Object Detection via Boundary Structure Segmentation

Alexander Toshev, Ben Taskar, and Kostas Daniilidis
GRASP Laboratory, University of Pennsylvania
Philadelphia, PA 19104, USA
{toshev, taskar, kostas}@cis.upenn.edu

Abstract

We address the problem of object detection and segmentation using holistic properties of object shape. Global shape representations are highly susceptible to clutter inevitably present in realistic images, and can be robustly recognized only using a precise segmentation of the object. To this end, we propose a figure/ground segmentation method for extraction of image regions that resemble the global properties of a model boundary structure and are perceptually salient. Our shape representation, called the chordiogram, is based on geometric relationships of object boundary edges, while the perceptual saliency cues we use favor coherent regions distinct from the background. We formulate the segmentation problem as an integer quadratic program and use a semidefinite programming relaxation to solve it. Obtained solutions provide the segmentation of an object as well as a detection score used for object recognition. Our single-step approach improves over state of the art methods on several object detection and segmentation benchmarks.

1. Introduction

In the past decade a multitude of different object representations have been explored, ranging from texture and local features to region descriptors and object shape. Although local features based on image gradients and texture perform relatively well for some object classes, many classes are not modeled sufficiently by local descriptors. For objects with distinctive shape local texture features provide weak description. In this paper we focus on the problem of exploiting global shape properties for object detection and relating those properties to object segmentation.

Shape is commonly defined in terms of the set of contours that describe the boundary of an object. Complementary to gradient- and texture-based representations, shape is more descriptive at a larger scale, ideally capturing the object of interest as a whole. Hence, a large number of global representations, such as curvature scale space, Fourier contour descriptors, Zernicke moments, etc., have been studied [25]. Unfortunately, such descriptions are very susceptible to clutter, due to spurious internal or background contours, and cannot be easily applied to real scenes. For this reason, a variety of local or semi-local descriptors have been studied, such as Shape Context [1] or PAS [5], which capture the shape of only a part of an object outline and are often integrated in an additional global description.

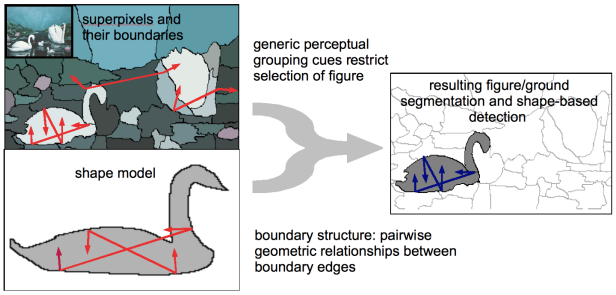

superpixels and their boundaries
generic perceptual grouping cues restrict selection of figure
resulting figure/ground segmentation and shape-based detection
shape model
boundary structure: pairwise geometric relationships between boundary edges

Figure 1. Boundary Structure Segmentation: Holistic shape matching in highly cluttered images with simultaneous object segmentation.

In this work we adhere to the Gestalt school's view that shape is perceived not simply as a collection of parts and propose a recognition method based on a *holistic* shape-boundary based representation. To apply a boundary-based representation in cluttered images, precise figure/ground segmentation is necessary to select the object boundaries for the computation of the shape descriptor. However, accurate automatic segmentation of the object from realistic clutter is often extremely difficult without *familiarity* of the target shape [18]. Evidence from human perception [19] suggests that familiarity plays a large role in figure/ground assignment. We propose the **Bo**undary **S**tructure **S**egmentation (BoSS) model, which addresses the problem of recognition and segmentation *simultaneously* in a *unified framework*. While matching the image with an object model, our method selects a set of foreground regions such that (i) their *global shape as a whole*, defined in terms of their boundary structure, resembles the shape of the object model, and (ii) the foreground represents a *coherent region* distinct from the background (see Fig. 1). The main contributions of our approach are threefold:

Shape description. We introduce a *global boundary-based shape representation*, called *chordiogram*, which is defined as the distribution of all geometric relationships (relative location and normals) between pairs of boundary edges – called chords – whose normals relate to the segmentation interior. This representation captures the *boundary structure* of a segmentation as well as the position of the *interior* relative to the boundary. Moreover, the chordiogram is translation invariant and robust to shape deformations.

Figure/ground Segmentation. We match the above boundary structure while *simultaneously* extracting figure/ground segmentation. This is possible due to the definition of the chordiogram, which relates the object boundaries to its interior. The perceptual grouping component of the segmentation model, which is defined in terms of configural cues of salient contours, color and texture coherence, and small perimeter prior, ensures that the detections constitute salient regions. More importantly, the joint matching and segmentation removes the irrelevant image contours during matching and allows us to obtain correct object detections and segmentation in highly cluttered images.

Inference. We pose BoSS in terms of selection of superpixels obtained via an initial over-segmentation, which is a hard combinatorial problem. We propose a concise formulation as an integer quadratic program, consisting of two terms – a boundary structure matching term defined over superpixel boundaries, and a perceptual grouping term defined over superpixels. The terms are coupled via linear constraints relating the superpixels with their boundary. The resulting optimization problem is solved using a Semidefinte Programming relaxation and yields shape similarity and figure/ground segmentation *in a single step*.

We achieve state-of-the-art results on two challenging object detection tasks – 94.3% detection rate at 0.3 fppi on ETHZ Shape Dataset [8] and 92.4% detection rate at 1.0 fppi on INRIA horses [6] as well as accurate object boundaries, evaluated on the former dataset.

