TR－A－0105

Computational Theory and Neural Network Models of Interaction Between Visual Cortical Areas

Mitsuo KAWATO, Toshio INUI,
Sadayuki HONGO and Hideki HAYAKAWA

1991. 3.22

ATR視聴覚機構研究所

〒619-02 京都府相楽郡精華町乾谷　☎07749-5-1411

ATR Auditory and Visual Perception Research Laboratories

Inuidani, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Telephone: +81-7749-5-1411
Facsimile: +81-7749-5-1408
Telex: 5452-516 ATR J

Computational Theory and Neural Network Models of Interaction Between Visual Cortical Areas

Mitsuo KAWATO, Toshio INUI,
Sadayuki HONGO and Hideki HAYAKAWA

Cognitive Processes Department
ATR Auditory and Visual Perception Research Laboratories,
Kyoto 619-02, Japan

Abstract

We develop a computational theory and a neural network model which coherently explains early, middle and high-level vision problems based on the anatomical structure and physiological functions of the visual cerebral cortices. Our computational theory is based upon a hierarchical and stochastic model of image generation with highly redundant, multiple representations at different description levels. We propose that feedforward neural connections from the lower to the higher visual areas provide approximated inverses of image generation, while feedback neural connections from the higher to the lower areas provide forward models of image generation. We propose a global, hierarchical model of interactions between several visual cortical areas, in which internal representations of the 3-D world in each area are specified. First, the solutions to several visual computational problems, such as boundary detection, motion, color, stereo and the shape from shading problem, are outlined in our general framework. In particular, the shape from shading problem will be dealt with in detail by a concrete neural network model and computer simulations. Second, brightness illusions, the Mach band and Craik O'Brien illusion are simulated by a neural network model based on our general framework with emphasis on the disappearance of the illusions under high contrast conditions. Finally, a learning algorithm called the "cross-covariance learning rule", with which the internal models of the visual world can be acquired in the visual cerebral cortices, is proposed.

Contents

1. Introduction

Marr (1982) postulated that the objective of early vision is to estimate the geometrical structures of visible surfaces in the 3-D world from a 2-dimensional image. During the 1980s, computational studies of vision were advanced considerably. Many were based on the following propositions made by Marr (1982):

1. There are several modules in early vision, each of which is specialized for processing different types of information from the 2D image (i.e., binocular disparity, shading, occluding contour, color, motion parallax and so on). The objective of these modules is to reconstruct the 2·1/2-D sketch which represents depth and/or orientation of visible surfaces within the viewer-centered coordinates.

2. Because the 2-D image data compared with the 3-D world is compressed through imaging process, computation in each module can not be properly solved unless some constraints to possible solutions, in other words, prior knowledge about visible surfaces, are given beforehand. In the brain, such computation is effectively and effortlessly performed by using assumptions such as smoothness of visible surfaces or rigidity of visible 3-D objects.

A mathematical problem is called well-posed when (1) there exist solutions, (2) the solution is unique, and (3) the solution depends on data of the problem continuously. If one or more of these conditions are not satisfied, the problem is called ill-posed. Poggio, Torre and Koch (1985) pointed out that vision can be regarded as an inverse process of the optics, and that it is an ill-posed problem. This inverse problem cannot be solved without additional information. According to Marr's theory, vision modules use some natural constraints. Mathematically speaking, vision modules must find the solution which is most consistent with both given image data and natural constraints. That is, the solution minimizes the cost function which consists of a data fitting term and a constraint term. Poggio, Torre and Koch (1985) coherently formulated many algorithms proposed in computational studies of vision within the framework of this standard regularization theory. Let S denote geometrical structure of a visible surface in the 3-D world, I the image intensity, and R the optics process. We then get:

$$I = R\mathrm{S}. \tag{1.1}$$

In standard regularization theory, the estimation of S is given so that the following objective function is minimized.

$$\|I - R\mathbf{S}\|^2 + \lambda\|P\mathbf{S}\|^2, \tag{1.2}$$

here P is the operator which represents constraints on the geometrical structure of a visible surface. λ is called a regularization parameter and determines the ratio of the first data fitting term and the second term regarding the constraint on geometry. λ is inversely proportional to the signal to noise ratio. For the simplest example, if the original image is degraded by adding white gaussian noise, λ is proportional to the noise variance. When both R and P are linear, then the objective function 1.2 becomes a quadratic form in S. Then, in some cases, one-shot algorithms, which calculate the required S from I using only feedforward computation, can be derived.

The limitation of the standard regularization theory was that discontinuities in the 3-D world could not explicitly be dealt with. Poggio and collaborators (Poggio, Torre & Koch, 1985) extended their theory to deal with the discontinuities by utilizing the coupled Markov random field (MRF) model which was proposed by Geman and Geman (1984).

Geman and Geman (1984) developed the coupled MRF as a prior knowledge in an image restoration problem from a degraded image. They formulated the problem as a maximum a posteriori (MAP) estimate. Let y denote an observed degraded image data, x the original image, and π probability. The MAP estimate is used to find the x which gives the maximum value for the following a posteriori probability under the condition of a given y.

$$\pi(x|y) = \frac{\pi(y|x)\pi(x)}{\pi(y)}. \tag{1.3}$$

This is the Bayes formula.

The MRF model of an image is defined as the model in which the state at each lattice point depends only on the state in its neighborhood. If we adopt this as a model of the original image in the image restoration problem, then the prior distribution $\pi(x)$ becomes the following Gibbsian.

$$\begin{aligned} \pi(x) &= (1/Z)\exp\{-U(x)\}, & (1.4) \\ U(x) &= \sum_{C \in \mathcal{C}} V_C(x). & (1.5) \end{aligned}$$

Here, Z is a normalizing constant called the partition function in statistical physics. In the above equations, the local property of the MRF is reflected in the fact that the total potential energy $U(x)$ is the summation of many local potential energies $V_C(x)$. Where C is the clique, a set of lattice points, every pair of whose distinct lattice points are neighbors. $\mathcal{C}$ is the set of cliques. The potential energy $V_C(x)$ is local in the sense that it depends only on the states of the lattice points within the clique C.

For this MRF model, the posterior distribution also becomes the Gibbsian and the posterior energy can be represented as follows:

$$U_P(x|y) = \sum_s \phi(x_s, y_s) + \sum_{C \in \mathcal{C}} V_C(x), \tag{1.6}$$

here s represents the lattice site. ϕ measures compatibility between the data y and the estimated original image x, and is local, that is, it depends only on (x_s, y_s) at a finite number of sites. The first term is the data energy which represents the discrepancy between the observed data y and the estimated image x. The second term corresponds to the a priori knowledge about the image. The first and the second terms of eq. 1.6 correspond to the first and the second terms of eq. 1.2, respectively. In this formulation the MAP estimation problem is transformed into finding the minimum of eq. 1.6. Hence, the standard regularization theory can be regarded as a special case of a more general MAP estimate (Poggio, Torre & Koch, 1985). Geman and Geman (1984) proposed the stochastic relaxation algorithm which can be used in conjunction with simulated annealing to find the global minimum of the energy.

One of the most outstanding contributions of Geman and Geman's 1984 paper was the introduction of the "line process" which is an imaginary stochastic process representing the object's boundary. By arranging the line process above the "intensity process" which represents the gray scale of the image, a hierarchical, stochastic model of the image was developed. In addition, the line processes may represent a discontinuity of several feature dimensions, including color, brightness, depth, and so on.

It turned out that the line process and the couped MRF are key tools not only for image restoration but also for early vision. In particular, this approach can be applied to integrating information across feature dimensions, that is, integration of different vision modules in middle vision (Poggio, Gamble & Little, 1988). However, the integration scheme in this model was rather simple. For example, the conditional probability of depth discontinuity becomes high when there exists a luminance edge (intensity discontinuity) at the same site.

Recently, algorithms which realize higher visual functions have been developed for artificial neural networks (see for example Rumelhart & McClelland, 1986). These functions include: association, learning and generalization. Furthermore, it is interesting that some kinds of optimization problems can be solved by neural networks. In such networks an energy function can be defined, and one can design the architecture of an analog neural network which minimizes a given energy function (e.g., Hopfield & Tank, 1985). In these models, the stable steady state of the network is the solution which minimizes the objective function.

In the next section, we will give the psychological, physiological and anatomical bases of the coupled MRF as a model of visual cerebral cortices. However, we must note that these experimental data support only such MRF behaviors as filling-in and discontinuity detection, and its local interaction property, but not stochastic calculation itself such as the Gibbs sampler. Actually, it is difficult to imagine that the exact stochastic calculation realized in the MRF is implemented in the brain even though the coupled MRF has mathematical similarities close to those of the Boltzmann machine (Ackley, Hinton & Sejnowski, 1985). However, most fortunately, the recent development of the theory of mean-field approximation of stochastic neural network models (Hopfield, 1982; Hopfield, 1984; Koch, Marroquin & Yuille, 1986; Peterson & Anderson, 1987; Iba, 1989; Hinton, 1989; Yuille, 1990) makes a mean-field approximation to the coupled MRF more biologically plausible. The mean-field-approximation neural network of the MRF has the following distinctive features. First, its dynamics described by ordinary differential equations possesses a Lyapunov function, in other word, an "energy". The energy value decreases as the state of the network changes according to its dynamics. It is guaranteed that the state converges to a "local" minimum of the energy. Second, the network has intrinsic recurrent connections. Third, these connections are local, which guarantees the Markov property. Finally, the neural field is isotropic, that is translation invariant.

In particular, Koch, Marroquin and Yuille (1986) showed that the coupled MRF model can be implemented approximately by an analogue neural network which has a close similarity to the Hopfield model (1984). This is a deterministic recurrent neural network model which can be regarded as the "mean field approximation" of the MRF and stochastic relaxation. In this model, the binary property of the line process is deliberately represented by the sigmoid activation function of artificial neurons.

The computational approach to early and middle vision using the coupled MRF is mathematically transparent and elegant, and very appealing as a parallel processing scheme from the engineering point of view. However, the following problems need to be resolved to further develop the MRF or its neural-network approximation as a computational model of biological visual systems. First, relaxation type neural networks which minimize energy have been rejected as realistic models of the brain because they require a number of iterations, and hence can not explain the brain's relatively fast calculations. Second, it is not apparent how high-level vision problems are managed by local models such as the MRF. For example, global and abstract information in an image needs to be processed to solve pattern recognition problems. However, by definition, the MRF is a local model and hence can not process global and abstract information in images. Third, learning rules of the MRF need to be developed so that internal models can be acquired from image examples without teaching about the 3-D world. Finally, computational theories and algorithms developed so far based on the MRF do not directly correspond to morphological structures of multiple visual cortical areas.

Marr (1982) pointed out that an information processing device (brain) must be understood at the following levels before one can be said to have understood it completely. (i) Computational theory. (ii) Representation and algorithm. (iii) Hardware implementation. Although in determining correct computational models and algorithms Marr himself took great advantages of mutual constraints which come from studies at one level on studies at other levels, successive studies in computational vision seem to place too much emphasis on the independence of the above three levels and pay only slight attention to hardware constraints. In this paper, we develop a computational theory and a neural network model which coherently explains early, middle and high-level vision problems based on the anatomical structures and physiological functions of the visual cerebral cortices.

2. Experimental support for a local model such as the coupled Markov random field model

2.1 Psychological evidence

In this section, we provide psychological evidence which supports the coupled Markov random field model for human visual processing.

The most prominent features of the coupled MRF of Geman and Geman (1984) are the filling-in of the intensity process and the detection of discontinuity by the line process. There are abundant psychological data which indicate the filling-in processes and the detection of discontinuity in the brain. Filling in of static retinal images, ganzfeld experiment results (Hochberg, Triebel &

Seaman, 1951), neon color spreading, and surface perception in sparse random dot stereogram are examples.

For example, we can perceive depth discontinuity when a pair of random-dot stereograms are fused binocularly in the brain. That is, there exists a process which represents the discontinuity explicitly in several different dimensions in the brain. The different dimensions are color, brightness, depth and so on. The other important point is that there exists a filling-in process. For example, we perceive a floating surface for random-dot stereograms. Furthermore, even for sparse random-dot stereograms, of up to 5% dot density, a surface can be perceived (Julesz, 1971, Wurger & Landy, 1988). Julesz and Chang (1976) found that we usually choose only the matches having the smallest disparity for an ambiguous stereogram in which there are many possible ways of matching its center. However, the particular match found can be biased by inserting unambiguously matchable dots at a particular disparity. This is an example of the filling-in process.

From the psychological experiment of the stabilized retinal image, it appears that information for the non-edge region is filled in by interpolation based on information around the edges. When we see a red disc which is stabilized on the retina, against a green background which is not stabilized, the color disappears in few milliseconds. And then the disc also disappears and we perceive only the homogeneous green background. This phenomenon indicates that color and contour cannot be perceived in the stabilized condition. Furthermore, it is suggested that color and luminance information are not transmitted to the higher visual center but are inferred from the information around the edges where the neurons are activated continuously. This process is called the filling-in process (Yarbus, 1967; Gerrits & Vendrik, 1970).

