Bottom-Up Perceptual Organization of Images into Object Part Hypotheses

Maruthi Narayanan and Benjamin Kimia

Brown University
School of Engineering
Providence, RI 02912
{maruthi_narayanan,benjamin_kimia}@brown.edu
http://vision.lems.brown.edu

Abstract. The demise of "segmentation-then-recognition" strategy led to a paradigm shift toward feature-based discriminative recognition with significant success. However, increased complexity in multi-class datasets reveals that local low-level features may not be sufficiently discriminative, requiring the construction and use of more complex structural features which are necessarily category independent. The paper proposes a bottom-up procedure for generating fragment features which are intended to be *object part hypotheses*. Suggesting that the demise of segmentation to generate a representation suitable for recognition was due to prematurely committing to a grouping option in the face of ambiguities, the proposed framework considers and tracks multiple alternate grouping options. This approach is made tractable by *(i)* using a *medial fragment* representation which allows for the simultaneous use of multiple cues, *(ii)* a set of transforms to effect grouping operations, *(iii)* a *containment graph* representation which avoids duplicate consideration of possibilities, and the estimation of the likelihood of a grouping sequence to retain only plausible groupings. The resulting hypotheses are evaluated intrinsically by measuring their ability to represent objects with a few fragments. They are also evaluated by comparison to algorithms which aim to generate full object segments, with results that match or exceed the state of art, thus demonstrating the suitability of the proposed mid-level representation.

1 Introduction

A significant progress bottleneck in the object recognition and multiview geometry applications in the 90's was attributed to the difficulties experienced by bottom-up segmentation, which could not produce segments sufficiently stable to be used as the basic matching units in these applications. A dramatic paradigm shift, relying not on fully segmented images but rather on a collection of features, driven by feature detectors [1] and descriptors, such as SIFT [2] and others [3], on the one hand, and significant process in learning methods on the other, led to a breakthrough in both of these applications, *e.g.*, on PASCAL for object recognition [4]. While modern approaches are now able to tackle a

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part I, LNCS 7572, pp. 257–271, 2012.

Fig. 1. A sampling of object part hypotheses for images from standard datasets shows recognizable fragments (in the actual fragment set the proportion of spurious fragments to recognizable ones is significantly higher, as expected). For the rightmost figure (Berkeley dataset) we only show the fragments that overlap with the ground truth object to show that the parts of the object are indeed covered by our hypotheses.

couple of hundred or so categories, experience with much larger datasets, *e.g.*, ImageNet [5] and denser datasets, *e.g.*, the Mammalian dataset [6] casts doubt on the *scalability* of these methods, as elaborated on below.

The main limiting factor concerns the *nature of the image representation* used. Recognition performance for methods that rely on the appearance of local patches drops exponentially with an increase in the number of categories, with a more severe drop rate when the database enjoys greater variation; see Figure 3 of [7]. Similarly, image classification rates drop exponentially, *e.g.*, from 35% to 7% when the number of categories goes from 200 to 10,000 [8]. The same drop in recognition performance can be observed if, instead of increasing the number of categories, the density of categories in the semantic space is increased: Performance drops to half in going from Caltech101 to the smaller but denser 72 category mammalian dataset [6]. Tatu *et al.* [9] showed that the image representation is a confounding factor in that the equivalence class of each HOG-based representation is enormous, spanning multiple categories (which does not show when comparing a limited number of categories, *e.g.,* an airplane against a potted plants.)

We posit that recognition success in these much larger datasets critically depends on the development of *more complex, more diagnostic features* to differentiate fine-grain category differences such as horses from donkeys, as compared to differentiating airplanes from potted plants or airplanes from the background in the smaller datasets. The construction of complex features when considering a large number of categories needs to take place in a bottom-up, category independent fashion [10]. Bottom-up grouping, however, requires tackling ambiguity, precisely the challenge faced by segmentation algorithms. Observe that classical bottom-up segmentation of pixels into regions, or edges into contours, involves effecting a sequence of grouping actions. Even if each of these is highly likely, a sequence of such actions is highly likely to fail if the number of operations is large enough. Instead, it is suggested here that *all* plausible grouping operations, rather than the best local grouping, be considered and tracked to generate alternate, conflicting but complementary outcomes. This form of *perceptual reasoning* seems intractable at first due to the apparent combinatorial explosion of possibilities. Nevertheless, it is shown here that through an effective representation, use of multiple cues, and avoidance of duplication, the grow rate is sufficiently tamed to allow reasonable partial groupings to form. A fraction of

these groupings represent meaningful object parts and it is argued that this is all the higher level processes really need; see Figure 1 for examples.

The bottom-up delineation of recognizable parts of objects has distinct advantages over bounding box or rectangular fragments [11]: *(i)* it removes the noise due to mixing foreground/background statistics, especially for complex, deformable, or articulating objects. *(ii)* it reveals the natural scale of the object that is important for selecting scale of features; *(iii)* representation and recognition of object parts allows for top-down full object segmentation; *(iv)* it allows for category-independent learning which is feasible over fragments but not over pixels; *(v)* allows for the discovery and learning of new objects; *(vi)* improved efficiency as bottom-up fragments are shared among multi-classes.

