The Role of Shape in Visual Recognition

Björn Ommer

Abstract Visual recognition requires a robust representation of typical object characteristics. Among all visual characteristics, shape plays a special role. It exhibits crucial invariance properties and captures the holistic structure of objects. However, shape cannot be extracted directly from an image, as it is an emergent property. Thus, representing shape is challenging, since it is related to several key problems of computer vision, such as grouping, segmentation, and correspondence problems. This paper reviews the development of shape in object recognition so far, discusses the reasons for the underlying developmental trends, and presents some promising recent contributions that point towards more accurate models of object structure.

1 Regularity, Structure, and Form

Our interaction with the world is constantly defined by the structure and characteristics of the objects around us, in particular by their form. This is only possible since our world exhibits an astounding degree of regularity. Let us now survey the prevalence of structure and the implications this has on cognition. Regardless, what scale we observe our universe on, order and regularity are evident everywhere. On a large scale, orderless clouds of matter condense due to gravitational attraction to form stars, stellar systems, and eventually galaxies consisting of billions of stars. On a scale that is directly accessible with our eyes we can, for instance, observe the complex, ordered patterns and forms exhibited by animals, plants, and non-living matter on earth. Examples are the symmetry and self-similarity featured by ferns, sea stars, or snowflakes. Finally, on an even smaller scale, the highly complex structure of DNA controls the development, functioning, and eventually the form of all living organisms. Moreover, the temporal domain features periodical structures such as the hydrologic cycle, the ever repeating seasons of the year, or our heartbeat.

It is astonishing that such complex, highly ordered structure even exists, since the second law of thermodynamics implies that the entropy (the degree of "disorder") of an isolated system—the universe in the most general case—is monotonically increasing. Moreover, not only the mere existence of order and structure, but its robustness to disrupting factors is as striking as it is necessary for the existence of life and our world as we know it. Consequently, it is self-evident that regularity and structure also play an important role in human thinking. Man has always been striving for a limited set of simple rules, laws, or relationships that, together with some

Björn Ommer
Heidelberg Collaboratory for Image Processing *(HCI)* & Interdisciplinary Center for Scientific Computing *(IWR)*, University of Heidelberg, e-mail: ommer@uni-heidelberg.de

S.J. Dickinson, Z. Pizlo (eds.), *Shape Perception in Human and Computer Vision*,
Advances in Computer Vision and Pattern Recognition,
DOI 10.1007/978-1-4471-5195-1_25,

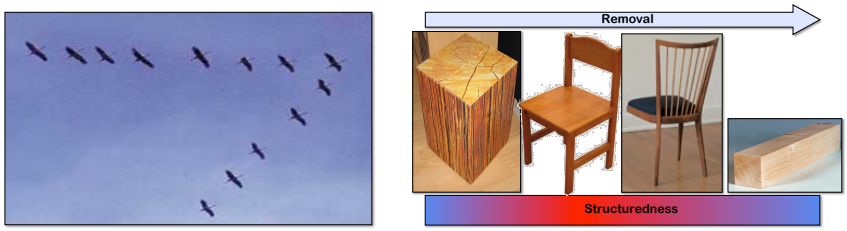
simple physical entities, would explain complex entities and thereby make the world comprehensible. When investigating these laws of nature, the scientific method has always exploited the regularity and order of our world. Seminal examples are the discovery of Newton's law of universal gravitation, which applies to apples as it does to extraterrestrial bodies like the moon and the prediction of the periodic table of (then mostly unknown) elements by Mendeleev. Consequently, only the regularity and order of our world makes it possible to learn from the past about the future, thus rendering learning and inference feasible.

