The Evolution of Object Categorization and the Challenge of Image Abstraction

Sven Dickinson
Department of Computer Science
University of Toronto

1 Introduction

In 2004, I was a guest at the Center for Machine Perception at the Czech Technical University. During my visit, a graduate student was kind enough to show me around Prague, including a visit to the Museum of Modern and Contemporary Art (Veletržní Palác). It was there that I saw the sculpture by Karel Nepraš entitled "Great Dialogue," a photograph of which appears in Figure 1. The instant I laid eyes on the sculpture, I recognized it as two humanoid figures seated and facing each other; when I've presented a 2-D image (Figure 1) of the sculpture to classroom students and seminar audiences, their recognition of the two figures was equally fast. What's remarkable is that at the level of local features (whether local 2-D appearance or local 3-D structure), there's little, if any, resemblance to the features constituting real 3-D humans or their 2-D projections. Clearly, the local features, in terms of their specific appearance or configuration, are irrelevant, for individually they bear no causal relation to humans. Only when such local features are grouped, and then *abstracted*, do the salient parts and configuration begin to emerge, facilitating the recognition of a previously unseen exemplar object (in this case, a very distorted statue of a human) from a known category (humans).

The process of image (or feature) abstraction begins with the extraction of a set of image features over which an abstraction can be computed. If the abstraction is parts-based (providing the locality of representation required to support object recognition in the presence of occlusion and clutter), the local features must be perceptually grouped into collections that map to the abstract parts. For the features to be groupable, non-accidental relations [152] must exist between them. While such relations could be appearance-based, such as color and texture affinity, appearance is seldom

Figure 1: The two shapes depicted in this statue clearly represent two humanoid figures seated and facing each other. At the level of local features, the figures are unrecognizeable. However, at a more abstract level, the coarse parts of the figures begin to emerge which, along with their relations, facilitate object categorization. The local features that constitute the abstract parts were not learned from training examples (they don't exist on a real human), nor were they grouped/abstracted using a prior target (human) model. This sculpture by Karel Nepraš, entitled "Great Dialogue," is found in the Museum of Modern and Contemporary Art (Veletržní palác), in Prague; image reproduced with permission.

generic to a category. Had the statue been painted a different color or textured with stripes or spots, for example, recognition would have been unaffected. Clearly, we require more powerful grouping cues that reflect the shape regularities that exist in our world – cues that have long been posited by the perceptual organization community [131, 265, 42, 43].

The ability to group together shape-based local features, such as contours or regions, is an important first step that has been acknowledged by shape-

based object recognition researchers since the 1960's [198]. However, the grouping of causally (i.e., non-accidentally) related features is necessary but not sufficient for object categorization. Returning to Figure 1, the grouping of the various local features that make up the torso of one of the figures is indeed an extremely challenging and important problem. Having recovered and grouped a set of salient shape features, a typical recognition system would proceed to establish one-to-one correspondence between salient image features (in the grouping) and salient model features. But herein lies the problem. Assuming a one-to-one correspondence between local image features, such as points, patches, contours, or even regions, constrains the model to be little more than a template of the image.

The true correspondence between the collection of local features making up the torso and the torso "part" on any intuitive model of a human lies not at the level of local image features but at a more abstract level of shape features. For example, one such abstraction of the seated human model is shown in Figure 2, which includes an elliptical part corresponding to the torso.[1] Under a one-to-one correspondence assumption, the myriad local features making up the statue torso (including many long, "salient" contours) must be abstracted before correspondence with the model torso can be established. It is important to note that this abstraction does not live explicitly in the image, i.e., it is not simply a subset of the grouped image features. And while such an abstraction clearly requires a model (in this case, an elliptical shape "prior"), the model assumes no object- or scene-level knowledge.

The problem of abstraction is arguably the most important and most challenging problem facing researchers in object categorization. This is not a new problem, but one which was far more commonly acknowledged (but no more effectively solved) by early categorization researchers whose models captured object shape at high levels of abstraction. Over the last four decades, our inability to effectively recover such abstractions from real images of real objects has led us to increasingly specific object recognition domains that require little or no abstraction. Understanding this evolution not only brings the abstraction problem into focus, but helps to identify the many important contributions made by categorization researchers over the last four decades.

[1] This is not meant to imply that the abstraction process is necessarily 2-D. Many, including Biederman [27] and Pizlo [184], would argue that such abstraction is 3-D. In that case, the ellipses in Figure 2 might be interpreted as the projections of ellipsoids.

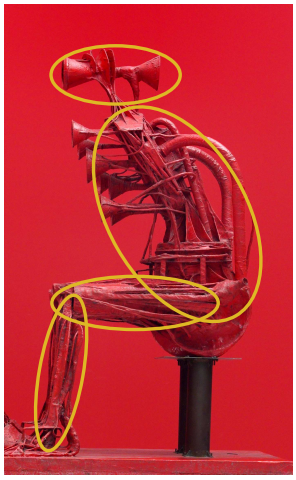


Figure 2: A shape abstraction of the seated humanoid on the left in Figure 1. Note that the boundaries of the shape abstraction do not map one-to-one to (or align well with) local features (e.g., contours) in the image.

2 Avoiding the Abstraction Problem: A Historical Trend

The evolution of object recognition over the past 40 years has followed a very clear path, as illustrated in Figure 3. In the 1970's, the recognition community focused on generic (alternatively, prototypical, categorical, or coarse) 3-D shape representations in support of object categorization. Objects were typically modeled as constructions of 3-D volumetric parts, such as generalized cylinders (e.g., [29, 2, 169, 45]), superquadrics (e.g., [176, 91, 229, 107, 238, 143, 144]), or geons (e.g., [27, 72, 74, 73, 24, 191, 40]). Figure 4 illustrates an example output from Brooks' ACRONYM system, which recognized both categories and subcategories from the constraints on the projections of generalized cylinders and their relations. The main challenge facing these early systems was the *representational gap* that existed between the low-level features that could be reliably extracted, and the abstract nature of the model components. Rather than addressing this representational gap through the development of effective abstraction mechanisms, the community effectively eliminated the gap by bringing the images closer to the models. This was accomplished by removing object surface markings and structural detail, controlling lighting conditions, and reducing scene clutter. Edges in the image could then be assumed to map

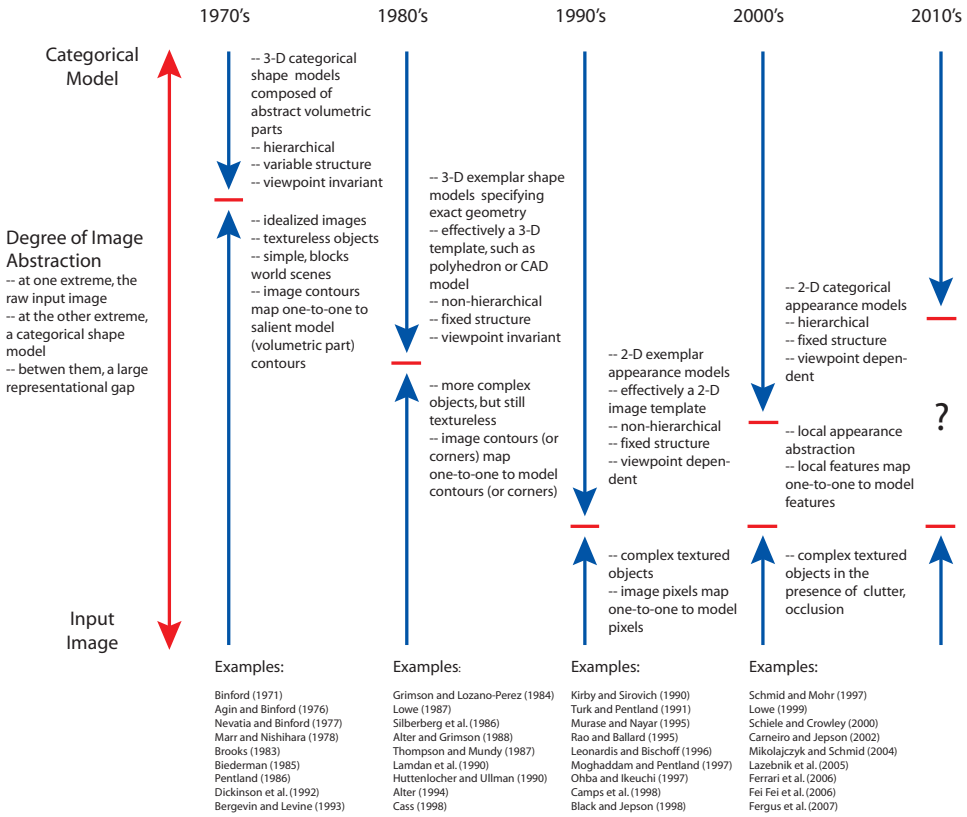
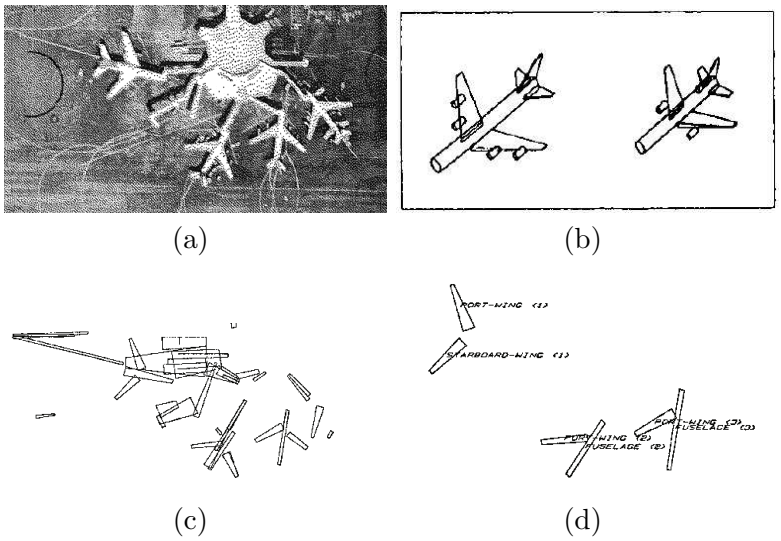


Figure 3: The evolution of object categorization over the past four decades (see text for discussion).

directly (one-to-one) to the occluding boundaries (separating figure from background) and surface discontinuities of the high-order volumetric parts making up the models.

The results left many unsatisfied, as the images and objects were often contrived (including blocks world scenes), and the resulting systems were unable to deal with real objects imaged under real conditions. Nevertheless, some very important principles emerged in the 1970's, many of which are being rediscovered by today's categorization community:

1. the importance of shape (e.g., contours) in defining object categories;
2. the importance of viewpoint-invariant, 3-D shape representations;
3. the importance of symmetry and other non-accidental relations in fea-



(a) (b) (c) (d)

Figure 4: Brooks' ACRONYM system [45] recognized 3-D objects by searching for the projections of their volumetric parts and relations: (a) input image; (b) 3-D models composed of generalized cylinders; (c) extracted ribbons from extracted edges; and (d) recognized objects (images courtesy of Rod Brooks).

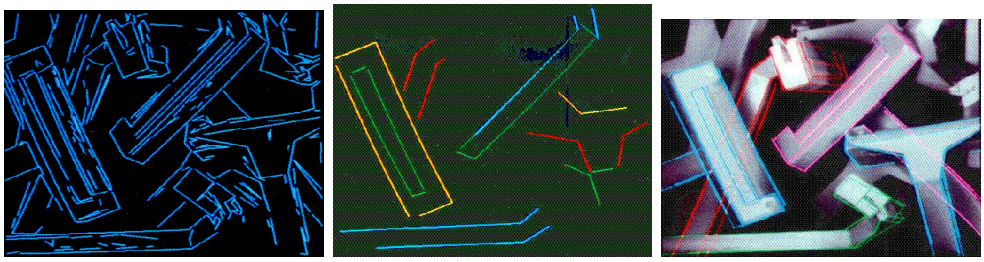
ture grouping;

4. the need for distributed representations composed of sharable parts and their relations to help manage modeling complexity, to support effective indexing (the process of selecting candidate object models that might account for the query), to support object articulation, and to facilitate the recognition of occluded objects;

5. the need for hierarchical representations, including both part/whole hierarchies as well as abstraction hierarchies;

6. the need for scalability to large databases, i.e., the "detection" or target recognition problem (as it was then known) is but a special case of the more general recognition (from a large database) problem, and a linear search (one detector per object) of a large database is unacceptable;

7. the need for variable structure, i.e., the number of parts, their identities, and their attachments may vary across the exemplars belonging to a category.



(a) (b) (c)

Figure 5: Lowe's SCERPO system [152] used perceptual grouping to prune hypothesized correspondences between image contours and polyhedral edges: (a) extracted edges; (b) extracted perceptual groups; and (c) detected objects and their poses (images courtesy of David Lowe).

