Invariance and Selectivity in the Ventral Visual Pathway

Stuart Geman
Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912
USA

Abstract

Pattern recognition systems that are invariant to shape, pose, lighting and texture are never sufficiently selective; they suffer a high rate of "false alarms". How are biological vision systems both invariant and selective? Specifically, how are proper arrangements of sub-patterns distinguished from the chance arrangements that defeat selectivity in artificial systems? The answer may lie in the nonlinear dynamics that characterize complex and other invariant cell types: these cells are *temporarily* more receptive to some inputs than to others (*functional connectivity*). One consequence is that pairs of such cells with overlapping receptive fields will possess a related property that might be termed *functional common input*. Functional common input would induce high correlation exactly when there is a match in the sub-patterns appearing in the overlapping receptive fields. These correlations, possibly expressed as a partial and highly local synchrony, would preserve the selectivity otherwise lost to invariance.
Keywords: correlation, synchrony, microcircuitry, nonlinearity, binding, vision

1 Introduction

Practical computer vision-systems answer practical questions: Is there a license plate in the image? What is the license plate number? Is there a defect in the chip geometry? How many faces are in the image? Who is in the image? Biological vision systems are less oriented towards a single question or set of questions and more oriented towards an ongoing process of image analysis. Indeed, real-world images have essentially infinite detail, which can be perceived only by a process that is itself ongoing and essentially infinite. The more you look, the more you see.

The implications of these remarks for biological vision systems are controversial. One extreme viewpoint is that, when faced with a complex image, brains construct an ever more elaborate data structure that simultaneously represents the richness of scene constituents and their inter-relationships. Scene analysis is the process of building something akin to a complex molecule whose atoms and bonds represent the multitude of constituents and relationships, possibly at a multitude of resolutions, that we perceive and reason about. This would be in the spirit of proposals by von der Malsbug [58] and Bienenstock [10, 11], and consistent with Grenander's proposition that patterns, in general, are best formulated as a relational composition of parts (Grenander [23], see also Fu [18]). At another extreme is the searchlight metaphor, whereby the

primary visual cortex serves as a kind of high-resolution buffer, and whereby image analysis is a process of selectively identifying parts in selected (attended) sub-regions. The process yields an annotated scene, "tree here, car there", through a highly directed search involving sequential and selective attention. This is more like models suggested by Treisman and Gelade [51], Marr [33], or Crick [13].

I propose to examine these fundamental biological questions from the perspective of the science of computer vision. This might appear misguided, given the evident shortcomings of artificial vision systems. But I would argue that the combination of great effort and modest progress in computer vision has in fact produced an important result: we know much more about what makes vision a hard problem than we did, say, twenty years ago. What exactly are the limitations of engineered vision systems? Where do they break down, and why? I contend that the basic limitations can be well articulated and that they lead to well focused questions that should be asked of biological vision systems.

As a preview, and as an introduction to the state of the art in computer vision, consider the practical problem of reading the identifying characters used to track wafers in semiconductor manufacturing. This is an example of the much-studied OCR (optical character recognition) problem. The highly automated semiconductor industry is the leading consumer of machine-vision products, with a broad range of applications where a computer equipped with a camera performs repetitive functions that are essential for a low-tolerance high-yield throughput. The OCR problem for wafer tracking is evidently difficult: many equipment manufacturers compete for a performance edge, yet the state of the art remains substantially short of human performance. This is despite best efforts to use neural networks, the latest developments in learning theory, or the latest techniques in pattern classification. Some years ago I worked on a team that developed a state-of-the-art reader for this application. Although the reader has been installed in over six thousand wafer-tracking machines, it is no exception to the rule that computer vision, even for constrained problems in controlled environments, is not yet competitive with human vision.

The difficulties begin with the patterned geometries that surround and often overlap with the identification markings (Figure 1), and they are compounded by other variables of presentation, including specularity and fluctuating contrast. Humans accommodate all of this effortlessly. In contrast, computer programs that can cope with the variability of the presentations of the characters suffer from "false alarms" (false de-

tections) between or overlapping the real characters, or in the structured backgrounds. Conversely, programs that are more selective (few or no false alarms) inevitably miss characters or make substitution errors. I propose that this *dilemma of invariance versus selectivity* is a central challenge in computer vision, and that unraveling the mechanism of its solution is a central challenge in understanding biological vision.

Figure 1: **OCR for wafer tracking.** Computer programs that accommodate the variability of character presentations are prone to false detections in the structured backgrounds.

Indeed, invariance is ubiquitous in biological vision systems. By computer vision standards, perception is astonishingly robust to coloring, texturing and contrast, as well as to pose parameters which define a nearly infinite-dimensional manifold for deformable objects. There is plenty of evidence for nearly invariant representations in the nervous system, starting with retinal circuits that perform nearly contrast-invariant calculations, through complex cells of V1 and V2 that exhibit some invariance to position and/or size, and into IT where cells with 10-, 20-, or even 30-degree receptive fields (cf. Sheinberg & Logothetis [46], Rolls & Tovee [42]) can be found to respond to an object presentation over a substantial range of poses and renderings. But then what of the *imprecision* that is the unavoidable companion to this invariance? What happens for example to the pose information? Can objects be correctly composed into

larger entities, at a later stage of processing, without taking into account their relative positions?

This dilemma of invariance versus selectivity is apparently related to the "binding problem": how does the nervous system signal the proper association of pieces to make a whole? Is it not the case that invariant representations, of evident value in and of themselves, nonetheless make poor building blocks? It is hard to imagine that the representations of complex, multi-part, deformable structures are not built out of invariant representations of their pieces. How then do we verify that the pieces are properly arranged? Perhaps relationships are themselves represented, explicitly. Relational units could then signal proper arrangements among constituents, but not unless there were either an unimaginable number of these or they too are invariant. The former is not biologically feasible, and the latter is not a solution, for how then is an invariant relational representation bound to an invariant representation of a constituent? We know from machine vision that the "backgrounds" of images and the contexts of objects are not anything like a noise process. They are instead made up of highly structured pieces that will conspire to mimic an object, be it as simple as a character or as complex as a face, if the pieces are allowed to come together without regard to sufficiently precise rules of relationship. How is it that biological vision systems are both invariant and selective?

I will argue that the answer lies in the microcircuitry of cortical neurons, which in essence serves as the local structure of a temporary but stable and globally configured representation. This point of view is in agreement with the molecule, but not the searchlight, metaphor. I will argue further that the proposed local dynamics are the natural, almost inevitable, consequence of the kinds of nonlinearities that are well-established and ubiquitous in neuronal processes of the ventral visual pathway.

2 Observations from Computer Vision

Most scientists entering the field of computer vision begin with an unrealistic optimism. What could be so hard? Build a detector for each entry in a library of objects of interest and annotate a scene: "This is here, that is there." In this section I will expand on the theme that much of the difficulty can be assigned to the competing requirements of invariance and selectivity, given the observation that what we call background is highly structured and, inconveniently, made up of much of the same stuff as the regions and

objects of real interest.

So far, machines cannot interpret images. There are practical successes, but these are characterized by specific goals for constrained scenarios, and are perhaps not in the direction of what we might call, loosely, image analysis. Just about nobody predicted that the problem would be this hard. To the contrary, as early as the 1960's, at which point one could imagine connecting a camera to a computer and writing software to interpret images, there was the feeling that a solution was within reach. Consider the 1966 "Summer Vision Project" of the MIT Artificial Intelligence Group (Papert [38]). The goal is succinctly captured in the abstract:

> *"The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of 'pattern recognition.'"*

One of the things that went wrong was with a key "sub-problem": segmentation. The idea was to define meaningful clusterings of pixels into regions and pieces, which could later be composed into meaningful objects. Nobody could have foreseen how hard this is. It is true that a big part of perception is in the determining of what goes with what, i.e. segmentation. But the rules of composition turn out to be subtle and circumstantial, depending on, among other things, lighting and texturing, the nature of the objects and surfaces that are being segmented, and the competing entities surrounding the area of interest. These rules have steadfastly resisted efforts to be systematically articulated, encapsulated, and turned into an if-then-style computer program.

The suspicion now is that the "knowledge engineering" approach, from the early days of AI, will not work; a vision system would require too much knowledge to articulate and organize. A more compelling approach, from a biological point of view, might be to design systems that *acquire* vision knowledge from examples. This is also appealing since, in a strict mathematical sense, there are statistical inference algorithms that can achieve provably *optimal* classification performance, given only a sequence of examples (e.g. raw images) and a corresponding sequence of correct classifications (e.g. "contains a face", or even "contains a face at such and such location"). This is the the-

ory of nonparametric inference (a.k.a. inductive inference or learning theory), whereby an *arbitrary* input/output relation is learned from examples (cf. Vapnik [56], Stone [48], Grenander [22], White [59], Geman et al. [21]). But convergence is asymptotic, which is to say that it takes effect as the sample size (number of image/output pairs) goes to infinity. So far, formulations of unconstrained vision problems in this manner have led to prohibitively slow convergence, in the sense of needing prohibitively large training sets in order to achieve interesting performance. As would be expected, this approach fares better on simplified problems involving isolated objects. An example is the problem of correctly classifying an image that consists only of one of the ten digits, handwritten. Artificial neural networks, implementing a nonparametric inference algorithm, and working from hundreds of thousands of examples, perform well on this task.