2.2 Physiological and anatomical evidence

In this subsection we present neurophysiological and anatomical support for a local model such as the coupled MRF model for the visual cerebral cortices. We use the word "local" in the sense that any cell with a local receptive field in the visual field has synaptic inputs only from cells which represent neighboring areas in the visual field.

One of the most prominent characteristics of visual areas is that the visual field is topographically represented in the cortex. Consequently, receptive field centers of the two neighboring neurons on the cortex are also close on the visual field. Anatomically, it is known that axons of the intrinsic neural connections within the same area extend typically only 2 to 3 mm horizontally to the cortex surface (Gilbert & Wiesel, 1983). Electrophysiologically, it was revealed by the cross-correlation method that the effective range of horizontal interaction is 2 to 3 mm, and then the interaction strength decreases dramatically with the distance (Ts'o, Gilbert & Wiesel, 1986). Toyama (1988) estimated that the inter-columnar interaction is an order of magnitude weaker than the intra-columnar interaction in V1. Consequently, if we regard the hypercolumn in V1 as a unit element for processing image data, the intrinsic neural connection in V1 can be said to provide only the nearest neighborhood interaction in MRF terminology.

This locality of the internal image model is considered a result of hardware limitation regarding connection numbers rather than a merit. Because the average number of neural connections (synaptic inputs) per neuron is about 1000 in the cerebral cortex and much smaller than the total number of neurons 10^{11}, it would be computationally very inefficient if the long-range image interaction were to be modeled by using only the intrinsic connection.

3. Forward and inverse models of image generation

Modelling of the image generation procedure can be done at many different description levels. At relatively low level, as expressed by Horn's image irradiance equation (Horn, 1977), image intensity can be determined from the depth and orientation of a visible surface, its reflectance and lighting condition. Description at higher levels is also possible. For example, image data could also be determined if the 3-D shapes, locations and velocities of objects arranged in the 3-D world are determined. We propose that a multiple-level description of the image generation procedure is used in the brain as described by the following image generation equation.

$$\begin{aligned} I(\mu, x, y, \lambda, t) &= R(\Delta G * I, dI, d^2I, v^{\perp}, sd, r(\lambda), L, md, \nu, C, A, V, N, O) \\ &= R(s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}) \\ &= R(\mathrm{S}) \end{aligned} \tag{3.1}$$

I represents light intensity of wavelength λ at location (x, y) on the left $(\mu = 0)$ or the right $(\mu = 1)$ retina at time t. The right-hand side of the equation describes image generation procedure by nonlinear function R. Here, $\Delta G * I$ is the convolution integral of the image with the Laplacian Gaussian function. Here, $\Delta G = \nabla^2(1/2\pi\sigma^2)\exp\{-(x^2 + y^2)/2\sigma^2\}$. dI and d^2I are the first and second derivatives of the image along with a specific direction. $v^{\perp}$ is a local velocity component in the direction with the maximum change of image intensity. *sd* is surface depth calculated from stereo disparity. $r(\lambda)$ shows the reflectance of points on the visible surface to the light of wavelength λ. L represents discontinuity such as occluding contours and junctions of different objects. L could be said to correspond to the line process. *md* is the depth and orientation of the visible surface calculated by various monocular cues. ν is the location of the light source and its wavelength distribution. C denotes 3-D locations of objects segregated by L. A represents various attributes of a distinct object such as color and texture. V is the velocity vector representing translation and rotation of objects. N is the velocity vectors of the body, head and eye of the observer. O represents memorized images of 3-D objects.

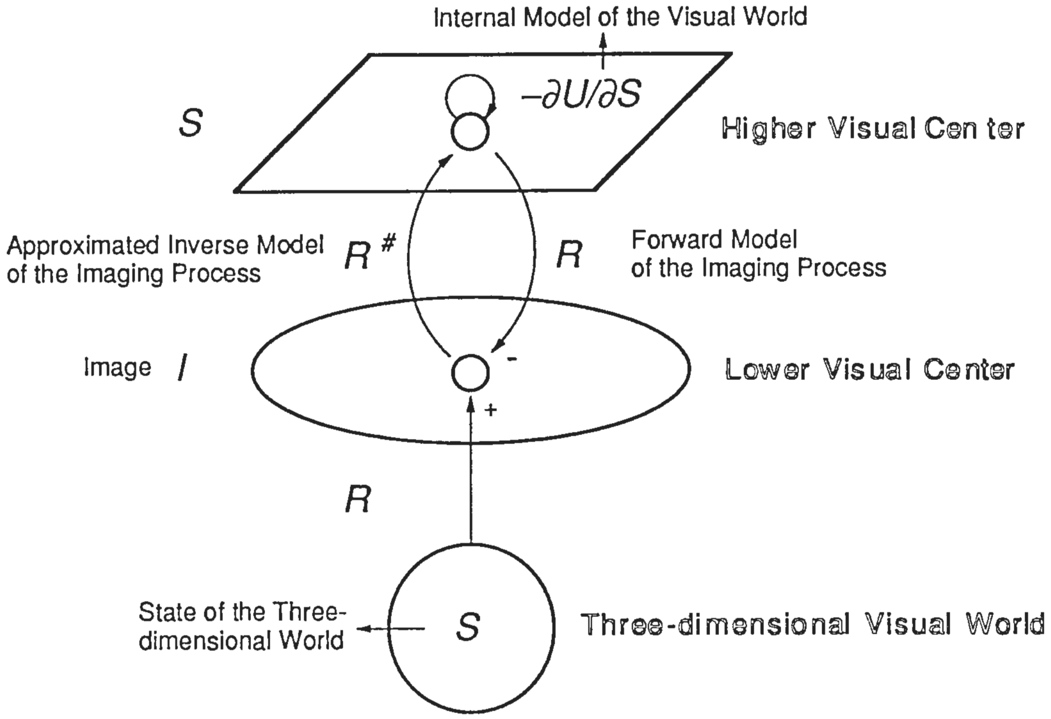
sd, *md*, $r(\lambda)$ and L together provide the 2·1/2-dimensional sketch of Marr. Estimation of V, N and O is high level vision. The estimation procedure from I to S can be called vision.

We here propose a computational model and an algorithm where the visual cerebral cortices solve vision problems based on approximated inverse models of R, forward models of R, and internal models of S. We assume that vision computation is essentially a MAP estimation. Let us denote the occurrence probability of S by $P(\mathrm{S})$, and denote the conditional probability of I for a given S by $P(I|\mathrm{S})$. Because of the local interaction property of the visual cerebral cortices supported by the physiological and psychological evidence described in the previous section, we assume that these two distributions are Gibbs distribution with the corresponding energies $U(\mathrm{S})$ and $U(I|\mathrm{S})$. According to the MAP estimation, the visual cortices find S which minimizes the following *a posteriori* energy.

$$\begin{aligned} U(\mathrm{S}|I) &= U(I|\mathrm{S}) + U(\mathrm{S}) \\ &= 1/2\|R^{\sharp}\{I - R(\mathrm{S})\}\|^2 + U(\mathrm{S}). \end{aligned} \tag{3.2}$$

Here $R^{\sharp}$ is an approximated inverse model of image generation process R. As is well known in early vision, generally there does not exist R^{-1} because inverse optics is an ill-posed problem. However, it is possible to consider $R^{\sharp}$ as an approximated inverse even when R^{-1} does not exist. Actually, many of the one-shot algorithms proposed in the computer vision field (see for example Marr, 1982) can be regarded concrete examples of $R^{\sharp}$. In 3.2, as is usual in Bayes formula, we neglected the constant term $U(I)$ on the left side.

We propose the global neural network model shown in Fig. 3..1 which can minimize the energy of eq. 3.2. This is a deterministic neural-network



Internal Model of the Visual World
$-\partial U/\partial S$
S
Higher Visual Cen ter
Approximated Inverse Model of the Imaging Process
$R^{\#}$
R
Forward Model of the Imaging Process
Image I
Lower Visual Center
-
+
R
State of the Three-dimensional World
S
Three-dimensional Visual World

Figure 3..1: Fundamental model of computational theory for visual cortical areas.

model which can be regarded as the mean-field approximation to the MRF as described in the introduction. Thus, it has the following properties. First, interactions between neurons are local. This implies not only that the intrinsic connections within each layer are local but also that connections between different hierarchical layers are topographic, and have quite restricted divergence and convergence. Second, the network is translation invariant. That is, for example, synaptic connection weights are uniform all over the visual field.

Dynamics of the network can be described as follows.

$$\begin{aligned} \mathbf{S}(0) &= R^{\sharp}(I) && (3.3) \\ d\mathbf{S}(t)/dt &= R^{\sharp}\{I - R(\mathbf{S})\} - \partial U(\mathbf{S})/\partial \mathbf{S} && (3.4) \end{aligned}$$

In Fig. 3..1, 2-dimensional image data I is represented in the lower visual center, and the representation of the visual world $\mathbf{S}$ is manipulated in the higher visual center. The model shown in Fig. 3..1 has mirror symmetry with respect to the lower visual center. If $\partial\{R^{\sharp}R(\mathbf{S})\}/\partial\mathbf{S} = E$ holds approximately, then eq. 3.4 provides the steepest descent method of eq. 3.2. Here E is the identity map.

Another interpretation of the dynamics 3.4 may also be possible, if $R^{\sharp}$ can be regarded as an approximation of the derivative of the inverse of R, that is, $(\partial R/\partial\mathbf{S})^{-1}$. In this case, the dynamics can be regarded as a continuous version of a Newton-like method. Let us illustrate this. First, a function which measures the difference between the real image and the reconstructed image is defined: $F(\mathbf{S}) = I - R(\mathbf{S})$. The problem finding $\mathbf{S}$ compatible with the image data I is equivalent to finding the zero of F. The well known iterative method for this problem, the Newton method, is as follows: $\mathbf{S}_{n+1} = \mathbf{S}_n - F'(\mathbf{S}_n)^{-1}F(\mathbf{S}_n) = \mathbf{S}_n + R'(\mathbf{S}_n)^{-1}\{I - R(\mathbf{S}_n)\}$. However, usually $R'(\mathbf{S})^{-1}$ can not be uniquely determined because of the ill-posedness of vision. A widely used alternative in this case is the following Newton-like method in which an approximation $R^{\sharp}$ of R'^{-1} is used: $\mathbf{S}_{n+1} = \mathbf{S}_n + R^{\sharp}(\mathbf{S}_n)\{I - R(\mathbf{S}_n)\}$. One shot algorithms represented by linear filters (for example Kersten, O'Toole, Sereno, Knil & Anderson, 1987, Hurlbert & Poggio, 1988) might be interpreted as examples of $R'(\mathbf{S})^{-1}$.

In both of the two interpretations, an essential condition of $R^{\sharp}$ is that $R^{\sharp}(0) = 0$. This is automatically met if $R^{\sharp}$ is a linear operator.

It is known that different visual cortical areas are connected both by feedforward and feedback neural connections (Pandya & Yeterian, 1988). The feedback neural connection from the higher visual center to the lower visual center in Fig. 3..1 provides an internal forward model of the imaging process

R. On the other hand, the feedforward neural connection from the lower visual center to the higher visual center provides an internal model of an approximated inverse $R^\sharp$ of R. Furthermore, an intrinsic neural connection within the higher visual center provides an internal model of $\mathbf{S}$ as the gradient $-\partial U(\mathbf{S})/\partial \mathbf{S}$ of the potential energy.

We explain model behavior according to eqs. 3.3 and 3.4 and Fig. 3..1. When a new image I is input, for example after saccade, a rough estimate $R^\sharp(I)$ of $\mathbf{S}$ is calculated in a one-shot manner by the feedforward neural connection. However, this estimate is not the MAP estimate. Following this initial calculation, the model starts relaxation calculation using the loop composed of the feedback and feedforward neural connections. First, estimation of image data $R(\mathbf{S})$ is calculated by the feedback connection from the estimation $\mathbf{S}$ in the higher visual center. Second, this estimation of image data is compared with the real image, and the error $I - R(\mathbf{S})$ is calculated. Third, this error transformed by the feedforward connection into $R^\sharp\{I - R(\mathbf{S})\}$ is fed back to the higher visual center. Fourth, the intrinsic connection in the higher visual center calculates the second term of eq. 3.4.

Relaxation type neural networks which minimize energy have been rejected as a realistic model of the brain (Marr, 1982) because they require a number of iterations and can not explain relatively fast calculation by the brain. However, the proposed model overcomes this shortcoming by one-shot calculation with the feedforward neural connection which gives $R^\sharp$. Because $R^\sharp(I)$ is a much better starting point than 0, the required number of relaxation iterations is much smaller. In other words, even if the computation time is severely limited, the network can obtain a fairly good solution from the MAP standard. Furthermore, if the feedforward connections can be regarded as a linear approximation of the derivative of the image generation, it is well known that the Newton-method is much quicker in convergence than the steepest descent method around the zero point.