2. Related Work

Due to the large volume of literature on recognition and segmentation, we review approaches closest to our work. Global shape descriptors, such as Fourier contour descriptors, Zernicke moments, Curvature Scale Space, etc. [25] have a long tradition in shape retrieval. However, they are applicable only for already segmented objects and cannot deal robustly with clutter. Semi-local shape descriptors have been proposed to address this limitation. Belongie et al. [1] introduce shape context as a histogram of contour edges, capturing parts of an object. To perform recognition with shape context one needs to integrate it in a global matching framework such as thin plate spline or voting, for example. To alleviate further the issues arising from clutter, Zhu et al. [26] select relevant object contours while matching shape contexts. Boundary fragments combined with a classifier and subsequent voting for object centers have been explored as well [17], [21]. These approaches are part-based and do not use global descriptors. Moreover, all of the above methods recover a set of object contours, but not the figure/ground organization of the image.

A different approach to shape-based recognition is to search for a set of image contours which best matches to a model. Ferrari et al. [8] search in a contour network for contour chains which resemble the model. In a subsequent work Ferrari et al. [5] define a descriptor for groups of adjacent contour segments and use it in conjuction with an SVM classifier. Lu et al. [14] explore particle filtering to search for a set of object contours. Felzenswalb and Schwarz [4] propose a hierarchical representation by decomposing a contour into a tree of subcontours and using dynamic programming to perform matching. Dynamic programming has been applied also by Ravishankar et al. [20] in a mutli-stage framework to search for a chain of object contours. All of the above approaches have to deal with a combinatorial search among image contours and have to decompose their inference into tractable subproblems, thus losing some of the global relationships between contours. On the contrary, we retain in our descriptor all relations between object boundaries to achieve a holistic representation. Although the above approaches recover some object contours, none of them recover full figure/ground organization.

Close interplay between segmentation and recognition has been studied by [23] who guide segmentation using part detections, but do not use global shape descriptors. Segment shape descriptors have been used by [10] for detection and segmentation. Leibe et al. [13] combine recognition and segmentation in a probabilistic framework. Recently, Gu et al. [11] use global shape features on image segments. However, segmentation is a preprocessing step, decoupled from the subsequent matching.

3. Boundary Structure Segmentation Model

For a given target object mask and image, the BoSS model extracts a region in the image such that: (i) it constitutes a perceptually salient figure/ground organization of the image and (ii) resembles the model in shape. In addition, BoSS provides a detection score for the particular object model. To define the BoSSmodel, we denote by $s \in \mathbb{R}^N$ a figure/ground segment indicator vector for an image partitioned into N superpixels: $s_i = 1$ if superpixel i belongs to the figure; -1 otherwise. We define our model over superpixels since this provides computational advantages, however it can be defined in the same way over pixels. We decompose the model into matching and perceptual grouping

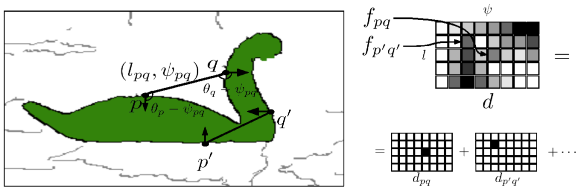


Figure 2. Left: Example of a configuration feature f_{pq} (see Sec. 3.1); Right: A chordiogram d of the figure segmentation (we plot only the length l and orientation ψ dimensions of the descriptor). d can be decomposed as the sum of the descriptors of individual chords (bottom right).

terms:

$$E_{BoSS}(s) = match(s, m) + group(s) \quad (1)$$

In the following, we describe our shape representation and the terms of the model.

3.1. Chordiograms as Shape Representation

To evaluate the similarity between a figure/ground segmentation and the model mask we use a *global boundary-based* shape descriptor, called the *chordiogram*. It is inspired by the Gestalt principle postulating that shape is perceived as whole [18], as well as by the success of contour-based shape descriptors [1].

To define a chordiogram, consider all possible pairs of boundary edges of a segmented object, called chords. Each chord captures the geometric configuration of two boundary edges, and their distribution can be used to describe the global shape. More precisely, for each chord (p, q), its configuration is described as: the length l_{pq} and the orientation ψ_{pq} of the vector connecting p and q as well as the orientations θ_p and θ_q of the normals to the segmentation boundary at p and q (see Fig. 2, left). The latter orientations are defined such that they point towards the object interior. Note that in this way we capture not only the boundary but also the object interior. Thus, the *configuration features* of a chord (p, q) can be written as: $f_{pq} = (\theta_p - \psi_{pq}, \theta_q - \psi_{pq}, l_{pq}, \psi_{pq})^T$, where the normal orientations are w. r. t. ψ_{pq}. We describe the set of all configurations, by defining the chordiogram d as a K-dimensional histogram of the above features for all chords:

$$d_k = \#\{(p, q) | f_{p,q} \in bin(k)\} \quad k = 1 \dots K \quad (2)$$

The lengths l_{pq} are binned together in a log space, which allows for larger shape deformation between points lying further apart, while all the angles are binned uniformly.

In terms of the definition of the pair configurations, the above descriptor is similar to Shape Context [1], which captures the relation of contour edges only to a fixed offset and is not global. The lack of an offset makes our descriptor translation invariant; however, it is not scale or rotation invariant. The descriptor is also inspired by Carlsson [2], which captures topological properties of set of points.

