Perceptual Grouping using Superpixels

Sven J. Dickinson[1], Alex Levinshtein[1], and Cristian Sminchisescu[2]

[1] University of Toronto
sven,babalex@cs.toronto.edu
[2] University of Bonn
cristian.sminchisescu@ins.uni-bonn.de

Abstract. Perceptual grouping plays a critical role in both human and computer vision. However, with the object categorization community's preoccupation with object detection, interest in perceptual grouping has waned. The reason for this is clear: the object-independent, mid-level shape priors that form the basis of perceptual grouping are subsumed by the object-dependent, high-level shape priors defined by a target object. As the recognition community moves from object detection back to object recognition, a linear search through a large database of target models is intractable, and perceptual grouping will be essential for sublinear scaling. We review two approaches to perceptual grouping based on grouping superpixels. In the first, we use symmetry to group superpixels into symmetric parts, and then group the parts to form structured objects. In the second, we use contour closure to group superpixels, yielding a figure-ground segmentation.

Keywords: perceptual grouping, superpixels, object categorization

1 Introduction

Perceptual grouping is a critical function in the human visual system, offering a powerful heuristic for grouping together causally related image features in support of both figure-ground segmentation and 3-D inference. In the mid-to-late 1990's, perceptual grouping was a thriving subcommunity in computer vision. However, over the past 10 years, there's been a steady decline in the number of perceptual grouping papers appearing in the computer vision community's main conferences. The reason for this is the reformulation of object recognition, historically cast as the problem of recognizing an object from a large database, as a detection problem, cast as the search for a particular target object.

The classical formulation of the object recognition problem, which defined the mainstream from the mid-1960's through to the late-1990's, was the recognition of an unexpected object from a database of objects. As illustrated in Figure 1(a), the feature extraction process began by extracting categorical or generic features, as the recognition community aspired to recognize categories, not exemplars. As far back as the seminal work of Roberts [16] in the mid-1960's, the recognition community understood that across the exemplars that belong to a category, shape is a more invariant property than appearance. As a result, the