4. Local, parallel and hierarchical model of visual cerebral cortices

In the 1980s, many cortical visual areas have been identified in the monkey cerebral cortex (Fig. 4..1a, Kaas, 1986). Figure 4..1b shows connection patterns between different visual areas (see for example Livingstone & Hubel, 1987a, b, Hubel & Livingstone, 1987, Zeki & Shipp, 1988).

For the primary visual area (V1) and the secondary visual area (V2), de-

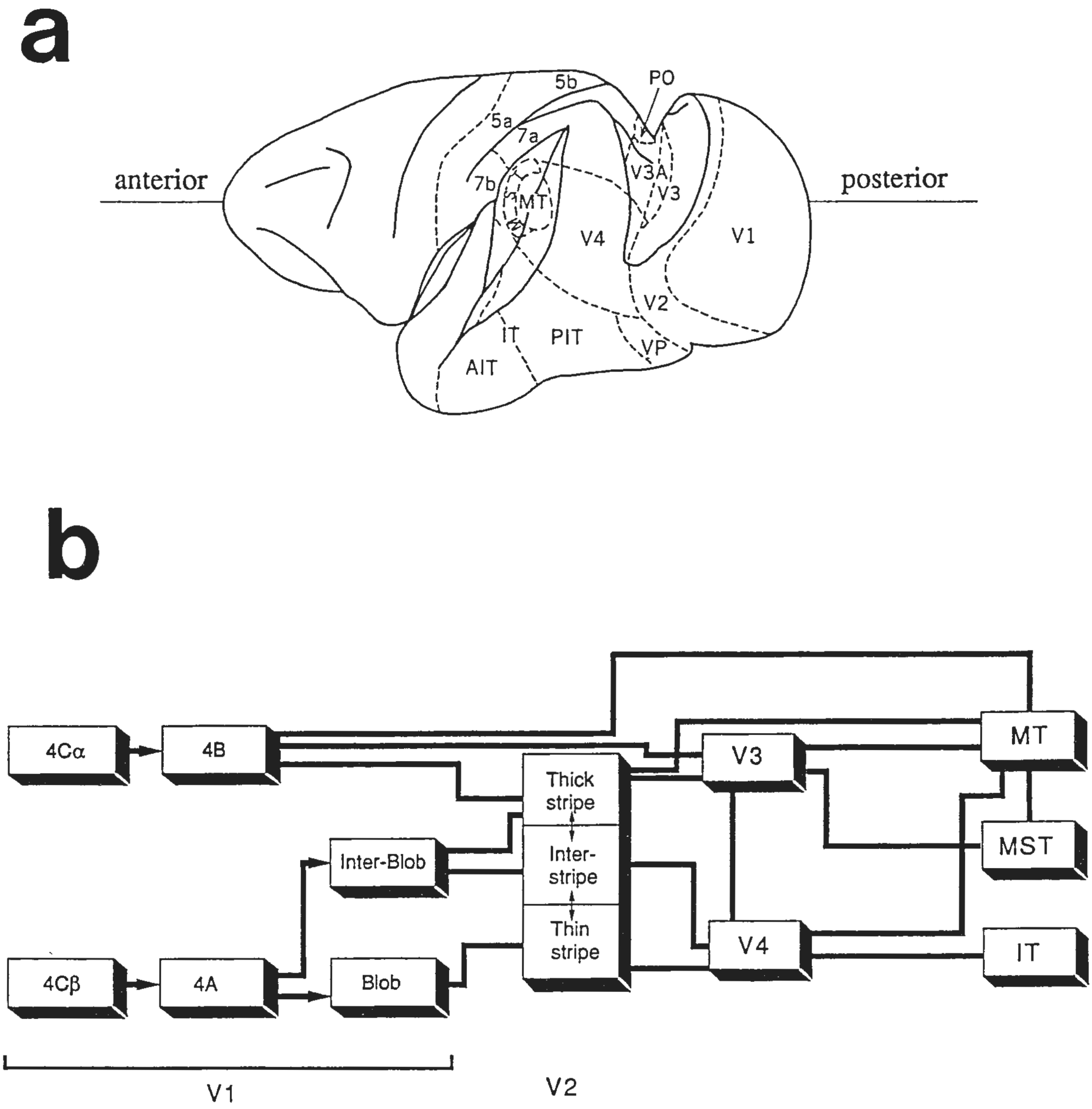


Figure 4..1: a. Visual areas in the Macaque monkey (modified from Kaas, 1986). b. Connection patterns between visual areas.

tails of neural-network structures and functions of subsystems in each area have been intensively studied. A block diagram of the visual cortical areas is shown in Fig. 4..1b. Connections with arrows show feedforward neural projections, which exist in the early part of the visual cortex. On the other hand, connections without arrows show reciprocal neural projections; there are feedforward and feedback interactions among the higher cortical areas.

In the following sections, we will try to answer the following questions based on our computational theory: (1) Why are there many cortical areas (V1, V2, V3, V4, MT, MST, IT) in the primate brain? (2) Why are almost all connections between visual areas reciprocal?

We first briefly illustrate what kinds of representations of the visual world are processed in each visual area, and then provide physiological and anatomical support for this model in the following subsections. The model shown in Fig. 3..1 is quite simplified in the sense that **S** is clamped and many visual areas are also clamped into a single layer. Fig. 4..2 shows in which cortical area each s_i in eq. 3.1 is represented, based on recent knowledge of neurophysiology and anatomy.

The following is only a brief summary of our model. Full accounts of this will be given in the following. $\Delta G * I$ is represented in layers 4Cβ and 4Cα of V1. dI and d^2I are represented in 4A and 4B of V1. $v^{\perp}$ is represented in 4B of V1. sd is represented in the thick stripe of V2. $r(\lambda)$ is represented in the thin stripe of V2. L is represented in the interstripe area of V2. md is represented in V3. ν, C and A are represented in V4. V is represented in MT. N is represented in MST. O is represented in IT. $\Delta G * I$ is represented in both layers 4Cβ and 4Cα. Similarly, dI and d^2I are represented in both 4A and 4B. In these double representations, the former locations have high sensitivity and high temporal resolution. The latter has high spatial resolution.

Parallel and hierarchical structures in Fig. 4..2 reflect the corresponding parallel and hierarchical natures of the image generation process expressed in eq. 3.1. That is, representations $s_1, \cdots, s_{14}$ to be estimated in vision, are roughly classified into three groups: those mainly related to colors, $r(\lambda)$, A and ν, those mainly related to shapes, L, C and O, and those mainly related to motion, $v^{\perp}$, sd, md, V and N. This corresponds to the three parallel, vertical information flows of the anatomical structure of Fig. 4..2.

There is also a striking hierarchy in the anatomical structure shown in Fig. 4..2 which corresponds to redundant representations of the imaging process at different description levels in eq. 3.1. For example, anatomical relationships between pairs of 4B of V1 and MT, interblobs of V1 and interstripes of

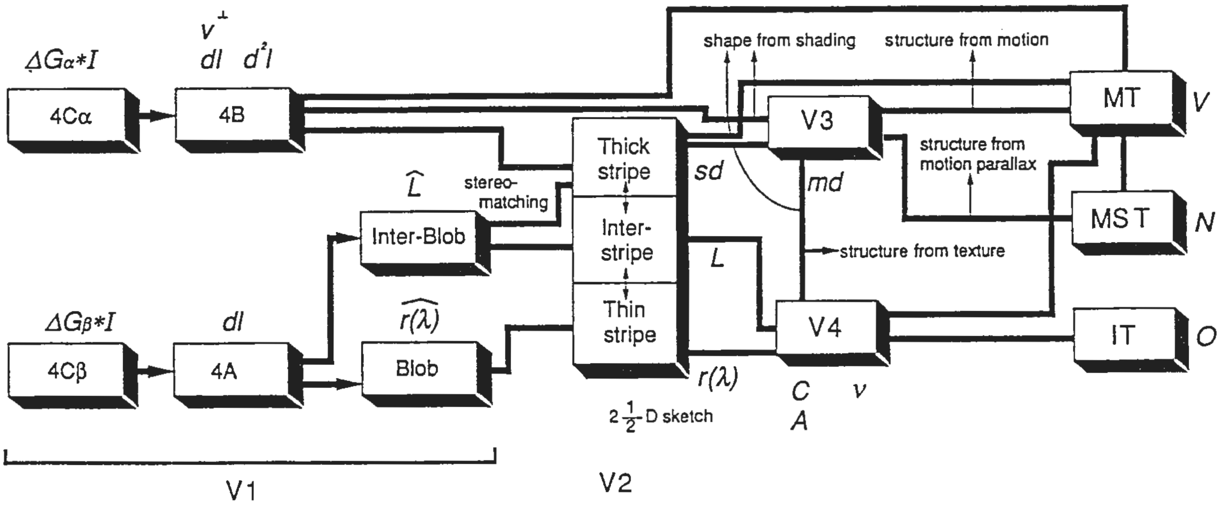


Figure 4..2: Local, parallel and hierarchical model of the visual world composed of the primary and higher visual cortices.

V2, thin stripes of V2 and V4, are hierarchical. A mathematical formulation of this hierarchy will be given in the next section.

4.1 Functions of the visual area 1 (V1)

We first describe physiological and anatomical support for our model regarding the primary visual cortex (V1). It is known that there are two parallel channels in the early visual system, which are called the magnocellular channel and the parvocellular channel, respectively. Magnocellular information from the retina goes through 4Cα to 4B in V1 (Livingstone & Hubel, 1984; Hubel & Livingstone, 1987; Tootel, Hamilton & Switkes, 1988, Hubel & Livingstone, 1990). The majority of the cells in layer 4B are simple (Livingstone & Hubel, 1984). Parvocellular information goes through 4Cβ to 4A and the interblobs and the blobs (Livingstone & Hubel, 1983; Livingstone & Hubel, 1987a, Tootel, Hamilton & Switkes, 1988, Michael, 1988). It is suggested from physiological studies that color and luminance information is processed in the blobs (Livingstone & Hubel, 1987b). Intensity discontinuities are processed at several different scales in 4B and the interblobs. Tootell, Silverman, Hamilton, Switkes and De Valois (1988) showed that neurons in the interblobs generally respond to stimuli with relatively higher spatial frequencies. This is one reason why we propose that the intermediate representation of the discontinuity is represented in the interblobs. Furthermore, there are binocular cells in 4B and in the interblobs (Hubel & Livingstone, 1987; Hubel & Livingstone, 1990). Thus, binocular disparity is also preprocessed in 4B and in the interblobs.

4.2 Functions of the visual area 2 (V2)

There are 3 subsystems in V2; the thick stripes, the thin stripes and the interstripes (Hubel & Livingstone, 1987, Tootell & Hamilton, 1989). These stripe subsystems communicate with each other. Furthermore, there are reciprocal connections between the thick stripes and 4B and the interblobs in V1 (Zeki & Shipp, 1988). The interstripes communicate with the interblobs in V1 and each other (Zeki & Shipp, 1988). From these connections and physiological data, it is suggested that the thick stripes compute the stereo depth (Hubel & Livingstone, 1987; Poggio, Gonzalez & Krause, 1988) from magnocellular (Livingstone & Hubel, 1987a, b) and parvocellular information (Schiller & Logotheris, 1990; Tyler, 1990). We propose that the interstripes compute physical discontinuities in the 3-D world, which do not directly correspond to mere intensity edges. This computation can be done by integrating information from several kinds of visual cues: intensity discontinuity information from the interblobs, color discontinuity from the thin stripes and stereo information from the thick stripes. The thin stripes in V2 compute spectral reflectance (Hubel & Livingstone, 1987). Recently, the neurons which respond to the subjective contour were found in V2 (Peterhans & von der Heydt, 1989; Von der Heydt & Peterhans, 1989). From our model, it is predicted that these neurons reside in the interstripes.

4.3 Functions of the visual area 3 (V3)

There is little decisive data to suggest computational functions of the neurons in V3 (but see Felleman & Van Essen, 1987). But structures of interconnections between V3 and other visual areas suggest our computational scheme for integration between higher order visual areas. There exist interconnections between V3 and V2, V4, MT and MST (Zeki & Shipp, 1988). We propose that V3 is the information processing center of monocular depth where monocular depth is represented, and that this calculation is executed through these four reciprocal connections. It is known that there are several cues for monocular depth perception. Computations of the surface orientation from various cues are called shape from shading, structure from motion, and structure from texture. The shape from shading problem is solved through the connections between 4B in V1 and V3, and between V4 and V3. The simple cells in 4B are supposed to detect the first and second directional derivatives of the intensity. The surface orientation, that is slant and tilt, can be calculated approximately from these derivatives and the illumination orientation and spectra of the light source which should be represented in V4 in our theory (see section 6 for a full description of the shape from shading problem).

In our model (see below) V4 represents the position and spectra of illumination source, configuration among figure parts, and texture. We suppose that structure from motion is calculated through connections between V3 and MT, and that structure from texture is calculated through connections between V3 and V4.

