1 Early Visual Learning

Shree K. Nayar
Columbia University

Tomaso Poggio
Massachusetts Institute of Technology

1. Intelligence and Learning

Cast as the balance between internal representation and external regularities, learning has long been a central issue in the understanding of intelligence. It may be argued however that until recently its indispensable role had not been entirely embraced in computer science. Indeed, the definition of intelligence that is frequently invoked in the field of artificial intelligence does not quite provide learning its deserved stature. For many years, the Turing test has served as an operational definition of intelligence, using which, most investigators in artificial intelligence have either directly or implicitly measured their goals and achievements: If a computer behaves in a way indistinguishable from a human, then it can be called intelligent, even if this "intelligence" has been painstakingly programmed by a highly skilled individual.

Consistent with Turing's implicit definition, intelligence was perceived thirty years ago as mainly reasoning, problem solving, theorem proving, or game playing (chess, for instance). In contrast, today, we understand intelligent lower animals better and comprehend the complexities of the tasks that our senses routinely solve. We have also realized through experience how intractable the problem of producing software can prove, and how much of it would be needed to replicate even the simplest aspects of intelligence. In light of these submissions, it may be argued that a somewhat altered definition of intelligence may be more suited to our present notions of the underlying problems. It seems to us that such a modified Turing test should emphasize the development and learning of perceptual, motor, and language competence. In particular, many would agree that a system should be declared intelligent only if it is capable of first acquiring and then exhibiting motor, perceptual, and language skills. This revised Turing test expresses, in an explicit manner, the central role of learning in the art and science of replicating intelligence.

That the brain is not hard-wired but rather can be modified by experience is one of the most remarkable features of the nervous system [7]. At the systems level, the view of how neurons of the cerebral cortex mediate perception and action is presently undergoing a radical paradigm shift. Rather than being fixed in configuration, the functional anatomy of the cortex appears to display activity-dependent plasticity on a time-scale sufficiently short to play a role in most perceptual and motor tasks. Even simple brain functions, such as the definition of visual acuity, appear to involve a short-term learning phase and hence seem to depend upon experience. Computational methods have supported these findings and have provided models in which perceptual tasks are learned by simple experiential modifications of networks of cells. Thus, activity-dependent and task-dependent plasticity presents itself as a plausible explanation to how networks of neurons accomplish perception and action.

As a result of these findings, in both computer science and neuroscience, learning has in recent years advanced to the forefront of research. From this revolution – exemplified by the area of neural networks and connectionism [6] – has emerged an intriguing trend. What we would call low-level learning appears to produce more interesting results than one would have anticipated just a few years ago. In other words, "simple" learning techniques based on statistics, pattern recognition, neural networks, and function approximation – all of which are intimately related to one another – have demonstrated their utility in a number of real problems and in a broad spectrum of applications [4] [5] [6].

2. Learning and Vision

The above arguments in favor of learning are directly applicable to the field of computational vision. The emphasis in machine vision has been the development of recovery methods and algorithms that seek to extract from images information that is pertinent to the task at hand. The trend has been to allow the vision practitioner to input all of his or her expert knowledge of the application into the system in the form of models, constraints, or parameters. This procedure may suffice in the case of small vision systems that are configured for highly structured tasks. However, versatility and scalability are desirable attributes in most vision systems. From this viewpoint, the manual input (hard-wiring) of models, constraints, and parameters has proven itself impractical. The only solution, therefore, is to incorporate learning capabilities within the vision system that allow it to both learn to see, as well as, learn by seeing.

Does this argument imply that a vision system that learns can thrive without the use of models and constraints? Quite the contrary. In most instances, the process of learning in the absence of physical and geometrical constraints turns out to be hopelessly underconstrained. Those models and constraints that are powerful in their generality, can serve as important foundations and guides during the process of learning. It is the less powerful ones with narrow scope that we desire the vision system to automatically acquire. For instance, in the context of object recognition, the perspective projection model of image formation would serve as a general *a priori* constraint that may be hard-wired into the system. In contrast, the geometry of a particular object can be viewed as specific knowledge that the system must be capable of learning without assistance.

