

# Spatial Competence via Self-Organisation: an Intersect of Perception and Development.

Mark W. Peters (markpeters@cse.unsw.edu.au)

Department of Artificial Intelligence, School of Computer Science and Engineering, University of New South Wales,  
Sydney, NSW 2052, Australia.

## Abstract

We address the question of how artificial systems and natural organisms develop spatial competence. Most artificial systems draw upon considerable sophisticated operator- or developer-originated knowledge about what in the world sensor signals represent. Natural systems do not have such sophisticated auxiliary sources of information. We are interested in how, despite this, they achieve perceptual organisation, and suspect that the methods they use will have generalisable effectiveness. We describe a process that creates coherent mappings between the physical world and the phenomenological realm, analogous to retinotopicity and sensory homuncularity in natural systems, and discuss its application to problems of higher dimensionality and higher levels of abstraction. Importantly, such a process, having proved successful in the perceptual robotics domain of our current interests, is likely to be found in other cognitive domains because its strengths lie in its ability to organise and implicitly summarise data in the absence of clues about what that data represents.

## Introduction

Most artificial systems draw upon considerable sophisticated operator- or developer-supplied knowledge when making assumptions about what lies on the external side of their sensor arrays. Natural systems do not have such sophisticated auxiliary, or *a priori*, sources of information. They must come to make the right assumptions completely autonomously (given the head start provided by their genetic endowment).

Several artificial self-organising low-level vision systems have been developed, e.g., by Linsker (1988). Generic to these systems is a dependence on a pre-existent orthogonal matrix of inputs, resulting from choices made by the operator or developer, or imposed by video or hardware standards. The data that these discrete vision systems process come conveniently packaged in this orthogonal matrix, which itself contains much implicit information about the outside world, but not all of it helpful.

Prokopowicz (1994) showed that there is no inherent value in the orthogonality of the input image, but that what is essential to a visual robot is an internal mapping between pixels and motor positions. His system, IRV, eschewing orthogonality, locates pixels relative to each other via motor commands in the form of a set of statements to the effect that, given motor movement  $M$ , the point in the environment that now appears in pixel  $B$ , will next appear in pixel  $A$ .

IRV is dependent on unequivocal knowledge of its motorium, and it is from this sound basis that it builds the organisation of its sensorium. Knowledge of the motorium takes the

form of direct control of the pan and tilt motors that govern its camera's direction of gaze. IRV therefore bears somewhat asymmetrical relationships to its sensorium (randomly ordered but subjected to modification) and its motorium (ordered, fixed, and used as a reference).

This asymmetry raises an interesting question: are the organisational principles used by IRV capable of learning *both* sensory organisation *and* motor organisation? And, if so, can they be used to evolve a system which might lie between any unknown sensorium-motorium coupling, and *learn* its way to a position of control over the whole apparatus?

The answer to these questions is positive if the sensorium can self-organise, since it can, in turn, be used to organise an unknown motorium, as has been shown, e.g., by van der Smagt (1995), whose work is in some ways a mirror image of IRV. And if so, this might help explain how immature animals learn to improve simultaneously both their coordination and perception, becoming spatially competent adults. This is one motivation for undertaking this research. The potential to develop robots that first configure themselves (and in so doing, adaptively learn how to perceive and behave) is a second motivator, but perhaps the most important motivation is that the problem characteristics can be cast in a context-independent form. The problem is that of learning, unsupervised, in an open loop, and in the absence of all meta-information, a succinct representation whose internal similarities aptly capture similarity in the unseen source of the data. The applicability of a solution therefore extends beyond the domain under current discussion.

We use the analogy of a jigsaw puzzle to illustrate the principle characteristic of our algorithm. The usual steps for solving a jigsaw puzzle can be characterised thus:

1. Locate corner and edge pieces
2. Coarsely group all pieces according to main colour and texture (e.g., 'sky')
3. Select pairs of high similarity
4. Fit pieces to nearest neighbours found in step 3.

