Figure-ground classification based on spectral properties of boundary image patches

Sudarshan Ramenahalli
Zanvyl Krieger Mind/Brain Institute
and Dept. of Elec and Computer Eng
Johns Hopkins University
Baltimore, MD 21218
email: sramena1@jhu.edu

Stefan Mihalas
Zanvyl Krieger Mind/Brain Institute
Johns Hopkins University
Baltimore, MD 21218
Current address:
Allen Institute for Brain Science
Seattle WA 98103
email:mihalas@jhu.edu

Ernst Niebur
Zanvyl Krieger Mind/Brain Institute
and Department of Neuroscience
Johns Hopkins University
Baltimore, MD 21218
email:niebur@jhu.edu

***Abstract*—Image understanding requires segregation of the visual scene into perceptual objects. Due to the projection of the three-dimensional world on two-dimensional sensor surfaces, objects closer to the observer occlude those which are more distant. At any given occlusion border, it is important to decide which side is the foreground (figure) and which is the background, a decision which is influenced both by global and local image contents. In this report, we focus on local cues. We randomly select small image patches located on figure-ground borders in complex natural scenes. Spectral anisotropy features are extracted from the patches and used to train a non-linear Support Vector Machine. Using data from two large image databases (LabelMe and BSDS300), the classifier achieves an accuracy near 70% per local patch on the task of deciding which side of an occlusion is the foreground. Although in many cases global influences are important for figure-ground segregation, we suggest that the low computational cost of local computation can make it a useful strategy for figure-ground segregation.**

I. Introduction

In a complex three-dimensional world, objects physically closer to the observer occlude those that are further away when they are projected two-dimensional photoreceptor array, the retina in the case of primates. An important step in visual perception is thus to decide which side of an occlusion boundary corresponds to the object closer to the observer (the figure), and which side is the background. This process is figure-ground assignment, also referred to as figure-ground organization [1], [2]. In some cases, the distance between a given object and the observer can be determined by stereoscopic information (disparity) but this is not always the case and other cues are typically available. They can be broadly classified as having local and global character, based on the size of the observation window relative to the figure. Some global cues are symmetry [3] and surroundedness [4]; examples of local cues are convexity [5], [6] and the presence of T-junctions [7] or extremal edges [8], [9]. Use of local cues typically requires less computational resources than global cues since only small image patches need to be processed for the former. In this report, we introduce a novel type of purely local image processing that, as we show, yields highly informative results about figure-ground organization.

Our method uses a feature vector derived from spectral properties of small image patches that straddle object boundaries. For the figure and ground parts of each patch, the feature vector encodes separately the distributions of spectral power orthogonal and parallel to the object boundary. In the next section, we will define the spectral anisotropies that we use. We use the feature vectors to train a Support Vector Machine which is then used to classify new sample patches (outside the training set).

Throughout this report, we will use the terms "figure" and "foreground" interchangeably to refer to the region of the local patch that is closer to the observer. Likewise, "ground" and "background" are interchangeably used to refer to the more distant region.

II. Spectral Anisotropies Close to Object Boundaries

We select image patches $\mathbf{p}(x, y)$ of size $N \times N$ that straddle the figure-ground edge. Half of the patch is part of the figure and the other half is part of the background, with pixels on the edge being part of neither; see eq. 2 for definition. The oriented energy spectrum of a side (figure or ground), parallel to the figure-ground boundary is defined as,

$$E_s(u, y) = |P(u, y)|^2 \tag{1}$$

where the y coordinate varies orthogonally to the figure-ground boundary and $P(u, y)$ is the discrete one-dimensional Fourier transform with respect to x, parallel to the figure-ground boundary. The subscript s denotes the side of the patch containing figure (f) or ground (g) as in,

$$s := \begin{cases} f & \text{if} \quad \text{y} \in (0, \frac{N}{2} - 1) \\ \\ g & \text{if} \quad \text{y} \in (\frac{N}{2} + 1, N - 1) \end{cases} \tag{2}$$

A Hamming window with its center at $(x = N/2, y)$ is used to reduce boundary artifacts. The average oriented energy of the figure side of the patch parallel to the figure-ground boundary

978-1-4673-3140-1/12/$31.00 ©2012 IEEE

is obtained as,

$$\overline{E}_{f\|}(u) = \frac{1}{K}\sum_{y=0}^{K-1} E_f(u,y) \tag{3}$$

where $K = \frac{N-1}{2}$.

