CHAPTER 1

# The Evolution of Object Categorization and the Challenge of Image Abstraction

Sven J. Dickinson

## 1.1 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 1.1. 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 (Fig. 1.1) 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, nonaccidental relations [152] must exist between them. Although such relations could be appearance-based, such as color and texture affinity, appearance is seldom 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

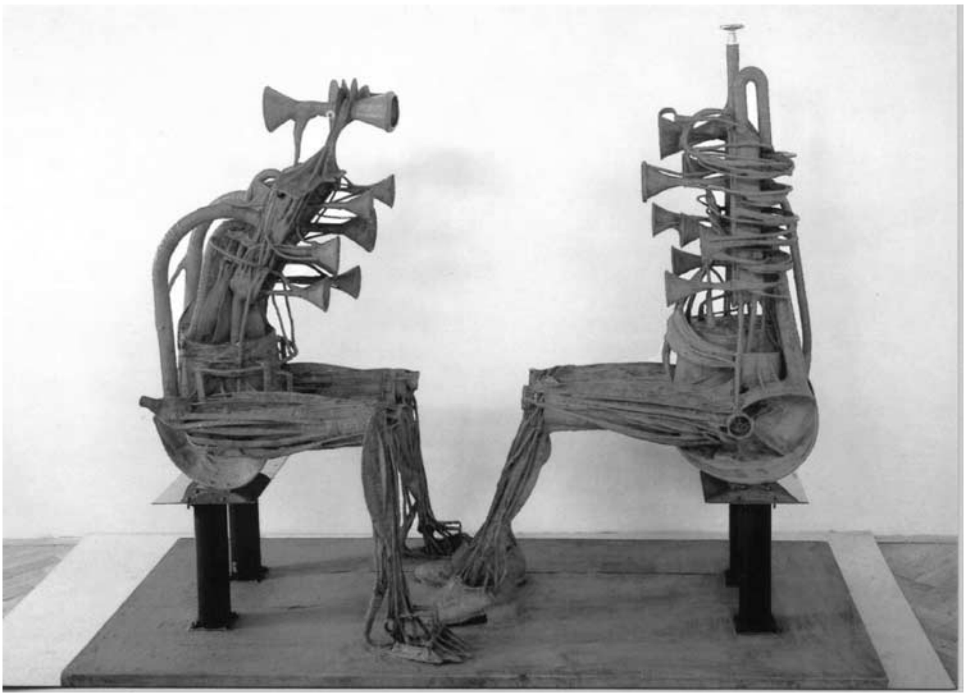


**Figure 1.1.** The two shapes depicted in this statue clearly represent two humanoid figures seated and facing each other. At the level of local features, the figures are unrecognizable. At a more abstract level, however, the coarse parts of the figures begin to emerge, which, along with their relations, 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. (See color plate 1.1.)

recognition researchers since the 1960s [198]. However, the grouping of causally (i.e., nonaccidentally) related features is necessary but not sufficient for object categorization. Returning to Figure 1.1, 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. The assumption that a one-to-one correspondence exists 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 1.2, which includes an elliptical part corresponding to the torso.[1] Under a one-to-one correspondence

[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 1.2 might be interpreted as the projections of ellipsoids.

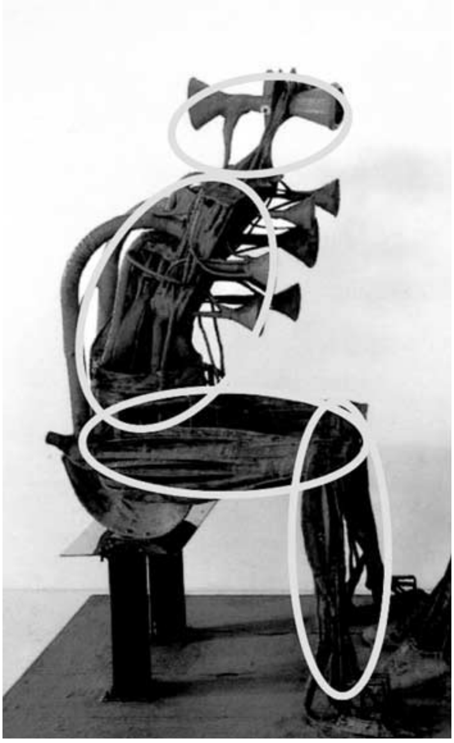


**Figure 1.2.** A shape abstraction of the seated humanoid on the left in Figure 1.1. Note that the boundaries of the shape abstraction do not map one-to-one to (or align well with) local features (e.g., contours) in the image. (See color plate 1.2.)