1.1 The Nature of Shape

Recognizing objects and dealing with them depends on their structure and characteristics. With our different senses we observe different modalities and, thus, different properties of objects. For visual perception the most important features are appearance and shape. Whereas appearance comprises aspects such as the reflectivity, color, and texture, shape represents the form or Gestalt of objects. Commonly shape is thought of as a feature of the object silhouette, e.g., the form of a boundary contour [25, 44, 57] or region [3], whereas appearance describes the properties of the face of the surface surrounded by the boundary. Thus, both can be seen as dual characteristics of an object, one being based on contour shape, the other on region appearance. Nevertheless, other notions of shape beyond the form of boundary contours have been utilized as well, such as the spatial configuration of patches in part-based models, e.g., [22, 33, 21, 42], or the spatial layout of landmark points in procrustes analysis [17]. Given a set of image patches or coordinates of landmark points, we need to combine all these distributed observations to obtain a representation of the object (this is the *binding problem* in perception [45]) and segregate them from spurious clutter. Individual local features typically do not contain sufficient information about an object and, thus, there is a large *semantic gap* [51] between local measurements and semantic concepts such as object categories. In this context, shape can be thought of as the "glue" that combines all local features by ensuring a sound overall spatial layout and thereby capturing the co-occurrence of all features. This spatial structure or geometric configuration of an object is commonly referred to as its *shape* [52, 29, 8, 17, 50]). Kendall [29] has given an informal definition of shape that has been aptly paraphrased by Dryden and Mardia [17]:

> "*Shape* is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object."

Visual object recognition requires then to solve the *correspondence problem*—features of a test image have to be matched against the descriptors of a learned representation, e.g., the complete boundary contour, patches, or keypoints. For an optimal assignment of query features to model features, local descriptor correspondences as well as the global spatial structure need to be handled at the same time [5]. The matching process is based on the assumption that objects do not scatter features arbitrarily in an image. This assumption is in turn founded on the structure and regularity of the visual world.



(a) (b)

Fig. 1 The emergence of shape. a) The triangular shape of the flock of birds is an emergent property that is not inherent in any of its components, i.e., no individual bird exhibits the characteristic of the triangle, only their ensemble does. b) When removing parts of the block of wood, structure starts to emerge and it persists even when individual parts such as the leg of the chair are lost. A further removal, however, destroys the structure again.

The Special Role of Shape: Invariance and Emergence

Among all visual characteristics, shape plays a special role. As indicated by Kendall's definition, shape is not only invariant to geometric transformations such as translation, rotation, and scaling. It is also invariant to changes of appearance, i.e., varying illumination, reflectivity, color, or texture. Therefore, shape is crucial for rendering vision robust to our ever-changing environment and it is key to enable recognition in adverse situations such as under low light.

Shape is, however, special in another respect. Whereas appearance can be directly perceived or measured by (semi-)locally observing brightness, color, or texture, shape cannot be captured directly. The shape of a hand is not immanent in any image pixel or edge; neither is it captured by individual photoreceptors or the receptive fields of retinal ganglion cells. Similarly, the triangular form of the flock of birds in Fig. 1(a) is not inherent in any single bird. So how can we represent shape, if it cannot be measured directly? Shape is an *emergent* property that only evolves from the ensemble of foreground stimuli once background clutter has been suppressed. Therefore, perception and modeling of shape directly depend on several other processes that are mutually interlinked. A *grouping* of foreground parts is needed to obtain object shape and *segmentation* segregates foreground from distractors. Grouping again consists of a data-driven bottom-up process and a top-down registration based on learned object models. As argued by Gestalt psychology [55], there exist cognitive processes of perceptual organization that follow the law of *Prägnanz* thereby seeking simple, robust groupings. Gestalt laws such as *good continuation* or *closure* yield a purely data-driven grouping that is directly based on the visual stimulus (Fig. 2 left). However, there are also complex grouping processes that require object knowledge and reasoning about them such as Fig. 2 right. These processes are in the spirit of cognitivism and they present a correspondence problem, i.e. registering parts of the stimulus to previously learned object models.

