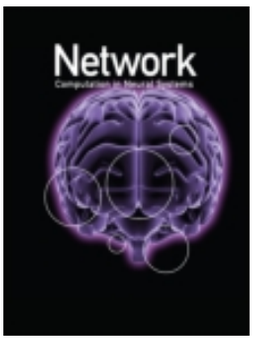

Network: Computation in Neural Systems

ISSN: 0954-898X (Print) 1361-6536 (Online) Journal homepage: https://www.tandfonline.com/loi/inet20

A forward-inverse optics model of reciprocal connections between visual cortical areas

Mitsuo Kawato, Hideki Hayakawa & Toshio Inui

To cite this article: Mitsuo Kawato, Hideki Hayakawa & Toshio Inui (1993) A forward-inverse optics model of reciprocal connections between visual cortical areas, Network: Computation in Neural Systems, 4:4, 415-422, DOI: 10.1088/0954-898X_4_4_001

To link to this article: https://doi.org/10.1088/0954-898X_4_4_001

Published online: 09 Jul 2009.

Network 4 (1993) 415–422. Printed in the UK

LETTER TO THE EDITOR

A forward-inverse optics model of reciprocal connections between visual cortical areas

Mitsuo Kawato†‡ , Hideki Hayakawa† and Toshio Inui§

† ATR Human Information Processing Research Laboratories, Kyoto 619-02, Japan
‡ Laboratory of Parallel Distributed Processing, Research Institute for Electronic Science, Hokkaido University, Sapporo, Hokkaido 060, Japan
§ Laboratory for Psychology, Faculty of Literature, Kyoto University, Kyoto 606, Japan

Received 11 October 1993

Abstract. We propose that the feedforward connection from the lower visual cortical area to the higher visual cortical area provides an approximated inverse model of the imaging process (optics), while the backprojection connection from the higher area to the lower area provides a forward model of the optics. By mathematical analysis and computer simulation, we show that a small number of relaxation computations circulating this forward-inverse optics hierarchy achieves fast and reliable integration of vision modules, and therefore might resolve the following problems. (i) How are parallel visual modules (multiple visual cortical areas) integrated to allow a coherent scene perception? (ii) How can ill-posed vision problems be solved by the brain within several hundreds of milliseconds?

Recent findings on multiple visual cortical areas (van Essen *et al* 1990) which represent distinct visual cues such as colour, motion and shape, and their parallel organization all the way through from the retina to the visual association cortices (Hubel and Livingstone 1987) pose a difficult computational problem: how are parallel visual modules (Julesz 1971) integrated to allow a coherent scene perception within a short time?

Visual images I are generated when light rays reflected from 3D objects in the visual world hit a 2D image sensor such as the retina, CCD or film. The imaging process R, which we call 'optics', compresses 3D objects into 2D images and thus loses information; hence a many-to-one mapping. Consequently, the early vision problems which estimate different aspects S of the geometrical structure in the 3D world from 2D images cannot be properly solved unless some constraints are given beforehand (Marr 1982, Poggio *et al* 1985) because they are one-to-many mappings. That is, the early vision problems are each computationally characterized as an inverse process of optics and *a priori* knowledge about the visual world is introduced as the constraint required. Accordingly, in many computational vision algorithms, the following sum J of two objective functions is minimized to find the best visual-world representation S which explains the image data I as well as satisfies the *a priori* knowledge (Ballard *et al* 1983, Poggio *et al* 1985):

$$J = \|R(S) - I\|^2 + \|Q(S)\|^2 \tag{1}$$

where the first term requires that the reconstruction of the image $R(S)$ from the representation S using the optics operator R be compatible with the real data I, and the second term imposes the *a priori* knowledge about the visual world, such as smoothness of the representation. Minimization is especially difficult when R or Q is strongly nonlinear;

0954-898X/93/040415+08$07.50

it can, however, be done by a kind of steepest descent method: the stochastic relaxation algorithm (Geman and Geman 1984, Poggio *et al* 1985) or its recurrent neural network (mean-field approximation) version (Koch *et al* 1986). However, a large number of iterations (usually more than a few hundred) is required, and no explanation exists for the typical visual information processing time in humans (100–400 ms) (Potter 1976, Inui and Miyamoto 1981) . Thus, hitherto, recurrent neural network models have been rejected as fast visual computational mechanisms (Marr and Poggio 1979, Thorpe and Imbert 1989, Rolls 1989).