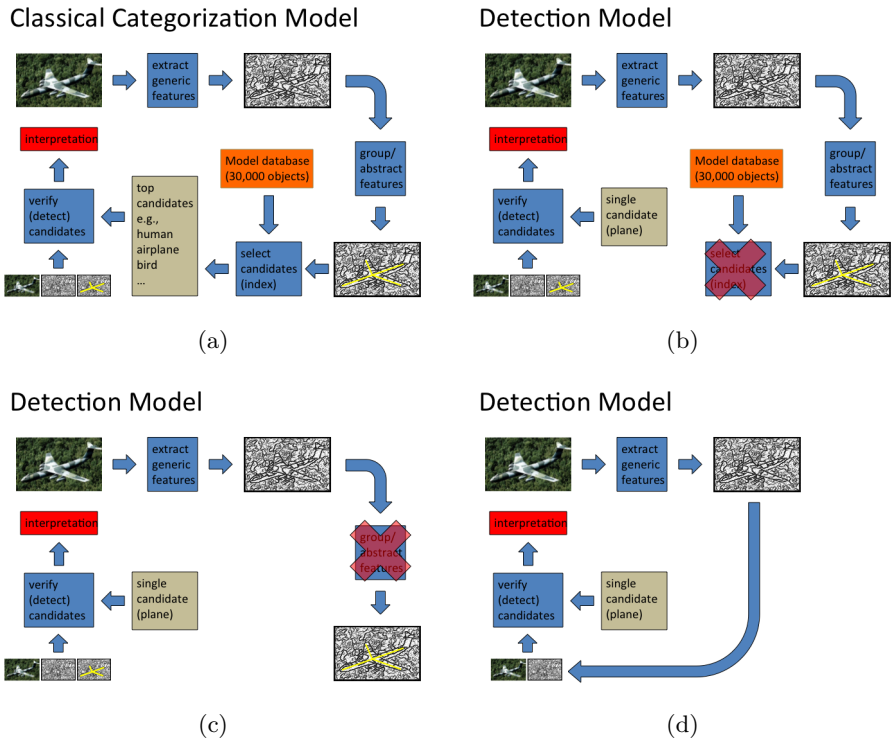

Fig. 1. The classical formulation of object recognition from a large database has given way to a more recent formulation of object recognition as target detection: (a) In the classical recognition model, the desire to extract shape features, considered more generic than appearance, began with edge detection. Because edgels were not discriminative, they were perceptually grouped and abstracted to form distinctive indexing structures that could prune a large database of objects down to a small number of promising candidates. (b) Over the past 10 years, the community has reformulated the recognition problem as object detection. Rather than verifying a number of candidates, the target candidate is known, rendering the process of indexing (or model selection) obsolete. (c) Without the need for domain-independent recovery, grouping, and abstraction of structure in order to prune a large database down to a small number of promising candidates, perceptual grouping is unnecessary. (d) As a result, verification (detection) can be applied directly to the ungrouped, low-level edge features.

majority of recognition systems from the mid-1960's to the late 1990's attempted to extract shape features, typically beginning with the extraction of edges, for at occluding boundaries and surface discontinuities, edges capture shape information. However, unlike today's distinctive local image features, e.g., SIFT [15], a local edgel carries very little information with which to index into a database of objects in an attempt to select a small number of promising object models that might account for the edgels.

The need for perceptual grouping in these early systems was critical, for only when the edgels were grouped into longer contours, perhaps parsed at high-curvature points, and grouped with other causally related contours, did distinctive indexing features emerge. Lowe's thesis [14] was the first to introduce computational models of perceptual grouping processes, e.g., proximity, collinearity, and parallelism, derived from image statistics. By grouping contour features into more distinctive groups (in Lowe's case, proximity followed by collinearity followed by parallelism), more discriminating indexing (using parallel lines instead of, say, triples of corners [7]) was possible. The more features were grouped, perhaps first into parts and then into multipart groups [6, 5], the more powerful the index and the fewer candidates that needed to be verified. Each candidate was verified, yielding a score (typically reflecting the degree to which a model could be aligned with image features), and the top-scoring candidate, if sufficiently strong, yielded the final interpretation.

The formulation of object recognition as the detection of a specific target object has dominated the recognition community over the past 10 years. As illustrated in Figure 1(b) and working backwards from the verification module, instead of having to verify a number of candidate object hypotheses, the detection problem identifies only a single hypothesis that needs to be verified (or detected). This, in turn, means that the indexing step, in which a large database of candidate objects is pruned down to a small set of candidates for verification, is superfluous, for the database effectively has a single object (target). Continuing to work our way backwards, as illustrated in Figure 1(c), if discriminative indexing features are not required to select promsing candidates, the perceptual grouping stage is also superfluous. Instead, as illustrated in Figure 1(d), the detector, i.e., verification, can be applied directly to the edgels, e.g., [4], to yield the final score, thereby short-circuiting the entire perceptual grouping process.

The existence of an object detector, representing a strong shape prior, eliminates the need for perceptual grouping, representing a much weaker, domain-independent shape prior. Unfortunately, as the categorization community moves from single object detection back to recognition from large databases, detection methods, typically formulated as template matching (or "sliding windows"), simply won't scale, and a linear search through thousands of templates is intractable, especially when an object can be viewed arbitrarily, it can articulate, and it can undergo significant within-class shape deformation. Verification (or detection) must be highly sublinear in the size of the database, demanding that discriminative indexing features be recovered *without knowledge of which object is being imaged*. Such domain-independent, bottom-up perceptual grouping is essential in the absence of an object prior.