In fact there have been many successes at solving practical and important machine vision problems. Still, the state of the art for artificial vision systems solving generic problems on generic scenes (say, "find all plants and animals in the following pictures") is very limited. Nothing approaches human performance.

This gap between human and machine performance on vision tasks brings to mind Turing's demanding test for artificial intelligence ([53]): a human, when interacting with a system through an interface such as a keyboard or microphone, can not determine whether it is another human or, instead, a computer program that is behind the interface. A version of the Turing Test, in this case administered by a machine, has been put to practical use in recent efforts to defeat computer programs that surf the web, masquerading as humans in order to gain access to email accounts, opinion polls, credit card numbers, etc. Increasingly, vulnerable websites are protected by so-called Human Interactive Proofs (HIP), which require the visitor to perform actions that are presumed to be peculiarly human. These actions are good "Turing Tests", in that current computer programs fail them. It is interesting that these proofs generally take the form of a vision task, such as identifying a string of characters embedded in a textured and structured background, or positioning a mouse on a face embedded in a field of face-like substructures (see for example Rui and Liu [43]). A pseudo-random number generator and a clever synthesis algorithm assure that the task is different with every visit. These HIP's work exactly because a computer program that might possess sufficient invariance to the variety of faces or characters will be defeated by a multitude of false targets in the structured backgrounds.

The dilemma of invariance versus selectivity is further illustrated, more concretely, by examining the mechanisms of some state-of-the-art vision systems. The hierarchical system proposed by Riesenhuber and Poggio ([40], [41], see also Tarr [49]) recognizes rigid and partially deformable objects with a high degree of pose invariance, and makes connections to specific biological structures and functions in the ventral visual pathway. It is based on an architecture reminiscent of Fukushima's Neocognitron [19], involving a feed-forward process through layers of units, with successively more invariance emerging from one layer to the next. In the Riesenhuber-Poggio model, invariance comes from the so-called MAX filter, whose output is the *maximum* of many precursor filters, each of which is typically (but not necessarily) linear. To see the connection to invariance, imagine that the precursor units represent simple matched filters for an object or object part, or even just a local edge element, at each of many translations over a limited receptive field (RF) area. The MAX filter will respond to the presentation of the object (or part or edge) with invariance to translation over the chosen receptive field. Obvious extensions can accommodate scale, rotation, and other pose parameters. The layers of the Riesenhuber-Poggio model alternate between being made up of units that compute MAX filters and units that compute linear filters, with each unit of each layer responding to a "receptive field" of outputs from previous layers. At the highest levels, units are responding invariantly to parts and objects.

Amit and (Donald) Geman ([6], [7]) also use MAX filters as a starting point for pose invariance. Additionally, their algorithms employ various mechanisms for lighting invariance, which is of course important for most practical applications. The work of Amit and Geman is motivated less by biological considerations and more by practical ones: what are the *computationally* most efficient strategies for detecting instances of (possibly deformable) objects within an unconstrained scene? The solution to this optimality problem can be approximated by a sequence of image-based calculations that depend on the results of previous calculations in such a way as to maximally reduce uncertainty. Thus the approach is sequential and not parallel. The image-based calculations are exactly of the MAX-filter type: Is there an end-stop in this vicinity? Is there a light-to-dark discontinuity in that vicinity? Many such queries can be performed in a very small amount of time. The algorithm produces a kind of map of queries, akin to a sequence of saccades, in which very little time is spent in the background and a great deal more time is spent on and around target objects.

Both algorithms can be tuned to detect essentially all instances of the target object.

This is their strength: invariance to pose, shape and other variables of presentation. On the other hand, as might be expected from our discussion of the tradeoffs between invariance and selectivity, the price for invariance is a loss of selectivity—multiple "false alarms" in structured backgrounds. The authors, of course, are well aware of the tradeoff. In D. Geman's view ([20]), a more computationally intense context-based algorithm, possibly cued by an earlier detection phase, would be needed to achieve high selectivity. Riesenhuber and Poggio [40], on the other hand, propose that selectivity might be achieved through a large repertoire of features and feature combinations. In essence, the proposal is that the coincidence of many features in an area of an image would be more-or-less diagnostic for a particular object. Mel [34] argues for a similar mechanism. In any case, the fact remains that existing object recognition systems based upon invariant features, or even feature conjunctions, are far less selective than their biological counterparts.

Where, exactly, is selectivity lost? A simple arrangement of two MAX filters already illustrates the problem. Suppose that each of these filters is tuned to detect a portion of a vertical straight line. As observed earlier, the MAX operation affords a degree of invariance whereby, in this example, horizontal shifts of a line within one of the RFs will have little effect on the corresponding output. Imagine further that the RFs are lined up vertically, so that a single extended vertical line passing through both RFs will produce a strong response from both filters; see Figure 2. Due to the invariance of the individual filters, their joint activation can be taken as evidence for the presence of an extended vertical line within a range of positions. Lines are parts of many things, and invariant line detectors therefore make useful components of a recognition system. The problem with this line detector is that it lacks selectivity. Two fragments of vertical lines, not necessarily collinear, will produce an output that is indistinguishable from an extended vertical line. This spurious response will happen often, e.g. in textures, across neighboring characters, and from other unexpected coincidences.

As pointed out earlier, the problem is generic: conjunctions of invariant representations of parts make for sloppy detectors of compositions. Some might argue that the problem is even more fundamental, having to do with the very nature of an interpretation of a visual scene. Many cognitive scientists hold that perception (indeed, all of cognition) is more a matter of building an elaborate hierarchical structure of relational compositions than one of merely labeling a blackboard or frame buffer with a set of identifications (cf. Fodor and Pylyshyn [15]). From this point of view, our

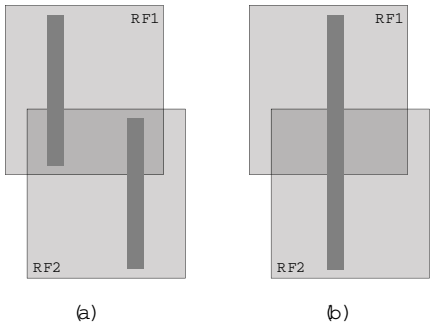

Figure 2: **Invariance vs. Selectivity.** Imagine two cells N1 and N2, with receptive fields RF1 and RF2, which signal vertical bars with invariance to horizontal position ("phase"). The situations in panels (a) and (b) are indistinguishable, given only the individual activity levels of N1 and N2.

dilemma of invariance versus selectivity is part of a larger problem of focusing mostly on objects, and not explicitly and equally on relationships. It is certainly true that the machine-vision community has focused largely on objects, *per se*.

3 Biological Vision: Functional Connectivity and Functional Common Input

MAX filters and other similar nonlinear combinations of linear filters have been proposed as models of the complex cells of V1 and V2. This connection suggests a biologically plausible solution to the invariance/selectivity dilemma. To illustrate this in the manner of a thought experiment, construct a MAX filter by starting with a single linear filter and then reproducing the linear filter at many locations (shifts). The MAX-filter output is then the maximum of the outputs of the linear filters, and its RF is the union of the RFs of the shifted linear filters:

$$y = \max_{\lambda \in \mathcal{L}} (\vec{a}^{\lambda} \cdot \vec{x})$$

where $\vec{x} = (x_1, \ldots, x_n)$ is the input to the receptive field (e.g. a vector of pixel intensities, or a vector of outputs from presynaptic neurons), y is the MAX filter response, and to each location $\lambda \in \mathcal{L}$ there corresponds a linear filter at λ given by $\vec{a}^{\lambda} = (a_1^{\lambda}, \ldots, a_n^{\lambda})$.

Consider now the *sensitivity* of the output, y, to the input from a particular pixel, i, in the RF. One natural measure of this sensitivity is the *derivative* of the MAX filter

output with respect to the input intensity, x_i, at pixel i. If the MAX filter were in fact a *linear filter*, then this derivative would be a constant, independent of the particular image visible across the RF. But the MAX operation is nonlinear and therefore the derivative (sensitivity) is a function of the image. What is the nature of this function? For a given input $\vec{x}$, denote by $\lambda^*(\vec{x})$ the location of the linear filter that achieves the maximum

$$\vec{a}^{\lambda^*(\vec{x})} \cdot \vec{x} = \max_{\lambda \in \mathcal{L}} (\vec{a}^{\lambda} \cdot \vec{x})$$

Notice that $\lambda^*(\vec{x})$ is piecewise constant, changing only at inputs $\vec{x}$ that produce ties. If we stay away from ties, then $\lambda^*(\vec{x})$ is constant and

$$\frac{\partial y}{\partial x_i} = a_i^{\lambda^*(\vec{x})},$$

which is zero if pixel i is not in the RF (support) of the filter located at $\lambda^*(\vec{x})$.[1] The MAX filter is thereby *insensitive* to intensity changes occuring outside of the RF of the maximally responding linear filter. In other words, there is a strong sense in which the MAX filter *commits* to a particular subset of inputs in its RF, and this commitment is stimulus-dependent. The *functional* connectivity changes with the stimulus.

