Fast 2D Border Ownership Assignment

Ching L. Teo
cteo@cs.umd.edu

Cornelia Fermüller
fer@umiacs.umd.edu

Yiannis Aloimonos
yiannis@cs.umd.edu

Computer Vision Lab, University of Maryland, College Park, MD 20742, USA

Abstract

A method for efficient border ownership assignment in 2D images is proposed. Leveraging on recent advances using Structured Random Forests (SRF) for boundary detection [8], we impose a novel border ownership structure that detects both boundaries and border ownership at the same time. Key to this work are features that predict ownership cues from 2D images. To this end, we use several different local cues: shape, spectral properties of boundary patches, and semi-global grouping cues that are indicative of perceived depth. For shape, we use HoG-like descriptors that encode local curvature (convexity and concavity). For spectral properties, such as extremal edges [28], we first learn an orthonormal basis spanned by the top K eigenvectors via PCA over common types of contour tokens [23]. For grouping, we introduce a novel mid-level descriptor that captures patterns near edges and indicates ownership information of the boundary. Experimental results over a subset of the Berkeley Segmentation Dataset (BSDS) [24] and the NYU Depth V2 [34] dataset show that our method's performance exceeds current state-of-the-art multi-stage approaches that use more complex features.

1. Introduction

Look at the two images in Fig. 1 with highlighted boundaries on the right. These are regions in the image where objects meet with one another or with the background. Humans are able to interpret complex scenes such as these and predict their approximate depth orderings with relative ease by integrating both bottom-up and top-down cues. In recent years, so-called boundary detectors have become very popular tools. These detectors use local cues, such as brightness, color, texture, gradients and simple features [24] in image patches to distinguish edge points likely at boundaries of surfaces from others. More recent approaches also include globalization processes using long-range relations of image points [2]. However, the image structure in the regions next to an occlusion edge can be used for more than

Figure 1. Example results of predicted boundaries (blue) and their ownership (red: foreground, yellow: background) from real-world images: BSDS (above) and NYU Depth V2 (below). Best viewed in digital copy.

boundary indication; it also encodes information about the relative depth about the edge's two adjacent regions, and to which of the regions the edge belongs to. It has been shown that image cues, such as the convexity of the edge [19], the edge junctions, contrast, or the gradient in the intensity and the texture carry this information [28]. In this paper, we focus on detecting classes of bottom-up cues that indicate *border ownership*, i.e. the information on which side of a boundary belongs to the foreground/object or the background, from 2D images. Determining border ownership is important from a computer vision perspective since it can be regarded as a preprocessing step for foreground-background segmentation [32], and is also closely related to selective attention [4]. From a biological viewpoint, neurophysiological findings from the visual cortex of macaque monkeys together with psychophysical experiments also suggest that the human visual cortex has specialized cells that perform some form of ownership prediction [36]. These mechanisms have been found in cortical areas V2 and V4 of monkeys [38], and they may be receiving feedback from higher cortical regions [4].

Fig. 1 shows example predictions using our proposed approach with their accuracy scores over two popular datasets: the Berkeley Segmentation (BSDS) and the NYU Depth V2 (NYU-Depth) [24, 34]. The prediction accuracy not only is state-of-the-art, but outperforms previous approaches [30, 22]. Our method exploits two novel features derived from findings in human psychophysics to determine the ownership of a boundary. The first one, known as *extremal edges* or image *folds* [20], captures how changes in the shading of pixels near real boundaries differ between foreground and background. It was shown in [29] that such folds exist in a variety of environments. So far this cue has not been exploited very efficiently for computer vision. [22] proposed to compute a measure based on the change of intensity perpendicular to previously detected edge points. Here we obtain the extremal edge cue from the principal components of grayscale image patches [18]. In order to adapt these patches better to the local shape of the edge, we adopt the framework of Sketch Tokens [23] and learn an orthonormal basis for each token class. As we show in §4.2, the top principal components that we retain encode not only extremal edges but also more complex structures such as T-junctions and parallel lines which are equally important cues for ownership assignment. We then derive *spectral features* that capture these local grayscale variations from the projections of these principle components. Intuitively spectral features are more important for close-up scenes. This is confirmed in our experiments, which show that the extracted spectral features from the NYU-Depth indoor dataset with structured lighting are more useful for assigning border ownership than those from the BSDS dataset, which consists of natural outdoor images.

The second feature detects Gestalt-like groupings of *mid-level* cues. Specifically, we introduce a new multi-scale grouping mechanism that implements the concept of contour closure, and common patterns such as radial and spiral textures. Since such patterns occur naturally in images, we expect the differences in the distribution of these patterns to be indicative of border ownership. Finally, by embedding these features within a Structured Random Forest (SRF), we are able to predict border ownership in *real-time*, $\approx$ 0.1s for a 320$\times$240 image. Notably, our method predicts both boundary *and* ownership together in a single step. Compared to previous works that considered border ownership determination as a separate step independent of boundary detection, our single-step approach is not only faster but also more accurate.

2. Related Works

Determining border ownership accurately in images involves several related works in computer vision which can be classified into two different areas: 1) depth ordering prediction and 2) object proposals. We briefly review each area in relation to the current work.

Depth ordering prediction. Perceiving ordinal depth from 2D images has been tackled as early as the classical "Blocks World" of Roberts [31]. Hoeim et al. [16] revisited the problem by combining numerous local and global cues: color, gradients, junctions, textures, sky above ground etc. into a large conditional random field (CRF) for recovering occlusion boundaries and depth ordering in a 2D image. The CRF weights were obtained from training data to ensure consistency of depth across different segments, which were merged in an iterative process from an initial over-segmentation. Along similar lines, Saxena et al. [33] imposed simple geometric constraints to estimate plane parameters related to the 3D location and orientation of each image patch to create a 3D pop-out of the image. Ren et al. [30] considered local convexity and junction cues and integrated them into a CRF to predict border ownership on Pb boundaries [24]. Leichter and Lindenbaum [22] followed up by computing distributions of ownership cues in ordinal depth: parallelity, image folds, lower-region etc. over curves, T-junctions and image segments. Stein and Hebert [35] further imposed motion constraints to detect occlusion boundaries consistently across video frames.