As stated earlier, the learning capabilities of neurons in the brain are becoming increasingly well-understood. Does it suffice then to use a connectionist architecture to solve all vision problems? In some instances, the answer is in the affirmative. However, we must acknowledge the fact that in many settings machine intelligence is not merely the enterprise of replicating human intelligence. Real-world applications, more often than not, demand machines to be far more quantitative than any known biological system. This calls for the creation and implementation of learning techniques and architectures that are complementary to biologically motivated ones.

Machine learning stands as an active research field all by itself [9] [8]. Over the years, this discipline has put forth a variety of methodologies, including inductive learning, connectionist learning, case-based learning, and genetic algorithms, to name a few. Is it not possible to view these results as a complete bag of tricks and simply cast visual learning as the art of selecting the appropriate machine learning technique for any given vision task? In some cases, such a strategy would work. However, visual perception brings with it its unique set of constraints and problems. For instance, the simultaneous occurrence of two objects in vision is not analogous to that of two voices in speech. The former could manifest the non-linear phenomenon of occlusion while the latter results in simple superposition. It may well be the case that the complexities inherent to visual perception require the creation of novel learning techniques. What a system needs to learn, and how the system learns it, has everything to do with the intrinsic properties of the data available to the system.

3. The Book at a Glance

This book focuses on visual learning techniques that are applied more or less directly to the signals provided by vision sensors. Hence, the use of the term "early" in the title of book; it is meant to reflect the stage of visual processing at which learning is invoked and does not connote learning in an infantile system, biological or artificial. Each chapter of the book presents recent results within this domain of signal-based visual learning. Each result is novel in either the underlying theory, the problem formulation, or the application that it seeks to solve. The chapters are designed to serve as technical overviews with extensive references to related work. It is hoped that this format will enable a novice to get quickly familiarized with the main issues in low-level visual learning, while providing an expert pointers to technical details. Additional reading material on the topic can be found in other recent proceedings and reports [3] [1] [2].

We begin with a chapter that brings to light recent psychophysical and electrophysiological results on the internal representation of three-dimensional objects. Pauls, Bricolo and Logothetis present a series of experimental data, obtained from macaque monkeys, that provide compelling evidence that the brain adopts a viewer-centered rather than an object-centered approach to visual representation. Single neuronal units in the inferior temporal cortex are found to exhibit high specificity, in that, they respond to particular views of three-dimensional objects. The response functions of these neurons are analyzed and shown to have unimodal bell-shaped structures. The authors argue that a small set of such neurons, each tuned to a different view of the same object, may together constitute a network that achieves view-invariant recognition via generalization. The authors report detailed experiments on the invariance of recognition to object position, scale, and orientation.

The type of network that Pauls *et al.* posit is the topic of the second chapter by Poggio and Beymer. They formulate the learning-from-example problem as one of multivariate function approximation from sparse data. The authors provide an overview of regularization networks and establish ties between such networks and classical techniques in pattern recognition as well as a number of popular neural network architectures. The notion of virtual examples is introduced as a means of solving the frequently encountered problem of insufficient training data. The authors apply their techniques to image analysis in vision and image synthesis in graphics, with particular emphasis on the pertinent problem of face recognition. Regularization networks are used to estimate parameters such as pose, expression, and face identity. The authors conclude with recent results on the detection of objects in cluttered scenes.

In the next chapter, a probabilistic framework for visual learning of objects is described by Pope and Lowe. The authors model three-dimensional objects as probability distributions that capture variations in visual appearance in image space. These distributions are organized at two levels. Large appearance variations are handled by partitioning the training data into clusters that correspond to different views. Within each cluster, smaller variations are represented as distributions that describe the presence (or absence) of features, their positions, and their physical attributes. The features used thus far include intensity edges, lines, junctions, parallel segments, and convex regions. Recognition is based on the relative positions and attributes of features in a novel image. Matching with a stored object database is done using the concept of probabilistic alignment.

