In: E. B. Goldstein (Ed.), *Encyclopedia of Perception Volume 2* (pp. 643-648). Thousand Oaks, CA: SAGE Publications, Inc.

Object Perception

The environment is full of elementary features such as shapes, colors, and textures. However, observers do not perceive these elements in isolation. Rather, they combine them into two-dimensional objects such as red circles or into three-dimensional objects such as cats and people. Broadly defined, object perception is the ability to combine elementary features into whole objects. The ability to perceive objects is an important precursor to recognize three-dimensional objects in the environment. Recognition can occur at different levels of specificity: Observers can identify an object as one they have seen before (e.g., their neighbour's pet cat), or they can categorize that object as belonging to a more general class (e.g., as an animal rather than a vehicle). This recognition process occurs very quickly and accurately, which allows observers to successfully interact with objects in their environment.

The basic problem of object perception is the changing nature of the visual input. Light is reflected from the surface of objects and picked up by millions of photoreceptors on the retinal layer at the back of each eye. The retina thus creates a two-dimensional image of the three-dimensional world. This retinal image is the input to the visual system. The input changes because observers and objects move relative to each other or lighting conditions vary. Features are extracted from this changing retinal image and subsequently used to perceive and recognize objects. Psychologists, neuroscientists, and computer scientists have all tackled this problem in perception. Consequently, different approaches have emerged to study object perception, which have ultimately lead to different theories of object recognition.

Structuralism and Gestalt Psychology

One of the earliest ideas about perception arose in the laboratory of Wilhelm Wundt, who established the first experimental psychology laboratory in 1879. Wundt believed that there were sensations evoked by physical stimulation of the corresponding sense, such as the sensation of heat, sound tone, or light intensity. He proposed that perception emerged from the combination of these pure sensations. He and his students trained themselves to systematically isolate and measure the quality and intensity of these sensations through a technique of introspection or self-observation. This approach was later called structuralism by Wundt’s student, Edward Titchener.

However, Max Wertheimer, Wolfgang Köhler, and Kurt Koffka found percepts that had no corresponding pure sensation, contrary to the tenets of structuralism. For example, Wertheimer found that briefly flashing two bars of light a short distance apart, one after the other, gave rise to the percept of a single bar moving from one location to another (which he called apparent motion or phi phenomenon). In apparent motion, there is no stimulus between the two bars to physically stimulate the retina and evoke a pure sensation of light. Consequently, Wertheimer and his colleagues founded an approach called Gestalt psychology that focused on characterizing percepts rather than sensations, and formulating the laws that govern their creation. These laws became the Gestalt principles of perceptual organization. One of the central laws is pragnanz, roughly meaning “good figure”, which states that a stimulus is perceived as simply as possible. Consider the stimulus illustrated on the top of Figure 1a. This figure is ambiguous because it can have more than one perceptual interpretation. It can be interpreted as two interlocking rings. A more complex, but equally valid, percept is of three shapes, shown in gray, abutting each other. Observers mostly report the first percept which is simpler. Other laws include proximity: Elements that are close together are grouped together as a percept; and similarity: Elements that are similar to each other (e.g., they have the same color) are grouped together. These laws help group elements which might otherwise be ambiguous into whole objects.