Object proposals. A recent trend in computer vision is to detect from an image, object-like regions in the foreground. Early works [1, 10] combined several "objectness" cues to train detectors. However, the applicability of such methods are limited as cue detection and integration is computationally expensive. Recently, Cheng et al. [3] introduced a surprisingly simple technique using binarized gradient norms of images that is able to produce high quality proposals at a fraction of the time of previous methods. The Gestalt concept of *closure* has been exploited by Nishigaki et al. [25, 37] in detecting object like regions via a mid-level grouping operator termed "image torque". Similarly, using a SRF based structured edge (SE) detector [8], Zitnick et al. [39] counts the number of contours that enter and exit a bounding box region to determine if there is enough closure within the proposed region.

Although many of these works have considered the border ownership problem implicitly in their problem formulation, it is often considered as an independent pixel-wise classification step over predicted input boundaries [30, 22] or segmentations [16, 35]. In order to ensure prediction consistency over larger scales, CRFs are often used at the expense of computation time. Our approach, by contrast, considers border ownership and boundary detection within a single SRF where consistency over multiple scales are enforced using structured output labels. Our approach is therefore self-contained: we predict both boundaries and border ownership in one single step unlike previous approaches that require further optimizations using a CRF. Consequently, our approach affords us to predict border

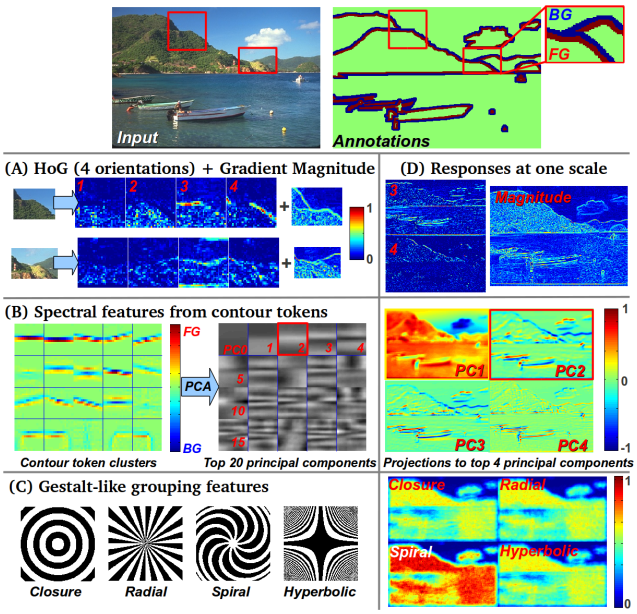

Figure 2. Border ownership cues used. (Top) Input image and annotations (red: foreground, blue: background) with example patches boxed. (Below) (A) Local shape (HoG + gradient magnitude) showing four discrete orientations, (B) Spectral features derived via PCA from 20 oriented token clusters (foreground at lower half) and their principal components with extremal edge cues in PC2 (boxed) and (C) Gestalt-like grouping target patterns: closure, radial, spiral and hyperbolic. (D) Corresponding responses at one scale for each of the features. See text for details.

ownership in real-time.

3. Approach

Our approach of determining border ownership via SRFs consists of two key components: 1) Features derived from ownership cues and 2) Imposing border ownership structure in the SRF. We describe these two components in the sections that follow.

3.1. Border ownership cues

We use some local cues reported in prior works [30, 22, 28, 26, 13, 9] that were shown to be important in determining border ownership and some new cues. Specifically, we use: 1) shape (convexity/concavity), 2) image folds or extremal edges derived from spectral properties of boundary patches and 3) Gestalt-like grouping features. In addition, our choice of features was influenced by how efficient we can extract them from local patches.

3.1.1 HoG-like descriptors

As reported in several previous works, shape cues such as local convexity and concavity of contours are important features that are indicative of foreground objects: the foreground ownership of a boundary tends to be on the concave side [26]. To capture this cue within a local patch, we construct a HoG-like descriptor [6] of image gradients where we quantize the gradient directions into 4 orientation bins. In addition, we use the gradient magnitude as an indicator for good boundary localization. The HoG-like descriptor of gradient orientations captures roughly the local shape of the patch, while its magnitude tells us how likely this patch should contain a real boundary. Notably, as shown in Fig. 2 (A), we see that the histograms for typical convex and concave patches are different. For efficiency, we compute these features in terms of "channels" [7] per image patch. Given a patch $\mathbf{P}$ of size $N \times N$, this results in a $N^2 \times 5$ feature vector per patch.

3.1.2 Extremal edges from PCA of contour tokens

Extermal edges, or image folds have been known for some time as one of the strongest border ownership cues [13, 9]. Huggins et al. [17] have shown that extremal edges can be reliably detected by computing the so-called shadow flow field in controlled environments. Recently, [29] have shown that extremal edges exists in natural images by performing a principal component analysis (PCA) of aligned oriented boundary image patches. Their key insight is that extremal edges account, after step edges, for most of the gray-level illumination variance at such regions. Motivated by this insight, we derived the basis functions using PCA oriented along so called contour fragments or Sketch Tokens [23] which are similar to shapemes [30] as shown in Fig. 2 (B). Since each contour token has an orientation determined by its foreground and background labels, we first orientate all patches so that the background and foreground occupy the top and bottom halves of the patch (using the center pixel as a reference) respectively. Clustering these orientated tokens produces a set of C token centers to which we then apply PCA over the S supporting patches, $\mathcal{P}_c = \{\mathbf{P}_1, \ldots, \mathbf{P}_S\}, c \in \{1 \ldots C\}$. By applying PCA over each $\mathcal{P}_c$, we learn a separate orthonormal basis corresponding to each token center. Specifically, given the $N^2 \times S$ data matrix $\mathbf{X}$ that contains at each column a vectorized (and demeaned) $\mathbf{P}_c$, we apply Singular Value Decomposition on its covariance matrix $\Sigma_{\mathbf{X}}$ to obtain a set of orthonormal basis spanned by the eigenvectors (columns) of $\mathbf{U}$:

$$\Sigma_{\mathbf{X}} = \mathbf{U}\mathbf{S}\mathbf{U}^{-1} \tag{1}$$

where we keep the top K eigenvectors, $u_k \in \mathbf{U}$, corresponding to the top K eigenvalues in $\mathbf{S}$ to obtain the projection matrix $\mathbf{W}_c = [u_1, \ldots, u_K]$. $\mathbf{W}_c$ represents a new basis that accounts for most of the variance per contour token center. As features, we reproject $\mathbf{X}$ to obtain $\mathbf{Y}_{K \times S} = \mathbf{W}_c^T \mathbf{X}$, the coordinates of each patch $\mathbf{P}_c$ in the new basis. This yields a feature vector of dimensions $N^2 \times K$. We show in

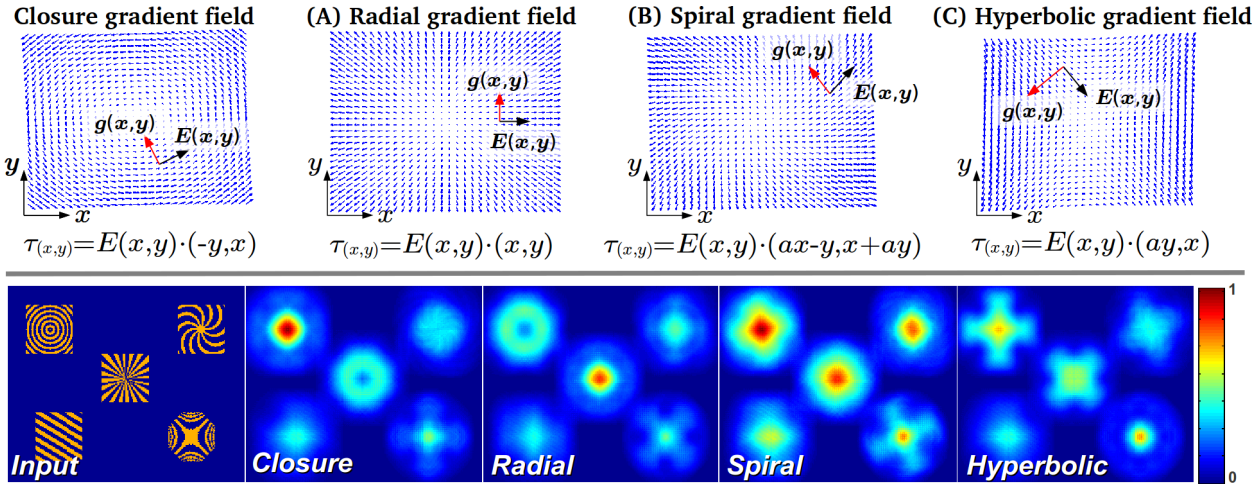

Figure 3. Generalizing the image torque for different Gestalt groupings. (Top) By rewriting $\tau_{(x,y)}$ in terms of a scalar product, we are able to generalize the image torque so that it becomes sensitive to: A) radial, B) spiral and C) hyperbolic patterns. (Bottom) Test toy image with different target patterns and their maximum responses over different scales. Notice the selective nature for each target pattern.

Fig. 2 (D-middle) the spectral features derived from the first four principal components (PC). Of note are the responses for PC2-PC4 which exhibit a large response only along real boundaries with positive values encoding foreground ownership and negative values encoding background ownership. In §4.2, we show further that PC2 exhibits the characteristics of extremal edges.

3.1.3 Gestalt-like grouping features

Gestalt psychologists have developed a set of well-known rules of "Gestalt" that suggests how humans perceive the world from 2D images. Gestalt rules deal with groupings of low-level features (e.g. edges), and can be regarded as a form of *mid-level* cue that captures the holistic properties of individual visual parts. These properties can then be used to organize these visual parts into more meaningful entities that serve as input to higher level processes: e.g. segmentation, recognition etc. In this work, we leverage on specific grouping patterns: 1) closure, 2) radial, 3) spiral and 4) hyperbolic (Fig. 2 (C)). Such patterns are useful for border ownership determination because foreground objects tend to exhibit different grouping patterns compared to the background [27], and such patterns have been observed in area V4 of macaques [11]. Closure, one of the strongest cues used in foreground object proposals tasks, is detected in this work by computing the image "torque" [25], $\tau_{\mathbf{P}}$, associated at each patch (Fig. 3 (Top-left)). The image torque is so-termed because it is analogous to the torque formulation known in physics, which is the cross-product between a tangential "force" vector $\vec{F}_q$ and its corresponding displacement vector $\vec{d}_{pq}$ where p denotes the center pixel in $\mathbf{P}$ and q an edge pixel in $\mathbf{P}$. The image torque for each edge point q is thus defined as $\tau_{pq} = \vec{F}_q \times \vec{d}_{pq}$. Summing up all $q \in \mathbf{P}$ and normalizing with the patch size yields $\tau_{\mathbf{P}}$:

$$\tau_{\mathbf{P}} = \frac{1}{2|N|} \sum_{q \in \mathbf{P}} \tau_{pq} = \frac{1}{2|N|} \sum_{q \in \mathbf{P}} \left(\vec{F}_q \times \vec{d}_{pq} \right) \quad (2)$$