Finally, shape is robust with respect to missing parts and clutter. As can be seen in Fig. 1(b), the operation of part removal creates structure and eventually annihilates it. Removing clutter lets structure (the shape of a chair) emerge. This structure is robust to further removal of object parts, but eventually it disappears and we are again

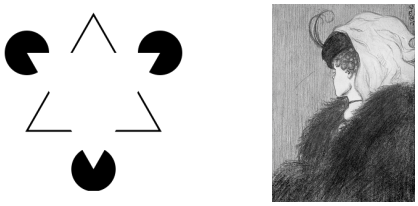


Fig. 2 Left: Kaizsa triangle, illusionary contours due to data-driven, bottom-up perceptual grouping. Right: young/old lady, ambiguous optical illusion due to top-down reasoning.

left with a mere block of wood. Robustness with respect to missing parts depends on the content of the parts. As argued by Attneave [2], points of high curvature are especially informative. [6] has proposed psychophysical experiments that underline this claim and Fig. 3 demonstrates how the recognition system of [47] approximates shape using a sparse representation with variable degree of detail.

2 Shape Representation for Visual Recognition

Computer-based object recognition has been actively pursued for half a century and a wide range of shape representations have been investigated. Over these years of research on shape models for visual recognition, several major trends evolved, disappeared, and reappeared again. Let us now review these broad movements and the influence they had on vision research.

2.1 The Days of Geometry: Blocks, Cylinders, and Acronyms

The first artificial object recognition systems entered the stage in the late 1950s, adopting ideas from signal processing, formal logic, and statistics and being tightly linked to the then newly proposed theme of artificial intelligence coined by John McCarthy and Marvin Minsky at the Dartmouth conference of 1956. 1963 can then be viewed as the real advent of the field when L.G. Roberts [46] presented his recognition system and proposed an edge detector, a line fitting, and a feature grouping procedure. To facilitate these first big steps into computer vision with the limited hardware resources of the day, significant simplifying assumptions were made. Systems were confined to a *blocks world* consisting of only polyhedral shapes on uniform background. While these restrictions enabled a sound theoretical investigation, they lead to vision algorithms that were founded on numerous unrealistic assumptions. Thus, later research tried to alleviate these restrictions by allowing for more and more complex scenes. Examples are Guzman's system [25] for recognizing 2-D curved object line drawings and Binford's *generalized cylinders* [7] that were taking curved shapes to 3-D. Based on the generalized cylinders, Brooks [10] constructed the symbolic reasoning system *ACRONYM* that utilized geometric constraints to prove the existence of parameterized configurations. Biederman [6] then proposed *geons*, a universally applicable dictionary of volumetric primitives for compositional recognition. To bridge the gap between 2-D images and the 3-D

Fig. 3 Shape is robust with respect to missing parts and shape information is predominantly concentrated at points of high curvature. Example sparse shape representation taken from [47].

world, Marr [37] introduced the primal sketch and the $2^1/_2$-D sketch. While many of these early systems were limited by requiring bottom-up extraction of object boundaries, Lowe's *SCERPO* system [35] directly searches for non-accidental combinations of edgels. A main theme of research in these days was model-based vision by posing recognition as a correspondence problem between a model and contours in the image, e.g., [27]. However, with *aspect graphs* the orthogonal movement of view-based approaches started in the 1970s (e.g. [30] and see [16] for a later bridge between aspect graphs and geons) although it was later discovered that this framework suffers from severe complexity issues. Moreover, *moment invariants* [26] received considerable interest in this era, but later this theme lost momentum due to limited representational power in case of only a single view.

All in all a main theme of the 1960s–1980s was geometry especially based on the shape of boundary contours. Moreover, representations were typically hierarchical and object centered.