Functional connectivity may appear at first to be a somewhat exotic phenomenon, resulting from the use of the MAX operation. But consider that *any* nonlinearity produces a stimulus-dependent derivative. Admittedly, idealizations like the MAX filter yield a particularly clean and easily interpreted form of this behavior, but the fact remains that a neuron is likely to be "listening" more intently to some of its inputs than others, and the distinguished inputs are likely to depend globally on the particular vector of pre-synaptic signals. Indeed, the idea of Hubel and Wiesel [27] of modeling a complex cell as a (nonlinear) pooling of simple cell outputs already points to this kind of behavior. This is apparent, for example, in the "energy models" of Adelson and Bergen [3]. An instance of these models, studied by Sakai and Tanaka [44], is the sum-of-squares model in which the complex cell output, as measured by firing rate, is the sum of squared outputs of two linear filters:

$$y = (\vec{a}^{\lambda} \cdot \vec{x})^2 + (\vec{a}^{\delta} \cdot \vec{x})^2$$

The filters $\vec{a}^{\lambda}$ and $\vec{a}^{\delta}$ might represent, for example, the same Gabor filter centered at two spatial locations, one a shift of the other. And the dot products, $\vec{a}^{\lambda} \cdot \vec{x}$ and $\vec{a}^{\delta} \cdot \vec{x}$,

[1] The derivative will typically be undefined at exact ties.

might represent the outputs of two classical simple cells, modeled as simple linear filters. Suppose that $\vec{a}^{\lambda}$ and $\vec{a}^{\delta}$ are localized with nearly non-overlapping supports. Then a given pixel i in the receptive field of the complex cell is likely to be represented in one of the filters (e.g. $|a_i^{\lambda}| > 0$) and not the other ($|a_i^{\delta}| \approx 0$). Now look at the influence of x_i on y:

$$\frac{\partial y}{\partial x_i} = 2a_i^{\lambda}(\vec{a}^{\lambda} \cdot \vec{x}) + 2a_i^{\delta}(\vec{a}^{\delta} \cdot \vec{x}) \approx 2a_i^{\lambda}(\vec{a}^{\lambda} \cdot \vec{x}) \tag{1}$$

Thus the functional connectivity between the model complex cell and the activity at pixel i is strong exactly when $\vec{x}$ matches the filter λ. The modeled cell selectively and circumstantially attends to a subset of its inputs, in much the same way as the MAX filter.

Most models of complex cells, and no doubt all biological complex cells, involve important *temporal* effects as well as nonlinearities that are not so neatly captured by a maximum operation or a sum of squares. But it is nevertheless the nature of invariant receptive field properties, by virtue of their nonlinearity, that the sensitivity of the cell output to the activity of a given input is a global function of the input pattern. What is more, as is illustrated in these simple examples, it is reasonable to conjecture that such cells focus their attention on a subset of their inputs that corresponds to a region in space containing a maximally stimulating pattern. Without much fuss, more sophisticated models of complex cell properties, such as "Nonlinear-Nonliner-Poisson" (NNP) models (Harrison et al. [25]), or the quadratic forms that arise from "Slow Feature Analysis" (Wiskott & Sejnowski [60]), can be tuned to just this kind of behavior.

It is perhaps not a stretch to suggest that in fact we should *expect* a focus of sensitivity, in exactly the manner demonstrated in these models, from any cell that we would think of as invariant for a target within an extended receptive field. After all, what do we mean by "invariant" if not, exactly, that the cell responds to a preferred pattern independently of certain pose parameters, and *indifferently to the details of structure and noise elsewhere in the receptive field*? The picture is complicated by contextual effects, saturation, extended ("non-classical") receptive fields (Vinje & Gallant [57]), and an important and subtle time course to the building of a representation (Lee et al. [30]). But the earlier responses, presumably reflecting more feed-forward than feed-back computation, *should* be largely a function of the target and not its background.

If this were the case, if functional connectivity were to operate in something of this manner, then it would provide a compelling biological mechanism for maintaining selectivity in a hierarchy of invariant cell types. To make the connection to selectivity, return to our thought experiment involving two complex cells, idealized as MAX filters, each tuned to vertical line segments, and with overlapping RFs situated one above the other as depicted in Figure 2. *Anatomically*, these filters have common inputs, but *functionally* the extent of common input depends on the particulars of the stimulus—see Figure 3. The size of the population of inputs that is *functionally* connected to *both* MAX filters is circumstantial. What circumstances promote common input? MAX filters functionally commit to the RFs of the particular linear filters that produce maximum outputs. The functional common input is therefore proportional to the overlap of the RFs of these selected linear filters, and in general these will overlap to the extent that they represent parts of the same structure. *Functional common input is promoted by the two vertical line segments being part of a single extended line.*

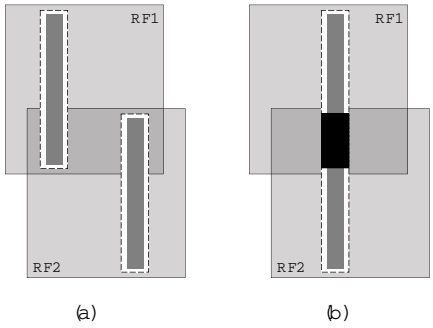

Figure 3: **Functional Common Input.** Same as previous Figure, with functional connectivity outlined by broken lines. *Anatomically*, common input is fixed, and proportional to the area of the intersection of the RFs. *Functionally*, common input depends on the stimulus. Area in black represents the functional common input, which is zero in (a) and maximized in (b) by an extended line segment.

This observation, that common input is likely to be signal-dependent and therefore not just an anatomical property, is not special to the MAX-filter model of complex cells, or for that matter to complex cells *per se*. Indeed, I have already pointed out that the *definition* of invariance is suggestive of a functional and local connectivity. This would mean that common input is circumstantial, rather than strictly anatomical, and it would follow that two invariant cells, with common anatomical input, would be functionally connected to many of the same pre-synaptic influences exactly when their

respective targets "fit together", or "agree", at the intersection of the respective RFs. In short, it is reasonable to speculate that the amount of functional common input to invariant cells with overlapping RFs is circumstantial, and in fact maximized under the particular circumstance that these cells are separately signaling patterns which are part of a single larger structure.[2]

Is circumstantial common input a readable and usable variable? Indeed, could it be that this variable, the *functional common input*, is readable and used to maintain the very selectivity that appeared lost to invariance?

4 Illustration

It goes without saying that relationships among parts or features can help in distinguishing one object from another, or an object from background. But how are relationships represented in neural systems? I have proposed a role for functional common input (fci), arising from the overlapping receptive fields of cells, or cell ensembles, that are tuned with some invariance to particular features, parts, or objects. A simple computer experiment, explored in this section, was devised to illustrate these ideas. The next section (§5) takes up the issue of "read out": how would fci be manifested in neural dynamics, and to what effect?

In brief, in these experiments two position-invariant model cells ("complex cells") of the maximum-filter type were constructed, one which responds to right eyes, and one which responds to left eyes. Multiple copies of these "complex cells" were situated in such a way that their collective receptive fields cover an entire image. Many left-right pairs of these position-invariant cells have overlapping receptive fields, as do the static ("simple cell") filters that feed them. Using an analytic measure of fci (see below), it was shown that, among left-right pairs of "complex cells" with high activities, those that have high fci tend to represent faces, whereas those with low fci tend to be "false alarms". In the presence of high activity in pairs of simulated "complex cells" with overlapping receptive fields, functional common input is evidence for a correct

[2]I am using the notion of a receptive field loosely and with some intended ambiguity. We can think of functional connectivity as establishing a temporary focus of sensitivity to a localized subset of the visual field, but what I really have in mind is a focus of sensitivity to a subset of feed-forward presynaptic neurons. Presumably, the two interpretations are related, and in fact one might argue that the former—commitment to a visual area—would result from a cascade of the latter—commitment to selected presynaptic neurons.

arrangement of parts.