In practice, we search over several scales $s \in \{5, 6, \cdots, N\}$ within $\mathbf{P}$ and retain the maximum torque response over all scales. An alternative derivation for $\tau_{\mathbf{P}}$ is to view the detection of closure patterns as detecting iso-contours corresponding to circles in the image. In general, we consider the patterns we want to detect as the iso-contours of some function f. For example circles are the iso-contours of the function $f(x, y) = x^2 + y^2$. We are interested in the tangent lines of these iso-contours, $g(x, y)$. Given the 2D gradient field, $\nabla f(x, y) = (f_x, f_y)$, the corresponding tangent vectors perpendicular to the gradient field are thus $g(x, y) = (-f_y, f_x)$. From the iso-contour equation of circles, it follows that the closure tangent vectors are $g(x, y) = (-y, x)$. Given an input test patch $\mathbf{P}$, we first determine its gradient field, denoted as $\nabla P(x, y) = (P_x, P_y), (x, y) \in \mathbf{P}$, and their edges (tangent vectors) as $E(x, y) = (-P_y, P_x)$. If a closure pattern exists in $E(x, y)$, then the edges must align well with tangent vectors $g(x, y)$. A simple measure of alignment for a point $(x, y) \in \mathbf{P}$ is thus the scalar product between $E(x, y)$ and $g(x, y)$:

$$\tau_{(x,y)} = E(x, y) \cdot g(x, y) = (-P_y, P_x) \cdot (-y, x) \quad (3)$$

which is equivalent to the definition of τ_{pq} for point q. Replacing τ_{pq} in eq. (2) with eq. (3) yields exactly the same results. The key insight from eq. (3) is that we are now able to modify $g(x, y)$ so that eq. (3) is sensitive to different patterns in the image. As we show in Appendix A, by writing different target iso-contour equations, we are able to detect different Gestalt patterns using the same formulation.

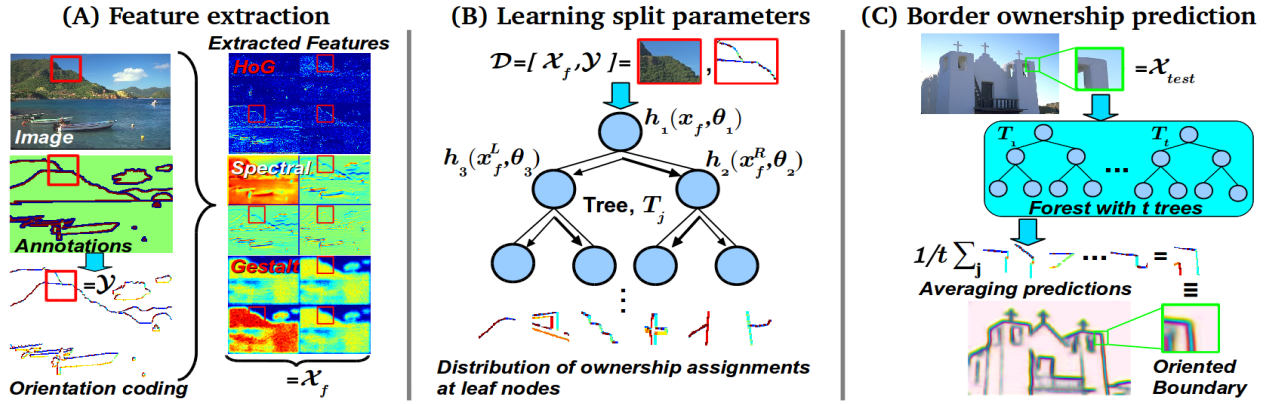

Figure 4. Training a SRF for border ownership assignment. (A) Example image with extracted features $x_f \in \mathcal{X}_f$ and ground truth annotations from the highlighted patch. We derive an orientation coding, $\mathcal{Y}$, from the annotations. (B) By mapping $\mathcal{Y}$ to discrete labels, we determine the optimal split parameters θ associated with each split function $h(x_f, \theta)$ that send features x_f either to the left or right child. The leaf nodes store a distribution of border ownership structured labels. (C) During inference, a test patch is assigned to a leaf node within a tree that contains a prediction of the border ownership. Averaging the prediction over all t trees yields the final ownership prediction. We then convert the orientation code into an oriented boundary (blue) that encodes the foreground (red) and background (yellow) predictions.

We show some sample responses using different $g(x, y)$ in Fig. 3 (Bottom) for four patterns: closure, radial, spiral and hyperbolic. For efficiency, we have implemented eq. (2) as a convolution operation so that their responses can be used directly as features of size $N^2 \times 4$ for training the SRF. Additionally, the responses of the Gestalt features for an example input image are shown in Fig. 2 (D-below). We note that because the background (e.g. sky) tends to be textureless, all the features have a small response. Notably, we observe that the strongest response occurs for the spiral pattern, which is localized in the forested foreground region.

3.2. Border ownership assignment via SRF

We use an extension of the Random Forest (RF) classifier [15], termed the Structured Random Forest (SRF). Similar to the RF, a SRF is an ensemble learning technique that combines t decision trees, $(T_1, \cdots, T_t)$, trained over random permutations of the data to prevent overfitting. The key difference is that in general, SRFs are able to learn a mapping between inputs of arbitrary complexity (e.g. strings, segmentations, relationships etc.) and similarly complex outputs. Due to their flexibility in representation, SRFs have been used successfully in a variety of computer vision tasks such as boundary detection [8] and semantic scene segmentation [21]. See [5] for a comprehensive review of RFs and their applications. In this work, we show that a SRF can be used as a border ownership classifier by imposing a spatial border ownership structure in the output labels (Fig. 4). Similar to [8], we assume that only the target output labels are structured (borders with ownership labels) while the inputs are non-structured (feature vectors derived from image patches).