In the present letter, we propose a hierarchical computational model for interaction between visual cortical areas which solves the above two problems. First, a fundamental hierarchical model is presented and several testable predictions are also made. Then a comprehensive model of interactions between visual areas is presented. As an example, the shape-from-shading problem is simulated using the proposed model.

Patterns of anatomical connections within visually-related areas must form the structural basis for solving these difficulties (Zeki and Shipp 1988). A hierarchical flow of connections is characterized by a specific organization of a laminar origin and termination of reciprocal cortico-cortical connections (Pandya and Yeterian 1988) (figure 1B). Rostrally-directed feedforward connections originate mainly from neurons in layer III and terminate in and around layer IV of the higher areas. In contrast, caudally-directed (backprojection) connections originate in layers V and VI and, to a lesser extent, in layer IIIa and terminate mainly in layer I. From single-cell recordings, it is shown that a high-level description S of the three-dimensional world is represented in higher visual cortices, e.g. the colour of an object irrespective of the illumination (V4), and the motion of an object as opposed to component motions (MT), (Movshon et al 1986). An intermediate representation $\hat{S}$ between abstract representation S and image I is represented in the primary visual cortex (V1). That is, the activity in V1 is more directly correlated with the raw image data.

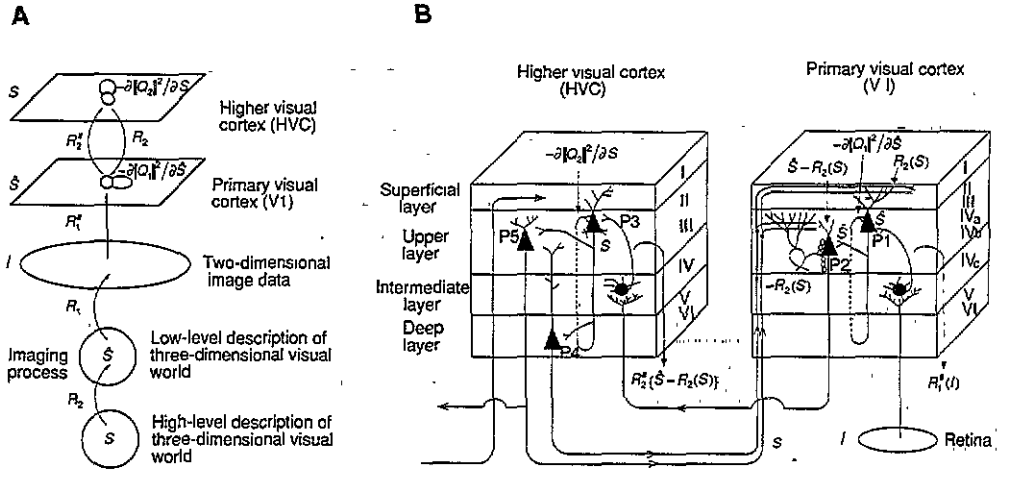

Figure 1. Fundamental forward-inverse optics model. **(A)** Model for reciprocal interactions between V1 and the higher visual cortex (HVC). In the lower half of the figure, the optics operation R in the outer world is decomposed into a lower and a higher part, R_1 and R_2. A model of this hierarchy in the brain is shown in the upper half of the figure. **(B)** Layered-neural-circuit model of the hierarchical interaction between V1 and HVC. Filled neurons are excitatory and a hollow neuron is inhibitory.