The subsequent chapter by Moghaddam and Pentland also addresses the use of a probabilistic framework for unsupervised learning of objects. Principal component analysis is used to first reduce the dimensionality of the training image set. The result is a low-dimensional eigenspace that is tuned to the objects of interest. Within the eigenspace, individual classes of objects form clusters that in an unsupervised setting need to be automatically identified. Two types of density estimates are proposed for modeling the clusters. One assumes a multivariate unimodal Gaussian model while the other is a mixture-of-Gaussians model for multimodal distributions. The authors present algorithms for automatically deriving the above structures from training data. The proposed scheme of probabilistic visual modeling is applied to the problems of face recognition, face coding, and gesture recognition.

Nayar, Murase, and Nene present the parametric eigenspace as a continuous representation of visual appearance. Given a vision task, its parameter space is coarsely sampled to obtain the training image set. Principal component analysis is used to achieve dimension reduction and the training images are projected to the resulting eigenspace to obtain a cluster of discrete points. The authors argue that, given the structured nature of our visual world, the clusters often lie on smooth continuous manifolds in eigenspace. These appearance manifolds are constructed from the discrete eigenspace projections by interpolation. Given a novel input projection, recognition and parameter estimation are done using efficient search algorithms that determine the closest appearance manifold point. The authors demonstrate the utility of parametric appearance manifolds in real-time applications, such as, object recognition and visual servoing.

The next chapter by Pomerleau describes the application of classical neural networks to the problem of vision-based autonomous navigation. The chapter stresses on the importance of careful design and implementation in visual learning. The significance of preprocessing of visual data to extract pertinent features while removing less-informative ones is described. It is shown how transformations can be applied to a single image to generate shifted and rotated ones, consistent with perspective projection, that can serve as additional training data. The author describes the implementation of "on-the-fly" learning where the network learns the mapping between steering direction and visual input while a human drives the vehicle. The autonomous navigation system has been demonstrated to function in a variety of settings, including, single-lane paved and unpaved roads, multi-lane and unlined roads, and obstacle ridden roads, at speeds up to 55 miles an hour.

The next contribution by Weng brings together a collection of existing tools such as dimension reduction using principal component analysis, hierarchical feature grouping, space partitioning trees, bayesian decision theory, and search techniques, into a single architecture for visual learning. The author argues that to attain the level of generality required to deal with arbitrary scenes, a system must be capable of self-organizing itself based on the visual information it encounters. The proposed system uses coarse-to-fine attention selection and a multi-level network to facilitate simultaneous recognition and segmentation. In the early levels of the network, the visual input is classified into rough categories that are grouped together and refined in subsequent levels. Experimental results are reported on the application of the proposed architecture to the problems of object recognition, motion understanding, and autonomous navigation.

Nelson explores the complexity of visual learning. It is argued that among the known techniques of computational learning, the only one that can effectively assemble the large amounts of information that comes with a typical vision problem is the construction of an indexed memory. While backpropagation, simulated annealing, reinforcement learning, and genetic algorithms are N^2 in complexity, a locally optimized indexed memory can be constructed in $N \log N$. For object recognition, the author advocates an associative memory organization coupled with the use of robust semi-invariant features that can be computed from images. To demonstrate these ideas, the author presents a memory-based recognition system for polyhedral objects.

In their chapter, Fayyad, Smyth, Burl, and Perona remind us of the importance of scalability in visual learning. While the previous chapters have explored either specific domains or small sets of training data, this one addresses the role of learning in the context of very large image databases (ones that typically include terabytes of information) that are common in the areas of astronomy, remote sensing, and medicine. The authors illustrate the need for efficient image processing, data classification, and data management. Techniques are implemented for varying focus-of-attention and extracting principal features for dimension reduction. A variety of decision tree based learning algorithms are employed for automatic rule-extraction and classifier construction. The applications discussed include the classification of volcanoes in 30,000 SAR images of Venus gathered by the spacecraft Magellan. As a future application, the authors hope to detect galaxies and stellar objects from enormous amounts of image data.