The idea of organizing pixels into fragments for recognition as a precursor to this work was proposed in [12]. The first practical use of multiple segmentation hypotheses was perhaps achieved in [13]. It has also been advanced in the form of taking advantage of segments formed from a hierarchical segmentation, *i.e.*, from a "segmentation tree" or a "region tree"' [14–16]. The drawback of this type of an approach is that erroneously formed segments can propagate from level to level, thus preventing the formation of the veridical segments. [17] used medial fragments randomly sampled in space and scale from the image shock graph to generate object part hypothesis which were then successfully used for object recognition, even without any learning. Alternatively, in the "Soup of Segments" approach [18–20] a diverse set of segments is generated by inducing *variations* in the segmentation procedure: *(i)* by using complementary segmentation algorithms, *(ii)* by changing segmentation parameters or seeding, and *(iii)* by merging adjacent fragments. In a more recent paper [19] a large number of independent binary min-cut segmentation problems are solved by starting from a sampling of seeds and terminals, with different extents of foreground bias. The resulting pool of figure segments are then processed and ranked using certain regularities typical of projections of real-world objects, and a learning scheme is used to sample a diverse subset. Another recent paper [20] samples a set of seed regions, obtained from a hierarchical segmentation and by varying parameters in a CRF framework, to generate a diverse set of regions that are guided toward object segmentations by learned affinity functions; a structured learning approach is then used to rank the regions so that the top-ranked regions are likely to correspond to different objects.

Our approach is distinct in several ways: First, previous work aims to generate full object hypotheses, while we aim to generate recognizable and meaningful object part hypotheses. Second, previous work generates multiple hypotheses by varying segmentation variables such as parameters, algorithms, seeding, *etc.,* while we systematically and exhaustively investigate all reasonable grouping options in a perceptual reasoning Gestalt framework, by asking questions like: what if these contour fragments are bridged across a gap? What if this contour fragment is spurious? These possibilities are all methodically tracked and the entire sets of options are maintained! This is an important distinction because while it is possible that viable fragments are not formed under the variations-of-variables scheme (e.g., object was too small and no seeds were initialized inside),

it is much less likely that viable fragment would avoid being considered due to the structural and exhaustive nature of the proposed scheme.

The overview of the paper is as follows. In Section 2 we show that superpixels and contour fragments lack sufficient representation bandwidth and instead propose *medial fragments*, which combines aspects of both. Section 3 casts perceptual grouping operations, such as gap completion, removal of spurious contours, and others, in the form of *transforms* on the medial fragments, with the distinct advantage of using both form and appearance in the transform. Section 4 introduces the containment graph as a mean of efficiently representing competing and alternative groupings, as a mid-level organization and a gateway to object categorization. Section 5 describes experiments and comparisons to related work.

2 Medial Fragments: Superpixels and Contours Won't Do

Two types of representations suggest themselves on which various grouping operations can operate: *superpixels* and *contour fragments*. It is argued below that neither is sufficient for the purpose of building multiple hypotheses in the face of realistic levels of ambiguity and clutter [12]. Instead we motivate and adopt *medial fragments* [12].

Superpixels are the basic unit of reasoning about object fragments resulting from grouping pixels into region fragments based on regional homogeneity [21–23], *etc.* One main difficulty with the boundaries of superpixels is that it is not clear which portions are meaningful boundaries and which are simply delimiters of the grouped pixels. These latter contours are more a product of the dynamics of the grouping process than an indicator of underlying structure, *e.g.*, the pink-yellow boundary in Figure 2b. The ambiguity created by the presence of these artifactual boundaries among meaningful boundaries limits the use of grouping based on contours.

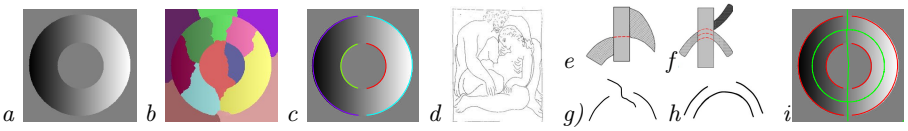


Fig. 2. (a,b) The use of superpixels is not appropriate in our approach because (i) artifactual boundaries (say the contour between pink and yellow regions) are not differentiated from actual boundaries, and (ii) the representation does not identify gap endpoints, required to identify completion candidates. (c) contour fragment endpoints identify potential gaps. (d) many image contours are not closed (e-f) appearance continuity is a powerful cue in grouping. The spatial interaction among contours prevents some gaps from being considered (g), while supporting others (h). (i) The shock graph pairs image contours.

Another fundamental difficulty in region-based representation is the inability to represent contours that end, Figure 2d, which among others, significantly limits the detection of gaps, *e.g.*, as present in the bottom and top portions of the inner and outer circles in the shaded torus of Figure 2a; It is difficult to

work with gaps in the superpixel representation, Figure 2b, because end-points of contours are not represented.

