Learning Spatio-Temporal Invariances

James V Stone*
Cognitive and Computing Sciences, University of Sussex, UK.
jims@uk.ac.sussex.cogs

Abstract

We present a neural network model for the *unsupervised* learning of high order visual invariances. The model is demonstrated on the problem of estimating *sub-pixel* stereo disparity from a temporal sequence of *unprocessed* image pairs. After learning on a given image sequence, the model's ability to detect sub-pixel disparity generalises, *without additional learning*, to image pairs from other sequences.

1 Introduction

The ability to learn high order visual invariances[1] - surface orientation, curvature, depth, texture, and motion - is a prerequisite for the more familiar tasks (e.g. object recognition, obstacle avoidance) associated with biological vision. This paper addresses the question: What strategies enable neurons to learn these invariances from a spatio-temporal sequence of images, without the aid of an external teacher?

The model to be presented is, in certain respects, similar to the IMAX models[2, 1, 12]. Unfortunately, the IMAX models suffer from several drawbacks. The IMAX merit function has a high proportion of poor local optima. In [2] this problem was ameliorated by using a hand crafted and biologically implausible weight-sharing architecture which which could not be used for other problem domains. In [1] temporally related input vectors were learned in a shift-invariant manner, and the "tendency to become trapped in poor local optima" (p367) was addressed by introducing a user-defined regularisation parameter to prevent weights from becoming too large. The IMAX models require storage of unit outputs over the entire training set, whereas a biologically plausible model should only use quantities that can be computed on-line. The stored outputs are required to evaluate the IMAX merit function and its derivative. This computationally expensive process requires large amounts of CPU time, which increases with the cube of the number of independent parameters implicit in the input data[12].

Although the model described in this paper is substantially different from the IMAX models, it shares with them a common assumption: Generic solutions to problems of modelling perception are derivable from an analysis of the types of

*The author is a joint member of the Schools of Biological Sciences, and Cognitive and Computing Sciences at the University of Sussex.

[1]The term invariance, as it commonly used in the literature, is somewhat misleading. The terms "invariance" and "parameter" are used here to refer to perceptually salient properties (e.g. depth).

BMVC 1994 doi:10.5244/C.8.67

spatial and temporal changes immanent in the structure of the physical world (see [11]). That is, a learning mechanism can discover high order parameters by taking advantage of quite general properties (such as spatial and temporal smoothness) of the physical world. These properties are not peculiar to any single physical environment so that such a mechanism should be able to extract a variety of high order parameters (e.g. size, position, 3D orientation and shape) via different sensory modalities (vision, speech, touch), and in a range of physical environments.

2 Learning Via Spatio-Temporal Constraints

Consider a sequence of images of an oriented, planar, textured surface which is moving relative to a fixed camera (see Figure 2). Between two consecutive image frames the distance to the surface changes by a small amount. Simultaneous with this small change in surface depth, a relatively large change in the intensity of individual pixels occurs. For example, a one-pixel shift in camera position can dramatically alter the intensity of image pixels, yet the corresponding change in the depth of an imaged surface is usually small. Thus *there is a difference between the rate of change of the intensity profile of an image and the corresponding rate of change of parameters associated with the imaged surface.* A high order parameter is therefore characterised by change over time, but the rate of this change is small, relative to that of the intensity of image pixels.

How can this characterisation of high order parameters be utilised? It is possible to constrain the outputs of a model so that the learning process gives rise to outputs which posses the general characteristics of a high order parameter. *An 'economical' way for a model to generate such a set of outputs is to adapt its connection weights so that the outputs specify some high order parameter implicit in the model's inputs.* That is, it is possible to place quite general constraints on the outputs, such that the 'easiest' way for a model to satisfy these constraints is to compute the value of a high order parameter. Such constraints determine neither which particular parameter should be computed, nor the output value for any given input. Instead, *they specify only that particular types of relations must hold between successive outputs.*

The rate of change of the output of a model can be measured in terms of the 'temporally local', or *short term*, variance associated with a sequence of output values. If it is to reflect the value of a high order parameter then the output value has to vary, and vary *smoothly* over time. Thus its short term variance should be small, relative to its *long term* variance.