In this paper, we briefly review our recent progress on two classical problems in perceptual grouping, each based on superpixels. We begin by describing a framework that first groups superpixels into symmetric parts, and then groups the symmetric parts into multipart structures [9]. Symmetry has played a prominent role in shape modeling for object recognition since the 2-D medial axis transform (MAT) of Blum [2] and the 3-D generalized cylinder (GC) of Binford

[1]. By detecting a set of symmetric parts and their attachments from a cluttered image of real objects, we recover a powerful shape index that can serve to prune a large database of objects down to a small number of promising candidates. In the second part of the paper, we address the classical problem of contour closure, i.e., finding a cycle of edgels in the image that separates figure from ground. We describe a framework that looks for groups of superpixels whose collective boundary has strong edgel support in the image [10]. The resulting shape boundary, or silhouette, can yield a structured, parts-based representation, e.g., [18], that can also be used to prune a large database down to a small number of promising candidates.

2 Symmetric Part Detection and Grouping

In [9], we introduced a novel approach to recovering the symmetric part structure of an object from a cluttered image, as outlined in Fig. 2. Drawing on the principle that a skeleton is defined as the locus of *medial points*, i.e., centers of maximally inscribed disks, we first hypothesize a sparse set of medial points at multiple scales by segmenting the image (Fig. 2(a)) into compact superpixels at different superpixel resolutions [11] (Fig. 2(b)). Superpixels are adequate for this task, balancing a data-driven component that's attracted to shape boundaries while maintaining a high degree of compactness. The superpixels (medial point hypotheses) at each scale are linked into a graph, with edges adjoining adjacent superpixels. Each edge is assigned an affinity that reflects the degree to which two adjacent superpixels represent medial points belonging to the same symmetric part (medial branch) (Fig. 2(c)). The affinities are learned from a set of training images whose symmetric parts have been manually identified. A standard graph-based segmentation algorithm applied to each scale yields a set of superpixel clusters which, in turn, yield a set of regularized symmetric parts (Fig. 2(d)).

In the second phase of our approach, we address the problem of perceptually grouping symmetric parts arising in the first phase. Like in any grouping problem, our goal is to identify sets of parts that are causally related, i.e., unlikely to co-occur by accident. Again, we adopt a graph-based approach in which the set of symmetric parts across all scales are connected in a graph, with edges adjoining parts in close spatial proximity (Fig. 2(e)). Each edge is assigned an affinity, this time reflecting the degree to which two nearby parts are believed to be physically attached. Like in the first phase, the associated, higher granularity affinities are learned from the regularities of attached symmetric parts identified in training data. Consequently, we explore two graph-based methods for grouping the detected parts. The first method is the same greedy approach that was used to cluster superpixels into parts. The second method employs parametric maxflow [8] to globally minimize an unbalanced normalized cuts criterion over the part graph. Both methods yield part clusters, each representing a set of regularized symmetric elements and their hypothesized attachments (Fig. 2(f)).

Our approach offers clear advantages over competing approaches. For example, classical multiscale blob and ridge detectors, such as [13] (Fig. 2(g)), yield