Furthermore, there exist reciprocal connections between V2 and V3 (Zeki & Shipp, 1988). As already mentioned, in our model, binocular depth is represented in V2 thick stripes. Therefore, consistency between monocular depth and binocular depth could be guaranteed through these connections. We believe that the system consisting of V2, V3 and V4, which interact with each other, is a consistency maintaining mechanism (Fig. 4..3). We suppose that the subjective contour is generated by the consistency maintaining mechanism which processes several kinds of visual cues and effects: contrast enhancement, occlusion, monocular depth. The effect of size constancy in the subjective contour demonstrated by Coren (1972) is a good example of the phenomena which should be realized by this consistency maintaining mechanism.

4.4 Functions of the visual area 4 (V4)

From several physiological and anatomical studies, it is supposed that position and spectra of illumination sources, and configuration among parts (especially occlusion), and texture are represented in V4. There are some experimental data which support these hypotheses. The computation of spectral reflectance is supported by the physiological studies in which color Mondrian patterns were used as stimuli (Zeki, 1983). From the physiological study of human vision (Lueck, Zeki, Friston, Deiber, Cope, Cunningham, Lammertsma, Kennard & Frackowiak, 1989), it was revealed that V4 is a center for color perception. On the other hand, some physiological findings suggest that both edge and texture information are processed (Desimone & Schein, 1987, Schein & Desimone, 1990), in addition to color (spectral reflectance) processing. Furthermore, as described below, the receptive field of the V4 neuron has a large suppressive surround. The property of the surround is the same in spectra and spatial frequency as the excitatory center. It was suggested that one of the functions of the neuron is to segregate a figure from a background by color and spatial frequency (Desimone, Schein, Moran & Ungerleider, 1985).

4.5 Functions of MT and MST

It is known that the velocity aperture problem is solved in MT (Movshon, Adelson, Gizzi & Newsome, 1986). There exists experiment data which suggests that the velocity of the observer is calculated in MST. The disparity-dependent direction-selective neurons (Roy & Wurtz, 1990) and expansion/contraction neurons (Tanaka, Fukada & Saito, 1989; Tanaka & Saito, 1989) are found in MST. It is believed that just as the expansion neurons could indicate the forward component of self-motion, the disparity-dependent direction-selective neurons could indicate the horizontal (rightward or leftward) component of self-motion (Roy & Wurtz, 1990).

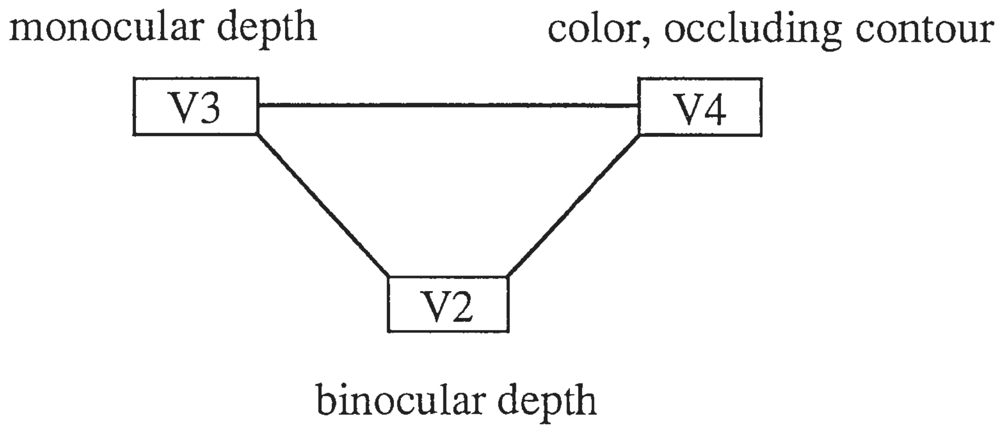


Figure 4..3: Consistency maintaining mechanism consisting of V2, V3, and V4.

4.6 2-and-a-half-dimensional sketch

Marr proposed a 2·1/2-D sketch as an intermediate representation of the depth and orientation of visible surfaces and their discontinuities not within the retinal coordinates but within the viewer-centered coordinates. Therefore, it should be stable irrespective of the eye-movement.

In our model, binocular depth is represented in the thick stripes, monocular depth is represented in V3, and edges are represented in the inter-stripes. Monocular depth is mainly visible surface orientation (and relative order of the surfaces in depth). Recently, it was found that the responses of some neurons in V3A for the visual stimulus are modulated by gaze-direction (Galletti & Battaglini, 1989). Galletti and Battaglini (1989) believed that these neurons play an important role in the transformation from the retina-centered coordinates to the head-centered coordinates. Therefore, the 2·1/2-D sketch corresponds to these representations in the three subsystems. This is very interesting because V2, which is a relatively lower center, plays a major role in expressing the 2·1/2-D sketch, which is considered as the integrative representation of the results of calculation in the vision modules (Fig. 4..4). In our scheme, the representation is the output of the consistency-maintaining mechanism which is constructed by V2, V3, and V4. Probably, it corresponds to the immediate perception of the world.

In the conventional framework shown in Fig. 4..4, the 2·1/2-D sketch is constructed after integration of outputs from several different modules. Thus, the 2·1/2-D sketch is represented in the latter stages of the visual information processing. However, in our model, the 2·1/2-D sketch is represented in a relatively earlier stage of the visual information processing. This is possible because of interactions between multiple cortical areas by feedforward and feedback connections.

4.7 Functions of the inferotemporal cortex (IT)

The inferotemporal cortex receives input information mainly from V4, which plays an important role in pattern perception and recognition. From several computational studies, 2-dimensional representations from several different viewpoints must be represented in IT. These representations could be calculated from the information represented in V4 as mentioned above. IT neurons participate in the short-term retention of visual features for behavioral use (Fuster, 1990).

According to Marr's theory, the representation for the object recognition

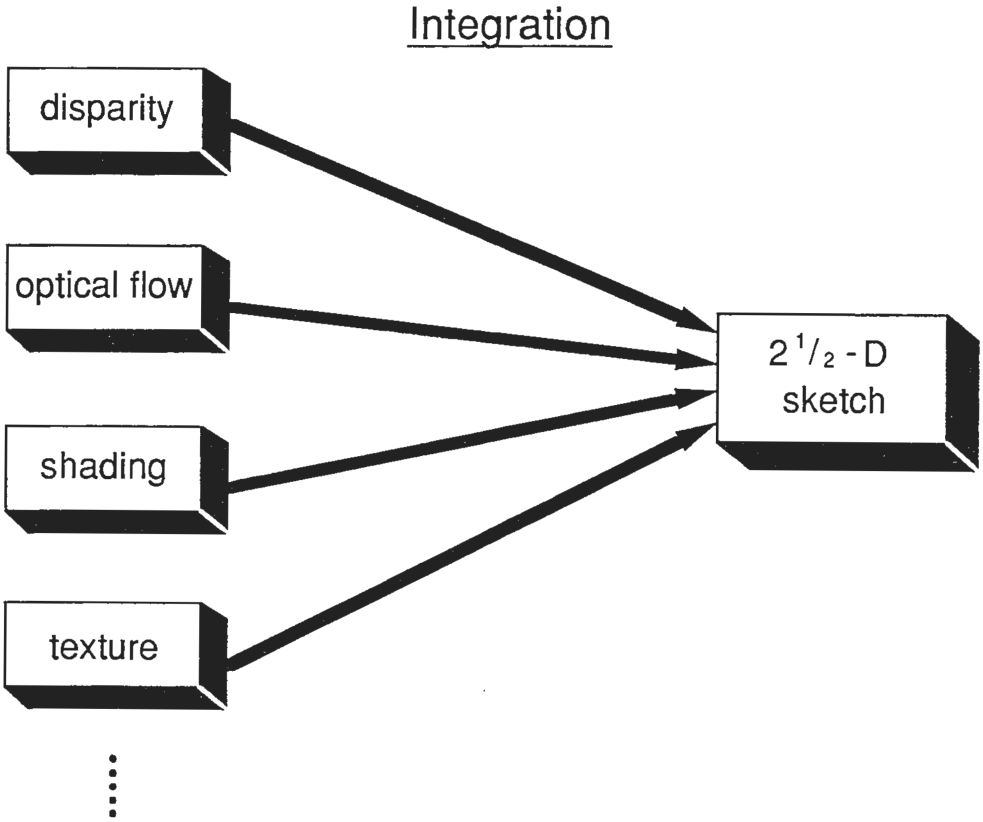


Figure 4..4: The visual information processing scheme proposed by Marr (1982). There are several vision modules in the early visual process. In the middle visual process, the outputs of the modules are integrated into one viewer-centered representation, the 2·1/2-D sketch.

is the 3-D model representation. Recently, from psychological experiments it appears that this is not the case for human visual memory. Instead, several 2-D representations obtained for several aspects of one object is the important representation for humans (Ullman & Basri, 1989). In accordance with the computational and psychological studies, it was found that there are many neurons in the IT and STS which respond to some specific aspect of the objects, including faces. Therefore, we believe that the appropriate representation for perception is the 2·1/2-D sketch, which is useful for object manipulation and navigation, and that the appropriate representation for recognition is some 2-D aspects.

5. Hierarchical interactions between cortical areas

5.1 Computational model of hierarchical interaction

The computational principle of the hierarchical structure pointed out in the previous section will be mathematically formulated with a simplified, hierarchical model shown in Fig. 5..1. In this model, the primary visual cortex estimates an intermediate representation $\hat{\mathrm{S}}$ between the higher level representation S and the image I. These two representations are hierarchical in the sense that $I = R(\mathrm{S})$, $I = R_1(\hat{\mathrm{S}})$, $\hat{\mathrm{S}} = R_2(\mathrm{S})$, that is, $R = R_1 R_2$ holds. Examples of pairs of S and $\hat{\mathrm{S}}$ in Fig. 4..2 are V in MT and $v^{\perp}$ in layer 4B of V1, L in the interstripe of V2 and $\hat{L}$ in the inter-blob of V1, and $r(\lambda)$ in the thin stripe of V1 and $\hat{r}(\lambda)$.

Similarly, as in the previous section, we can calculate the MAP estimate of S and $\hat{\mathrm{S}}$ by maximizing the following *a posteriori* probability.

$$P(\mathrm{S}, \hat{\mathrm{S}}|I) = \frac{P(I|\hat{\mathrm{S}}, \mathrm{S})P(\hat{\mathrm{S}} \wedge \mathrm{S})}{P(I)} \tag{5.1}$$

Here, we consider a purely hierarchical model of image generation where S can generate an image only via $\hat{\mathrm{S}}$ as shown in the lower half of Fig. 5..1. Then, instead of eq. 5.1, we need only minimize the following probability.

$$P(I|\hat{\mathrm{S}})P(\hat{\mathrm{S}}|\mathrm{S})P(\mathrm{S}) \tag{5.2}$$

Correspondingly, the following potential energy is given.

$$U(I|\hat{\mathrm{S}})+V(\hat{\mathrm{S}})+U(\hat{\mathrm{S}}|\mathrm{S})+U(\mathrm{S}) = 1/2\|R_1^{\sharp}(I)-\hat{\mathrm{S}}\|^2+V(\hat{\mathrm{S}})+1/2\|R_2^{\sharp}\{\hat{\mathrm{S}}-R_2(\mathrm{S})\}\|^2+U(\mathrm{S}) \tag{5.3}$$

