Contour Context Selection for Object Detection: A Set-to-Set Contour Matching Approach

Qihui Zhu[1], Liming Wang[2], Yang Wu[3] and Jianbo Shi[1]

[1]Department of Computer and Information Science, University of Pennsylvania
qihuizhu@seas.upenn.edu, jshi@cis.upenn.edu

[2]Department of Computer Science and Engineering, Fudan University
wanglm@fudan.edu.cn

[3]Institue of Artificial Intelligence and Robotics, Xi'an Jiaotong University
ywu@aiar.xjtu.edu.cn

Abstract. We introduce a shape detection framework called *Contour Context Selection* for detecting objects in cluttered images using only one exemplar. Shape based detection is invariant to changes of object appearance, and can reason with geometrical abstraction of the object. Our approach uses salient contours as integral tokens for shape matching. We seek a maximal, holistic matching of shapes, which checks shape features from a large spatial extent, as well as long-range contextual relationships among object parts. This amounts to finding the correct figure/ground contour labeling, and optimal correspondences between control points on/around contours. This removes *accidental alignments* and does not hallucinate objects in background clutter, without negative training examples. We formulate this task as a set-to-set contour matching problem. Naive methods would require searching over 'exponentially' many figure/ground contour labelings. We simplify this task by encoding the shape descriptor algebraically in a linear form of contour figure/ground variables. This allows us to use the reliable optimization technique of Linear Programming. We demonstrate our approach on the challenging task of detecting bottles, swans and other objects in cluttered images.

1 Introduction

We study the problem of object detection in natural images using shape. Visual objects can be represented at a variety of levels from signal (filter responses) to symbol (object parts). Our approach focuses on representation of the shape that is closer to the symbol level, which would allow abstract geometrical reasoning of the object. Shape description is invariant to color, texture, and brightness changes, which could enable significant reduction in the number of training examples, and increase of accuracy of the detection.

Object detection using shape alone is not an easy task. Most shape matching algorithms are susceptible to *accidental alignment*: hallucinating objects in background clutter. To avoid foreground (surface marking) and background clutter, shape descriptors are often computed within a window of limited spatial extent. Local window features are discriminative enough for detecting objects such as faces, cars and bicycles. However, for many simple objects, such as swans, mugs or bottles, local shape features are insufficient.

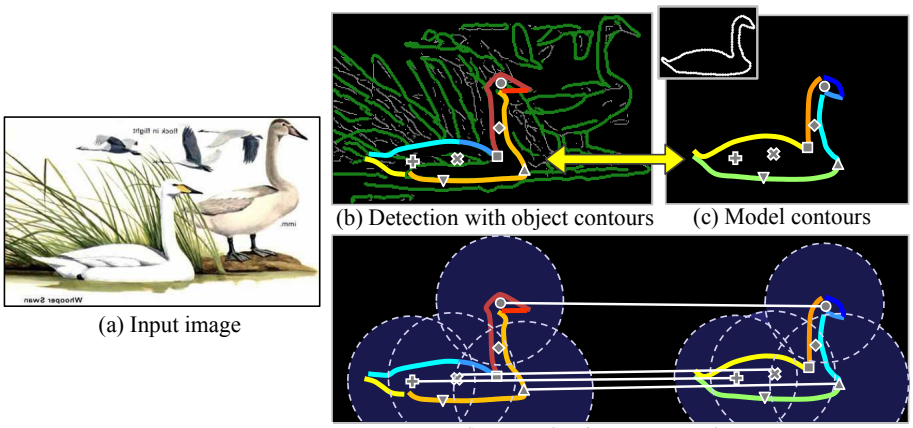
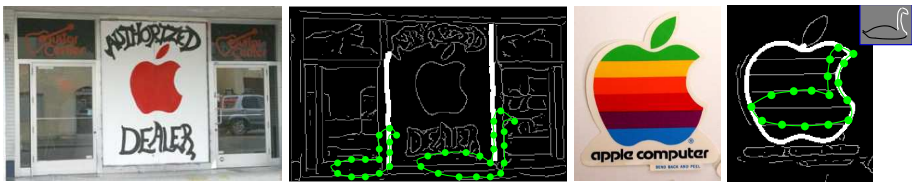


Fig. 1. Using a single line drawing object model shown in (c), we detect object instances in images with background clutter in (a) using only shape. Bottom-up contour grouping provides tokens of shape matching. Long salient contours in (b) provide distinctive shape descriptions, allowing both efficient and accurate matching. Image and model contours, shown by different colors in (b) and (c), do not have one-to-one correspondences. We formulate shape detection as a set-to-set matching task in (d) consisting of: (1) correspondences between control points, and (2) selection of contours that contribute contextual shape features to those control points, within a circular neighborhood.

To overcome this *accidental alignment* problem, we propose a shape detection approach called *Contour Context Selection* consisting of the following the key ingredients:

1. We detect salient contours using bottom-up segmentation or contour grouping. Long contours are more distinctive, and maintaining contours as integral tokens for matching removes many false positives due to accidental alignment.
2. We break the model shape into its informative semantic parts, and explicitly check which subset of model shape parts is matched. Missing critical model parts can signal an accidental alignment between the image and model.
3. We seek holistic shape matching. We measure shape features from a large spatial extent, as well as long-range contextual relationships among object parts. Accidental alignments of holistic shape descriptors between image and model are unlikely.