2.2 The Dawn of Appearance

As a response to setbacks of geometric approaches based on object boundary shape and with improvements in computational resources, the 1990s saw the rise of appearance methods. By applying principle component analysis to the intensity image, Turk and Pentland removed noisy dimensions and obtained *eigenfaces* [53]. More general eigenspace representations were analyzed by [39] and in the comparison by [11] template matching was superior to keypoint geometry. However, global image transformations such as translation, scaling, or illumination changes have to be removed in a preprocessing stage before applying appearance models such as the PCA-based approach. Therefore, sliding window procedures [48] or cascaded evaluation [54] are typically used. Moreover, the holistic object representation (the complete object is represented as one appearance patch of intensity values) leads to models of high complexity and renders them fragile with regard to variability in spatial structure as is the case for articulated objects. To address the latter problem, *deformable template matching* has been introduced [58] and prototypical deformations have been captured by *active appearance models* [12]. This approach compensates for variations in the spatial structure by applying a global transformation when matching templates. Another solution that is currently very popular are *part-based models* in spirit of the approach by Fischler and Elschlager [24]. These models represent an object as consisting of a number of specific parts that feature

characteristic spatial structure which can be modeled using a graph [31], a joint constellation model of all parts [22], or with probabilistic Hough voting [33].

In retrospective it can be observed how early contributions in the era of appearance models have abandoned spatial structure and shape only to see it reappear a few years later to handle articulation. Nevertheless, the main focus has been on appearance and compared to the previous geometric period, shape representation has become significantly more coarse, e.g., part-based models sampled at few interest points. Moreover, the view-based paradigm and shallow structures have dominated.

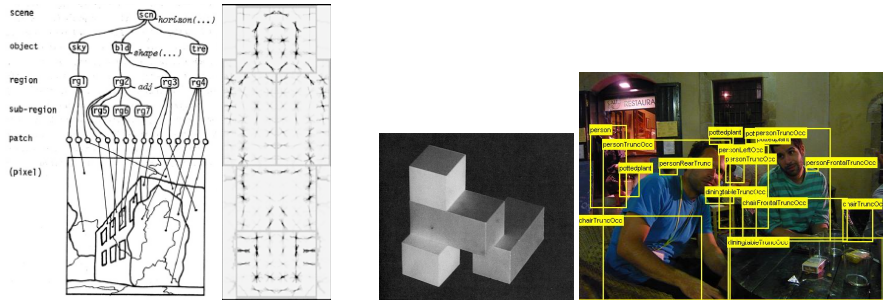
2.3 Textons Everywhere

The turn of the millennium clearly marks the advent of powerful semi-local feature descriptors and interest point detectors. Compared to appearance patches that represent objects as a matrix of intensities or colors, these features gained invariance to geometric deformations, illumination changes, and noise by histogramming over edge pixels and their orientations, thereby again picking up the idea of *textons* [28]. The influential *SIFT* features introduced by David Lowe in [34] were followed by numerous other descriptors such as *shape context* [4] and *histograms of oriented gradients (HOG)* [14]. Popular object representations built upon these descriptors were *bag-of-feature models* [13], and models based on probabilistic latent semantic analysis such as [49]. These approaches model only feature co-occurrence and completely disregard spatial structure. By evaluating separate bag-of-features in cells of a regular grid, spatial pyramid kernels were used in [32] to add rigid grid-like structure to this framework. In effect this led again to a classical rigid template matching approach—this time however with bag-of-features over image sites replacing the intensity values of image pixels. To obtain additional flexibility to geometric deformations, Felzenszwalb et al. [20] combined rigid, regular-grid-like templates with part-based models. All of these template-based approaches utilize sliding windows. However, scanning over all locations and scales and evaluating a classifier is not only computationally costly but also lacks psychophysical motivation. These issues are tackled by voting methods such as [36, 43].

Texton features have successfully addressed the invariance issues of appearance patches. The potential of these powerful descriptors inspired early models like bag-of-features that abandoned spatial structure altogether, which returned later on again. However, compared to the geometric era, the spatial models were fairly simple, i.e. rigid templates, subsequently extended by star-shaped part models in the currently popular approach of [20]. All in all the root filter of the currently successful approach of [20] is a mere texton template—the whole object is represented as a spatially varying texture (cf. Fig. 4(b)), same being true for the parts as well.