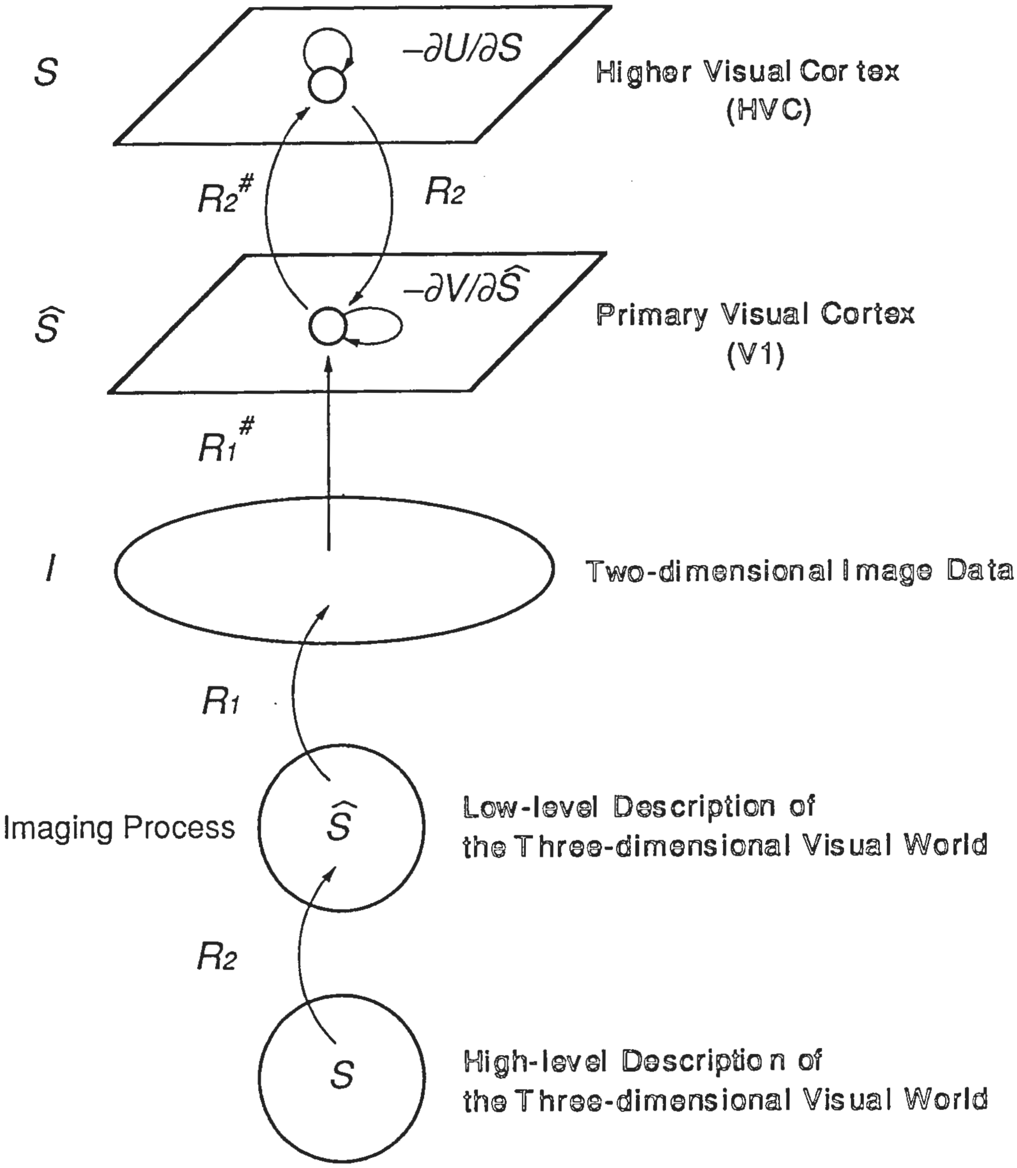

S
$-\partial U/\partial S$
Higher Visual Cortex (HVC)
$R_2^{\#}$
R_2
$\widehat{S}$
$-\partial V/\partial \widehat{S}$
Primary Visual Cortex (V1)
$R_1^{\#}$
I
Two-dimensional Image Data
R_1
Imaging Process
$\widehat{S}$
Low-level Description of the Three-dimensional Visual World
R_2
S
High-level Description of the Three-dimensional Visual World

Figure 5..1: Fundamental computational model of the hierarchical structure composed of the primary visual cortex and the higher visual cortex.

$R_1^\sharp$ and $R_2^\sharp$ are approximated inverse models of R_1 and R_2, respectively. We can derive the following dynamics of the neural network model shown in Fig. 5..1 as an approximated steepest descent method or a Newton-like method of the above energy eq. 5.3.

$$\begin{aligned}
\hat{\mathrm{S}}(0) &= R_1^\sharp(I) && (5.4)\\
d\hat{\mathrm{S}}(t)/dt &= R_1^\sharp(I) - \hat{\mathrm{S}} - \partial V/\partial\hat{\mathrm{S}} + R_2(\mathrm{S}) - \hat{\mathrm{S}} && (5.5)\\
\mathrm{S}(0) &= R_2^\sharp R_1^\sharp(I) && (5.6)\\
d\mathrm{S}(t)/dt &= R_2^\sharp\{\hat{\mathrm{S}} - R_2(\mathrm{S})\} - \partial U(\mathrm{S})/\partial\mathrm{S} && (5.7)
\end{aligned}$$

In the previous section, we stated that our model has only local interaction. Local interaction does not only imply locality in the visual field, but also in any visual representation of any area. For example, the orientation column in V1 could be viewed as an anatomical necessity to provide a meaningful internal model of the visual world under the condition of a limited number of connections. Furthermore, it is known that intrinsic connection is also local for higher visual areas. Consequently, we obtain the following prediction. *If a specific description of the visual world is represented in some area of the higher visual cortex, it is represented in columnar structure within that area.*

The locality of the interaction exists not only for the intrinsic neural connection but also for the feedforward and feedback neural connections between different visual areas. Thus, if the visual cerebral cortex were to possess only a single layer, it would not be able to describe the visual world with a global and abstract description, and hence would not be able to cope with middle or high-level vision problems. However, the real cerebral cortex and our model attain a global and abstract model by overlaying the local models many times as described in this subsection.

V1, V2, V3, V4 and MT are distinct vision modules, but nevertheless tightly connected. Thus, the overall visual system is interactive and complex. In our computational theory, the V2-V3-V4 complex forms an important and interesting system. This system estimates surface orientation and depth from different types of visual information. It is believed that this system is a consistency-maintaining-mechanism, and that perception corresponds to the activity pattern of the whole neural system. Of course, we can focus on specific dimensions (or attributes). Furthermore, we can control the activity of the different subsystems. For example, *a priori* estimation of the illumination source is usually biased toward the upper direction (Ramachandran, 1988), but it can be changed by intention. We are interested in how the overall resulting activity changes within the V2-V3-V4 complex when the activity pattern in one module is modified by attention.

The mathematical framework developed for interactions between different hierarchical modules has important implications also for interactions between parallel modules such as those within the V2-V3-V4 complex. Recently, interactions between vision modules are intensively investigated through psychological experiments (e.g. Bülthoff & Mallot, 1988, Stevens & Brookes, 1988, Buckley, Frisby & Mayhew, 1989, Cavanagh, 1987). These studies indicate that an output from one module depends on outputs from other modules. Our theory again solves the interaction problem by a rough first estimate by one shot algorithm and then gradual and iterative modification of the estimate by relaxation computation. Even if two modules are in a parallel relationship rather than a hierarchical relationship, either one module is dominant or the visual signal reaches one of the modules earlier than the other. Then, the dominant and early module sends its estimate to the other in a one-shot manner. Consistency of the outputs of the two modules will then be met by successive communications between them. This general architecture resolves the combinatorial explosion problem in vision caused by the large number of different visual cues (Ballard, Hinton & Sejnowski, 1983).

5.2 Layered-neural-circuit model for hierarchical interaction

Fig. 5..2 shows a layered-neural-circuit model for the hierarchical model described in the previous subsection. This model takes into account layered structures of both the primary visual cortex and the higher visual cortex (HVC).

The image data impinging on the retina is transformed into $R_1^\sharp(I)$ by feedforward calculation as shown in the first term of eq. 5.5. This calculation is done by the retinal neural network, the lateral geniculate nucleus, and the cortical network between pyramidal cells in the upper layer (P1 in Fig. 5..2) and stellate cells in the IVC layer of V1. The intrinsic recurrent connections shown by broken curves of Fig. 5.5 realize the third term $-\partial V/\partial \hat{\mathrm{S}}$ of eq. 5.5 which corresponds to the internal model of $\hat{\mathrm{S}}$. Within this recurrent loop, there exist interneurons which realize nonlinear calculations. Pyramidal cells P1 receive synaptic inputs $R_2(\mathrm{S})$ (the fourth term of eq. 5.5) by feedback neural connections from the pyramidal cells in the deep and upper layers of HVC (P4 and P5 in Fig. 5..2). While receiving these three kinds of synaptic inputs,

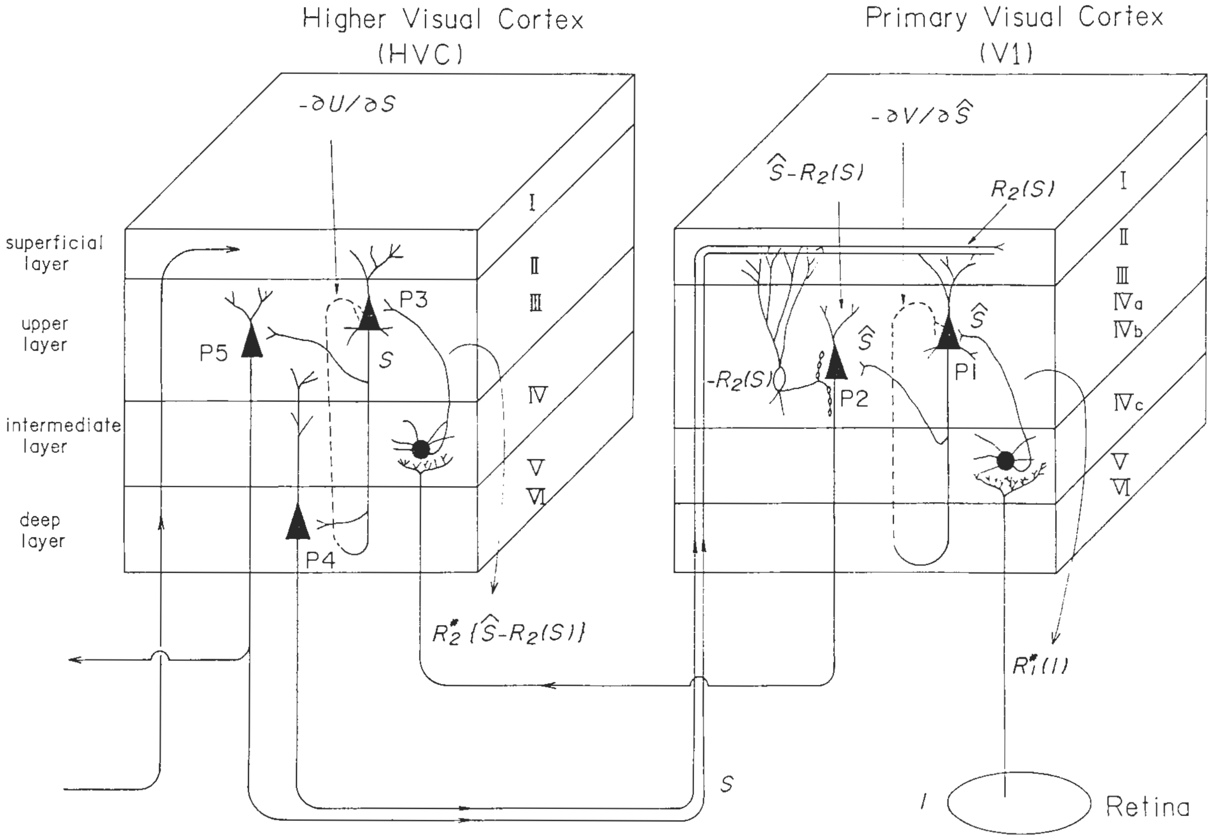


Figure 5..2: Layered-neural-circuit model of the hierarchical structure composed of the primary visual cortex and the higher visual cortex.

P1 estimates $\hat{\mathbf{S}}$ based on eq. 5.5. Pyramidal cells P2 in the upper layer of V1 receive two kinds of inputs: $\hat{\mathbf{S}}$ from P1, and feedback input $-R_2(\mathbf{S})$ from HVC via nonpyramidal inhibitory interneurons. P2 calculates $\hat{\mathbf{S}} - R_2(\mathbf{S})$ based on these two inputs and sends it to HVC by the feedforward neural connections. This corresponds to the content inside the curly bracket of eq. 5.7.

The feedforward signal calculated by P2 is then transformed into the synaptic input $R_2^\sharp\{\hat{\mathbf{S}} - R_2(\mathbf{S})\}$ to P3 in the upper layer of HVC by stellate cells in the intermediate layer of HVC. This is the first term of eq. 5.7. The second term of eq. 5.7, which gives an internal model of $\mathbf{S}$, is calculated by the intrinsic recurrent loop within HVC, shown by the broken curve in Fig. 5..2. P3 in HVC estimates $\mathbf{S}$ based on eq. 5.7. P4 and P5 in HVC receive synaptic inputs regarding $\mathbf{S}$ from P3, and send them back to the superficial layer of V1 by the feedback neural connection.

We think that our detailed neural circuit model is compatible with anatomical knowledge of the cerebral cortex (see for example Houser, Vaughn, Hendry, Jones & Peters, 1984; Parnavelas, 1984). However, it might be necessary to interpose interneurons between the feedback connections and the inhibitory interneurons in the upper layer.

One can see that V1 and HVC execute essentially the same computation while using very similar hardware. Although we explained the hierarchical neural network model as a model of V1 and HVC, the same model can be applied to any two visual areas which are arranged hierarchically (e.g. MT and MST). A multi-layer hierarchical structure is constructed by cascading several numbers of the unit network shown in Fig. 5..2. Fig. 4..2 contains several of these multi-layer hierarchical structures.

5.3 Examples of approximate inverse-optics operations

In our theory, existence of approximated inverse optics is essential. In this section we give several examples of the inverse-optics operations.