Our *Contour Context Selection* reduces to finding a maximal, holistic matching between a *set* of image contours and a *set* of model parts. It searches over *figure/ground* labeling of the image and model contours, and *correspondences* between them. It is important to note that, in general, image contours and model contours do not correspond one-to-one. The holistic matching occurs only by considering a set of 'figure' contours together. To formulate this *set-to-set* matching task, we define control points sampled on and around the image and model contours. We compute shape features on the control points from the 'figure' contours within a large neighborhood (see Fig. 1). The task is to find the correct figure/ground contour labeling, such that there is an optimal one-to-one



(a) Accidental alignment (b) Missing critical parts

Fig. 2. Typical false positives can be traced to two causes: (1) Accidental alignment shown in (a). Our algorithm prunes it by exploiting contour integrity, *i.e.* requiring contours to be whole-in/whole-out. Contours violating this constraint is marked in white on the image. (2) Missing critical object parts indicates that the matching is a false positive. In (b), after removing the accidental alignment to the apple logo outline (marked in white), only the body can find possible matches and the neck of the swan is completely missing shown at the top-right corner of (b). Our approach rejects this type of detection by checking missing critical model contours after joint contour selection.

matching of the control points. This set-to-set matching potentially requires searching over exponentially many choices of figure/ground labeling. We simplify this task by encoding the shape descriptor algebraically in a linear form of contour selection variables, allowing to use the reliable optimization technique of Linear Programming.

This paper is organized as follows. Section 2 introduces the basic concept and formulation of Contour Context Selection. We present the computational solution for this framework using Linear Programming (LP) in Section 3. Section 4 describes related works and comparisons. Section 5 demonstrates our approach on the challenging task of detecting non-rectangular and wiry objects, followed by the conclusion in Section 6.

2 Shape Detection as a Set-to-Set Contour Matching Problem

Our goal is to detect objects in images using a *single* model and identify correspondences between the image and the model.

We use salient contours, extracted from bottom-up contour grouping, as tokens for image-model shape matching. Shape matching with contours instead of isolated edges has several advantages. Long salient contours are more distinctive, which leads to efficiency of the search as well as the accuracy of shape matching. Furthermore, by requiring the entire contour to match objects as a whole, we remove accidental alignment causing false postive detections (see Fig. 2 (a) for an example).

Using contour grouping as the starting point of shape matching carries risk as well. Contours could be mis-detected, or accidentally leak to background clutter. A good contour grouping algorithm is essential for shape matching. We utilize the approach in [1] which has demonstrated good performance in cluttered images detecting reliable contours. Furthermore, these contours groups are not disjoint, providing multiple hypotheses at places where contours can potentially leak to other objects (*e.g.* junctions).

To evaluate shape matching, one needs to measure the accuracy of alignment, and more importantly, determine *which* parts are aligned. For simple shapes, missing a small

but critical object part can indicate a complete mismatch, see Fig. 2 (b). In this work, we manually divide the model into contours which corresponds to distinctive parts. Just as image contours, we require model contours to be matched as a whole.

The computational task of shape matching thus consists of parallel searches over image contours and model contours to obtain the maximal match of the image and model shapes. We cast the shape detection as the following problem:

Set-to-set contour matching. Given an image $\mathcal{I}$ and a model $\mathcal{M}$ represented by two sets of long salient contours:

- Image: $\mathcal{I} = \{C_1^I, C_2^I, \dots C_{|\mathcal{I}|}^I\}$, C_k^I is the k^{th} contour;
- Model: $\mathcal{M} = \{C_1^M, C_2^M, \dots, C_{|\mathcal{M}|}^M\}$, C_l^M is the l^{th} contour.

we would like to select the maximal contour subsets $\mathcal{I}^{sel} \subseteq \mathcal{I}$ and $\mathcal{M}^{sel} \subseteq \mathcal{M}$, such that object shapes composed by $\mathcal{I}^{sel}$ and $\mathcal{M}^{sel}$ match (see Fig. 1 for an image example).

Once this set-to-set matching is solved, to quantify shape matching we measure

- *which set of model contours are matched;*
- *how well the matched contours are aligned.*

The final classification cost function combines the following two terms:

$$C_{classification} = C_{config} \cdot C_{align} \tag{1}$$

where C_{config} evaluates the configuration of the matched model contours, and C_{align} measures quality of their alignment defined later in the following sections.

The main technical difficulty is that the image and model contours do not have one-to-one correspondence. Contours detected from bottom-up grouping and segmentation are different from the semantically meaningful contours in the model. However, as a whole they will have matches (see Fig. 1). Set-to-set contour matching bridges this semantic gap between the bottom-up grouping and the top-down model.