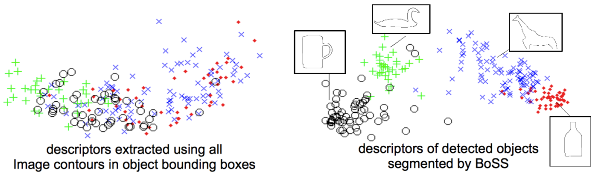


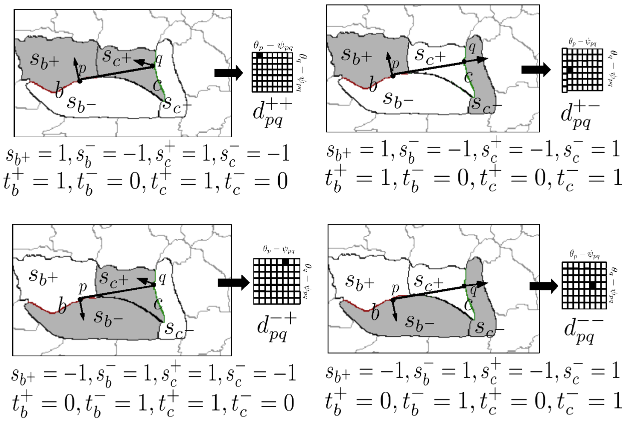
Figure 3. The top 2 principal components of chordiograms computed using PCA for objects in the ETHZ Shape dataset (see Sec. 5). (We omit the class 'Applelogos' for the sake of cleaner illustration).

Another important difference is that we capture the contour orientation relative to object interior. Orienting the boundary normals with respect to the interior contributes to better discrimination, for example, between concave and convex structures (configurations f_{pq} and $f_{p'q'}$ respectively in Fig. 2), which otherwise would be indistinguishable. The discriminative power can be seen on the right side of Fig. 3, where objects of four different types are well separated using chordiograms, provided we compute it on segmented objects. If, however, we use all image contours inside the object bounding box, we obtain cluttered descriptors (Fig. 3, left), which are much harder to separate. This motivates the coupling of the chordiogram with figure segmentation, as explained next. This coupling allows us to use descriptor support which covers the whole object, thus the descriptor is used globally.

3.2. Boundary Structure Matching

The matching term in Eq. (1) compares the chordiograms of the model and an image segmentation. To formalize the matching model, we need to express the descriptor as a function of the object segmentation s. It will prove useful to provide an equivalent definition to Eq. (2). Suppose the contribution of a chord (p, q) to the descriptor is denoted by a chord descriptors $d_{pq} \in \{0, 1\}^K$: $(d_{pq})_k = 1$ iff $f_{pq} \in bin(k)$. Then Eq. (2) can be expressed as a linear function of the chord contributions: $d = \sum_{p,q} d_{pq}$ (see Fig. 2, right). Hence, if we can express the selection of chord descriptors as a function of s, then we can express the chordiogram in terms of s. The main difficulty in the selection of chords lies, as we can see in Fig. 4, in the fact that each chord can result in four different configuration features depending on the position of the object interior with respect to the chord edges: each edge has two possible normal orientations depending on the object interior.

To relate this phenomenon to the figure/ground segmentation, we express the descriptor in terms of a selection of segment boundaries, which are related to the figure in the image by assigning the boundaries to the segments comprising the figure. This is motivated by the idea of figure/ground organization of the image, where the figure is defined as re-



$s_{b^+} = 1, s_b^- = -1, s_c^+ = 1, s_c^- = -1$
$t_b^+ = 1, t_b^- = 0, t_c^+ = 1, t_c^- = 0$

$s_{b^+} = 1, s_b^- = -1, s_c^+ = -1, s_c^- = 1$
$t_b^+ = 1, t_b^- = 0, t_c^+ = 0, t_c^- = 1$

$s_{b^+} = -1, s_b^- = 1, s_c^+ = 1, s_c^- = -1$
$t_b^+ = 0, t_b^- = 1, t_c^+ = 1, t_c^- = 0$

$s_{b^+} = -1, s_b^- = 1, s_c^+ = -1, s_c^- = 1$
$t_b^+ = 0, t_b^- = 1, t_c^+ = 0, t_c^- = 1$

Figure 4. Suppose, b is the common boundary between superpixels s_{b^+} and s_{b^-}; c is the boundary between s_{c^+} and s_{c^-}. If b and c are selected as object boundaries, there are four possible selections of the neighboring superpixels and thus four possible configurations of the chord (p, q). The selection s can be equivalently represented in terms of the indicator variables t of the boundary segments b and c, as shown under the diagrams for each case.