Curve fragments are often viewed as a precursor to figure-ground segregation and for formulating the Gestalt notion of good continuation, Figure 2c, *e.g.*, using elastica [24, 25] or using the Euler Spiral [26]. Good continuation is used to disambiguate grouping choices. However, a contour representation cannot take into account the interaction of nearby contours which may invalidate a completion, Figure 2g, or which continuations group the wrong sides of a figure, Figure 2e. This representation also cannot strongly encourage the groupings in Figure 2h, as it should because it cannot represent "*figural continuity*", a very powerful cue in disambiguating conflicting groupings. Figural continuity requires good continuation of *a pair of contours* that are bound together and that continue together. Finally, another proposed powerful Gestalt cue is *appearance continuity*, which states that two fragments' grouping depends on the continuity of their appearance, Figure 2f. This requires a representation of the region bound between two contour fragments, clearly lacking from a contour representation.

Medial fragments: [12] in contrast *(i)* differentiate meaningful contours from contours that delimit regions; *(ii)* represent end-points, *(iii)* retain boundary fragments and pair them and *(iv)* represent the region between a pair of boundary fragments. Formally, consider the set of image contour fragments $C_i, i = 1, \cdots, ...N$ obtained from some contour extraction process. The shock graph [27–29] is a data structure similar to the medial axis, but which is reinterpreted as the locus of singularities, or shocks, thus refining the classification of medial points to sources, sinks, branches, which are isolated and represent the shock graph nodes, and others which connect these through monotonically flowing curves; representing shock graph links. Denote the shock graph links by $S_j, j = 1, \cdots, M$, with each S_j arising as a result of interaction of waves from a pair of contour fragments or isolated points. The region spanned by the propagating waves from the pair of boundaries that quenched at the shock segment, Figure 4g-h, is the zone of influence associated with the shock segment. Each pixel uniquely belongs to some zone of influence. Each shock segment, the two associated contour fragments, and the region bound between these is jointly referred to as a *Medial Fragment*. We refer to the medial fragments that initially arise from the contour fragments, without any grouping, as *Atomic Fragments*. Atomic fragments do not contain any contours by construction, and are analogous to superpixels, except that real and artifactual contours are explicitly represented, Figure 3 and form the basic unit of perceptual reasoning.

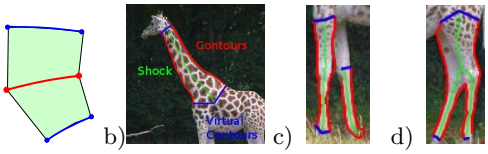


Fig. 3. A sketch of a fragment (a), is also sketched on top of an image in (b). Real examples are shown in (c,d).

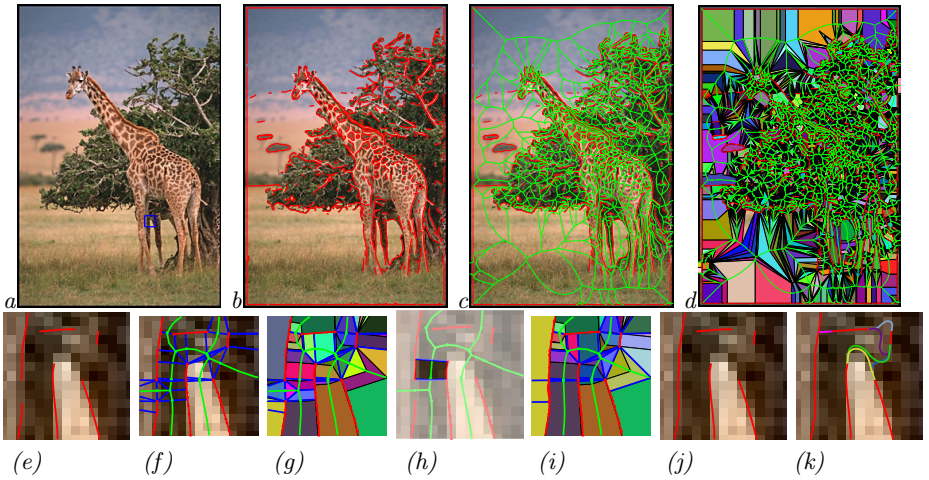


Fig. 4. (a) The original image, (b) its contour fragments in red, (c) the shock graph in green (the dynamics and types of shocks are not shown, but are very important), (d) the resulting medial fragments, (e) contour fragments of the zoomed area in (a), (f) composite contour and shock graph, (g) medial fragments, (h) one medial fragment highlighted, (i) the application of Gap I transform from Figure 5; (j) the resulting contours; (k) The next set of potential completions

3 Fragment Transforms

The projection of objects onto images undergo an onslaught of visual transformations, as illumination, object pose, viewing distance or direction, *etc.* vary. In many cases contours extracted from images do not map to a complete object silhouette: gaps and clutter edges abound or internal contours due to 3D folds and self-occlusion, and contours due to partial occlusions. Our goal in this paper is to obtain parts whose shape and appearance is distinct and recognizable. Thus, gaps must be completed, spurious contours removed, fragments across occlusions connected, *etc.* This paper argues that there are significant advantages to casting these operations as transforms on the medial fragment as opposed to grouping contour fragments or regions. A sequence of medial transforms then represents a perceptual grouping operation. We illustrate these below.