The 1980's ushered in 3-D models that captured the exact shape of an object. Such models, inspired by CAD models, were effectively 3-D templates, e.g., [106, 225, 116, 152, 153, 240, 60, 5, 57, 61]. Figure 5 illustrates an example output from Lowe's SCERPO system, which recognized a 3-D polyhedral template of an object from non-accidental groupings of features comprising its projection. Provided that such models could be acquired for a real object (requiring considerable overhead), the community found that it could build object recognition systems capable of recognizing real (albeit restricted) objects – a very important development indeed. While object models were still viewpoint-invariant (since they were 3-D), hierarchical representations became less common as the models became less coarse-to-fine. This time, the representational gap was eliminated by bringing the model closer to the imaged object, requiring the model to capture the exact geometry of the object. Moreover, since the presence of texture and surface markings seriously affected the search complexity of these systems, once again the objects were texture-free, so that a salient image edge mapped to (for example) a polyhedral edge. Again, there was dissatisfaction, as the resulting systems were unable to recognize complex objects with complex surface markings. Moreover, the overhead required to construct a 3-D model, either by hand or automatically from image data, was significant.

It is important to note that while both the above generations of systems assumed a one-to-one correspondence between salient image features and model features, there was a dramatic redefinition of the problem from category recognition to exemplar recognition. In earlier systems, the bottom-up recovery of high-level volumetric parts and their relations, forming powerful

indexing structures, meant that models could accommodate a high degree of within-class shape variation. However, as the scope of indexing structures later retreated to individual lines, points, or small groups thereof, their indexing ambiguity rose dramatically, and extensive verification was essential to test an abundance of weak model hypotheses. The need for 3-D model alignment, as a prerequisite for verification, required that models were essentially 3-D templates that modeled the shape of an exemplar rather than a category (although some frameworks supported the articulation of rigid parts). Still, at the expense of backing down from the more challenging categorization problem, recognition had begun to penetrate real industrial domains, providing real solutions to real problems.

Most object recognition systems up to this point employed 3-D models and attempted to recognize them in 2-D images (3-D from 2-D). However, a number of researchers, e.g., [102, 23, 213, 251, 47, 263, 75, 20], began to study the invariant properties of views and their application to view-based 3-D object recognition (2-D from 2-D). Inspired by the early aspect graph work of Koenderink and van Doorn [129], a large community of researchers began to explore the properties of aspect graphs in support of view-based object recognition [118, 132, 185, 76, 206, 233, 74, 73, 79, 70, 77, 101, 100, 217, 230]. While view-based methods were gaining momentum, they still lagged behind the 3-D from 2-D methods, which were now shifting toward the use of geometric invariants to enable recognition from larger object databases [136, 165, 94].

In the early 1990's, a number of factors led to a major paradigm shift in the recognition community, marking the decline of 3-D shape models in favor of appearance-based recognition. Faster machines could now support the high throughput needed to accommodate the multitude of image templates required to model a 3-D object. Moreover, no 3-D modeling (including software and trained personnel) was required for model acquisition; a mere turntable and camera would suffice. More importantly, by focusing on the explicit pixel-based appearance of an object, the complex, error-prone problem of segmentation could be avoided. For the first time, recognition systems were constructed that could recognize arbitrarily complex objects, complete with texture and surface markings, e.g., [128, 250, 166, 193, 142, 162, 170, 49, 32]). Figure 6 illustrates an example output from Murase and Nayar's appearance-based (view-based) 3-D object recognition system, which used PCA and nearest-neighbor search to drastically reduce the complexity of image correlation over a large database.

This time, the representational gap was eliminated by bringing the models all the way down to the image, yielding models that were images them-

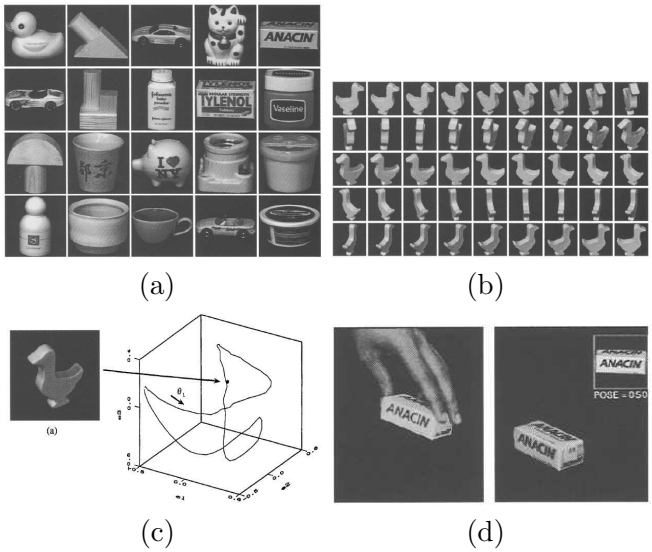
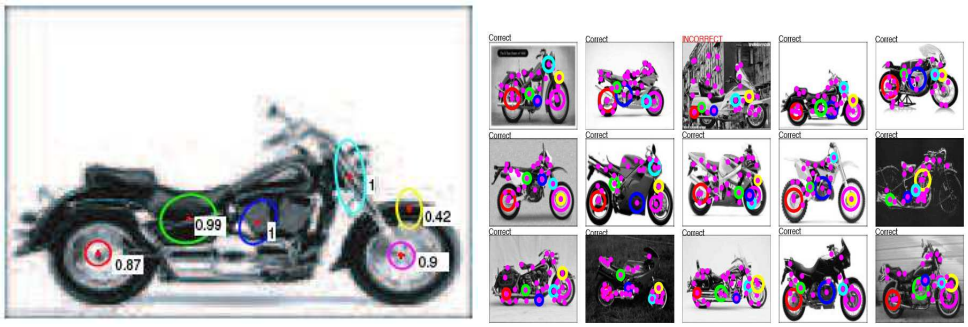


Figure 6: Murase and Nayar's appearance-based (view-based) recognition system [166]: (a) a database of objects; (b) a dense set of views is acquired for each object; (c) the views trace out a manifold in low-dimensional space, with each view lying on the manifold; (d) recognizing a query object (images reproduced from [166] with permission of the *International Journal of Computer Vision*, Springer).

selves. The resulting systems could therefore recognize only exemplar objects – specific objects that had been seen at training time. Despite a number of serious initial limitations of this approach, including difficulties in dealing with background clutter, illumination change, occlusion, translation, rotation, and scaling, the approach gained tremendous popularity, and some of these obstacles were overcome [142, 49, 21, 141, 19]. But the templates were still global, and invariance to scale and viewpoint could not be achieved.

To cope with these problems, the current decade (2000's) has seen the appearance model community turn to the same principles adopted by their shape-based predecessors: a move from global to local representations (parts), and the use of part representations that are invariant to changes in translation, scale, image rotation, illumination, articulation, and viewpoint, e.g., [154, 155, 261, 262, 50, 1, 53, 52, 161, 137, 130, 210]. While early systems characterized collections of such features either as overly rigid geometric configurations or, at the opposite extreme, as unstructured "bags", later systems, e.g., [209, 256, 51, 52, 89, 90, 82, 87, 189], added pairwise spatial



(a) (b)

Figure 7: Learning scale-invariant parts-based models from examples (Fergus et al. [87]): (a) Learned motorcycle model with ellipses representing part covariances and labels representing probability of occurence; (b) example model detections in query images, with colored circles representing matched part hypotheses (images reproduced from [87] with permission of the *International Journal of Computer Vision*, Springer).

constraints, again drawing on classical shape modeling principles from the 1970's and 1980's. For example, Figure 7 illustrates the system of Fergus et al. [87], in which a scale-invariant, parts-based object model is learned from a set of annotated training examples and is used to detect new instances of the model in query images. Unlike the 1970's and 1980's, today's systems are applied to images of cluttered scenes containing complex, textured objects. Yet something may have been lost in our evolution from shape to appearance, for today's appearance-based recognition systems are no more able to recognize yesterday's line drawing abstractions than were yesterday's systems able to recognize today's images of real objects.

Like the 1990's, today's models have been brought close to the image. But this trend is clearly reversing and starting to swing back. And unlike the previous three decades, the representational gap has not been completely eliminated. The scope of a local feature has expanded from a single pixel to a scale-invariant patch. Moreover, the patch representation encodes not the explicit pixel values, but rather a weak abstraction of these values (e.g., the gradient histograms found in SIFT [155] or the radial distribution of mass found in Belongie et al.'s shape context [22]). The increased level of abstraction offered by these local features supports an increased amount of within-class variation of a category's appearance. This proved to be sufficient to handle some restricted categories whose exemplars do indeed share

the same local features. Such categories, including cars, faces, people, and motorcycles, can be characterized as geometrically regular configurations of recurring, distinctive, local features. However, such categories are likely to be the exception rather than the rule, for local features are seldom generic to a shape category. In fact, for most categories, it's quite possible for two exemplars to not share a single local appearance-based feature.

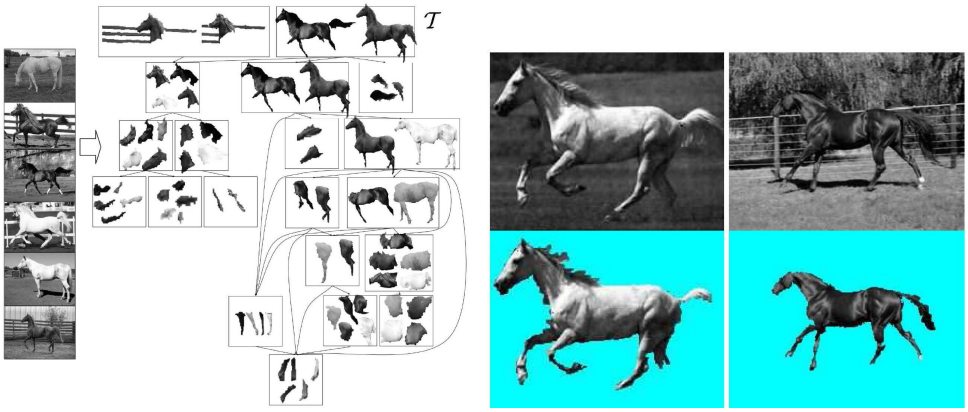
If one extrapolates this upward trajectory in (decreasing) feature specificity, one might first predict a return to those image contours that encode the shape (occluding boundaries or surface discontinuities) of an object – features that are far more generic to a category than appearance.[2] Yet the cost of more generic features is their increased ambiguity, for a small fragment of contour (e.g., resulting from a curve partitioning process that parses contours at curvature discontinuities or inflections) carries very little category-specific information. As proposed decades earlier, the solution lies in grouping together causally related, nearby contours into more distinctive structures.

How distinctive depends entirely on the problem. In a detection (or target recognition) task, for which model selection is provided, the need for complex, bottom-up contour grouping to yield distinctive indexing structures is absent in the presence of a strong template; rather, only minimal grouping is required to test a particular model. This is precisely the approach taken in recent work, e.g., [167, 172, 87, 145, 88], which builds relational models of contour fragments in support of object detection. However, in a more general recognition task, more ambitious domain-independent grouping is essential, which clearly introduces additional complexity. To help manage this complexity, feature hierarchies have re-emerged, in combination with powerful learning tools, to yield exciting new categorization frameworks [7, 6, 179, 41, 241, 92, 171, 4, 242, 273].[3] Figure 8 illustrates the system of Todorovic and Ahuja [242], in which a region-based hierarchical object model is learned from training examples and used to detect new instances of the model in query images.

But what of the more general categorization problem of recognition from a large database? Continuing our trajectory of working with image contours, we will have to group them into larger, more distinctive indexing structures

[2] In all fairness, appearance-based methods (based on explicit pixel values) implicitly encode both shape and non-shape information, but cannot distinguish between the two. Hence they are less invariant to changes in appearance when shape is held constant.

[3] In fact, Tsotsos [247, 248, 249] proved that such hierarchies are essential for managing the complexity of visual recognition.



(a) (b)

Figure 8: Learning hierarchical segmentation tree-based models from examples (Todorovic and Ahuja [242]): (a) learned hierarchical tree-union model (right) from examples (left), capturing the recursive containment and spatial layout of regions making up the model; (b) example model detections (below) in query images (above) (images reproduced from [242], Copyright ©2008 IEEE).

that can effectively prune a large database down to a few candidates.[4] If we want our models to be articulation invariant, then our indexing structures will map naturally to an object's parts. And if we want to reduce the dimensionality of the parts to allow part sharing across categories, then we somehow have to boost the power of our indexing structures to offset the increased ambiguity of our parts. That means grouping parts together until the resulting indexing structures are sufficiently powerful. Interestingly enough, this is exactly the original framework proposed in the 1970's, meaning that if our prediction holds, we will have come full circle. If we do revisit this paradigm, we will do so with vastly faster machines, more powerful inference and search algorithms, and a desire to learn representations rather than handcraft them. But has this convergence of machine learning and object categorization led to deeper representational insight?

The trend over the last four decades is clear. Rather than developing mechanisms for image and shape abstraction that are required to bridge the

[4]Indexing can take many forms, including hashing, e.g., [136, 93, 94], e.g., decision trees (including kd-trees) [118, 102, 20, 214, 215], and coarse-to-fine model hierarchies, e.g., [45]. All assume that the query object is unknown and that a linear search of the database is unacceptable (or intractable).