Specifically, the experiments were performed on the photograph shown in the upper-left panel of Figure 4. Two filters were generated from the pixel data on and around the right and left eyes, respectively, of one of the women in the photograph (upper-right panel). Pixels along the bridge of the nose contribute to both filters. Let $\vec{\tilde{r}} = (\tilde{r}_1, \ldots, \tilde{r}_n)$ be the pixel gray levels in the chosen rectangle surrounding the woman's right eye, and let $\vec{\tilde{l}} = (\tilde{l}_1, \ldots, \tilde{l}_m)$ be the pixel gray levels in the chosen rectangle surrounding her left eye. The heavily lined rectangle surrounding the right eye has 20 rows and 35 columns, so $n = 700$. The lightly lined rectangle surrounding the left eye has 20 rows and 26 columns, so $m = 520$. The utility of the filters are improved by normalization: Let $\vec{r} = (r_1, \ldots, r_n)$ and $\vec{l} = (l_1, \ldots, l_m)$ be, respectively, the normalized versions of $\vec{\tilde{r}}$ and $\vec{\tilde{l}}$ ($\frac{1}{n}\sum r_k = \frac{1}{m}\sum l_k = 0$ and $\frac{1}{n}\sum r_k^2 = \frac{1}{m}\sum l_k^2 = 1$).

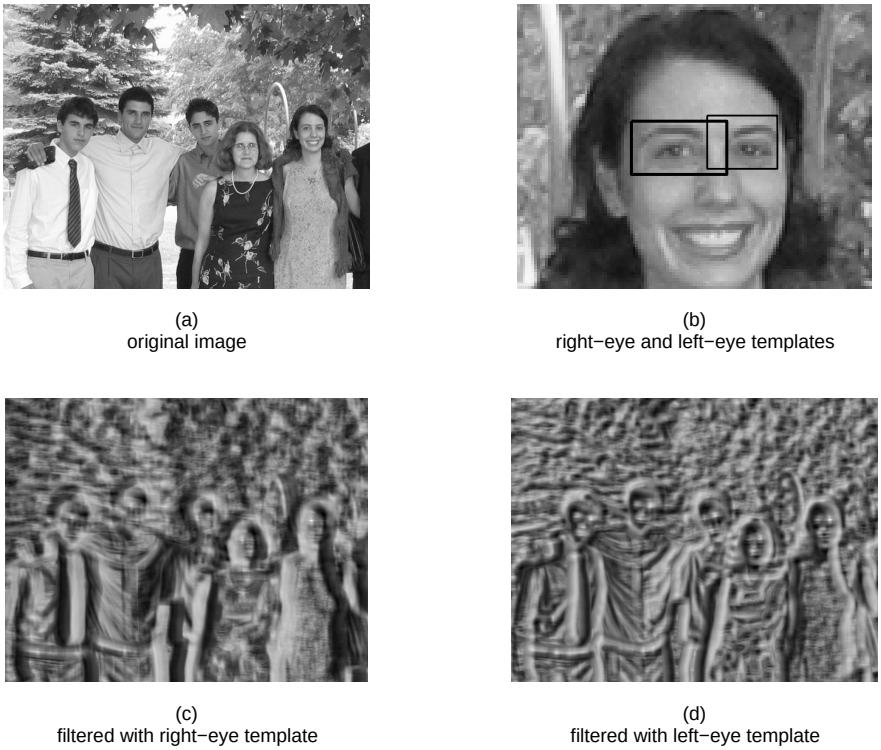

Figure 4: **Left-eye and Right-eye Filters.** (a) Picture used in experiments (615×795 pixels). (b) Left-eye (20×26) and right-eye (20×35) templates chosen from one of the faces in the picture. (c) Correlation of right-eye template at every location of the original image. (d) Correlation of left-eye template at every location of the original image.

Either filter can be applied at any location in the image. Let λ designate a location. Imagine situating the heavily lined (right-eye) rectangle with its upper-left corner at λ, and let $\vec{\tilde{x}}^{\lambda} = (\tilde{x}_1^{\lambda}, \ldots, \tilde{x}_n^{\lambda})$ be the gray-level image values within the rectangle that correspond to the right-eye filter values $r_1, \ldots, r_n$.[3] Here again it is better to work with the normalized data, $\vec{x}^{\lambda} = (x_1^{\lambda}, \ldots, x_n^{\lambda})$, having mean zero and mean square one. The right-eye filter response at λ (call it y^{λ}) is just the correlation coefficient between $\vec{r}$ and $\vec{x}^{\lambda}$:

$$y^{\lambda}(\vec{x}^{\lambda}) = \frac{1}{n}\vec{r} \cdot \vec{x} = \frac{1}{n}\sum_{k=1}^{n} r_k x_k \tag{2}$$

which is between -1 and 1, on account of $\vec{r}$ and $\vec{x}$ being normalized. The bottom-left panel in Figure 4 shows the filter response at every location in the image. The analogous display, for the left-eye filter, is in the bottom-right panel.

As discussed already in §2, a position-invariant right-eye filter (a highly idealized "complex cell") can be constructed by simply taking the maximum filter value over a designated "receptive field". (There are other methods, possibly more realistic in terms of neural hardware, for nonlinearly combining filter values and accomplishing basically the same thing.) In Figure 5, the large square containing part of the face of the center person encloses a 70×70 image region. Thinking of this as a receptive field (RF), we can define a model position-invariant cell (which we will call, loosely, a complex cell) by maximizing the response over all right-eye filters with supports contained entirely in this RF. The actual maximum is achieved at the heavily lined rectangle surrounding the right eye. An example of a complex cell tuned to left eyes is illustrated in the same figure. The large square containing part of the woman's face also surrounds a 70×70 "receptive field", with the maximizing left-eye filter indicated by the lightly lined rectangle.

Consider now pairs of these right-eye/left-eye complex cells, with receptive fields overlapping and situated side by side: the 70×70 RF of the right-eye cell on the left, and the 70×70 RF of the left-eye cell on the right. The *union* of the two receptive fields forms a rectangle that is 70 by 110 pixels—the RF of the right-eye cell occupies the first 70 columns, the RF of the left-eye cell occupies the last 70 columns, and thirty columns are common to both RFs. The large rectangle in Figure 5, on the face of the second person from the left, is an example, and the locations of the maximizing

[3] In §3, it was convenient to "fix $\vec{x}$" and "move the filter" by λ—hence the notation $\vec{a}^{\lambda}$; here it is more convenient to fix the filter, $\vec{r}$ or $\vec{l}$, and "move the image" by λ—hence the notation $\vec{x}^{\lambda}$.

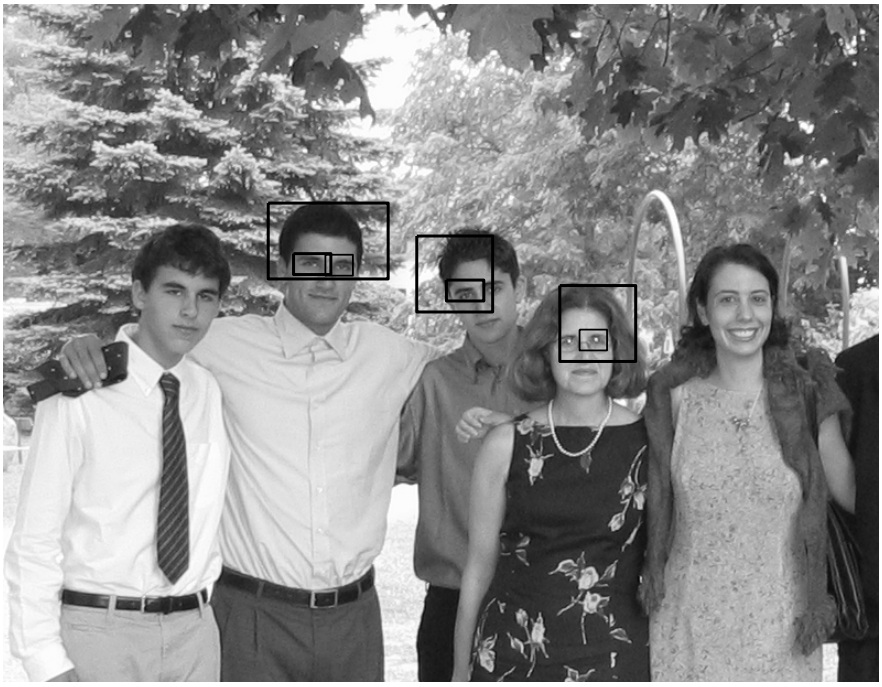

Figure 5: **"Complex cells" and complex cells with overlapping receptive fields.** Position-invariant left-eye and right-eye detectors are built by maximizing filter response over square "receptive fields". Smaller, lightly lined rectangle shows the position of the maximum left-eye filter response within the square covering part of the women's face (second from right). Smaller, heavily lined rectangle shows the position of the maximum right-eye filter response within the square covering part of the man's face (middle). Large rectangle over part of the man's face (second from left) outlines *combined* and *overlapping* right-eye and left-eye receptive fields, and shows the positions of the maximum right-eye and left-eye filter responses.

right-eye and left-eye filters are indicated by the smaller heavily lined and lightly lined rectangles, respectively. The model complex cells are responding to the man's right and left eyes.

