(1994). *Encyclopedia of Human Behavior*, V. S. Ramachandran, (ed.),
San Diego, CA: Academic Press, Vol. 4, pp. 503-512.

# VISUAL REPRESENTATION: FROM FEATURES TO OBJECTS

Michael J. Tarr

*Yale University*



## GLOSSARY

**Frame of Reference** The coordinate system used for encoding the spatial positions of features located within visual input.

**Object Representation** A mental structure that encodes selected properties of a given physical object, most often including shape information.

**Perceptual Organization** Processes by which spatially discrete local features within visual input are grouped to form meaningful structures, such as surfaces and objects.

**Structural Description** An object representation composed of three-dimensional parts and explicit relationships between them in a frame of reference based on the object.

**View-Based Representation** An object representation composed of images, each depicting the characteristic appearance of an object from a particular viewpoint in a frame of reference based on the observer.

**Visual Search** The process by which a set of features defining a unique element within a scene is located using both preattentive and serial attention mechanisms.

Light enters our eyes as an undifferentiated array of luminance points. However, because most of the light within the optic array is reflected off of surfaces and objects in the environment, inherent within this array is a tremendous amount of information about the structure of the world around us. Reconstructing this structure, that is, recovering the layout of the scene, is perhaps the most important goal of human visual perception. Accomplishing this involves many complex processes that work towards converting patterns of light and dark into *visual representations*. Much like the physical world, visual representations are coherent, organized structures that correspond to surfaces, objects, and scenes. It is these mental representations, not the patterns of light that fall upon the retina, that form the basis of our visual percepts and cognition, being used in mechanisms as varied as object recognition, visual reasoning, linguistic descriptions of the environment, and categorization.

## I. FEATURE PROCESSING

### A. Preattentive Vision

It is generally understood that our visual system does rudimentary feature analysis at the earliest stages of processing. In particular, the existence of cortical cells differentially sensitive to stimulus orientation indicates that a series of orientation-selective filters are used to locate the position and orientation of local edges within visual input. Likewise, a variety of other early visual pathways apparently serve as filters to locally compute properties such as size and color. At this stage, the output of visual processing may be considered a set of *feature maps* that independently code for the value of a given feature at each position within a scene. For instance, given the letter "A" as visual input, a feature map for orientation would encode a distinct local orientation at each spatial position covered by the "A" — creating an array of oblique left, oblique right, horizontal, and possibly indeterminate local features (see Figure 1). In a similar manner, maps may be formed for color, size, motion, depth, as well as other properties of the scene.

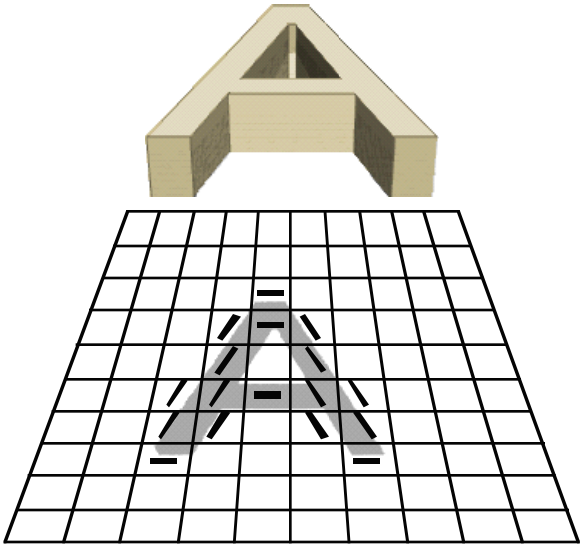

**Fig. 1.** A feature map for orientation given the letter "A" as an input. Dark bars in each cell indicate the computed local orientation for that position. The shadow of the "A" is provided for clarity and is not part of a feature map.

Evidence from visual search tasks indicates that different locations within an individual feature map are processed in parallel. For instance, imagine trying to locate a single vertical bar target among randomly placed distractors composed of oblique bars of the same shape and color (see Figure 2a); the time to perform this task is independent of the number of distractors present in the display. The target is said to "pop-out." That is, determining the presence of a target based on a feature difference within a single map does not require the application of a serial attention mechanism and, hence, is *preattentive*. Indeed, a similar "pop-out" effect has been found for target/distractor feature differences across color, size, motion direction, simple shape, and many other properties — each suggesting that there exists an independent feature map for a given perceptual dimension.