Whereas the above chapters are concerned with the problems of recognition and navigation, the one by Bhanu, Wu, and Lee addresses the problem of image segmentation. Existing segmentation algorithms rely on an expert to use his or her instincts to tune the parameters of the algorithm until a reasonable segmentation is obtained. For instance, in the case of outdoor scenes, current segmentation techniques would need to be continuously tuned based on environmental conditions, the properties of the imaging devices used, and the time of day. The authors describe an adaptive segmentation system that uses a feedback loop consisting of a machine learning subsystem, a typical image segmentation algorithm, and a performance evaluation component. The learning subsystem is based on genetic adaptation and uses a combination of a classical genetic algorithm and hill climbing. Segmentation results for outdoor color imagery are presented.

Greenspan addresses the problem of learning and recognizing image textures. Gabor wavelet decomposition is used to represent textures in frequency and orientation space and extract a small number of pertinent features. A few different non-parametric supervised learning schemes are used to learn the visual properties of the textures. For classification, k-nearest neighbor and multi-layered perceptron based algorithms have been implemented. The recognition process is shown to be rotation-invariant due to the orientation sensitive representation used. The author reports experiments on the robustness of classification in the case of structured and unstructured textures, the scalability of the developed system, and the use of texture recognition for image segmentation.

Departing from mainstream vision problems, the final chapter by Salganicoff, Rucci, and Bajcsy addresses the problem of learning based control of a robot manipulator that relies on both visual and tactile information. The two sensory inputs are found to be consistent in a way that permits them to share information with one another. For instance, a particular visual event can lead to a sequence of tactile events and vice versa. This interplay allows one form of stimulus to corroborate as well as complement the other. The authors use a neural network to enable the manipulator to shift its gaze direction and foveate based on visual and tactile stimuli. In addition, an active learning approach is proposed for automatically generating object grasping strategies based on range scanner and tactile sensor inputs.

Needless to say, the chapters of this book by no means serve as a complete survey of low-level visual learning. It only embodies a coarse sampling of some of the representative paradigms that have emerged in recent years. Furthermore, of equal importance are learning issues related to higher levels of visual representation and perception. This is the topic of a companion book edited by K. Ikeuchi and M. Veloso that summarizes recent results on symbolic methods for visual learning. It is hoped that the two books together will provide a useful overview of the present state of visual learning and aid in the investigation of this relatively unexplored area of vision research.

References

[1] Y. Aloimonos, R. S. Michalski, and A. Rosenfeld, "Machine Vision and Learning: Research Issues and Directions," *Report on the NSF/ARPA Workshop on Machine Vision and Learning*, Harpers Ferry, WV, Oct. 1992.

[2] B. Bhanu and T. Poggio, Editors, Special section on "Learning in Computer Vision," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. 16, No. 9, Sept. 1994.

[3] K. Bowyer and L. Hall, Editors, *Proc. of AAAI Fall Symposium on Machine Learning in Computer Vision: What, Why, and How?,* Raleigh, NC, Oct. 1993.

[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis,* Wiley, NY, 1973.

[5] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* Second edition, Academic Press, San Diego, CA, 1990.

[6] S. Haykin, *Neural Networks: A Comprehensive Foundation,* Maxwell Macmillan International, NY, 1994.

[7] E. R. Kandel, J. H. Schwartz, and T. J. Jessell, Editors, *Principles of Neural Science,* Third edition, Appleton & Lange, Norwalk, CN, 1991.

[8] P. Langley, *Elements of Machine Learning,* Morgan Kaufmann Publishers, San Fransisco, CA, 1996.

[9] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, Editors, *Machine Learning: An Artificial Intelligence Approach,* Vols. 1 & 2, Morgan Kaufmann, San Fransisco, CA, 1986.