5.3.1 Detection of intensity discontinuity

For detecting intensity discontinuity, the hierarchical structure consisting of the following three layers is important; $4C\beta$, the interblobs in V1 and the interstripes in V2. It is assumed that these three layers represent $\Delta G * I$, $\hat{L}$, and L, respectively. The first and the third representations correspond to the intensity process and the line process in the coupled MRF, respectively. L, which is the intensity discontinuity, is represented by simple cells (with polarity) and complex cells (without polarity), which are different in preferred orientation. $\hat{L}$ corresponds to zero-crossing obtained by Laplacian Gaussian convolution. It does not correspond exactly to physical edges. The forward computation from $\hat{L}$ to L could be performed as follows: The neuron in the second layer ($\hat{L}$) is excited by the neuron which has similar preferred orientations, but the center of the receptive field is displaced along the preferred orientation. Simultaneously, it is inhibited by the neuron whose center of the receptive field is displaced perpendicularly to the preferred orientation regardless of the preferred orientation. Such an operation realizes an inverse model of optics, which takes into account the a priori knowledge about physical discontinuity in the 3-D world, that is "continuity of discontinuity".

5.3.2 Derivation of velocity field

Kersten, O'Toole, Sereno, Knil and Anderson (1987) trained a neural network model to solve the aperture problem in motion perception. Wang, Mathur and Koch (1989) proposed a biologically realistic neural network model for motion perception. In these studies, candidates of approximated inverse optics can be found.

In our model, velocity computation should be done through the connections from 4B in V1 to MT. We explain only feedforward calculation, that is, the inverse optics operation. The optimal direction of motion and the center of the receptive field of the neuron in 4B are denoted by ϕ_k and (x, y), respectively. The velocity component along the optimal direction is $v_{xy}^{\perp}(\phi_k)$. Then, the forward computation to obtain a rough estimation of the true velocity V at each position (x, y) is done with the following equation:

$$V_{xy}(\phi_k) = \sum_{(i,j)\in c} \sum_{k} v_{ij}^{\perp}(\phi_k) \cos(\phi_k - \phi) \tag{5.8}$$

Here, $V_{xy}(\phi)$ is a component of the estimated velocity along the direction ϕ. The first summation of the right side is summation of outputs of all the neurons inside the neighborhood c of the point (x, y). We can calculate a rough estimate of the true velocity vector which is somewhat independent of the local gray level only by this feedforward calculation.

5.3.3 Estimation of spectral reflectance

The surface spectral reflectance is computed through the hierarchical structure consisting of blobs in V1, thin stripes in V2, and V4. Hurlbert & Poggio (1988) computed the inverse transformation (Moore-Penrose pseudoinverse) from the image intensity to the surface reflectance. It appeared that the filter consists of a small excitatory center with a large inhibitory surround. The filter is very similar to the receptive field found in V4, the spatial feature of which is the large silent suppressive surround (Desimone, Schein, Moran & Ungerleider, 1985). The surround spectral property is the same as the excitatory center. In general, the surface spectral reflectance is piecewise constant, while the source illumination intensity changes gradually over the surface. Therefore, one of the functions of the large surround is to suppress the effect of the illumination.

5.3.4 Stereoscopic matching and determination of the disparity

The correspondence problem must be solved to calculate the stereo disparity. The possible forward calculation is Marr-Poggio's second algorithm which is a kind of one-shot algorithm (Marr & Poggio, 1979). It solves the correspondence between left and right images based on the zero-crossing (Marr & Hildreth, 1980) under the smoothness constraint of the depth of the surface. This computation would be done between interblobs in V1 and thick stripes in V2. In this algorithm, the ambiguity of correspondence is reduced through a coarse-to-fine processing cycle.

Recently, some psychological evidence has been obtained. Some evidence supports the algorithm, and some does not. Mitchison (1988) showed the evidence that the coarse-scale features guide the choice of a particular set of correspondence. Some psychophysical experiments suggest that the centroids, peaks, and troughs of the activity pattern resulting from the convolution of the image with the receptive fields of visual neurons, are also used as primitives other than the zero-crossings (Nishihara, 1988; Daugman, 1988; Bülthoff & Mallot, 1987, 1988; Legge & Gu, 1989). Thus, considering these psychophysical experiments, the following regularization functional proposed by Poggio, Torre and Koch (1985) may be an appropriate choice.

$$\int \{[\Delta G * (I(0,x,y) - I(1,x+sd(x,y),y))]^2 + \lambda(\nabla sd)^2\} dxdy \qquad (5.9)$$

6. Shape from shading: a model of monocular depth perception

6.1 Local shading analysis by Pentland

"Shape from Shading" in computer vision (e.g., Horn, 1975; Horn & Brooks, 1989), which is used to recover the 3-D shapes of objects from the intensity of 2-D image data, is equivalent to one element of "monocular depth perception" in human vision. However, it is not apparent how humans recover 3-D shape information. In this section, we discuss this problem at the three levels of explanation proposed by Marr (1982): computational theory, representation and algorithm, and hardware implementation.

When constructing a model of perceiving shape from shading, representation of the model's input and output is very important. In human vision, looking into a slanting flat plane from a window causes the shape perception of a front parallel plane. Lighting condition variations do not effect the perception of shape from shading (see Fig. 6..1, for example). Therefore, it is quite appropriate to use the derivatives of image intensity for the input representation rather than the image intensity itself. On the other hand, we choose as the output the surface orientation represented by slant θ and tilt ϕ whose viewing direction is parallel to the z-axis (see Fig. 6..2). Therefore, from the point of view of these representations, the "Local Shading Analysis method" (Pentland, 1984, 1986) is well-suited to a model of perceiving shape from shading in human vision.

The "Local Shading Analysis method" is an approximated scheme, which is based on three assumptions: (1) the surface has Lambertian umbilic points and relatively low slant value [surface attribute assumption], (2) the z-component of the illuminant direction is relatively large [illuminant direction assumption], (3) the surface curvature κ is common throughout the region [surface curvature assumption]. Under these assumptions, the tilt ϕ becomes the image direction in which the second derivative of image intensity d^2I is the greatest and the slant θ is approximated as shown in eq. 6.1. Here, the surface curvature κ is determined by applying the constraint that the resulting slant θ must satisfy the inequality $0 \leq \cos\theta \leq 1$.

$$\theta = \cos^{-1}\{\kappa(|\nabla^2 I/I| - \kappa^2)^{-1/2}\} \tag{6.1}$$

However, because the algorithm of the "Local Shading Analysis method" works only under the assumptions described above, we believe that this method can not fully account for human monocular depth perception in the real world.

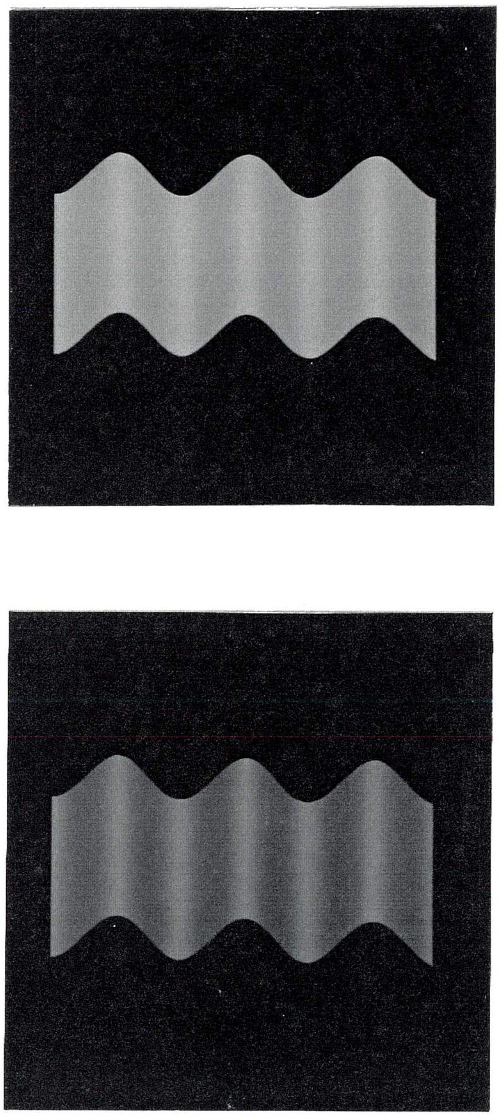


Figure 6..1: Shaded image of a smoothly curved surface with two different lighting conditions.

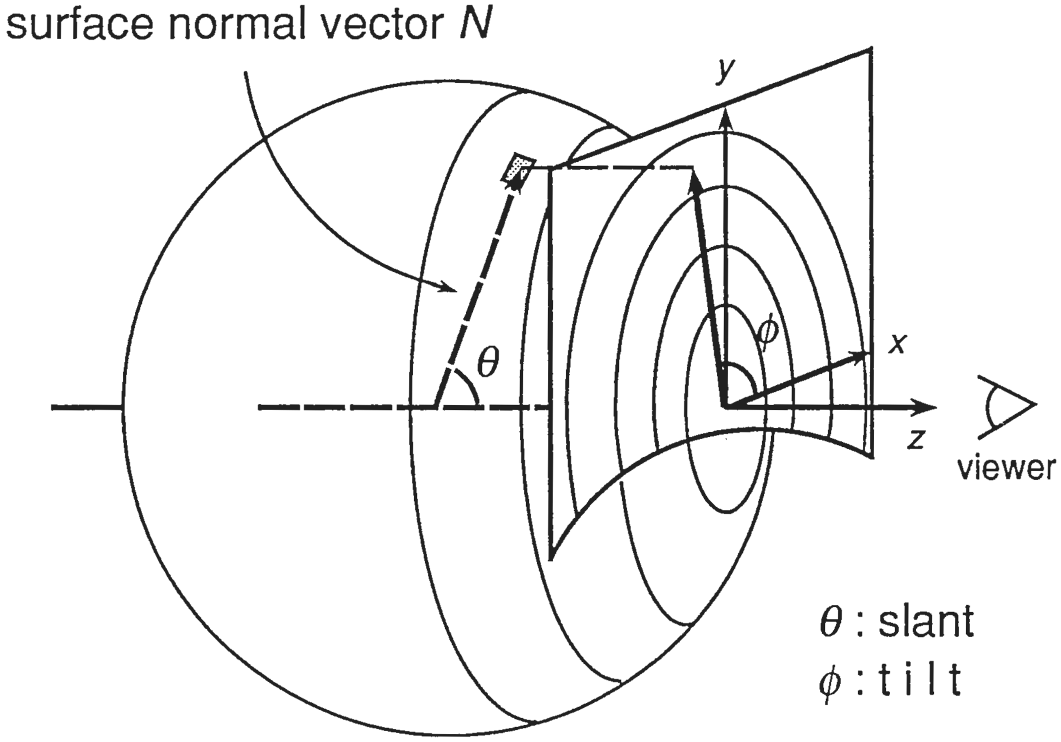


Figure 6..2: Representation of surface orientation using slant θ and tilt ϕ

To solve this problem, we used the forward and inverse models of image generation proposed in the previous section, because the forward model R (optics) can compensate for the shortcomings of the inverse model $R^{\#}$ (approximated inverse optics) like the "Local Shading Analysis method".

6.2 Algorithm and neural network model

Based on the forward and inverse models of image generation, we constructed a detailed model of perceiving shape from shading. Practically, the perception of shape from shading requires two visual calculations that respectively estimate the surface orientation N and the illuminant direction ν (see Pentland, 1982 for example). In this section, assuming that the illuminant direction is known, we propose an algorithm that estimates the surface orientation N (represented by slant θ and tilt ϕ).

At the start, the first and second derivatives of the image, along with various directions, are calculated from the image data impinging on the retina. Then, the local surface orientation N is calculated from these derivatives. Fig. 6..3 shows the fundamental computational model of the hierarchical structure composed of the primary visual cortex and the higher visual cortex (V3). In this model, the first and second derivatives of the image intensity are represented at the primary visual cortex and the surface orientation N is represented at the higher visual cortex.