### B. Conjunction Search

*1. Feature-Integration*. In contrast to preattentive processing, a somewhat different situation exists when features must be combined across maps. For instance, imagine trying to locate a green-vertical target among randomly placed green-oblique and red-vertical distractors (see Figure 2b); the time to perform this task is dependent on the number of distractors present in the display, that is, increasing the number of distractors increases the time to locate the target. Therefore, determining the presence of a target based on a *conjunction* of features across two or more maps requires a serial item-by-item search. Furthermore, if a serial attention mechanism is used to locate a conjunction target and is moved randomly from location to location without repetition, then on the average the target should be found after searching through one-half of the items in the display. In contrast, displays with no target present should result in a search in which every distractor item is examined. Empirical evidence bears out this prediction, with no-target displays producing response times approximately twice as long as target-present displays with the same number of distractors. Thus, while individual perceptual dimensions may be processed in parallel, it is only through the application of serial attention to discrete locations that coherent percepts may be formed.

This role for serial search is consistent with earlier models of serial attention that center around the "spotlight" metaphor. The fundamental assumption of such models is that serial attention has the characteristics of a spotlight: it selects items within the spotlight, it is indivisible over spatial location, must be spatially moved from one location to another, and may be expanded or contracted to cover greater or smaller regions. However, while recent evidence suggests that there are many conditions under which this metaphor does not apply, questions about how attention is allocated do not significantly alter its role in the integration of features across maps.

*2. Guided Search*. While double conjunction searches (e.g., a conjunction between color and orientation) support the concept of serial search that combines information across feature maps, there is also evidence that additional information is used in guiding which items are actually examined. For instance, imagine a triple conjunction search in which the target is a large-green-vertical bar and the distractors are

small-red-vertical, small-green-oblique, and large-red-oblique. In such a search, the target shares exactly one feature with each of the distractors; the time to perform this task is actually less than that found for a double conjunction search using the same number of distractors. Moreover, the time to locate a target increases at a slower rate as the number of distractors is increased when compared to the rate of increase for double conjunctions. A simple serial search model would not predict these results: rather, it would predict that regardless of the number of features shared between the target and the distractors, once features must be combined across maps, a serial search must be used to locate the target. Consequently, the integration of features can not simply be a matter of searching item by item. Thus, more complex visual searches may be *guided* in some manner so as to reduce the overall number of items searched.

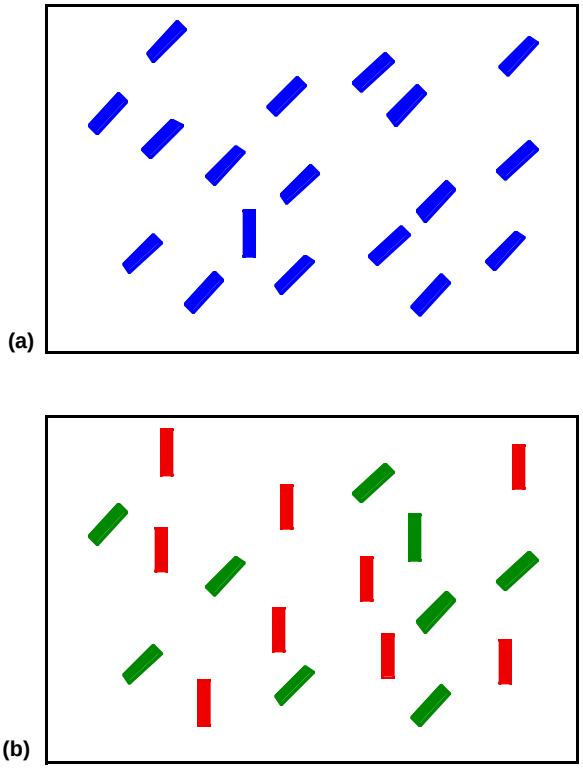


**Fig. 2.** Visual search tasks. (a) A preattentive search task: Find the vertical bar among the oblique bars. (b) A conjunction search task: Find the green-vertical bar among the green-oblique and red-vertical bars.

Another type of triple conjunction search yields evidence that helps to illuminate some of the mechanisms underlying complex visual search. For instance, imagine trying to locate a large-green-vertical target among small-green-vertical, large-green-oblique, and large-red-vertical distractors. In this search, the target shares two features with each of the distractors; the time to perform this task is comparable to that found for a double conjunction searches using the same number of distractors and the rate of increase per additional distractors is comparable as well. These results indicate that the information available to guide serial search mechanisms is dependent on the *relative* number of features present in the target as compared to the distractors. In both a double conjunction search and a triple conjunction search with double-feature distractors, the target differs from each distractor by only a single feature. However, in a triple conjunction search with single-feature distractors, the target differs from each distractor by at least two features. More recent evidence suggests that such examples reflect a more general principle in which visual searches become easier as the target and the distractors become less similar.