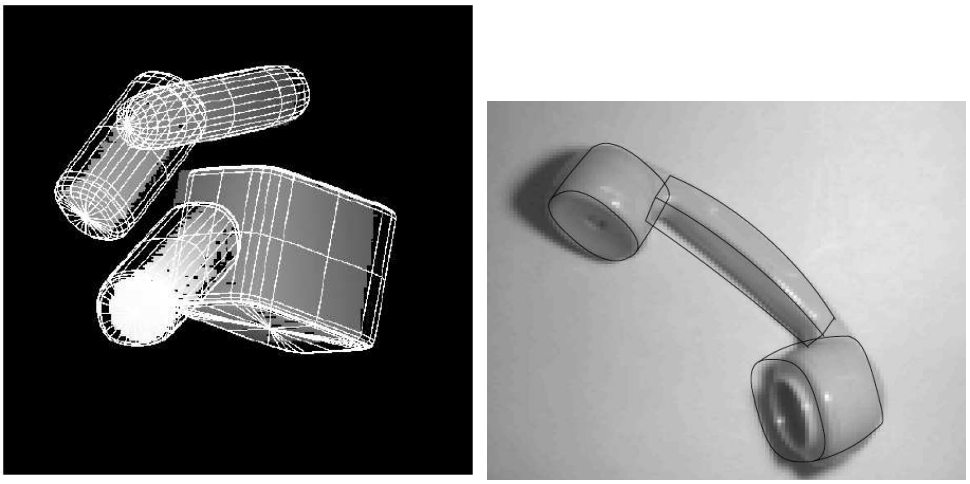
representational gap between our favorite "salient" image features and true categorical models, we have consistently and artificially eliminated the gap, originally by moving the images up the abstraction hierarchy (simulating the abstraction) and later by moving the models down the abstraction hierarchy (making them less categorical). Driven by a desire to build recognition systems that could solve real problems, the evolution of recognition from category to exemplar was well-motivated. But the community is clearly headed back toward categorization. And although our models are slowly creeping back up the abstraction hierarchy, image features are still tightly coupled to model features, and the critical problem of abstraction continues to receive little attention. Until this important problem is addressed, progress in more general categorization seems unlikely.

3 The Abstraction of Shape

In the 1970's, there was no shortage of abstract shape representations. For example, Binford's generalized cylinder (GC) [29] (see Figure 4) was a powerful, symmetry-based part model whose complexity was unbounded.[5] To manage the complexity of bottom-up shape recovery, a number of restrictions were introduced that arguably strengthened the fundamental role of symmetry. Such restrictions included, for example, a straight axis, a homogeneous sweep function, a linear sweep function, or a rotationally symmetric cross-section [169, 2, 159, 45, 253, 272, 188, 31, 133, 186, 160, 30, 157, 59]. While an abstract object model composed of restricted generalized cylinders and their spatial relations could support powerful categorization, recovering such parts and relations from images of real objects was the stumbling block. When image contours mapped one-to-one to the occluding boundaries or surface discontinities of restricted GC's, such recovery was indeed possible. However, when the projected model contours were not a subset of the observed image contours, part recovery was not possible. Abstraction mechanisms for mapping observed image contours to abstract model contours were simply not available at that time.

In the 1980's, two powerful symmetry-based, volumetric shape abstractions emerged, still founded on the symmetry axis concept, but each taking a very different approach to restricting the complexity of the generalized cylinder. Superquadric ellipsoids [17, 176, 229, 107, 91, 143, 144, 71] provided a rich set of deformations with a small set of parameters. While most

[5]The generalized cylinder is defined by axis, cross-section, and sweep functions, each of which can be arbitrarily complex.



(a) (b)

Figure 9: Two powerful 3-D shape abstractions evolved from the generalized cylinder (GC), each retaining a powerful symmetry property and each restricting the GC in different way: (a) superquadric ellipsoids [17] were typically recovered from 3-D range data (from Leonardis et al. [144], Copyright ©1997 IEEE); (b) geons [27] were typically recovered from 2-D image data (from Pilu and Fisher [181]; image reproduced with permission of Springer).

successful superquadric ellipsoid recovery lived in the range data domain (see Figure 9(a)), where a surface model could often be abstracted from a cloud of 3-D shape points, their recovery from 2-D images was far more challenging and far less successful, again lacking the abstraction mechanisms to map observed image contours to abstract model contours [69].

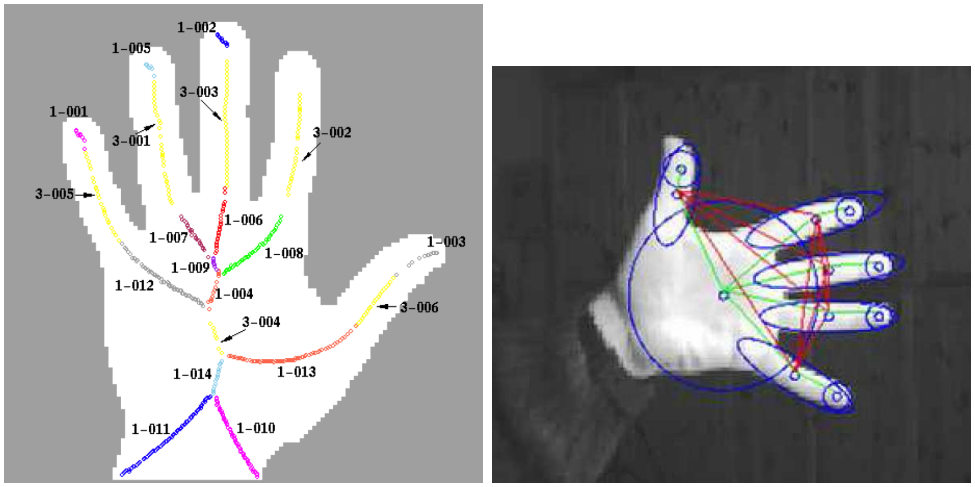
Biederman's geons [27, 28] represented a qualitative partitioning of the space of GC's according to simple dichotomous and trichotomous properties that humans could distinguish effortlessly. While this psychological theory launched a subcommunity of categorization researchers to develop computational models for geon recovery [25, 24, 191, 192, 268, 66, 181] (see Figure 9(b)), including more general qualitative volumetric part models [72, 74, 73, 67], they again faced the same challenge as their GC ancestors: salient image contours do not necessarily map one-to-one to abstract model contours. While proponents of GC's, restricted GC's, superquadric ellipsoids, and geons were well motivated in attempting to model an object's abstract parts and their relations, their assumption that the features comprising these models could be directly observed in the image was unrealistic.

Instead of pursuing the abstraction mechanisms that would allow such modeling frameworks to be further explored, the frameworks were abandoned due to their inability to recognize real images of real objects.

Blum's medial axis transform (MAT) [35, 36, 37] is a 2-D axial symmetry-based shape description which, like Binford's generalized cylinder which followed it, spawned an entire shape subcommunity. Just as geons imposed a qualitative partitioning of the space of generalized cylinders, shock graphs [223] (see Figure 10(a)) imposed a qualitative partitioning on the branches of the medial axis transform. And just as geons inspired a community of geon-based recognition systems, shock graphs inspired a community of shock graph-based recognition systems [223, 174, 212], while 3-D medial surfaces (analogous to medial axes) led to medial surface graph-based recognition systems [224] (the mathematics, algorithms, and applications of medial representations are detailed in [222]). However, just as GC-based systems assumed that salient contours in the image map one-to-one to contours generated by the (GC) model, the medial axis-based community assumed that salient contour points (i.e., *all* points on an object's silhouette) map one-to-one to points generated by the (MAT) model. For either framework to succeed on real images of real objects, image abstraction must yield a set of abstract contours which, in turn, map to the features of an abstract categorical model.

One might ask to what extent abstract shape recovery can be purely bottom-up. Part models like the unrestricted GC in 3-D and a medial branch in 2-D impose no mid-level shape priors (i.e., shape constraints) to help regularize their recovery from real images of real objects. As a consequence, too much stock is placed in features arising from simple bottom-up segmentation, such as contours or regions, and the assumption that they map one-to-one to salient model features. Looking back at the geon-based recognition systems, geons provided a powerful set of regularizing constraints on the data, but were never used as the basis for an image abstraction process. Shock graphs offered a similar set of constraints, but have also not yet been effectively used as the basis for image abstraction, although there have been efforts to regularize, in a bottom-up sense, the MAT [8, 236, 255, 83, 10].

The mid-level shape prior of symmetry has been used, albeit with limited success, as a basis for such image abstraction. Multi-scale blobs, including the work of Crowley [62, 63], Lindeberg [148, 149], Blostein and Ahuja [34], and Shokoufandeh et al. [220, 218] all employ ridge and/or blob models (see Figure 10(b)) as symmetry-based mid-level part constraints. While the models provide excellent regularization and have led to powerful hierarchical shape representations for recognition, they have not been successfully recov-



(a) (b)

Figure 10: Two powerful 2-D qualitative shape abstractions: (a) the shock graph (from Siddiqi et al. [223]), whose parts represent a qualitative partitioning of Blum's medial axis transform [35], and whose edges span adjacent parts, directed from larger to smaller (image reproduced from [223] with permission of the *International Journal of Computer Vision*, Springer); (b) the blob graph (from Shokoufandeh [218]), whose parts capture elongated symmetric structure at different scales, and whose edges capture parent-child (topological=green) and sibling (geometric=red) relationships between parts (image reproduced from [218] with permission of *Computer Vision and Image Understanding*, Elsevier).

ered from textured objects. Moreover, it's not clear whether they provide a rich enough shape description, for unlike shock graphs or geons, the parts cannot bend or taper.

Symmetry is an invariant in all these approaches, and has its roots in Gestalt psychology. The above symmetry-based abstractions (superquadrics, geons, and shock graphs) represent but a small fraction of a much broader community working with symmetry-based shape models for object categorization, both in 3-D, e.g., [164, 177, 178, 239], and in 2-D, e.g., [147, 200, 211, 95, 274, 3, 126, 271, 9]. While symmetry provides a basis for perceptual grouping of image structure, such as contours, it's important to realize that in general (at least for real images of real objects), the groups may not be in a form suitable for direct matching to an abstract model. Rather, the groups must be abstracted and regularized to yield a set of abstract features

that only then can be matched to an abstract model. It is highly unlikely for such features to exist explicitly in the image; rather, they must be inferred from appropriate groupings of local features that we can extract. For example, in Figure 2, the elliptical contour defining any of the part abstractions does not exist explicitly in the image, but rather defines the extent of an elliptical cluster of local features that are causally related.

Finally, despite the community's focus on exemplar (or restricted category) recognition over the last 15-20 years, it is important to acknowledge that there has been an active community who is committed to the problems of shape abstraction and categorization. Apart from the symmetry-based frameworks described above, important shape abstractions include shape contexts [22], inner-distance [150] in 2-D and shape distributions [173] in 3-D, multi-scale boundary-based methods, e.g., [97], parameterized blob models, e.g., [122], deformable models, [18, 86, 269], and articulated models, e.g., [119, 194, 85, 264]. It is also important to acknowledge the work of the content-based image retrieval (CBIR) community [226]. While they have focused less on the problems of segmentation, grouping, and shape abstraction, they have focused on powerful (typically global) abstractions of appearance that are appropriate for many image retrieval tasks. One such image abstraction consists of simply representing an entire image at low (e.g., 32×32) resolution, the minimum resolution at which human subjects can correctly interpret images of natural scenes [244]. If the model database contains enough correctly labeled, low-resolution model images (e.g., 80,000,000 in [244]), the space of real images of real objects can be sampled densely enough to facilitate surprisingly effective, albeit restricted, forms of object categorization.

4 The Abstraction of Structure

Successful image abstraction into a set of abstract shape primitives is only part of the problem, for a recovered primitive configuration may still not match the configuration of the correct model. For example, a chair having one leg has a very different part configuration than a chair having four legs. In fact, for most categories, part structure is variable and must somehow be parameterized. While the current recognition community is focused on strategies that assume a one-to-one correspondence between local image and model features, the need for models that accommodate variable structure was acknowledged long ago by, for example, Fu [96] and Rosenfeld [180, 202]. Brooks' ACRONYM system also parameterized structure, allowing

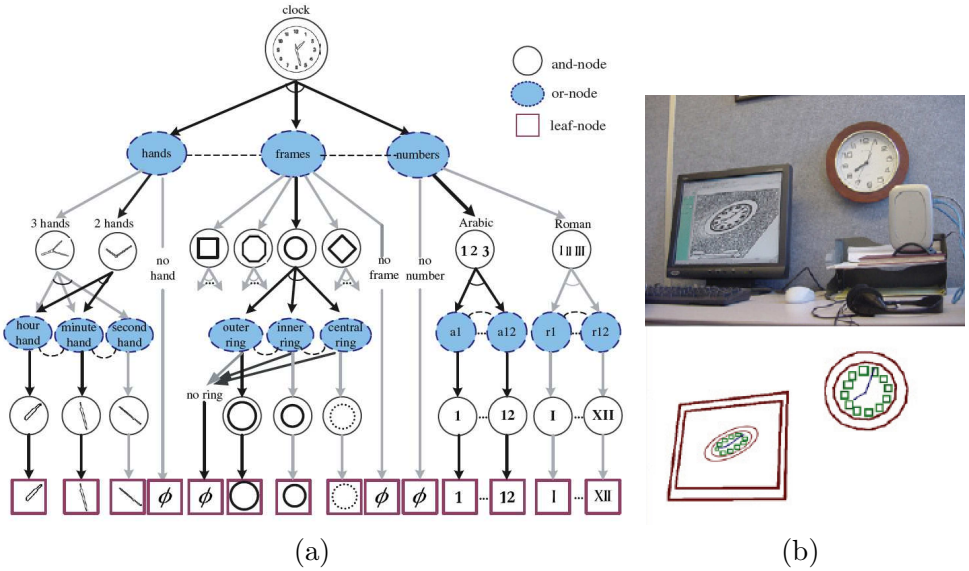


Figure 11: Parameterizing structure represents a form of structural abstraction, supporting less brittle object categories having variable structure. Drawing on the classical shape grammars of the 60's and 70's and on stochastic grammars from the computational linguistics community, Zhu and Mumford have rekindled interest in shape grammars. In this example, taken from [275], the clock grammar shown in (a) is used to recognize the two clocks (real and display) in the image shown in (b) (images reproduced from [275], courtesy of Now Publishers).

variable numbers of parts in an elaborate constraint manipulation system [45]. As the need to model structural variability is once again acknowledged, grammar-based methods, including AND/OR graphs, are beginning to re-emerge, for example, in work by Zhu and Mumford [275] (see Figure 11), Jin and Geman [123], and by Levinshtein et al. [146]. To the extent that the rewrite rules in a grammar can model the coarse-to-fine appearance of an object,[6] a grammar can provide an effective structural abstraction [146]. Other exciting categorization frameworks that support structural variability are also emerging, such as the hidden state shape models (HSSMs) of Wang et al. [259].