Here we define $\hat{N}$ as an intermediate representation which is between the higher-level representation N and the image I. The feedforward connection from the retina to the primary visual cortex provides the calculation of $\hat{N}_1$ and $\hat{N}_2$. $\hat{N}_1$ is composed of directions η_i and the first derivative of the image dI along with the directions η_i as

$$\hat{N}_1 = \begin{pmatrix} \eta_1 & dI(\eta_1) \\ \eta_2 & dI(\eta_2) \\ \vdots & \vdots \\ \eta_{2n} & dI(\eta_{2n}) \end{pmatrix}. \tag{6.2}$$

$\hat{N}_2$ is composed of directions η_i and the second derivative of the image d^2I along with the directions η_i as

$$\hat{N}_2 = \begin{pmatrix} \eta_1 & d^2I(\eta_1) \\ \eta_2 & d^2I(\eta_2) \\ \vdots & \vdots \\ \eta_n & d^2I(\eta_n) \end{pmatrix}. \tag{6.3}$$

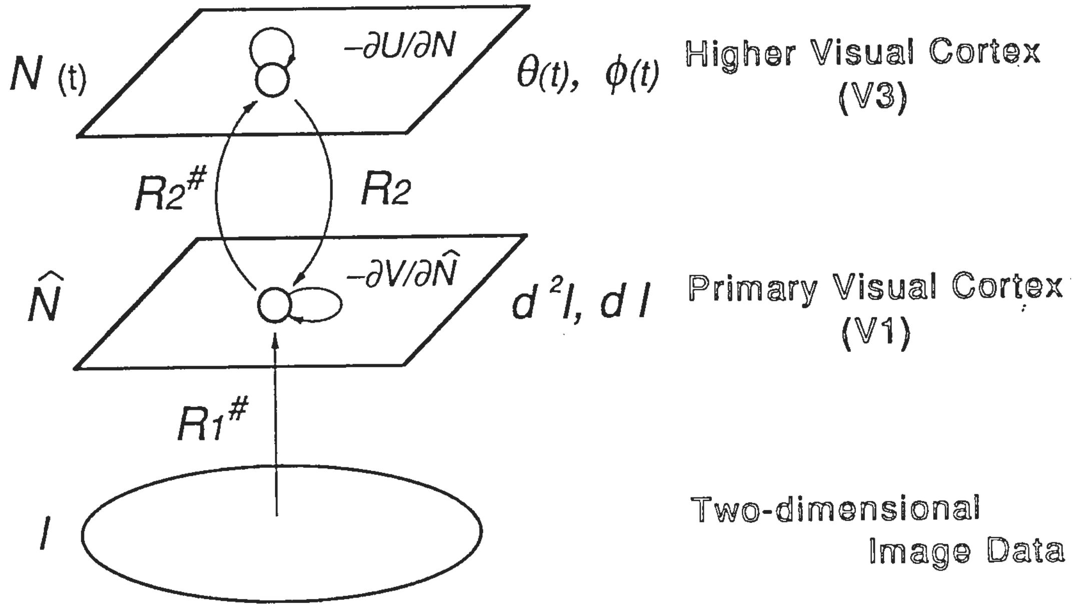

N (t)
$-\partial U/\partial N$
$\theta(t)$, $\phi(t)$
Higher Visual Cortex (V3)
$R_2^{\#}$
R_2
$\hat{N}$
$-\partial V/\partial \hat{N}$
$d^2 I$, $d I$
Primary Visual Cortex (V1)
$R_1^{\#}$
I
Two-dimensional Image Data

Figure 6..3: Fundamental computational model of the hierarchical structure for perceiving shape from shading.

Here, $\eta_{n+i} = \eta_i + \pi$ and η_i $(i \leq n)$ satisfies $0 \leq \eta_i < \pi$. $R^{\#}_{11}$ and $R^{\#}_{12}$ are the approximated inverses that respectively calculate dI and d^2I as

$$\begin{aligned} R^{\#}_{11} &= \sum_{\epsilon} G(\epsilon,\sigma)\partial/\partial x' G(x',y'+\epsilon,\sigma), \quad (6.4) \\ R^{\#}_{12} &= \sum_{\epsilon} G(\epsilon,\sigma)\partial^2/\partial x'^2 G(x',y'+\epsilon,\sigma). \quad (6.5) \end{aligned}$$

Here, $G(x,\sigma) = (2\pi\sigma^2)^{(-1/2)}exp(-x^2/2\sigma^2)$ and $G(x,y,\sigma) = (2\pi\sigma^2)^{(-1)}exp(-(x^2 + y^2)/2\sigma^2)$. (x',y') is the coordinate system after a rotational transformation through angle η_i in the image plane, and σ is the standard deviation of the Gaussian function. The approximated inverse models $R^{\#}_{11}$ and $R^{\#}_{12}$ include a "smoothness constraint" because the Gaussian function of these models smoothes image I. The Gaussian function has been used as the main element of the optimal edge detector in the intensity image (Torre & Poggio, 1986; Canny, 1986). Eq. 6.4 corresponds to the odd symmetrical receptive field of a simple cell and eq. 6.5 corresponds to the even symmetrical receptive field of a simple cell in the primary visual cortex. Fig. 6..4 shows the filter shapes of $R^{\#}_{11}$, $R^{\#}_{12}$.

Next, we describe an algorithm in which the higher visual cortex interacts with the primary visual cortex. Since surface orientation N is composed of slant θ and tilt ϕ, it is equivalent to a vector with a radius of unit length in a spherical coordinate (see Fig. 6..2) :

$$N(t) = (1,\theta,\phi). \quad (6.6)$$

Slant θ satisfies $0 \leq \theta < \pi/2$, tilt ϕ satisfies $0 \leq \phi < 2\pi$ and t represents the time from the beginning of the estimation.

We explain how to estimate the surface orientation N in the approximated calculation $R^{\#}_2$. The approximated inverse optics $R^{\#}_2$ calculates the initial surface orientation $N(0)$ from $\hat{N}_2$ as

$$N(0) = R^{\#}_2(\hat{N}_2). \quad (6.7)$$

This calculation by the network model is realized between the primary visual cortex and the higher visual cortex (see Fig. 6..5). $2n$ neurons exist at the unit of the higher visual cortex, and the population coding of the activation of these neurons represents surface slant θ and tilt ϕ (see Fig. 6..6). In these figures, the activation level of the i-th excited neuron Θ_i is determined by the second derivative of the image $d^2I(\eta_i)$, and the location of the excited neuron Φ_i is

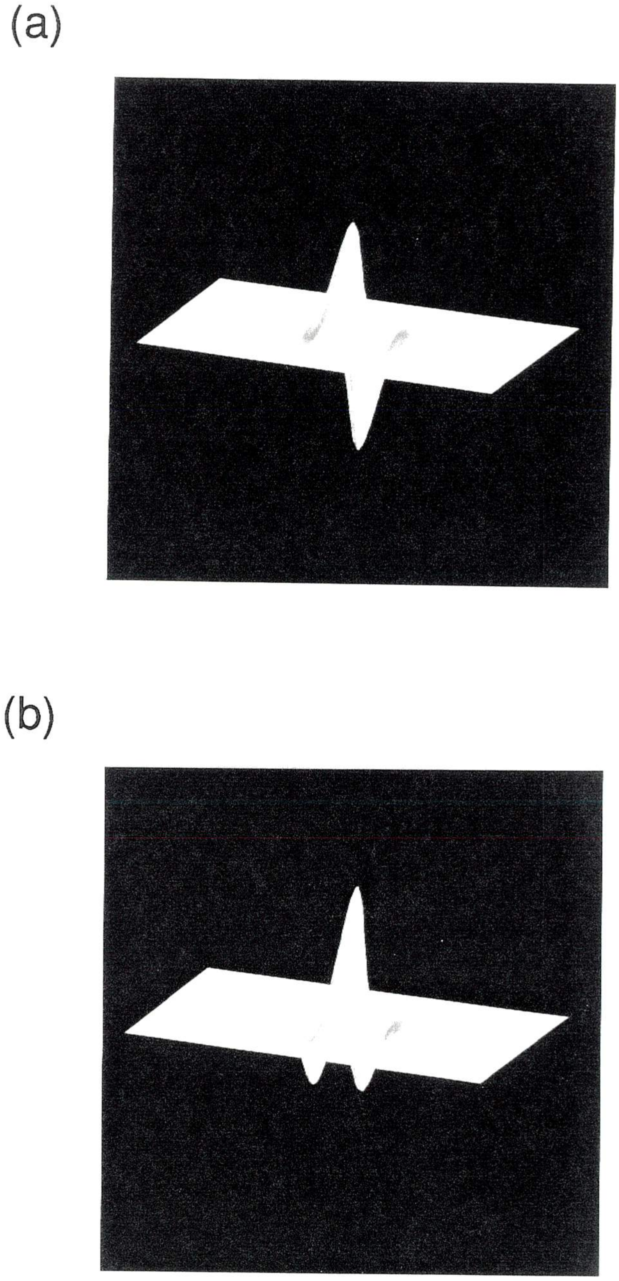


Figure 6..4: Filters, along with a specific direction, that function as $R^{\#}_{11}$ (a) and $R^{\#}_{12}$ (b).

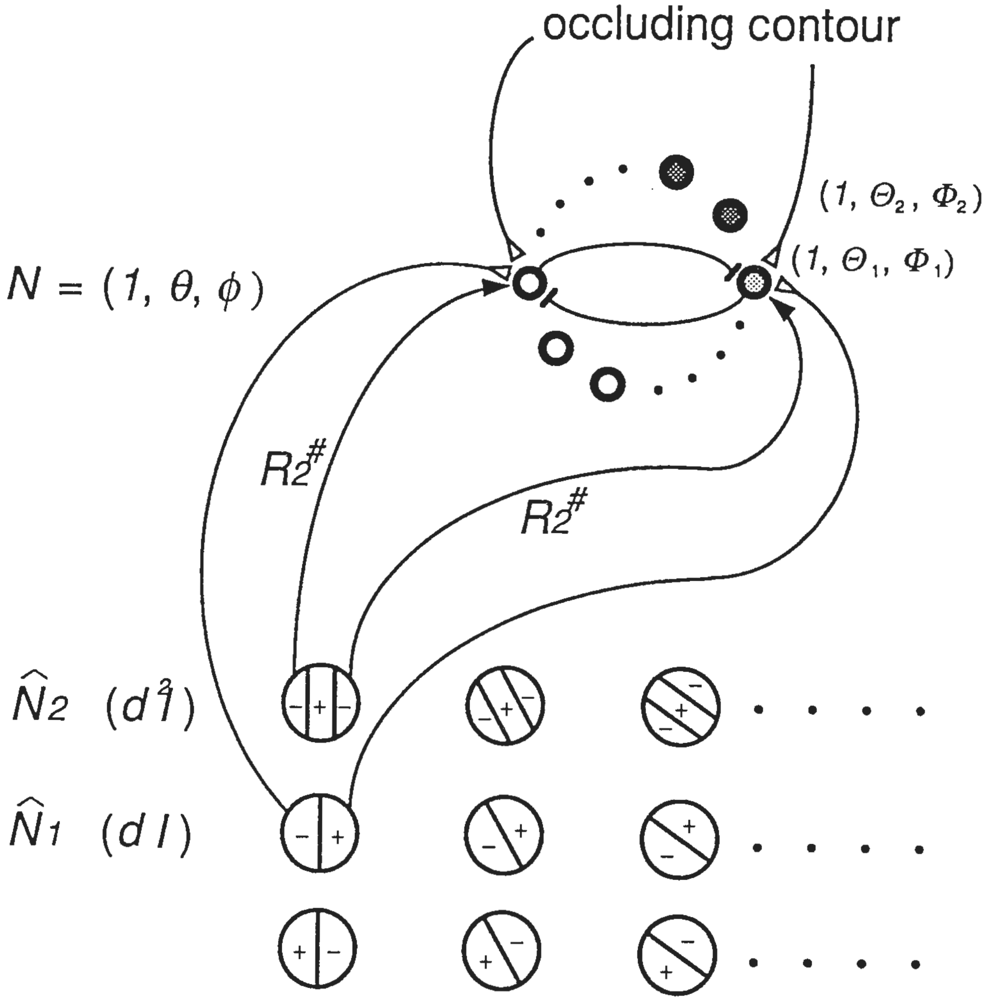


Figure 6..5: Network model between the primary visual cortex and the higher visual cortex. The gray level of each neuron at the higher visual cortex represents the activation level.