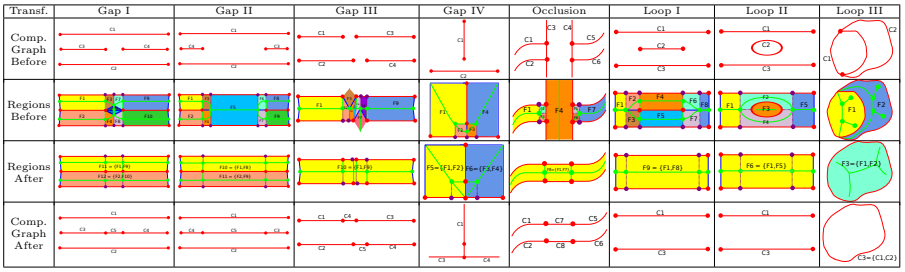


Fig. 5. A partial list of transforms, arranged in columns

Gaps are those portions of veridical contour which did not make it through the contour extraction process. As such, each gap leaves behind a pair of *end-points*, or a single end-point if the gap is at a junction. The remedy is simple: complete the contour between two end-points by interpolation. However, which of the many pairing of end points should be completed? To aid in the selection process, we augment the well-known Gestalt cue of *contour continuity* with two other proposed cues: *figural continuity* and *appearance continuity*. The medial fragment representation makes this possible, since the template in the first column of Figure 5 will not form if there are intervening contours. Specifically *(i)* contour continuity is formulated in by Euler Spiral energy [26] with variables $\{\gamma, \kappa_0, l\}$, where γ, κ_0, and l represent derivative of curvature, initial curvature, and arc-length of the completion curve. The parameters of the energy function are learned under supervised learning using a generalization of the Geisler model [30]; *(ii)* figural continuity is formulated by comparing the radius of the medial axis for fragments F1 and F9, and F2 and F10, respectively; *(iii)* Appearance continuity ensures that the appearance of fragments F1 and F9, and F2 and F10, respectively is contiguous; we have not implemented the latter cue at the moment. The first two cues associate a likelihood with each GAP I Transform. For a recent perceptual study of gap completion see [31].

The remaining two types of gaps are essentially the same modulo differences in gap width and end-point arrangement. The fourth type of gap arises from a junction, but it is essentially handled in the same way. There is also an dis-occlusion transform which is essentially a gap in the figure.

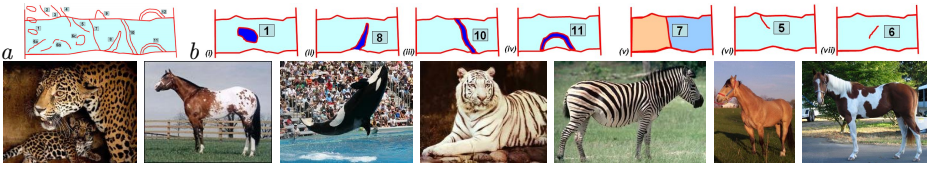


Fig. 6. (a)We use a sketch of a normative object part, the horizontal cyan strip, to systematically derive all possible ways a spurious contour can interfere with its formation, leading to seven principal cases, as in(b). These can be easily recognized in natural images (c), *e.g.*, leopard spots, markings off a whale, tiger stripes, *etc.* Our set of loop transforms correspond to the removal of their effect so that object part hypotheses can form.

Spurious Contours are those contour fragments that do not belong to the object part boundaries, although they often explain the 3D local form, object texture, reflectance, *etc.* They are only spurious in the sense that they interfere with the formation of an object part. A syntactic account of how spurious contours can interfere with the formation of object part boundaries in seven canonical cases is shown in Figure 3a-b. On all cases the shock graph has a loop corresponding to the interfering element which can be removed by the application of one of the three loop transforms shown in Figure 5.

The remaining transforms are similarly intuitive embodiments of known Gestalt grouping operations and will not be further explained here. It is possible that

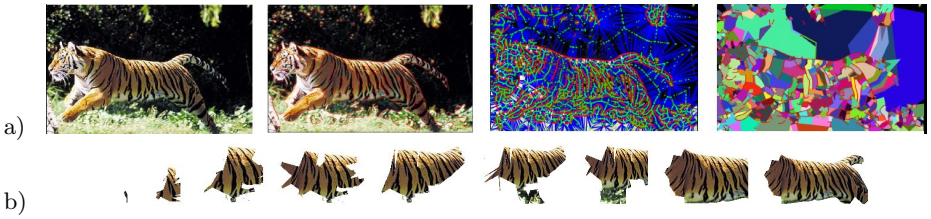


Fig. 7. a) a tiger, its contours, the full shock-contour graph, and fragments. b) A manual sequence of transforms leads to a recognizable object part.

additional transforms arise with additional experience with this approach, but the ones listed are the major ones. This language of transforms is powerful enough to be able to organize tiger parts from the very challenging tiger image, Figure 7.