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; that is, it is not simply a subset of the grouped image features. Moreover, although 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 that 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 recover effectively 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.

## 1.2 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 1.3. In the 1970s, the recognition community focused on generic

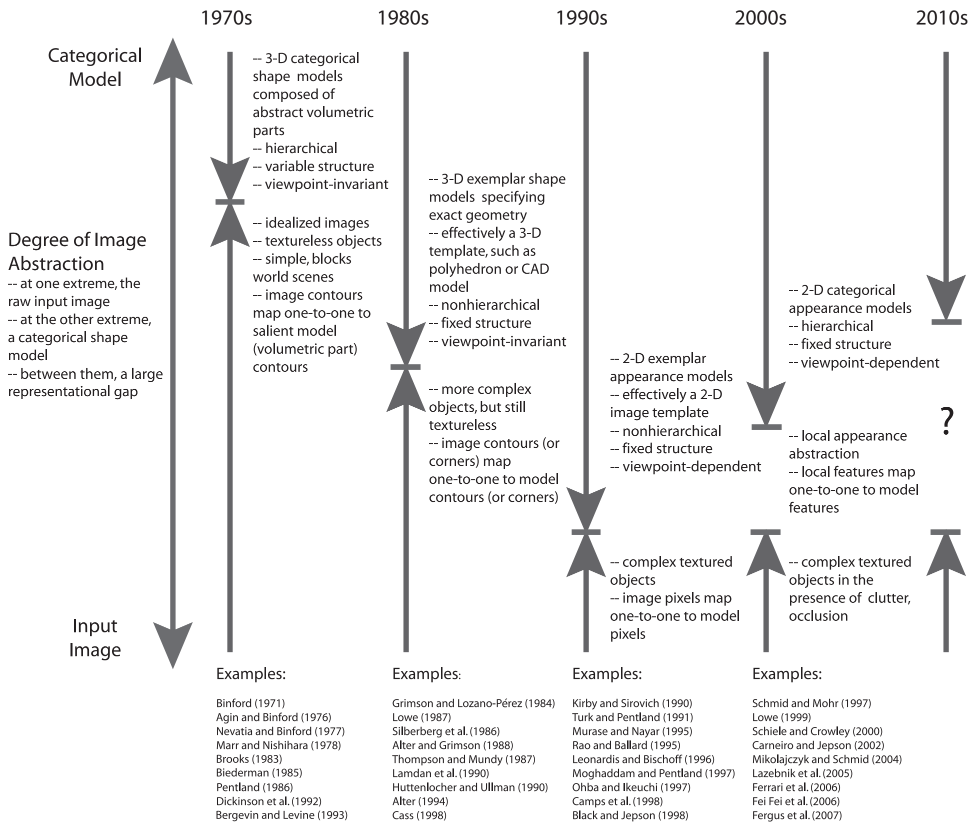
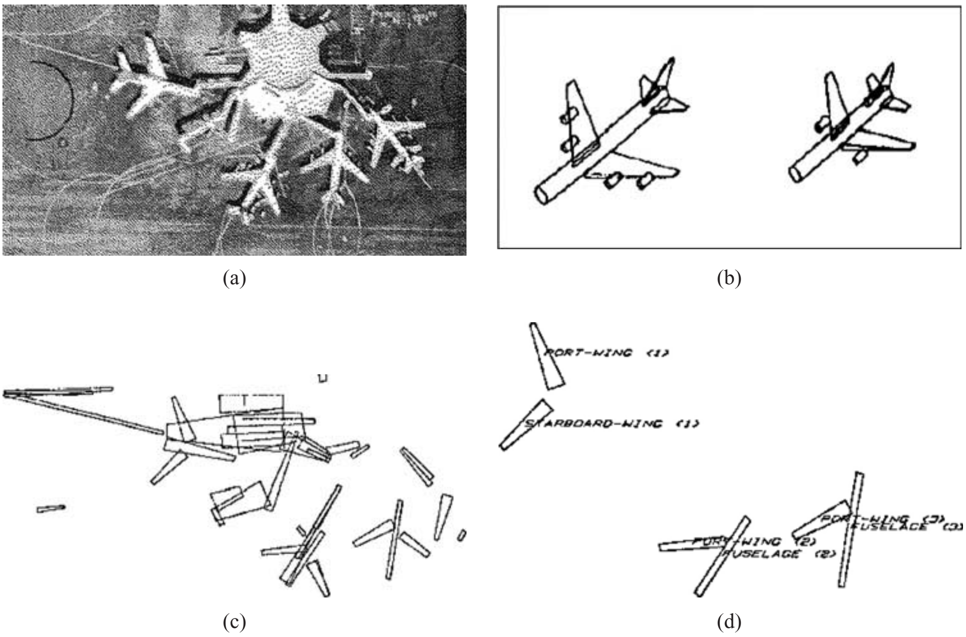