Each panel in Figure 6 is an example of a combined RF of the type shown in Figure 5. Thus the rectangles are each 70×110 pixels, and include to the left and right the 70×70 pixel receptive fields of complex cells tuned to right and left eyes respectively. An ensemble of these combined RFs (1702 in all) was obtained by situating the upper-left corner of a 70×110 pixel rectangle at all row addresses that are multiples of 15 and all column addresses that are multiples of 15. Each panel in Figure 6 shows the

positions of the maximizing filters for the model complex cells—heavily lined rectangles for the right-eye filters, and lightly lined rectangles for the left-eye filters.

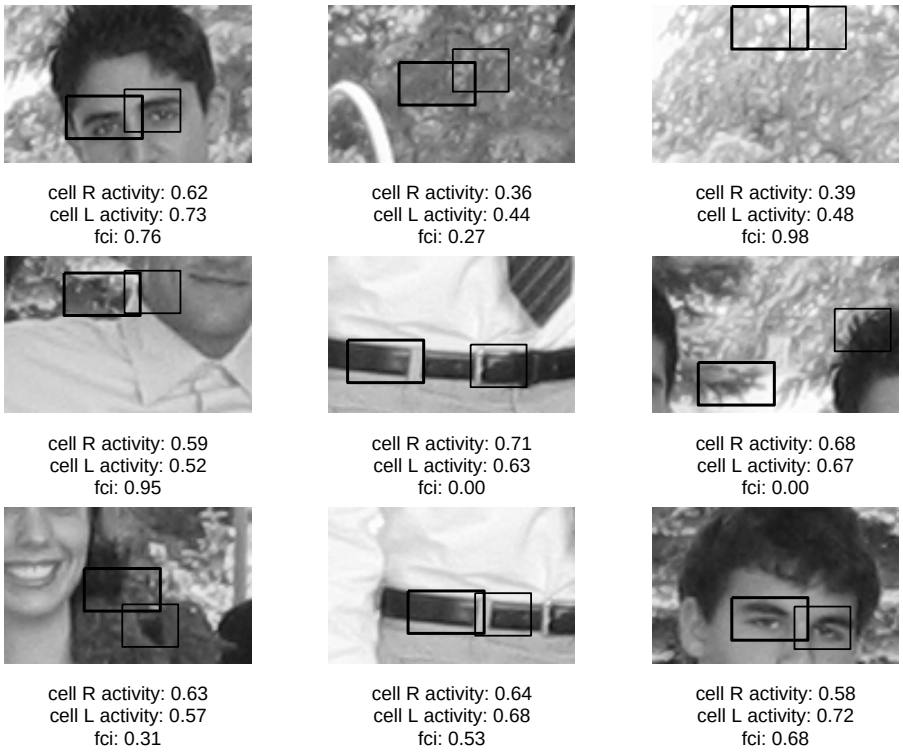

Figure 6: **Examples of combined receptive fields.** Each panel from overlapping right-eye and left-eye receptive fields. Heavily lined rectangle is the position of the maximizing right-eye filter (with correlation labeled "cell R activity") and lightly lined rectangle is the position of the maximizing left-eye filter (with correlation labeled "cell L activity"). Agreement in the region of overlap of the maximizing filters is partly captured by the functional common input ("fci"), as described in text.

Underneath each panel are the corresponding response levels of the two complex cells—"cell R" refers to the cell that is position invariant to right eyes, and "cell L" refers to the cell that is position invariant to left eyes. In each case, the response ("activity") is just the output of the maximizing filter. And "fci" is a measure of functional common input, derived as follows.

Fix a right-eye type complex cell, and fix $\lambda = \lambda^*$, the location of the maximum right-eye filter response within the receptive field. Consider the functional connectivity of this model cell to the input associated with location i in the cell's RF. If i is not in the

support of the right-eye filter located at λ^*, then by virtue of the maximum operation the cell is functionally disconnected from the input: the output of this idealized complex cell is unchanged by changes at i. If, on the other hand, i is in the support of the right-eye filter located at λ^*, then there is a contribution of

$$r_\alpha x_\alpha^{\lambda^*}$$

to the overall cell activity, where $\alpha = \alpha(\lambda^*, i)$ is the coordinate of the filter ($\vec{r}$) and image ($\vec{x}^{\lambda^*}$) vectors corresponding to location i. One measure, then, of functional connectivity to i would be r_α, the multiplier of the normalized activity at i.

Define the functional connectivity of a left-eye type complex cell to input associated with location i to be, by analogy, zero if i is not in the support of the maximizing left-eye filter, and l_β if, otherwise, i is in the support, where $\beta = \beta(\gamma^*, i)$ and γ^* is the location of the maximum filter response. The *product* of these connectivity measures (namely

$$r_{\alpha(\lambda^*,i)} l_{\beta(\gamma^*,i)}$$

if i is in the support of both maximizing filters, and zero otherwise) is then a measure of the degree to which input from location i is functionally common to both cells. Whether both coefficients are large and positive (common excitatory influence) or large and negative (common inhibitory influence), activity at location i can be expected to promote statistical dependence between the activities of the two cells. Finally the *sum* over i within the combined RF of these products is a measure of functional common input, and is labeled "fci" in Figure 6. This amounts to the inner product, over the intersection of the left- and right-eye maximizing filters, of the normalized filter coefficients. For ease of interpretation, the value is rescaled so that the maximum fci, over all combined RFs, is one.

By this construction, fci is a measure of both overlap *and agreement* between the maximizing left-eye and right-eye filters. Compare the second panel (top row, middle column) with the fourth panel (second row, first column), in Figure 6. Functional common input is substantially larger in the fourth panel than in the second panel (.95 versus .27, respectively), but the area of overlap is essentially the same (152 versus 154 pixels, respectively). Evidently, in the second panel the left- and right-eye filters are not well matched within their region of intersection, and evidently the quality of a match varies with the spatial relationship between the maximizing filters. In the framework of our model cells, the *sensitivities* of the left- and right-eye "complex cells," to the inputs

representing their combined receptive fields, do not match up for this particular image - there is very little *functional* common input. *Anatomically*, there is a constant and substantial common input, by virtue of the 70×30 pixels common to the overlapping 70×70 pixel receptive fields.

In a system designed to detect pairs of eyes belonging to a face, it is to be expected that fci will improve selectivity—many combined RFs with strong responses to both left-eye and right-eye filters violate the spatial relationship between the maximizing filters that would be characteristic of a face. In these cases fci is likely to be small or zero: either the filters are not well matched within their intersections, or there is no intersection at all. One way to demonstrate the utility of fci is to compare performance of a simple face detection system based, on the one hand, solely on the activities of cells R and L, to one which is based, on the other hand, both on cell activities *and* fci. Consider, for example, the two "Receiver Operating Characteristic" (ROC) curves for face detection drawn in Figure 7. The dashed-line curve is derived from the combined complex-cell activities, as measured by the product of the individual cell activities. A face detection system is built from this combined activity by introducing a threshold, T: when the combined activity is greater than T, we declare a face in the combined RF. Otherwise, we declare no face. The ROC curve is the relationship between the detection probability ($Pr\{\text{face detected}|\text{face in combined RF}\}$) and the false alarm probability ($Pr\{\text{face detected}|\text{no face in combined RF}\}$), derived by varying the threshold from $-\infty$ (at which point every RF is declared to have a face) to $+\infty$ (at which point no RF is declared to have a face). The probabilities are estimated, empirically, by running the detection algorithm on an ensemble of combined RFs, in this case the 1702 RFs obtained by stepping through the image in increments of 15 rows and 15 columns.

Thresholding, instead, on the three-way product of activity in cell R, activity in cell L, and fci[4] produces the solid-line ROC curve in Figure 7. Evidently, and not surprisingly, at any given detection probability the false alarm rate is dramatically reduced by taking fci into account, as a proxy for relative positioning. (And bear in mind that the overall false alarm rate is particularly sensitive to the false alarm probability, since by far most RFs do not contain a face.) Indeed, fci *alone* gives comparable performance to activity alone, as can be demonstrated by building an ROC curve by thresholding directly on fci (Figure 7, dotted line).

[4]Products, rather than sums, produce ROC curves that are unaffected by arbitrary scaling of the activity and fci variables.

There is perhaps a danger that the simplicity of these experiments hides the complexity of the vision task. Could a hierarchy of part-whole compositions of the type built here form the basis for a state-of-the-art vision system? Not exactly, or at least not without a great deal more invariance at the level of feature detection, a mechanism for reversing incorrect local matches ("top-down processing"), and a mechanism for resolving competing global interpretations. The intention, instead, is to illustrate the importance of relational information, and the possibility that relational information is captured, in part, by what would appear to be a meaningful physiological parameter—functional common input.