We note that steps 1 and 4 both depend on the form (edges) of the pieces, whereas steps 2 and 3 depend only on their content (colour and texture). However, difficult problems are often devoid of convenient *a priori* 'form' clues (analogous to jigsaw piece edges, or location in an orthogonal matrix) because these are the artifacts of a prior organisation - precisely what we do not have, need to derive, and cannot assume. We must instead rely solely on a

method analogous to steps 2 and 3, grouping according to similarity.

### Methods

Rather than simply arranging pieces of an image, our current task is to arrange inputs that carry those pieces of information to us. Once arranged, the inputs will convey any image correctly. To extend the analogy, it is as if, having completed the jigsaw, we are able to note the location of each piece and the precise unique shape of that piece. Using that information alone we are able to assemble any other jigsaw from the same cutter, without seeing its image.

We define a sensory input as a line of communication that provides information about a single point in the world. This information, its signal, varies over time if that part of the world changes. An input also has an internal location in a context made by the set of all inputs.

There are three ways in which any two inputs can be said to be related:

**Environmental Proximity** – the points in the world to which the inputs refer are close together;

**Behavioural Proximity** – over time the values of the inputs vary proportionally and in time with each other;

**Geometrical Proximity** – the internal locations (coordinates) assigned to the inputs are close together.

It is the relationships of these three kinds of proximity that form the subject of this paper. The input to our algorithm is behavioural data; the output is geometrical data. We show that there is sufficient isomorphism between unavailable environmental data and available behavioural data such that the latter can be used to construct a coherent geometrical representation of the former. From this observation we ultimately wish to bootstrap a learning visuo-motor agent (see Figure 1).

As the goal of this work is to *derive* order for a set of inputs, it is inappropriate to *assume* or *inherit* order from elsewhere. Therefore we must ignore the organisation inherent in the standard video matrix. Similarly, we do not depend on any framing concepts such as up, down, left, right, or any alignments with the direction of fall, or the robot's base or motor axes. In fact, we randomly scrambled the locations of all pixels (c.f. Prokopowicz 1994). Doing so means that our geometrical arrangement of inputs remains immune to configuration peculiarities and is determined only by the *content* of the signals.

One can draw no inference about the organisation of any visual system on the basis of one static image. If there is no change in the input, any applied ordering produces spurious geometrical arrangements, bearing no relation to the envi-

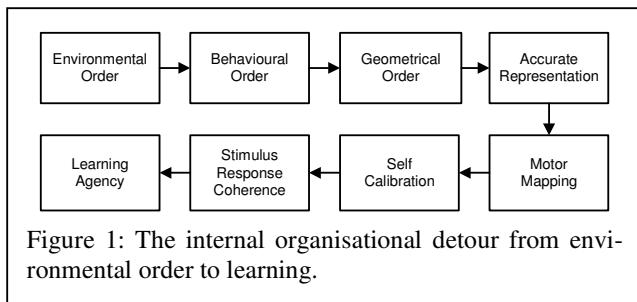


Figure 1: The internal organisational detour from environmental order to learning.