The average oriented energy of the figure side orthogonal to the edge, $\overline{E}_{f\perp}$, is computed analogously (with the one-dimensional Fourier transform now performed on the y co-ordinates, orthogonal to the edge), as are the corresponding quantities (parallel and orthogonal to the edge) for the background, $\overline{E}_{g\|}$ and $\overline{E}_{g\perp}$.

The total oriented spectral energy (a scalar) in the frequency band bounded by u_1 and u_2 and parallel to the figure-ground boundary is defined as,

$$T_{f\|} = \int_{u_1}^{u_2} \overline{E}_{f\|}(u)du \tag{4}$$

For the orthogonal direction, equations 1-4 are modified accordingly. The total oriented spectral energy in the figure orthogonal to the figure-ground edge ($T_{f\perp}$) is computed analogously from $\overline{E}_{f\perp}$. The corresponding quantities on the ground halves of the patches, total oriented spectral energy parallel ($T_{g\|}$) and orthogonal ($T_{g\perp}$) to the figure-ground edge are defined completely analogously.

We can now define the 4–dimensional feature vectors used for the classification process. Their elements are the total energy parallel and orthogonal to the figure-ground edge and the vectors are thus defined as,

$$\mathbf{f} = \begin{bmatrix} T_{f\perp} & T_{f\|} & T_{g\perp} & T_{g\|} \end{bmatrix}^T \tag{5}$$

Clearly, the first two dimensions originate from the figure and the last two from the ground. Odd and even numbered indices correspond to the orthogonal and parallel orientations, respectively.

III. Support Vector Machines

Support Vector Machines (SVMs) are a supervised binary classification technique [10], [11]. The basic idea is that the classification hyperplane is determined by maximizing its distance from the nearest data points (*support vectors*) on either side of the hyperplane.

Let $\{\mathbf{f}_i : i = 1,...,n\}$ be the training data set, where $\mathbf{f}_i \in \mathbf{R}^d$. Let the labels on each sample be $y_i \in \{+1,-1\}$, indicating one of the two classes to which the sample belongs. An SVM finds weights, $\mathbf{w}$ and a bias, b, that minimize the Euclidean norm $\|\mathbf{w}\|$ such that, for all data points ($\mathbf{f_i}, y_i$),

$$y_i(\langle \mathbf{w}, \mathbf{f}_i\rangle + b) \geq 1 \tag{6}$$

The support vectors are the feature vectors $\mathbf{f}_i$ that lie exactly on the decision boundary (closest to the classification hyperplane). Hence, for support vectors, we have $y_i(\langle \mathbf{w}, \mathbf{f}_i\rangle + b) = 1$. In the case of nonlinear SVMs, for instance, for SVMs with radial basis function (RBF) kernels, the inner product is replaced by an appropriate kernel function.

After training, test feature vectors are presented to the SVM. The class a given test vector $\mathbf{z}_j$ is assigned to is computed as sign($\langle \mathbf{w}, \mathbf{z}_j\rangle + b$). More explicitly, $\mathbf{z}_j$ is classified as from class ($y_j = +1$) if,

$$\text{sign}(\langle \mathbf{w}, \mathbf{z}_j\rangle + b) = +1 \tag{7}$$

and from class (-1) if this expression is -1.

Finding optimal weights, $\mathbf{w}$ and bias b requires solution of the following optimization problem:

$$\underset{\mathbf{w},b,\xi}{min} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i \tag{8}$$

$$\text{subject to} \quad y_i(\mathbf{w}^T\phi(\mathbf{f}_i) + b) \geq 1 - \xi_i, \tag{9}$$

$$\xi_i \quad \geq 0 \tag{10}$$

where the function $\phi(.)$ maps the lower dimensional feature vector $\mathbf{f}_i$ into a higher dimensional space, and ξ_i are slack variables. The bounding box (or penalty term) parameter $C > 0$ determines the accepted rate of misclassification, with lower values leading to more misclassification. Finally, $K(\mathbf{f}_i, \mathbf{f}_j) = \phi(\mathbf{f}_i)^T\phi(\mathbf{f}_j)$ is the kernel function. We use radial basis functions with a nonlinear kernel defined as,

$$K(\mathbf{f}_i, \mathbf{f}_j) = \exp(-\gamma\|\mathbf{f}_i - \mathbf{f}_j)\|^2), \ \gamma > 0 \tag{11}$$

where γ is the parameter that controls the width of the kernel. For more detailed discussions of SVM techniques see refs. [12], [13].