[6]Recall Marr's famous coarse-to-fine representation of a human with a single cylinder at the most abstract end of the modeling spectrum, and a detailed configuration down to the fingers at the least abstract end [158].

There have been other approaches to dealing with variable structure. One approach is to develop representations for structural abstraction, so that two configurations (or graphs) which have similar structure have similar structural abstractions. For example, in the domain of hierarchical structures, representing coarse-to-fine feature hierarchies that are ubiquitous in computer vision, Shokoufandeh et al. [219] draw on spectral graph theory to compute a low-dimensional abstraction of a directed acyclic graph (DAG). The eigenvalues of a DAG's antisymmetric adjacency matrix characterize the degree distribution of the graph's nodes, and are combined to define a low-dimensional vector description of the "shape" of a graph. This structural abstraction forms a basis for both indexing (nearest-neighbor search) as well as matching, unifying these two important problems in a common representational framework. However, the approach implicitly assumes that an effective grouping process has been used to generate the hierarchy.

In the absence of a strong set of grouping rules, a given set of extracted features may give rise to an exponential number of possible abstraction hierarchies. For example, consider the number of region adjacency graphs that can be abstracted from a single region adjacency graph by contracting an edge (merging two regions) to yield a new graph, and repeating. However, in a supervised setting, two exemplar graphs, which may not share a single node in correspondence, may be known to represent the same category. Keselman [127] searched for the *lowest common abstraction* of a set of exemplar graphs, i.e., the most informative graph derivable from each of the exemplars. While the technique was able to abstract a categorical model from a set of examples for which input feature correspondence might not exist, the method provided little insight into generating appropriate abstractions for a single image, as it was too dependent on "evidence" provided by other images known to belong to the same category.

Keselman's approach sought to group features to support one-to-one correspondence at some higher level, representing a many-to-many correspondence at the original level. An alternative strategy, proposed by Demirci et al. [64], computes an explicit many-to-many node correspondence between the original edge-weighted graphs. Drawing on recent work from the graph embedding community, the graphs are embedded with low distortion into a vector space, where shortest-path distances between nodes in a graph are reflected in the geometric distances between the nodes' corresponding points in the embedded space. The resulting points are matched many-to-many using the Earth Mover's Distance (EMD) algorithm, with the computed flows specifying the many-to-many node correspondences between the original graphs.

The approach cleverly transforms an intractable combinatorial problem into a tractable geometric problem, but it relies on assigning an appropriate set of edge weights in the original graphs. If an edge weight is small, the two nodes (spanned by the edge) are embedded nearby to one another, and will likely receive flow from a common source in the EMD solution. A small edge weight can be thought of as a high probability that the features are non-accidentally related. Hence the approach is only as successful as the perceptual grouping heuristics used to generate the graph and its edge weights. Moreover, the many-to-many solution yields only corresponding collections, not corresponding abstractions. Still, it does acknowledge the need to overcome the popular assumption that for every salient image feature, there exists a corresponding model feature. In fact, the need for many-to-many matching is acknowledged by a growing subcommunity. Important approaches have been proposed based on graph-edit distance [46, 212, 199], spectral methods [48], tree-union [241], association graph methods [175], and the emerging grammar-based methods mentioned earlier.

Finally, a discussion of shape and structural abstraction is incomplete without a reference to functional object descriptions [201, 99, 267, 254], which can be thought of as the highest form of shape abstraction. Many categories of objects exhibit a high degree of shape and structural variability, and that for such categories, explicit geometric models are too brittle. One might argue that even shape grammars, in their attempt to explicitly encode the possible structural variations, might be too unwieldy if the within-class structural variability is too high. In response to such categories, Stark and Bowyer [231, 232] proposed a number of functional predicates, designed to test the essential functional features of a model whose structural variability could be infinite. However, while these features could be described as functional, e.g., a chair provides horizontal support, vertical support, foot clearance, etc., such functional primitives could, in fact, be thought of as highly abstract geometric primitives (which were, in fact, computed from 3-D shape data). Rivlin et al. [197] suggested that reasoning about function first requires the extraction of high-order shape primitives, and that the mapping from shape primitives to functional primitives was many-to-one. Putting these two ideas together, such a many-to-one mapping could be considered a further form of shape abstraction. Since mechanisms for high-order primitive shape extraction were not available, not to mention the ability to further abstract shape to the level of functional primitives, functional models lost popularity during the mid-1990's.

5 Segmentation, Grouping, and the Role of Models: Beyond Target Recognition

Today's categorization community has clearly acknowledged the deficiency of appearance-based region segmentation to correctly separate figure from ground, and the need for some sort of prior knowledge to overcome this deficiency. While the perceptual organization community sought to inject mid-level, object-independent knowledge into the process, there has been a recent tendency to bypass mid-level knowledge and inject object-dependent knowledge into the process, e.g., [39, 252, 140, 270, 234, 151, 163, 243, 266, 260, 38, 111] (see Figure 12). Cast as a knowledge-based (or top-down) segmentation problem, it's important to note that this bears a close resemblance to classical target (or model-based) recognition, in which an individual object model (whether exemplar or category) is used to constrain image segmentation. In any classical target recognition task, the target was typically aligned with its detected instance, with the alignment defining (as a by-product) a figure/ground separation (including a parsing of the object into parts if the representation was parts-based). While target recognition was an important problem, particularly in military applications, knowing exactly which object to search for in the image, and hence which constraints to apply, was considered a special case of the more general problem of recognition from a large database.

Today's knowledge-based systems are superior to their early predecessors, particularly in terms of their ability to learn such constraints from training data. But once again the problem of mid-level shape abstraction has been avoided through the use of overly strong model assumptions. If the image to be labelled or segmented can contain any of 10,000 object categories (with arbitrary viewpoint, scale, articulation, occlusion, etc.), such techniques clearly don't scale up, and an indexing mechanism is required to prune all but a few promising candidates. Local features will have to be grouped and abstracted into mid-level primitives to support articulation, occlusion, and within-class shape deformation. Given a small vocabulary of such primitives, primitive extraction defines a tractable recognition task in its own right. When an extracted ("recognized") primitive is combined with a few other nearby primitives, they together yield a highly distinctive indexing structure. Somehow, the more specific detection problem has, in recent years, drawn the community's attention away from the more general categorization problem. But in doing so, the need for mid-level, generic parts and relations in the presence of a strong, top-down model is greatly

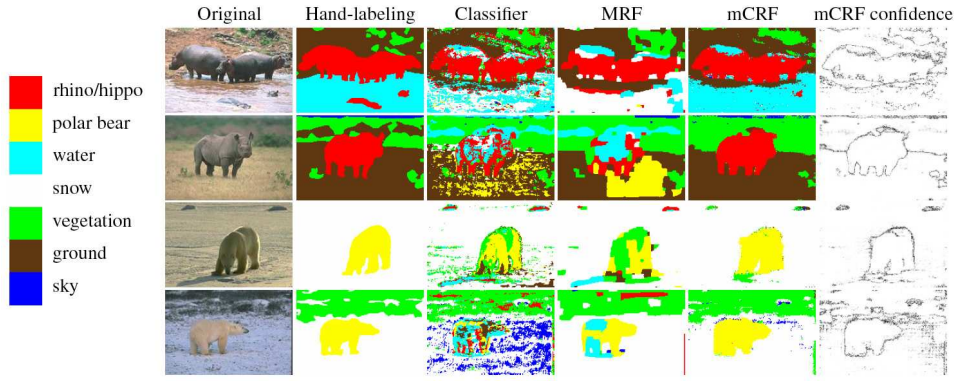


Figure 12: Image labeling results from the multiscale conditional random field approach of He et al. [110]. From labeled training images, knowledge of an object's appearance, position, size, and background can be learned and applied to the segmentation of test images (image reproduced from [110], Copyright ©2004 IEEE).

diminished, just as it was in the classical verification-oriented recognition systems of the 1980's.

While the detection problem may have drawn attention away from the bottom-up perceptual grouping and mid-level shape abstraction problems, it has nonetheless led to some very powerful abstraction mechanisms based on local image statistics. Such statistics, computed over some appropriate area, can take the form of distributions of semi-local features, with overlapping spatial support and some degree of spatial flexibility, ranging from total flexibility in a bag-of-features model [78] to multiple levels of spatial flexibility in a multiresolution, multilevel histogram pyramid [104]. Multilevel bag-of-feature models are by no means the only way to build representations with multiple levels of spatial selectivity. Biologically inspired architectures like HMAX [216] or convolutional neural networks [138] construct hierarchical representations by alternating successive layers of convolution (template matching to prototypes) and rectification (max pooling operations over afferent unit responses) in order to progressively build descriptors with increased invariance to scale, image rotation, and position. Local image statistics can also be used as inputs for training predictive models (complex, potentially multivalued, or one-to-many mappings) of 3D abstractions, such as planar structures [112] or human poses [26, 228, 227, 125].

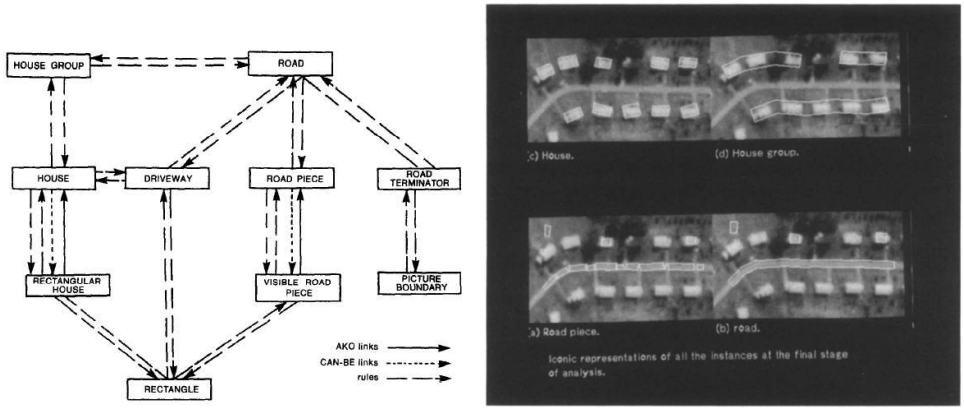
While these approaches focus primarily on appearance and less on shape,

they clearly acknowledge the need for flexible, abstract models that are robust to within-class variation. However, since they operate in a detection environment, the problem of segmentation is typically avoided by simply running the detector at all locations, orientations, and scales, until the image statistics inside a window match that of the target. Such an approach, which effectively tries all possible segmentations (over which the feature distribution is computed), simply does not scale up to either general viewpoint invariance or recognition from large databases. Perceptual grouping mechanisms (perhaps based on feature statistics, e.g., [196, 195]) must first group together causally related features into parts without regard to object class. Only then should shape statistics over the part (i.e., feature group) be computed, offering a powerful part abstraction mechanism.

6 Expanding Model Scope: Objects to Scenes

During the golden years of DARPA-funded image understanding research in the US, much of the object recognition community was devoted to *knowledge-based vision* systems (sometimes called *expert vision systems* [203], or *context-based vision systems*) which exploited scene-specific, contextual knowledge to provide additional evidence with which to disambiguate poorly segmented objects (see Figure 13). Early seminal work by researchers such as Tenenbaum and Barrow [237] and Hanson and Riseman [108] in the late 1970's popularized the integration of segmentation and interpretation. Over the next 10-15 years, the knowledge-based vision community manually constructed models that mapped functional or semantic relationships between objects in a scene to geometric relationships among their projections in the image. The resulting systems were applied to such diverse problems as aerial photo interpretation, e.g., [117, 109, 124], autonomous road following, e.g., [68], mechanical system image analysis, e.g., [44], medical image analysis, e.g., [246, 103], perceptual grouping, e.g., [207], or more general contexts, e.g., [235].