3 The Learning Method

The general strategy just described can be implemented using a multi-layer neural network model with a single linear output unit u. The output of u at each time t is z_t. The cumulative output $\tilde{z}_t$ of u is a temporal exponentially weighted sum of outputs z_t. We can obtain the desired behaviour in z by altering the connection weights such that z has a large long-term variance V, and a small short-term variance U. Thus, maximising V/U maximises the variance of z over a long interval,

whilst simultaneously minimising its variance over relatively short intervals.

These requirements can be embodied in a merit function F, which can then be maximised with respect to the inter-unit connection weights of the model. The output of an output unit is $z_t = \sum_j w_{ij}\, y_j$, where w_{ij} is the value of a weighted connection from the jth to the ith unit, and y_j is the output of the jth unit. The merit function F is defined as:

$$F = log\, \frac{V}{U} = log\, \frac{\sum_{t=1}^{T} (\overline{z} - z_t)^2}{\sum_{t=1}^{T} (\tilde{z}_t - z_t)^2} \tag{1}$$

Where V is the variance of z, and U is the short-term variance of z. The mean output is of a unit is $\overline{z}$. The cumulative output $\tilde{z}_t$ of a unit is an exponentially weighted sum of outputs z over a period ϕ preceding t:

$$\tilde{z}_t = \sum_{\tau=t-\phi}^{t-1} \lambda^{t-\tau} z_\tau \tag{2}$$

The 'integral' of λ has unit area, $\sum_{\tau=t-\phi}^{t-1} \lambda^{t-\tau} = 1$, where $0 \leq \lambda \leq 1$. The learning algorithm consists of computing the derivative of F with respect to every weight in the model, and using a conjugate gradient method to locate a maximum in F. This method requires every input (image pair) to be presented twice at the input layer of units, for each learning iteration (line search). The required derivatives of F with respect to weights between successive layers of units are computed using the chain rule (but not the learning method) described in [8]. The value of λ was set such that the half-life h of λ was 32 time steps, and $\phi = 4h$. The initial weights were set to random values between ± 0.3.

In contrast to the IMAX models discussed in the introduction, the model does not require the states of all output units over the entire set of inputs to be stored. Instead, the function F, and the derivatives of U and V, can be computed incrementally at each time step. Thus the quantities required to alter weights (in order to maximise F) can be computed on-line.

3.1 Model Architecture

As shown in Figure 1, the model consists of three layers of units. Every unit in each layer is fully connected to every unit in the following layer. The first layer consists of 20 linear input units, arranged in two rows of 10 units. The second layer consists of 10 units, each of which has an output $y = tanh(X)$, where X is the total input to a second-layer unit from units in the input layer. The input to the ith unit is $X_i = \sum_j (w_{ij} x_j + \theta_j)$, where w_{ij} is the value of a weighted connection from the jth to the ith unit, and x_j is the output of the jth input unit. All and only units in the second layer have a bias weight θ from a unit with constant output of 1. This bias weight is adapted in the same way as all other weights in the model. The output layer consists of a single linear unit, as described in the previous section.

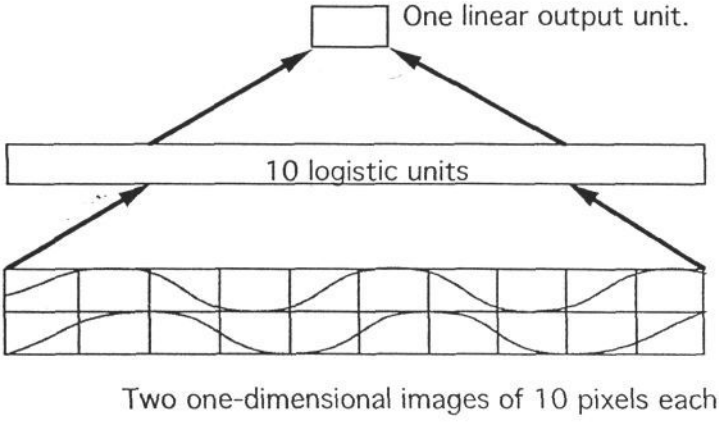

Figure 1: The neural network model architecture.

3.2 The Input Data

The input consisted of a temporal sequence of stereo images with *sub-pixel disparities*. The sequence was derived from a planar surface which was both translating and oscillating in depth in a sinusoidal manner (see Figure 2). This moving surface was used to generate an ordered set of 1000 stereo pairs of one dimensional 10-pixel images. The disparity of successive images in this set varied sinusoidally between ± 1 according to the depth of the imaged surface. The sinusoid had a period of 1000 time steps to correspond with the set of 1000 image pairs.