1970s
1980s
1990s
2000s
2010s
Categorical Model
Degree of Image Abstraction
-- at one extreme, the raw input image
-- at the other extreme, a categorical shape model
-- between them, a large representational gap
Input Image
-- 3-D categorical shape models composed of abstract volumetric parts
-- hierarchical
-- variable structure
-- viewpoint-invariant
-- idealized images
-- textureless objects
-- simple, blocks world scenes
-- image contours map one-to-one to salient model (volumetric part) contours
-- 3-D exemplar shape models specifying exact geometry
-- effectively a 3-D template, such as polyhedron or CAD model
-- nonhierarchical
-- fixed structure
-- viewpoint-invariant
-- more complex objects, but still textureless
-- image contours (or corners) map one-to-one to model contours (or corners)
-- 2-D exemplar appearance models
-- effectively a 2-D image template
-- nonhierarchical
-- fixed structure
-- viewpoint-dependent
-- complex textured objects
-- image pixels map one-to-one to model pixels
-- 2-D categorical appearance models
-- hierarchical
-- fixed structure
-- viewpoint-dependent
-- local appearance abstraction
-- local features map one-to-one to model features
-- complex textured objects in the presence of clutter, occlusion
?
Examples:
Binford (1971)
Agin and Binford (1976)
Nevatia and Binford (1977)
Marr and Nishihara (1978)
Brooks (1983)
Biederman (1985)
Pentland (1986)
Dickinson et al. (1992)
Bergevin and Levine (1993)
Examples:
Grimson and Lozano-Pérez (1984)
Lowe (1987)
Silberberg et al. (1986)
Alter and Grimson (1988)
Thompson and Mundy (1987)
Lamdan et al. (1990)
Huttenlocher and Ullman (1990)
Alter (1994)
Cass (1998)
Examples:
Kirby and Sirovich (1990)
Turk and Pentland (1991)
Murase and Nayar (1995)
Rao and Ballard (1995)
Leonardis and Bischoff (1996)
Moghaddam and Pentland (1997)
Ohba and Ikeuchi (1997)
Camps et al. (1998)
Black and Jepson (1998)
Examples:
Schmid and Mohr (1997)
Lowe (1999)
Schiele and Crowley (2000)
Carneiro and Jepson (2002)
Mikolajczyk and Schmid (2004)
Lazebnik et al. (2005)
Ferrari et al. (2006)
Fei Fei et al. (2006)
Fergus et al. (2007)

**Figure 1.3.** The evolution of object categorization over the past four decades (see text for discussion).

(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 1.4 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 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



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

**Figure 1.4.** 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.

emerged in the 1970s, 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 nonaccidental relations in feature grouping;
4. the need for distributed representations composed of shareable 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 and abstraction hierarchies;
6. the need for scalability to large databases – that is, 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; and
7. the need for variable structure – that is, the number of parts, their identities, and their attachments may vary across the exemplars belonging to a category.