While the idea that domain-specific knowledge must play an important role was widely adopted by those seeking solutions to practical problems, there were those who dismissed knowledge-based vision systems as ad hoc or overly specific. Such systems were typically very slow, and largely unsuccessful, for the problems they addressed were often extremely difficult. Moreover, encoding domain knowledge in a system required significant overhead, and the fact that knowledge of one domain rarely lent itself to the next domain did not encourage their widespread adoption. When appearance-based



(a) (b)

Figure 13: The SIGMA image understanding system of Hwang et al. [117]: (a) a contextual model of a suburban housing development; (b) detecting instances of the contextual model's components in an image. Contextual constraints can help overcome poor segmentation and spurious hypotheses (images reproduced from [117] with permission of *Computer Vision, Graphics, and Image Processing*, Elsevier).

vision appeared on the scene in the early 1990's, coupled with a decreasing DARPA presence, knowledge-based vision systems all but disappeared. But they would make a comeback, just as shape-based categorization has re-emerged in the mainstream.

The first hints at a return to knowledge-based vision have come in the form of encoding scene context using simple spatial image statistics *without* requiring explicit segmentation or grouping, e.g., [58, 245, 110, 134] (see Figure 14). While appropriate for the recognition of very broad contexts, e.g., label a scene as containing people vs. furniture, the lack of segmentation prevents the localization of individual objects that might comprise a more specific context. Moreover, contexts whose components differ in shape (rather than appearance) cannot be addressed without segmentation. Still, unlike earlier generations of context-based or knowledge-based systems, contextual knowledge is learned automatically from annotated training examples, making such a framework far easier to transport to other domains. In more recent work, the community is returning to the classical ideas of combining object recognition and 3-D scene understanding, with notions of scene context capturing high-level relationships among objects, e.g., the fact that cars are typically found on roads and people are not typically detected

(a) (b)

Figure 14: From a set of labeled training images, Torralba [245] learns the correlation between context and object properties based on low-level image feature statistics computed over annotated training images: (a) test images for which $p(vehicles|v_c) < 0.05$, where v_c is a vector of contextual features derived from statistics computed over low-level image features; (b) test images for which $p(vehicles|v_c) > 0.95$ (image reproduced from [245] with permission of the *International Journal of Computer Vision*, Springer).

on the sides of a building [113].

When an object's identity is ambiguous, it makes little sense to ignore the contextual clues offered by nearby objects. If a context is too strictly specified, i.e., components appear at particular locations, scales, and orientations in an image, then context-based vision amounts to little more than brittle object detection. Interacting with a complex environment means explicitly recognizing its objects whose number, location, orientation (in both the image and in depth), scale, shape, and appearance may be highly variable. The fact that objects are coupled to form contexts (or environments, such as streetscapes, dining rooms, offices, or an army barracks) means that large object databases can be partitioned into smaller, context-specific databases. Determining what context (database) you're looking at requires that you first segment, group, and index into a set of contexts. The most likely context, in turn, defines the context-specific database with which to constrain the recognition of the unidentified or ambiguous objects in the scene. Since contextual indexing lies at the object level, and since the objects making up a context are typically categorical, we're back to the same problem of shape and configuration abstraction in support of object recognition. Contexts are to their component objects as objects are to their component parts. In either case, we must start with the segmentation, abstraction, and grouping

Figure 15: Learning structured appearance models for image annotation (from Jamieson et al. [121]): (a) Sample input image–caption collection, where each image contains hundreds of local (SIFT) features (yellow crosses). From the input training collection, associations between structured subsets of local features and particular nouns are learned (discovered); (b) sample output, where one of the objects learned during training (representing the Maple Leafs logo) is detected (shown with red features and green relationships in a yellow box), and annotated with its name ("Maple Leafs").

of an object's parts.

Finally, on the topic of increasing model scope, one might imagine moving beyond purely visual (shape-based or appearance-based) models toward more semantic models, coupling visual and semantic information. Learning the associations between visual features and nouns in training image captions has led to augmented models that combine visual appearance with an object name, supporting automatic image annotation, e.g., [56, 15, 11, 16, 13, 33, 12, 258, 55, 120, 121, 54, 14, 190] (see Figure 15). But the linguistic knowledge exploited by such approaches is minimal, and the unrealized potential for object names to invoke powerful semantic knowledge to guide image segmentation and scene interpretation is enormous. While some efforts, e.g., [115, 114], have exploited the restricted semantic information captured in WordNet's IS-A hierarchy [84], little work has tapped into more general semantic knowledge that may be captured through statistical analysis of text corpora.

7 Managing Search Complexity: The Case for 3-D Models

Early categorization systems typically modeled objects in 3-D, and sought to infer 3-D model features from 2-D image features. For in the words of the distinguished vision researcher, Jan-Olof Eklundh, "We need to build vision systems that look at the world, not at images." When 3-D models gave way to appearance models in the early 1990's, the need for viewpoint invariance in a recognition system translated from a single viewpoint-invariant 3-D model to a dense collection of 2-D views. But the appearance-based community was not the first to propose view-based 3-D object recognition, for the view-based [102, 23, 213, 251, 47, 263, 75, 20] and aspect graph-based [129, 118, 132, 185, 76, 206, 233, 74, 73, 70, 77, 101, 100, 217, 230] communities had studied both the theoretical as well as the practical implications of the approach.

The cost of transforming the 3-D recognition problem into a 2-D one is significant [79]. Plantinga and Dyer [185] proved that for a *rigid* polyhedron with n faces, the complexity of its aspect graph in terms of the number of distinct configurations of features observed as the viewing sphere is traversed, is $O(n^9)$. For articulating or deformable categories, the complexity becomes even more prohibitive. One way out of this dilemma is to apply the classical parts-based approach, using aspects to model not the entire objects, but rather the views of a small number of 3-D parts which, in turn, could be combined to form an infinite number of 3-D objects [74, 73, 70]. Thus, the number of views is fixed, and independent of the number of objects in the database. Moreover, since the parts are simple (having low complexity in terms of surfaces), their aspect graphs are similarly simple.

The evolution of appearance-based categorization saw a curious movement away from viewpoint invariance (which would reduce the number of views required to model a 3-D object) to focusing on particular views of categories, such as the sides of cars, the fronts of faces, the sides of motorcycles, or the sides of horses (effectively ignoring all but a few of the views required to model a 3-D object). The earlier view-based recognition assumption of having to recognize all possible views of a 3-D object did not translate into a multitude of detectors, one per view class. Surprisingly, only recently is the current categorization community beginning to return to the roots of view-based 3-D recognition, seeking view-based descriptions which offer greater viewpoint invariance.

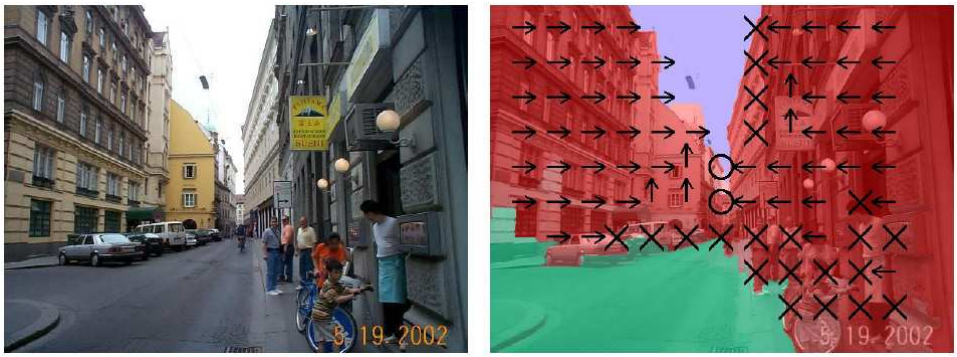
We are starting to see representations that bear a strong resemblance

to the aspect graphs of the 1980's and 1990's, e.g., [204, 135, 208]. And as the community returns to shape modeling (as reflected in recent work on contour-based representations, e.g., [221, 172, 87, 145]), it may well rediscover that for complex, articulating shape models, the number of aspects is intractable. The problem is further compounded when the community moves beyond object detectors to the more general problem of unexpected object recognition. The road back to 3-D is once again on the horizon, with exciting new work on learning to infer 3-D shape from 2-D appearance [112] (see Figure 16), and on the perception of 3-D shape from 2-D contours using a 3-D compactness constraint [184].[7] As contours once again become fashionable, one might expect a return to the problem of 3-D shape-from-contour. And as we move from the narrow problem of detection toward the more general problem of unexpected object recognition, the need to extract local (occlusion resistant) viewpoint- and articulation-invariant indices will focus shape-from-contour at the part level (rather than at the object level). After 40 years, we may once again be faced with the problem of recovering a vocabulary of higher-order, 3-D part models and their relations. And once again, the major challenge will be the abstraction of such part models from real images of real objects.

8 Identifying Our Shortcomings: The Need for New Benchmarks

Early work in categorization was rarely evaluated thoroughly, but rather demonstrated on a small set of anecdotal images. The weaknesses of the approaches were rarely discussed or illustrated in detail, and one was left to wonder on what domains a reported method might be successful. It was not the case that early categorization researchers did not appreciate the importance of evaluation. Rather, a number of factors conspired to make systematic evaluation a challenge: 1) the models were primarily 3-D and a standard representation had not been adopted by the community; 2) image and/or model databases were unavailable; and 3) computing power was

[7] Our ability to perceive 3-D objects is largely unaffected by the absence of such "direct" depth cues as binocular disparity or motion [183, 184]. This suggests that perceptual grouping operates on a single 2-D image, and while adding more images will always improve 3-D interpretation, it will not change the way shapes are found in, and abstracted from, the images. While this chapter has focused on the case of categorization (and abstraction) from a single 2-D image, it does not imply that binocular reconstruction and structure from motion, when available, cannot contribute to the processes of 3-D shape recovery and abstraction.



(a) (b)

Figure 16: Recovering 3-D surface layout from a single image (from Hoeim et al. [112]): (a) input image; (b) recovered surface layout: colours reflect class labels (green=support, red=vertical, blue=sky), while subclass labels are indicated by markings (left/up/right arrows for planar left/center/right, 'O' for porous, 'X' for solid) (image reproduced from [112] with permission of the *International Journal of Computer Vision*, Springer).

extremely limited, with a single image taking minutes or hours to interpret.

Not until the 1990's, with the advent of appearance-based recognition (where the model was the image) and faster machines, did evaluation benchmarks begin to emerge. The most prominent was the Columbia COIL-100 database [168], followed by the ETH database [139], and more recently the Amsterdam Library of Object Images [98], Caltech-101 [81], Caltech-256 [105], the MIT "tiny" image database [244], the LabelMe database [205], the ESP dataset [257], and a variety of other databases contained in the PASCAL Object Recognition Database Collection [78]. These databases were long overdue, and provide a means for more uniform evaluation and comparison of our work. They also provide a wealth of real-world images with which to automatically learn object models.

Categorization algorithm evaluation has improved dramatically since the early 1990's. But while we can now compare each other's algorithms on a standard dataset, it's not clear whether these datasets reflect the strengths and weaknesses of our algorithms [187, 182]. While early databases tested invariance to viewpoint (since they exhaustively enumerated a large set of views of an object), they did not test invariance to scale, image rotation, occlusion, significant articulation, clutter, or significant within-class shape deformation. Conversely, while more recent databases test significant

within-class appearance and shape deformation, they do not systematically test invariance to scale, image rotation, occlusion, clutter, and articulation. It's not that such transformations do not exist in the database collections, it's that they are not systematically parameterized so that our algorithms' failure modes can be clearly identified.

In some sense, our benchmarks are either too simple or too complex. When they're too simple, we run the risk of ignoring important invariance goals in our categorization system design because such goals are not reflected in the data. When a system reports good results on such a dataset, we have no way of knowing how it will fare under conditions not reflected in the dataset. Conversely, when the benchmarks are too complex, the invariance goals become obfuscated by the data. When a system reports good (or better) results on such a dataset, we don't know which conditions are handled well and which are not. The performance indicators simply don't yield critical insight into what aspects of the problem we need to improve on.

The community clearly needs a dataset that isolates the various conditions that we need to address. Such a database may take the form of a sequence of image suites, progressing from exemplars imaged under very controlled conditions to categories imaged under very challenging conditions, with a full spectrum of suites in between. For example, "suite-0" might fix a number of object imaging conditions, e.g., single scale, fixed illumination, fixed articulation, fixed appearance (i.e., an exemplar), no occlusion, and no clutter. The only free parameter would be viewpoint. Suite-0 would therefore be used to evaluate your algorithm's invariance to viewpoint change, and nothing else. Next up would be suite-1, which fixes all conditions except, for example, image scale, enabling you to evaluate the scale-invariance of your algorithm. Each successive suite, in turn, would test a different condition. Moreover, each condition would be systematically parameterized, so that where you fail on a particular suite would tell you exactly *how* invariant you are to that suite's condition(s). Early databases, such as COIL-100 [168] and the Amsterdam Image Library [98] parameterized viewpoint and illumination, while one recent database [65], created from the COIL-100 database, systematically parameterizes degree of occlusion.