How might such information be represented preattentively so as to guide the serial stage in which features are combined? One possibility is that there exists an additional preattentive map that encodes the overall activation of all feature maps — an approximate sum of the features in common between the target and that distractor at each location. For a given location, the larger this sum is relative to other locations and noise inherent within the map, the more likely is the item at that location to be the target. Therefore, when the feature difference between the target and the distractors is great, the information provided by this map will allow the serial search mechanism to ignore some distractors and overcome any noise so as to quickly locate the target; in contrast, when the feature difference is smaller, the information provided by this map will not facilitate excluding many of the distractors or locating the target above the noise, consequently it will be less reliable in guiding the serial mechanism. Thus, preattentive mechanisms guide serial search so that the positions examined are not random; however, this guidance varies in efficiency depending on the featural relationship between the target and the distractors.

### C. Combining Features

Beyond visual search tasks, there is the more general question of how disparate features within a given object are combined to form a coherent percept. One final piece of evidence adds further weight to the hypothesis that serial attention is necessary to integrate features. Imagine very briefly viewing a display containing a green “X,” a red “S,” and a yellow “T” adjacent to each other; observers often report seeing *illusory conjunctions* of features, for instance, a red “X” or a yellow “S”. Apparently, without the benefit of time to allocate serial attention, features corresponding to a given spatial location are not *bound* together, that is, they may be combined incorrectly. The binding together of features across preattentive maps is one of the primary functions of serial attention. Moreover, as

will be discussed in the following sections, a similar "binding problem" occurs at many levels of visual processing.

One highly speculative, but intriguing, neural mechanism for possibly solving the binding problem may involve the temporal responses of the neurons that serve as feature detectors throughout visual processing. While the concept of encoding stimulus properties through the firing pattern of populations of neurons is accepted in audition (i.e., to encode a sound with a frequency higher than the firing rate for individual neurons, a group of neurons may fire in a "volley" sequence that collectively indicates the frequency), it is relatively novel in vision. The essential idea is that neurons that encode individual feature properties for a given spatial location may be bound together by synchronizing their firing to a common rate. In this model there is no single neuron that represents the complex structure formed by a conjunction of features; instead, it is a population of neurons and their *temporal pattern* that represent this information. This hypothesis suggests at least one clue to the neural basis of serial attention: rather than simply being located in a particular brain structure, attention may involve mechanisms that enable representations of elements of an object or scene to be combined by the imposition of a common temporal code. As we shall see, this is a problem that must be solved not only for how features processed in parallel at each spatial position are combined, but also for the equally important issue of how features across *different* spatial locations are combined to reconstruct the more complex structures that are the basis of visual perception.

## II. PERCEPTUAL ORGANIZATION

Subsequent to the feature analysis performed by early visual processing, the layout of the scene is represented as an array of spatially discrete local features. While such representations implicitly encode the structure of surfaces and objects depicted within the array, it is the goal of perceptual organization to make these and other structures explicit. Two distinct kinds of processes may be available to accomplish this: first *grouping* processes may be used to identify commonalties among local features indicating that they should be combined into more complex structures; second, *segmentation* processes may be used identify differences among local features indicating that they should be separated into discrete structures. Both kinds of processes may be said to be *inferential* in that they operate without the benefit of complete information about the scene (for example, viewing a house partially obscured by trees), instead relying on inferences based on partial data and regularities inherent in the physical world.

Perceptual organization is central to vision, as well as other perceptual modalities, in that it is only by first organizing and marking structure within perceptual input that more complex representations may be created. Indeed, the mechanisms that carry out this task are some of the most sophisticated and impressive to be found in human cognition — a fact that is obscured by the many salient demonstrations of perceptual organization. However, while perceptual organization forms a cornerstone of visual processing, it is perhaps one of the most poorly understood mechanisms of perception. Conversely, perceptual organization is also a mechanism that has generated a wealth of phenomena. Therefore, one common starting point for its study has been to enumerate its specific principles of operation.

### A. Principles of Grouping

The most fundamental principles of perceptual organization were established by the Gestalt school of psychology in the early part of the century. All of these principles are based on the idea that complexity should be minimized, that is, mechanisms of perceptual organization will "prefer" simpler interpretations over complex interpretations. The operation of this principle may be easily illustrated by several demonstrations among the many available phenomena associated with principles of organization. For instance, upon inspecting Figure 3a, most observers will report perceiving three rows of spheres rather than four columns. This illustrates the principle of *proximity*, whereby, all else being equal, elements that are spatially nearer to each other are more likely to be grouped together to form a larger structure. Likewise, upon inspecting Figure 3b, most observers will report perceiving four columns composed of either spheres or cubes rather than four rows of alternating elements. This illustrates the principle of *similarity*, whereby, again all else being equal, elements that are similar in shape, color, or other visual properties are more likely to be grouped together. Finally, upon inspecting Figure 3c, most observers will report perceiving lines that extend from disc to disc forming a triangle. This illustrates the principle of *good continuation*, whereby collinear edges or boundaries are grouped together.