IV. Data and Methods

We used two image datasets available on the Internet, the MIT LabelME database [14] and the Berkeley Segmentation Data Set (BSDS300) [15] to generate sets of image patches. Patch selection and processing followed closely the procedures described in ref. [9].

The BSDS300 consists of 300 images, all with a resolution of 481×321 pixels. All but five of the images have human-generated segmentation borders around perceived objects which mostly coincided with the actual figure-ground boundary. Most images were marked by different human observers and consequently have multiple segmentation maps with different numbers of segmented regions in each map. For each image, we chose the segmentation map that had the smallest number of segmented parts. We selected patches of size $N \times N = 33 \times 33$ straddling these borders at random locations along the borders, according to the following procedure. The 2-dimensional pixel locations along the borders were stored in an indexed list, $\mathbf{x}_i, i = 1,...,M$. Indices were drawn randomly from this list (without replacement) and the selected pixel locations then defined the centers of the selected patches. In this way, the center of a patch was randomly located on the contours extracted from segmentation map. We collected 1475 image patches from the BSDS300 dataset, at the rate of 5 patches per image (from all images except the five that did not have segmentation maps). Patches were then rotated such that the figure portion occupied the

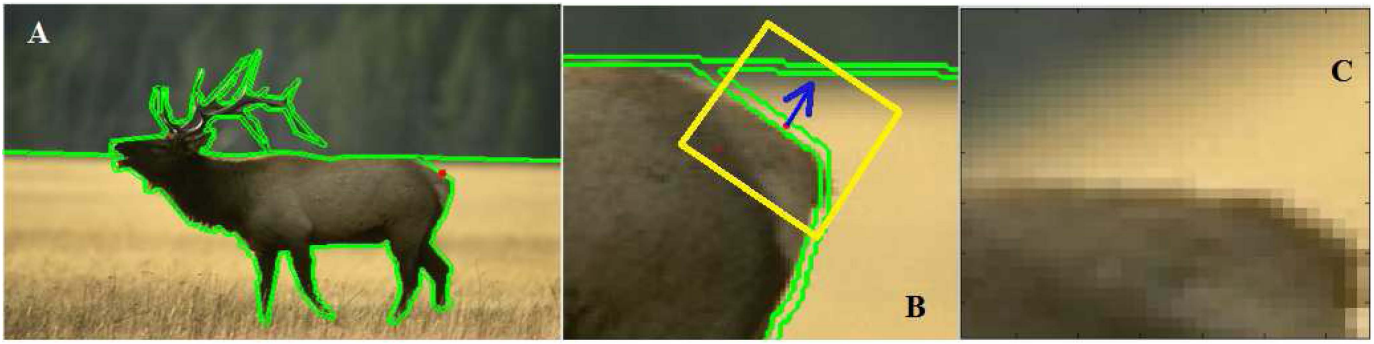

Fig. 1. Patch extraction. (A) Original image with human-drawn segmentation contours (green) overlaid). The location of the patch to be extracted is marked by a red dot. (B) An image patch in its original orientation. The yellow box indicates the area around the patch center that will be part of the rotated patch. The blue arrow points towards the background. (C) The image patch aligned in standard orientation, with the figure part at the bottom and the background part at the top.

Category	Number of Images	Number of Patches
Indoor	199	524
Beach	138	480
Office	62	204
Street	120	340
Forest	64	213
Total	**585**	**1761**

TABLE I
LABELME DATASET: NUMBER OF IMAGES AND PATCHES BY IMAGE CATEGORY

bottom half and the background the top-half of the rotated patch. Subsequently, they were converted to 8-bit gray scale (range $[0, 255]$). Figure 1A shows an example image with a segmentation contour overlaid, and Fig. 1B,C an example of the patch extraction process.

The MIT LabelMe database consists of user-contributed images on which objects have been labeled by users but not segmentation contours. The large number of images in this data base allowed us to minimize the effect of biases such as illumination, foreground and background types, color *etc.* by selecting images from several categories. We chose images from five scene types: office environment, indoor scene, street, beach, and forests. Due to the heterogeneous nature of the database, the sizes of the images varied between 256×256 and 2048×1500 pixels. Since no contour segmentation is available for this image set, we selected the centers of patches to be extracted by hand. In total, 1761 image patches were collected from this database, with a varying number of patches from each image, and the size of each patch the same as for the BSDS dataset (33×33). Details of image and patch selection are given in table I. Patches were then turned in the standard orientation and converted to grayscale, again as described for the BSDS300 data set.