Let us denote the input as $\mathcal{X}_f$ composed of features $x_f \in \mathcal{X}_f$ derived from a training patch $\mathbf{P}$. The target output is a structured label $\mathcal{Y} = \mathbb{Z}^{N \times N}$ that contains the *orientation* coded annotation of the border ownership. Using a 8 way local neighborhood system, this amounts to 8 different possible orientations of border ownership (Fig. 4 (A-bottom)) that each decision tree will predict. The goal of training a SRF (or a RF in general) is to determine, for the i^{th} split (internal) node, the parameters θ_i associated with a binary split function $h(x_f, \theta_i) \in \{0, 1\}$ so that if $h(\cdot) = 1$ we send x_f to the left child or to the right child otherwise. We define $h(x_f, \theta_i)$ to be an indicator function with $\theta_i = (k, \rho)$ and $h(x_f, \theta_i) = \mathbf{1}[x_f(k) < \rho]$, where k is the feature dimension corresponding to one of the features described in §3.1. Following [12], we select at most $\sqrt{k}$ feature elements for evaluation. ρ is the learned decision threshold that splits the data $\mathcal{D}_i \subset \mathcal{X}_f \times \mathcal{Y}$ at node i into $\mathcal{D}_i^L$ and $\mathcal{D}_i^R$ for the left and right child nodes respectively. ρ is based on maximizing a standard information gain criterion M_i:

$$M_i = H(\mathcal{D}_i) - \sum_{o \in \{L,R\}} \frac{|\mathcal{D}_i^o|}{|\mathcal{D}_i|} H(\mathcal{D}_i^o) \qquad (4)$$

We use the Gini impurity measure: $H(\mathcal{D}_i) = \sum_y c_y(1 - c_y)$ with c_y denoting the proportion of features in $\mathcal{D}_i$ with ownership label $y \in \mathcal{Y}$. For non-structured $\mathcal{Y}$, computing eq. (4) is straightforward. In the case of structured labels, we first compute an intermediate mapping $\Pi : \mathcal{Y} \mapsto \mathcal{L}$ of structured labels into discrete labels $l \in \mathcal{L}$ following [8] that allows us to compute eq. (4) directly. $\mathcal{L}$ is a set of labels that corresponds to different types of possible contour token centers (see §3.1.2), and this means that we can reuse the results from the feature extraction step during training for added efficiency.

The process is repeated with the remaining data $\mathcal{D}^o, o \in \{L, R\}$ at both child nodes until a terminating criterion is satisfied. Common terminating criteria are: 1) maximum depth of tree d_t is reached, 2) a minimum input $|\mathcal{D}|$ is achieved or 3) the gain in M_i is too small. The leaf nodes of each tree after training thus contain the predicted local ownership orientation decision y (Fig. 4 (B)). Note that unlike the RF, where a prediction is performed independently per pixel, the SRF enforces spatial consistency in the structured labels at the leaf nodes so that the final predictions do not change too much along boundaries. In order to account for scale variations, we further sample patches from three (original, half and double) different resolutions of the input image. During inference, we sample test patches densely (at the original resolution) over the entire image and classify them using all t decision trees in the SRF. The final ownership label at each pixel is determined by averaging the predicted orientation labels across all t trees, producing an orientation code that we convert directly into an oriented boundary representation (Fig. 4 (C)).

4. Experiments

4.1. Datasets, baselines and evaluation procedure

We evaluate the performance of border ownership assignment over two publicly available datasets containing real world images: 1) The Berkeley Segmentation Dataset (BSDS) [24] and 2) The NYU Depth V2 (NYU-Depth) dataset [34]. For BSDS, we use a separate subset of 200 labeled images (obtained from the training subset of BSDS-300) that contains ownership annotations. As this dataset was used by the two baseline approaches: 1) Global-CRF of Ren et al. [30] and 2) 2.1D-CRF of Leichter and Lindenbaum [22], the results we report in §4.3 are directly comparable. We use the same test/train split as both baselines, with 100 images for training and 100 images for testing. The NYU-depth dataset consists of 1449 RGB-Depth images taken from a variety of indoor environments. The training set consists of 795 images while the remaining 654 images are used for testing. All images in the dataset are hand annotated with 1000+ object class labels. Following [14], we select the top 35 most frequent object labels (excluding flat surfaces such as walls, floors and ceilings) in order to automatically generate a large number of ownership labels along the boundaries of these objects, using the depth information to produce the ground truth labels for the entire dataset. Compared to BSDS, where only 36.1% of boundary pixels have ownership annotations, we increase the annotation density to nearly 50% in NYU-Depth. Several examples of the input data, ground truths and results are shown in Fig. 5. Full results and videos that show real-time ownership assignment in cluttered kitchen scenes are available in the supplementary material.

Notation	Description	Value
N	patch size	16
C	number of contour token clusters	20
K	principal components used	5
t	number of trees	16
d_t	maximum tree depth	64

Table 1. Parameters used for training the SRF.

We report the same accuracy evaluation metric used in [30] and [22], where we count the number of correctly classified border ownership pixels against the ground truth. This is computed via a bipartite graph matching to determine the closest correspondences between the predicted border ownership pixels and the ground truth. Predictions that were not matched are not considered. Following [22], we set this threshold to 0.75% of the image diagonal. The parameters used for training the SRF are the same for both datasets and we summarize them in Table 1.

4.2. Comparing spectral components

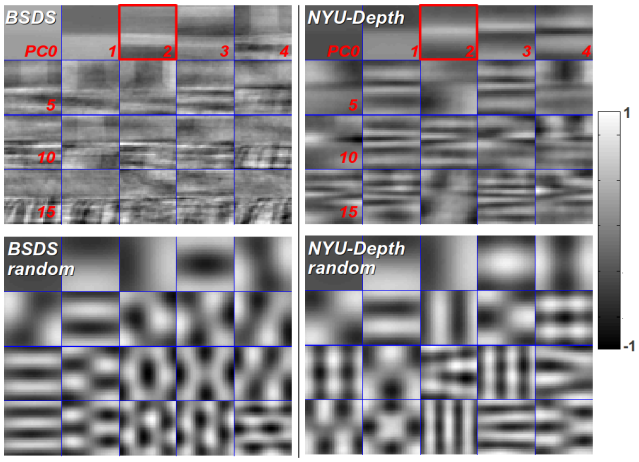

Figure 6. Top 20 principal components for BSDS (left) and NYU-Depth (right) for a particular token cluster center. (Bottom row) Components derived from random patches in each dataset.

