of the subband coefficients 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 due to transitions between regions. Pixels are then classified into smooth and non-smooth classes. Non-smooth pixels are further classified on the basis of dominant orientation as horizontal, vertical, +45°, −45°, and complex (i.e., no dominant orientation).

Adaptation and Perceptual Tuning

The segmentation algorithm combines 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 estimate the major segments, and then to obtain accurate and precise localization of the border between regions.

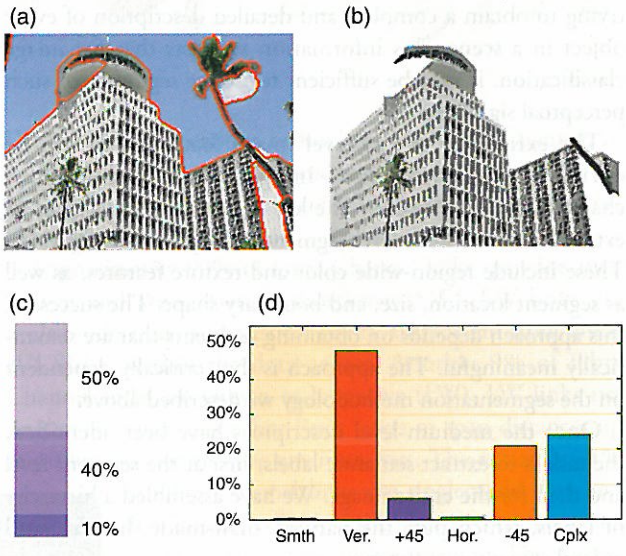

Figure 2 Once we segment an image (a), we can analyze segments (b) to identify medium-level descriptors such as color composition (c) and texture composition (d). To further reduce the dimensionality of the feature vectors, we map the dominant colors onto a small set of prototypical colors (i.e., black, white, gray, red, green, etc.).

The border-refinement approach illustrates the adaptive nature of the algorithm. We estimate the texture characteristics (color-composition feature vector) of each pixel using a small window and compare them to localized estimates of the texture characteristics of each of the adjacent regions using a larger window. We then use the similarity metric to classify the pixel as part of one of the regions. A spatial constraint is added to ensure region smoothness. We repeat this procedure for all pixels on the borders of non-smooth regions. A few iterations are necessary for convergence. The key to adapting to the spatial variations of texture characteristics is that the window sizes progressively decrease as the algorithm converges.

Several critical parameters of texture features and the segmentation algorithm can be determined by subjective tests.[5,6] These include thresholds for classifying smooth and non-smooth pixels, determining the dominant orientation, and identifying 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. Experimental results demonstrate that this perceptual tuning leads to significant improvements in segmentation performance.

Bridging the Semantic Gap

Recent subjective experiments have identified important semantic categories that people use for image organization and retrieval.[7] Two important dimensions in human similarity perception are "natural" versus "man-made," and "human" versus "non-human." It was also found that certain cues, such as "sky," "water," "mountains," etc., have an important influence in human image perception. Rather than trying to obtain a complete and detailed description of every object in a scene, this information suggests that for image classification, it may be sufficient to isolate segments of such perceptual significance.

The extraction of low-level image features that can be correlated with high-level image semantics remains a challenging task, however. The key to bridging this gap is the extraction of medium-level segment descriptors (see figure 2). These include region-wide color and texture features, as well as segment location, size, and boundary shape. The success of this approach depends on obtaining segments that are semantically meaningful. The approach is thus critically dependent on the segmentation methodology we described above.

Once the medium-level descriptors have been identified, the task is to extract semantic labels, first at the segment level and then for the entire image. We have assembled a hierarchy of labels, which puts the natural, man-made, human, and animal categories at the top.

To demonstrate the effectiveness of the proposed approach, we conducted a set of simple experiments with a database of approximately 1600 photographs, focusing initially on the natural versus man-made dimension. The images were segmented using the segmentation algorithm described above, and the resulting segments were labeled manually to be used as the ground truth for supervised learning and testing. This activity resulted in approximately 4000 labeled segments, 80% of which were used for training and the rest for testing.

We used the Fisher linear discriminant method for training and classification. In our initial experiment, we used only spatial-texture and color-composition features. Our results showed a recall rate (the ratio of correctly labeled segments to the total number of relevant segments in the database) and precision (the ratio of correctly labeled segments to the total number of segments that the algorithm assigned to the label) that compare favorably to the methods in the literature (see table).

Results of Classification Testing

	Natural	Man-Made
Precision	88%	66%
Recall	87%	69%

We are currently in the process of identifying the correct descriptors for discriminating between categories farther down in the hierarchy (i.e., sky, water, mountains, buildings, etc.). Overall scene interpretation will be based on probabilistic layout models. **oe**

Thrasyvoulos Pappas is associate professor and Dejan Depalov is a PhD candidate in the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL; and Junqing Chen is imaging scientist at Unilever Research, Trumbull, CT. For questions, contact Pappas at 847-467-1243; 847-491-4455 (fax); or pappas@ece.northwestern.edu.

References

1. J. Chen et al., IEEE Trans. Image Processing 14[10], p. 1524 (2005).

2. T. Pappas, J. Chen, and D. Depalov, "Perceptually based techniques for image segmentation, semantic classification, and retrieval," IEEE Signal Processing Mag., to appear.

3. J. Portilla and E. Simoncelli, Int. J. Computer Vision 40[10], p. 49 (2000).

4. T. Pappas, IEEE Trans. Signal Processing SP-40[4], p. 901 (1992).

5. J. Chen and T. Pappas, Proc. SPIE 5666, p. 227 (2005).

6. www.ece.northwestern.edu/~pappas/research/texture_perception_test.html

7. A. Mojsilović and B. Rogowitz, IEEE Trans. Multimedia 6, p. 828 (2004).

Acknowledgments

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