Chapter VIII. Towards a biologically plausible computational model of ventral visual cortex

We have now come a long way since our initial steps towards defining the proble of visual recognition. We started with characterizing the spatial and temporal statistics of natural images (Lecture 2). We explored how neurons along ventral visual cortex respond to a variet of different stimulus conditions (Lectures 3, 5, 7, 8). We described the recognition impairments that arise through cortical lesions (Lecture 4) and the effect of applying currents to the neural circuitry (Lecture 9). We would like to put all of these separate bits and pieces of data into a coherent framework to rigorously understand how neuronal circuits help us recognize objecst. Here we summarize some of the initial steps towards a theoretical understanding of the computational principles behind transformation-invariance visual recognition in the primate cortex.

8.1. Defining the problem

We start by defining what needs to be explained and the necessary constraints to solve the problem. A theory of visual object recognition, implemented by a computational model, should be able to explain the following phenomena and have the following characteristics:

1. *Selectivity*. The primate visual system shows a remarkable degree of selectivity and can differentiate among shapes that appear to be very similar at the pixel level (e.g. arbitrary 3D shapes created from paperclips, different faces, etc.). Critical to object recognition, a model should be able to discriminate among physically similar but distinct shapes.
2. *Transformation tolerance*. A trivial solution to achieve high selectivity would be to memorize all the pixels in the object. The problem with this type of algorithm is that it would not tolerate any changes in the image. An object can cast an infinite number of projections onto the retina. These changes arise due to changes in object position with respect to fixation, object scale, plane or depth rotation, changes in contrast or illumination, color, occlusion and others. The importance of combining selectivity and tolerance has been emphasized by many investigators (e.g. (Rolls, 1991; Olshausen et al., 1993; Logothetis and Sheinberg, 1996; Riesenhuber and Poggio, 1999; Deco and Rolls, 2004b; Serre et al., 2007b) among others).
3. *Speed*. Visual recognition is very fast, as emphasized by many psychophysical investigations (Potter and Levy, 1969; Kirchner and Thorpe, 2006; Serre et al., 2007a), scalp EEG measurements (Thorpe et al., 1996) and neurophysiological recordings in humans (Liu et al., 2009) and monkeys (e.g. (Richmond et al., 1983; Keysers et al., 2001; Hung et al., 2005) among others). This speed imposes an important constraint to the number of computational steps that the visual system can use for pattern recognition (Rolls, 1991; Serre et al., 2007b).

4. *Generic*. We can recognize a large variety of objects and shapes. Estimates about the exact number of objects or object categories that primates can discriminate vary widely depending on several assumptions and extrapolations (e.g. (Standing, 1973; Biederman, 1987; Abbott et al., 1996; Brady et al., 2008)). Certain types of shapes may be particularly interesting, they may have more cortical real estate associated with them, they could be processed faster and could be independently impaired. For example, there has been extensive discussion in the literature about faces, their representation and how they can be different from other visual stimuli. Yet, independently of precise figures about the number of shapes that primates can discriminate and independently also of whether natural objects and faces are special or not, it is clear that there exists a generic system capable of discriminating among multiple arbitrary shapes. For simplicity and generality, we focus first on this generic shape recognition problem. Face recognition, or specialization for natural objects versus other shapes constitute interesting and important specific instantiations and sub problems of the general one that we try to address here.

5. *Implementable in a computational algorithm*. A successful theory of visual object recognition needs to be described in sufficient detail to be implemented through computational algorithms. This requirement is important because the computational implementation allows us to run simulations and hence to quantitatively compare the performance of the model against behavioral metrics. The simulations also lend themselves to a direct comparison of the model's computational steps and neurophysiological responses at different stages of the visual processing circuitry. The algorithmic implementation forces us to rigorously state the assumptions and formalize the computational steps; in this way, computational models can be more readily compared than "armchair" theories and models. The implementation can also help us debug the theory by discovering hidden assumptions, bottlenecks and challenges that the algorithms cannot solve or where performance is poor. There are multiple fascinating ideas and theories about visual object recognition that have not been implemented through computational algorithms. These ideas can be extremely useful and helpful for the field and can inspire the development of computational models. Yet, we emphasize that we cannot easily compare theories that can be and have been implemented against other ones that have not.

6. *Restricted to primates*. Here we restrict the discussion to object recognition in primates. There are strong similarities in visual object recognition at the behavioral and neurophysiological levels between macaque monkeys (one of the prime species for neurophysiological studies) and humans (e.g. (Myerson et al., 1981; Logothetis and Sheinberg, 1996; Orban, 2004; Nielsen et al., 2006; Kriegeskorte et al., 2008; Liu et al., 2009).

7. Biophysically plausible. There are multiple computational approaches to visual object recognition. Here we restrict the discussion to models that are biophysically plausible. In doing so, we ignore a vast literature in Computer Vision where investigators are trying to solve similar problems without direct reference to the cortical circuitry. These engineering approaches are extremely interesting and useful from a practical viewpoint. Ultimately, in the same way that

computers can become quite successful at playing chess without any direct connection to the way humans play chess, computer vision approaches can achieve high performance without mimicking neuronal circuits. Here we restrict the discussion to biophysically plausible algorithms.