Figure 3c also demonstrates the efficacy of perceptual organization. In this display, not only are collinear boundaries grouped together, but there is an overall perception of an opaque triangle floating above three discs, each partially occluded by one point of the triangle. While edges of the triangle appear quite vivid, there is actually no physical basis for inferring a boundary between each of the discs (prove this by covering the discs). Such illusory edges are referred to as *subjective contours* and may function exactly like physically-based edges in many perceptual phenomena, including optical illusions, further perceptual organization, and object perception. Thus, not only does perceptual organization mark structure within visual

input, but it can serve to "fill in the gaps" in degraded or occluded objects.

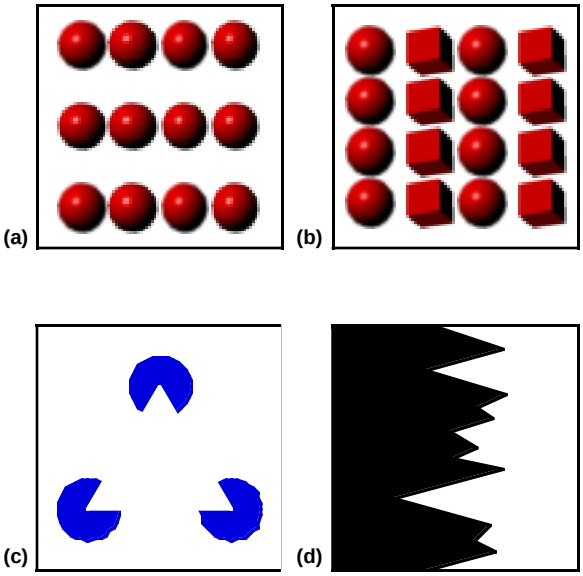


**Fig. 3.** Principles of perceptual organization. (a) Proximity: Spheres are grouped into rows based on the proximity between elements. (b) Similarity: Spheres and cubes are grouped into columns based on the similarity between elements. (c) Good Continuation: Collinear boundaries of triangular gaps in each disc are grouped to form continuous contours. Panel (c) also illustrates subjective contours; notice that a complete triangle is perceived, even though at some points there is no physical basis for inferring a boundary. (d) Figure-Ground Segregation: Depending on which side is considered the figure, a range of light or dark mountains may be perceived.

### B. Segmentation

While grouping processes are used for combining features, segmentation processes are used for segregating features and feature assemblies into discrete structures such as objects and parts. Segmentation often operates in conjunction with grouping, so for instance, two adjacent textures may be segregated into separate regions based on the similarity or proximity of elements within each texture. Another example of segmentation is illustrated in Figure 3d. This figure has two distinct interpretations: either a dark range of mountains with a plateau in the center or a light pair of mountains with several valleys in between (whatever the initial perception, the alternate may be easily perceived by attending to the half that is seen as sky). In each case, one half of the figure is considered to be most salient, called the *figure*, and the remaining half is considered to be background, or simply the *ground*. Notice that when one half is perceived as figure, the other half is always perceived as ground — only one interpretation is available at a time. This process of making one object or set of objects stand out from the rest of the scene, referred to as *figure-ground segregation*, is often considered a prerequisite for recognition mechanisms that match individual objects to stored object representations.

Another form of segmentation that may facilitate recognition is the segregation of objects into discrete parts. As will be discussed in the next section, part-based object representations offer a natural description of complex objects and categories. The decomposition of a given object into parts is consistent across different observers, indicating that a common mechanism is used to arrive at such a description. One possibility is that surfaces and objects are divided into parts by defining part boundaries at the points of greatest curvature within inward facing concavities along a surface or object contour. For instance, imagine a dumbbell shape; the part boundary will be located at the vertical midpoint of the dumbbell. This simple principle has a great deal of explanatory power for how parts are determined. In Figure 3d, it may be used to account for the boundaries between mountains: when the dark half is perceived as the figure, the divisions between the peaks are located in the lightly colored valleys; when the light half is perceived as the figure, the divisions shift and are located at the darkly color valleys — always corresponding to the points of greatest curvature along inward facing concavities.