gions which 'own' their boundary [18]. More precisely, we consider a set $\mathcal{B}$ of potential object boundaries, $b \in \mathcal{B}$ being a common boundary between two neighboring superpixels b^+ and b^-. Further, if b is selected as part of the object boundary, then b is 'owned' by the one neighboring superpixel (b^+ or b^-), which lies in the object interior. This can be expressed using auxiliary variables $t_b^k \in [0, 1]$, $k \in \{+, -\}$ (see Fig. 4):

$$t_b^+ = \begin{cases} 1 & s_{b^+} = 1 \text{ and} \\ & s_{b^-} = -1 \\ 0 & \text{otherwise} \end{cases} \quad t_b^- = \begin{cases} 1 & s_{b^+} = -1 \text{ and} \\ & s_{b^-} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The first case for both variables corresponds to b^+ being part of the object and b^- part of the background; the second is the opposite case. Then we can differentiate the aforementioned four cases for a configuration of a chord (p, q). Suppose, p and q lie on boundary segments b and c respectively. Then the chord descriptor of (p, q) resulting by selecting b^k and c^l as foreground is denoted by d_{pq}^{kl}, $k, l \in \{+, -\}$. This allows us to express the global chordiogram in terms of the chordiograms of the descriptors of individual chords based on selected boundaries t:

$$d(t) = \sum_{b,c \in \mathcal{B}} \sum_{p \in c, q \in b} \sum_{k,l \in \{+,-\}} d_{pq}^{kl} t_b^k t_c^l \tag{4}$$

After we have parameterized the chordiogram, we chose to compare it with the model using L_1 distance:

$$match(t, m) = ||d^m - d(t)||_1 \tag{5}$$

subject to the constraints (3) and $t = \begin{pmatrix} t^+ \\ t^- \end{pmatrix} \in \{0, 1\}^{2N}$.

3.3. Perceptual Grouping

Optimizing the matching term in Eq. (5) will result in a figure, defined by s and boundaries t, of maximal shape similarity to the model. However, we need to assure that s represents a *perceptually salient segmentation*, i. e. the resulting figure should be a coherent region distinct from the background. If we denote by $w_{e,g}$ the similarity between the appearance of superpixels e and g, then we can express the above condition by the standard graph-cut score:

$$-s^T W s = -1^T W 1 + 2 \sum_{\substack{e \in \text{figure} \\ g \in \text{ground}}} w_{e,g} \tag{6}$$

where the first term is constant. We also expect that the *most selected superpixel boundaries are supported by edge response in the image*, i. e. they are are not hallucinated. For a boundary segment b, we denote by c_b the percent of the pixels of b *not covered by image edges* extracted using thresholded Pb [16]. Then the boundary cost is defined as

$$c^T(t^+ + t^-) = \sum_{b \in \mathcal{B}} c_b t_b^+ + \sum_{b \in \mathcal{B}} c_b t_b^- \tag{7}$$

Finally, we combine both costs:

$$group(s, t) = -\beta s^T W s + \gamma c^T (t^+ + t^-) \tag{8}$$

for $s \in \{-1, 1\}^N$ and $t^+, t^- \in \{0, 1\}^N$.

The total cost minimized by the BoSS model combines costs from Eq. (5) and Eq. (8)

$$\min_{s,t} ||d^m - d(t)||_1 - \beta s^T W s + \gamma c^T (t^+ + t^-) \tag{9}$$

$$\text{s. t.} \quad t_b^+ - t_b^- = 1/2(s_{b^+} - s_{b^-}) \quad \forall b \in \mathcal{B} \tag{10}$$

$$t_b^+ t_b^- = 0 \quad \forall b \in \mathcal{B} \tag{11}$$

$$s \in \{-1, 1\}^N, \quad t_b^+, t_b^- \in \{0, 1\} \tag{12}$$

Constraints (10) and (11) are equivalent to constraints (3), which can be easily verified for all four possible integer values of the variables t_b^+, t_b^- and variables s_{b^+}, s_{b^-}.

In summary, the matching cost operates on the boundary indicators t, while the grouping cost is expressed in terms of superpixel indicators s. Both costs are made consistent via coupling constraints, which ensure that the resulting figure segmentation resembles in shape the given model and represents meaningful grouping in the image.

Example We examine the contribution of each term of the model on one concrete example presented in Fig. 5. By using only the matching term we are able to localize the object and obtain a rough mask, which however extends the back of the horse and ignores its legs (first column). The inclusion of the superpixel grouping bias helps to remove some of the erroneous superpixels above the object which have a different color than the horse (second column). Finally, if we add the boundary term, it serves as a sparsity regularization on t and results in a tighter segmentation (third column). Thus, the incorrect superpixels above the horse get removed, since they contain hallucinated boundaries not supported by edge response. Additionally, it recovers some of the legs, since they exibit strong edge response along their boundary.

4. Optimization via Semidefinite Program

The optimization of the integer quadratic program in Eq. (9) is NP-hard. We chose Semidefinite Programing (SDP) to obtain relaxed solutions. For this purpose, we introduce two variables, which bring the quadratic terms of Eq. (9) into linear form: $T = tt^T, S = ss^T$. This allows us to state the SDP relaxation as follows:

$$\begin{aligned} \min_{S,T,s,t} \; & \|d^m - d(T)\|_1 - \beta tr(W^T S) + \gamma c^T(t^+ + t^-) \\ \text{s. t.} \quad & t_b - t_{m+b} = 1/2(s_{b^+} - s_{b^-}) \quad \forall b \in \mathcal{B} && (13) \\ & T_{b,m+b} = 0 \quad \forall b \in \mathcal{B} && (14) \\ & diag(S) = 1_N && (15) \\ & t_b = T_{b,b}, t_{m+b} = T_{m+b,m+b} \forall b \in \mathcal{B} && (16) \\ & \begin{pmatrix} T & t \\ t^T & 1 \end{pmatrix} \succeq 0, \begin{pmatrix} S & s \\ s^T & 1 \end{pmatrix} \succeq 0 && (17) \end{aligned}$$

where N is the number of superpixels and $m = |\mathcal{B}|$ the number of boundaries.

