Image recognition: Visual grouping, recognition, and learning

Joachim M. Buhmann*†, Jitendra Malik‡, and Pietro Perona§

*Institut für Informatik, Universität Bonn, 53117 Bonn, Germany; ‡Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720; and §Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125

Vision extracts useful information from images. Reconstructing the three-dimensional structure of our environment and recognizing the objects that populate it are among the most important functions of our visual system. Computer vision researchers study the computational principles of vision and aim at designing algorithms that reproduce these functions. Vision is difficult: the same scene may give rise to very different images depending on illumination and viewpoint. Typically, an astronomical number of hypotheses exist that in principle have to be analyzed to infer a correct scene description. Moreover, image information might be extracted at different levels of spatial and logical resolution dependent on the image processing task. Knowledge of the world allows the visual system to limit the amount of ambiguity and to greatly simplify visual computations. We discuss how simple properties of the world are captured by the Gestalt rules of grouping, how the visual system may learn and organize models of objects for recognition, and how one may control the complexity of the description that the visual system computes.

Image Recognition

Our eyes provide us with an abundance of information about the outside world. Thanks to vision we become aware of the objects and living beings that surround us and represent their form and properties in our brains. Computer vision researchers aim at reproducing this capability in machines.

Vision is difficult. The images of a human head and a melon are very similar if taken with the same illumination, whereas two images of the same head taken under different lighting conditions are extremely different. Yet, we have no problem in telling which is which. The image of a tree is composed of an intricate pattern of lights and darks, greens, yellows, and browns and yet we are able to perceive it as a single object and simultaneously to perceive the leaves and branches that compose it. It is obvious from these examples that the metric in the world of images, i.e., a naive distance measure in the extremely high-dimensional space of image intensities, is not very informative for extracting concepts from images. Different objects may produce the same image and, vice versa, the same object may give rise to very different images depending on viewpoint and lighting conditions.

Knowledge of the world is vital in resolving these difficulties and taking advantage of whatever little information an image can provide. Many visual illusions demonstrate that the visual system is built to take educated guesses on the nature of stimuli.

Grouping

Humans have a remarkable ability to organize their perceptual input; instead of a collection of values associated with individual photoreceptors, we perceive a number of visual groups, usually associated with objects or well-defined parts of objects. This ability is equally important for machine vision.

Perceptual grouping was first studied in the context of human vision by researchers in the Gestalt school of psychology in the early part of the 20th century. They pointed out several factors that could be used to group together parts of an image that most likely arise from a single object in the scene. Similarity of color or texture are very powerful: humans readily form groups from parts of an image that are uniform in color, such as a connected red patch, or uniform in texture, such as a plaid region. Contour fragments in an image are linked together if they exhibit "good continuation," i.e., can be linked to form a smoothly curving extended contour. These are sound probabilistic inferences in a world where objects tend to have parts with coherent color and texture and are bounded by smooth contours. A survival advantage would accrue to those animals that had incorporated such factors in their visual processing.

Grouping also may be based on high-level, abstract relationships. Regions in the two-dimensional image that are symmetric about an axis can be grouped as projections of three-dimensional objects, such as a vase. An arrangement of four lines in proper positions can be interpreted as a nose, eyes, mouth configuration and grouped as a face.

A number of competing formalisms, such as Markov Random Fields (1), layer approaches (2), and cut techniques drawn from spectral graph theory (3) are being explored as models for grouping. Shi and Malik (3) formulate visual grouping as a graph partitioning problem. The nodes of the graph are the entities that we want to partition; for example, in image segmentation, they will be the pixels; in video segmentation, they will be space-time triplets. The edges between two nodes correspond to the strength with which these two nodes belong to one group, again in image segmentation, the edges of the graph will correspond to how much two pixels agree in luminance, color, texture, etc.; whereas in motion segmentation, the edges describe the similarity of the motion. Intuitively, the criterion for partitioning the graph will be to minimize the sum of weights of connections across the groups and maximize the sum of weights of connections within the groups.

