Commentary on the article "Banishing the Homunculus" by Horace Barlow

David Mumford
Department of Mathematics, Harvard University

Feedback and the Homunculus

The problem of the homunculus is usually presented in the following setting: an information processing agent is described which has senses and effectors. The input from the senses is led to an intial processing stage, where significant features are extracted; these are led to a higher stage of processing, and to another, perhaps multi-modal integrative stage. At some point a decision is made about what the stimulus 'means' to the agent and now the process is reversed. The decision leads to a choice in a large plan, which in turn is translated into a current step in a finer grain plan. Eventually, specific muscle/motor commands are issued to the effectors. In this architecture, one has the feeling that the essential component of thought has not been analyzed and that at the decision stage, there is still the need for a little man to look at the refined description of the sensory input, to think it over and decide how he wants to modify his master plan. If not that, then we seem to be thrown back on a Rod Brooks-style finite state automata at the top level, and this seems awfully stupid compared to our image of ourselves.

Unfortunately, much of the theorizing of neurophysiologists and psychologists has reinforced this architectural view of what the brain does. Time and time again, we see flowcharts in which the retina sends data to the LGN which sends data to V1 which sends data to V2 and MT leading, for instance to the dorsal and ventral pathways through posterior cortex, then to frontal lobe, etc. One example which is explicitly in the homunculus tradition is that a window of attention is superimposed on the low-level image pyramid and that this window is *copied* into IT for analysis and, hopefully, object recognition. This approach would then say that the role of V1 is to provide a massive filter bank, extracting important local features of the 2-space-dimensional and 1-time-dimensional visual signal. Although more and more properties of V1 single-cell recordings are complicating this model, it is still the dominant paradigm.

I would like to make a hypothesis for the role of V1 which is very different from the above and which deflates substantially the role of the homunculus (without, however, explaining the computations needed to replace him). This

hypothesis is that the role of V1 in the neocortex is not as the first processing stage for vision, but as *the one and only high-resolution visual buffer in the neocortex, which is involved in all computations which need high-resolution, such as recognizing objects and discovering differences between memories and current stimuli, when details are crucial.*

We have to be careful about what this means and how it might be possible. Firstly, feedback pathways are absolutely essential in order that V1 play a continuing role in the high level processing as well as the low level processing of an image. These are, of course, found in rich abundance in the cortex. Secondly, it does imply that V1 single cell responses will be affected by high level aspects of the image, but it does not mean that V1 is doing object recognition all by itself. Some integrative mechanism which forms temporary cell assemblies, linking V1, V2, V4 and IT for example would be needed to coordinate cell activity during such a computation. In this integrated activity, the role of V1 (and perhaps the LGN) is to provide the buffer in which detailed visual structure is identified and placed into complex global structures, involving lighting, depth, object classes and the like. Recent experiments [Zipser et al, 1994] can be interpreted as lending support to this hypothesis, but the crucial tests have not been done.

Where is the homunculus in such a model? Clearly, the little guy can no longer sit in a small room at the top of the information processing pyramid. He must now have a hundred TV cameras checking out all the low level buffers, the intermediate areas where global structures are tracked and the higher areas where multi-modal interpretations are weighed. Do we really need him then?

Markov Random Fields and Sparse Coding

Sparse coding, e.g. the hypothesis that one or a small number of cell's firing signifies the presence of grandmother, is very attractive, but a clear computational reason for believing in it has been elusive, as Barlow points out. I would like to argue that the computational model provided by Markov Random Fields presents one possible explanation. Most of the analysis of sparse coding has been done using information theory in an abstract setting divorced from details about the world. A totally different approach is to start with the premise that the brain's task is to infer hidden facts about the state of the world and that how to accomplish this task depends on the nature of the probability space defined by possible world states. I will argue that if this space is appproximated by a Markov Random Field, such inferences can only be done efficiently using sparse coding.