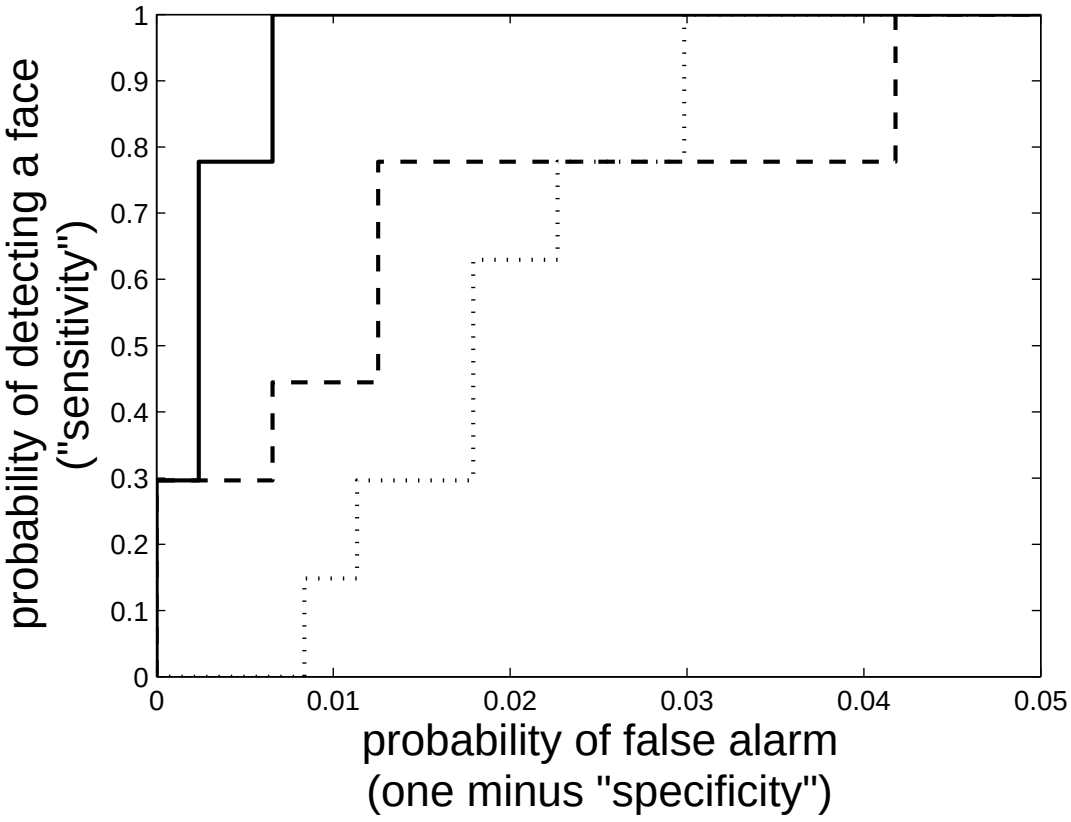

Figure 7: **Receiver Operating Characteristic for three face-detection systems.** Performance based on activity levels from pairs of invariant right-eye-detecting and left-eye-detecting units (dashed line), compared with performance based on activity levels *and* functional common input (solid line). The dotted line depicts performance based solely on functional common input.

5 Partial Synchrony

Although the role, and to an extent the very existence, of fine-temporal structure in cortical activity is controversial, most physiologists would agree that, everything else being equal, i.e., fixing the *firing rates* of two cells, the extent of synchronous activity between these two cells will reflect the extent to which their presynaptic populations overlap. In short, common input promotes synchrony. And most physiologists would also agree that, as a variable, synchrony is "readable" in that cells that fire synchronously are more likely to promote super-threshold activity in a common postsynaptic cell—"coincidence detection", or what Abeles [2] termed, more precisely, "coincidence advantage". (See Azouz & Gray [8], Mikula & Niebur [35], and Tsodyks & Markram [52] for models and experimental results that lead to this conclusion, at least within certain ranges of parameters.) Indeed, precisely these effects have already been demonstrated in vivo, for instance in retinal/LGN/striate cortical circuitry: More retinal common input promotes more fine-temporal thalamic synchrony (Dan et al. [14]); more thalamic synchrony promotes more activity in postsynaptic cells of the striate cortex (Alonso et al. [4]). And within the sub-columnar microcircuitry of cortex, there are tight divergent/convergent anatomical loops, which apparently produce precisely timed postsynaptic events in layer 2/3 (see Yoshimura et al. [61]).

What sorts of *global* data structures are consistent with the local mechanisms of functional common input, synchrony, and coincidence detection? Presumably, a cell representing a piece of a larger whole would participate simultaneously in many of these local circuits, by virtue of having different, and in some cases even disjoint, functional common inputs with each of many other cells. A curve, for example, would induce activities in a collection of complex cells, and these activities would reflect the topology of the curve through a succession of *local* dependencies (partially synchronous activities) that define neighbors in a *one-dimensional* representation—see Figure 8. Next-layer recipient cells, activated by local synchronous activities, would inherit and preserve this topology by the very same mechanism of functional common input. This kind of neighborhood relationship is not transitive. Depending upon the stimulus, cells 'A' and 'B' might share a large population of inputs, as might cells 'B' and 'C', yet cells 'A' and 'C' might have little or no common input. This creates a *temporary* topology: 'A' and 'B' are neighbors; 'B' and 'C' are neighbors; 'A' and 'C' are *not* neighbors.

More generally, the picture that emerges is one of a temporary topological repre-

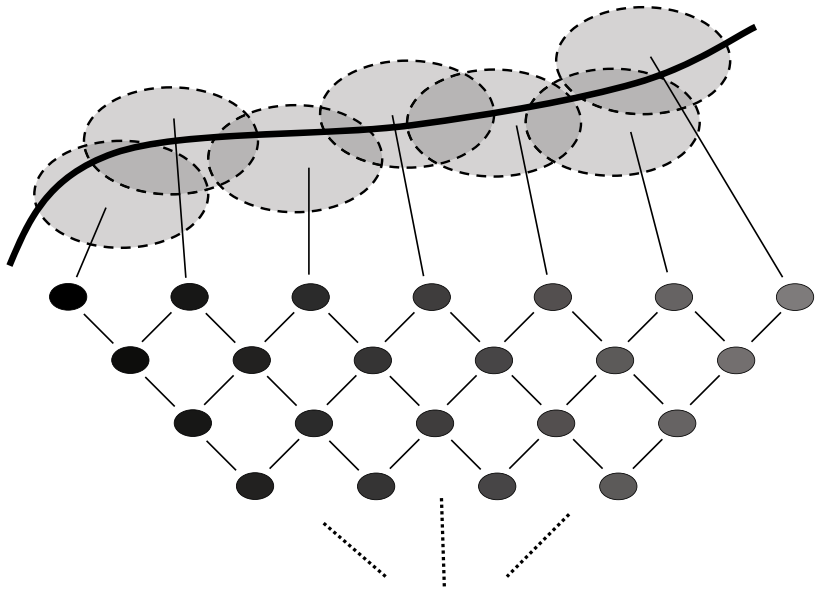

Figure 8: **Topological Representation of a Curve.** Cartoon version of a temporary arrangement of neurons and neuronal connections, held together by bottom-up and top-down loops of functional common input, partial synchrony, and coincidence detection. The segments of the curve passing through intersections of the layer 1 RFs (intersections of the dotted ovals) represent functional common inputs to the layer 1 cells. More generally, the proposal is that perception amounts to a construction of a temporary topological subgraph of a much richer and more permanent graph of anatomical connections. Presumably, there are important lateral components as well (not shown), perhaps tying together the representations within each level through similar loops of diverging and converging temporary connections.

sentation, glued together by diverging and converging circumstantial connectivity patterns, something like the correlation structure proposed by von der Malsburg [58] or the braid structure proposed by Bienenstock [10]: *divergence* from inputs that are functionally common to a collection of two or more cells rendered partially synchronous by virtue of these shared inputs; *convergence* from a collection of partially synchronous presynaptic cells to postsynaptic cells sensitive to well-timed cooperative inputs. Abeles [1] seems to have been among the first to recognize the potential computational advantage of these loops of diverging and converging connections. (See also Izhikevich et al. [28] for a recent computational theory based upon a similar principle). The picture that I have in mind, however, is somewhat different from that of the cell assembly that operates in phase with a more-or-less global rhythm, or by a lock-step synchronous activity across an entire sub-population (as discussed for example by Abeles [1, 2],

Bienenstock [10], Crick [13], and Shastri & Ajjanagadde [45]). The topology being proposed here is local and hierarchical, defined at any one level by overlapping pieces and across levels by part-whole relationships.