Before we present evaluation results of the approach, we first perform an analysis of the spectral components produced by applying PCA over clustered token patches in both the indoor (NYU-Depth) and outdoor (BSDS) datasets. We show in Fig. 6 a visual comparison of the top 20 principal components (PC) obtained from one token cluster center: horizontal with the background at the top half and the foreground at the lower half of each patch, baselined against components derived from random patches (bottom row). In both datasets, we sampled 500,000 patches. We make four observations. First, the top component (PC1) is the same for both BSDS and NYU-Depth, which is a step edge. The second component (PC2, boxed in Fig. 6) exhibits the distinctive signature of extremal edges: with a shading on the lower-half (foreground) and no shading in the top-half

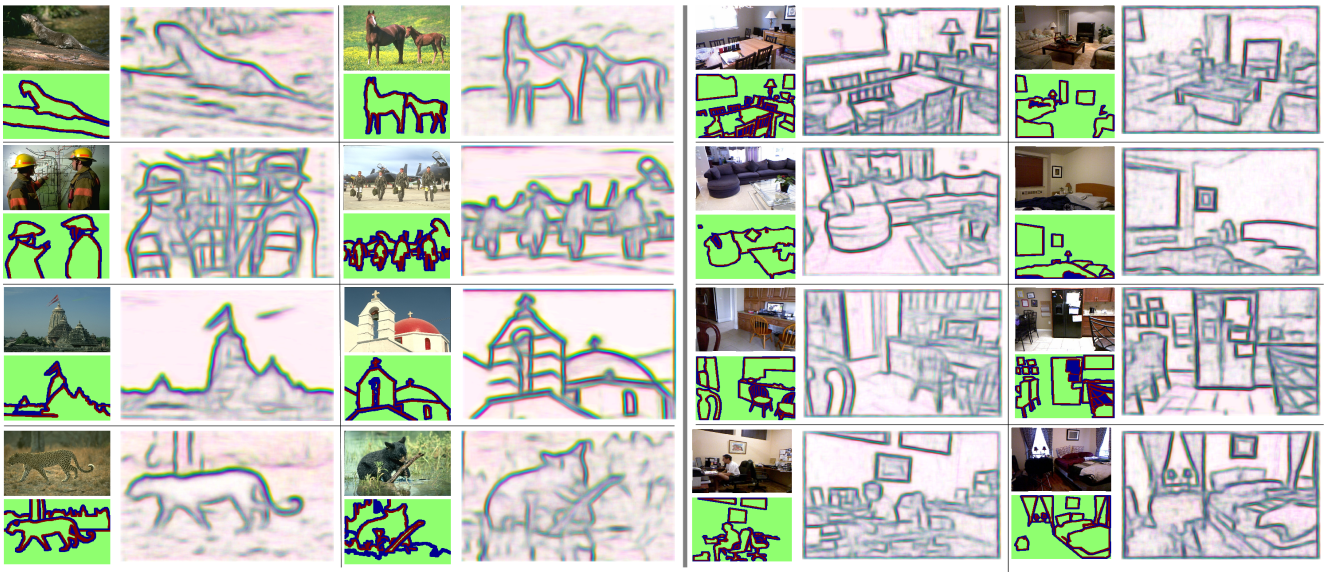

Figure 5. Example results from both BSDS (left panel) and NYU-Depth (right panel) datasets. Eight results per dataset: (Top-left counterclockwise): images, ground truth labels (red: foreground, blue: background) and ownership prediction (red: foreground, yellow: background, blue: boundaries). Best viewed in digital copy.

(background). This confirms the observations made by Ramenahalli et al. [29] on the basis of a much smaller number of images (585), and confirm that extremal edges are present across different scenes and environments. Second, we note that the intensity variation in PC2 from NYU-Depth appears "smoother" across the foreground region compared to BSDS. This seems to indicate that extremal edges are more stable in the indoor NYU-Depth dataset. One possible explanation would be that the structured lighting in indoor environments supports the existence of extremal edges better than the diffused lighting common in outdoor situations. Third, we note that other ownership cues such as T-junctions and parallel structures are also captured within the top PCs of both datasets (e.g. PC6 and PC9). Finally, as none of the PCs from random patches exhibit the signature of extremal edges (or other ownership cues), this further confirms that the spectral features we use are unique along true object boundaries.

4.3. Results

We perform a series of quantitative ablation studies over different features sets in both datasets and compared their performance with the baselines Global-CRF and 2.1D-CRF in the BSDS dataset. In a second experiment, we also applied the basis functions learned from NYU-Depth (indoor) over the BSDS dataset in order to validate our observations in §4.2 that the spectral components from the indoor NYU-Depth scenes are more informative than those obtained from BSDS (outdoor). The full results are summarized in Table 2. We show the contribution for individual features, as well as the improvements when the feature is used with other cues. As a point of reference, we note that for BSDS, we are classifying over 18,000 pixels, while we are approaching 2,500,000 pixels for NYU-Depth. Finally, since our approach predicts boundaries in addition to ownership, we evaluate its boundary prediction accuracy in a third experiment (Table 3).

Feature set	BSDS	NYU-Depth
HoG	72.0%	66.0%
+ Spectral (no contour tokens)	73.1% (72.0%)	67.0% (65.6%)
+ Spectral (contour tokens)	74.0% (72.3%)	68.1% (66.7%)
+ Gestalt patterns	74.4% (72.7%)	**68.4**% (66.7%)
All features + Spectral (NYU)	**74.7**% (72.8%)	-
Global-CRF [30]	69.1%	-
2.1D-CRF [22]	68.9%	-

Table 2. Border ownership prediction accuracy for various ablations compared with the baselines (last two rows). '+' denotes the addition of new features to those above the current row. Numbers in parenthesis denote the use of the single feature for prediction.