We propose that the backprojection connections provide a forward model of the optics

process, while the feedforward connections between the two areas provide an approximated inverse model of that optics process. The simple forward-inverse optics model of figure 1 will be extended to the more realistic model shown in figure 2. Although there exists no unique inverse optics operation, by taking account of the two terms in equation (1) to some extent, it is always possible to derive some approximate inverse optics operations, which compute a rough estimation of S from the image data I by one-shot calculation, in the form of feedforward neural connections. See the appendix for a derivation using linearization and, for examples, see Kersten *et al* (1987), Wang *et al* (1989) and Hurlbert and Poggio (1988). These inverse optics computations are only approximately valid under very restricted conditions; thus, if only they were to be used, the brain would not be able to generally execute correct vision computations. On the other hand, such computations by themselves can solve simple and easy vision tasks.

When new image data impinges on the retina, a rough estimate of the higher representation is first calculated by using only the feedforward connections. This higher representation is then transformed back to an intermediate representation by the backprojection connections, and then compared with the current representation in V1 to calculate the error. The error is filtered by the approximated inverse operation and sent again to the higher visual cortex to modify the higher representation. Intrinsic connections within V1 and higher visual cortices (Gilbert and Wiesel 1983) make the estimates more compatible with *a priori* knowledge about the structures in the 3D visual world. Figure 1B gives a detailed explanation of the proposed relaxation computation based on the known laminar structures.

P1 neurons in the upper layer of V1 receive three kinds of synaptic inputs, from layer IVc, via recurrent intrinsic connections (broken curve) and from HVC. Their states $\hat{S}$ change according to the following dynamics:

$$\mathrm{d}\hat{S}(t)/\mathrm{d}t = R_1^{\sharp}(I) - \hat{S} - \partial\|Q_1(\hat{S})\|^2/\partial\hat{S} + R_2(S) - \hat{S}. \tag{2}$$

The image data impinging on the retina is transformed into $R_1^{\sharp}(I)$ by the retinal circuit, the lateral geniculate nucleus, and the synaptic weights in layer IVc. The intrinsic recurrent connections produce the third term $-\partial\|Q_1(\hat{S})\|^2/\partial\hat{S}$ which renders $\hat{S}$ compatible with the *a priori* knowledge about $\hat{S}$. P1 neurons receive synaptic inputs $R_2(S)$ by the backprojection neural connections from the pyramidal cells in the deep (P4) and upper layers (P5) of the HVC. States S of the P3 neurons in the upper layer of the HVC change according to the following dynamics:

$$\mathrm{d}S(t)/\mathrm{d}t = R_2^{\sharp}\{\hat{S} - R_2(S)\} - \partial\|Q_2(S)\|^2/\partial S. \tag{3}$$

P2 neurons in the upper layer of V1 calculate $\hat{S} - R_2(S)$ based on two kinds of inputs: $\hat{S}$ from P1, and feedback input $-R_2(S)$ from HVC via inhibitory interneurons. The feedforward signal calculated by P2 is then transformed into the synaptic input $R_2^{\sharp}\{\hat{S} - R_2(S)\}$ to P3 in the upper layer of HVC by stellate cells in the intermediate layer of HVC. $-\partial\|Q_2(S)\|^2/\partial S$, which implements the *a priori* knowledge about S, is calculated by the intrinsic recurrent loop within HVC, shown by the broken curve. P4 and P5 in HVC receive synaptic inputs regarding S from P3, and send them back to the superficial layer of V1 by the backprojection connections.

The proposed computation finds the maximum *a posteriori* estimation of the 3D world, which best accounts for the image data based on the *a priori* knowledge. In a quite general situation, we can theoretically guarantee convergence to the (local) optimal solution (Kawato *et al* 1991, Hayakawa *et al* 1992, Wada and Kawato 1993). The required iteration

number is dramatically reduced and the relaxation reaches its equilibrium during the inter-saccade period because the initial guess calculated by the approximate inverse optics is much better than no initial guess. Although several previous theories (Grossberg and Mingolla 1985, Fukushima 1986, Carpenter 1987, Harth *et al* 1987, Thorpe and Imbert 1989, Rolls 1989, Mumford 1992, Finkel and Sajda 1992) postulate similar functional roles to the forward optics model for backprojection connections, the present theory is unique in ascribing an approximate inverse optics model to feedforward connections and introducing forward-inverse optics into the relaxation calculation. They are essential to solve the above difficulties.