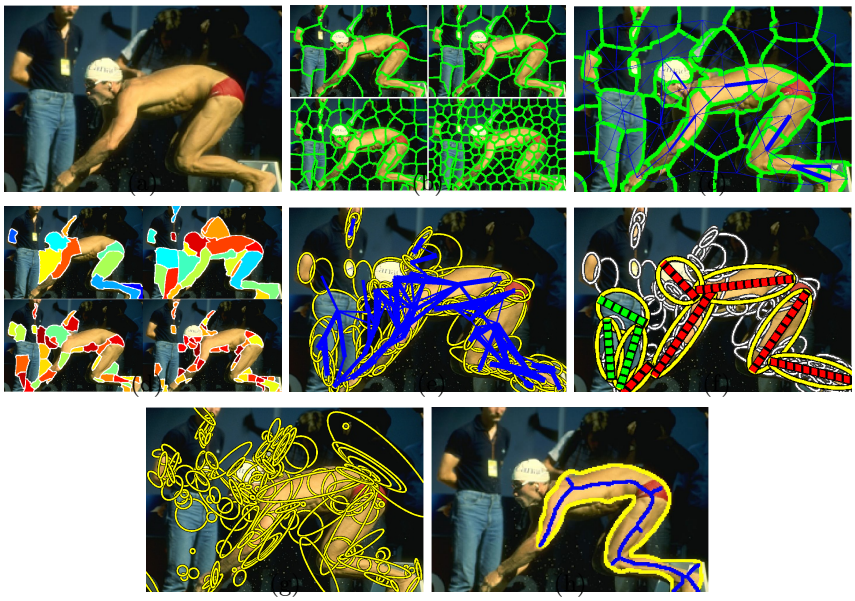

Fig. 2. Overview of our approach for multiscale symmetric part detection and grouping: (a) original image; (b) set of multiscale superpixel segmentations (different superpixel resolutions); (c) the graph of affinities shown for one scale (superpixel resolution); (d) the set of regularized symmetric parts extracted from all scales through a standard graph-based segmentation algorithm; (e) the graph of affinities between nearby symmetric parts (all scales); (f) the most prominent part clusters extracted from a standard graph-based segmentation algorithm, with abstracted symmetry axes overlaid onto the abstracted parts; (g) in contrast, a Laplacian-based multiscale blob and ridge decomposition, such as that computed by [13], shown, yields many false positive and false negative parts; (h) in contrast, classical skeletonization algorithms require a closed contour which, for real images, must be approximated by a region boundary. In this case, the parameters of the N-cuts algorithm [17] were tuned to give the best region (maximal size without region undersegmentation) for the swimmer. A standard medial axis extraction algorithm applied to the smoothed silhouette produces a skeleton (shown in blue) that contains spurious branches, branch instability, and poor part delineation.

many spurious parts, a challenging form of noise for any graph-based indexing or matching strategy. And even if an opportunistic setting of a region segmenter's parameters yields a decent object silhouette (Fig. 2(h)), the resulting skeleton may exhibit spurious branches and may fail to clearly delineate the part structure. From a cluttered image, our two-phase approach recovers, abstracts, and groups a set of medial branches into an approximation to an object's skeletal part structure, enabling the application of skeleton-based categorization systems to more realistic imagery. Details of the approach can be found in [9].

Some qualitative results are shown in Figure 3. Proceeding left to right, top to bottom, we see excellent part recovery and grouping for the starfish, the plane, the windmill, and the runner, respectively. In the case of the windmill, a

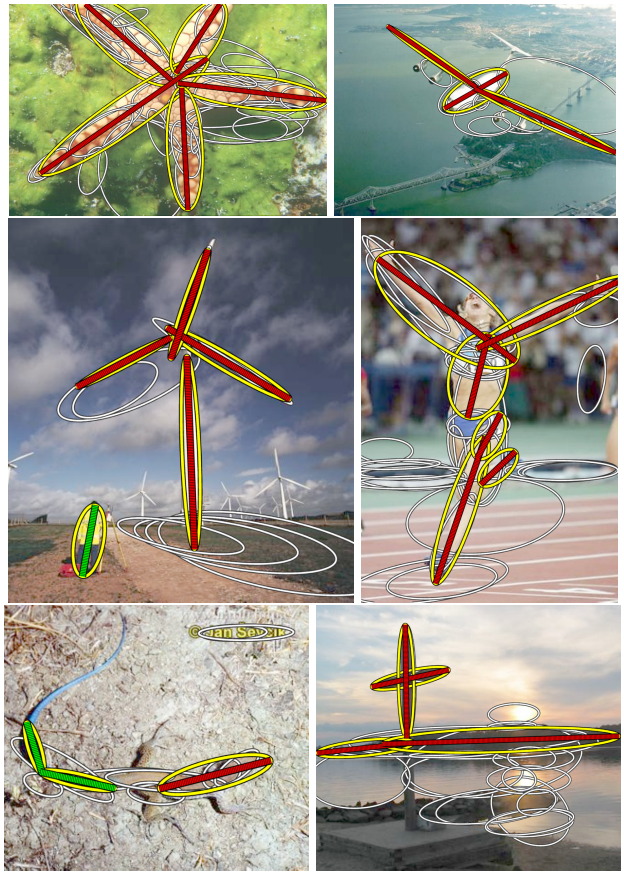

Fig. 3. Detected medial parts and their clusters.

second, singleton cluster, representing the entire body of the human, is recovered; however, the distant windmills are not recovered, for their scale is smaller than the smallest superpixel scale. The final two figures represent failure modes. In the case of the lizard, the curved symmetric tail is oversegmented into piecewise linear symmetric parts. In the case of the lake scene, the symmetric parts making up the horizon tree line are incorrectly grouped with the dock structure due to a lack of apparent occlusion boundary between the dock structure and the tree line parts.