The above problem was obtained from problem (9) in two steps. First, we relax the constraints $T = tt^T$ and $S = ss^T$ to $T \succeq tt^T$ and $S \succeq ss^T$ respectively, which by Schur complement are equivalent to (17). Second, we weakly enforce the domain of the variables from the constraint (12). The $-1/1$-integer constraint on s is expressed as diagonal equality constraint on the relaxed S (see Eq. 15), which can be interpreted as bounding the squared value of the elements of s by 1. The $0/1$-integer constraint (see Eq. (16)) is enforced by requiring that the diagonal and the first row of T have the same value. Since $T = tt^T$, this has the meaning that the elements of t are equal to their squared values, which is true only if they are 0 or 1. Finally, the coupling constraints (10) and (11), one of which is quadratic, naturally translate to linear constraints (13) and (14).

Discretization Discrete solutions are obtained by thresholding s. Since s has N elements, there are at most N different discretizations, all of which are ranked using their distance to the model. If a threshold results in a set of several disconnected regions, we consider all possible subsets of this set. The algorithm outputs the top 5 ranked non-overlapping masks. Note that we are capable of detecting several instances of an object class since they result in several disconnected regions which are evaluated independently.

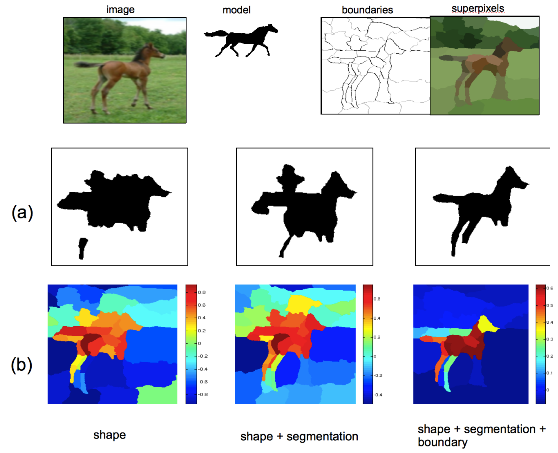


Figure 5. (a) For an input image and model, as shown in the first row, our algorithm computes an object segmentation displayed in (a) row. We present three solutions by using only the matching term from Eq. (5) in first column; the matching term together with the superpixel segmentation prior (first cost in Eq. 8) in second column; and the whole cost function consisting of the matching, segmentation and the boundary term in third column. (b) We also show for the three cost combinations the relaxed values of the segmentation variable s.

Implementation Details We use chordiograms with 4 log bins for the distance feature with the largest bin equal to the diameter of the model. For all angles we use 8 equally spaced bins, resulting in 2048-dimensional descriptor.

To obtain superpixels we oversegment the image using NCuts [3] with $N = 45$ segments. The grouping cues used to define the affinity matrix W^{pixels} are color and intervening contours [24] based on Pb [16]. To define the segmentation term (8) in our model we can use any affinity matrix. We choose to use the same grouping cues as for segmentation above. For each pair of superpixels k and l we average the pixel affinities to obtain an affinity matrix over the superpixels: $W_{kl}^{superpixels} = \frac{1}{a_k a_l} \sum_{p \in k, q \in l} \widehat{W}_{pq}^{pixels}$, a_k and a_l being the number of pixels contained in k and l respectively. Above, $\widehat{W}^{pixels}$ is obtained from the top N eigenvectors E of W^{pixels}: $\widehat{W}^{pixels} = E \Lambda E^T \approx W^{pixels}$, where Λ are the corresponding eigenvalues. This low-rank approximation represents a smoothed version of the original matrix and reduces the noise in the original affinities. Finally, the weights of the term in Eq. (9) were chosen to be $\beta = 0.01$ and $\gamma = 0.6$ on five images from ETHZ dataset and held constant for all experiments.

For the optimization we use SeDuMi [22]. To compute the number of variables in the SDP, one can assume that each superpixel has at most C neighboring superpixels. Hence we obtain $m = CN$ boundary variables. The total variable number in the relaxed problem is bounded by $N^2 + C^2N^2 \in O(N^2)$. In our experiments, we have $N = 45$ and the value of C is less than 5 which results in less than 200 boundary segment variables. The segmentation of an image takes $5 - 15$ secs on a 3.50 GHz processor.

5. Experiments

Detection In this section we present object detection results on two datasets. The ETHZ Shape Dataset [7] consists of 255 images of 5 different object classes. The images are highly cluttered – in the background as well as internal spurious contours – and the objects vary in scale. The second dataset, INRIA horses, has 340 images, half of which contain horses. This dataset presents challenges not only in terms of clutter and scale variation, but also in articulation, since the horses are in different poses.

We apply BoSS on both datasets with same parameters (see sec. 4). We use hand-drawn object outlines as shape models. In particular, we use one model per class for the ETHZ Shape Dataset and 6 horse models representing different poses for the INRIA horse dataset (see Fig. 7 and 9). For each image and model we run BoSS over several scales[1] to produce detection and segmentation hypotheses and score them based on the output of the matching from eq. (5). We use non-maximum suppression – for every two hypotheses, whose bounding boxes overlap by more than 50%, we retain the one with the higher score and discard the other one.