8. *Restricted to the visual system*. The visual system is not isolated from the rest of the brain and there are plenty of connections between visual cortex and other sensory cortices, between visual cortex and memory systems in the medial temporal lobe and between the visual cortex and frontal cortex. It is likely that these connections also play an important role in the process of visual recognition, particularly through feedback signals that incorporate expectations (e.g. the probability that there is a lion in an office setting is very small), prior knowledge and experience (e.g. the object appears similar to another object that we are familiar with), cross-modal information (e.g. the object is likely to be a musical instrument because of the sound). To begin with and to simplify the problem, we restrict the discussion to the visual system.

8.2. Visual recognition goes beyond identifying objects in single images

We emphasize that visual recognition is far more complex than the identification of specific objects. Under natural viewing conditions, objects are embedded in complex scenes and need to be separated from their background. How this segmentation occurs constitutes an important challenge in itself. Segmentation depends on a variety of cues including sharp edges, texture changes and object motion among others. Some object recognition models assume that segmentation must occur prior to recognition. There is no clear biological evidence for segmentation prior to recognition and therefore this constitutes a weakness in such approaches. We do not discuss segmentation here (see (Borenstein et al., 2004; Sharon et al., 2006) for recent examples of segmentation algorithms).

Most object recognition models are based on studying static images. Under natural viewing conditions, there are important cues that depend on the temporal integration of information. These dynamic cues can significantly enhance recognition. Yet, it is clear that we can recognize objects in static images and therefore many models focus on the reduced version the pattern recognition problem using static objects. Here we also focus on static images.

We can perform a variety of complex tasks that rely on visual information that are different from identification. For example, we can put together images of snakes, lions and dolphins and categorize them as animals. Categorization is a very important problem in vision research and it also constitutes a formidable challenge for computer-based approaches. Here we focus on the question of object identification.

8.3. Modeling the ventral visual stream – Common themes

Several investigators have proposed computational models that aim to capture some of the essential principles behind the transformations along the primate ventral visual stream. Before discussing some of those models in more detail, we start by providing some common themes that are shared by many of these models.

The input to the models is typically an image, defined by a matrix that contains the grayscale value of each pixel. Object shapes can be discriminated quite well in grayscale images and, therefore, most models ignore the added complexities of color processing (but eventually it will also be informative and important to add color to these models). Because the focus is often on the computational properties of ventral visual cortex, several investigators often ignore the complexities of modeling the computations in the retina and LGN; the pixels are meant to coarsely represent the output of retinal ganglion cells or LGN cells. This is of course one of the many oversimplifications in several computation models given that we know that images go through a number of transformations before retinal ganglion cells convey information to the LGN and on to cortex (Meister, 1996).

Most models have a hierarchical and deep structure that aims to mimic the approximately hierarchical architecture of ventral visual cortex (Felleman and Van Essen, 1991; Maunsell, 1995). The properties of deep networks has received considerable attention in the computational world, even if the mathematics of learning in deep networks that include non-linear responses is far less understood than shallow counterparts (Poggio and Smale, 2003). It seems that neocortex and computer modelers have adopted a *Divide and Conquer* strategy whereby a complex problem is divided into many simpler tasks.

Most computational models assume, explicitly or implicitly, that cortex is cortex, and hence that there exist canonical microcircuits and computations that are repeated over and over throughout the hierarchy (Riesenhuber and Poggio, 1999; Douglas and Martin, 2004; Serre et al., 2007b).

As we ascend through the hierarchical structure of the model, units in higher levels typically have larger receptive fields, respond to more complex visual features and show an increased degree of tolerance to transformations of their preferred features.

8.4. A panoply of models

We summarize here a few important ideas that have been developed to describe visual object recognition. The presentation here is neither an exhaustive list nor a thorough discussion of each of these approaches. For a more detailed discussion of several of these approaches, see (Ullman, 1996; LeCun et al., 1998; Riesenhuber and Poggio, 2002; Deco and Rolls, 2004a; Serre et al., 2005b).

Straightforward template matching does not work for pattern recognition. Even shifting a pattern by one pixel would pose significant challenges for an algorithm that merely compares the input with a stored pattern on a pixel-by-pixel fashion. As noted at the beginning of this chapter, a key challenge to recognition is that an object can lead to infinite number of retinal images depending on its size, position, illumination, etc. If all objects were always presented in a standardized position, scale, rotation and illumination, recognition would be considerably easier (DiCarlo and Cox, 2007; Serre et al., 2007b). Based on this notion, several approaches are based on trying to transform an incoming object into a canonical prototypical format by shifting, scaling and rotating objects (e.g. (Ullman, 1996)). The type of transformations required is usually rather complex, particularly for non-affine transformations. While some of these problems can be overcome by ingenious computational strategies, it is not entirely clear (yet) how the brain would implement such complex calculations nor is there currently any clear link to the type of neurophysiological responses observed in ventral visual cortex.