2.4 Half a Century of Evolution—A Critique

Looking back on the development of shape models for visual recognition, some interesting trends become apparent. There have obviously been orthogonal movements as well, but these could be seen as the mainstream developments of the field.



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

Fig. 4 a) Full-up scene interpretation from the 1970s, [40] and b) currently popular, template-like texton models [20]. c) Benchmark problems of the early days, e.g., blocks world and d) present day recognition benchmarks such as PASCAL VOC [19].

Complexity of Shape Models: Whereas the early years saw a focus on rich object shape and scene models [40], currently popular representations such as [20] describe objects as a mere texton. After only a few years of development, Roberts [46] had invented many of the key components of modern recognition systems in 1963. Some 15 years later, models that contained almost anything up to a complete scene interpretation had been proposed [40], Fig. 4(a). Moving another 30 years forward in time and comparing these rich models of the 1970s with the currently popular, template-like texton models (e.g. Fig. 4(b)) this could be seen as a great setback. However, the judgment depends on the vantage point and requires further discussion. So what went wrong, what right, and why?

Real World Benchmarks and Performance: Although the representation of shape has become less intricate, there has been a dramatic improvement in performance. Whereas blocks world (Fig. 4(c)) and other early scenarios used for system evaluation were artificial and simplistic, present day benchmarks made significant steps towards the real world recognition challenge, cf. Fig. 4(d). Rather than detecting blocks in front of uniform background, multi-scale detection of diverse object categories in cluttered natural scenes [19] has become a main theme, thus dealing with difficult problems such as large intra-class variability, many categories, segmentation of clutter, and multi-scale detection. Despite this positive development it should however be noted that several of the simpler problems in less realistic scenes are still unsolved, that benchmarks such as [19] are also only caricatures of reality, and, most importantly, they are blending numerous unsolved problems of vision and do not allow to evaluate the progress on individual subproblems.

Dimensionality and Flexibility: While there has been a trend towards less complex shape models, the complexity of the low-level descriptors has increased enormously. Simple parametrized surfaces were followed by PCA applied to appearance templates before texton features became popular and increased dimensionality from 128-D (SIFT) over 10 000-D [20] to over 160 000-D in [15]. Dealing with this high dimensionality became only possible by adopting landmark contributions from machine learning and pattern recognition such as kernel methods. However, given the

limited amount of training and test data, curse of dimensionality is a serious concern in light of these developments. Nevertheless, there are also very promising developments. Compared to simple condition-action-rules such as the production rules of [40], machine learning has lead to flexible systems that automatically adapt to training data.

3 Quo Vadis?

Visual object recognition has made great progress, especially in terms of the realism of its benchmark problems, the flexibility of the developed systems, and the retrieval performance. However, there has been a shift in mainstream research to focus on much coarser and less accurate shape representations than in the early days and on high-dimensional low-level descriptors. Many reasons including practicability (complex low-level features are readily available), universality (coarse structure models make less restricting assumptions), and feasibility (simpler structure models can be easily adapted from the literature) have led to this trend. Nevertheless, using textons to represent complete objects and their shape is obviously only a very crude approximation. In effect the rich spatial structure of shape is basically treated as a mere texton, cf. Fig. 4(b).

3.1 Shape: Representing Statistical Dependencies between Parts

We have seen that shape is an emergent property that captures statistical dependencies between local features by aggregating descriptors, e.g., those that lie along object boundaries. However, commonly used part-based methods such as propabilistic Hough voting [33, 36] fail to model these dependencies and simply treat heavily overlapping features that are sampled close to another as being independent. Voting then sums over the mutually dependent feature votes. The same critique also applies to sliding windows based on texton templates such as the popular approach [20]. By utilizing a linear classifier to combine the cells of the rootfilter (non-linear classification is not feasible due to complexity), mutual dependencies cannot be learned. Consequently, the two most common approaches to visual object detection—voting and sliding window texton templates—are treating objects to be a mere sum of their parts, cf. [56]. This assumption is against the fundamental conviction of Gestalt theory that the whole object is different from the sum of its parts [55] and that shape emerges from all constituents by explicitly capturing mutual part relationships.