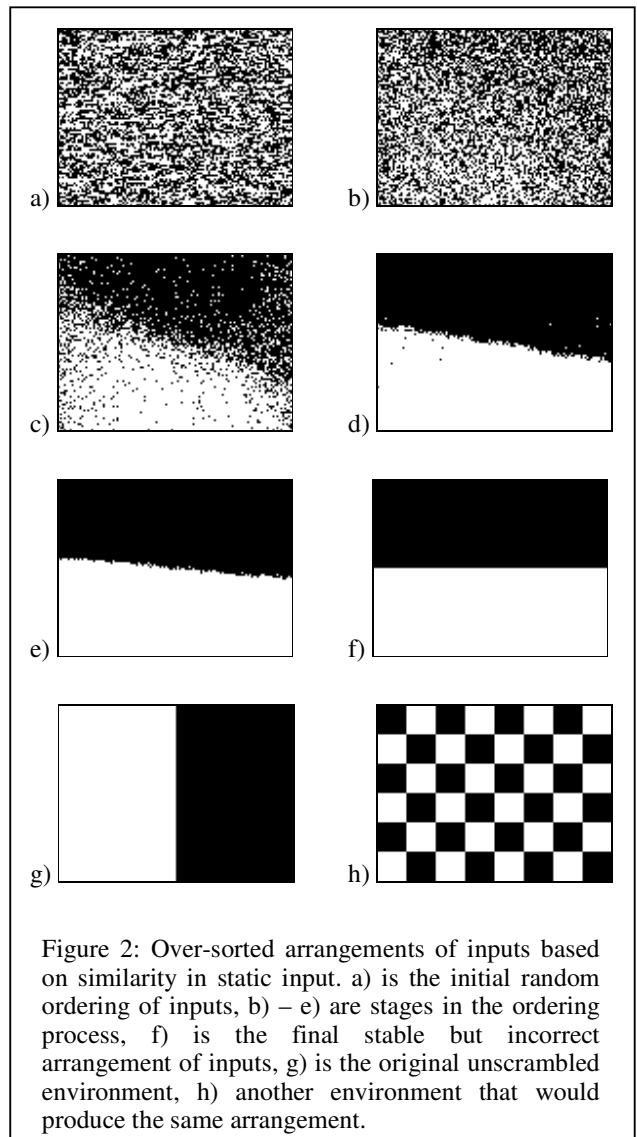
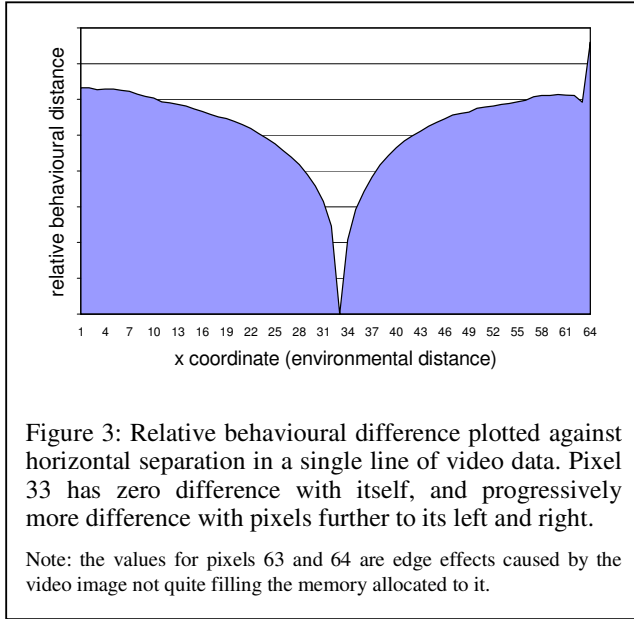


Figure 2: Over-sorted arrangements of inputs based on similarity in static input. a) is the initial random ordering of inputs, b) – e) are stages in the ordering process, f) is the final stable but incorrect arrangement of inputs, g) is the original unscrambled environment, h) another environment that would produce the same arrangement.

ronment (see Figure 2) as there is too much ambiguity present in the relationships between pixels. It is as if the jigsaw had many pieces of the same shape and colour, which can just as well be set in many different arrangements. So, the criterion for organising the array of inputs can only be behavioural proximity (common histories between inputs, or similarity over time, not just in a snapshot).

We first tried to use natural environmental movement as the source of change. This proved to be ineffective, as it is difficult to find convenient places where a camera can be set up in the confidence that sufficient movement will be detected by all pixels. The next approach was to use the camera's pan-tilt mount to continuously change the direction of gaze of the camera, but this was curtailed due to fear that the motors would overheat. The most effective solution was to dispense with a camera completely, and simply connect a VCR, tuned to a television station, directly into the computer. This provided a continuous source of all kinds of 'first person' movement: pans, tilts,



zooms, tracking; and plenty of natural ‘third person’ movement too.