2.1 Solution for set-to-set contour matching

Our solution to the set-to-set matching problem includes three essential components:

Control point correspondence. While contour themselves do not correspond in one-to-one, their shape information can be evaluated at nearby *control points*, and those control points could have one-to-one correspondences. Suppose control points $\{p_1, p_2, \dots, p_m\}$ are sampled from the image and $\{q_1, q_2, \dots, q_n\}$ are sampled from the model. We define the correspondence matrix $(U^{cor})_{m \times n}$ from the image to the model as:

$$U_{ij}^{cor} = \begin{cases} 1, & \text{if } p_i \text{ matches } q_j \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Note that these control points can be located anywhere in the image, not limited to contours. Computing dense point correspondences is unnecessary. Instead, rough matching of some control points is sufficient to select and match contour sets $\mathcal{I}^{sel}$ and $\mathcal{M}^{sel}$.

Feature representation: holistic shape features. The important question is, what will be the appropriate shape feature for matching these control points, and how to compute

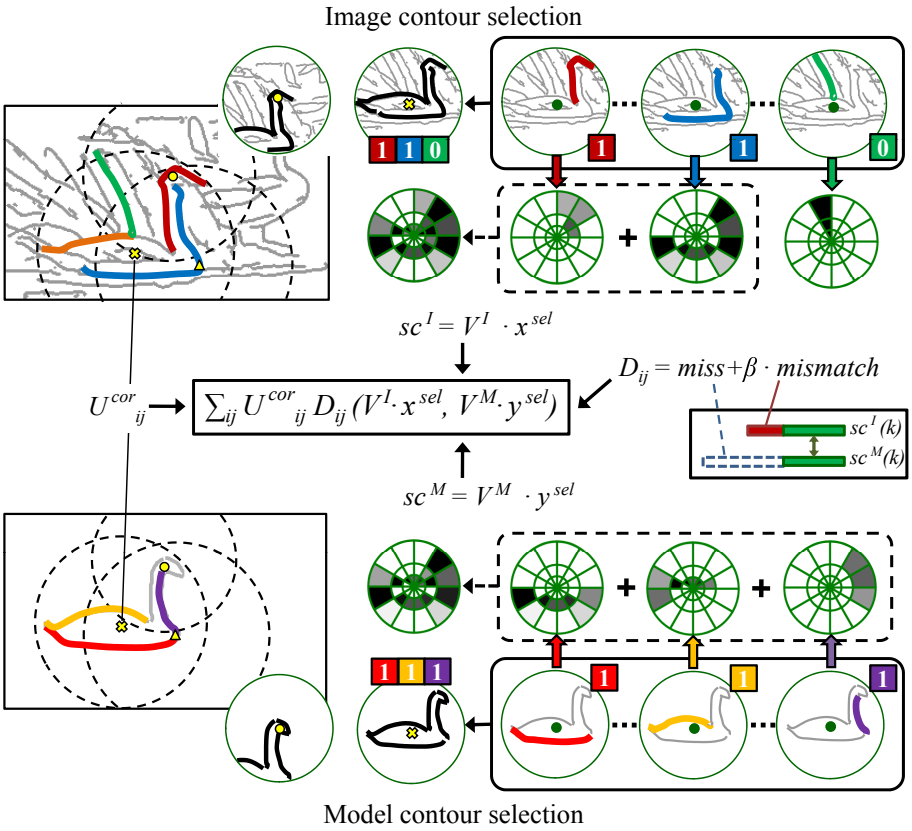


Fig. 3. Illustration of our computational solution for set-to-set contour matching on shape detection example from Fig. 1. The top and the bottom row shows the image and model contour candidate sets marked in gray. Each contour contributes its shape information to nearby *control points* in the form of Shape Context histogram, shown on the right. By selecting different contours (x^{sel}, y^{sel}), each control point can take on a set of possible Shape Context descriptions (sc^I, sc^M). With the correct contour selection in the image and model (marked by colors), there is a one-to-one correspondence U^{cor}_{ij} between (a subset of) image and model control points (marked by symbols). This is a computationally difficult search problem. The efficient algorithm we developed is based on an encoding of Shape Context description (which could take on exponentially many possible values) using linear algebraic formulation on the contour selection indicator: $sc^I = V^I \cdot x^{sel}$. This leads to the Linear Programming optimization solution.

shape dissimilarity D_{ij}. Comparing D_{ij} requires the feature to be common in the image and the model. Since there do not exist one-to-one correspondences between contours, the feature description is more appropriate on the contour set or global shape level rather than on the individual contour level. We propose a holistic shape representation at the control points covering not only nearby contours but also faraway contours (see Fig. 3).

The holistic shape representation immediately poses the problem of *figure/ground selection* since figure/ground segmentation is unknown and the shape feature is likely to include both foreground and background contours. Unknown segmentation introduces great difficulties to any shape features with a *fixed* context. A fixed context feature cannot adapt to the combinatorial possibilities of figure/ground labeling, each generating different contexts. Without the correct segmentation, background clutter and contours from other objects can corrupt the shape feature. Our strategy is to adjust the context of the holistic shape features during matching depending on the figure/ground selection. Therefore, we are able to select the right features and determine the figure/ground segmentation simultaneously.