The grey-level of surface elements was chosen randomly from the range (0,1). The surface grey-levels were smoothed (using a Gaussian with a standard deviation of 10 surface elements), and normalised to have zero mean and unit variance.

At each time step t (see Figure 2), each image pair was constructed by sub-sampling the intensity values on the moving surface. The grey-level of each image pixel was derived from the mean of 10 contiguous surface elements. This allows image pairs with sub-pixel disparities to be generated. For example, if members of a pair sample from surface regions which are separated by one surface element (=0.1 of a pixel width) then the images have a disparity of 0.1 pixels. For disparities in the range ± 1 this means that image pairs with a total of 21 disparities can be obtained. Note that adjacent pixel grey-levels were de.ived from adjacent, non-overlapping surface regions.

In order to simulate the translation of the surface, the first image $I1_t$ of a pair was moved along the surface by 20 surface elements (=2 image pixels) at each time step. The second image $I2$ of a pair was aligned with $I1$, and then shifted along the surface according to the current disparity value.

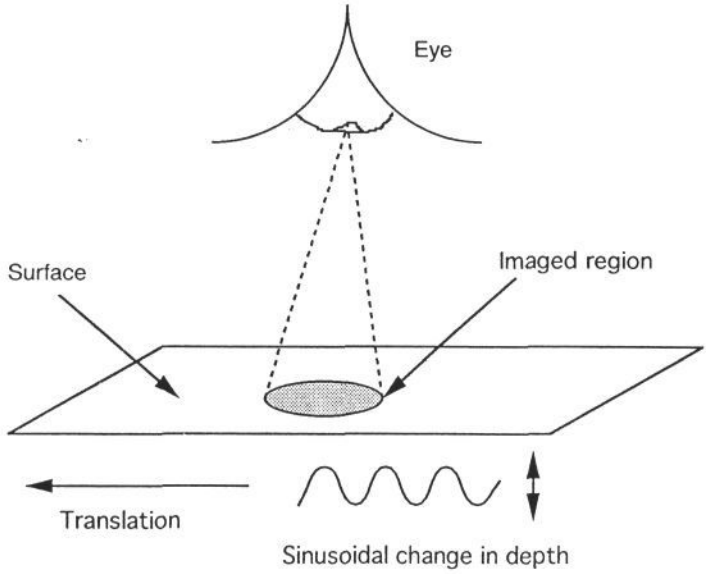

Figure 2: Schematic diagram of how surface moves over time. The surface depth varies sinusoidally as the surface translates at a constant velocity parallel to the image.

4 Results

The model was tested on stereo pairs of images (see Figures 1..3). The system converges reliably, giving correlation magnitudes between output z and disparity of not less than 0.9. After 100 conjugate gradient iterations the correlation r between the output z and disparity was $r = -0.957$ (see Figure 3). The effect of altering the number of units in the middle layer was minimal. Similar results were obtained with as few as four units in the middle layer.

Generalisation: If the model has learned disparity (and not some spurious correlate of disparity) then it should generalise to new stereo sequences, *without any learning of these sequences.* Accordingly, the model was tested with a sequence consisting of 1000 stereo pairs. These were obtained from a new surface constructed in the same manner as was used for the original data set. During testing, the disparity varied sinusoidally between ± 1 as before. However, rather that deriving consecutive image pai s from neighbouring surface regions, each image pair was derived from a random point on the surface. This tested the ability of the system to estimate disparity independently of the particular grey-level profiles in each image pair. Using these data the correlation was $r = -0.916$.

Note that the rate at which disparity varies has no effect on this test correlation, because the output z does not depend upon previous values of z. Thus the model has not only learned to detect disparity of a single data set. Nor has it learned disparity which varies at only one sinusoidal frequency.