Compositionality

We cannot measure shape directly in an image. Neither are joint models of all parts such as constellation models [22] feasible for the usually large quantities of parts. How can we then model part dependencies effectively in a way that lets shape emerge? To bridge the large gap between local features and holistic object shape,

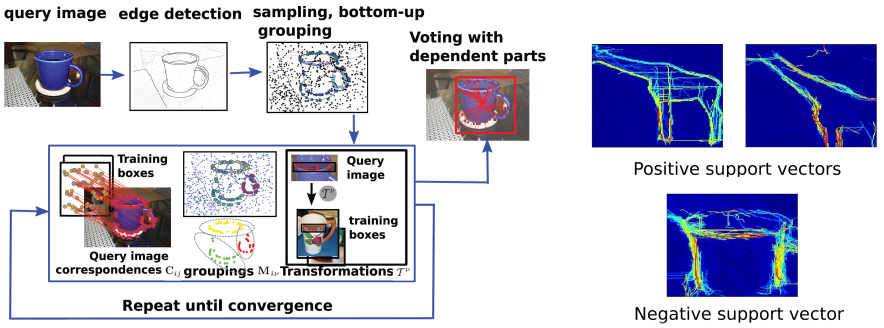


Fig. 5 a) Compositions by grouping dependent parts and solving the correspondence problem [56]. b) Learning discriminative joint placements of contours yields object shape (sample support vectors for giraffes) [57].

hierarchical approaches have been proposed. These were highly popular in the early days when only weak features such as edges or geometric primitives were used. Hierarchies then lost momentum with the arrival of powerful features when some vision problems could be addressed on the level of features without reasoning about more complex object structure (e.g. bag-of-features). Recently, however, compositional methods have shown to be effective in combining local descriptors in hierarchies that culminate in a holistic representation of object structure with all its flexibility. *Compositionality* refers to the prominent ability of human cognition to represent entities as hierarchies of meaningful and generic parts. As demonstrated by Biederman [6] the atomic constituents are usually much simpler than the scenarios described by their compositions. Moreover, these parts are generic so that they can be used for representing numerous object categories, thus being essential for the flexibility of human cognition. The power of compositionality is not rooted in the atomic parts but stems from modeling the dependencies between the parts. In particular, seeking *non-accidental* [35] part relationships renders vision robust with regard to clutter and object variability. Written language can for instance be represented with just a mere 26 letters, where meaning is not inherent in individual characters but only results from their compositions, i.e., words and sentences.

Based on these ideas, a compositional system for category-level recognition has been presented in [41]. Using the Gestalt laws of perceptual organization, candidate compositions are formed. Then a discriminative strategy is employed to retain only characteristic compositions. This unsupervised discovery of mid-level discriminative compositions [41, 42] establishes a layer of intermediate abstractions in the resulting hierarchy. In [42] a graphical model combines multiple layers of compositions and scene context while learning follows a Bayesian approach and is based on cross-validation. Whereas these methods learn the compositional structure without supervision, poselets [9] have followed-up on these ideas by requiring additional supervision information for labeling object specific compositions. Fidler et al. [23] have build a hierarchy of constellation models to speed-up multi-class classification.