A critical issue in exploring alternative perceptual organization is to associate a likelihood to each transform sequence so that only plausible cases can be explored. Assuming independence among transforms, the likelihood a transform sequence forms a valid fragment is

$$p(T_1, T_2, ..., T_N) = \prod_i^N p(T_i) \tag{1}$$

where $p(T_i)$ represents the likelihood of a transform and $p(T_1, T_2, ..., T_N)$ is a transform sequence. See supplementary section for further technical details.

4 Perceptual Reasoning with the Containment Graph

The successful manual sequence of transforms in Figure 7 is one among numerous many: Initially and after applying a transform sequence, there are typically a number of transforms which are applicable next. This leads to an exponential search space which quickly becomes impractical. Two key observations reduce the search space to a manageable size: *(i)* not all transform sequences are plausible, and *(ii)* there is a significant duplication in that two distinct transform sequences can result in the same fragment. We propose the idea of a *containment graph* to implement both.

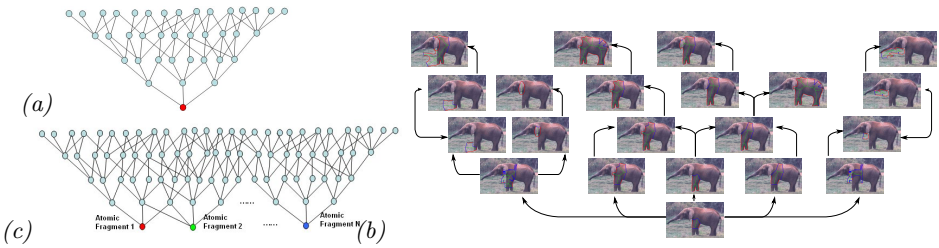


Fig. 8. (a) The graph of grouping possibilities involving a single fragment. (b) An example graph for a fragment of an elephant. (c) The union of graphs from all fragments is the **containment graph**.

First, the state of an image at any point in the course of applying transforms is represented by a node, and the set of applicable transforms at each stage are represented by links. Observe that two transform sequences applied to two distinct fragments can lead to the same state, especially when the two fragments are adjacent or the same. Figure 8b. The containment graph is constructed in a breath first manner and effectively represents inherent grouping ambiguity by representing each grouping hypothesis as a node. The containment graph nodes are our proposed mid-level representation to be used by high level processes.

Second, just as it is unwise to commit to a single grouping option at each grouping stage, it is equally unwise to keep all the options indiscriminately against mounting evidence. The application of each transform sequence represents an "investment based on faith" since raw, proximal data is being manipulated to generate a possibly more structured and regular representation, *e.g.*, contour completion across a gap may close a shape. The likelihood defined in Equation 1 is therefore capped at a minimum value to limit the exploration of highly unlikely transformation sequences. This leads to numerous terminal nodes in the containment graph.

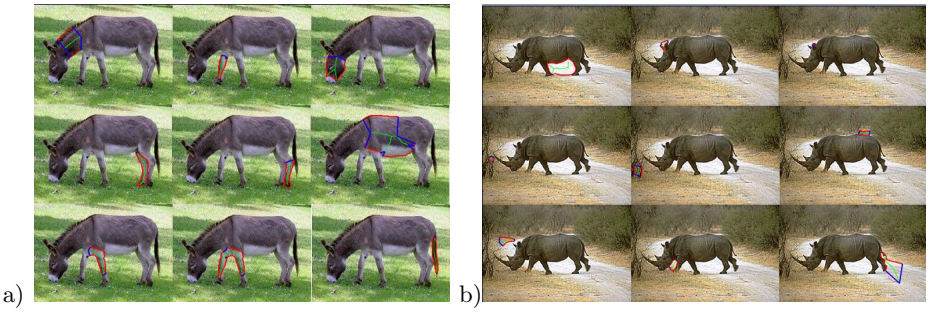


Fig. 9. Selection of nodes from the containment graph: a) donkey fragments b) rhino fragments

The value of using the containment graph can be measured by probing the total number of nodes explored with and without a containment graph. We expect that as the number of contour fragments increases so does the relative value of the containment graph. Figure 10 shows the relative number of nodes

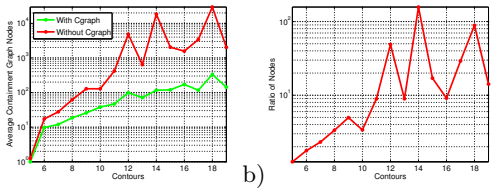


Fig. 10. a) Average Number of Nodes with and without the containment graph versus the number of contours b) Ratio of Nodes with and without containment graph

explored for a coarse scale version of the Weizmann Segmentation Dataset [32]. It is evident that the containment graph reduces the search space between an order to two orders of magnitude for 10-20 contours. The value of the containment graph at a finer scale when the number of fragments exceed beyond 20 cannot be explored because it is impossible to run the experiments without the containment graph in any reasonable amount of time!