The 1980s 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 1.5 illustrates an example output from Lowe's SCERPO system, which recognized a 3-D polyhedral template of an

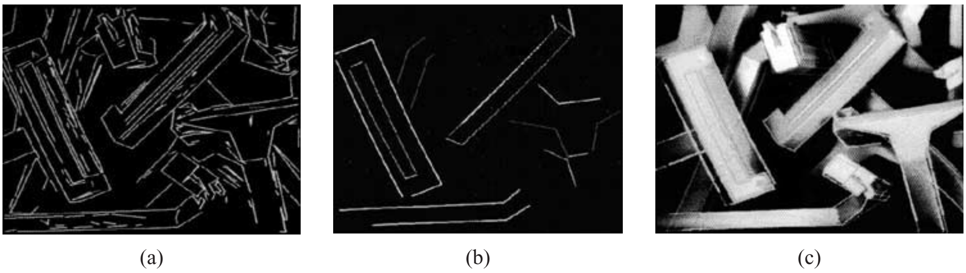
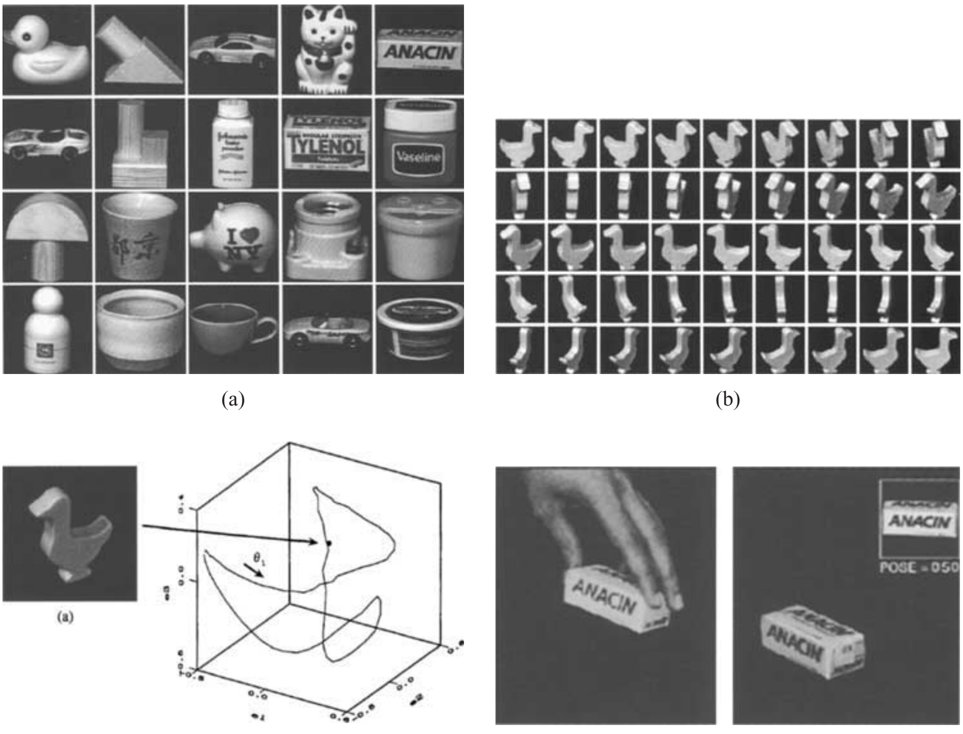


**Figure 1.5.** Lowe's SCERPO system [152] used perceptual grouping to prune hypothesized correspondences between image contours and polyhedral edges: (a) extracted edges; (b) extracted perceptual groups; and (c) detected objects and their poses. Images courtesy of David Lowe. (See color plate 1.5.)

object from nonaccidental 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. Although 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, which required the model to capture the exact geometry of the object. Moreover, because 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, because 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 although both the generations of systems just discussed 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