A number of approaches are based on decomposing an object into its component parts and their interactions. The idea behind this notion is that there could be a small dictionary of object parts and a small set of possible interactions that act as building blocks of all objects. Several of these ideas can be traced back to the prominent work of David Marr (Marr and Nishihara, 1978; Marr, 1982) where those constituent parts were based on generalized cone shapes. The artificial intelligence community also embraced the notion of structural descriptions (Winston, 1975). In the same way that a mathematical function can be decomposed into a sum over a certain basis set (e.g. polynomials or sine and cosine functions), the idea of thinking about objects as a sum over parts is attractive because it may be possible and easier to detect these parts in a transformation-invariant manner (Biederman, 1987; Mel, 1997). In the simplest instantiations, these models are based on merely detecting a conjunction of object parts, an approach that suffers from the fact that part rearrangements would not impair recognition but they should (e.g. a house with a garage on the roof and the chimney on the floor). More elaborate versions include part interactions and relative positions. Yet, this approach seems to convert the problem of object recognition to the problem of object part recognition plus the problem of object parts interaction recognition. It is not entirely obvious that object part recognition would be a trivial problem in itself nor is it obvious that *any* object can be uniquely and succinctly described by a universal and small dictionary of simpler parts. It is not entirely trivial how recognition of complex shapes (e.g. consider discriminating between two faces) can be easily described in terms of a structural description of parts and their interactions. Computational implementations of these structural descriptions have been sparse (see however (Hummel and Biederman, 1992)). More importantly, it is not entirely apparent how these structural descriptions relate to the neurophysiology of the ventral visual cortex (see however (Vogels et al., 2001)).

A series of computational algorithms, typically rooted in the neural network literature (Hinton, 1992), attempt to build deep structures where inputs can be reconstructed (for a recent version of this, see e.g. (Hinton and Salakhutdinov, 2006). In an extreme version of this approach, there is no information loss along the deep hierarchy and backward signals carry information capable of re-creating arbitrary inputs in lower visual areas. There are a number of interesting applications for such "auto-encoder" deep networks such as the possibility of performing dimensionality reduction. From a neurophysiological viewpoint, however, it seems that the purpose of cortex is precisely the opposite, namely, to lose information in biologically interesting ways. It is not clear why one build an entire network to copy the input (possibly with fewer units). In other words, as emphasized at the beginning of this chapter, it seems that a key goal of ventral visual cortex is to be able to extract biologically relevant information (e.g. object identity) in spite of changes in the input at the pixel level.

Particularly within the neurophysiology community, there exist several "metric" approaches where investigators attempt to parametrically define a space of shapes and then record the activity of neurons along the ventral visual stream in response to these shapes (Tanaka, 1996; Brincat and Connor, 2004; Connor et al., 2007). This dictionary of shapes can be more or less quantitatively defined. For example, in some cases, investigators start by presenting different shapes in search of a stimulus that elicits strong responses. Subsequently, they manipulate the "preferred" stimulus by removing different parts and evaluating how the neuronal responses are modified by these transformations. While interesting, these approaches suffer from the difficulties inherent in considering arbitrary shapes that may or may not constitute truly "preferred" stimuli. Additionally, in some cases, the transformations examined only reveal anthropomorphic biases about what features could be relevant. Another approach is to define shapes parametrically. For example, Brincat and colleagues considered a family of curvatures and modeled responses in a six-dimensional space defined by a sum of Gaussians with parameters given by the curvature, orientation, relative position and absolute position of the contour elements in the display. This approach is intriguing because it has the attractive property of allowing investigators to plot "tuning curves" similar to the ones used to represent the activity of units in earlier visual areas. Yet, it also makes strong assumptions about the type of shapes preferred by the units. Expanding on these ideas, investigators have tried to start from generic shapes and use genetic algorithms whose trajectories are guided by the neuronal preferences (Yamane et al., 2008). What is particularly interesting about this approach is that it seems to be less biased than the former two. The key limitation here is the recording time and this type of algorithm, particularly with small data sets, may converge onto local minima or even not converge at all. Genetic algorithms can be more thoroughly examined in the computational domain. For example, investigators can examine a huge variety of computational models and let them "compete" with each other through evolutionary mechanisms (Pinto et al., 2009). To guide the evolutionary paths, it is necessary to define a cost function; for example, evolution can be constrained by rewarding models that achieve better performance in certain recognition tasks. This type of approach can lead to models with high performance (although it also suffers from difficulties related to local minima). Unfortunately, it is not obvious that better performance necessarily implies any better approximation to the way in which cortex solves the visual recognition problem.

8.5. Bottom-up hierarchical models of the ventral visual stream

A hierarchical network model can be described by a series of layers $i = 0,1,...,N$. Each layer contains $n(i) \times n(i)$ units arranged in a matrix. The activity of each unit in each layer can be represented by the matrix $\mathbf{x}_i$ ($\mathbf{x}_i \in \mathbb{R}^{n(i)\times n(i)}$). In several models, $x_i(j,k)$ (i.e., the activity of unit at position *j,k* in layer *i*) is a scalar value interpreted as the firing rate of the unit. The initial layer is defined as the input image; $\mathbf{x}_0$ represents the (grayscale) values of the pixels a given image.

Equations 1 and 2 above constitute the initial steps for many object recognition models and capitalize on the more studied parts of the visual system, the pathway from the retina to primary visual cortex. The output of Equation 2, after convolving the output of center-surround receptive fields with a Gabor function, can be thought of as a first order approximation to the edges in the image. As noted above, our understanding of ventral visual cortex beyond V1 is far more primitive and it is therefore not surprising that this is where most models diverge. In a first order simplification, we can generically describe the transformations along the ventral visual stream as:

$$\mathbf{x}_{i+1} = f_i(\mathbf{x}_i) \qquad \textbf{Equation 11.1}$$

This assumes that the activity in a given layer only depends on the activity pattern in the previous layer. This assumption implies that at least three types of connections are ignored: (i) connections that “skip” a layer in the hierarchy (e.g. synapses from the LGN to V4 skipping V1); (ii) top-down connections (e.g. synapses from V2 to V1 (Virga, 1989)) and (iii) connections within a layer (e.g. horizontal connections between neurons with similar preferences in V1 (Callaway, 1998)) are not included in **Equation 11.1**.