### C. Non-Accidental Properties

By what general principles does perceptual organization operate? While it is often stated that "simpler" interpretations are preferred, it is difficult to arrive at a precise definition of what constitutes simple. One alternative is that preferred configurations of features are those that correspond to properties of meaningful physical structures, e.g., surfaces and objects. Consequently, perceptual organization may operate by exploiting regularities in input that are *non-accidental* in that they arise due to regularities that are likely to occur in the physical world. A non-accidental property may defined as a configuration of features for which the probability of occurring due to chance is essentially zero; or conversely, as a configuration for which the number of times it arises due to a coherent structure in the physical world is large compared to the times it arises through chance. Thus, it is to the advantage of the perceiver to consistently infer that particular feature configurations correspond to structure.

The principle of non-accidentalness may be used to account for the grouping phenomena discussed earlier. For example, the principle of proximity may be explained by the fact that if two features are adjacent in an image, they are likely to be adjacent in the three-dimensional world. Because objects are composed of coherent surfaces and not random points, adjacent features are likely to sit on the same surface, and,

consequently, should be grouped together. A similar explanation may be offered for the principle of similarity: because the textures of surfaces tend to be coherent, features that are similar are likely to have arisen from the same surface. Finally, good continuation may explained by the fact that surfaces tend to be smooth over large areas; therefore, collinear edges within an image are far more likely to correspond to a single boundary than to an accidental alignment of disconnected surface features. This explanation also accounts for the perception of subjective contours; in such instances, it is more likely that apparent gaps in objects and the alignment between gap boundaries arise because of occlusion by a single, coherent structure, than because of accidental collinearity between irregularly shaped forms.

Non-accidentalness has been used to enumerate several additional properties of perceptual organization. In all such cases, a preference for particular feature configuration is based on the likelihood that the two-dimensional configuration reflects the presence of a similar property in the three-dimensional, physical world. Many of these "non-accidental properties" are closely related to the principles of grouping: for instance, proximity of endpoints reflects a common termination point at a corner; collinearity and curvilinearity reflect features that are elements of a common edge or contour in three-dimensions; and, parallelism reflects edges that are parallel in three-dimensions and therefore likely to denote the boundaries of an object. Finally, the property of symmetry about an axis has been identified as reflecting structures that are symmetrical in three-dimensions. As will be discussed in the following section, non-accidental properties may form the basis for recovering the structure of parts within objects.

## III. OBJECT REPRESENTATION

### A. Properties of the Representation

Following the perceptual organization of visual input, the layout of the scene is represented as a collection of coherent surfaces and objects. Indeed, there are many situations in which this representation may be sufficient for successful task completion. For example, navigating a hallway may only require detecting the presence of obstacles; likewise, grasping an object may only require noting its size and rough shape. In contrast, many of our interactions with the environment necessitate not only a knowledge of the physical properties of the scene, but the recognition, including both identification and categorization, of objects within the scene. In order to achieve recognition, two fundamental problems must be addressed: first, object representations must be *recovered* from perceptually organized visual input; second, recovered object representations must be *matched* to like-format representations stored in memory. Given this model of object recognition, it is the properties of the representation that will determine the efficiency and robustness of both recovery and matching. Therefore, this section begins with an introduction to four dimensions along which specific formats for object representations may vary from each other. It should be emphasized that these dimensions are independent of one another, so that a format formed by any combination of values along each is theoretically possible. However, as discussed in the latter part of this section, in practice, particular settings tend to co-occur, forming two major families of object representations.

*1. Frames of Reference*. In order to encode object shape, as well as other spatial properties, object representations employ a coordinate system in which the spatial positions of features are defined. Such a coordinate system is commonly referred to as a *frame of reference*. Generally, two distinct types of reference frames have been proposed for object representations: *object-centered* and *viewer-centered* (a third type of reference frame, environment-centered, is rarely posited for object representations, but has been implicated in spatial representations that are used for navigation). Object-centered and viewer-centered reference frames may be differentiated by whether the coordinate system is based on the object to be represented, or, alternatively, is based on the viewer (encompassing possibly the viewer's retina, head, or body). This distinction is illustrated in Figure 4. In an object-centered frame of reference (top to lower-left panels), the coordinate system moves with a rotation of the object; therefore, the description of the object is not altered by changes in viewpoint. In contrast, in a viewer-centered frame of reference (top to lower-right panels), the coordinate system does not move with a rotation of the object, but instead, remains based on the position of the viewer; therefore, the description of the object is altered by changes in viewpoint.

Given that successful recognition depends on a match between a stored object representation and perceptual input, the reference frame used will have consequences for how recognition mechanisms compensate for discrepancies in orientation. The fundamental property of an object-centered frame is that it is orientation independent: so long as the frame used to represent perceptual input is the same as that used for the corresponding object representation, the descriptions will match regardless of viewpoint. By comparison, the fundamental property of a viewer-centered frame is that it is orientation dependent: because changes in the orientation of either the viewer or the object will alter the description of perceptual input, the perceiver must normalize the orientation of the input to match the viewpoint of the corresponding object representation. This normalization may be accomplished by a mental transformation process in

which the orientation of the perceptual input is modified by a transformation analogous to a physical rotation.