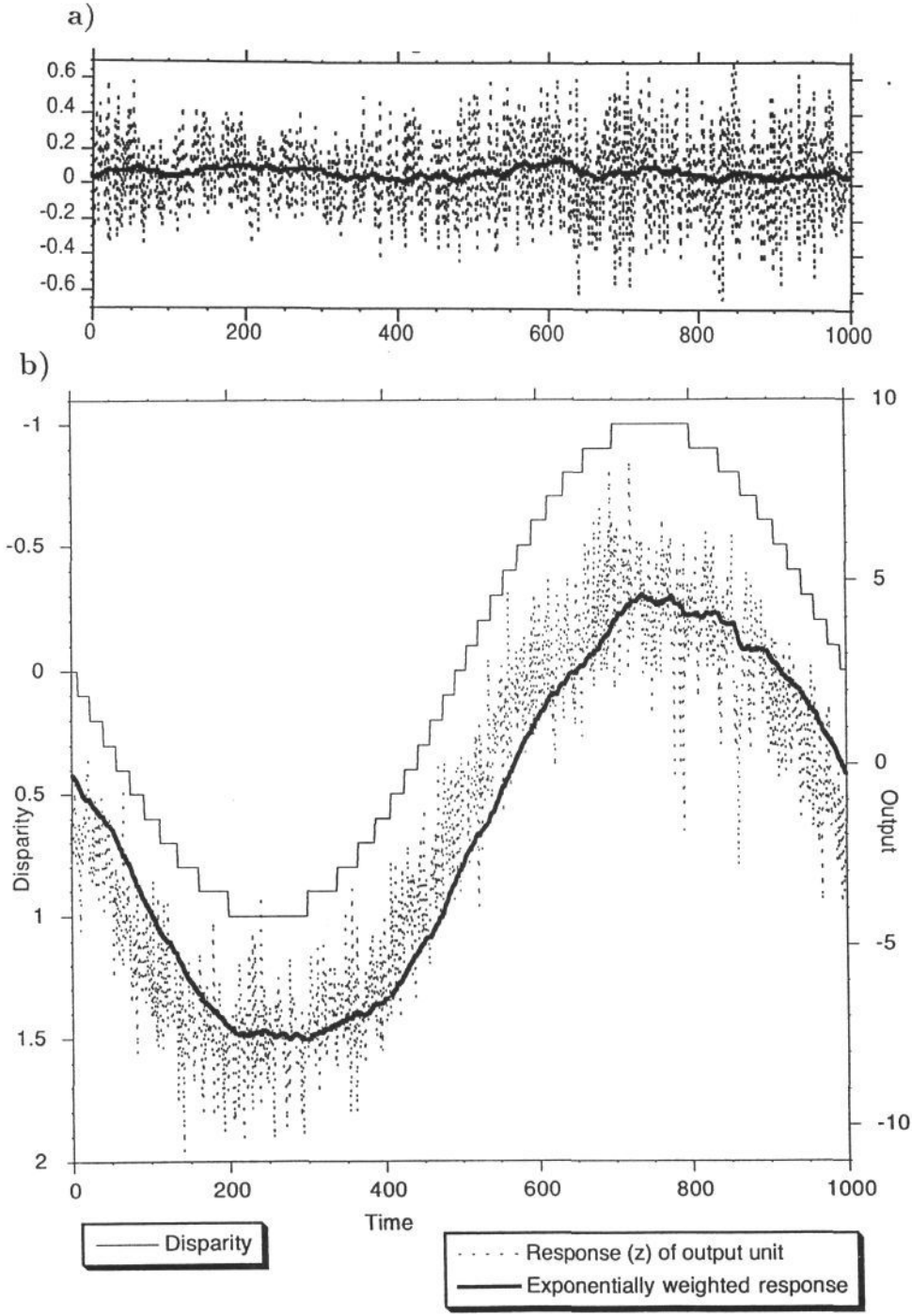

Figure 3: Graphs of time t versus output z (a) before and (b) after learning.

Convergence: The magnitude of the correlation between unit output and disparity was $|r| < 0.9$ only for simulations with a half-life $h < 2$ time steps (see Figure 4). For $h \geq 2$ the rate of convergence, defined as $1/$(the number of iterations required such that $|r| > 0.5$), was proportional to h^2, where h is the half-life of λ.

These results indicate that the function F has a very low proportion poor local maxima. This, in turn, suggests that the 'energy landscape' defined by F is relatively smooth, allowing it to be traversed by simple search techniques. Simulations using simple gradient ascent produce comparable results in about 500 iterations. More importantly, it suggests that maxima are reliably associated with model weight values which enable the detection of high order parameters of inputs.