Using video input we can show that the behavioural proximity between any two inputs is monotonically related to their environmental proximity (see Figure 3). So, behavioural proximity implies environmental proximity. Curiously, once the significant noise has been removed this relationship is closely approximated by the elegant curve:

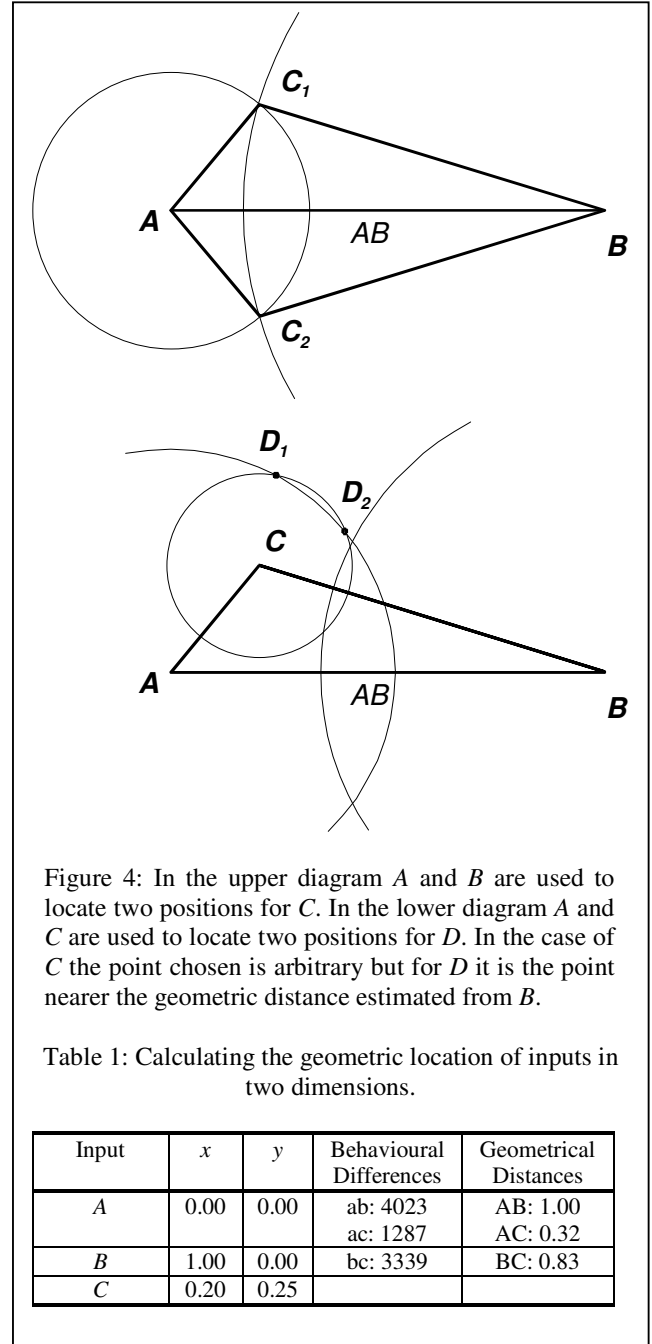
$$y = 1 - 1/e^x \quad (1)$$

We wonder whether this relationship might be a general one, since it seems to convey the diminishing probability, having found something at one location, of finding something similar at other increasingly distant locations. Armed with the knowledge of the relationship’s monotonicity, we can now describe the two-dimensional version of the method in detail.

We start by randomly selecting one input, *A*, and searching for others with high behavioural proximity to *A*. Behavioural proximity can be defined by a threshold of, say, 15% of the maximum possible behavioural difference between inputs. Once we have found enough similar inputs (the number required is the number of dimensions we intend our representation to have, plus 2) we observe them for a while, in order to determine all mutual behavioural proximities.