Matching constraint: contour integrity. Contour selection implies that we restrict each contour to be an integral unit in matching. For each contour $C_k^I = \{p_1^{(k)}, p_2^{(k)}, ..., p_c^{(k)}\}$ where $p_i^{(k)}$'s are edge pixels, there are only two choices: either all the edge pixels $p_i^{(k)}$ participate in the matching, or none of them are included. Partially matched contours are not allowed. The same constraint is applied to model contours C_l^M as well. We introduce contour selection indicators $x^{sel} \in \{0,1\}^{|I|\times 1}$ in the *entire* test image and $y^{sel} \in \{0,1\}^{|M|\times 1}$ in the model defined as

$$(\textbf{Image contour selection}) \qquad x_\ell^{sel} = \begin{cases} 1 & \text{Contour } C_\ell^I \text{ is selected} \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

$$(\textbf{Model contour selection}) \qquad y_\ell^{sel} = \begin{cases} 1 & \text{Contour } C_\ell^M \text{ is selected} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

The constraint of contour integrity makes matching robust to accidental alignment.

2.2 Context Sensitive Shape Features

Now we are ready to introduce the holistic shape representation called *Context Sensitive Shape Features* determined by the figure/ground selection of the contours induced by x^{sel} and y^{sel}. We choose Shape Context (SC) [2] as our basic shape feature descriptor. Measuring global shape requires the scope of SC to be large enough to cover the *entire* object. Define sc_i^I to be the vector of SC histogram centered at control point p_i, *i.e.* $sc_i^I(k) = \#$ of points in bin k. We introduce a contribution matrix V_i^I with size (#bin)×(#contour) to encode the contribution of each contour to each bin of sc_i^I:

$$V_i^I(k,l) = \#\text{ of points in bin } k \text{ from contour } C_l \tag{5}$$

Similar notations sc_j^M and V_j^M are defined for SC at control point q_j in the model.

The key observation is that shape features sc_i^I will be *different* depending on context x^{sel}, *i.e.* they are not fixed. Since each contour can have 2 choices, either selected or not selected, there exists 2^n possible contexts – exponential in the number of contours n. One advantage of histogram type of features such as Shape Context is that the exponentially many combinations of contexts can be written in a simple linear form:

$$sc_i^I(k) = \sum_l V_i^I(k,l) \cdot x_l^{sel} = (V^I \cdot x^{sel})_k \tag{6}$$

This allows us to cast the complex search as an optimization problem later.

Our goal is to find x^{sel} and y^{sel} such that they produce similar shape features: $V_i^I \cdot x^{sel} \approx V_j^M \cdot y^{sel}$. We evaluate and compare these two features by the context sensitive dissimilarity:

$$(\textbf{Context sensitivity}) \qquad D_{ij}(sc_i^I, sc_j^M) = D_{ij}(V_i^I \cdot x^{sel}, V_j^M \cdot y^{sel}) \tag{7}$$

The shape dissimilarity D_{ij} not only depends on the local attributes of p_i and q_j, but more importantly, on the context given by x^{sel} and y^{sel}. Matching object shapes boils down to minimizing D_{ij}, which is a combinatorial search problem on x^{sel} and y^{sel}.

2.3 Contour Context Selection Cost

Finding set-to-set contour matching finally becomes a joint search over correspondences U^{cor} and contour selection x^{sel}, y^{sel} by minimizing the following cost:

$$\min_{U^{cor}, x^{sel}, y^{sel}} C_{align}(U^{cor}, x^{sel}, y^{sel}) = \frac{1}{c} \sum_{i,j} U_{ij}^{cor} D_{ij}(V^I x^{sel}, V^M y^{sel}) \tag{8}$$

$$\text{s.t.} \quad U^{cor} \in \text{GeoSet} \tag{9}$$

where $c = \sum_{i,j} U_{ij}^{cor}$ is the number of control point correspondences. Correspondences U^{cor} from different object parts should have geometric consistency. We use a star model for checking global geometric consistency. Each correspondence (p_i, q_j) can predict an object center c_{ij}. For the correct set of correspondences, all the predicted centers should form a cluster, *i.e.* close to their average $center(U^{cor}) = \sum c_{ij} U_{ij}^{cor} w_{ij} / \sum U_{ij}^{cor} w_{ij}$, where w_{ij}'s are the weights on correspondences. Thus correspondences U^{cor} satisfying the geometric consistency constraint can be expressed as:

$$\text{GeoSet} = \{\|center(U^{cor}) - c_{ij} U_{ij}^{cor}\| \leq d_{max} \text{ if } U_{ij}^{cor} \neq 0\} \tag{10}$$

where d_{max} is the maximum distance allowed for deviation from the center.

What is the right matching cost $D_{ij}(V_i^I \cdot x^{sel}, V_j^M \cdot y^{sel})$*?* Recall that our problem is to search for the maximal 'common' subsets from the image and model contours such that their shapes are similar. This maximal condition on the contour subsets places additional requirement on the shape dissimilarity D_{ij}. A straightforward cost function, such as the L_1-norm: $D_{ij}(V_i^I \cdot x^{sel}, V_j^M \cdot y^{sel}) = \|V_i^I \cdot x^{sel} - V_j^M \cdot y^{sel}\|$, will simply result in the trivial solution which chooses empty sets from both sides (*i.e.* $x^{sel} = \mathbf{0}$, $y^{sel} = \mathbf{0}$). In fact all the norms as well as χ^2 distance suffer from the same problem.