On the ETHZ Shape Dataset we achieve 89.2%/90.5% detection rate at 0.3/0.4 fppi using Pascal criterion[2] and 93.4%/94.2% under 20% overlap criterion[2], as reported in Table 1 and Fig. 6. As shown in Fig. 7, our method is capable of detecting objects of various scales in highly cluttered images, even under occlusion (image 1, 6), as well as multiple instances (images 8,12). The major sources for incorrect detections are accidental alignments with background contours (image 16) and partially incorrect boundaries (in image 15 the mug is correctly detected, but glued to a background segment).

On INRIA Horses dataset, we achieve state of the art detection rate of 92.4% at 1.0 fppi (see Fig. 8). Examples of detections of horses in different poses, scales and in cluttered images are shown in Fig. 9.

Reranking In order to compare with approaches on ETHZ Shape dataset which use supervision, we use weakly labeled data to rerank the detections obtained from BoSS. We use only the labels of the training images to train a classifier but not the bounding boxes. This classifier can be used to rerank new hypotheses obtained from BoSS.

More precisely, we use half of the dataset as training and the other half as test (we use 5 random splits). We use BoSS to mine for positive and negative examples. The top detection in a training image using a model which represents the label of that image is considered a positive example; all other detections are negative examples. The chordiograms of these examples are used as features to train one-vs-all SVM [12] for each class. During test time, each detection is scored using the output of the SVM corresponding to the model used to obtain this detection. Note that this is a different setup of supervision which requires less labeling – while we need one hand-drawn model per class to obtain detections via BoSS, we do not use the bounding boxes but only the labels of the training images to score them. We argue that the effort to obtain a model is constant while segmenting images by hand is much more time consuming.

The results are shown in Table 1. The weak supervision leads to 94.3%/96.0% detection rate under Pascal criterion, which is an improvement of approx. 5% over BoSS. It is attributed to the discriminatively learned weights of the chordiogram's bins. This corresponds to discriminatively learning object shape variations and builds on the power of BoSS to deal with clutter.

Segmentation In addition to the detection results, we evaluate the quality of the detected object boundaries and object masks. For evaluation of the former we follow the test settings of [7][3]. We report recall and precision of the detected boundaries in correctly detected images in Table 2. We achieve higher recall at higher precision compared to [7]. This is mainly result of the fact that BoSS attempts to recover a closed contour and in this way the complete object boundary. These statistics show that the combination of shape matching and figure/ground organization results in precise boundaries ($> 87\%$ for all classes except Giraffes). The slightly lower results for Giraffes is due to the legs which are not fully captured in the provided class models. We also provide object mask evaluation as percentage of the image pixels classified incorrectly by the detected mask (see Table 2). For all classes we achieve less than 6% error, and especially classes with small shape variation such as Bottles and Applelogos we have precise masks ($< 3\%$ error).

6. Conclusion

We introduce a model for joint object segmentation and detection. It is based on a global boundary-based shape descriptor, the chordiogram, which captures the boundary structure and relates it to the interior of an object. This allows us to combine the shape matching with figure segmentation and thus to deal with highly cluttered images. The model, solved using single-step optimization, achieves state of the art results on two detection benchmarks.

[1] For ETHZ Shape dataset we use 7 different scales, such that the scale of the model, defined as the diameter of its bounding box, range from 100 to 300 pixels. Similarly, for INRIA Horse dataset we used 10 scales ranging from 55 to 450 pixels.

[2] Pascal criterion: the intersection of the hypothesis and ground truth bounding boxes overlap more than 50% with the union of both; 20% overlap detection criterion: the intersection of the hypothesis and ground truth bounding boxes overlap more than 20% with the each of them.

[3] A detected boundary point is considered a true positive if it lies within t pixels of a ground truth boundary point, where t is set to 4% of the diagonal of the ground truth mask. Based on this definition, one computes recall and precision.

	Algorithm	Apple logos	Bottles	Giraffes	Mugs	Swans	Average
20% overlap	BoSS[†]	**95.5%/95.5%**	**96.4%/96.4%**	**93.4%/95.6%**	**84.8%**/86.4%	**97.0%/97.0%**	**93.4%/94.2%**
	Lu et. al [14][†♯]	92.5%/92.5%	95.8%/95.8%	86.2%/92.0%	83.3%/92.0%	93.8%/93.8%	90.3%/93.2%
	Fritz et. al [9][*]	-/89.9%	-/76.8%	-/90.5%	-/82.7%	-/84.0%	-/84.8%
	Ferrari et. al [6], [7][†]	84.1%/86.4%	90.9%/92.7%	65.6%/70.3%	80.3%/83.4%	90.9%/93.9%	82.4%/85.3%
Pascal criterion	BoSS[†]	95.5%/95.5%	**96.4%**/96.4%	81.3%/84.6%	75.8%/78.8%	97.0%/97.0%	89.2%/90.5%
	BoSS + reranking[*]	**100%/100%**	96.3%/**97.1%**	86.1%/91.7%	90.1%/91.5%	**98.8%/100%**	**94.3%/96.0%**
	Maji et. al [15][*]	95.0%/95.0%	92.9%/96.4%	89.6%/89.6%	93.6%/**96.7%**	88.2%/88.2%	91.9%/93.2%
	Gu et. al [11][*]	90.6%/-	94.8%/-	79.8%/-	83.2%/-	86.8%/-	87.1%/-
	Ravishankar et. al [20][†○]	95.5%/97.7%	90.9%/92.7%	**91.2%/93.4%**	**93.7%**/95.3%	93.9%/96.9%	93.0%/95.2%