5 Experiments

We present experimental results along two distinct directions, one on the ability of the method to segment figure from ground, as in traditional algorithms, and one on the ability of the method to generate object parts. While the former is a forced evaluation of the algorithm outside its native scope (we aim to generate object parts not full objects), nevertheless we have found that in fact our fragmentation results match, and at times, exceed the quality of full object segments produced by existing techniques. Second, in the absence of an existing procedure to evaluate algorithms which produce object fragments, we introduce an approach for evaluating fragmentation algorithm results.

Evaluation of Segmentation. Algorithms are tested against two popular and public ally available segmentation datasets for which manual segmentation is available: *(i)* Berkeley Segmentation Dataset and *(ii)* MSRC Dataset. The standard measure to compare two fragments F_i and $\bar{F}_{\bar{i}}$ is the *Jaccard Index* $J(F_i, \bar{F}_{\bar{i}}) = \frac{|F_i, \cap \bar{F}_{\bar{i}}|}{|F_i, \cup \bar{F}_{\bar{i}}|}$, also known as *overlap* in the object recognition literature. Given a ground truth fragment F_i, a set of captured fragments $\bar{\mathcal{F}} = \{\bar{F}_1, \bar{F}_2, \cdots, \bar{F}_{\bar{N}}\}$ is gauged by the *Best Fragment Score (**BFS**)* [19, 20].

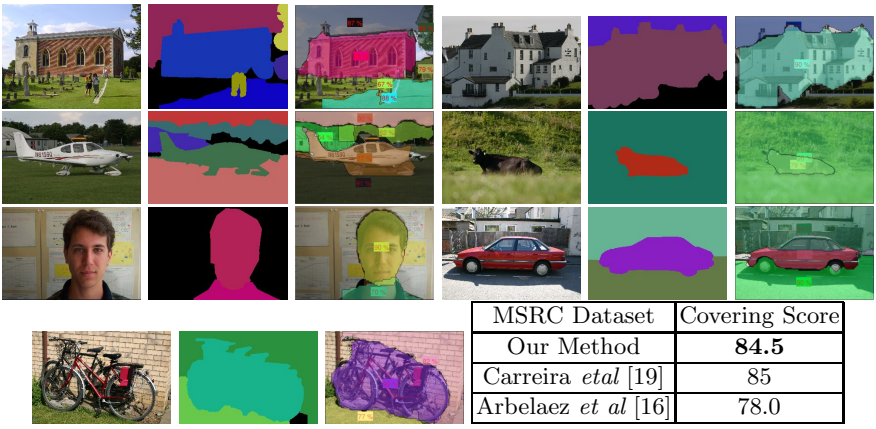
$$BFS(F_i, \bar{\mathcal{F}}) = \max_{\bar{F}_{\bar{i}} \in \bar{\mathcal{F}}} J(F_i, \bar{F}_{\bar{i}}). \tag{2}$$

Two sets of fragments, say ground truth fragments $\mathcal{F} = \{F_1, F_2, \cdots, F_N\}$ and computed fragments $\bar{\mathcal{F}} = \{\bar{F}_1, \bar{F}_2, \cdots, \bar{F}_{\bar{N}}\}$ are compared using the notion of *covering* [16], measuring how well computed fragments $\bar{\mathcal{F}}$ cover ground truth fragments in $\mathcal{F}$,

$$Covering(\mathcal{F}, \bar{\mathcal{F}}) = \frac{\sum_{i=1}^{N} BFS(F_i, \bar{\mathcal{F}})|F_i|}{\sum_{i=1}^{N} |F_i|} = \frac{\sum_{i=1}^{N} |F_i| \max_{\bar{F}_{\bar{i}} \in \bar{\mathcal{F}}} J(F_i, \bar{F}_{\bar{i}})}{\sum_{i=1}^{N} |F_i|} \tag{3}$$

Figure 11 shows the best fragment per annotated object for a selection from MSRC images with a comparison to [19, 16]. Figure 12 illustrates our results on BSDS with a comparison to [20, 19].

When evaluating any segmentation algorithm, it is also important to asses the number of segments produced. Ideally one would like as few as possible fragments, while still maintaining segmentation quality. Our current scheme for reducing the number of fragments is very simple at this stage: in a visitation schedule of fragments we remove all fragments whose overlap with retained fragments exceeds a threshold. For example, for the BSDS300 Test dataset the object



MSRC Dataset	Covering Score
Our Method	**84.5**
Carreira *etal* [19]	85
Arbelaez *et al* [16]	78.0

Fig. 11. The best object part hypotheses for MSRC examples and Covering Score

BSDS TestSet	Recall@50% **BFS**
Our Method	**87%**
Endres *etal* [20]	84%
CPMC [19]	75%

Fig. 12. The best object part hypotheses for examples from the Berkeley dataset

recall at the pascal criterion is 86% with 8600 fragments. More complex clustering algorithms together with a ranking procedure (as used in [19, 20]) should reduce this at least another order of magnitude or more, altough we are now exploring this. The fragments produced by our approach match or slightly exceed the state of the art in segmenting figure from ground, even though the focus of our approach is not really figure ground segregration but rather generating object part hypotheses. The proper evaluation of this aspect motivates the introduction of a new measure.