The subindex *i* in the function *f* indicates that the transformation from one layer to another is not necessarily the same. A simple form that *f* could take is a linear function:

$$\mathbf{x}_{i+1} = \mathbf{W}_i\mathbf{x}_i \qquad \textbf{Equation 11.2}$$

where the matrix $\mathbf{W}_i$ represents the linear weights that transform activity in layer *i* into activity in layer *i+1*. Not all neurons in layer *i* need to be connected to all neurons in layer *i+1*; in other words, many entries in **W** can be 0. This simple formulation fins some empirical evidence; for example, Hubel and Wiesel proposed that the oriented filters in primary visual cortex could arise from a linear summation of the activity of neurons in the lateral geniculate nucleus with appropriately aligned center-surround receptive fields (Hubel and Wiesel, 1962). Unfortunately, neurons are far more intricate devices and non-linearities abound in their response properties. For example, Hubel and Wiesel also described the activity of so-called complex cells that are also orientation tuned but show a non-linear response as a function of spatial frequency or bar length.

It is tacitly assumed by most modelers that there exist general rules, often summarized in the epithet “cortex is cortex”, such that only a few such transformations are allowed. One of the early models that aimed to describe object recognition, inspired by the neurophysiological findings of Hubel and Wiesel, was the neocognitron (Fukushima, 1980) (see also earlier theoretical ideas in (Sutherland, 1968)). This model had two possible operations, a linear tuning function (performed by “simple” cells) and a non-linear OR operation (performed by "complex" cells). These two operations were alternated and repeated through the multiple layers in the deep hierarchy. This model demonstrated the feasibility of such linear/non-linear cascades in achieving scale and position tolerance in a letter recognition task. Several subsequent efforts (Olshausen et al., 1993; Wallis and Rolls, 1997; LeCun et al., 1998; Riesenhuber and Poggio, 1999; Amit and Mascaro, 2003; Deco and Rolls, 2004b) were inspired by the Neocognitron architecture.

One such effort to expand on the computational abilities of the Neocognitron in the computational model developed in the Poggio group (Riesenhuber and Poggio, 1999; Serre et al., 2005b; Serre et al., 2007b). This model is characterized by a purely feed-forward and hierarchical architecture. An image, represented by grayscale values, is convolved with Gabor filters at multiple scales and positions to mimic the responses of simple cells in primary visual cortex. Like other efforts, the model consists of a cascade of linear and non-linear operations. These operations come in only two flavors in the model: a tuning operation and soft-max operation. Both operations can be expressed in the following form:

$$x_{i+1}[k] = g\left(\frac{\sum_{j=1}^{n} w[j,k]\ x_i^p[j]}{\alpha + \left(\sum_{j=1}^{n} x_i^q[j] \right)^r} \right)$$ **Equation 11.3**

where $x_{i+1}[k]$ represents the activity of unit *k* in layer *i+1*, $w[j,k]$ represents the connection weight between unit *j* in layer *i* and unit *k* in layer *b+1*, *p, q, r* are integer constants, *a* is a normalization constant and *g* is a nonlinear function (e.g. sigmoid). Depending on the values of *p*, *q* and *r* different interesting behaviors can be obtained. In particular, taking *r=1/2*, *p=1*, *q*=2, leads to a *tuning operation*:

$$x_{i+1}[k] = g\left(\frac{\sum_{j=1}^{n} w[j,k]\ x_i[j]}{\alpha + \sqrt{\sum_{j=1}^{n} x_i^2[j]}} \right)$$ **Equation 11.3'**

The responses of the unit are controlled by the weights **w**. As emphasized above, tuning is a central aspect of any computational model of visual recognition, allowing units along the hierarchy to respond to increasingly more elaborate features. Taking **w**=1, *p=q+1*, *r=1*, leads to a softmax operation, particularly for large values of *q*:

$$x_{i+1}[k] = g\left(\frac{\sum_{j=1}^{n} x_i^{q+1}[j]}{\alpha + \sum_{j=1}^{n} x_i^q[j]} \right)$$ **Equation 11.3''**

Of note, the unit with response $x_{i+1}[k]$ shows similar response tuning to the units with response $x_i[j]$ for $j = 1,...,n$. Yet, the higher-level unit shows a stronger degree of tolerance to those aspects of the response that differentiate the responses of different units with similar tuning in layer *i*. For example, different units in layer *i* may show identical feature preferences but have slightly different receptive fields. A winner-take-all unit in layer i+1 that takes as input the responses of those units would inherit the same feature preferences but would reveal a larger receptive and tolerate changes in the position of the feature within the larger receptive field. Both operations can be implemented through relatively simple biophysical circuits (Kouh and Poggio, 2004).

This and similar architectures have been subjected to several tests including comparison with psychophysical measurements (e.g. (Serre et al., 2007a)), comparison with neurophysiological responses (e.g. (Deco and Rolls, 2004b; Lampl et al., 2004; Hung et al., 2005; Serre et al., 2005b; Cadieu et al., 2007) and quantitative evaluation of performance in computer vision recognition tasks (e.g. (LeCun et al., 1998; Serre et al., 2005a; Mutch and Lowe, 2006)).