Of course the idea that the brain builds an accurate interpretation out of purely feed-forward computations is too simple. Vision engineers know that accurate segmentation, for example, is impossible without top-down calculations representing a feed-back of contextual information and prior expectation. And neuro-anatomists know that most visual areas receive more fibers from feed-back connections than they send through feed-forward connections. But it is possible that the same dynamical principles apply to the feed-back pathways, and that a particular pattern of activity is selected and reinforced by virtue of highly circumstantial, functional, common inputs from the *top down*. The picture of a temporary connectivity made up of divergent/convergent loops would be the same whether viewed from the bottom up or the top down. This opens the door to much the same mechanisms of functional common input, synchrony, and coincidence detection in the feed-back pathway (Figure 8). Perhaps the two, the feed-forward and feed-back pathways, work towards a consensus, whereby expectation and context are resolved with otherwise ambiguous sensory signals. Carpenter & Grossberg [12], Mumford [36, 37], and Ullman [54] have proposed similar interpretations of feed-forward/feed-back dynamics.

These ideas bring us back to the molecule metaphor, in which visual perception is akin to a process of building an elaborate and extended representation out of neuronal activities, with an image-dependent topology that is realized through explicit and temporary connections. Functional common input would bind the pieces—the atoms—through statistical dependence, most likely partial synchrony. In any one realization the connectivity would be sparse, reflecting something of the topology of the visual scene, albeit with a superimposed hierarchical structure, and perhaps elaborated by abstract conjunctions—folds—that are not literally local ("same", "parallel", and so-on). The remarkable connectivity of the brain (some estimate the graph-theoretic "diameter" to be smaller than 10—"six degrees of separation"), as revealed for example by the studies of van Essen et al. [55, 31], would facilitate this sort of representation by providing a rich anatomical scaffolding that makes close functional pairings possible even among physically distant cells.

This proposal, of representation through commitment to a subgraph of the anatomical graph, raises more questions than it answers, including, especially:

- What kinds of neuronal mechanisms could supply rapid and reversible changes in connectivity, and in particular, a robust commitment to selected subsets of presynaptic cells? In other words, by what mechanisms is a connectivity "subgraph" chosen from the presumably far richer graph of *anatomical* connections? Possibly, this could arise from the nonlinear dynamics of the neurons themselves, as in models discussed above, or possibly synaptic changes, on a very short time scale, are involved. Indeed, there is now plenty of evidence for both. Recent experiments by Polsky et al. [39] indicate that dendritic nonlinearities, *by themselves*, could support the kind of subset selection process idealized in the MAX filter, even without a precursory layer of "simple cells". And experiments by Markram et al. [32] reveal the rapid and reversible changes in synaptic plasticity anticipated by von der Malsburg in his Correlation Theory [58]. Either way, the effect would be the same: functional connectivity leads to (circumstantial) functional common input leads to partial synchrony.

- I have postulated that partial synchrony, i.e. near-synchronous spikes in response to a temporary and partial sharing of inputs, signals a correct arrangement of parts in the receptive fields of two invariant cells. How would the two signals, firing rate and partial synchrony, interact to influence a postsynaptic cell's activity, and in particular, how sensitive might a cell be to partial synchrony, *per se*? A simple product of "fci" (modeling synchrony) and filter outputs (modeling firing rates) was enough to substantially improve selectivity in the experiments in §4. Are there mechanisms, perhaps involving inhibitory inter-neurons, for sharpening this effect (e.g. "synchrony by competition" as proposed by Tiesinga and Sejnowski [50])? Two recent experiments demonstrate striking nonlinearities that bear on the issue of signal to noise: The experiments of Azouz & Gray [8] and those of Polsky et al. [39] both demonstrate a "super-additive" interaction, indicating that the response to two well-timed events can be substantially greater than the sum of individual responses.

What are the prospects for experimental exploration of these ideas? It is possible that the foundation—functional common input inducing partial synchrony—can be put to the test with available technologies. I have already remarked that this kind of synchrony would be non-transitive, and in this sense, local. Compare this to the idea of a synfire chain, in which *every pair* of a population of neurons is correlated. If these

correlations are instead local, then we can hardly expect to see evidence for them in a randomly selected pair of cells responding to a randomly selected image. On the other hand, we might be able to exploit the notion of composition through overlapping and matching representations (as in §4) to devise a more directed search in a challenging, but perhaps feasible, experiment. In brief, the proposal would this:

1. Fit a suitable nonlinear model (e.g. an NNP-type model—see Harrison et al. [25]) to each of two simultaneously recorded complex cells with overlapping receptive fields.

2. Record the time course of some *analytic* measure of the functional common input. The input/output models built for the pair of complex cells would amount to two explicit spike-activity functions, say $f(\vec{x})$ and $g(\vec{x})$, of the intensity image $\vec{x}$, as measured over the union of the receptive fields. (Possibly $\vec{x}$ would first be normalized, or otherwise pre-processed, as in §4.) One natural analytic measure of functional common input, similar to the one used for the experiments in §4, would be the sum of the products of the partial derivatives, i.e. the inner product of the gradients: $\nabla f(\vec{x}) \cdot \nabla g(\vec{x})$.

3. Record, as well, the time course of the *significance* of synchronous events in, say, the trailing one-hundred milliseconds[5].

A correlation between the time courses of functional common input and synchrony would be good evidence for a readable signal carrying information about the relative phase (position) of target patterns in the cells' receptive fields. And given that the computations can probably be managed in real time, "on line", it might be possible to improve the efficiency of the endeavor by driving the cells in the direction of synchrony, perhaps through a gradient procedure analogous to the ones suggested by Földiák [16] and Földiák et al. [17], or perhaps "by hand" through patterns with pieces that excite the respective cells and fit together coherently. In any case, the receptive-field patterns that actually produce high measures of functional common input could be examined for evidence of a properly aligned composition of individual target patterns.

[5]There is a *dynamic programming* algorithm, based on the method of "spike jitter" (see Hatsopoulos et al. [26] or Amarasingham [5]), that is computable in real time and corrects for possible phasic inputs that masquerade as significant synchrony (see Harrison & Geman [24]).

6 Summary

The rich local connectivity of the brain, together with the largely topological inter-area organization, suggests that assemblies of cells with a multitude of anatomically shared inputs might be commonplace. Yet by virtue of the nonlinearity of neuronal dynamics, a given cell at a given instant may or may not be sensitive to a given presynaptic input. Sensitivity, or *functional connectivity*, depends globally on the entire presynaptic activity pattern. It follows that the degree of *functional* common input, as opposed to the degree of anatomical common input, is circumstantial, depending in particular on the collective pattern of presynaptic activities. If common input promotes synchronous events ("partial synchrony"), then synchrony is evidently also circumstantial and therefore carries information about presynaptic activities.

What kind of information would be signaled by synchronous spikes, in say a pair of neurons? If these neurons represent target patterns, with invariance to some parameters of pose, then by little more than the definition of invariance we should expect these neurons to be functionally connected to inputs representing areas in and around their respective stimulating target patterns. If this were the case, then two such cells, sharing a significant population of common presynaptic neurons, would have a strong *functional* common input exactly when the poses of the respective target patterns were consistent, i.e. when the patterns were aligned and coherent. By these considerations, synchronous spikes ought to signal local consistencies of arrangements, as argued for by Singer [47] and others.

Horace Barlow [9] suggested that learning amounts to building a hierarchy of accommodations of "suspicious coincidences". A recognition device capable of signaling the presence of parts (e.g. edge-type discontinuities, contours, strokes, arrangements of strokes, hands, and so on), say invariantly to location or other pose parameters, would be otherwise unprepared for the coincidences of *arrangements* of parts induced by their participation in one or another compound entity (a stroke, a letter, an arm, and so on). In the absence of a model for the composition, the arrangement of parts is highly suspicious. More than a century before, Laplace [29], in an essay about probabilities, contemplated some of these same matters, and concluded similarly that we group pieces into a familiar whole precisely because the alternative represents an unlikely coincidence. Possibly, a departure from independence in the outputs of local collections of cells, manifested explicitly by an abundance of near-synchronous spikes, signals the

suspicious coincidences that drive both the learning of a hierarchy of reusable parts as well as the construction of a temporary internal representation of this hierarchy.

In that these observations derive from basic physiological principles of cortical microcircuitry, they are not necessarily limited to vision *per se*. It is possible that these same mechanisms of functional connectivity, functional common input, and partial synchrony support highly dynamic and organized topological representation throughout the cortex.

Maybe "binding by synchrony" is more mainstream than one might first think. Maybe it is a logical conclusion from a chain of mainstream ideas. There is nothing radical about the idea of common input producing synchrony, or for that matter the idea that synchronous spikes make for a particularly effective stimulus (coincidence detection). An apparent leap is the idea of functional connectivity, and its consequence, functional common input. But functional connectivity amounts to little more than nonlinearity. The real leap is in the proposition that invariant cells functionally connect to their target patterns, and that the resulting functional common input creates a strong and readable signal for composition.