We assign nominal locations (i.e., Cartesian coordinates) to two of these inputs, starting by assigning, say, (0, 0) to *A*, (1, 0) to *B*. Other inputs will be arranged relative to these two according to their mutual behavioural proximities. For instance, in Table 1, the geometric distances of *C* (0.2, 0.25) have been calculated from the behavioural differences *ab*, *ac*, *bc*, and the geometric distance  $AB = 1$ , as follows:

$$AC = \frac{AB \cdot f(ac)}{f(ab)} \quad BC = \frac{AB \cdot f(bc)}{f(ab)} \quad (2)$$



from which the Cartesian coordinates of *C* are easily derived. For *C* we arbitrarily choose one of the two points, *C*<sub>1</sub>, *C*<sub>2</sub>, that satisfy the geometry. For subsequent inputs we follow essentially the same procedure: comparing a short-list of inputs already set, and selecting from them those exhibiting most behavioural proximity, then using formulae of the same form as (2), calculating two possible locations for the candidate input. Of these two we choose the one nearer the distance estimated using the behavioural proximity of another reference input (see Figure 4).

Note that relative to environmental space, our geometric space may appear skewed, inverted, scaled, or even bent. This is not significant since the actual mapping is still monotonic and, minimally, needs only provide the motorium with a monotonically coherent representation of the environment. It is therefore sufficient. A second learning process, e.g., van der Smagt (1995) must be implemented to map motor commands from this geometrical representation back to the world.

## Results

After only one pass the process converges closely upon the environmental order it cannot directly observe. In Figure 5 we present the resulting geometric location plotted against environmental location. Note that, even at this stage, a previously scrambled image passed through the new rearrangement of inputs is perfectly comprehensible to the naked eye, exhibiting relatively minor distortions as objects translate through the image plane. Any robotic control system that operates non-ballistically would be able to use such input to iteratively position either itself or a mechanical arm relative to observed objects.

The process can be enhanced in a number of ways. A short list may be maintained for each input, containing other inputs, which have so far been found to show greatest behavioural proximity to it. After the first pass through all inputs, this list can be used for fine-tuning of locations at a local level, for all subsequent passes. Extending the time over which observations of behavioural proximity are made reduces noise effects. Raising the threshold for inclusion in triangulation calculations also improves accuracy, since it constrains selection to near neighbours, whose environmental-behavioural proximity relationships are less ambiguous, due to the proximal steepening of the curve in Figure 3.

Incorrect triangulations early on in the process have strong detrimental effects on subsequent organisation. It is therefore worthwhile allowing the earlier observations to extend for longer periods than later ones, and to use a much higher behavioural proximity threshold.

The entire process may be analogous to the extension of axons from one area of the nervous system to another. Those that first reach a, say, cortical destination are the most free to take a location, and the least likely to change their location. Later axons will seek out areas most corresponding to their own signals, but will exert negligible attraction on such areas simply because of the weight of numbers already there.

## Applications

Spatial mappings derived in the manner just described obviously do not possess the neat orthogonality of standard pixel arrays. Standard grid-like neighbourhoods and all convolutions that depend on them are therefore inapplicable to data in this form. However, non-orthogonal, neighbour-independent methods have been found for edge detection (Prokopowicz & Cooper 1995), motion detection (Prokopowicz 1994), the location of centroids in both artificial (Peters & Sowmya 1996, 1997, 1998b) and natural (Sparks, Lee & Rohrer 1990) systems, and the calculation of spatially located interest metrics (Peters 1998).

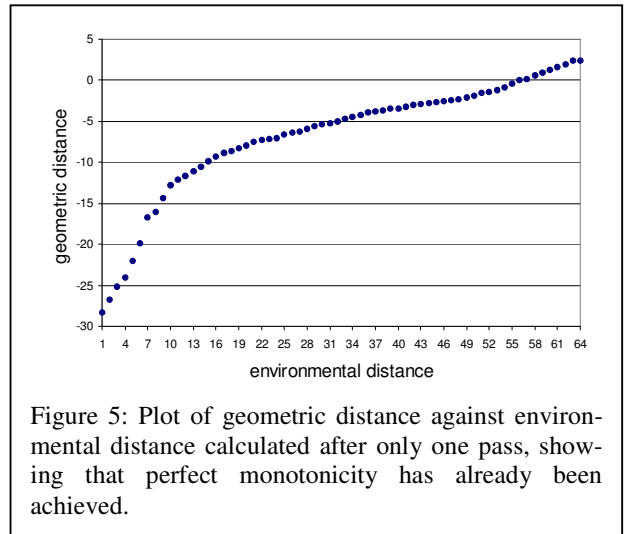


Figure 5: Plot of geometric distance against environmental distance calculated after only one pass, showing that perfect monotonicity has already been achieved.