As the suites progress toward the "human vision" suite, exemplars would give way to categories, rigid objects would give way to deformable objects, and uniform backgrounds would give way to cluttered backgrounds. Categories in earlier suites would exhibit very little within-class appearance or shape deformation, and in later suites would exhibit significant structural variability. Further suites could then combine conditions, leading to many

subsets of conditions which might tease out limitations of particular algorithms. To evaluate your algorithm would then amount to starting at suite-0 and reporting your results on each suite up to the conditions you claim to be invariant to. Such a set of suites would need to be designed by a consortium with no prior disposition to a particular recognition paradigm. In fact, to be paradigm invariant, 3-D data of the imaged objects should also be provided, allowing for the automatic construction of 3-D models which some may prefer over view-based models.

The existence of such a set of suites would allow our algorithms to evolve in a clear direction, ever more invariant to increasingly challenging conditions, but never losing sight of the need to address the fundamental conditions. Without the carefully designed intermediate suites, testing on only the most challenging suites which combine many conditions (akin to today's popular databases) may contribute little to our understanding of categorization. If such databases become more performance- than diagnostic-oriented, they may, in fact, end up distracting the categorization community from focusing on those particular issues that deserve attention. It is here that we can take a cue from our human vision colleagues, as the problem of designing proper experiments to test the performance of a vision system and to evaluate competing models has existed for a long time in the form of psychophysics. The first formal presentation of psychophysical methods can be found in Fechner [80]. A recent review that emphasizes the use of signal detection theory can be found in Macmillan and Creelman [156], and examples of the application of psychophysical methodology to the study of 3-D shape perception is presented by Pizlo [184].

9 Conclusions

The problem of object categorization has been around since the early 1970's. The legacy left by that original community was a set of rich object representations that modeled the coarse, prototypical, 3-D shape of an object. While important concepts such as viewpoint invariance, hierarchical representations, structural variability, indexing, and symmetry are rooted in this early work, the lack of image abstraction mechanisms restricted these systems to contrived images of contrived scenes. Instead of incrementally building on these rich representational ideas, models became gradually stronger, first in terms of shape and then appearance, thereby avoiding the need for image abstraction mechanisms. The resulting recognition systems began to be useful, first solving real exemplar-based industrial recognition problems under

tightly controlled conditions, and more recently solving real exemplar-based recognition problems in the real world.

Having made enormous progress on the problem of exemplar recognition, the community is now eager to return to the categorization problem. But the gradual redefinition of the recognition problem from categories to exemplars, followed by a representational movement from shape to appearance, has unfortunately displaced a rich history of categorization from our community's memory. The sudden popularity of object recognition in the early 2000's is due in part to the fact that an image can now be mapped to a set of very distinctive local feature vectors without having to engage in the classical, unsolved problems of segmentation and grouping. This has drawn a new generation of computer vision and machine learning researchers into the ring. Our progress will clearly benefit from both the increased popularity of the problem as well as the influx of new techniques from other communities. However, a much smaller portion of this new community will have witnessed the evolution of categorization, contributing further to the separation of the categorization community from its roots.

Today's categorization community has moved quickly to apply exemplar-based appearance models to more categorical tasks. Ultimately, these are destined to fail, for local appearance is seldom generic to a category. This is reflected in a recent shift back to shape, along with a recent rediscovery of the importance of viewpoint invariance.[8] This is a very positive development, for our computers, our inference engines, our ability to deal with uncertain information, and our ability to learn a system's parameters rather than hand-code them represent enormous improvements over previous generations. As such, our return to earlier problems will lead to vastly more effective solutions. Without a doubt, we are heading in the right direction again.

But one invariant has survived the pendulum-like journey of our community: our tendency to avoid the difficult problem of image (or shape) abstraction. Once we acknowledge this important problem, we must be patient and not expect results too soon. We must understand the history of the research in our community, building on important representational ideas and concepts from the past, and not being dismissive of earlier work just because it did not deal with real images. Each generation of categorization researchers has made important contributions and we must incrementally build on the foundations laid by our predecessors. When we do develop

[8]Note that viewpoint invariance is equally important in the study of human vision, where it is usually referred to as "shape constancy."

solutions, they must be carefully evaluated under controlled conditions that can provide us with the most constructive feedback. Finally, we must reconnect with our human vision colleagues, so that we can maximally benefit from their research on the most impressive categorization system of them all: the human vision system.

10 Acknowledgements

I am indebted to the following individuals who provided thoughtful feedback on and corrections to this chapter: Narendra Ahuja, Ronen Basri, Gustavo Carneiro, Larry Davis, Afsaneh Fazly, Gertruda Grolinger, Mike Jamieson, Allan Jepson, Anatoliy Kats, Yakov Keselman, Alex Levinshtein, David Lowe, Diego Macrini, Stefan Mathe, Zygmunt Pizlo, Pablo Sala, Stan Sclaroff, Linda Shapiro, Ali Shokoufandeh, Kaleem Siddiqi, Cristian Sminchisescu, Suzanne Stevenson, Babak Taati, Alireza Tavakoli Targhi, Sinisa Todorovic, John Tsotsos, and Steve Zucker. My sincerest thanks to you all.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 26(11):1475–1490, 2004.

[2] G. Agin and T. O. Binford. Computer description of curved objects. *IEEE Transactions on Computers*, C-25(4):439–449, 1976.

[3] N. Ahuja and J.-H. Chuang. Shape representation using a generalized potential field model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):169–176, 1997.

[4] N. Ahuja and S. Todorovic. Learning the taxonomy and models of categories present in arbitrary images. In *Proc. IEEE Int. Conf. Computer Vision*, Rio de Janeiro, Brazil, 2007.

[5] T. D. Alter. 3-d pose from 3 points using weak-perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):802–808, 1994.

[6] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

[7] Y. Amit, D. Geman, and K. Wilder. Joint induction of shape features and tree classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:1300–1305, 1997.

[8] J. August, K. Siddiqi, and S. W. Zucker. Ligature instabilities in the perceptual organization of shape. *Computer Vision and Image Understanding*, 76(3):231–243, 1999.

[9] X. Bai and L. J. Latecki. Path similarity skeleton graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, XX(YY), 2007.

[10] X. Bai, L. J. Latecki, and W. Liu. Skeleton pruning by contour partitioning with discrete curve evolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):449–462, 2007.

[11] K. Barnard, P. Duygulu, N. de Freitas, and D. Forsyth. Object recognition as machine translation - part 2: Exploiting image database clustering models. In *Proceedings, European Conference on Computer Vision*, 2002.

[12] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[13] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, and D. Forsyth. The effects of segmentation and feature choice in a translation model of object recognition. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[14] K. Barnard and Q. Fan. Reducing correspondence ambiguity in loosely labeled training data. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[15] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proceedings, International Conference on Computer Vision*, 2001.

[16] K. Barnard and P. Gabbur. Color and color constancy in a translation model for object recognition. In *Eleventh Color Imaging Conference*, 2003.

[17] A. H. Barr. Superquadrics and angle-preserving transformations. *IEEE Computer Graphics and Applications*, 1(1), 1981.

[18] R. Basri, L. Costa, D. Geiger, and D. Jacobs. Determining the similarity of deformable shapes. *Vision Research*, 38:2365–2385, 1998.

[19] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

[20] J. S. Beis and D. Lowe. Indexing without invariants in 3d object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):1000–1015, 1999.

[21] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible lighting conditions? In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–277, 1996.

[22] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[23] J. Ben-Arie. The probabilistic peaking effect of viewed angles and distances with application to 3-D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(8):760–774, August 1990.

[24] R. Bergevin and M. D. Levine. Part decomposition of objects from single view line drawings. *CVGIP: Image Understanding*, 55(1):73–83, January 1992.

[25] R. Bergevin and M. D. Levine. Generic object recognition: Building and matching coarse 3d descriptions from line drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:19–36, January 1993.

[26] D. Beymer and T. Poggio. Image representations for visual learning. *Science 28*, 272(5270):1905–1909, June 1996.

[27] I. Biederman. Human image understanding: Recent research and a theory. *Computer Vision, Graphics, and Image Processing*, 32:29–73, 1985.

[28] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

[29] T. O. Binford. Visual perception by computer. In *Proceedings, IEEE Conference on Systems and Control*, Miami, FL, 1971.

[30] T. O. Binford and T. S. Levitt. Evidential reasoning for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):837–851, 2003.

[31] T. O. Binford, T. S. Levitt, and W. B. Mann. Bayesian inference in model-based machine vision. In *UAI*, pages 73–96, 1987.

[32] M. J. Black and A. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.

[33] D. Blei and M. Jordan. Modeling annotated data. In *Proc. Int'l Conf. Research and Development in Information Retrieval*, 2003.

[34] D. Blostein and N. Ahuja. A multiscale region detector. *Computer Vision, Graphics, and Image Processing*, 45:22–41, 1989.

[35] 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.

[36] H. Blum. Biological shape and visual science. *J. Theor. Biol.*, 38:205–287, 1973.

[37] H. Blum and R. N. Nagel. Shape description using weighted symmetric axis features. *Pattern Recognition*, 10:167–180, 1978.

[38] E. Borenstein and J. Malik. Shape guided object segmentation. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 969–976, 2006.

[39] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *Proceedings, European Conference om Computer Vision*, pages 109–124, 2002.

[40] D. Borges and R. Fisher. Class-based recognition of 3d objects represented by volumetric primitives. *Image and Vision Computing*, 15(8):655–664, 1997.

[41] G. Bouchard and B. Triggs. Hierarchical part-based visual object categorization. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 710–715, 2005.

[42] K.L. Boyer and S. Sarkar. Perceptual organization in computer vision: status, challenges, and potential. *Computer Vision and Image Understanding*, 76(1), 1999.

[43] K.L. Boyer and S. Sarkar. *Perceptual organization for artificial vision systems*. Kluwer, Boston, 2000.

[44] M. Brand. Physics-based visual understanding. *Computer Vision and Image Understanding*, 65(2):192–205, 1997.

[45] R. Brooks. Model-based 3-D interpretations of 2-D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):140–150, 1983.

[46] H. Bunke. On a relation between graph edit distance and maximum common subgraph. *Pattern Recognition Letters*, 18(8):689–694, 1997.

[47] J. Burns, R. Weiss, and E. Riseman. View variation of point-set and line-segment features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(1):51–68, January 1993.

[48] T. Caelli and S. Kosinov. An eigenspace projection clustering method for inexact graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(4):515–519, 2004.

[49] O. Camps, C. Huang, and T Kanungo. Hierarchical organization of appearance based parts and relations for object recognition. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–691, Santa Barbara, CA, 1998.

[50] G. Carneiro and A. D. Jepson. Phase-based local features. In *Proceedings, European Conference on Computer Vision*, pages 282–296, 2002.

[51] G. Carneiro and A. D. Jepson. Flexible spatial models for grouping local image features. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 747–754, 2004.

[52] G. Carneiro and A. D. Jepson. Flexible spatial configuration of local image features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2089–2104, 2007.

[53] G. Carneiro and D. Lowe. Sparse flexible models of local features. In *Proceedings, European Conference on Computer Vision*, pages 29–43, 2006.

[54] G. Carniero, A.B. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[55] G. Carniero and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[56] M. La Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the World Wide Web. In *Proc. of IEEE Workshop on Content-based Access of Image and Video Libraries*, June 1998.

[57] T. A. Cass. Robust affine structure matching for 3d object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1265–1274, 1998.

[58] H. Cheng and C. A. Bouman. Multiscale bayesian segmentation using a trainable context model. *IEEE Transactions on Image Processing*, 10(4):511–525, 2001.

[59] J.-H. Chuang, N. Ahuja, C.-C. Lin, C.-H. Tsai, and C.-H. Chen. A potential-based generalized cylinder representation. *Computers and Graphics*, 28:907–918, 2004.

[60] D. T. Clemens and D. W. Jacobs. Space and time bounds on indexing 3d models from 2d images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1007–1017, 1991.

[61] M. S. Costa and L. G. Shapiro. 3d object recognition and pose with relational indexing. *Computer Vision and Image Understanding*, 79(3):364–407, 2000.

[62] J. Crowley and A. Parker. A representation for shape based on peaks and ridges in the difference of low-pass transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2):156–169, March 1984.

[63] J. Crowley and A. C. Sanderson. Multiple Resolution Representation and Probabilistic Matching of 2–D Gray–Scale Shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):113–121, January 1987.

[64] M. F. Demirci, A. Shokoufandeh, Y. Keselman, L. Bretzner, and S. Dickinson. Object recognition as many-to-many feature matching. *International Journal of Computer Vision*, 69(2):203–222, 2006.

[65] T. Denton, J. Novatnack, and A. Shokoufandeh. Drexel object occlusion repository (door). Technical Report DU-CS-05-08, Drexel University, 2005.

[66] S. Dickinson, R. Bergevin, I. Biederman, J.-O. Eklundh, A. Jain, R. Munck-Fairwood, and A. Pentland. Panel report: The potential of geons for generic 3-D object recognition. *Image and Vision Computing*, 15(4):277–292, April 1997.

[67] S. Dickinson, H. Christensen, J. Tsotsos, and G. Olofsson. Active object recognition integrating attention and viewpoint control. *Computer Vision and Image Understanding*, 67(3):239–260, September 1997.

[68] S. Dickinson and L. Davis. A flexible tool for prototyping alv road following algorithms. *IEEE Journal of Robotics and Automation*, 6(2):232–242, April 1990.

[69] S. Dickinson and D. Metaxas. Integrating qualitative and quantitative shape recovery. *International Journal of Computer Vision*, 13(3):1–20, 1994.