It may be helpful to review what Markov Random Fields (MRF's) are. Let us assume that we have a large collection $\{X_v\}$ of variables, which could be

light intensities at different pixels, edge or filter strengths, lighting directions, Boolean variables expressing things like 'granny present' or 'patient has bacterial infection'. Some of these may be known in the current situation, some unknown, but all vary from time to time. We want to use statistics to estimate them, so we need to make a probability space out of the set of all possible sets of values $\{X_v = a_v\}$ they may have, i.e. give a big table of probabilities of all such assignments. As is well-known, this is impossible to do by storing the table, which gets vastly too large to store for even tens of variables. MRF's provide one of the few ways to effectively define probabilities on such ensembles. What makes the concept natural is that MRF's may be defined in 2 equivalent ways: by an abstract requirement of certain conditional independencies or by a simple concrete formula for the probability of each assignment. It works by assuming that the variables can be thought of as the nodes of a graph, in which edges join any pair of variables *which are directly dependent*. The important point is that the graph is sparse: each variable has only a relatively few edges. (By relative, I mean, e.g. that if there are a hundred million variables, then perhaps the maximum number of edges at one node – the degree of the graph – should be bounded by ten thousand.) The abstract definition of a MRF is that if some set of variables $\{X_w\}_{w \in W}$ is fixed and if any path in the graph joining X_a and X_b must cross W, then X_a and X_b are to be conditionally independent, given $\{X_w\}_{w \in W}$. The concrete definition says that

$$-\log(Pr(\{X_v = a_v \text{ all } v\}) = \sum E_C(\{a_v\}_{v \in C}) + \text{ cnst}$$

where C ranges over the 'cliques' of the graph, the clusters of totally connected nodes and E_C, called the 'energy', is a local interaction term, involving only the variables in the clique.

The basic hypothesis of those who apply MRF's to modeling thought – whether in speech recognition, vision or medical expert systems – is that MRF's are a good approximation to the true probability distibution of the random variables of the world. This is the strongest form of the hypothesis. It is certainly possible to believe in modified or weaker forms of the hypothesis: to extend MRF's with pointers (as in the point process models of [Ripley & Kelly, 1971]), or to merely ask that this approximation is usable even when inaccurate, or to fall back on saying that the part of the world with enough independencies to be MRF-like is the only one we can think about.

If we accept the above *MRF hypothesis*, then what are its neural implications? Clearly the brain must be able to learn and store the local energy terms E_C in some way. Given that the set of neurons in the brain is also a relatively sparsely connected graph, this suggests that some kind of rough correspondence exist between the two graphs. Let's look at two possibilities:

Consider the extreme case where there is no correspondence. This means that

we have a fully distributed representation of each hidden variable X_v in the firing patterns of millions of neurons. To express a direct link between two such variables, which does not involve any other variables whose representations overlap extensively with the first two, we must extract some invariant from the whole firing pattern which signifies the individual variable. This would seem to be a decoding problem as hard as calculating X_v in the first place.

The other extreme version is to make the MRF graph and the neuronal graph isomorphic. This is the ultimate grandmother cell theory, with 'grandmother synapses' as well, i.e. two cells linked if and only if the corresponding random variables are connected by an edge in the MRF. An example of such a theory is Hopfield's symmetric weight neural net theory. Less extreme would be a theory in which a small number (e.g. 100) of neurons carried the value a_v. This is a sparse coding theory in the sense that some small cluster of cells somewhere expresses the presence of grandmother. It seems consistent not only with Hopfield-style nets, but with other theories in which individual pulses carry information such as the theory of synfire chains [Abeles 1991]. I believe that accepting the MRF hypothesis drives you strongly to some form of sparse coding.

References

1. M.Abeles, 1991, *Corticonics*, Cambridge Univ. Press.

2. B.Ripley & F.Kelly, 1977, Markov point processes, *J.London Math. Soc.*, **15**:188-192.

3. K.Zipser, V.Lamme, T.S.Lee & P.Schiller, 1994, A role for primate striate cortex in cue-independent scene segmentation (submitted to *Nature*).