We introduces the *joint selection cost* for D_{ij} which balances the maximal requirement on the match of contour sets and the quality of the match. We seek to match as many model contours as possible while the difference between image and model contours is small. Before describing the details, we first introduce a few variables. Set

- $sc_j^{\mathcal{MF}} = V_j^M y^{full}$ to be the shape context centered at model point q_j selecting the full model, where $y^{full} = \mathbf{1}_{|\mathcal{M}| \times 1}$ means selecting all model contours;

– $sc_i^I = V_i^I x^{sel}$ to be the shape context with selection x^{sel} on image at p_i;
– $sc_j^M = V_j^M y^{sel}$ to be the shape context with selection y^{sel} on model at q_j.

We use $sc_j^{\mathcal{MF}}(k)$, $sc_i^I(k)$, $sc_j^M(k)$ to denote the k^{th} bin in the shape context.

Our joint selection cost consists of two terms: *miss* and *mismatch* (see Fig. 3). To match as many model contours as possible, the following difference between the number of matched points and that of full model points should be minimized:

$$\text{miss}_k^{(ij)} = sc_j^{\mathcal{MF}}(k) - \min(sc_i^I(k), sc_j^M(k)) \tag{11}$$

Here $\min(sc_i^I(k), sc_j^M(k))$ counts the number of matched contour points between the image and model in shape context bin k.

The above term $\text{miss}_k^{(ij)}$ alone does not measure how well the selected image contours match to the selected model contours. To ensure the matching quality, the amount of difference between the number of image and model contour points in all shape context bins needs to be minimized:

$$\text{mismatch}_k^{(ij)} = |sc_i^I(k) - sc_j^M(k)| \tag{12}$$

By combining Eq. (11) and Eq. (12), we have the following dissimilarity:

$$D_{ij} = \frac{\sum_k [\text{miss}_k^{(ij)} + \beta \cdot \text{mismatch}_k^{(ij)}]}{\sum_k sc_j^{\mathcal{MF}}(k)} \tag{13}$$

where $\beta > 1$ is a factor balancing the two types of costs. We use $\sum_k sc_j^{\mathcal{MF}}(k)$ to normalize the cost D_{ij} such that it is invariant to the number of contour points.

3 Computational Solution via Linear Programming

Direct optimization of contour context selection cost function Eq. (8) is a hard combinatorial search problem. The shape dissimilarity $D_{ij}(V^I \cdot x^{sel}, V^M \cdot y^{sel})$ can only be evaluated given correspondences U^{cor}. However, finding the correct correspondences U^{cor} requires x^{sel} and y^{sel}. Therefore, the inference problem becomes circular. We approximate this joint optimization by breaking the loop into two steps: *single point figure/ground labeling* and *joint contour selection*. The first step focuses on finding reliable correspondences U^{cor} (maybe sparse) by matching image contours to the whole model. The second step selects contours simultaneously from both image contours labelled as figure and all the model contours being matched, based on the correspondences computed in the first step. In both steps, we optimize the cost function by relaxing it as an instance of Linear Programming (LP).

3.1 Single Point Figure/Ground Labeling

Our first step discovers all potential control point correspondences U_{ij} and computes the correspondent figure/ground labeling x^{sel} for them. We fix $y^{sel} = 1$ to encourage matching to the full model as much as possible. Partial matches are undesired since the