Recognition of Object Classes

Thanks to vision we can recognize reliably people, animals, and inanimate objects from a safe distance. Recognition can happen at multiple levels of abstraction: Fido, a poodle, a friendly dog, a medium-sized mammal, an animal. Recognizing an object requires associating an image with a memory of that object called a model. Models are usually not innate: typically we do not recognize things that we have not seen before (human faces and snakes are rare exceptions), therefore we must construct models from our daily visual experience. We are, however, good at generalization; we will recognize a person as such even if we have never met that specific person before.

What is the structure of object models? Because our visual experience consists of images, the simplest model of an object

This paper is a summary of a session presented at the fifth annual German-American Frontiers of Science symposium, held June 10–13, 1999, at the Alexander von Humboldt Foundation in Potsdam, Germany.

†To whom reprint requests should be addressed. E-mail: jb@informatik.uni-bonn.de.

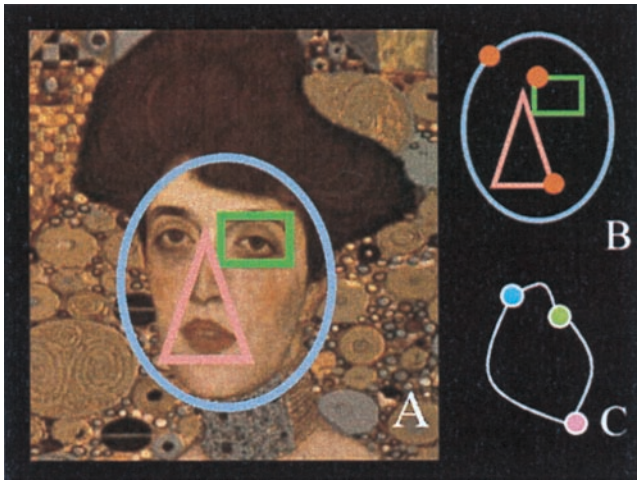

Fig. 1. Object model consisting of parts and their relationship.

consists of a collection of images of that object. In this case matching an image to a model is simple: measure the minimum difference (or distance) between the current image of the object and each of the model images. This corresponds to calculating the distance between one point (the test image) and a cloud of points (the model) in a high-dimensional space (as many dimensions as pixels in the image). Unfortunately this method is impractical: a reasonable sample of all the possible views of an object (six degrees of freedom encoding the point of view) under all possible lighting conditions (infinite degrees of freedom) would involve so many images that both storing the model and searching for the right match would be hopeless. Fortunately, significant redundancy in such models allows us to reduce both storage and computational requirements. One may think of the model as a thin low-dimensional surface embedded in a high-dimensional space; the surface is parameterized by the point of view and the lighting conditions. The challenge is representing this surface efficiently; the aim is to calculate the distance in the low-dimensional local coordinates of the surface rather than the high-dimensional image space. A combination of warping and principal component analysis (4, 5) works well for verifying the identity of human faces under controlled lighting conditions when the position of the head is known. These methods are not adequate for handling large deformations and occlusion of part of the object or when the object's location in the image is unknown.

More sophisticated techniques are required to represent models of object classes where there is a high degree of variability between one object and another belonging to the same class. Representing independently elements of the appearance of local image features and their mutual position in the image allows one to write probabilistic models of objects belonging to a class. Models of this kind (6) recently have been demonstrated to recognize data as different as handwriting and human faces (see Fig. 1). Conditioning on the position of a putative feature allows one to generate hypotheses on where the rest of an object is located in the image, which allows searching for an object efficiently in a manner that resembles attentional searching in biological visual systems.

Learning and Vision

When we recover the geometry of the world, or recognize objects, we are fitting models to the data provided by our eyes. These models are formed from our experience in two ways. In supervised learning, a teacher specifies class labels "this is a tiger," for example, and images. In unsupervised learning, which is the norm in biological systems, the process has to be driven by an attempt to find good internal representations given the statistics of natural images (7).