A Compositional Shortcut

Compositional hierarchies are ideal for representing object structure by modeling relationships between parts. However, we cannot merely stack an arbitrary number of layers on top of each other and expect a functioning hierarchy. Noise and other disturbances at the feature level can be amplified by successive representation layers. Consequently, a recent development has been to avoid arbitrarily deep hierarchies by iteratively optimizing a single layer of compositions. In [56] this is achieved by integrating compositionality into Hough voting. Rather than incorrectly assuming parts to be independent, dependent parts are grouped while solving the correspondence problem and forcing all parts within the resulting compositions to agree on a concerted object hypothesis, Fig. 5(a). As a result three key problems of vision are addressed jointly, i) grouping object parts into meaningful compositions, ii) establishing correspondences between query object and training samples, and iii) foreground/background segregation. To avoid bottom-up grouping altogether, [57] applies maximum margin multiple instance learning to obtain a dictionary of meaningful contours. Shape is then represented by learning the consistent joint placement of all these contours, Fig. 5(b). Object detection and the assembling of their shape are addressed simultaneously. Contour co-activation captures part dependencies and a discriminative approach yields consistent joint placements of all model contours. The dual problem of shape-based compositional region grouping has been addressed in [38]. Finally, compositionality and shape are not limited to representing individual objects. [1] has presented a video parsing approach to abnormality detection. They parse complete scenes by establishing a set of shapes that jointly represent all the foreground, thereby taking interactions between object shapes into account.

4 Conclusion and Outlook

Among all visual characteristics, shape is of crucial importance. Shape exhibits important invariance properties, unites heterogeneous scattered features, and captures the holistic structure of objects. Being an emergent property, shape cannot be measured directly, thus rendering its representation challenging. Consequently, a large body of vision research has focused on modeling object structure during the last half century. Broad trends during this time were i) a geometric era with an emphasis on spatial structure, boundary contours, hierarchies, and model-based approaches, ii) appearance models with comparably coarse shape representation, shallow structures, and a view-based paradigm, and recently iii) an era of powerful texton-based features with bag-of-features, part-based models, and texton templates. Over the years the performance of vision systems, the complexity of recognition benchmarks, and the flexibility of the learning algorithms has increased, significantly. Compared to the early days there is, however, an emphasis on relatively coarse models of object shape (to the point of treating shape as a spatially varying texture) and a trend towards ever increasing dimensionality (addressed in [18]). Moreover, there has been

a back and forth of interest in and complexity of shape models. The arrival of new features has typically first led to an increased interest in low-level representation followed by a later reemphasis of shape. Finally, hierarchical models based on compositionality have recently shown great potential for bridging the gap between local features and holistic shape. They capture non-accidental part dependencies to model structure and they have addressed key problems of vision such as top-down grouping, foreground/background segregation, and the correspondence problem.

References

1. B. Antic and B. Ommer. Video parsing for abnormality detection. In *ICCV*, 2011.
2. F. Attneave. Some informational aspects of visual perception. *Psych Rev*, 61(3), 1954.
3. R. Basri and D. Jacobs. Recognition using region correspondences. *IJCV*, 25:8–13, 1995.
4. S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):pp.509–522, 2002.
5. A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *CVPR*, pages 26–33, 2005.
6. I. Biederman. Recognition-by-components: A theory of human image understanding. *Psych Rev*, 94(2):pp.115–147, 1987.
7. T. O. Binford. Visual perception by computer. In *IEEE Conf on Systems and Control*, 1971.
8. F. L. Bookstein. Size and shape spaces for landmark data in two dimensions. *Statistical Science*, 1(2):pp.181–222, 1986.
9. L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.
10. R. A. Brooks. Symbolic reasoning among 3-D models and 2-D images. *AI*, 17(1–3), 1981.
11. R. Brunelli and T. Poggio. Face recognition: Features versus templates. *PAMI*, 15, 1993.
12. T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *ECCV*, 1998.
13. G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV*. *Workshop on Statistical Learning in Computer Vision*, 2004.
14. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.
15. T. Deselaers and V. Ferrari. Global and efficient self-similarity for object classification and detection. In *CVPR*, pages 1633–1640, 2010.
16. S. J. Dickinson, A. Pentland, and A. Rosenfeld. 3-D shape recovery using distributed aspect matching. 14(2):pp.174–198, 1992.
17. I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis*. John Wiley, 1998.
18. A. Eigenstetter and B. Ommer. Visual recognition using embedded feature selection for curvature self-similarity. In *NIPS*, 2012.
19. M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. PASCAL VOC'06, 2006.
20. P. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.
21. P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):pp.55–79, 2005.
22. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, pages 264–271, 2003.
23. S. Fidler, M. Boben, and A. Leonardis. A coarse-to-fine taxonomy of constellations for fast multi-class object detection. In *ECCV*, 2010.
24. M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, c-22(1):pp.67–92, 1973.
25. A. Guzman. Analysis of curved line drawings using context and global information. *Machine Inteligence*, 6:325–376, 1971.