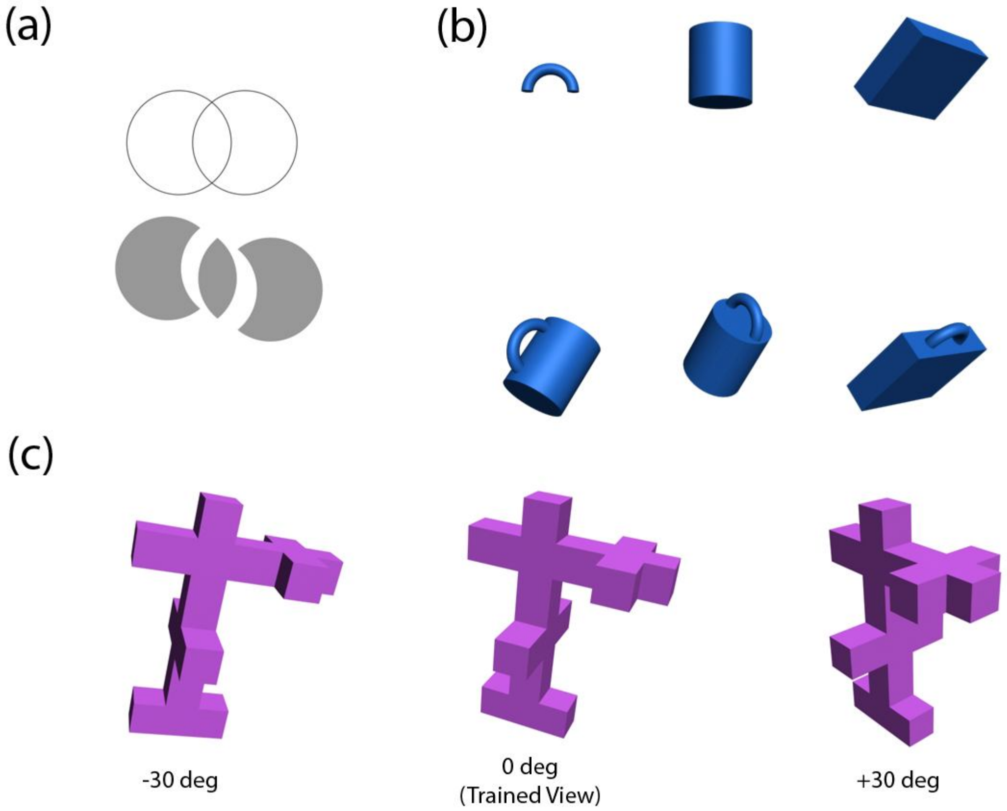

Figure 1. (a) An ambiguous figure to illustrate the law of pragnanz. (b) Three of the 36 geons proposed by Biederman. The combination of these geons and their spatial relationship form different objects. (c) An example of Tarr and Pinker's novel object seen from different viewpoints.

Modern Approaches to Object Perception

In his influential book, David Marr popularized a computational approach to object perception and recognition. For Marr, the ability to perceive and recognize objects is essentially a problem in information processing. He conceptualized a series of stages. At each stage, information is transformed into a different representation and transmitted to the next stage. His main stages included a primal sketch, a two-and-a-half dimensional (2.5D) sketch, and finally a 3D model. The primal sketch transforms light intensities on the retinas into edges and regions. The 2.5D sketch then groups these edges and regions into visible surfaces. This representation encodes information about color, texture, and even distance to the observer, at each point on the surface. The representation at this stage is 2.5D because of the additional depth information that is encoded. Finally, a full 3D mental model of an object is reconstructed from the 2.5D sketch. This model specifies object parts relative to each other (e.g., the hand is connected to the wrist; the wrist is connected to the arm; the arm is connected to the shoulder; and so on). This spatial coordinate system is called an object-centered reference frame. Marr's computational approach helped link the Gestalt psychologists' various principles of perceptual organization to other visual functions, such as grouping parts into 3D models. He also popularized the idea that object perception is about reconstructing the 3D scene.

Another modern approach that built on the early ideas of object perception is Anne Treisman's feature integration theory. She suggested that attention is needed to group elements into percepts, and proposed that this grouping process occurred in two stages. First, there is a pre-attentive stage in which elementary features, such as color and line orientation at different positions on the retinas, are automatically analyzed. Then in a focused attention stage, these features are combined into objects. This combination process requires attention to integrate features belonging to an object. Without attention, the visual system sometimes incorrectly combines features from different objects leading to illusory conjunctions. For example, if a red "X" is flashed with a blue "K", participants may report seeing a red "K" if they are not attentive.

Theories of Object Recognition

The stable percept of an object is an important step towards recognizing that object but one challenge is the changing visual input. Different theories of object recognition have therefore been proposed to explain how observers solve what is known as the invariance problem. This problem states that observers perceive the same object, and can identify or categorize it, despite possible differences in the retinal images of that object. For example, when observers see an object from two different viewpoints, the 2D image created on the retina is different between those two views. Similarly, different lighting conditions create different retinal images of the same object. Researchers agree that the underlying representation that supports this kind of robust recognition is critical; however, the nature of this representation remains a key issue in object recognition. Thus, theories of object recognition generally fall into two broad classes which differ on the features used for the object representation.

Structural Description Theories

Proponents of structural description theories propose that objects are represented by parts and their spatial relationships, which together form a structural description of an object. These descriptions discard an object’s color and texture, for example, as the appearance of surface properties change with changes in viewing conditions (e.g., a change in lighting can change how color appears to an observer). The basic idea is that the same structural description can be recovered or otherwise derived from different retinal images of the same object. This robustness remains an appealing aspect of structural description theories despite the loss of surface information. Structural description theories have also been referred to as part-based or edge-based theories, given their reliance on parts and edges.