The set of inputs need not be restricted to just one source. Multiple cameras can be set up to supply incoming data, and the algorithm can still be applied. This is thus a solution to some forms of the data fusion problem. There may be a need to adjust signals to normalise both range and sign if sensor devices are of sufficiently different design. Additionally, derivatives may be used instead of the original signal.

Given that the geometric arrangement of inputs depends solely on environmental changes it will evolve to a space-invariant state, effectively creating the inverse function for the optical distortion and sampling bias of the anterior parts of the system. Note that it should not be assumed that sampling patterns would be fixed during operation. Peters & Sowmya (1998a) have shown that there are good reasons for changing sampling density and bias according to data received *during* operation.

Having overcome space-variance, the system is able to produce a relatively metrical memory map of its surrounding space, which can then be used to note changes that take place in the environment even while the system is looking in another direction. It can also be used to calibrate a motor mapping by following a program of random moves, and observation of the resultant changes.

Much self-calibration research has been directed at finding the minimum set of prerequisites for calibration. Pollefeys & Van Gool (1997) have shown that a system requires a minimum of three images (of the same scene, with common points identified) to complete its calibration. Other approaches include developing a system to derive the function that converts given pixel locations to given spatial locations (Sharma & Srinivasa 1996, Srinivasa). There has been very little work attempting to deal with varying camera parameters, though this has been shown to be possible in the absence of skew (Pollefeys, Koch, & Van Gool 1997).

These self-calibration solutions can only be used in systems that are already somewhat organised. They require:

1. certainty that the world has not changed between successive images
2. an organised image-producing infrastructure that enables recognition of vision primitives such as corners or, alternatively, operator input of veridical example data
3. a clear distinction between calibration and operation phases.

The algorithm we have introduced requires none of these, though it should be noted that it is unlikely to match the precision of other techniques. Instead, by reducing prerequisites, the algorithm becomes useful in situations where others are simply inapplicable (e.g., when more than one parameter changes, when there is skew, when *all* camera parameters are unavailable or unforeseeable).

### Implications

This method provides a new interpretation of Hebbian learning (Hebb 1949) in which inputs (neurons) with behavioural proximity migrate or extrude efferent processes towards each other, rather than the traditional interpretation in which neurons somehow strengthen an explicit mutual connection.

If the human vision system achieves its spatial organisation and high levels of visual acuity via self-organising means similar to those described, then we would expect to see large differences between mature adult vision and neonate vision. This is because the method depends on continuous environmentally meaningful, temporally varying visual input, and this is not available to us until we are born.

Evidence from tests of infants reveals that they do indeed exhibit inferior visual acuity and inferior contrast sensitivity, just as we would expect. Development of visual acuity is rapid until approximately six months, but continues until twelve months, by which time normally developing infants achieve 20/20 vision (see Figure 6). Contrast sensitivity improves at a similar rate.

Several neurophysiological studies demonstrate that labile neuronal level mappings are quite common. Merzenich & Kass (1982) showed how cortical real estate recently made vacant is co-opted by remaining afferent lines, whose coherent convergence in new areas of the cortex can only be explained by the similarity of the signals they convey, not a pre-established developmental arrangement.

The patterning of ocular dominance columns discovered by Hubel & Wiesel (1977) is another example, where inputs from anatomically quite separated origins (the two eyes) seek out common destinations in the cortex, and actually terminate in areas of less than 0.4 mm diameter.

Moreover, their later work with Stryker (Hubel, Wiesel & Stryker 1978) showed that within this geometrical arrangement there exist still finer convergences, those of the line orientation preference neurons.