Method	BSDS-500	NYU-Depth
Our approach	0.73,0.74,0.76	0.63,0.64,0.60
gPb-owt-ucm [2]	0.73,**0.76**,0.73	0.63,0.66,0.56
SE [8]	0.73,0.75,**0.77** (SE-SS)	**0.65**,**0.67**,**0.65** (SE-RGB)

Table 3. Boundary prediction accuracy. The numbers reported in each cell are [ODS, OIS, AP] following [2]. Results for gPb-owt-ucm and SE are reproduced from [8].

Ablation studies of different features. The first four rows in Table 2 summarize the mean accuracy of border ownership assignment when different combinations of feature sets are used. The general trend is that with more cues used, the ownership prediction improves for both datasets. We note that the results confirm the usefulness of learning sep-

arate basis functions corresponding to different contour token centers (third row), where there is around 1% improvement in accuracy over the case where no contour tokens are used (second row). For the latter, we simply learned a basis over 8 ownership orientations. We also show the contribution of individual features in parenthesis. Of interest is that Gestalt-like features perform on par with spectral features in the NYU-Depth dataset while they have a larger individual influence in BSDS. A likely explanation is that most indoor man-made objects are *textureless* compared to outdoor environments. Additional experiments with more controlled environments have to be done to confirm this hypothesis.

Applying NYU-depth (indoor) spectral features to BSDS dataset. In the second experiment, we applied the basis functions obtained from NYU-Depth to the BSDS dataset. This results in a slight improvement to 72.8% of its individual contribution. Due to this small degree of improvement, more experiments with a more careful selection of indoor patches should be performed to confirm our hypothesis in §4.2. Nonetheless, we note that combining NYU-Depth spectral features with other features yield the best overall prediction accuracy for BSDS (74.7%) in all experiments.

Comparison with state-of-the-art. The prediction accuracy of the proposed SRF border ownership assignment outperforms previous state-of-the-art results: 1) Global-CRF and 2) 2.1D-CRF by at least 2% even using simple HoG-like (shape) features in the BSDS dataset. The performance when all features are combined is even more significant: $> 5\%$ or around 900 pixels that were reclassified correctly. Compared to 2.1D-CRF with a reported mean run-time of 15s, inference using the SRF is $\approx$100 times faster (0.1s).

Boundary prediction accuracy. Our approach (using all features) produces reasonable *boundary* (not ownership) predictions that are comparable with state-of-the-art boundary detectors: gPb-owt-ucm [2] and structured edges (SE) [8] when evaluated over the larger BSDS-500 [2] and NYU-Depth datasets (Table 3). Since our approach evaluates test patches at the original resolution without any depth information, we compared the closest variants of SE: SE-SS (single scale) and SE-RGB (no depth) in BSDS-500 and NYU-Depth respectively. Ablations of features produce insignificant deviations from these results, which shows that the proposed features are more suitable for ownership than boundary prediction. Furthermore, these results are even more significant since our approach is trained on a smaller subset of ownership labels in both datasets.

5. Conclusions

We have presented a fast approach for border ownership assignment that outperforms two current state-of-the-art approaches using CRFs. The approach exploits the speed and flexibility in the representation of Structured Random Forests so that ownership structure is imposed in the final output labels of each decision tree. We have also developed novel feature representations that capture perceptually salient ownership cues: 1) extremal edges and 2) Gestalt-like groupings. For extremal edges, we first learn separate basis functions clustered at contour token centers to capture local shape better. Re-projecting the input image into the new basis produces a set of spectral features in which the top components capture a variety of ownership cues including extremal edges. For detecting Gestalt-like groupings, we reformulated a recently introduced closure operator (the image torque) so that it generalizes to a variety of grouping patterns in the image.

As border ownership assignment is one of the key steps for depth perception, we plan to extend this work by adding in more cues: motion, focus, other Gestalt-like groupings (e.g. symmetry) and higher-level cues for scene understanding (e.g. semantic labels). By making this efficient border ownership assignment code available[1], we also provide a tool to the community that others can explore in tasks such as segmentation and recognition.

A. Generalizing the image torque to other patterns

Following the notations used in §3.1.3, we write down the following iso-contour functions for: 1) radial $f_r(x,y)$, 2) spiral $f_s(x,y)$ and 3) hyperbolic $f_h(x,y)$:

$$
\begin{aligned}
f_r(x,y) &= \operatorname{atan}\left(\frac{y}{x}\right) \Rightarrow \nabla f_r(x,y) = \begin{pmatrix} \frac{y}{\sqrt{x^2+y^2}} \\ \frac{-x}{\sqrt{x^2+y^2}} \end{pmatrix} \\
f_s(x,y) &= x^2 - ay^2 \Rightarrow \nabla f_s(x,y) = \begin{pmatrix} x \\ -ay \end{pmatrix} \\
f_h(x,y) &= \ln(\sqrt{x^2+y^2}) - a \operatorname{atan}\left(\frac{y}{x}\right) \Rightarrow \\
&\nabla f_h(x,y) = \frac{1}{x^2+y^2}\begin{pmatrix} ax - y \\ x + ay \end{pmatrix}
\end{aligned}
\tag{5}
$$

which leads to the following expressions for the tangent vectors $g(x,y)$:

$$
\begin{aligned}
g_r(x,y) &= (x,y) \\
g_s(x,y) &= (ax - y, x + ay) \\
g_h(x,y) &= (ay, x)
\end{aligned}
\tag{6}
$$

for some values of $a = \{\frac{1}{3}, 1, 3\}$. Substituting the corresponding $g(x,y)$ from eq. (6) in eq. (3) enables us to compute the alignment of the target pattern in the image.