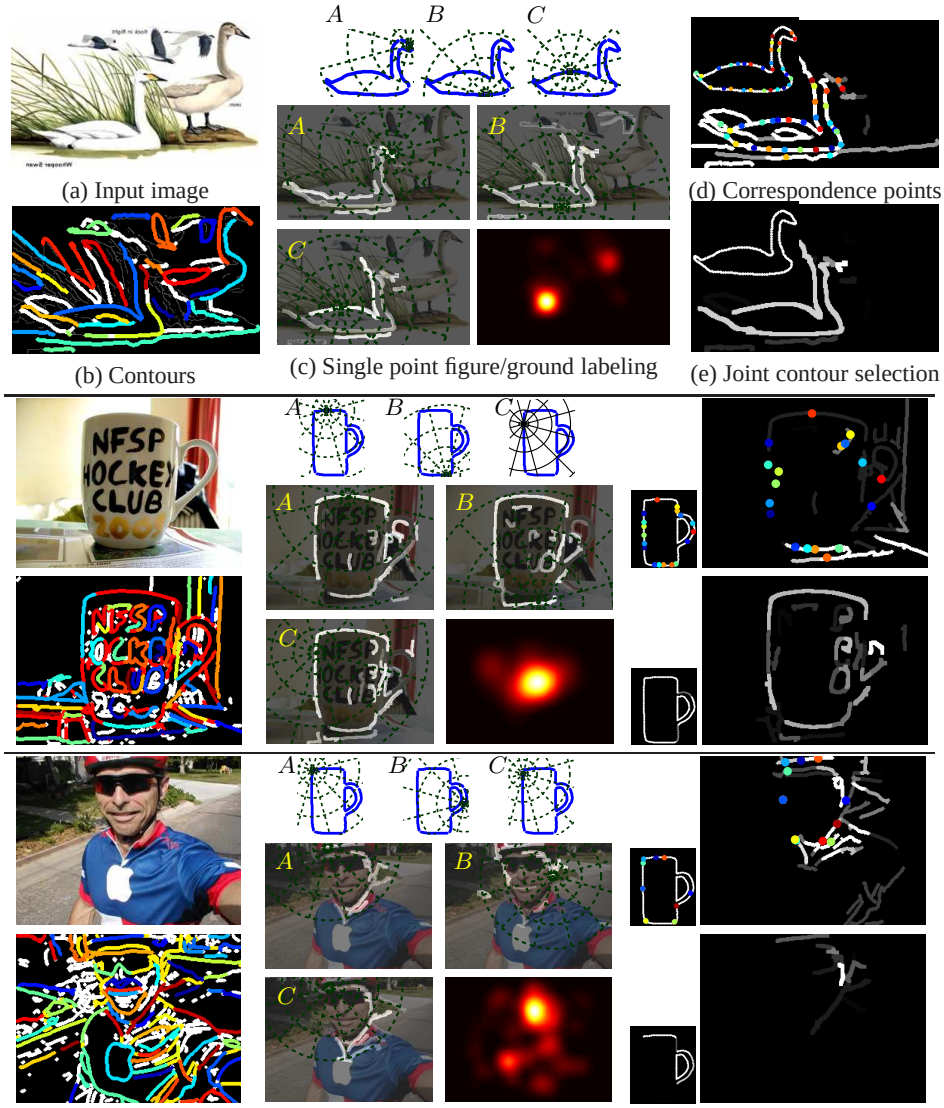


Fig. 4. Illustration of Contour Context Selection for shape detection. From the input image (a), we detect long salient contours shown in (b). For possible control point correspondences in (c), we select foreground contours whose global shape configuration most resembles to the model, with selection x^{sel} shown in gray scale (the brighter, the larger value of x^{sel}). Voting map for pruning geometrically inconsistent correspondences is shown at the right bottom corner of (c). (d) shows the consistent correspondences marked by different colors using the hypothesized correspondences. The optimal joint image-model contour selection is shown in (e). Note in the last example, model selection allow us to detect false match between the mug and the face.

correspondences they produce are much less reliable. Therefore, only the $mismatch$ term in Eq. (13) is applied and hence the dissimilarity D_{ij} reduces to L_1-norm:

$$\min_{x^{sel}} \quad \|V_i^I \cdot x^{sel} - V_j^{\mathcal{M}} \cdot 1\|_1 \tag{14}$$

This cost will not collapse to zero because model contour selection is fixed ($y^{sel} = 1$).

Brute force approach of the above problem is formidable even for mid-size problems (20-30 contours). We compute an approximate solution by continuous relaxation on binary variable x^{sel} and add the constraint $0 \leq x^{sel} \leq 1$ thereafter. Since the norm in the cost function is L_1[1], this leads to an instance of Linear Programming (LP) which can be computed very efficiently.

Correspondences found from single point figure/ground labeling might not satisfy geometric consistency Eq. (10). Therefore, we enforce geometric consistency by pruning hypotheses of control point correspondences via a voting procedure [3]. Each image control point can predict an object center using the best match to model control points computed by Eq. (14). These predictions generate votes weighted by the shape dissimilarity, which accumulate to a voting map. We extract object centers from the local maxima and further backtrace the voting centers to identify consistent correspondences.

3.2 Joint Contour Selection

We have obtained the rough correspondences U^{cor} from the previous step. We optimize the contour selection cost Eq. (8) *w.r.t.* x^{sel}, y^{sel} to prune false positives and detect objects. The outcome includes both the matching cost C_{align} and model contours actually matched, indicated by y^{sel}. Both of them can be used to prune false positives. Note that it is not required to have a complete correspondence set U^{cor} since the cost Eq. (13) has been normalized by the number of correspondences.

LP can also be used to solve Eq. (8) for contour context selection by relaxing x^{sel} and y^{sel} to real value vectors. Eq. (13) and Eq. (8) translate to the following problem:

$$\min_{x^{sel}, y^{sel}} \sum_{U_{ij}^{cor}=1} \{\frac{1}{N_i} \sum_k [sc_i^{\mathcal{MF}}(k) - \min(sc_i^I(k), sc_j^M(k))] + \frac{\beta}{N_i} \|sc_i^I - sc_j^M\|_1\}$$
$$\text{s.t.} \;\; sc_i^I = V_i^I \cdot x^{sel}, \;\; sc_j^M = V_j^M \cdot y^{sel}$$

where $N_i = \sum_k sc_i^{\mathcal{MF}}(k)$ is a normalization constant and $\min(x, y)$ computes the elementwise min of vectors x and y. The two terms in the summation are miss and $\mathrm{mismatch}$ in Eq. (13) respectively. The above problem can be relaxed to an instance of LP by adding slack variables $s_{ijk} \geq sc_i^I(k)$ and $s_{ijk} \geq sc_j^M(k)$ for $\min(sc_i^I(k), sc_j^M(k))$.