Finding the best model is a compromise between fitting the data and minimizing the complexity of the model. In science the preference for simple theories over complex ones is known as Occam's razor and often is treated as a matter of aesthetics. However, for a visual organism that is constantly engaged in model construction, constructing better theories, i.e., ones with greater predictive power, is a matter of life and death!

A mathematical justification for preferring simple models or theories in accordance with available data arises from statistical learning theory (8). Flexible models with many degrees of freedom adapt to stochastic fluctuations in the data, whereas overly simple models cannot represent essential aspects of the signal. Vapnik and Chervonenkis (8) estimate how much the expected performance/risk of a selected solution, i.e., the best solution on the available data, deviates from the optimal solution in the model class. This optimal solution with minimal expected risk most often does not minimize the costs on the available data. Uniform convergence of empirical risk to expected risk is a necessary and sufficient mathematical condition for learning.

Image analysis is inherently multiscale; segmentation, grouping, or classification have to be performed at the appropriate scales of resolution. The foliage of trees in a photograph of a forest might appear homogeneous at low resolution but a closer look at high resolution reveals differences in leaf shapes that generate tree-specific foliage textures. Statistical learning theory relates the complexity of models to the amount of available data, i.e., the appropriate image scale. In image segmentation these scales denote the spatial resolution of segments, the fuzziness of segment boundaries, and the number of segments, respectively. These scales have to be coupled by an underlying inference principle. The well-known stochastic optimization algorithm simulated annealing or its deterministic variants provide a computational temperature as a control parameter to couple these scales (9). These algorithms can be tuned for real-time applications and for just-in-time processing with limited resources as they occur in robotics, vision-based surveillance, and inspection.

Challenges in Vision

The enormous growth of computational power in the last decade has propelled computer vision to a stage that renders complex tasks feasible at an affordable cost. Numerous products have emerged that are based on computer vision. Examples are content-based search of images and videos in databases, intelligent surveillance systems, vision-guided autonomous vehicles, fingerprint/face/iris recognition.

However, success on general image recognition is still faraway. This problem covers questions from low-level image processing up to high-level semantic image interpretation and object classification. Pixels have to be grouped into segments, segments are composed to form object models with geometric or appearance information, and finally, these models have to be classified into appropriate object categories. The grand challenge of vision, particularly image recognition, lies in constructing a unified framework for modeling image content with appropriate semantic abstraction levels.

1. Geman, S. & Geman, D. (1984) *IEEE Trans. Pattern Anal. Machine Intelligence* **6,** 721–741.
2. Wang, J. Y. A. & Adelson, E. H. (1994) *IEEE Trans. Image Proc.* **3,** 625–638.
3. Shi, J. & Malik, J. (1997) in *Proceedings of the IEEE Conference: Computer Vision and Pattern Recognition '97* (IEEE Press, Los Alamitos, CA), pp. 731–737.
4. Sirovich, L. & Kirby, M. (1987) *J. Opt. Soc. Am. A* **4,** 519–524.
5. Vetter, T. & Poggio, T. (1997) *IEEE Trans. Pattern Anal. Machine Intelligence* **19,** 733–742.
6. Burl, M., Weber, M. & Perona, P. (1998) in *Proceedings of the 5th European Conference on Computer Vision '98*, eds. Burkhardt, H. & Neumann, B. (Springer, New York), LNCS 1407, Vol. II, pp. 628–641.
7. Zhu, S. C. & Mumford, D. (1997) *IEEE Tr. Pattern Anal. Machine Intelligence* **19,** 1236–1250.
8. Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory* (Springer, New York).
9. Hofmann, T., Puzicha, J. & Buhmann, J. (1998) *IEEE Trans. Pattern Anal. Machine Intelligence* **20,** 803–818.