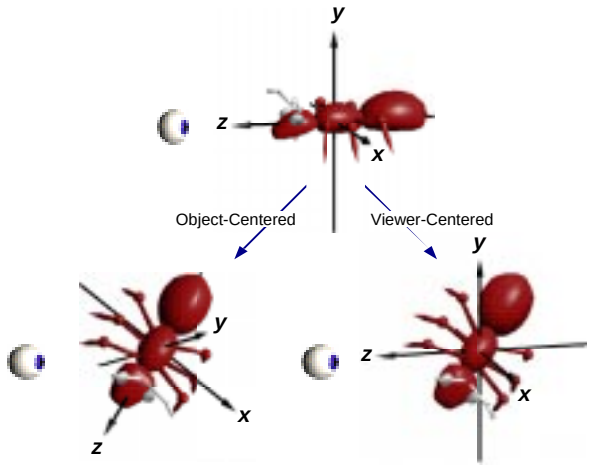


**Fig. 4.** Object-centered and viewer-centered frames of reference. In an object-centered frame (top to lower-left), the coordinate system is based on the object, moving with the ant regardless of the position of the viewer. In contrast, in a viewer-centered frame (top to lower-right), the coordinate system is based on the viewer, remaining with the viewer regardless of the position of the ant. Illustration by Scott Yu.

*2. Dimensionality*. A second property of an object representation is the number of *spatial dimensions* encoded, varying from two to three dimensions. While physical objects are three-dimensional, the pattern of light that impinges upon the retina is two-dimensional. Because it is widely held that early visual processes use a variety of cues to recover depth, it is likely that object representations include at least three-dimensional distances between surface points. This relative depth coding has been referred to as "two-and-one-half-dimensional." In contrast, a complete three-dimensional reconstruction of the object would include information about the structure of all surfaces of an object, including those not present in the current viewpoint. One consequence of this is that three-dimensional representations may only be formed by exposure to many different views of an object or by inferences about unobserved structure from a limited number of views.

*3. Parts*. One of the most salient properties of physical objects is that they are composed of *parts*. Likewise, object representations may employ one of several schemes to encode the part structure of objects. Generally, parts have been defined in two ways: parts may be represented by *volumetric primitives* that are three-dimensional and capture the approximate shape of a given part of an object (see Figure 5a); alternatively, as discussed in the previous section, parts may be delineated by a principle for locating the boundaries that specify where one part ends and another begins. These mechanisms are not mutually exclusive; the latter may be used to establish the locations of individual parts within an object, and the former may then be used to capture the shape of each part.

One of the most appealing aspects of using volumetric primitives to represent parts is that they facilitate the *coarse coding* of object shape. To understand why this is desirable, envision a model of object representation in which every variation in the shape of every part is encoded as a distinct entity. In such as scheme, even the most minute change in shape would result in a new part being represented, thereby producing an astronomical number of stored representations. Moreover, it would be difficult, if not impossible, to recognize two variants of the same object class as members of the same category (i.e., a kitchen chair and a desk chair). Even worse, it would be difficult to identify two variants of the *same object* — slight differences in shaping leading to completely separate representations. In contrast to this hypothetical model, a model employing a restricted number of volumetric primitives can represent many parts in an efficient manner. Not only does this reduce the amount of distinct information that must be represented, but it suggests that the same primitive may be used to approximate many variations in part shape. For instance, in Figure 5a, the primitive used to represent the fuselage of the plane may subsume many differently shaped fuselages. Therefore, many variations on this particular plane, as well as many somewhat differently shaped planes, may be categorized as the same type of object.

While volumetric primitives provide an elegant solution to the problem of representing parts, they also present two obstacles: First, primitives must be recovered from visual input with some consistency over changes in both part shape and position; second, it is unclear as to whether a restricted set of volumetric primitives is sufficient to capture all of the meaningful variations in part shape that may be encountered. These limitations may constrain the role of part-based representations in human object recognition.