We give some circumstantial evidence and testable predictions. First, anatomical studies demonstrated that feedback connections in V1 contact inhibitory neurons (see a hollow neuron in V1 of figure 1B) as well as excitatory neurons (Johnson and Burkhalter 1991), in the manner supporting the required connection circuit in figure 1B. Second, in the steady state, even if the layer III pyramidal neurons in the higher visual cortex fire vigorously, layer IV stellate neurons should be silent in the latter half of the inter-saccadic interval. For they do not receive inputs because in V1 there is no discrepancy between the intermediate representation $\hat{S}$ and the reconstructed representation $R(S)$. This counter-intuitive prediction has indirect support from statistical analysis of the temporal firing patterns of neurons in IT. Richmond and Optican (1987) showed that all of the neurons analysed had principal components that were either phasic or tonic in response to two-dimensional patterns. This suggests that analysed neurons receive major inputs either from layer IV stellate neurons or the pyramidal neurons. The prediction could be directly tested by laminar markings of recorded cells. Finally, temporal patterns of firing must be quite different between simple images and complicated images which are ambiguous or difficult to interpret; e.g. a simple square with uniform colour and luminance as opposed to overlapping transparent squares. The receptive field of examined neurons should be chosen in the midst of the pattern. Stellate neurons must stop firing 150–200 ms after stimulus presentation in the former case due to 'filling-in' between pyramidal neurons, but not in the latter case.

Fast and reliable integration of vision modules can be achieved in the forward-inverse architecture shown in figure 2 containing multiple visual areas, multiple hierarchical levels (e.g. V1 $\Longleftrightarrow$ V2 $\Longleftrightarrow$ V4 $\Longleftrightarrow$ IT) and parallel information flows. The model also encompasses previous experimental and conceptual studies, as well as our speculative working hypotheses of representations by different visual areas and the functions of their connections. Whether these hypotheses are well supported by experimental data or not is not critical to our theory. What is essential is that somewhat different representations of the visual world are encoded by different areas and each area cannot determine its output unless all connected areas provide information to it.

A major difficulty in module-integration is that each representation in a different area cannot be simultaneously or independently estimated without knowledge of other representations. In most of the previous models providing integration of different vision modules, integration occurs after the calculation of each module, leading to combinatorial explosion problems and illusatory conjunction problems (Ballard *et al* 1983, Malsburg 1988). Untying the intermingled relationship in this chicken-and-egg problem requires the modular hierarchical structure of distinct modules and the usage of forward and approximate inverse optics. That is, each module initially estimates a rough and approximate representation based on default assumptions of other modules' outputs through approximated inverse optics. Then, the backprojection connections transform abstract and distinct representations in higher visual cortices into image-like representations at the lower visual cortex, where they are compared for the purpose of inconsistency detection. This error