Acknowledgment. I would like to thank Elie Bienenstock, Donald Geman, and Matthew Harrison for sharing their ideas and shaping mine. Supported by Army Research Office contract DAAD19-02-1-0337, National Science Foundation grant DMS-0427223, and National Science Foundation grant IIS-0423031 as part of the NSF/NIH Collaborative Research in Computatonal Neuroscience Program.

References

[1] M. Abeles. *Local cortical circuits: An electrophysiological study*. Springer-Verlag, Berlin, 1982.

[2] M. Abeles. *Corticonics: Neuronal circuits of the cerebral cortex*. Cambridge University Press, Cambridge, UK, 1991.

[3] E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. *J. of the Optical Society of America A*, 2:284–299, 1985.

[4] J.M. Alonso, W.M. Usrey, and R.C. Reid. Precisely correlated firing in cells of the lateral geniculate nucleus. *Nature*, 383:815–819, 1996.

[5] A. Amarasingham. *Statistical methods for the assessment of temporal structure in the activity of the nervous system*. PhD thesis, Brown University, Division of Applied Mathematics, 2004.

[6] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

[7] Y. Amit and D. Geman. A computational model for visual selection. *Neural Computation*, 11:1691–1715, 1999.

[8] R. Azouz and C.M. Gray. Adaptive coincidence detection and dynamic gain control in visual cortical neurons in vivo. *Neuron*, 37:513–523, 2003.

[9] H.B. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.

[10] E. Bienenstock. A model of neocortex. *Network: Computation in Neural Systems*, 6:179–224, 1995.

[11] E. Bienenstock. Composition. In A. Aertsen and V. Braitenberg, editors, *Brain Theory - Biological Basis and Computational Theory of Vision*, pages 269–300. Elsevier, 1996.

[12] G.A. Carpenter and S. Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 37:54–115, 1987.

[13] F. Crick. Function of the thalamic reticular complex: The searchlight hypothesis. *Proc. Natl. Acad. Sci. USA*, 81:4586–4590, 1984.

[14] Y. Dan, J.M. Alonso, W.M. Usrey, and R.C. Reid. Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus. *Nature Neuroscience*, 1:501–507, 1998.

[15] J.A. Fodor and Z.W. Pylyshyn. Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28:3–71, 1988.

[16] P. Földiák. Stimulus optimisation in primary visual cortex. *Neurocomputing*, 38-40:1217–1222, 2001.

[17] P. Földiák, D.K. Xiao, C. Keysers, R. Edwards, and D.I. Perrett. Rapid serial visual presentation for the determination of neural selectivity in area stsa. *Progress in Brain Research*, 144:107–116, 2003.

[18] K. S. Fu. *Syntactic Pattern Recognition and Applications*. Prentice-Hall, Englewoods Cliffs, 1982.

[19] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980.

[20] D. Geman. Coarse-to-fine classification and scene labeling. In D. Denison et al., editor, *Nonlinear Estimation and Classification*, Lecture Notes in Statistics, pages 31–48. Springer-Verlag, New York, NY, 2003.

[21] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1991.

[22] U. Grenander. *Abstract Inference*. Wiley, New York, NY, 1981.

[23] U. Grenander. *General Pattern Theory: a mathematical study of regular structures*. Oxford University Press, Oxford, 1993.

[24] M. Harrison and S. Geman. An exact jitter method using dynamic programming. Technical Report APPTS #04-3, Brown University, Division of Applied Mathematics, 2004. http://www.dam.brown.edu/ptg/REPORTS/04-03.pdf.

[25] M. Harrison, S. Geman, and E. Bienenstock. Using statistics of natural images to facilitate automatic receptive field analysis. Technical Report APPTS #04-2, Brown University, Division of Applied Mathematics, 2004. http://www.dam.brown.edu/ptg/REPORTS/04-02.pdf.

[26] N.G. Hatsopoulos, S. Geman, A. Amarasingham, and E. Bienenstock. At what time scale does the nervous system operate? *Neurocomputing*, 52-54:25–29, 2003.

[27] H.D. Hubel and T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154, 1962.

[28] E.M. Izhikevich, J.A. Gally, and G.M. Edelman. Spike-timing dynamicss of neuronal groups. *Cerebral Cortex*, 14:933–944, 2004.

[29] P.S. Laplace. *A philosophical essay on probabilities (Essai Philosophique sur les Probabilités)*. Dover, New York, 1951. (F.W. Truscott & F.L. Emory, Trans., original work published 1814).

[30] T.S. Lee, C.F. Yang, R.D. Romero, and D. Mumford. Neural activity in early visual cortex reflects behavioral experience and higher-order perceptual saliency. *Nature Neuroscience*, 5:589–597, 2002.

[31] J.W. Lewis and D. Van Essen. Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. *J. of Comparative Neurology*, 428:112–137, 2000.

[32] H. Markram, Y. Wang, and M.V. Tsodyks. Differential signaling via the same axon of neocortical pyramidal neurons. *Proc. Natl. Acad. Sci.*, 95:5323–5328, 1998.

[33] D. Marr. *Vision*. W.H. Freeman, San Francisco, CA, 1982.

[34] B.W. Mel. Seemore: Combining color, shape and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation*, 9:777–804, 1997.

[35] S. Mikula and E. Niebur. The effects of input rate and synchrony on a coincidence detector: analytical solution. *Neural Computation*, 15:539–547, 2003.

[36] D. Mumford. On the computational architecture of the neocortex: I. The role of the thalamo-cortical loop. *Biological Cybernetics*, 65:135–145, 1991.

[37] D. Mumford. On the computational architecture of the neocortex: II. The role of cortico-cortical loops. *Biological Cybernetics*, 66:241–251, 1992.

[38] S. Papert. The summer vision project. Technical Report Memo AIM-100, Artificial Intelligence Lab, Massachussets Institute of Technology, 1966.

[39] A. Polsky, B.W. Mel, and J. Schiller. Computational subunits in thin dendrites of pyramidal cells. *Nature Neuroscience*, 7:621–627, 2004.

[40] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999.

[41] M. Riesenhuber and T. Poggio. Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, 12:162–168, 2002.

[42] E.T. Rolls and M.J. Tovee. The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. *Experimental Brain Research*, 103:409–420, 1995.

[43] Y. Rui and Z. Liu. Artifacial: Automated reverse turing test using facial features. In *ACM Mulitmedia Systems Journal*, May 2004.

[44] K. Sakai and S. Tanaka. Spatial pooling in the second-order spatial structure of cortical complex cells. *Vision Research*, 40:855–871, 2000.

[45] L. Shastri and V. Ajjanagadde. From simple associations to systematic reaoning: A connectionist representation of rules, variables and dynamic bindings. *Behavioral and Brain Sciences*, 16:417–494, 1993.

[46] D.L. Sheinberg and N.K. Logothetis. Noticing familiar objects in real world scenes: The role of temporal cortical neurons in natural vision. *Journal of Neuroscience*, 21:1340–1350, 2001.

[47] W. Singer. Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, 24:49–65, 1999.

[48] C.J. Stone. Consistent nonparametric regression. *Annals of Statistics*, 5:595–620, 1977.

[49] M.J. Tarr. Pandemonium revisited. *Nature Neuroscience*, 2:932–935, 1999. (News and Views on: Riesenhuber & Poggio, Hierarchical Models of Object Recognition in Cortex, same issue).

[50] P.H.E. Tiesinga and T.J. Sejnowski. Rapid temporal modulation of synchrony by competition in cortical interneuron networks. *Neural Computation*, 16:251–275, 2004.

[51] A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[52] M.V. Tsodyks and H. Markram. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proc. Natl. Acad. Sci.*, 94:719–723, 1997.

[53] A.M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.

[54] S. Ullman. Sequence-seeking and counter streams: A computational model for bi-directional information flow in the visual cortex. *Cerebral Cortex*, 5:1–11, 1995.

[55] D.C. Van Essen, C.H. Anderson, and D.J. Felleman. Information processing in the primate visual system: an integrated systems perspective. *Science*, 255:419–423, 1992.

[56] V.N. Vapnik. *Estimation of dependences based on empirical data*. Springer-Verlag, New York, NY, 1982.

[57] W.E. Vinje and J.L. Gallant. Natural stimulation of the nonclassical receptive field increases information transmission efficiency in v1. *Journal of Neuroscience*, 22:2904–2915, 2002.

[58] C. von der Malsburg. The correlation theory of brain function. Technical report, Max-Planck Institute for Biophysical Chemistry, Department of Neurobiology, Göttingen, Germany, 1981. (Reprinted in *Models of Neural Networks II*, 1994, E. Domany, J.L. van Hemmen, and K. Schulten, eds., Berlin, Springer).

[59] H. White. Connectionists nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3:535–549, 1990.

[60] L. Wiskott and T.J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14:715–770, 2002.

[61] Y. Yoshimura, J.L.M. Dantzker, and E.M. Callaway. Excitatory cortical neurons form fine-scale functional networks. *Nature*, 433:868–873, 2005.