3 Contour Closure

In this section, we review our framework for efficiently searching for optimal contour closure; details can be found in [10]. Fig. 4 illustrates an overview of our approach to computing contour closure. Given an image of extracted contours

(Fig. 4(a)), we begin by restricting contour closures to pass along boundaries of superpixels computed over the contour image (Fig. 4(b)). In this way, our first contribution is to reformulate the problem of searching for cycles of contours as the problem of searching for a subset of superpixels whose collective boundary has strong contour support in the contour image; the assumption we make is that those salient contours that define the boundary of the object (our target closure) will align well with superpixel boundaries. However, while a cycle of contours represents a single contour closure, our reformulation exploits a mechanism to encourage superpixel subsets that are spatially coherent.

Spatial coherence is an inherent property of a cost function that computes the ratio of perimeter to area. We modify the ratio cost function of Stahl and Wang [19] to operate on superpixels rather than contours, and extend it to yield a cost function that: 1) promotes spatially coherent selections of superpixels; 2) favors larger closures over smaller closures; and 3) introduces a novel, learned gap function that accounts for how much agreement there is between the boundary of the selection and the contours in the image. The third property adds cost as the number and sizes of gaps between contours increase. Given a superpixel boundary fragment (e.g., a side of a superpixel) representing a hypothesized closure component, we assign a gap cost that's a function of the proximity of nearby image contours, their strength, and their orientation (Fig. 4(c)). It is in this third property that our superpixel reformulation plays a second important role – by providing an appropriate scope of contour over which our gap analysis can be conducted.

In our third contribution, the two components of our cost function, i.e., area and gap, are combined in a simple ratio that can be efficiently optimized using parametric maxflow [8] to yield the global optimum. The optimal solution yields the largest set of superpixels bounded by contours that have the least gaps (Fig. 4(d)). Moreover, parametric maxflow can be used to yield the top k solutions (see [3], for example). In an object recognition setting, generating a small set of such solutions can be thought of as generating a small set of promising shape hypotheses which, through an indexing process, could invoke candidate models that could be verified (detected). The use of such multiscale hypotheses was shown to facilitate state-of-the-art object recognition in images [12].

In Figure 5, we illustrate results of our superpixel closure (SC) method. In the case of the carriage, swimmer, plane, golfer, baseball player, plane, and spider, we see that the algorithm nearly correctly segments figure from background, and is able to capture the deep concavities of the object, which is particularly visible with the spider. In the case of the horse, elephant, and giraffe, we see evidence of undersegmentation due to the properties of the objective function that we're optimizing. In each case, there are false boundaries (e.g., horizon) that can increase the area of the figure without introducing additional gap. In other words, if the algorithm can follow a gap-free contour that yields a larger area, e.g., following the contour between ground and sky in the giraffe image, it will do so, yielding a bias towards compact objects.