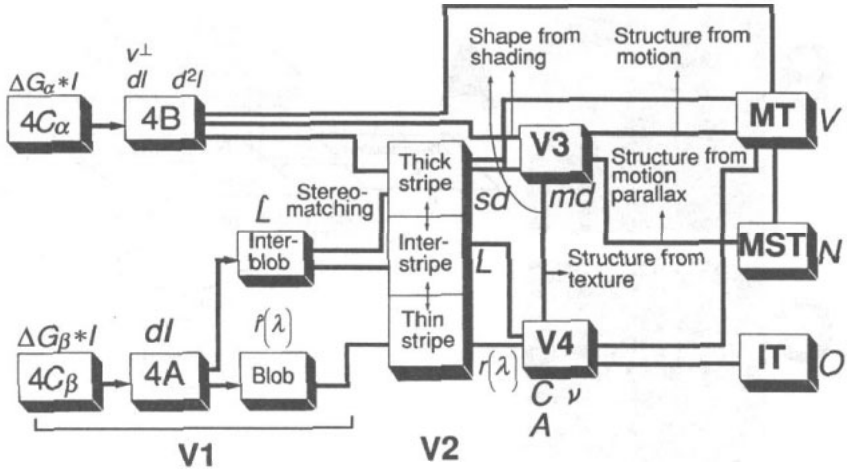

Figure 2. The forward-inverse optics model adapted to a parallel and hierarchical structure of visual cortices. We do not literally propose that each area represents purely a physically identifiable quantity. Connections without arrows show reciprocal neural connections. Intensive integration of colour, stereo, shape and motion could take place by using discontinuity (Poggio *et al* 1988) possibly represented in interstripes of V2. Definitions of symbols are as follows. $\Delta G*I$: convolution integral of the image with the Laplacian Gaussian function. dI and d^2I: first and second derivatives of the image along with specific directions. $v^{\perp}$: local velocity component in the direction with the maximum change of image intensity. *sd*: surface depth calculated from stereo disparity. $r(\lambda)$: reflectance of points on the visible surface of a light of wavelength λ. *L*: discontinuities such as occluding contours and junctions of different objects. *md*: depth and orientation of the visible surface calculated by various monocular cues. ν: location of the light source and its wavelength distribution. *C*: 3D locations of objects segregated by *L*. *A*: various attributes of a distinct object such as colour and texture. *V*: velocity vector representing translation and rotation of objects. *N*: velocity vectors of the body, head and eyes of the observer. *O*: memorized images of 3D objects.

is again filtered by the approximate inverse optics and several modules' outputs are made more compatible.

As an illustrative example, we selected the so-called shape-from-shading problem where one needs to guess the three-dimensional structure of an image from shading information only. The task is to simultaneously estimate the surface orientation, discontinuity in 3D objects such as occluding contours or edges, and light source direction only from image intensity data such as that in figure 3B. The model in figure 3A is comprised of three modules, for surface orientation (middle: *N* module), discontinuity detection (left: *L* module), and light source direction (right: ν module), which might schematically correspond to V3, V2, and V4 in figure 2, respectively. The following *a priori* knowledge about the visual world was used for each vision module: surface orientation is smooth other than on ridgelines in the *N* module, discontinuity (like ridgelines) is continuous (Geman and Geman 1984) in the *L* module, and light source is uniform for the whole image in the ν module. Approximated inverse optics computations in each module are as follows: the surface orientation is constrained orthogonal to the detected discontinuity within the image plane and the tilt estimation is constrained to its previous value (*N*); discontinuity is simply detected from the image intensity gradient (*L*); and light source direction is estimated from a blurred image (ν).

Within 20 iterations, which is biologically plausible, the model estimated the ridge lines and occluding contours (figure 3C), the light source direction (figure 3D arrow) and the