Our interest is not to find the contour separating figure from ground, a task that we assume has been completed, but to determine which side of the contour corresponds to the figure and which to the ground. Therefore, we trained the SVM with a set of patches with positive (correct) and negative (inversed) figure-ground assignment. For each patch, the 4-dimensional feature vector $\mathbf{f}$ described in Section II was computed. Pilot experiments (results not shown) indicated larger differences between figure and ground sides in the higher than in the lower spatial frequencies. We therefore chose $u_1 = 3$ and $u_2 = 8$ in Equation 4. Features were centered and scaled to have unit standard deviation. We use radial basis function (RBF) kernels, see eq. 11. A subset of vectors, $\mathbf{f}'$, were selected as correctly assigned and inserted into a matrix such that each column corresponds to a feature and each row to a sample. A set, of the same size, of inversely assigned patches was constructed by interchanging the indices corresponding to figure and ground. Training of the SVM then proceeded as described in Sec. III, applying Sequential Minimal Optimization (Matlab, Natick MA). A fraction of 5% of the training examples were allowed to violate the Krush-Kuhn-Tucker conditions [13].

The analysis was carried out for LabelME and BSDS databases separately. The patch databases (1761 patches for LabelME, 1475 patches for BSDS300) were divided into training and test sets. Two thirds of the patches were used for training and the remaining one third for testing, and selection for these subsets was random. The training patches are further divided into positive examples and negative examples (50:50 ratio) by switching the indices of features as described. Appropriate class labels, *positive* $= +1$ and *negative* $= -1$ were assigned.

The performance of the classifier with RBF kernels depends on two key parameters, γ which determines the width of the RBF kernel, and C which controls the accepted misclassification rate, see eqs. 8-11. To find the best values of the parameters, a grid search was performed with initial values of both γ and C in the range $[10^{-5}, 10^{5}]$. For each pair of values, we train the classifier with ten fold cross-validation. The trained model with the best cross validation score is used to further refine the values of (γ, C) pairs.

Optimal parameters were determined in a grid search with initial values of both γ and C in the range $[10^{-5}, 10^{5}]$ and initial step size of 1 in the exponent. For each pair of values, we train the classifier with ten fold cross-validation. The

Database	# Samples	γ_{opt}	C_{opt}	CV score	Accuracy
LabelME	587	0.759	1.777	84.07%	67.12%
BSDS	491	0.687	2.282	88.72%	69.25%

TABLE II
SVM RESULTS. FOR BOTH DATABASES (COLUMN 1), WE SHOW THE NUMBER OF IMAGE PATCHES IN THE TEST SET (COL. 2) AND THE OPTIMAL PARAMETERS γ_{opt} (COL. 3) AND C_{opt} (COL. 4). COLUMN 5 SHOWS THE CROSS VALIDATION (CV) SCORES AND COLUMN 6 THE PERCENTAGE OF CORRECT FIGURE-GROUND ASSIGNMENTS.

Database	1 pixel	2 pixel	3 pixel
LabelME	65.76%	66.61%	67.12%
BSDS	67.21%	68.84%	68.02%

TABLE III
ACCURACY WITH PATCHES SHIFTED BY 1, 2 AND 3 PIXELS (SEE TEXT).

trained model with the best cross validation score is used to further refine the values of (γ, C) pairs on a finer grid, where each parameter is systematically varied in small increments (step size 0.1 in the exponent). Once we obtain the optimal parameter values (γ_{opt},C_{opt}), they are used to train the full training set to get our final classifier model.

V. RESULTS AND DISCUSSION

The optimal values (γ_{opt},C_{opt}) were slightly different for the two databases. They are shown in Table II together with the respective cross-validation scores, defined as the accuracy of classification obtained on the training set after the 10-fold cross-validation.

Performance of the trained classifiers was assessed by running them on test data. Our test datasets consisted of 491 samples for the BSDS database and 587 samples for the BSDS database, the results are shown in Table II. We also ran the classifiers after training on the entire patch databases comprising both training and test data (1761 patches from LabelMe, 1475 from BSDS300). As expected, performance was increased, reaching 74.29% and 73.59% accuracy for BSDS300 and LableMe datasets, respectively.