Unit Receptive Fields: An analysis of the response characteristics of units revealed the following observations. Units in the middle layer have stable outputs only over a small range of disparities. This is consistent with the response properties of disparity-sensitive neurons in the primary visual cortex with narrow tuning profiles[7]. In contrast, the output unit has a stable response over every ('small') interval within a wide range of disparities (see Figure 3), and may correspond to a neuron "with extended and reciprocal excitatory and inhibitory responses"[7](p.749).

5 The Frustration of Learning

Within each simulation, the value of of the half-life h of λ is constant. However, if the value of the half-life h is decreased then the cumulative output $\tilde{z}_t$ of u becomes an increasingly good approximation to the output z_{t-1}. Recall that the derivative of U with respect to each weight is used to reduce $(\tilde{z}_t - z_t)^2$. Adjusting the model's weights so as to reduce the difference d_t between z_t and $\tilde{z}_t \approx z_{t-1}$ creates a form of frustration [6]. That is, the benefits of decreasing d_t are cancelled by the costs of *increasing* corresponding differences $d_\tau (\tau \neq t)$ of other times. For small values of h, each consecutive pair of outputs attempts to reduce its difference via changes in the model weights, without regard to the effects on the differences of pairs at other times. The 'temporal myopia' associated with small h values obstructs the discovery of a set of weights which minimises d for all times. For larger values of h, the difference between z_t and z_{t-1} may not be reduced in *every* learning iteration (because now $\tilde{z}_t \not\approx z_{t-1}$). That is, for large h, the effect of minimising U may be to *increase* d_t, if this enables many $d_\tau : t \neq t$ for other times (and therefore U) to be decreased. This increase in d_t could correspond to a step in the wrong direction (downhill) on an energy function associated with a lower temperature. Thus, a high h value permits local extrema associated with low temperatures to be avoided. By using a large h value, the final differences between z_t and z_{t-1} are usually smaller than would be obtained in a frustrated system (i.e. with a small value of h).

The parameter h acts somewhat like an *annealing parameter*, smoothing out local maxima in F at high values of h. As h approaches infinity so U approaches V, and therefore $V/U \approx 1$ for any weight values. As in [5, 4, 9], at high 'temperatures' the energy function defined by F is convex, and there exists a single maximum. As the temperature (h) is reduced, the energy function becomes increasingly non-convex. Each of a series of decreasing temperatures is associated

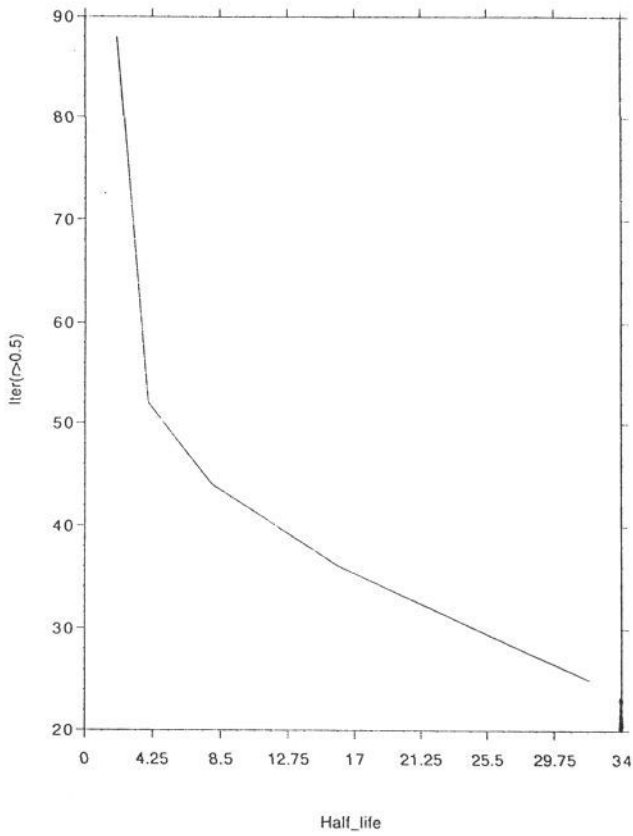

Figure 4: Graph of half-life h versus number of iterations k required to obtain $|r| > 0.5$. A graph of $k(|r| > 0.5)$ versus h^{-2} yields a linear fit with a correlation of 0.993 for $h = \{2, 4, 8, 16, 32\}$.

with an increasingly non-convex energy function. But, in each case, the maximum of the current energy function can be used as the starting point for the search for the maximum of the next function (associated with the new, lower temperature). This annealing method has not been empirically tested, and, within each simulation reported here, the value of h is constant.