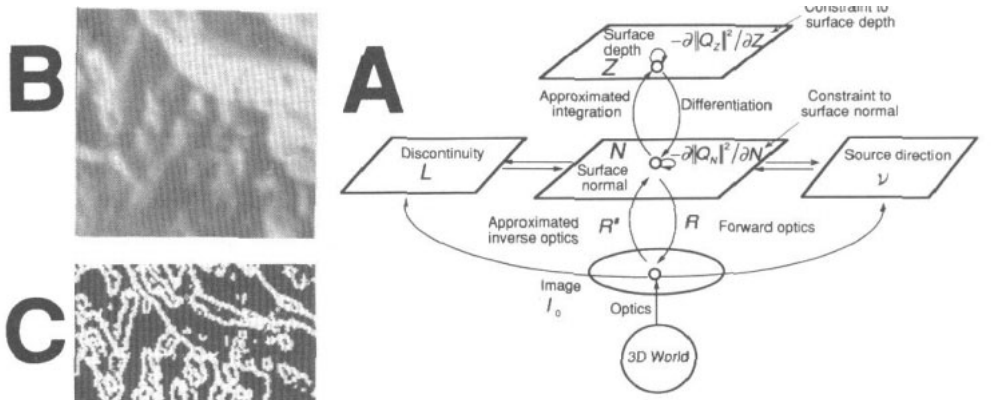

Figure 3. Simulation results of shape-from-shading problem. **(A)** Forward-inverse optics model for simultaneous estimation of surface orientation, discontinuity location and direction, and light source direction in the shape-from-shading problem. *The model is comprised of three* modules: the discontinuity module (left), the surface orientation module (middle), and the light source direction module (right). **(B)** A natural intensity image of a mountain landscape. **(C)** Detected discontinuity locations. **(D)** Reconstructed intensity data with the estimated direction of the light source (arrow). **(E)** Depth map reconstructed from the surface orientation shown as a grey level (brighter: higher; darker: lower). **(F)** Depth map reconstructed from the surface orientation shown as a perspective view.

depth (figures 3E, F) from the intensity data (figure 3B) for a natural mountain landscape image. The reconstructed intensity image (D) was fairly close to the given intensity data (B) and the estimated surface was quite smooth (F) other than at discontinuity locations (C). We also confirmed that surface orientation, discontinuity locations, and light source direction were accurately estimated for synthesized images of spheres, ellipsoids and polyhedrons (Hayakawa *et al* 1992). Furthermore, the model was much faster and much more stable than conventional computer vision algorithms (Horn 1977). The forward-inverse optics architecture, therefore, under quite general conditions, can simultaneously estimate different representations of the 3D world within a small number of iterations by integrating different vision modules.

This work is supported by Human Frontier Science Project Grants to Mitsuo Kawato and Toshio Inui.

Appendix

For simplicity, we show that a linear approximated inverse optics operator $R^\sharp$ is obtained by linearizing the original nonlinear objective function (1) around a particular representation S_0. Let us define $S = S_0 + \Delta S$. Then, equation (1) can be approximated to the first order of ΔS as follows:

$$J \cong \|R'(S_0)\Delta S - \{I - R(S_0)\}\|^2 + \|Q'(S_0)\Delta S + Q(S_0)\|^2 \tag{4}$$

where R' and Q' are derivatives of R and S. Then, S which gives the minimum of the modified objective function can be calculated by a simple linear, one-shot calculation as follows:

$$\begin{aligned} R^\sharp(I) &\equiv S_0 + \Delta S \\ &= S_0 + \{R'(S_0)^T R'(S_0) + Q'(S_0)^T Q'(S_0)\}^\dagger \\ &\quad \times \{R'(S_0)^T (I - R(S_0)) - Q'(S_0)^T Q(S_0)\} \end{aligned} \tag{5}$$

where † denotes the Moore–Penrose pseudo inverse matrix.

References

[1] Ballard D H, Hinton G E and Sejnowski T J 1983 Parallel visual computation *Nature* **306** 21–6
[2] Carpenter G and Grossberg S 1987 A massively parallel architecture for a self-organizing neural pattern recognition machine *Comp. Vision Graphics Image Proc.* **37** 54–115
[3] Finkel L H and Sajda P 1992 Object discrimination on depth-from-occlusion *Neural Computation* **4** 901–21
[4] Fukushima K 1986 A neural network model for selective attention in visual pattern recognition *Biol. Cybern.* **55** 5–15
[5] Geman S and Geman D 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images *IEEE Trans. Pattern Analysis Machine Intelligence* **PAMT-6** 721–41
[6] Gilbert C D and Wiesel T N 1983 Clustered intrinsic connections in cat visual cortex *J. Neurosci.* **3** 1116–33
[7] Grossberg S and Mingolla E 1985 Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading *Psychol. Rev.* **92** 173–211
[8] Harth E, Unnikrishnan K P and Pandya A S 1987 The inversion of sensory processing by feedback pathways: A model of visual cognitive functions *Science* **237** 184–7
[9] Hayakawa H, Nishida S and Kawato M 1992 Shape from shading using discontinuity in image *ATR Technical Report* **TR-A-0148** (Kyoto: ATR)
[10] Horn B K P 1977 Understanding image intensities *Artificial Intelligence* **8** 201–31