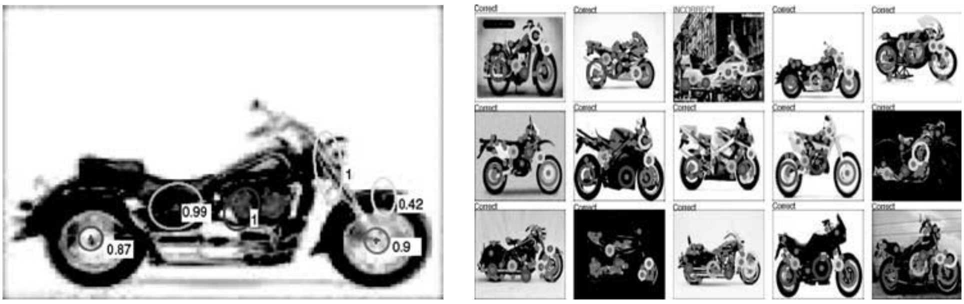
(a) (b)

(c) (d)

**Figure 1.6.** 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.

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]. Although 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 1990s, 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 1.6 illustrates an example output from the appearance-based (view-based) 3-D object recognition system of Murase and Nayar, which used PCA and nearest-neighbor search to reduce drastically the complexity of image correlation over a large database.



(a) (b)

**Figure 1.7.** Learning scale-invariant parts-based models from examples (from Fergus et al. [87]): (a) Learned motorcycle model with ellipses representing part covariances and labels representing probability of occurrence; (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. (See color plate 1.7.)

This time, the representational gap was eliminated by bringing the models all the way down to the image, which yielded models that were images themselves. The resulting systems could therefore recognize only exemplar objects, which were 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 (2000s) 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]). Whereas early systems characterized collections of such features as either 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 constraints, again drawing on classical shape modeling principles from the 1970s and 1980s. For example, Figure 1.7 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 systems of the 1970s and 1980s, 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 models of the 1990s, today's models have been brought close to the image; however, this trend is clearly reversing and starting to swing back. 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 the shape context of Belongie et al. [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, which clearly introduces additional complexity, is essential. To help manage this complexity, feature hierarchies have reemerged, in combination with powerful learning tools, to yield exciting new categorization frameworks [7, 6, 179, 41, 241, 92, 171, 4, 242, 273].[3] Figure 1.8 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 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. Moreover,

[2] In all fairness, appearance-based methods (based on explicit pixel values) implicitly encode both shape and nonshape 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.

[4] Indexing can take many forms, including hashing (e.g., [136, 93, 94]), decision trees (including kd-trees) e.g. [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).

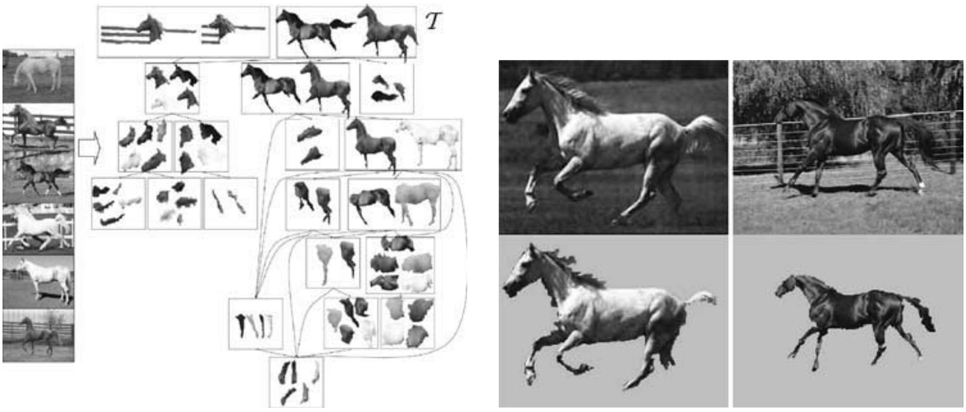


**Figure 1.8.** Learning hierarchical segmentation tree-based models from examples (from Todorovic and Ahuja [242]): (a) learned hierarchical tree-union model (right) from examples (left), capturing the recursive containment and spatial layout of regions making up the model; (b) example model detections (below) in query images (above). Images reproduced from [242], ©2008 IEEE with permission.

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 1970s, 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 to handcraft them. Yet 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 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. 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.

## 1.3 The Abstraction of Shape

In the 1970s, there was no shortage of abstract shape representations. For example, Binford’s generalized cylinder (GC) [29] (see Fig. 1.4) was a powerful, symmetry-based