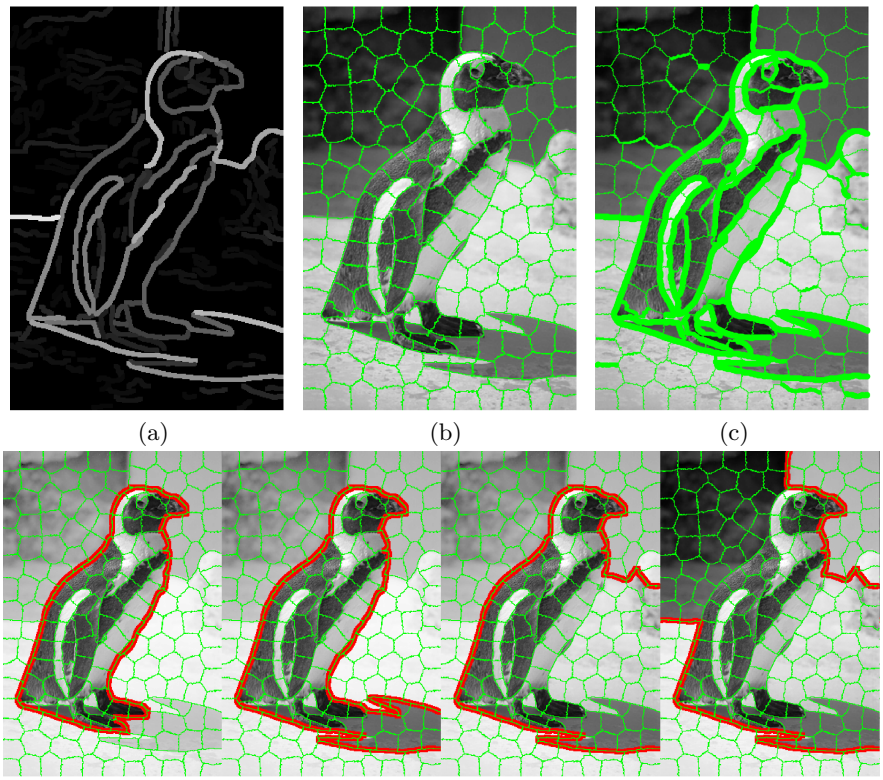

Fig. 4. Overview of our approach for image closure: (a) contour image – while we take as input only this contour image, we will overlay the original image in the subsequent figures to ease visualization; (b) superpixel segmentation of contour image, in which superpixel resolution is chosen to ensure that target boundaries are reasonably well approximated by superpixel boundaries; (c) a novel, learned measure of gap reflects the extent to which the superpixel boundary is supported by evidence of a real image contour (line thickness corresponds to the amount of agreement between superpixel boundaries and image contours); (d) our cost function can be globally optimized to yield the largest set of superpixels bounded by contours that have the least gaps. In this case the solutions, in increasing cost (decreasing quality), are organized left to right.

4 Conclusions

The perceptual grouping of contours according to symmetry and closure has been long been a problem of interest to human and computer vision researchers. However, both problems have traditionally been solved by first extracting contours and then grouping the contours, leading to prohibitive combinatorial complexity. We have explored both these problems from the dual standpoint of region-based processing, where regions are compact superpixels that minimize