The selected model contours from joint contour selection form a shape configuration that are actually matched to image contours. Because the number of object model contours is typically very limited (usually 6 to 8), we can manually specify a dictionary of all possible configurations of true positives, *i.e.* setting C_{config} in Eq. (1) to be 0/1.

[1] Besides L_1, other distance functions such as L_2 and χ^2 for shape context can also be used. However, the relaxations will be computationally much more intensive.

Detection of model contours with bad configurations, e.g. missing critical parts, are rejected. This configuration checking together with the matching cost C_{align} can prune most of the false positives while preserving true positives.

4 Related Works and Discussion

Salient contours and their configurations are more distinctive than individual edge points for shape matching. Ferrari *et al.* ([4],[5]) represent objects by learning a codebook of Pairs of Adjacent Segments, which are consecutive roughly straight contour fragments. They achieve detection using a bag-of-words approach. Shotton *et al.* [6] learn a boosted contour-based shape features for object detection. These efforts utilize mostly short contour fragments, and therefore have to rely on many training examples to boost the discriminative power of shape features. In contrast, our work takes the advantage of contour grouping such as [1] to detect long salient contours, capturing more global geometric information of objects. More importantly, we constrain these long contours to act as a whole unit, *i.e.* they can either be entirely matched to an object, or entirely belong to the background. This characteristic makes shape matching more immune to accidental alignment to background clutter. Similar properties are exploited by grouping based verification approaches [7], and the recent work by Felzenswalb *et al.* [8].

From a broader perspective, our recognition framework is based on shape matching, which has a long history in vision. A large amount of research has been done on different levels of shape information. Early works [9,10] focused on silhouettes which are relatively simple for representing shape. Silhouette based approaches are limited to objects with a single closed contour without any interior edges with occlusions. Objects in real images are more complex, and may be embedded in heavy clutter. Efforts on dense matching of the edge points often focus on spatial configurations of key points, such as geometric hashing [11], decision tree [12] and Active Shape Models [13]. However, keypoints alone are insufficient to distinguish objects shapes in cluttered images [2].

Feature representation and shape similarity measurement are the key issues for matching. Shape Context [2] uses spatial distribution of edges points relative to a given point to describe shape. Inner Distance Shape Context (IDSC) refines it to account for articulated objects [14]. We build our basic shape feature representation on Shape Context, with contour as the unit. A much larger context window covering the whole object enables our approach to capture global shape configurations. We introduce a novel contour selection mechanism to extract global shape features against background clutter.

5 Experiments

We demonstrate our detection approach using only one hand-drawn model without negative training images, To evaluate our performance, we choose the challenging ETHZ Shape Classes [5] containing five diverse object categories with 255 images in total. Each image has one or more object instances. All categories have significant scale changes, illumination changes and intra-class variations. Moreover, many objects are surrounded by extensive background clutter and have interior contours. We have the same experimental setup as [5], using only a single hand-drawn model for each class

	Apple logos	Bottles	Giraffes	Swans	Mugs
Our precision/recall	**49.3%/86.4%**	**65.4%/92.7%**	**69.3%/70.3%**	**31.3%/93.9%**	25.7%/83.4%
Precision/recall in [5]	32.6%/86.4%	33.3%/92.7%	43.9%/70.3%	23.3%/93.9%	40.9%/83.4%

Table 1. Comparison of precision. We compare the precision of our approach to the precision in [5] at the same recall (lower recall in [5]). We convert the result of [5] reported in DR/FPPI into P/R since the number of images in each class is known. Our performance is significantly better than [5] in four out of five classes. The other class "Mugs" have some instances that are too small to be detected by contour grouping. Note that we did not use magnitude information which plays an important role in [5].

and all 255 images as test set. To adapt to large scale variance, we generate multiple models by resizing the original one to 5 to 8 scales for each class.

We first use contour grouping proposed in [1] to generate long salient contours from images. Contours can have overlaps due to multiple possible groupings at junctions. Large window shape context for contour selection has 12 polar angles, 5 radial bins and 4 edge orientations. Moreover, blurring on bins [3] is used to increase the robustness of shape context to deformation and inner-class variations. This refinement can also be encoded into contribution matrices V^I, V^M as well. LPs arising from single point figure/ground labeling as well as joint contour selection are solved efficiently by using off-the-shelf toolbox SeDuMi. Single point figure/ground labeling for each hypothesized correspondence is computed within 0.2 sec. After selecting the figure contours, votes for object center weighted by shape dissimilarity are collected into a voting map. We extract local maximums in the voting map above certain threshold to generate object hypotheses. Since the correct object scale is unknown beforehand, voting is performed in a multiscale fashion, with non-maximum suppression on both position and scale.