Table 1. Detection rates at 0.3/0.4 false positives per image, using the 20% overlap and Pascal criteria. We achieve state of the art results on all categories under the first detection criterion. Under the Pascal criterion, we achieve state of the art rates on the dataset as well. For Applelogos, Swans and Bottles, the results are equal to the ones using the weaker criterion. This is due to the exact localization, which can be achieved when segmenting the object. For Giraffes and Mugs results are slightly lower due to imperfect segmentation (some segments leak into the background or miss parts) – the detections which are correct under the weaker 20% overlap criterion, are not counted as correct which Pascal criterion. However, there are correctly segmented objects under the Pascal criterion which are ranked lower. The employed reranking helps to recover some of them. ([†] use only hand labeled models. [*] use strongly labeled training data with bounding boxes, while we use weakly labeled data in the reranking, i. e. no bounding boxes. [♯] considers in the experiments only at most one object per image and does not detect multiple objects per image. [○] uses a slightly weaker detection criterion than Pascal.)

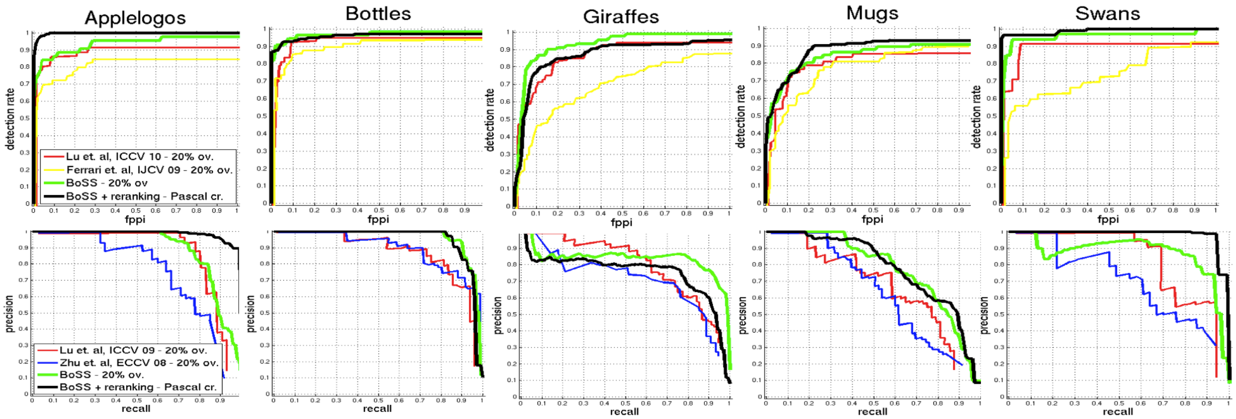


Figure 6. Results on ETHZ Shape dataset. Top: detection rate vs false positives per image; bottom: precision recall curves. Results using BoSS are shown using 20% overlap as well as after reranking using the stricter Pascal criterion. Both consistently outperform other approaches, evaluated using the weaker 20% overlap criterion.

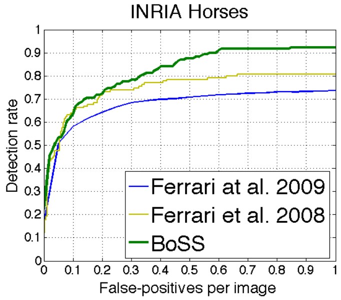


Figure 8. Detection rate vs false positives per image (fppi) for our and other approaches on INRIA Horse dataset.

Method	Det. rate
BoSS	92.4%
[15]	85.3%
[5]	80.8%
[7]	73.8%

	boundary precision/recall		pixel error
	BoSS	Ferrari et. al [7]	BoSS
Applelogos	91.8%/97.5%	91.6%/93.9%	1.6%
Bottles	90.3%/92.5%	83.4%/84.5%	2.7%
Giraffes	76.8%/82.4%	68.5%/77.3%	5.9%
Mugs	86.5%/90.5%	84.4%/77.6%	3.6%
Swans	85.8%/87.6%	77.7%/77.2%	4.9%

Table 2. Precision/recall of the detected object boundaries and pixel classification error of the detected object masks for ETHZ Shape dataset.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4), 2002.

[2] S. Carlsson. Order structure, correspondence and shape based categories. In *International Workshop on Shape, Contour and Grouping*, 1999.