8.6. Top-down signals in visual recognition

In spite of the multiple simplifications, the success of bottom-up architecture in describing a large number of visual recognition phenomena suggest that they may not be a bad first cut. As emphasized above, bottom-up architectures constitute only an approximation to the complexities and wonders of neocortical computation.One of the several simplifications in bottom-up models is the lack of top-down signals. We know that there are abundant back-projections in neocortex (e.g. (Felleman and Van Essen, 1991; Callaway, 2004; Douglas and Martin, 2004)). The functions of top-down connections have been less studied at the neurophysiological level but there is no shortage of computational models illustrating the rich array of computations that emerge with such connectivity. Several models have used top-down connections to guide attention to specific locations or specific features within the image (e.g. (Olshausen et al., 1993; Itti and Koch, 2001))(Tsotsos, 1990; Deco and Rolls, 2005; Rao, 2005; Compte and Wang, 2006; Chikkerur et al., 2009). The allocation of attention to specific parts of an image can significantly enhance recognition performance by alleviating the problems associated with image segmentation and with clutter.

Top-down signals can also play an important role in recognition of occluded objects. When only partial object information is available, the system must be able to perform object completion and interpret the image based on prior knowledge. Attractor networks have been shown to be able to retrieve the identity of stored memories from partial information (e.g. (Hopfield, 1982)). Some computational models have combined bottom-up architectures with attractor networks at the top of the hierarchy (e.g. (Deco and Rolls, 2004b)).

During object completion, top-down signals could play an important role by providing prior stored information that influences the bottom-up sensory responses. Several proposals have argued that visual recognition can be formulated as a Bayesian inference problem (Mumford, 1992; Rao et al., 2002; Lee and Mumford, 2003; Rao, 2004; Yuille and Kersten, 2006; Chikkerur et al., 2009). Considering three layers of the visual cascade (e.g. LGN, V1 and higher areas), and denoting activity in those layers as $\mathbf{x}_0$, $\mathbf{x}_1$ and $\mathbf{x}_h$ respectively, then the probability of obtaining a given response pattern in V1 depends both on the sensory input as well as feedback from higher areas:

$$P(\mathbf{x}_1|\mathbf{x}_0) = \frac{P(\mathbf{x}_0|\mathbf{x}_1)P(\mathbf{x}_1|\mathbf{x}_h)}{P(\mathbf{x}_0|\mathbf{x}_h)} \qquad \textbf{Equation 11.8}$$

where $P(\mathbf{x}_1|\mathbf{x}_h)$ represents the feedback biases conveying prior information. An intriguing idea proposed by Rao and Ballard argues that top-down connections provide predictive signals whereas bottom-up signals convey the difference between the sensory input and the top-down predictions (Rao and Ballard, 1999).

8.7. Applications

There is no shortage of applications where automatic or semi-automatic algorithms are being explored in computer vision. Here are a few examples:

(A) Intelligent content-based search. Searching for images in the web by content will open the doors to a large number of applications. Facebook users can already experiment with prototypes that let them search for people. One may be able to look for images that are similar to a search query in terms of content. Blind people may be able to point their phones and find out where they are and how to navigate.
(B) Prototype cars that can navigate automatically rely heavily on algorithms to detect pedestrians, other cars, other objects and road conditions.
(C) ATM machines may be able to recognize their customers. Cars and houses may recognize their owners.
(D) Security screening in places like airports may benefit from automatic recognition systems.
(E) Several clinical problems are based on pattern recognition and computers may soon help doctors to make informed decisions based on their understanding of patterns.

8.8. Computer vision tasks

Algorithms have been developed to address several interrelated problems in machine vision. While some of the boundaries are blurred in several applications, it is useful to think about the following tasks:

(A) Object detection. For example, a digital camera may require detecting the presence and location of a face in an image for focusing. Face detection may thrive without solving the problem of recognition.
(B) Object segmentation. In natural images, it may be of interest to separate an object from the background. For example, it may be important to detect the location of a tumor in an image. Or to detect the presence of a tank in a camouflaged image.
(C) Object recognition. Recognizing objects can often be thought of as associating the image with labels. These labels may refer to the identity of

the object (e.g. given a face, who is it?) or the object's category (e.g. is there an animal in this image?).

(D) Object verification. In some cases, it may be of interest to evaluate whether two images are the same or not.

8.9. Object segmentation

Given a natural scene, humans (and other species) are quite good at being able to characterize and localize different objects embedded in complex backgrounds. The fact that this is not a trivial problem is highlighted by the ubiquitous use of camouflage in the animal world. Particularly for objects that are not moving, matching colors, contrast and textures can help animals avoid predators or at least buy sufficient time for escape. Basic aspects of segmentation may depend on adequately detecting edges. However, more complex problems often involve a deeper understanding of the interrelationships among different object parts. A typical case involves recognizing a zebra as a whole animal as opposed to thinking of each stripe as a separate object.