6 Discussion

The model discovers high order invariances by computing a non-linear function of components of each input vector. Informally, this often amounts to evaluating *relations* between sets of components of each input image.

The stereo disparity task learned is a hyper-acuity task. That is, the amount of disparity is smaller than the width of any single receptor (pixel). Specifically, 19 of the 21 disparities were less than one pixel. Members of a stereo pair which have a sub-pixel disparity differ in terms of the local slope and curvature of their intensity profiles, and not necessarily in terms of the positions of the peaks and troughs in these profiles. Thus, detecting disparities of less than one pixel requires more than the construction of a pixel-to-pixel correspondence between members of a stereo pair. It requires comparisons of relations-between-pixels in one image with relations-between-pixels in the other image. The resultant meta-relation that specifies the amount of disparity in each pair is, in statistical terms, of a high order.

It is widely accepted that learning relies upon the temporal proximity of learned events. Classical conditioning only occurs if the conditioned stimulus is presented within a small interval before the unconditioned stimulus; reinforcement learning typically relies upon events which are temporally proximal. There is evidence that the 'body maps' in the sensory cortex can be modified by altering the temporal relations between inputs from adjacent sensory regions[3]. Given the importance of time in these different types of learning, it does not seem unreasonable to suggest that time also plays a critical role in the learning of perceptually salient parameters.

As has been demonstrated here, learning to extract perceptual parameters from continuously changing visual inputs relies upon the temporal continuity of parameter values. For example, a sequence of images of a rotating cup can generate rapid changes in pixel values, but the orientation, curvature and depth of the imaged surfaces usually changes relatively slowly over time. Given such a sequence, any learning system that did not take advantage of the temporal continuity of invariances would be discarding a powerful and general heuristic for discovering important properties of the physical world.

7 Conclusion

Conventional low-level computer vision techniques rely upon the assumption that a parameter value is invariant over some region of space(see [10]). The model described in this paper assumes that useful parameters vary relatively slowly over time. When presented with a sequence of images, the model discovered precisely those parameters which describe the behaviour of the imaged surface through time.

The model may lend itself to the self-organised construction of hierarchical systems, in which successive layers compute increasingly higher order parameters, with the highest layers performing object recognition.

Acknowledgements: Thanks to R Lister, S Isard, T Collett, A Bray, J Budd and C North for comments on drafts of this paper, and to Harry Barrow for useful discussions. This research was supported by a Joint Council Initiative grant awarded to J Stone, T Collett and D Willshaw.

References

[1] S Becker. Learning to categorize objects using temporal coherence. *Neural Information Processing Systems*, pages 361–368, 1992.

[2] S Becker and GE Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 335:161–163, 1992.

[3] SA Clark, A Allard, WM Jenkins, and MM Merzenich. Recpetive fields in the body-surface map in adult cortex defined by temporally correlated inputs. *Nature*, 332:444–445, March 1988.

[4] R Durbin and D Willshaw. An analogue approach to the travelling salesman problem using an elastic net method. *Nature*, 326(6114):689–691, 1987.

[5] JJ Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proc Nat Ac Sci*, 81:3088–3092, 1984.

[6] S Kirkpatrick, CD Gelat, and MP Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[7] GF Poggio. Cortical neural mechanisms of stereopsis studied with dynamic random-dot stereograms. *Cold Spring Harbour Symposia on Quantitative Biology*, LV:749–756, 1990.

[8] DE Rumelhart, GE Hinton, and RJ Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[9] JV Stone. The optimal elastic net: Finding solutions to the travelling salesman problem. *ICANN92*, pages 170–174, 1992.

[10] JV Stone. Shape from local and global analysis of texture. *Phil. Trans. Roy. Soc. Lond.(B)*, 339(1287):53–65, January 1992.

[11] JV Stone. Computer vision: What is the object? In *Prospects for AI, Proc. Artificial Intelligence and Simulation of Behaviour, Birmingham, England. IOS Press, Amsterdam.*, pages 199–208, April 1993.

[12] RS Zemel and GE Hinton. Discovering viewpoint invariant relationships that characterize objects. *Technical Report, Dept. of Computer Science, University of Toronto, Toronto, ONT MS5 1A4, 1991*, 1991.