Can a single neurophysiological process explain all of these examples of convergence of signals carrying similar data? And can the same process by used to explain the obvious homuncularity of the somatosensory cortex, discovered by Penfield & Jasper (1954)? We feel that our process could well explain *how* such mappings develop, and also gives a

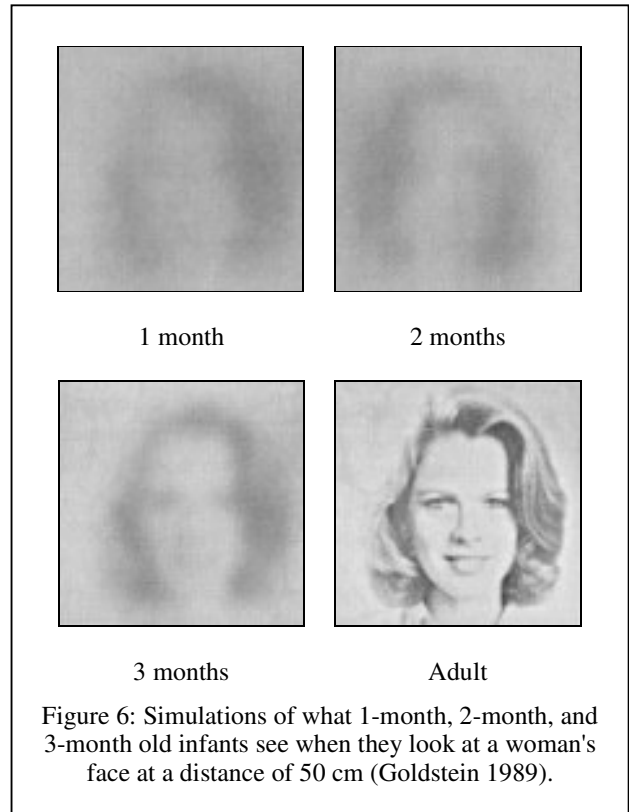


Figure 6: Simulations of what 1-month, 2-month, and 3-month old infants see when they look at a woman's face at a distance of 50 cm (Goldstein 1989).

clue as to *why*. Many researchers have looked at the brain from the standpoint of computational and physiological economy in various forms, both spatial and temporal. Dong & Atick (1995) concentrate on the potential usefulness of inferred decorrelation functions in the lateral geniculate nucleus, and use time-varying images as a way of explicating the statistics of the processes they infer. We showed how similar decorrelation could bypass massive computational problems in motion tracking (Peters & Sowmya 1996). Michison (1991) speculated that there are evolutionary constraints forcing economy in the wiring of the cortex. Such constraints would naturally cause neurons carrying similar signals to converge rather than extend axons through the finite space of the brain.

The advantage of convergence according to behavioural proximity is that if two signals are normally so similar that one can be taken to imply the other, it is considerably more economical to arrange them so that they are literally, phenomenologically and neuro-anatomically, next door neighbours, rather than attempting to construct an explicit relationship or rule which expresses their near-identity.

Such a process could have originated far back in the early days of evolutionary development of the central nervous system, as an economical organising principle, yet now be useful at any level of abstraction where a strong implication relation needs to be physically instantiated.

### Discussion

The methods described are used to organise a set of one-dimensional inputs into a two-dimensional arrangement.

However, there is no limit on the dimensionality of the problem or its solution. Similar techniques could be used to organise inputs in arrangements having three or more dimensions, or to organise inputs with paired values, or tuples.

Both traditional connectionist and symbolic approaches have been applied to similar problems. They both share the shortcoming of attempting to represent complex information in explicit form, which leads to vast computational demands. A connectionist approach is to explicitly represent relationships between inputs by weights. This becomes impractical when dealing with even quite constrained vision problems involving only, say, 128 by 96 pixels. The number of weights generated for an input array of this size is 150,994,944. The symbolic approach also has great difficulty in representing relationships between large numbers of inputs. Its forte is in later stages when visual data has been condensed and summarised in a form amenable to symbolic representation and

logical manipulation, but it is not practical to apply a symbolic approach to the raw data in order to get to this position.

By representing environmental relationships implicitly, in a non-orthogonal arrangement the data requirement is no longer explosive (i.e., only 12,288 for 128 by 96). The complexity is reduced from order  $N^2$  to order  $N$ .