The first viable structural description theory for human object perception was proposed by David Marr and Keith Nishihara. According to their theory, object parts (e.g., a cat’s leg) are represented by 3D primitives called generalized cones, which specified arbitrary 3D shapes with a set of parameters. For example, a cylinder can be produced by taking a circular cross-section and sweeping it along a straight line. The circle traces out a cylinder with the line forming the main axis of that cylinder. By comparison, a rectangular cross-section sweeps out the surface of a brick. More complex 3D shapes can similarly be produced by sweeping different 2D cross-sections across different axes.

One of the challenges faced by Marr and Nishihara was how 3D generalized cones can be recovered from 2D images. They suggested that an object's bounding contour—the outline of an object in a picture—could be used to find the axes of its main parts. These axes could then be used to derive generalized cones and their spatial configuration. Recognition could then proceed by matching the structural description recovered from the image to those stored in visual memory. Thus, Marr and Nishihara try to solve the invariance problem by recovering view-invariant 3D models from images.

Following Marr and Nishihara's seminal 1978 work, Irving Biederman proposed another influential structural description theory in the mid 1980s —Recognition by Components (RBC). Biederman argues that objects are mentally represented by a set of 36 components and their spatial relationship. He called these geons, for "geometrical ions". Geons are a subset of the generalized cones proposed by Marr and Nishihara, three of which are shown on the top of Figure 1b. The combination of these geons into structural descriptions can be used to create familiar objects like a mug, a pail, or a briefcase, as shown in the bottom of Figure 1b.

RBC theory builds on Marr and Nishihara's structural description theory in two innovative ways. First, unlike generalized cones, geons only differed qualitatively from each other. For example, a geon's axis can only be straight or curved whereas generalized cones can, in principle, have any degree of curvature. Biederman's second innovation was to propose a more direct means to recover geons from images. According to RBC theory, geons are recovered from non-accidental properties. These are properties of edges in an image (e.g., lines) that are associated with properties of edges in the world. To understand non-accidental properties, consider seeing a box from many different viewpoints. From most views, observers see three sides of the box which terminates in a "Y"-junction at a corner. This two-dimensional junction is an example of a non-accidental property, and it is associated with a three-dimensional corner. From a few viewpoints, the Y-junction is not visible in the image such as when observers see one side of the box. It is therefore ambiguous whether such an image is a three-dimensional box or a two-dimensional rectangle. However, these "accidental" viewpoints are much less likely to be encountered relative to non-accidental viewpoints. Thus, non-accidental properties are highly robust (though not entirely invariant) to changes in viewpoint, viewing distance, and illumination.

Biederman suggested that two to three geons are enough to represent many objects. He also stated that RBC theory accounts for recognizing objects at what psychologists call the basic level. For example, "dog" is at the basic level; "poodle" is at a more specific subordinate level; and "animal" is at a more general superordinate level. To recognize objects at the subordinate level requires qualitatively different representations than the structural descriptions proposed by RBC theory. Biederman, for example, suggested that face recognition is based on the representation of fine metric details, such as the distance between eyes, rather than geons.

To solve the invariance problem, Biederman proposed a principle of componential recovery: Objects can be identified if their component geons can be identified. As geons are recovered from highly stable non-accidental properties, the same geon can be recovered from many different viewpoints, viewing distances, and illumination. Perhaps the strongest evidence for RBC theory comes from Biederman's contour deletion studies. He and his colleagues took line drawings of everyday objects like a cup. They then deleted contours that could be used to recover geons (e.g., at junctions) or deleted the same amount of contours from other sections that could not be used to recover geons. They found that observers had no problem naming line drawings with preserved junctions but had difficulty naming line drawings with deleted junctions.

Image-based Theories

In contrast to structural description theories, proponents of image-based theories posit that objects are represented as measurements of features that preserve many aspects of an object as they appears to an observer. For example, the visual system measures features such as color, texture, and even shading patterns on a surface. It can also encode the spatial location of these features. These measurements constitute a view of an object, which depict its appearance under specific viewing conditions much like a picture taken by a camera from a fixed viewpoint. If a viewing condition changes (e.g., the observer walks to a different location to view an object), these measurements also change. In effect, object perception is supported by representations that are copies of the retinal images. For example, whereas structural descriptions discard surface features (e.g., color), an image-based representation would encode all visible features. Image-based theories have also been called view-based and appearance-based theories.