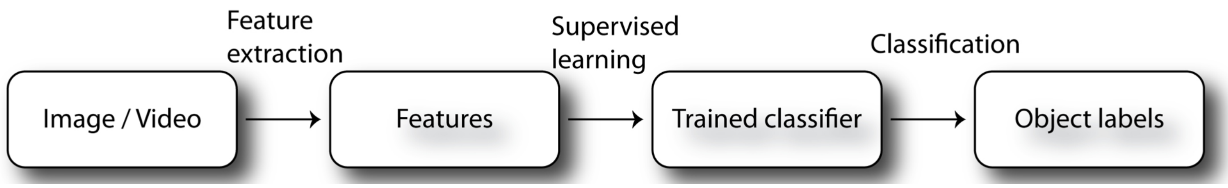
Some algorithms require recognition prior to segmentation while other algorithms use segmentation to guide recognition in complex scenes. To avoid this chicken-and-egg dilemma, it is tempting to speculate that certain aspects of bottom-up recognition and segmentation could occur (or at least) start independently of each other, using overlapping neuronal circuits. Top-down signals may then combine segmentation and recognition in synergistic fashion. For examples of object segmentation algorithms see (Borenstein et al., 2004).

8.10. A general scheme for object recognition

Figure 12.1 illustrates a typical approach in computer vision efforts. Consider a series of N labeled images ($\boldsymbol{x}_i, y_i$) where i=1,…,N, $\boldsymbol{x}$ is a matrix representing the image and y is a label (e.g. face present or not). A set of features f is extracted from the images: $\boldsymbol{f}_i=g(\boldsymbol{x}_i)$. Those features may include properties such as edges, principal components, etc. How those features are chosen is one of the key aspects that differentiates computer vision algorithms. A supervised learning scheme is then used to learn the map between those features and labels (Poggio and Smale, 2003; Meyers and Kreiman, 2011; Singer and Kreiman, 2011). For example, a support vector machine (SVM) classifier with a linear kernel may be used to learn the structure of the data and labels. A cross-validation procedure is followed by separating the data into a training set and a test set to avoid overfitting. After training, the algorithm is evaluated with the images in the test set. By using different algorithms applied to the same data, the merits of alternative approaches can be quantitatively compared.

8.11. A successful example: digit recognition

Figure 12.1. A general scheme for object recognition. Features are extracted from an image (or video). Those features are used to train a classifier via supervised learning. The resulting classification boundary is used with novel images (different from the ones used during training) to assign object labels to images.



Recognizing hand-written digits constitutes an example where computers have reached high accuracy, almost comparable to human levels (e.g. (LeCun et al., 1998)). Figure 12.2 shows an example of the errors made by an early attempt at recognizing hand-written digits. The overall error rate of this algorithm was <2%. Several of those errors are not trivial to recognize and humans could make mistakes as well.

8.12. Image recognition competitions

There are several computer vision competitions with large data sets consisting of labeled images. One such competition is called the Imagenet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2014). The 2014 instantiation of the object classification part of this challenge included 1000 object classes, 1,281,413 images for training (732-1300 images per class) and 100,000 images for testing (100 images per class). This competition also includes other tasks beyond classification including object detection and localization. To give an idea of performance, the winning team in the object classification part of the challenge achieved an error rate slightly above 6%. This is quite impressive considering that there were 1000 classes. It should be noted, though, that the results of these competitions are often reported in a somewhat strange way by allowing the models 5 changes to get it right and reporting the results as correct if any of those 5 predictions are correct. This makes it a bit more difficult to directly compare against human performance (Borji and Itti, 2014). Another aspect of machine vision that has also been highlighted is the difficulty in interpreting how the machine classifies objects and

Figure 12.2. Example of digit recognition mistakes by the algorithm in LeCun et al 1988. Below each digit, the image shows the true label and the computer label.

investigators have reported puzzling examples where minimal changes to an image drastically change the predicted class (Szegedy et al., 2014). With that said, the results are still quite remarkable and they show rapid progress in teaching machines to recognize objects.

8.7. The road ahead

If you can do it, a computer can do it too. Significant progress has been made over the last decade in teaching computers to perform multiple tasks that were traditionally thought to be the domain of humans. Any desktop computer can play chess competitively and the best computers can beat the world's chess champion. IBM's Watson has thrived in the trivia-like game of Jeopardy. And while imperfect, Siri and related systems are making enormous strides in becoming the world's best assistants.

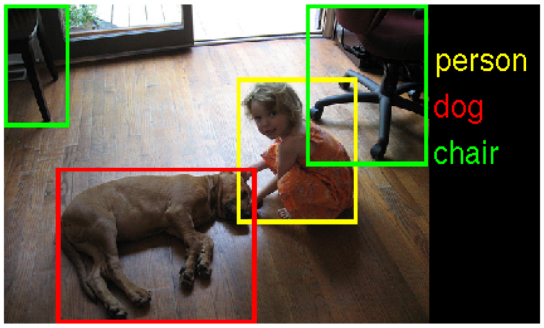
In the domain of vision, computational algorithms are already able to perform certain tasks such as recognizing digits in a fully automatic fashion at human performance level and demonstrate reasonable performance in other tasks such as detecting faces for focusing on in digital cameras. In several other tasks, humans still outperform the most sophisticated current algorithms but the gap between machines and humans in vision tasks is closing rapidly. Here we provide an overview of several computer vision systems, particularly in the context of pattern recognition problems and describe what machine vision systems can and cannot do,

Significant progress has been made towards describing visual object recognition in a principled and theoretically sound fashion. Yet, the lacunas in our understanding of the functional and computational architecture of ventral visual cortex are not small. The preliminary steps have distilled important principles of neocortical computation including deep networks that can divide and conquer complex tasks, bottom-up circuits that perform rapid computations, gradual increases in selectivity and tolerance to object transformation.