In summary, we demonstrate a robust, adaptive, self-organising system with minimal dependencies, practical usefulness in robotics, strong biological plausibility and explanatory power, computational economy, and potentially broad application due to its context independence.

### Acknowledgements

In composing these ideas we have benefitted much from discussions with Jim Franklin, Christian Killin, Tim Lambert, Arthur Ramer, Mark Reid, and Charles Willock.

### References

- Dong, D. W., & Atick, J. J. (1995). Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 6, 2, 159-178.
- Goldstein, E. B. (1989). *Sensation and Perception* (3<sup>rd</sup> ed.). Pacific Grove: Brooks/Cole Publishing.
- Hebb, D. O. (1949). *The organization of behavior: a neurophysiological theory*. New York: Wiley.
- Hubel, D. H., & Wiesel, T. N. (1977). Functional architecture of Macaque monkey visual cortex. *Proceedings of the Royal Society of London, Series B*, 198, 1-59.
- Hubel, D. H., Wiesel, T. N., & Stryker, M. P. (1978) Anatomical demonstration of orientation columns in macaque monkey. *Journal of Comparative Neurology*, 177, 361-380.
- Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, March, 105-117.
- Merzenich, M. M., & Kaas, J. H. (1982). Reorganisation of mammalian somatosensory cortex following peripheral nerve injury. *Trends in Neuroscience*, 5, 434-436.
- Michison, G. (1991) Neuronal branching patterns and the economy of cortical wiring. *Proceedings of the Royal Society of London, Series B*, 245, 151-158.
- Penfield, W., & Jasper, H. H. (1954). *Epilepsy and the functional anatomy of the human brain*. Boston: Little, Brown.
- Peters, M. W. (1998). Towards artificial forms of intelligence, creativity, and surprise. *Proceedings of the Twentieth Meeting of the Cognitive Science Society, 1998*. Madison, Wisconsin.
- Peters, M. W., & Sowmya, A. (1996). Active vision and adaptive learning. *Proceedings of Intelligent Robots and Computer Vision XV* (pp. 413-424). Boston: SPIE Volume 2904.
- Peters, M. W., & Sowmya, A. (1997). WRAITH: ringing the changes in a changing world. *Proceedings of the Fourth Conference of the Australasian Cognitive Science Society*. Newcastle, Australia.
- Peters, M. W., & Sowmya, A. (1998a). A real-time variable sampling technique for active vision: DIEM. *Proceedings of the International Conference on Pattern Recognition, 1998*. Brisbane, Australia.
- Peters, M. W., & Sowmya, A. (1998b). Autonomous multi-domain attention control. *Proceedings of the Pacific Rim International Conference on Artificial Intelligence, 1998*. Singapore.
- Pollefeys, M., & Van Gool, L. (1997). A stratified approach to metric self-calibration. *Proceedings of Computer Vision and Pattern Recognition, 1997*.
- Pollefeys, M., Koch, R., & Van Gool, L. (1997). Self-calibration and metric reconstruction in spite of varying and unknown camera parameters.
- Prokopowicz, P. N., (1994). *The development of perceptual integration across eye movements in visual robots*. Doctoral dissertation, Intelligent Perception and Action Laboratory, Northwestern University.
- Prokopowicz, P. N., & Cooper, P. R. (1995). The Dynamic Retina: contrast and motion detection for active vision. *International Journal of Computer Vision*, 16, 191-204.
- Sharma, R., & Srinivasa, N. (1996). Saccade control for an active vision stereo head using a learned spatial representation. *Proceedings of the 1996 IEEE International Symposium on Intelligent Control* (pp. 91-96).
- Sparks, D. L., Lee, D., & Rohrer, W. H. (1990). Population coding of the direction, amplitude, and velocity of saccadic eye movements in the superior colliculus. *Proceedings of the Cold Spring Harbor Symposia on Quantitative Biology, The Brain*. 55 (pp. 805-811).
- van der Smagt, P. (1995). *Visual robot arm guidance using neural networks*. Doctoral dissertation, Facultiet Wiskunde en Informatica, University of Amsterdam.