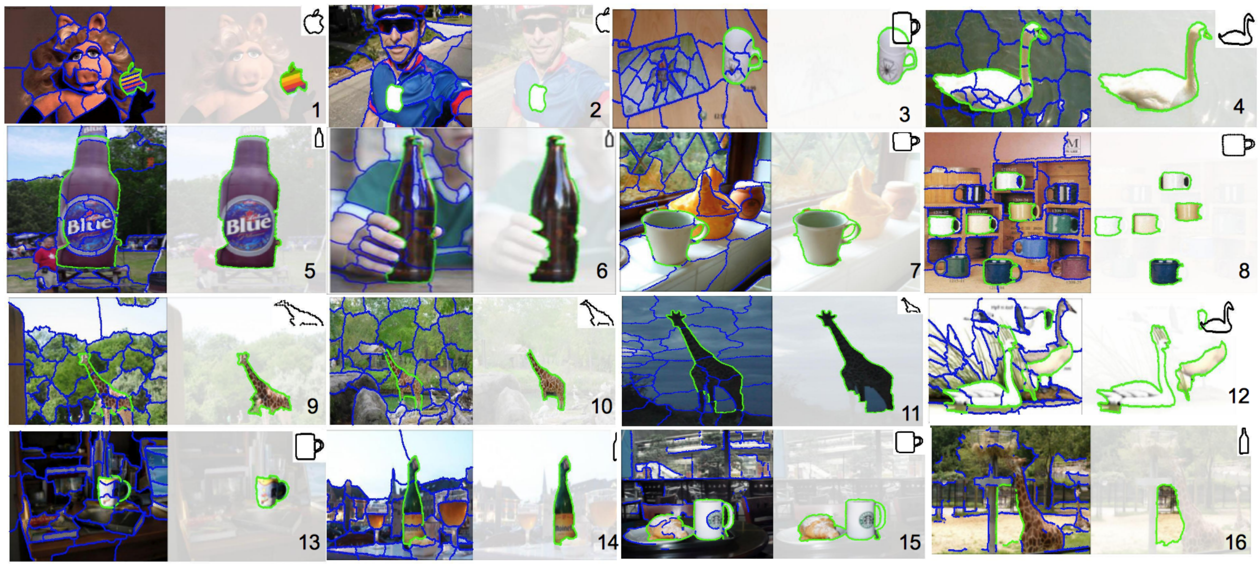


Figure 7. Example detection on ETZ Shape dataset. For each example, we show on the left side the selected superpixel boundaries, and on the right the selected object mask.

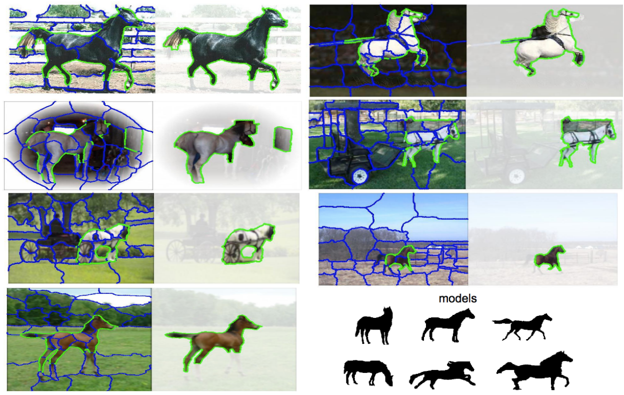


Figure 9. Examples of detections for INRIA horses dataset. For each image we show the selected superpixel boundaries on the left and the detected object segmentation on the right. Bottom right: 6 models used in the experiments.

[3] T. Cour, F. Benezit, and J. Shi. Spectral segmentation with multiscale graph decomposition. In *CVPR*, 2005.

[4] P. Felzenszwalb and J. Schwartz. Hierarchical matching of deformable shapes. In *CVPR*, 2007.

[5] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of adjacent contour segments for object detection. *TPAMI*, 2008.

[6] V. Ferrari, F. Jurie, and C. Schmid. Accurate object detection with deformable shape models learnt from images. In *CVPR*, 2007.

[7] V. Ferrari, F. Jurie, and C. Schmid. From images to shape models for object detection. *IJCV*, 2009.

[8] V. Ferrari, T. Tuytelaars, and L. V. Gool. Object detection by contour segment networks. In *ECCV*, 2006.

[9] M. Fritz and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. In *CVPR*, 2008.

[10] L. Gorelick and R. Basri. Shape based detection and top-down delineation using image segments. *IJCV*, 83(3), 2009.

[11] C. Gu, J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. In *CVPR*, 2009.

[12] T. Joachims. Making large-scale svm learning practical. In *Advances in Kernel Methods - Support Vector Learning*, 1999.

[13] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1-3), 2008.

[14] C. Lu, L. J. Latecki, N. Adluru, X. Yang, and H. Ling. Shape guided contour grouping with particle filters. In *ICCV*, 2009.

[15] S. Maji and J. Malik. Object detection using a max-margin hough transform. In *CVPR*, 2009.

[16] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE TPAMI*, 2004.

[17] A. Opelt, A. Pinz, and A. Zisserman. A boundary-fragment-model for object detection. In *ECCV*, 2006.

[18] S. Palmer. *Vision science: Photons to phenomenology*. 1999.

[19] M. Peterson and B. Gibson. Must Figured-Ground Organization Precede Object Recognition? An Assumption in Peril. *Psychological Science*, 5(5), 1994.

[20] S. Ravishankar, A. Jain, and A. Mittal. Multi-stage contour based detection of deformable objects. In *ECCV*, 2008.

[21] J. Shotton, A. Blake, and R. Chipolla. Contour-based learning for object detection. In *ICCV*, 2005.

[22] J. F. Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 1999.

[23] S. Yu and J. Shi. Object-specic figure-ground segregation. In *CVPR*, 2003.

[24] S. X. Yu and J. Shi. Multiclass spectral clustering. In *ICCV*, 2003.

[25] D. Zhang and G. Lu. Evaluation of mpeg-7 shape descriptors against other shape descriptors. *Multimedia Systems*, 9(1), 2003.

[26] Q. Zhu, L. Wang, Y. Wu, and J. Shi. Contour context selection for object detection: A set-to-set contour matching approach. In *ECCV*, 2008.