26. M. K. Hu. Visual pattern recognition by moment invariants. *Trans on Inf Theory*, 8(2), 1962.
27. D. P. Huttenlocher and S. Ullman. Object recognition using alignment. In *ICCV*, 1987.
28. B. Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981.
29. D. Kendall. Shape manifolds, procrustean metrics and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):pp.81–121, 1984.
30. Jan J. Koenderink and A. J. van Doorn. The singularities of the visual mapping. *Biological Cybernetics*, 24:51–59, 1976.
31. M. Lades, J. C. Vorbrüggen, J. M. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42:pp.300–311, 1993.
32. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006.
33. B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV. Workshop on Stat Learn in Comp Vision*, 2004.
34. D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.
35. D. G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer, 1985.
36. S. Maji and J. Malik. Object detection using a max-margin hough transform. In *CVPR*, 2009.
37. D. Marr. *Vision*. W. H. Freeman, San Francisco, CA, 1982.
38. A. Monroy and B. Ommer. Beyond bounding-boxes: Learning object shape by model-driven grouping. In *ECCV*, pages 580–593, 2012.
39. H. Murase and S. K. Nayar. Visual learning and recognition of 3-d objects from appearance. *IJCV*, 14(1):pp.5–24, 1995.
40. Y. Ohta, T. Kanade, and T. Sakai. An analysis system for scenes containing objects with substructures. In *Intl Joint Conf on Pattern Recognition*, pages 752–754, 1978.
41. B. Ommer and J. M. Buhmann. Learning compositional categorization models. In *ECCV*, LNCS 3953, pages 316–329, 2006.
42. B. Ommer and J.M. Buhmann. Learning the compositional nature of visual object categories for recognition. *PAMI*, 32(3):501–516, 2010.
43. B. Ommer and J. Malik. Multi-scale object detection by clustering lines. In *ICCV*, 2009.
44. A. Opelt, A. Pinz, and A. Zisserman. Incremental learning of object detectors using a visual shape alphabet. In *CVPR*, pages 3–10, 2006.
45. A. Revonsuo and J. Newman. Binding and consciousness. *Consciousness and Cogn*, 8, 1999.
46. L. G. Roberts. *Machine Perception Of Three-Dimensional Solids*. PhD thesis, MIT, 1963.
47. J. Schlecht and B. Ommer. Contour-based object detection. In *BMVC*, 2011.
48. H. Schneiderman and T. Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *CVPR*, pages 45–51, 1998.
49. J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their localization in images. In *ICCV*, pages 370–377, 2005.
50. C. G. Small. *The Statistical Theory of Shape*. Springer, New York, NY, 1996.
51. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *PAMI*, 22:1349–1380, 2000.
52. D. W. Thompson. *On Growth and Form*. Dover, 1917.
53. M. Turk and A. Pentland. Eigenfaces for recognition. *J Cogn Neurosci*, 3(1):pp.71–86, 1991.
54. P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages 511–518, 2001.
55. M. Wertheimer. Untersuchungen zur Lehre von der Gestalt I. Prinzipielle Bemerkungen. *Psychologische Forschung*, 1:47–58, 1922.
56. P. Yarlagadda, A. Monroy, and B. Ommer. Voting by grouping dependent parts. In *ECCV*, pages 197–210, 2010.
57. P. Yarlagadda and B. Ommer. From meaningful contours to discriminative object shape. In *ECCV*, 2012.
58. A.L. Yuille, P.W. Hallinan, and D.S. Cohen. Feature extraction from faces using deformable templates. *IJCV*, 8(2):pp.99–111, 1992.