In stark contrast with the pathway from the retina to primary visual cortex, we do not have a quantitative description of the feature preferences of neurons along the ventral visual pathway. And several computational models do not make clear, concrete and testable predictions towards systematically characterizing ventral visual cortex at the physiological levels. Computational models can perform several complex recognition tasks and compete against non-biological computer vision approaches. Yet, for the vast majority of recognition tasks, they still fall significantly below human performance.

The next several years are likely to bring many new surprises in the field. We will be able to characterize the system at unprecedented resolution at the experimental level and we will be able to evaluate sophisticated and computationally intensive theories in realistic times. In the same way that the younger generations are not surprised by machines that can play chess quite competitively, the next generation may not be surprised by intelligent devices that can "see" like we do.

Figure 12.3. Example labeled image from the validation set in the 2013 ILSVRC competition. [Image source: http://www.image-net.org/challenges/LSVRC/2013/]



8.13. References

Abbott LF, Rolls ET, Tovee MJ (1996) Representational capacity of face coding in monkeys. Cerebral cortex 6:498-505.

Amit Y, Mascaro M (2003) An integrated network for invariant visual detection and recognition. Vision research 43:2073-2088.

Biederman I (1987) Recognition-by-components: A theory of human image understanding. Psychological Review 24:115-147.

Borenstein E, Sharon E, Ullman S (2004) Combining Top-Down and Bottom-Up Segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, DC.

Borji A, Itti L (2014) Human vs. computer in scene and object recognition. In: CVPR.

Brady TF, Konkle T, Alvarez GA, Oliva A (2008) Visual long-term memory has a massive storage capacity for object details. Proceedings of the National Academy of Science U S A 105:14325-14329.

Brincat SL, Connor CE (2004) Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nature Neuroscience 7:880-886.

Cadieu C, Kouh M, Pasupathy A, Connor C, Riesenhuber M, Poggio T (2007) A model of V4 shape selectivity and invariance. Journal of Neurophysiology 98:1733-1750.

Callaway EM (1998) Local circuits in primary visual cortex of the macaque monkey. Annu Rev Neurosci 21:47-74.

Callaway EM (2004) Feedforward, feedback and inhibitory connections in primate visual cortex. Neural Netw 17:625-632.

Chikkerur S, Serre T, Poggio T (2009) A Bayesian inference theory of attention: neuroscience and algorithms. In: (MIT-CSAIL-TR, ed). Cambridge: MIT.

Compte A, Wang XJ (2006) Tuning curve shift by attention modulation in cortical neurons: a computational study of its mechanisms. Cerebral cortex 16:761-778.

Connor CE, Brincat SL, Pasupathy A (2007) Transformation of shape information in the ventral pathway. Current opinion in neurobiology 17:140-147.

Deco G, Rolls ET (2004a) Computational Neuroscience of Vision. Oxford Oxford University Press.

Deco G, Rolls ET (2004b) A neurodynamical cortical model of visual attention and invariant object recognition. Vision research 44:621-642.

Deco G, Rolls ET (2005) Attention, short-term memory, and action selection: a unifying theory. Prog Neurobiol 76:236-256.

DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends Cogn Sci 11:333-341.

Douglas RJ, Martin KA (2004) Neuronal circuits of the neocortex. Annu Rev Neurosci 27:419-451.

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cerebral cortex 1:1-47.

Fukushima K (1980) Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics 36:193-202.

Hinton G (1992) How neural networks learn from experience. Scientific American 267:145-151.

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313:504-507.

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. PNAS 79:2554-2558.

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology 160:106-154.

Hummel JE, Biederman I (1992) Dynamic binding in a neural network for shape recognition. Psychol Rev 99:480-517.

Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast Read-out of Object Identity from Macaque Inferior Temporal Cortex. Science 310:863-866.

Itti L, Koch C (2001) Computational modelling of visual attention. Nat Rev Neurosci 2:194-203.

Keysers C, Xiao DK, Foldiak P, Perret DI (2001) The speed of sight. Journal of Cognitive Neuroscience 13:90-101.

Kirchner H, Thorpe SJ (2006) Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. Vision research 46:1762-1776.

Kouh M, Poggio T (2004) A general mechanism for tuning: gain control circuits and synapses underlie tuning of cortical neurons. In: (Memo MA, ed). Cambridge: MIT.

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126-1141.

Lampl I, Ferster D, Poggio T, Riesenhuber M (2004) Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. J Neurophysiol 92:2704-2713.

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc of the IEEE 86:2278-2324.

Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A Opt Image Sci Vis 20:1434-1448.

Liu H, Agam Y, Madsen JR, Kreiman G (2009) Timing, timing, timing: Fast decoding of object information from intracranial field potentials in human visual cortex. Neuron 62:281-290.