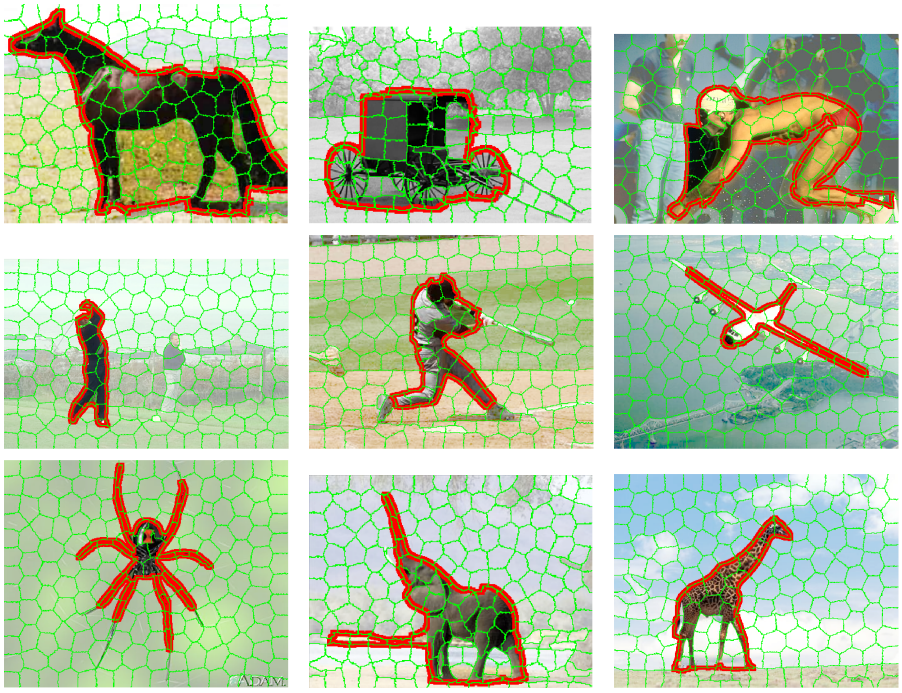

Fig. 5. Example Results of Superpixel Closure

undersegmentation. On the case of symmetry-based grouping, the superpixels represent deformable maximally inscribed disks (medial points), and we learn to group them when they belong to the same symmetric part. In the case of closure-based grouping, the superpixels represent "chunks" of boundary, and when the right subset of superpixels is found, those chunks of boundary will form a closure with minimal gap. The perceptual grouping of superpixels in both cases yields discriminative shape structures that can support effective indexing, as the community moves from detection back to recognition from large databases.

5 Acknowledgements

This research was sponsored, in part, by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0060. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either express or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein. This research was also sponsored by NSERC.

References

1. T. O. Binford. Visual perception by computer. In *Proceedings, IEEE Conference on Systems and Control*, Miami, FL, 1971.
2. H. Blum. A Transformation for Extracting New Descriptors of Shape. In Weiant Wathen-Dunn, editor, *Models for the Perception of Speech and Visual Form*, pages 362–380. MIT Press, Cambridge, 1967.
3. J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.
4. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.
5. S. Dickinson, A. Pentland, and A. Rosenfeld. From volumes to views: An approach to 3-D object recognition. *CVGIP: Image Understanding*, 55(2):130–154, 1992.
6. S. Dickinson, A. Pentland, and A. Rosenfeld. 3-D shape recovery using distributed aspect matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):174–198, 1992.
7. Daniel P. Huttenlocher and Shimon Ullman. Recognizing solid objects by alignment with an image. *Int. J. Comput. Vision*, 5(2):195–212, 1990.
8. V. Kolmogorov, Y.Y. Boykov, and C. Rother. Applications of parametric maxflow in computer vision. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007.
9. A. Levinshtein, S. Dickinson, and C. Sminchisescu. Multiscale Symmetric Part Detection and Grouping. In *IEEE International Conference on Computer Vision*, September 2009.
10. A. Levinshtein, C. Sminchisescu, and S. Dickinson. Optimal contour closure by superpixel grouping. In *ECCV*, pages 480–493, 2010.
11. Alex Levinshtein, Adrian Stere, Kiriakos N. Kutulakos, David J. Fleet, Sven J. Dickinson, and Kaleem Siddiqi. Turbopixels: Fast superpixels using geometric flows. *PAMI*, 31(12):2290–2297, 2009.
12. F. Li, J. Carreira, and C. Sminchisescu. Object Recognition as Ranking Holistic Figure-Ground Hypotheses. In *CVPR*, June 2010.
13. Tony Lindeberg and Lars Bretzner. Real-time scale selection in hybrid multi-scale representations. In *Scale-Space*, volume 2695 of *Springer LNCS*, pages 148–163, 2003.
14. D. G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Norwell, MA, USA, 1985.
15. David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
16. L. Roberts. Machine perception of three-dimensional solids. In J. Tippett et al., editors, *Optical and Electro-Optical Information Processing*, pages 159–197. MIT Press, Cambridge, MA, 1965.
17. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
18. Kaleem Siddiqi, Ali Shokoufandeh, Sven J. Dickinson, and Steven W. Zucker Y. Shock graphs and shape matching. *International Journal of Computer Vision*, 35:13–32, 1999.
19. J.S. Stahl and Song Wang. Edge grouping combining boundary and region information. *IEEE Transactions on Image Processing*, 16(10):2590–2606, Oct. 2007.