As noted in Section IV, we assume marking of the contour separating the figure and the ground has been completed. We were concerned about possible artifactual results due to slight but systematic misalignment between the actual figure-ground border and the human-generated contour. We therefore applied the already trained SVM model (γ_{opt},C_{opt}) on new data sets generated by shifting the figure and ground parts away from the figure-ground boundary by $1-3$ pixels. We found that this essentially did not change the results, see Table. III.

VI. CONCLUSION

A combination of local and global cues can be used to obtain information about Figure-Ground Organization in complex scenes. While global information is clearly necessary for figure-ground segregation in some situations as shown both in primate neurophysiology [16] and in computational studies [17], we are here interested in the contribution of local cues.

Our results show that the novel measure introduced here, anisotropy of spatial frequency power along and perpendicular to the object boundary, is highly predictive of figure-ground relationships. We further show that this measure is suitable for use in complex natural scenes by nonlinear SVMs with radial basis function kernels, reaching an average of nearly 70% correct classification per patch. If independence between patches can be assumed, high performance can thus be obtained from a small number of patches along a contour at very low computational cost. Furthermore, we speculate that performance could be increased even more by combining this measure with other local features, *e.g.,* extremal edges [9]. We analyze in a separate report [18] why such surprisingly high performance can be achieved with purely local cues.

ACKNOWLEDGMENT

Work supported by Office of Naval Research grant N000141010278 and NIH R01EY016281-02.

REFERENCES

[1] E. Rubin, "Visuell wahrgenommene figuren," *Kobenhaven: Glydenalske Boghandel*, 1921.
[2] ——, *Figure and Ground: in Visual Perception*, S. Yantis, Ed. Psychology Press, 2001.
[3] P. Bahnsen, "Eine Untersuchung uber Symmetrie und Asymmetrie bei visuellen Wahrnehmungen," *Zeitschrift fur Psychologie*, vol. 108, pp. 129–154, 1928.
[4] S. E. Palmer, *Vision Science-Photons to Phenomenology*. Cambridge, MA: MIT Press, 1999.
[5] G. Kanizsa and W. Gerbino, "Convexity and symmetry in figure-ground organization," *Vision and Artifact*, pp. 25–32, 1976.
[6] H.-K. Pao, D. Geiger, and N. Rubin, "Measuring convexity for figure/ground separation," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2, 1999, pp. 948 –955 vol.2.
[7] F. Heitger, L. Rosenthaler, R. V. D. Heydt, E. Peterhans, and O. Kübler, "Simulation of neural contour mechanisms: from simple to end-stopped cells," *Vision Research*, vol. 32, no. 5, pp. 963 – 981, 1992.
[8] S. Palmer and T. Ghose, "Extremal edges: A powerful cue to depth perception and figure-ground organization," *Psychological Science*, vol. 19, no. 1, pp. 77–84, 2008.
[9] S. Ramenahalli, S. Mihalas, and E. Niebur, "Extremal edges: Evidence in natural images," in *Information Sciences and Systems (CISS), 2011 45th Annual Conference on*, march 2011, pp. 1 –5.
[10] V. N. Vapnik, *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.
[11] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, T. Dietterich, Ed. MIT Press, 2002, vol. 98, no. 462.
[12] C.-w. Hsu, C.-c. Chang, and C.-j. Lin, "A practical guide to support vector classification," *Bioinformatics*, vol. 1, no. 1, pp. 1–16, 2010.
[13] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
[14] B. Russell, A. Torralba, and W. Freeman, "Labelme: The open annotation tool." [Online]. Available: http://labelme.csail.mit.edu/
[15] C. Fowlkes, D. Martin, and J. Malik, "Local figure-ground cues are valid for natural images," *Journal of Vision*, vol. 7, no. 8, 2007.
[16] H. Zhou, H. Friedman, and R. Von Der Heydt, "Coding of border ownership in monkey visual cortex," *Journal of Neuroscience*, vol. 20, no. 17, pp. 6594–6611, 2000.
[17] E. Craft, H. Schütze, E. Niebur, and R. Von Der Heydt, "A neural model of figure-ground organization," *Journal of Neurophysiology*, vol. 97, no. 6, pp. 4310–4326, 2007.
[18] S. Ramenahalli, S. Mihalas, and E. Niebur, "Spectral anisotropy provides information for figure-ground organization in natural images," (in preparation).