Evaluation of Fragmentation: The obvious distinction between evaluation of Figure-Ground segregration algorithms and those that are expected to generate object parts is that in the latter the fragments are not necessarily expected to cover the entire object with a single fragment. The expectation from a good "fragmentation algorithm" is that each object would be covered with as few fragments as possible, and as accurately as possible. First a fragment $\bar{F}_{\bar{i}}$, should participate in describing an object F_i only if it accurately describes a portion of it, *i.e.,* participating fragments are expected to have high precision, $\frac{|F_i \cap \bar{F}_{\bar{i}}|}{|\bar{F}_{\bar{i}}|} > \tau_p$

Fig. 13. An illustration of evaluating the value of fragments to describe objects from a sample of BSDS

where τ_p is a constant close to 1, *e.g*, $\tau_p = 0.95$. Second a set of high precision fragments $\{\bar{F}_{\bar{i}_1}, \bar{F}_{\bar{i}_2}, ..., \bar{F}_{\bar{i}_K}\}$ can accurately describe an object if their union has high percent object recall, $\frac{|F_i \cap (\bigcup_{k=1}^{K} \bar{F}_{\bar{i}_k})|}{|F_i|}$. Third, since any object can be covered well with a very large number of fine-grained fragments, good coverage needs to be counterbalanced with the number of fragments used to describe the object. Thus, our approach to the evaluation of fragments trades off cumulative percent object recall versus the number of fragments used to describe the object, as measured on a pool of high precision fragments. Since the selection of which subset of fragments should be used to explain the object is combinatorially large we take a greedy approach to it: The best fragment in the pool of high precision fragments is selected as the first fragment, the set of object pixels explained by this fragment is removed from contributing to future recall values and the remaining part of the object is probed for coverage by the remaining high precision fragments. In other words, given that l fragments have been selected, the $l+1^{th}$ fragment is selected by

$$\min_{j \notin \{i_1, i_2, \cdots, i_l\}} \frac{|[F_i - (\bigcup_{k=1}^{l} \bar{F}_{\bar{i}_k})] \cap F_j|}{|F_i - (\bigcup_{k=1}^{l} \bar{F}_{\bar{i}_k})|} = \min_{j \notin \{i_1, i_2, \cdots, i_l\}} |F_i \cap \bar{F}_{\bar{j}} - (\bigcup_{k=1}^{l} \bar{F}_{\bar{i}_k}) \cap \bar{F}_{\bar{j}}|$$

This procedure is illustrated in Figure 13

Figure 14 shows the fragmentation score averaged over the entire BSDS300 Test dataset. Observe that the best high precision fragment of [19] covers slightly more of the object than our best high precision fragment does. However, as more fragments are included, our fragments outperform those of [19]. This is somewhat expected as the stated goal of [19] is to provide figure ground segmentation through many tries, while our stated goal is to produce object part fragments.

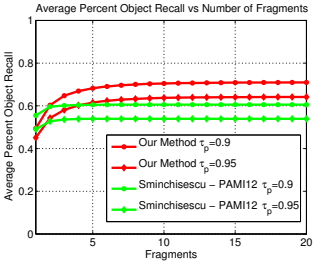


Fig. 14. Average Fragmentation Measure at $\tau_p = 0.90, 0.95$ for Test Set of BSDS300

6 Conclusion

We have presented a novel mid-level representation of an image as a hierarchical set of fragments, some of which delineate object parts. The approach generates these hypotheses by simultaneously representing alternative and conflicting grouping options. The approach is tractable due to the use of a medial representation of fragments which uses multiple cues simultaneously and due to the use of a containment graph which avoids duplication and which only explores viable options. It is clearly shown above that the fragments generated by our approach contain full object segments, matching or exceeding the current methods, although the primary focus is on producing fragments that represent object parts effectively. The framework has a great deal of unexplored potential, *(i)* the full set of transforms has not been implemented yet, and *(ii)* not all the likelihood functions use the full set of cues available. We plan to explore recognition strategies based on these fragments. These developments are expected to significantly increase the performance of this approach.

Acknowledgments. The support of NSF grant 1116140 and the Brown University Center for Computation and Visualization is gratefully and the acknowledged.

References

1. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.J.V.: A comparison of affine region detectors. IJCV 65(1-2), 43–72 (2005)
2. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. Journal of Computer Vision 60(2), 91–110 (2004)
3. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. 27(10), 1615–1630 (2005)
4. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1627–1645 (2010)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, Florida, USA. IEEE Computer Society Press (2009)
6. Fink, M., Ullman, S.: From Aardvark to Zorro: A benchmark for mammal image classification. International Journal of Computer Vision 77(1-3), 143–156 (2008)
7. Griffin, G., Perona, P.: Learning and using taxonomies for fast visual categorization. In: CVPR 2008. IEEE Computer Society (2008)
8. Deng, J., Berg, A.C., Li, K., Fei-Fei, L.: What Does Classifying More Than 10,000 Image Categories Tell Us? In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 71–84. Springer, Heidelberg (2010)
9. Tatu, A., Lauze, F., Nielsen, M., Kimia, B.B.: Exploring the representation capabilities of the HOG descriptor. In: ICCV Workshops, pp. 1410–1417. IEEE (2011)