Logothetis NK, Sheinberg DL (1996) Visual object recognition. Annual Review of Neuroscience 19:577-621.
Marr D (1982) Vision. San Francisco: Freeman publishers.
Marr D, Nishihara HK (1978) Representation and recognition of the spatial organization of three-dimensional shapes. Proc R Soc Lond B Biol Sci 200:269-294.
Maunsell JHR (1995) The brain's visual world: representatioin of visual targets in cerebral cortex. Science 270:764-769.
Meister M (1996) Multineuronal Codes in Retinal Signaling. PNAS 93:609-614.
Mel B (1997) SEEMORE: Combining color, shape and texture histogramming in a neurally inspired approach to visual object recognition. Neural Computation 9:777.
Meyers EM, Kreiman G (2011) Tutorial on Pattern Classification in Cell Recordings. In: Understanding visual population codes (Kriegeskorte N, Kreiman G, eds). Boston: MIT Press.
Mumford D (1992) On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol Cybern 66:241-251.
Mutch J, Lowe D (2006) Multiclass Object Recognition with Sparse, Localized Features. In: CVPR, pp 11-18. New York.
Myerson J, Miezin F, Allman J (1981) Binocular rivalry in macaque monkeys and humans: a comparative study in perception. Behavioral Analysis Letters 1:149-159.
Nielsen KJ, Logothetis NK, Rainer G (2006) Discrimination strategies of humans and rhesus monkeys for complex visual displays. Current Biology 16:814-820.
Olshausen BA, Anderson CH, Van Essen DC (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. Journal of Neuroscience 13:4700-4719.
Orban GA, Van Essen, D., Vanduffel, W. (2004) Comparative mapping of higher visual areas in monkeys and humans. Trends in Cognitive Sciences 8:315-324.
Pinto N, Doukhan D, DiCarlo JJ, Cox DD (2009) A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Comput Biol 5:e1000579.
Poggio T, Smale S (2003) The mathematics of learning: dealing with data. Notices of the AMS 50:537-544.
Potter M, Levy E (1969) Recognition memory for a rapid sequence of pictures. Journal of experimental psychology 81:10-15.
Rao RP (2004) Bayesian computation in recurrent neural circuits. Neural Comput 16:1-38.
Rao RP (2005) Bayesian inference and attentional modulation in the visual cortex. Neuroreport 16:1843-1848.
Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience 2:79-87.
Rao RPN, Olshausen BA, Lewicki MS, eds (2002) Probabilistic Models of the Brain: Perception and Neural Function. Cambridge: MIT Press.

Richmond B, Wurtz R, Sato T (1983) Visual responses in inferior temporal neurons in awake Rhesus monkey. Journal of Neurophysiology 50:1415-1432.
Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nature Neuroscience 2:1019-1025.
Riesenhuber M, Poggio T (2002) Neural mechanisms of object recognition. Current opinion in neurobiology 12:162-168.
Rolls E (1991) Neural organization of higher visual functions. Current opinion in neurobiology 1:274-278.
Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang S, Karpathy A, Khosla A, Bernstein M, Berg A, Fei-Fei L (2014) ImageNet Large Scale Visual Recognition Challenge. In: CVPR: arXiv:1409.0575, 2014.
Serre T, Wolf L, Poggio T (2005a) Object Recognition with Features Inspired by Visual Cortex. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). San Diego: IEEE Computer Society Press.
Serre T, Oliva A, Poggio T (2007a) Feedforward theories of visual cortex account for human performance in rapid categorization. PNAS 104:6424-6429.
Serre T, Kouh M, Cadieu C, Knoblich U, Kreiman G, Poggio T (2005b) A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. In, pp CBCL Paper #259/AI Memo #2005-2036. Boston: MIT.
Serre T, Kreiman G, Kouh M, Cadieu C, Knoblich U, Poggio T (2007b) A quantitative theory of immediate visual recognition. Progress In Brain Research 165C:33-56.
Sharon E, Galun M, Sharon D, Basri R, Brandt A (2006) Hierarchy and adaptivity in segmenting visual scenes. Nature 442:810-813.
Singer J, Kreiman G (2011) Introduction to Statistical Learning and Pattern Classification. In: Visual Population Codes (Kriegeskorte N, Kreiman G, eds). Boston: MIT Press.
Standing L (1973) Learning 10,000 pictures. Quarterly Journal of Experimental Psychology 25:207-222.
Sutherland NS (1968) Outlines of a theory of visual pattern recognition in animals and man. Proc R Soc Lond B Biol Sci 171:297-317.
Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks. In: International Conference on Learning Representations.
Tanaka K (1996) Inferotemporal cortex and object vision. Annual Review of Neuroscience 19:109-139.
Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. Nature 381:520-522.
Tsotsos J (1990) Analyzing Vision at the Complexity Level. Behavioral and Brain Sciences 13-3:423-445.
Ullman S (1996) High-Level Vision. Cambridge, MA: The MIT Press.
Virga A, Rockland, KS (1989) Terminal Arbors of Individual "Feedback" Axons Projecting from Area V2 to V1 in the Macaque Monkey: A Study Using Immunohistochemistry of Anterogradely Transported Phaseolus vulgaris-leucoagglutinin. The Journal of Comparative Neurology 285:54-72.

Vogels R, Biederman I, Bar M, Lorincz A (2001) Inferior temporal neurons show greater sensitivity to nonaccidental than to metric shape differences. J Cogn Neurosci 13:444-453.

Wallis G, Rolls ET (1997) Invariant face and object recognition in the visual system. PROGRESS IN NEUROBIOLOGY 51:167-194.

Winston P (1975) Learning structural descriptions from examples. In: The psychology of computer vision (Winston P, ed), pp 157-209. London: McGraw-Hill.

Yamane Y, Carlson ET, Bowman KC, Wang Z, Connor CE (2008) A neural code for three-dimensional object shape in macaque inferotemporal cortex. Nature neuroscience 11:1352-1360.

Yuille A, Kersten D (2006) Vision as Bayesian inference: analysis by synthesis? Trends Cogn Sci 10:301-308.