Acknowledgments

This work was funded by the support of the European Union under the Cognitive Systems program (project PO-

[1] Code, data and supplementary material are available at: http://www.umiacs.umd.edu/~cteo/BOWN_SRF

ETICON++), the National Science Foundation under INSPIRE grant SMA 1248056, and by DARPA through U.S. Army grant W911NF-14-1-0384. We thank the reviewers for their constructive feedback and I. Leichter for initial help with the BSDS dataset.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *PAMI*, 34(11):2189–2202, 2012. 2

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 33(5):898–916, 2011. 1, 7, 8

[3] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*, pages 3286–3293, 2014. 2

[4] E. Craft, H. Schütze, E. Niebur, and R. von der Heydt. A neural model of figure-ground organization. *J. Neurophysiology*, 97(6):4310–4326, 2007. 1

[5] A. Criminisi and J. Shotton. *Decision forests for computer vision and medical image analysis*. Springer, 2013. 5

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 3

[7] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *PAMI*, 36(8):1532–1545, 2014. 3

[8] P. Dollár and C. L. Zitnick. Fast edge detection using structured forests. *PAMI*, 2015. 1, 2, 5, 7, 8

[9] N. Dorfman, D. Harari, and S. Ullman. Learning to perceive coherent objects. In *CogSci*, pages 394–399, 2013. 3

[10] I. Endres and D. Hoiem. Category-independent object proposals with diverse ranking. *PAMI*, 36(2):222–234, 2014. 2

[11] J. L. Gallant, C. E. Connor, S. Rakshit, J. W. Lewis, and D. C. Van Essen. Neural responses to polar, hyperbolic, and cartesian gratings in area v4 of the macaque monkey. *J. Neurophysiology*, 76(4):2718–2739, 1996. 4

[12] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006. 5

[13] T. Ghose and S. E. Palmer. Extremal edges versus other principles of figure-ground organization. *Journal of Vision*, 10(8):3, 2010. 3

[14] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *CVPR*, pages 564–571, 2013. 6

[15] T. K. Ho. Random decision forests. In *ICDAR*, pages 278–282, 1995. 5

[16] D. Hoiem, A. A. Efros, and M. Hebert. Recovering occlusion boundaries from an image. *Int'l J. Computer Vision*, 91(3):328–346, 2011. 2

[17] P. S. Huggins, H. F. Chen, P. N. Belhumeur, and S. W. Zucker. Finding folds: On the appearance and identification of occlusion. In *CVPR*, pages 718–725, 2001. 3

[18] P. S. Huggins and S. W. Zucker. Representing edge models via local principal component analysis. In *ECCV*, pages 384–398. 2002. 2

[19] W. Kanizsa and W. Gerbino. Convexity and symmetry in figure-ground organization. *Vision and artifact*, pages 25–32, 1976. 1

[20] J. J. Koenderink and A. J. Van Doorn. The singularities of the visual mapping. *Biological cybernetics*, 24(1):51–59, 1976. 2

[21] P. Kontschieder, S. R. Bulo, H. Bischof, and M. Pelillo. Structured class-labels in random forests for semantic image labelling. In *ICCV*, pages 2190–2197, 2011. 5

[22] I. Leichter and M. Lindenbaum. Boundary ownership by lifting to 2.1 d. In *ICCV*, pages 9–16, 2009. 2, 3, 6, 7

[23] J. J. Lim, C. L. Zitnick, and P. Dollár. Sketch tokens: A learned mid-level representation for contour and object detection. In *CVPR*, pages 3158–3165, 2013. 1, 2, 3

[24] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *PAMI*, 26(5):530–549, 2004. 1, 2, 6

[25] M. Nishigaki, C. Fermüller, and D. DeMenthon. The image torque operator: A new tool for mid-level vision. In *CVPR*, pages 502–509, 2012. 2, 4

[26] S. E. Palmer. *Vision Science: Photons to phenomenology*, volume 1. MIT press Cambridge, MA, 1999. 3

[27] S. E. Palmer and J. L. Brooks. Edge-region grouping in figure-ground organization and depth perception. *J. Exp Psychol: Hum Percep Perform*, 34(6):1353–1371, 2008. 4

[28] S. E. Palmer and T. Ghose. Extremal edge: A powerful cue to depth perception and figure-ground organization. *Psychological Science*, 19(1):77–83, 2008. 1, 3

[29] S. Ramenahalli, S. Mihalas, and E. Niebur. Extremal edges: Evidence in natural images. In *Conf. on Information Sciences and Systems*, pages 1–5, 2011. 2, 3, 7

[30] X. Ren, C. C. Fowlkes, and J. Malik. Figure/ground assignment in natural images. In *ECCV*, pages 614–627. 2006. 2, 3, 6, 7

[31] L. Roberts. *Machine perception of 3-d solids*. PhD thesis, MIT, 1965. 2

[32] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. on Graphics*, 23(3):309–314, 2004. 1

[33] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 31(5):824–840, 2009. 2

[34] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760, 2012. 1, 2, 6

[35] A. N. Stein and M. Hebert. Occlusion boundaries from motion: Low-level detection and mid-level reasoning. *Int'l J. Computer Vision*, 82(3):325–357, 2009. 2

[36] R. von der Heydt, T. Macuda, and F. T. Qiu. Border-ownership-dependent tilt aftereffect. *J. Opt. Soc. Am. A*, 22(10):2222–2229, Oct 2005. 1

[37] Y. Xu, Y. Quan, Z. Zhang, H. Ji, C. Fermüller, M. Nishigaki, and D. Dementhon. Contour-based recognition. In *CVPR*, pages 3402–3409, 2012. 2

[38] H. Zhou, H. S. Friedman, and R. Von Der Heydt. Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience*, 20(17):6594–6611, 2000. 1

[39] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, pages 391–405. 2014. 2