*4. Spatial Relations*. Knowledge of parts alone is often inadequate for uniquely recognizing an object; for example, a cup and a bucket may be composed of the same two parts, a curved handle and an open-ended cylinder. In order to differentiate between objects sharing parts, the *spatial relations* between parts must be specified. One method of encoding spatial relations is to represent them implicitly. For instance, by preserving the spatial configuration of visual input within a two- or three-dimensional image array, the spatial positions of each of the parts are preserved. Because the relative positions of the parts are known, spatial relationships between them may be computed, although only by inspecting the image to determine a desired relation. Alternatively, spatial relations may be represented explicitly; that is, there may be a particular

value assigned to the relationship between a part and a reference point. Explicit spatial relations may be highly specific, for instance, by using numerical measures of distance and angle, or categorical, for instance by using relations that correspond to spatial terms used in language (i.e., "above," "below," and "on"). As with volumetric primitives, using a restricted set of more general spatial relations offers advantages in terms of coarse coding and reducing the overall information load.

The choice of a reference point in explicit spatial relations suggests an additional property of object representations: whether spatial relations are encoded hierarchically or with regard to a single reference point for the entire object. For instance, [torso]->[arm]->[upperarm,lowerarm,hand]->[palm]->[fingers] is an example of a part hierarchy, while positioning all of these parts relative to the center of the torso is an example of a single-point representation. While descriptions using hierarchical spatial relations are likely to be somewhat more complex than single point arrangements, they offer certain advantages: first, at the highest level of the hierarchy the description of an object may be quite simple, including only a small number of diagnostic parts; second, hierarchical representations capture the fact that objects frequently contain parts of many different sizes; and, third, hierarchical representations facilitate the encoding of articulated parts in which spatial relations between two adjacent parts vary over a restricted range. This latter point is crucial for representing many natural objects; for example, the spatial relationship between our upper and lower arms is not fixed, but rather should be represented as a range of acceptable configurations.

**B. Families of Object Representation**

Combining variations of these four properties produces a wide range of possible formats for object representations. However, current research indicates that properties of object representations apparently cluster into two families: structural descriptions and view-based representations.

Object representations based on *structural descriptions* generally share the following defining properties: they are object-centered, composed of volumetric primitives, encode complete three-dimensional information, and use explicit categorical spatial relations between parts. Additionally, they often contain a hierarchical description in which smaller parts are located relative to adjacent relatively larger parts. Figure 5a illustrates a structural description for an airplane (although the representation can only be displayed two-dimensionally from one orientation and without explicit spatial relations). Two attributes are characteristic of structural-descriptions. First, because they encode information in an object-centered frame of reference, visual tasks relying on structural descriptions are unaffected by changes in orientation, size, and position, as well as mirror-reflection of the object. Object identification tasks in which performance is invariant over variations in viewpoint are typically interpreted as evidence for structural descriptions. Second, because structural descriptions use volumetric primitives that produce a coarse coding of object shape, they are suitable for determining the basic categories of objects (e.g., "car" or "person"), but not specific individuals (e.g., "'65 Mustang" or "Bob").

One variant on structural description models uses a highly restricted class of volumetric primitives to represent parts. The complete set of primitives is generated by combining attributes of the cross section and the axis of three-dimensional volumes that reflect any of several different non-accidental properties. Each combination forms a unique three-dimensional volume, such as a brick or a cone. For example, a brick would be defined by edges terminating at common points at each of the corners and parallel edges along each of the faces. This approach offers a possible solution to the problem of consistently recovering volumetric primitives from visual input: because the primitives are based on combinations of non-accidental properties, once a given configuration of properties is identified, the corresponding volumetric primitive may be inferred. Furthermore, this scheme has the advantage of basing part representations on features that are known to exist within visual input; therefore, the structure of these primitives is based on regularities of physical objects, thereby possibly having greater generality for adequately capturing variations in part shape.

One alternative to structural descriptions is the *view-based* family of object representations. View-based representations generally share the following defining properties: they are viewer-centered, use a boundary locating principle to define parts, encode two-dimensional information plus depth, and implicitly encode spatial relations. View-based representations depict the appearance of an object from a single viewpoint (sometimes called a "view"). Figure 5b illustrates several view-based representations for an airplane; each image differs from the others in viewpoint, representing different surfaces, parts, and object shape. Two attributes are characteristic of view-based representations. First, because they encode information in a viewer-centered frame of reference, visual tasks relying on view-based representations are affected by changes in the orientation, as well as possibly size and position, of objects. Identification tasks in which performance is systematically related to viewpoint are typically interpreted as evidence for view-based representations. Second, because view-based representations preserve the original viewpoint and shape of parts within objects, they are suitable for making fine discriminations between objects from within a category (e.g., differentiating between two faces and or two models of cars).