[70] S. Dickinson and D. Metaxas. Using aspect graphs to control the recovery and tracking of deformable models. *International Journal of Pattern Recognition and Artificial Intelligence*, 11(1):115–142, 1997.

[71] S. Dickinson, D. Metaxas, and A. Pentland. The role of model-based segmentation in the recovery of volumetric parts from range data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(3):259–267, March 1997.

[72] S. Dickinson, A. Pentland, and A. Rosenfeld. A representation for qualitative 3-D object recognition integrating object-centered and viewer-centered models. In K. Leibovic, editor, *Vision: A Convergence of Disciplines*. Springer Verlag, New York, 1990.

[73] 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.

[74] 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.

[75] S. Dickinson, D. Wilkes, and J. K. Tsotsos. A computational model of view degeneracy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):673–689, 1999.

[76] D. Eggert and K. Bowyer. Computing the orthographic projection aspect graph of solids of revolution. *Pattern Recognition Letters*, 11:751–763, 1990.

[77] D. Eggert, K. Bowyer, C. Dyer, H. Christensen, and D. Goldgof. The scale space aspect graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1114–1130, November 1993.

[78] M. Everingham, A. Zisserman, C. K. I. Williams, L. J. Van Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkó, S. Duffner, J. Eichhorn, J. D. R. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A. J. Storkey, S. Szedmák, B. Triggs, I. Ulusoy, V. Viitaniemi, and J. Zhang. The 2005 pascal visual object classes challenge. In *MLCW*, pages 117–176, 2005.

[79] O. Faugeras, J. Mundy, N. Ahuja, C. Dyer, A. Pentland, R. Jain, K. Ikeuchi, and K. Bowyer. Why aspect graphs are not (yet) practical for computer vision. *Computer Vision, Graphics and Image Processing: Image Understanding*, 55(2):212–218, 1992.

[80] G. Fechner. *Elements of Psychophysics*. Holt, Rinehart & Winston, New York, 1860/1966.

[81] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *IEEE Workshop on Generative-Model Based Vision*, Washington, D.C., 2004.

[82] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[83] J. Feldman and M. Singh. Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences*, 103:18014–18019, 2006.

[84] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[85] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[86] P. F. Felzenszwalb. Representation and detection of deformable shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):208–220, 2005.

[87] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *International Journal of Computer Vision*, 71(3):273–303, 2007.

[88] V. Ferrari, F. Jurie, and C. Schmid. Accurate object detection with deformable shape models learnt from images. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[89] V. Ferrari, T. Tuytelaars, and L. J. Van Gool. Simultaneous object recognition and segmentation by image exploration. In *Proceedings, European Conference om Computer Vision*, pages 40–54, 2004.

[90] V. Ferrari, T. Tuytelaars, and L. J. Van Gool. Simultaneous object recognition and segmentation from single or multiple model views. *International Journal of Computer Vision*, 67(2):159–188, 2006.

[91] F. Ferrie, J. Lagarde, and P. Whaite. Darboux frames, snakes, and super-quadrics: Geometry from the bottom up. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(8):771–784, 1993.

[92] S. Fidler and A. Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[93] P. Flynn and A. Jain. 3D object recognition using invariant feature indexing of interpretation tables. *CVGIP:Image Understanding*, 55(2):119–129, March 1992.

[94] D. A. Forsyth, J. L. Mundy, A. Zisserman, C. Coelho, A. Heller, and C. Rothwell. Invariant descriptors for 3d object recognition and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):971–991, 1991.

[95] A. R. J. François and G. G. Medioni. Generic shape learning and recognition. In *Object Representation in Computer Vision II, International Workshop, Cambridge, UK*, pages 287–320, 1996.

[96] K.S. Fu. *Syntactic Methods in Pattern Recognition*. Academic Press, 1974.

[97] Y. Gdalyahu and D. Weinshall. Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1312–1328, 1999.

[98] J.-M. Geusebroek, G. Burghouts, and A. Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005.

[99] J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979.

[100] Z. Gigus, J. F. Canny, and R. Seidel. Efficiently computing and representing aspect graphs of polyhedral objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):542–551, 1991.

[101] Z. Gigus and J. Malik. Computing the aspect graph for line drawings of polyhedral objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(2):113–122, 1990.

[102] C. Goad. Special purpose automatic programming for 3D model-based vision. In *Proceedings, DARPA Image Understanding Workshop*, pages 94–104, Arlington, VA, 1983.

[103] L. Gong and C. A. Kulikowski. Composition of image analysis processes through object-centered hierarchical planning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):997–1009, 1995.

[104] K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In *Proceedings, IEEE Con-*

ference on Computer Vision and Pattern Recognition, pages 19–25, 2006.

[105] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[106] W. Grimson and T. Lozano-Pérez. Model-based recognition and localization from sparse range or tactile data. *International Journal of Robotics Research*, 3(3):3–35, 1984.

[107] A. Gupta and R. Bajcsy. Volumetric segmentation of range images of 3d objects using superquadric models. *CVGIP: Image Understanding*, 58(3):302–326, 1993.

[108] A. Hanson and E. Riseman. Visions: A computer vision system for interpreting scenes. In A. Hanson and E. Riseman, editors, *Computer Vision Systems*, pages 303–334. Academic Press, New York, NY, 1978.

[109] D. Harwood, R. Prasannappa, and L. Davis. Preliminary design of a programmed picture logic. In *Proceedings, Image Understanding Workshop*, volume 2, pages 745–755. Science Applications International Corp., 1988.

[110] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 695–702, 2004.

[111] X. He, R. S. Zemel, and D. Ray. Learning and incorporating top-down cues in image segmentation. In *Proceedings, European Conference on Computer Vision*, pages 338–351, 2006.

[112] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1), October 2007.

[113] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 2137–2144, 2006.

[114] A. Hoogs and R. Collins. Object boundary detection in images using a semantic ontology. In *AAAI*, 2006.

[115] A. Hoogs, J. Rittscher, G. Stien, and J. Schmiederer. Video content annotation using video analysis and a large semantic knowledgebase. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[116] D. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212, 1990.

[117] V. Hwang, L.S. Davis, and T.Matsuyama. Hypothesis integration in image understanding systems. *Computer Vision, Graphics and Image Processing*, 36(3):321–371, 1986.

[118] K. Ikeuchi and T. Kanade. Automatic generation of object recognition programs. *Proceedings of the IEEE*, 76:1016–1035, 1988.

[119] S. Ioffe and D. Forsyth. Probabilistic methods for finding people. *International Journal of Computer Vision*, 43(1):45–68, 2001.

[120] M. Jamieson, S. Dickinson, S. Stevenson, and S. Wachsmuth. Using language to drive the perceptual grouping of local image features. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[121] M. Jamieson, A. Fazly, S. Dickinson, S. Stevenson, and S. Wachsmuth. Learning structured appearance models from captioned images of cluttered scenes. In *Proceedings, International Conference on Computer Vision*, 2007.

[122] A. D. Jepson, D. J. Fleet, and M. J. Black. A layered motion representation with occlusion and compact spatial support. In *Proceedings, European Conference on Computer Vision*, pages 692–706, 2002.

[123] Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 2145–2152, 2006.

[124] D. M. McKeown Jr., W. A. Harvey, and L. E. Wixson. Automating knowledge acquisition for aerial image interpretation. *Computer Vision, Graphics, and Image Processing*, 46(1):37–81, 1989.

[125] A. Kanaujia, C. Sminchisescu, and D. N. Metaxas. Semi-supervised hierarchical models for 3d human pose reconstruction. In *Proceedings,*

IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2007.

[126] R. A. Katz and S. M. Pizer. Untangling the blum medial axis transform. *International Journal of Computer Vision*, 55(2-3):139–153, 2003.

[127] Y. Keselman and S. Dickinson. Generic model abstraction from examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1141–1156, 2005.

[128] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for thecharacterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[129] J. Koenderink and A. van Doorn. The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32:211–216, 1979.

[130] J. J. Koenderink and A. J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.

[131] K. Koffka. *Principles of Gestalt Psychology*. Harcourt, Brace, New York, 1935.

[132] D. Kriegman and J. Ponce. Computing exact aspect graphs of curved objects: Solids of revolution. *International Journal of Computer Vision*, 5(2):119–135, 1990.

[133] D. J. Kriegman and J. Ponce. On recognizing and positioning curved 3-d objects from image contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(12):1127–1137, 1990.

[134] S. Kumar and M. Hebert. Discriminative random fields. *International Journal of Computer Vision*, 68(2):179–201, 2006.

[135] A. Kushal, C. Schmid, and J. Ponce. Flexible object models for category-level 3d object recognition. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, 2007.

[136] Y. Lamdan, J. Schwartz, and H. Wolfson. Affine invariant model-based object recognition. *IEEE Transactions on Robotics and Automation*, 6(5):578–589, October 1990.

[137] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, 2005.

[138] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[139] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 409–415, 2003.

[140] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *Proceedings, British Machine Vision Conference*, Norwich, UK, September 2003.

[141] A. Leonardis and H. Bischof. Robust recognition using eigenimages. *Computer Vision and Image Understanding*, 78(1):99–118, 2000.

[142] A. Leonardis and H. Bischoff. Dealing with occlusions in the eigenspace approach. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 453–458, San Francisco, CA, June 1996.

[143] A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation of range images as the search for geometric parametric models. *International Journal of Computer Vision*, 14(3):253–277, 1995.

[144] A. Leonardis, A. Jaklic, and F. Solina. Superquadrics for segmenting and modeling range data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(11), 1997.

[145] M. Leordeanu, M. Hebert, and R. Sukthankar. Beyond local appearance: Category recognition from pairwise interactions of simple features. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[146] A. Levinshtein, C. Sminchisescu, and S. Dickinson. Learning hierarchical shape models from examples. In *Proceedings, EMMCVPR*, pages 251–267, 2005.

[147] M. Leyton. A process-grammar for shape. *Artif. Intell.*, 34(2):213–247, 1988.

[148] T. Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *International Journal of Computer Vision*, 11:283–318, 1993.

[149] T. Lindeberg. Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):117–154, 1998.

[150] H. Ling and D. W. Jacobs. Shape classification using the inner-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):286–299, 2007.

[151] L. Liu and S. Sclaroff. Deformable model-guided region split and merge of image regions. *Image and Vision Computing*, 22(4):343–354, 2004.

[152] D. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Norwell, MA, 1985.

[153] D. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:355–395, 1987.

[154] D. Lowe. Object recognition from local scale-invariant features. In *Proceedings, International Conference on Computer Vision*, pages 1150–1157, 1999.

[155] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[156] N.A. Macmillan and C.D. Creelman. *Detection theory: a user's guide*. Lawrence Erlbaum, Mahwah, NJ, 2005.

[157] W. B. Mann. *Three Dimensional Object Interpretation of Monocular Grey-Scale Images*. PhD thesis, Stanford University, 1995.

[158] D. Marr. *Vision*. W. H. Freeman, San Francisco, CA, 1982.

[159] D. Marr and H. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Royal Society of London*, B 200:269–294, 1978.

[160] G. Medioni and K. Rao. Generalized cones: Useful geometric properties. *Computer Vision, Graphics and Image Processing*, 10(3):185–208, October 1992.

[161] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[162] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[163] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–333, 2004.

[164] P. Mulgaonkar, L. Shapiro, and R. Haralick. Matching "sticks, plates and blobs" objects using geometric and relational constraints. *Image and Vision Computing*, 2(2):85–98, 1984.

[165] J. Mundy and A. Zisserman, editors. *Geometric Invariance in Computer Vision*. MIT Press, Cambridge, MA, 1992.

[166] H. Murase and S. Nayar. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[167] R. Nelson and A. Selinger. A cubist approach to object recognition. In *Proceedings, IEEE International Conference on Computer Vision*, Bombay, January 1998.

[168] S. Nene, S. Nayar, and H. Murase. Columbia object image library (coil-100). Technical Report CUCS-006-96, Columbia University, February 1996.

[169] R. Nevatia and T. O. Binford. Description and recognition of curved objects. *Artificial Intelligence*, 8:77–98, 1977.

[170] K. Ohba and K. Ikeuchi. Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9):1043–1048, 1997.

[171] B. Ommer and J. M. Buhmann. Learning the compositional nature of visual objects. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[172] A. Opelt, A. Pinz, and A. Zisserman. Incremental learning of object detectors using a visual shape alphabet. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 3–10, 2006.

[173] R. Osada, T. A. Funkhouser, B. Chazelle, and D. P. Dobkin. Shape distributions. *ACM Transactions on Graphics*, 21(4):807–832, 2002.

[174] M. Pelillo, K. Siddiqi, and S. Zucker. Matching hierarchical structures using association graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11):1105–1120, November 1999.

[175] M. Pelillo, K. Siddiqi, and S. W. Zucker. Many-to-many matching of attributed trees using association graphs and game dynamics. In *IWVF*, pages 583–593, 2001.

[176] A. Pentland. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28:293–331, 1986.

[177] A. Pentland. Automatic extraction of deformable part models. *International Journal of Computer Vision*, 4:107–126, 1990.

[178] A. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modeling and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):715–729, 1991.