Precision vs. recall (P/R) curve is used for quantitative evaluation. To compare with the results in [5] which is evaluated by detection rate (DR) vs. false positive per image (FPPI), we translate their results into P/R values. We choose P/R instead of DR/FPPI because DR/FPPI depends on the ratio of the number of positive and negative test images and hence is biased. Our final results on this dataset can be seen in Fig. 5. Results of the two steps of our framework are both evaluated. Single point figure/ground labeling only uses matching cost as the final evaluation for detection, while joint contour selection uses both matching cost and the detected shape configuration. Compared to the latest result in [5], our performance is considerably better on four classes out of five. We also compare voting using simple local shape context with our first step of contour selection. Contour selection greatly improves detection performances (see Fig. 5).

Our shape matching algorithm can reliably extract and select contours of object instances in test images, robust to background clutter and missing contours. Image results of detection with selected object and model contours are demonstrated in Fig. 6.

6 Conclusion

We introduce a novel shape based recognition framework called *Contour Context Selection*. We construct context sensitive shape features depending on selected contours

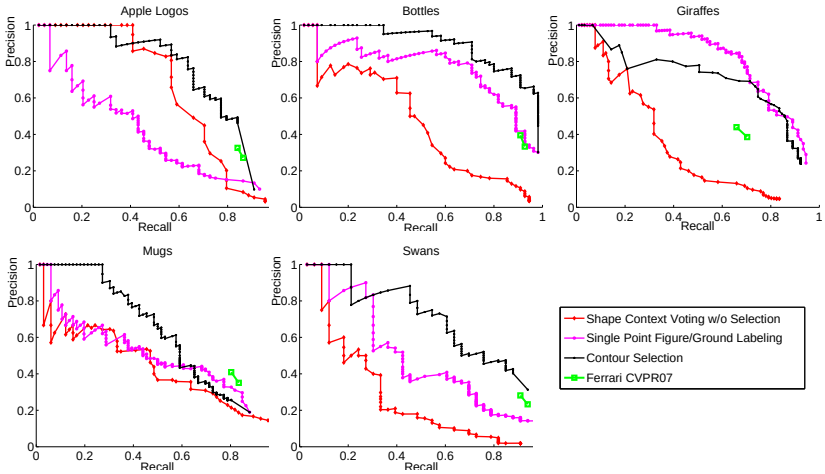


Fig. 5. Precision vs. recall curves on five classes of ETHZ Shape Classes. Our precisions on "Apple logos", "Bottles", "Giraffes" and "Swans" are considerably better than results in [5]: **49.3%/32.6%** (Apple logos), **65.4%/33.3%** (Bottles), **69.3%/43.9%** (Giraffes) and **31.3%/23.3%** (Swans). Also notice that we boost the performance by large amount compared to local shape context voting without contour selection.

and propose a method to search the best match. Joint selection on both image and model contours ensures detection to be robust to background clutter and accidental alignment. We are able to detect object in cluttered images using only one training example. Experiments on hard object detection task demonstrate promising results.

Acknowledgment. This work is partially supported by NSF grants: NSF-IIS-04-47953 (CAREER), NSF-IIS-03-33036(IDLP) and NSF-IIS004-31070. We would like to thank Alexander Toshev and Praveen Srinivasan for helpful discussions.

References

1. Zhu, Q., Song, G., Shi, J.: Untangling cycles for contour grouping. In: ICCV. (2007)
2. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. PAMI (2002)
3. Wang, L., Shi, J., Song, G., Shen, I.F.: Object detection combining recognition and segmentation. In: ACCV. (2007)
4. Ferrari, V., Fevrier, L., Jurie, F., Schmid, C.: Groups of adjacent contour segments for object detection. PAMI (2007)
5. Ferrari, V., Jurie, F., Schmid, C.: Accurate object detection with deformable shape models learnt from images. In: CVPR. (2007)
6. Shotton, J., Blake, A., Cipolla, R.: Contour-based learning for object detection. In: ICCV. (2005)
7. Amir, A., Lindenbaum, M.: Grouping-based nonadditive verification. PAMI (February 1998)
8. Felzenszwalb, P.F., Schwartz, J.D.: Hierarchical matching of deformable shapes. In: CVPR. (2007)

False positives pruned by model contour selection | Failure cases

Fig. 6. Examples of contour context selection on model and image contours in ETHZ Shape Classes. The first five rows show detected objects from image with significant background clutter. In the last row, the first four cases are false positives successfully pruned by our algorithm due to bad configurations of selected model contours. The last two are failure cases. Each image only displays one detected object instance.

9. Zahn, C.T., Roskies, R.S.: Fourier descriptors for plane closed curves. IEEE Trans. of Computing (March 1972)
10. Gdalyahu, Y., Weinshall, D.: Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes. PAMI (1999)
11. Lamdan, Y., Schwartz, J.T., Wolfson, H.J.: Affine invariant model-based object recognition. IEEE Trans. Robotics and Automation **6** (1990) 578–589
12. Amit, Y., Wilder, K.: Joint induction of shape features and tree classifiers. PAMI (1997)
13. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer Vision and Image Understanding **61**(1) (1995) 38–59
14. Ling, H., Jacobs, D.W.: Using the inner-distance for classification of articulated shapes. In: ICCV. (2005)