10. Dickinson, S.: The evolution of object categorization and the challenge of image abstraction. In: Dickinson, S., Leonardis, A., Schiele, B., Tarr, M. (eds.) Object Categorization: Computer and Human Vision Perspectives, pp. 1–37. Cambridge University Press (2009)
11. Vidal-Naquet, M., Ullman, S.: Object recognition with informative features and linear classification. In: ICCV, Nice, France, pp. 281–288 (2003)
12. Tamrakar, A., Kimia, B.B.: Medial visual fragments as an intermediate image representation for segmentation and perceptual grouping. In: Proceedings of CVPR Workshop on Perceptual Organization in Computer Vision, p. 47 (2004)
13. Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image. In: ICCV 2005: Proceedings of the Tenth IEEE International Conference on Computer Vision, pp. 654–661. IEEE Computer Society (October 2005)
14. Todorovic, S., Ahuja, N.: Extracting subimages of an unknown category from a set of images. In: CVPR 2006, pp. 927–934. IEEE Computer Society (2006)
15. Todorovic, S., Ahuja, N.: Unsupervised category modeling, recognition, and segmentation in images. IEEE Trans. Pattern Anal. Mach. Intell. 30(12), 2158–2174 (2008)
16. Arbelaez, P., Maire, M., Fowlkes, C.C., Malik, J.: From contours to regions: An empirical evaluation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, Florida, USA, pp. 2294–2301. IEEE Computer Society Press (2009)
17. Ozcanli, O.C., Kimia, B.B.: Generic object recognition via shock patch fragments. In: Rajpoot, N.M., Bhalerao, A. (eds.) Proceedings of the British Machine Vision Conference, September 10-13, pp. 1030–1039. Warwick Print, Coventry (2007)
18. Malisiewicz, T., Efros, A.A.: Improving spatial support for objects via multiple segmentations. In: British Machine Vision Conference, BMVC (September 2007)
19. Carreira, J., Sminchisescu, C.: CPMC: Automatic object segmentation using constrained parametric min-cuts. IEEE Trans. Pattern Anal. Mach. Intell. (2012)
20. Endres, I., Hoiem, D.: Category Independent Object Proposals. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 575–588. Springer, Heidelberg (2010)
21. Ren, X., Malik, J.: Learning a classification model for segmentation. In: ICCV 2003: Proceedings of the Ninth IEEE International Conference on Computer Vision, pp. 10–17. IEEE Computer Society (2003)
22. Ahuja, N., Todorovic, S.: Connected segmentation tree - a joint representation of region layout and hierarchy. In: CVPR 2008. IEEE Computer Society (2008)
23. Malisiewicz, T., Efros, A.A.: Improving spatial support for objects via multiple segmentations. In: British Machine Vision Conference, BMVC (September 2007)
24. Mumford, D.: Elastica and computer vision. In: Algebraic Geometry and Its Applications, pp. 491–506. Springer (1994)
25. Williams, L., Jacobs, D.: Stochastic completion fields: A neural model of illusory contour shape and salience. Neural Computation 9, 849–870 (1997)
26. Kimia, B.B., Frankel, I., Popescu, A.M.: Euler spiral for shape completion. IJCV 54, 159–182 (2003)
27. Kimia, B.B., Tannenbaum, A.R., Zucker, S.W.: Toward a Computational Theory of Shape: An Overview. In: Faugeras, O. (ed.) ECCV 1990. LNCS, vol. 427, pp. 402–407. Springer, Heidelberg (1990)
28. Kimia, B.B., Tannenbaum, A.R., Zucker, S.W.: Shapes, shocks, and deformations, I: The components of shape and the reaction-diffusion space. IJCV 15(3), 189–224 (1995)

29. Giblin, P.J., Kimia, B.B.: On the intrinsic reconstruction of shape from its symmetries. PAMI 25(7), 895–911 (2003)
30. Geisler, W.S., Perry, J.S., Super, B.J., Gallogly, D.P.: Edge co-occurrence in natural images predicts contour grouping performance. Vision Research 41, 711–724 (2001)
31. Narayanan, M., Kimia, B.: To complete or not to complete: Gap completion in real images. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 47–54 (June 2012)
32. Alpert, S., Galun, M., Basri, R., Brandt, A.: Image segmentation by probabilistic bottom-up aggregation and cue integration. In: CVPR 2007. IEEE Computer Society (2007)
33. Maire, M., Arbelaez, P., Fowlkes, C., Malik, J.: Using contours to detect and localize junctions in natural images. In: CVPR 2008, pp. 1–8. IEEE Computer Society (2008)
34. Tamrakar, A., Kimia, B.B.: No grouping left behind: From edges to curve fragments. In: ICCV 2007: Proceedings of the Eleventh IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil. IEEE Computer Society (October 2007)