[179] B. Perrin, N. Ahuja, and N. Srinivasa. Learning multiscale image models of 2d object classes. In *Proceedings, Asian Conference on Computer Vision*, pages 323–331, 1998.

[180] J. L. Pfaltz and A. Rosenfeld. Web grammars. In *Proceedings, International Joint Conference on Artificial Intelligence*, pages 609–620, 1969.

[181] M. Pilu and R. B. Fisher. Recognition of geons by parametric deformable contour models. In *Proceedings, European Conference on Computer Vision*, pages 71–82, 1996.

[182] N. Pinto, D. D. Cox, and J. J. DiCarlo. Why is real-world visual object recognition hard? *PLoS Computational Biology*, 4(1:e27), 2008.

[183] Z. Pizlo. Perception viewed as an inverse problem. *Vision Research*, 41:3145–3161, 2001.

[184] Z. Pizlo. *3D shape: its unique place in visual perception*. MIT Press, Cambridge, MA, 2008.

[185] H. Plantinga and C. Dyer. Visibility, occlusion, and the aspect graph. *International Journal of Computer Vision*, 5(2):137–160, 1990.

[186] J. Ponce. Straight homogeneous generalized cylinders: differential geometry and uniqueness results. *International Journal of Computer Vision*, 4(1):79–100, January 1990.

[187] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In *Toward Category-Level Object Recognition*, pages 29–48, 2006.

[188] J. Ponce and D. Chelberg. Finding the limbs and cusps of generalized cylinders. *International Journal of Computer Vision*, 1(3):195–210, 1987.

[189] T. Quack, V. Ferrari, B. Leibe, and L. Van Gool. Efficient mining of frequent and distinctive feature configurations. In *Proceedings, International Conference on Computer Vision*, Rio de Janeiro, Brasil, October 2007.

[190] A. Quattoni, M. Collins, and T. Darrell. Learning visual representations using images with captions. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[191] N. Raja and A. Jain. Recognizing geons from superquadrics fitted to range data. *Image and Vision Computing*, 10(3):179–190, April 1992.

[192] N. Raja and A. Jain. Obtaining generic parts from range images using a multi-view representation. *CVGIP:Image Understanding*, 60(1):44–64, July 1994.

[193] R. P. N. Rao and D. H. Ballard. An active vision architecture based on iconic representations. *Artif. Intell.*, 78(1-2):461–505, 1995.

[194] J. Rehg, D. Morris, and T. Kanade. Ambiguities in visual tracking of articulated objects using two- and three-dimensional models. *International Journal of Robotics Research*, 22(6):393–418, 2003.

[195] X. Ren, C. C. Fowlkes, and J. Malik. Scale-invariant contour completion using conditional random fields. In *Proc. 10th Int'l. Conf. Computer Vision*, volume 2, pages 1214–1221, 2005.

[196] X. Ren, C. C. Fowlkes, and J. Malik. Figure/ground assignment in natural images. In *Proc. 9th Europ. Conf. Comput. Vision*, volume 2, pages 614–627, 2006.

[197] E. Rivlin, S. Dickinson, and A. Rosenfeld. Recognition by functional parts. *Computer Vision and Image Understanding*, 62(2):164–176, 1995.

[198] 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.

[199] A. Robles-Kelly and E. R. Hancock. Graph edit distance from spectral seriation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):365–378, 2005.

[200] H. Rom and G. Medioni. Hierarchical decomposition and axial shape description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):973–981, October 1993.

[201] E. Rosch, C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[202] A. Rosenfeld. *Picture Languages: Formal Models for Picture Recognition*. Academic Press, 1979.

[203] A. Rosenfeld. Expert vision sytems: some issues. *Computer Vision, Graphics, and Image Processing*, 34(1):99–102, April 1986.

[204] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision*, 66(3):231–259, 2006.

[205] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, to appear, 2008.

[206] M. Sallam and K. Bowyer. Generalizing the aspect graph concept to include articulated assemblies. *Pattern Recognition Letters*, 12:171–176, 1991.

[207] S. Sarkar and K. L. Boyer. Using perceptual inference networks to manage vision processes. *Computer Vision and Image Understanding*, 62(1):27–46, 1995.

[208] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *Proceedings, IEEE International Conference on Computer Vision*, 2007.

[209] B. Schiele and J. L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.

[210] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[211] S. Sclaroff and A. Pentland. Modal matching for correspondence and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(6):545–561, June 1995.

[212] T. Sebastian, P. N. Klein, and B. Kimia. Recognition of shapes by editing their shock graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):550–571, 2004.

[213] M. Seibert and A. Waxman. Adaptive 3-D object recognition from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):107–124, 1992.

[214] K. Sengupta and K. L. Boyer. Organizing large structural modelbases. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(4):321–332, 1995.

[215] K. Sengupta and K. L. Boyer. Modelbase partitioning using property matrix spectra. *Computer Vision and Image Understanding*, 70(2):177–196, 1998.

[216] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007.

[217] I. Shimshoni and J. Ponce. Finite-resolution aspect graphs of polyhedral objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):315–327, 1997.

[218] A. Shokoufandeh, L. Bretzner, D. Macrini, M. F. Demirci, C. Jönsson, and S. Dickinson. The representation and matching of categorical shape. *Computer Vision and Image Understanding*, 103(2):139–154, 2006.

[219] A. Shokoufandeh, D. Macrini, S. Dickinson, K. Siddiqi, and S. W. Zucker. Indexing hierarchical structures using graph spectra. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1125–1140, 2005.

[220] A. Shokoufandeh, I. Marsic, and S. Dickinson. View-based object recognition using saliency maps. *Image and Vision Computing*, 17(5-6):445–460, 1999.

[221] J. Shotton, A. Blake, and R. Cipolla. Contour-based learning for object detection. In *Proceedings, International Conference on Computer Vision*, pages 503–510, 2005.

[222] K. Siddiqi and S. Pizer. *Medial Representations: Mathematics, Algorithms, and Applications*. Springer Verlag, New York, 2008.

[223] K. Siddiqi, A. Shokoufandeh, S. Dickinson, and S. Zucker. Shock graphs and shape matching. *International Journal of Computer Vision*, 30:1–24, 1999.

[224] K. Siddiqi, J. Zhang, D. Macrini, A. Shokoufandeh, S. Bioux, and S. Dickinson. Retrieving articulated 3-d models using medial surfaces. *Machine Vision and Applications*, to appear, 2007.

[225] T. Silberberg, D. A Harwood, and L. S. Davis. Object recognition using oriented model points. *Compututer Vision, Graphics, and Image Processing*, 35(1):47–71, 1986.

[226] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[227] C. Sminchisescu, A. Kanaujia, and D. Metaxas. "bm^3e: Discriminative density propagation for visual tracking". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, November 2007.

[228] C. Sminchisescu, A. Kanaujia, and D. N. Metaxas. Learning joint top-down and bottom-up processes for 3d visual inference. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 1743–1752, 2006.

[229] F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(2):131–146, 1990.

[230] T. Sripradisvarakul and R. Jain. Generating aspect graphs for curved objects. In *Proceedings, IEEE Workshop on Interpretation of 3D Scenes*, pages 109–115, Austin, TX, 1989.

[231] L. Stark and K. Bowyer. Achieving generalized object recognition through reasoning about association of function to structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1097–1104, 1991.

[232] L. Stark and K. Bowyer. Function-based generic recognition for multiple object categories. *CVGIP: Image Understanding*, 59(1):1–21, January 1994.

[233] J. Stewman and K. Bowyer. Direct construction of the perspective projection aspect graph of convex polyhedra. *Computer Vision, Graphics, and Image Processing*, 51:20–37, 1990.

[234] A. J. Storkey and C. K. I. Williams. Image modeling with position-encoding dynamic trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):859–871, 2003.

[235] T. M. Strat and M. A. Fischler. Context-based vision: Recognizing objects using information from both 2d and 3d imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1050–1065, 1991.

[236] A. Telea, C. Sminchisescu, and S. Dickinson. Optimal inference for hierarchical skeleton abstraction. In *Proceedings, International Conference on Pattern Recognition*, pages 19–22, 2004.

[237] J. M. Tenenbaum and H. G. Barrow. Experiments in interpretation-guided segmentation. *Artificial Intelligence*, 8(3):241–274, June 1977.

[238] D. Terzopoulos and D. Metaxas. Dynamic 3D models with local and global deformations: Deformable superquadrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):703–714, 1991.

[239] D. Terzopoulos, A. Witkin, and M. Kass. Symmetry-seeking models and 3D object recovery. *International Journal of Computer Vision*, 1:211–221, 1987.

[240] D. Thompson and J. Mundy. Three-dimensional model matching from an unconstrained viewpoint. In *Proceedings, IEEE International Conference on Robotics and Automation*, pages 4:208–220, 1987.

[241] S. Todorovic and N. Ahuja. Extracting subimages of an unknown category from a set of images. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 927–934, 2006.

[242] S. Todorovic and N. Ahuja. Unsupervised category modeling, recognition, and segmentation in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear.

[243] S. Todorovic and M. C. Nechyba. Dynamic trees for unsupervised segmentation and matching of image regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1762–1777, 2005.

[244] A. Torralba, R. Fergus, and W. T. Freeman. Tiny images. Technical report, Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, 2007.

[245] A. B. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003.

[246] J. K. Tsotsos. Knowledge organization and its role in representation and interpretation for time-varying data: the alven system. *Computational Intelligence*, 1:16–32, 1985.

[247] J. K. Tsotsos. A 'complexity level' analysis of immediate vision. *International Journal of Computer Vision (Marr Prize Special Issue)*, 2(1):303–320, 1988.

[248] J. K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Sciences,*, 13(3):423–445, 1990.

[249] J. K. Tsotsos. On the relative complexity of passive vs. active visual search. *International Journal of Computer Vision*, 7(2):127–141, 1992.

[250] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[251] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006, October 1991.

[252] S. Ullman, M. Vidal-Naquet, and E. Sali. Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, 5(7):1–6, 2002.

[253] F. Ulupinar and R. Nevatia. Perception of 3-D surfa¡ces from 2-D contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:3–18, 1993.

[254] L. Vaina and M Jaulent. Object structure and action requirements: A compatibility model for functional recognition. *International Journal of Intelligent Systems*, 6:313–336, 1991.

[255] M. van Eede, D. Macrini, A. Telea, C. Sminchisescu, and S. Dickinson. Canonical skeletons for shape matching. In *Proceedings, International Conference on Pattern Recognition*, pages 64–69, 2006.

[256] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[257] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *ACM Conference on Human Factors in Computing Systems*, pages 319–326, 2004.

[258] S. Wachsmuth, S. Stevenson, and S. Dickinson. Towards a framework for learning structured shape models from text-annotated images. In *HLT-NAACL03 Workshop on Learning Word Meaning from Non-Linguistic Data*, 2003.

[259] J. Wang, V. Athitsos, S. Sclaroff, and M. Betke. Detecting objects of variable shape structure with hidden state shape models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):477–492, 2008.

[260] W. Wang, I. Pollak, T.-S. Wong, C. A. Bouman, M. P. Harper, and J. Mark Siskind. Hierarchical stochastic image grammars for classification and segmentation. *IEEE Transactions on Image Processing*, 15(10):3033–3052, 2006.

[261] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 2101–2108, 2000.

[262] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proceedings, European Conference on Computer Vision*, pages 18–32, 2000.

[263] D. Weinshall and M. Werman. On view likelihood and stability. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):97–108, February 1997.

[264] I. Weiss and M. Ray. Recognizing articulated objects using a region-based invariant transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1660–1665, 2005.

[265] M. Wertheimer. Laws of organization in perceptual forms. In W. Ellis, editor, *Source Book of Gestalt Psychology*. Harcourt, Brace, New York, NY, 1938.

[266] J. M. Winn and N. Jojic. Locus: Learning object classes with unsupervised segmentation. In *Proceedings, International Conference on Computer Vision*, pages 756–763, 2005.

[267] P. Winston, T. Binford, B. Katz, and M. Lowry. Learning physical description from functional descriptions, examples, and precedents. In *Proceedings, AAAI*, pages 433–439, Palo Alto, CA, August 1983.

[268] K. Wu and M. D. Levine. 3-d shape approximation using parametric geons. *Image and Vision Computing*, 15(2):143–158, 1997.

[269] Y. N. Wu, Z. Z. Si, H. F. Gong, and S. C. Zhu. Active basis for deformable object modeling, learning and detection. *International Journal of Computer Vision*, to apear, 2008.

[270] S. X. Yu and J. Shi. Object-specific figure-ground segregation. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–45, 2003.

[271] M. Zerroug and G. G. Medioni. The challenge of generic object recognition. In *Object Representation in Computer Vision, International Workshop, New York City, NY, USA*, pages 217–232, 1994.

[272] M. Zerroug and R. Nevatia. Volumetric descriptions from a single intensity image. *International Journal of Computer Vision*, 20(1/2):11–42, 1996.

[273] L. Zhu, Y. Chen, and A. Yuille. Unsupervised learning of probabilistic grammar-markov models for object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear, 2008XX.

[274] S. Zhu and A. L. Yuille. Forms: a flexible object recognition and modelling system. *International Journal of Computer Vision*, 20(3):187–212, 1996.

[275] S.-C. Zhu and D. Mumford. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, 2(4):259–362, 2007.