The image-based approach emerged in the early 1990s as a result of two key developments. First, there was an accumulation of evidence which indicated that observers were highly sensitive to viewing conditions, contrary to the predictions of structural description theories. For example, Michael Tarr and Steven Pinker trained observers to recognize block-like objects, which they had never seen before, from a specific viewpoint. Tarr and Pinker then rotated these objects away from their familiar trained view and tested observers' ability to recognize them. The effect of depth rotation on the appearance of these novel objects is illustrated in Figure 1c. Tarr and Pinker found that observers responded more slowly and made more mistakes to novel views relative to trained views. Second, Tomaso Poggio, Shimon Edelman, and Heinrich Bülthoff developed computational models which learned to match novel images of an object to trained images of that object. Importantly, these models do not reconstruct an object's 3D structure. Rather, they store a collection of views, and match a novel image to a candidate stored image based on the similarity of the two images. In addition to providing a theoretically-rich framework for understanding object recognition, the behaviours of these models were similar to the behaviours of humans.

Tarr and Pinker synthesized these developments to address the invariance problem in their multiple-views theory of object recognition. They posited the idea that a 3D object can be represented as a linked collection of experienced 2D views. To explain how novel views are recognized, they further postulated a time-consuming mental rotation process, analogous to physically rotating an object, to transform a novel view until it matched a stored view. Observers could mentally rotate the input image until it matched a stored view. Accordingly, the time to perform this operation and how accurately it was performed should be proportional to the difference in viewpoint between the input view and stored view. Thus, Tarr and Pinker predicted that performance should depend on the orientation difference between the novel view and the nearest view encoded in memory. This is what they found in their experiments. Poggio and his colleagues proposed another normalization mechanism called view interpolation. In this model, stored views are used to capture possible appearances of an object. The stored views allowed the model to interpolate, or estimate, novel views of an object for recognition purposes. The idea is analogous to fitting a smooth curve through known data points, and then predicting unknown data from this curve.

Finally, according to image-based theories, parts also play a role in object perception. However, unlike structural description theories, these parts are fragments extracted from specific images of objects from a particular class (e.g., different cats) rather than being pre-specified (e.g., as a small set of geons). An image has a very large number of fragments which can be of any size. However, only a very small proportion of the possible fragments are useful for object recognition. Shimon Ullman uses a statistic called mutual information to find these informative fragments. This measure gives the probability that an object is present in an image if a fragment is present in the image. For example, if a human eye is present in the image, then there is a good chance that a human face is also present in that image. In this way, mutual information maximizes the ability for a set of fragments to distinguish between different objects, e.g., cats from dogs and horses. The set of informative fragments for a class is related to regularities that exist across members of that class. For example, cats are more likely than horses to have pointy ears and whiskers. There is both behavioural and neural evidence that human observers use fragments in object perception.

Two versus Three Dimensions

Structural description models are often conceptualized as 3D models because the primitives can be explicitly three dimensional, as in Marr and Nishihara’s generalized cones. Likewise, image-based models are often conceptualized as 2D models because of its analogy to a camera. However, neither class of theories need to be strictly two or three dimensional. For instance, Biederman argued that geons capture 2D non-accidental properties which do not drastically change with viewing conditions rather than 3D structure per se. Similarly, Tarr suggested that information about depth at points on a visible surface can be a feature that is measured by the visual system and used for object perception. Shading cues, for example, can provide this additional depth information. However, depth information can only be recovered from visible surfaces so the representation is neither strictly two dimensional nor strictly three dimensional—akin to Marr’s 2.5D sketch. This issue of dimensionality points toward a common ground for structural description and image-based theories of object recognition.

Quoc C. Vuong

See also Bayesian Approach to Perception; Face Perception; Inference in Perception; Vision: Cognitive Influences; Visual Scene Statistics

Further Readings

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

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco, CA: W H Freeman.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B, 200,* 269-294.

Peissig, J. J., & Tarr, M. J. (2007). Visual object recognition: Do we know more now than we did 20 years ago? *Annual Review of Psychology, 58,* 75-96.

Peterson, M. A. Object perception. (2001). In E. B. Goldstein (Ed.), *Blackwell Handbook of Perception,* Chapter 6, pp. 168-203. Oxford: Blackwell Publishers.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature, 343*, 263-266.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21*, 233-282.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*, 97-136.

Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Science, 11*, 58-64.