[11] Hubel D H and Livingstone M S 1987 Segregation of form, color and stereopsis in primate area 18 *J. Neurosci.* **7** 3378–415
[12] Hurlbert A and Poggio T 1988 Synthesizing a color algorithm from examples *Science* **239** 482–5
[13] Inui T and Miyamoto K 1981 The time needed to judge the order of a meaningful string of pictures *J. Exp. Psychol.: Human Learning & Memory* **7** 393–6
[14] Johnson R R and Burkhalter A 1991 Feedback connections in visual cortex contact inhibitory neurons *Soc. Neurosci. Abstr.* **17** 844
[15] Julesz B 1971 *Foundations of Cyclopean Perception* (Chicago: University of Chicago Press)
[16] Kawato M, Inui T, Hongo S and Hayakawa H 1991 Computational theory and neural network models of interaction between visual cortical areas *ATR Technical Report* **TR-A-0105** (Kyoto: ATR)
[17] Kersten D, O'Toole A, Sereno E, Knil D and Anderson J 1987 Associative learning of scene parameters from images *Appl. Opt.* **26** 4999–5006
[18] Koch C, Marroquin J and Yuille A 1986 Analog 'neural' networks in early vision *Proc. Natl Acad. Sci. USA* **83** 4263–7
[19] Marr D 1982 *Vision* (San Francisco: Freeman)
[20] Marr D and Poggio T 1979 A computational theory of human stereo vision *Proc. R. Soc. London* B **204** 301–28
[21] Movshon J A, Adelson E H, Gizzi M S and Newsome W T 1986 The analysis of moving visual patterns *Pattern Recognition Mechanisms* ed C Chagas, R Gattass and C Gross (New York: Springer) pp 117–51
[22] Mumford D 1992 On the computational architecture of the neocortex. II. The role of cortico-cortical loops *Biol. Cybern.* **66** 241–51
[23] Pandya D N and Yeterian E H 1988 Architecture and connections of cortical association areas *Cerebral Cortex, 4, Association and Auditory Cortices* ed A Peters and E G Jones (New York: Plenum) pp 3–61
[24] Poggio T, Torre V and Koch C 1985 Computational vision and regularization theory *Nature* **317** 314–9
[25] Poggio T, Gamble E B and Little J J 1988 Parallel integration of vision modules *Science* **242** 436–40
[26] Potter M C 1976 Short-term conceptual memory for pictures *J. Exp. Psychol.: Human Learning & Memory* **2** 509–22
[27] Richmond B J and Optican L M 1987 Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. II. Quantification of response waveform *J. Neurophysiol.* **57** 147–61
[28] Rolls E T 1989 Functions of neuronal networks in the hippocampus and neocortex in memory *Neural Model of Plasticity: Theoretical and Empirical Approaches* ed J H Byrne and W O Berry (New York: Academic) pp 240–65
[29] Thorpe S J and Imbert M 1989 Biological constraints on connectionist modelling *Connectionism in Perspective* ed R Pfeifer, Z Schreter, F Fogelman-Soulie and L Steels (Amsterdam: North-Holland) pp 63–92
[30] van Essen D C, Felleman D J, DeYoe E A, Olavarria J and Knierim J 1990 Modular and hierarchical organization of extrastriate visual cortex in the macaque monkey *Cold Spring Harbor Symposia on Quantitative Biology* **LV** 679–96
[31] von der Malsburg C 1988 Pattern recognition by labeled graph matching *Neural Networks* **1** 141–8
[32] Wada Y and Kawato M 1993 A neural network model for arm trajectory formation using forward and inverse dynamics models *Neural Networks* **6** in press
[33] Wang H T, Mathur B and Koch C 1989 Computing optical flow in the primate visual system *Neural Computation* **1** 92–103
[34] Zeki S and Shipp S 1988 The functional logic of cortical connections *Nature* **335** 311–7