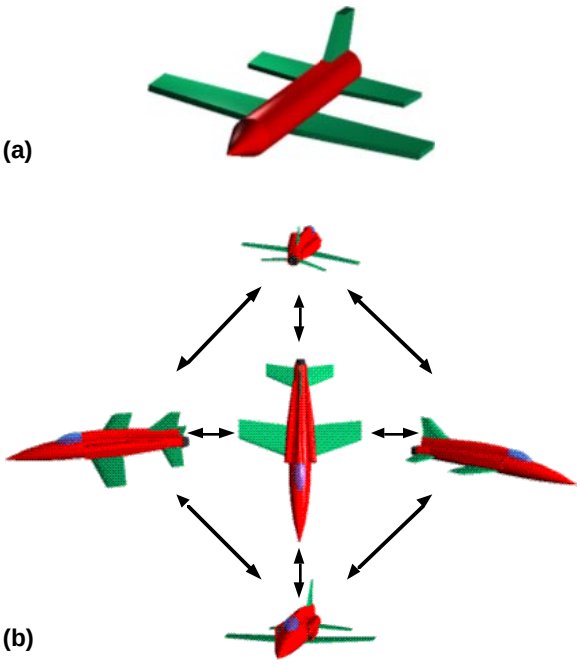


**Fig. 5.** Families of object representation. (a) A structural description of an airplane composed of volumetric primitives. (b) A view-based representation of an airplane composed of multiple viewer-centered images. Illustration by Scott Yu.

One variant on view-based representations uses *multiple-views* to represent the three-dimensional structure of objects. While a view-based representation depicts an object from a single viewpoint, a more comprehensive representation may be formed by linking a set of non-overlapping, or characteristic, views. For instance, in Figure 5b, five characteristic views of the airplane cover many of the part and surface configurations that are likely to be observed. A multiple-views model also introduces several associated mechanisms. First, in order to achieve a match between a view of an object and the same object observed in a non-matching orientation, a mental transformation is used to *align* the object with the stored view. Mental rotation is a well-documented visual process that produces systematically longer response times for greater angles of rotation, a pattern consistent with the systematic effects of orientation found in identification tasks. Second, mechanisms are required for determining which viewpoints of an object should be instantiated as views. Research suggests that familiarity with a given viewpoint and the geometry of an object both play a role in this process. Therefore, the greater the number of times an object appears at a given viewpoint and the more that viewpoint does not overlap with stored views, the more likely it is that a view will be stored for that orientation. Indeed, the most familiar and visually salient viewpoint of an object, referred to as *canonical*, is the orientation from which an object is typically imagined and most easily identified.

Converging evidence from studies of object recognition, human memory, and human neuropsychology indicates that both structural descriptions and view-based representations are used in visual cognition. One robust finding has been systematic effects of object orientation on recognition times. The use of view-based representations is supported by naming tasks in which response times increase with distance from a familiar view of an object, rather than just the upright orientation. In contrast, the use of structural descriptions is supported by somewhat different recognition tasks in which responses times are unaffected by object orientation, for instance, categorizing objects. In human memory, conscious, or "explicit," memory for previous exposure to an object is impaired by changes in orientation, size, or mirror-reflection, indicating that explicit memory is mediated by representations that are view-based. However, unconscious, or "implicit," memory for previous exposure to an object is unimpaired by changes in size or mirror-reflection, indicating that implicit memory is mediated by structural descriptions. Finally, in human neuropsychology, there is a dissociation between the recognition of words and the recognition of visually similar objects, such as faces. This finding has been interpreted as evidence for two types of representations, one in which there is an extensive decomposition of parts into a structural description, and one in which there is little part decomposition, but detailed information about the appearance of the object. Together these results point to two systems for object representation: a structural description system most suited to categorizing objects according to their parts, and, a view-based system most suited to recognizing objects from a specific viewpoint.

**BIBLIOGRAPHY**

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*, 115-147.

Cooper, L. A., & Schacter, D. L. (1992). Dissociations between structural and episodic representations of visual objects. *Current Directions in Psychological Science*, *1*(5), 141-146.

Farah, M. J. (1992). Is an object an object an object? Cognitive and neuropsychological investigations of domain-specificity in visual object recognition. *Current Directions in Psychological Science*, *1*(5), 164-169.

Osherson, D. N., Kosslyn, S. M., & Hollerbach, J. M. (Eds.). (1990). *Visual Cognition and Action*. Cambridge, MA: The MIT Press.

Pinker, S. (Ed.). (1985). *Visual Cognition*. Cambridge, MA: The MIT Press.

Plaut, D. C., & Farah, M. J. (1990). Visual object representation: Interpreting neurophysiological data within a computational framework. *Journal of Cognitive Neuroscience*, *2*(4), 320-343.

Tarr, M. J., & Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science*, *1*(42), 253-256.

Treisman, A. (1990). Features and objects in visual processing. In I. Rock (Ed.), *The Perceptual World* (pp. 97-110). New York, NY: W. H. Freeman and Company.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 419-433.