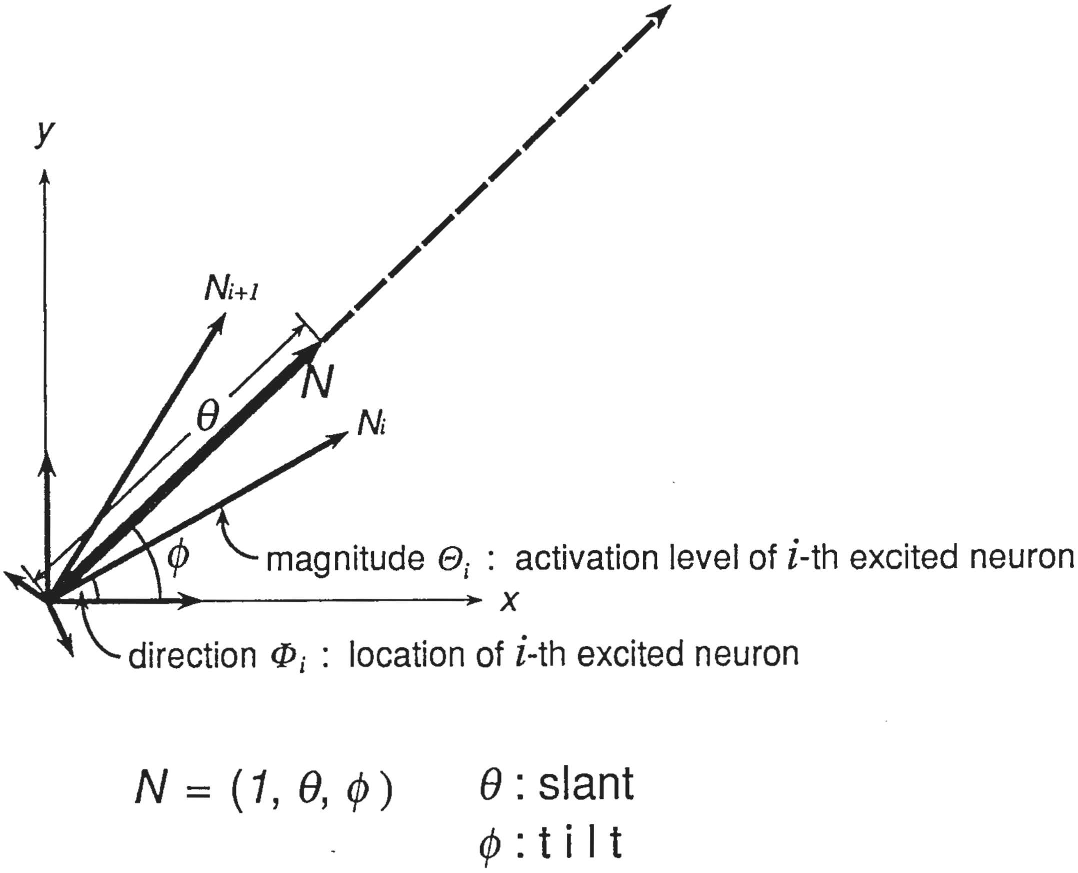

y
N_{i+1}
θ
N
N_i
ϕ
magnitude Θ_i : activation level of i-th excited neuron
x
direction Φ_i : location of i-th excited neuron
$N = (1, \theta, \phi)$
θ : slant
ϕ : tilt

Figure 6..6: Population coding of the activation of the neurons at the higher visual cortex. Broken thick line shows the summation of vector N_i and solid thick line shows the normalized summation of vector N_i.

determined by the information about the boundary given by the occluding contour or self-shadow and the first derivative of the image $dI(\eta_i)$.

The value Θ_i is calculated by using the "Local Shading Analysis method" as

$$\begin{aligned}\Theta_i &= \cos^{-1}\{(\min|d^2I(\eta_i)|/|d^2I(\eta_i)|)^{1/3}\} \\ &\quad (i = 1, 2, \cdots, n).\end{aligned} \tag{6.8}$$

Here, $\min|d^2I(\eta_i)|$ is the minimum of $|d^2I(\eta_i)|$ along with the direction η_i in the image plane. Eq. 6.8 is an approximated calculation of the equation $\cos^{-1}(|(\rho\lambda z_0\kappa)/d^2I(\eta_i)|^{1/3})$ based on the assumption that the surface has Lambertian umbilic points and the surface curvature κ is common throughout the region. Here, ρ is the albedo of the surface, λ is the intensity of the illuminant, and z_0 is the z-component of the illuminant direction.

The value Φ_i is ambiguous depending on the surface type (convex or concave surface): It is either η_i or $\eta_i + \pi$. The ambiguity is resolved using information about the boundary caused by the occluding contour or self-shadow and about the sign of $dI(\eta_i)$. We first choose the value Φ_i in the neighborhood of the boundary caused by the occluding contour or self-shadow that satisfies $V \cdot e(\Phi_i) \geq 0$. Here, V is a vector normal to the boundary from the lighted towards the shadowed side in the image plane. $e(\Phi_i)$ is a unit vector with the direction Φ_i. The value Φ_i has an ambiguity with π change at the point of the depth extremum, which could be estimated from the intensities at the two occlusion points (Todd & Reichel, 1989). Then, the value Φ_i is checked by both the real data and the estimated data of the sign of $dI(\eta_i)$. The algorithm utilizes the fact that general objects have convex surfaces in the neighborhood of the occluding contour in the real world and that the sign of $dI(\eta_i)$ depends on the surface type: convex surface or concave (Pentland, 1986). Therefore, this model can estimate the surface orientation of arbitrary smooth objects with convex and concave surfaces.

Next, the initial surface orientation $N(0)$ is estimated by applying the population coding. In rectangular coordinates, the activation level of the i-th excited neuron Θ_i and the location of the excited neuron Φ_i are represented by a vector N_i (see Fig 6..6) :

$$N_i = \begin{pmatrix} \cos\Phi_i \cdot \Theta_i \\ \sin\Phi_i \cdot \Theta_i \end{pmatrix}. \tag{6.9}$$

Therefore, the initial surface orientation $N(0)$ corresponds to population vector of N_i vectors $(i = 1, 2, \cdots, n)$ as

$$N(0) = n^{-1/2} \sum_{i=1}^{n} N_i \tag{6.10}$$

Now, following this initial estimation $N(0)$, the model starts the relaxation calculation as

$$dN(t)/dt = R_2^{\#}\{\hat{N}_2 - R_2(N)\} - \partial U(N)/\partial N \tag{6.11}$$

The forward model calculates $R_2(N)$ from the estimated surface orientation $N(t)$, and error $\hat{N}_2 - R_2(N)$ is calculated. Then $R_2^{\#}\{\hat{N}_2 - R_2(N)\}$ is calculated by the inverse model $R_2^{\#}$, and is sent to the higher visual cortex. The second term of eq. 6.11 is caused by the internal model $U(N)$. In this case, we use the membrane potential as the internal model energy $U(N)$.

Here, we assume that the unit vector of the illuminant direction is (x_s, y_s, z_s) and that the surface type is Lambertian. Therefore, the forward model R_2 calculates the estimation of $dI(\eta_i), d^2I(\eta_i)$ using the equation $I = x_s \sin\theta \cos\phi + y_s \sin\theta \sin\phi + z_s \cos\theta$ and the filters $R_{11}^{\#}$, $R_{12}^{\#}$.

6.3 Simulation results

Using the above algorithm, we estimated the surface orientation of the profile of a sphere and an ellipsoid from their intensity images. The images have 8-bit resolution in the gray level and their size is 512×512 pixels. The radius of the sphere in the images is about 125 pixels. The long-axis length of the ellipsoid parallel to the viewing direction is 250 pixels and its short-axis length is 125 pixels. The standard deviation of the two filters $R_{11}^{\#}$, $R_{12}^{\#}$ is 15 pixels. The coefficient of each term of the relaxation calculation, as shown in eq. 6.11, is 0.2. We judge that the calculation has converged when the correction factor of slant θ is less than 0.02 radians.

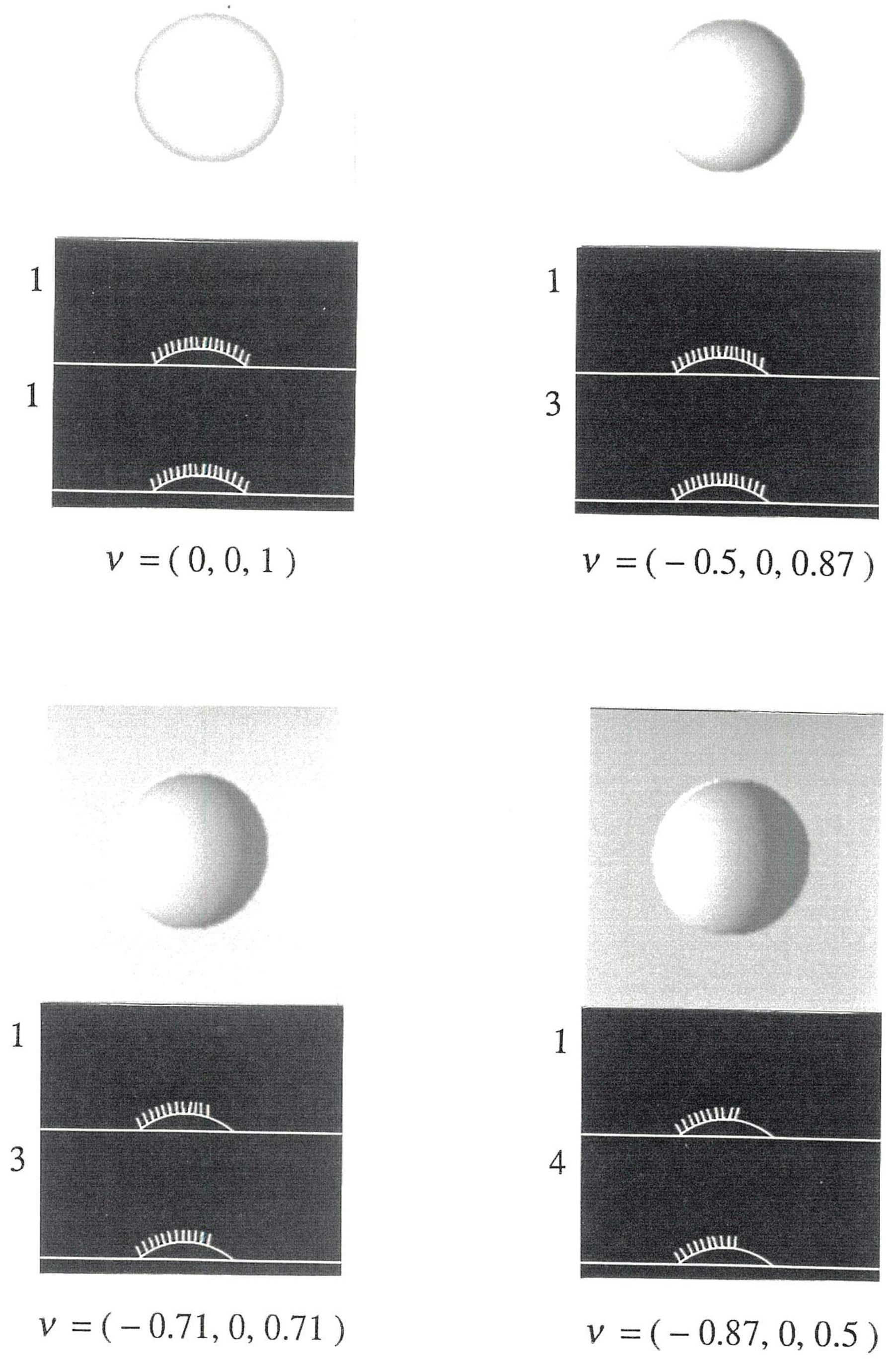
Fig. 6..7 shows the images of the sphere for four illumination conditions and the estimated surface normals of the horizontal profile of the sphere. Similarly, Fig. 6..8 shows the case of the ellipsoid. Each curved line at the middle and bottom of the images is a center part of the profile of the true surface shape, and the short lines represent the estimated surface normals. The reason why the part near the boundary is not reconstructed is that the convolution of the image with $R_{11}^{\#}$ or $R_{12}^{\#}$ can not adequately be done when these filters intersect with the boundary. The middle of each image shows the first estimated surface normals at $t = 0$ and the bottom of each image shows the final estimated surface normals when the calculation has converged. Each numeral in Fig. 6..7 and Fig. 6..8 represents the number of iterations. The illuminant direction ν is defined under the assumption that the x-axis is parallel to the horizontal direction of the image and the y-axis is parallel to the vertical.

Fig. 6..7 (c),(d) and Fig. 6..8 (c),(d) do not satisfy the "illuminant direction assumption" because the z-component of the illuminant direction in the images is too small. And the surface curvature κ in the images of Fig. 6..8 does not satisfy the "surface curvature assumption". However, surface normals in all these cases could be recovered to reasonable accuracy within 10 iterations.

In our model, the boundary caused by the occluding contour or self-shadow plays a very important role. This is because the boundary not only determines the area where the surface orientation is estimated by this model but also resolves the problem of tilt ambiguity which is indispensable for the estimation of surface orientation.

Next, we discuss the possible implementation of hardware between the primary and higher visual cortices. Strictly speaking, from the point of view of hardware implementation, it is not appropriate to use the "Local Shading Analysis method" as the calculation of the approximated inverse optics $R_2^{\#}$. However, since Eq. 6.8 is almost equivalent to a sigmoid function, it is expected that $R_2^{\#}$ can readily be implemented by feedforward neural networks such as multi-layer perceptrons. How the optics R_2 is implemented depends on how the illuminant direction is estimated and represented. If these problems can be solved, the interaction between the surface orientation and the illuminant direction is simply represented by the inner product when the surface type is Lambertian. The differential of the estimated intensity dI, d^2I, can be realized by a lateral connection between neurons. Also, it is possible that the approximated inverse optics $R^{\#}$ and the "smoothness constraint" can be learned in the neural network (see Lehky & Sejnowski, 1989 for example).

Furthermore, the local parallel and hierarchical model of the visual world based on this computational model can well explain psychological examples of shape from shading (Ramachandran, 1988; Todd & Reichel, 1989): the inversion of convex and concave surfaces, the effect of subjective contours, ordinary structures of surfaces. In particular, we regard the interaction between V3 and V4 as important in perceiving shape from shading, because we believe that monocular depth is represented in V3, and the location of the light-source and 3-D locations of objects are represented in V4 (see section 3). Although we do not have decisive physiological support for our neural network model in V3, the small number of iteration by our model (typically less than 10)



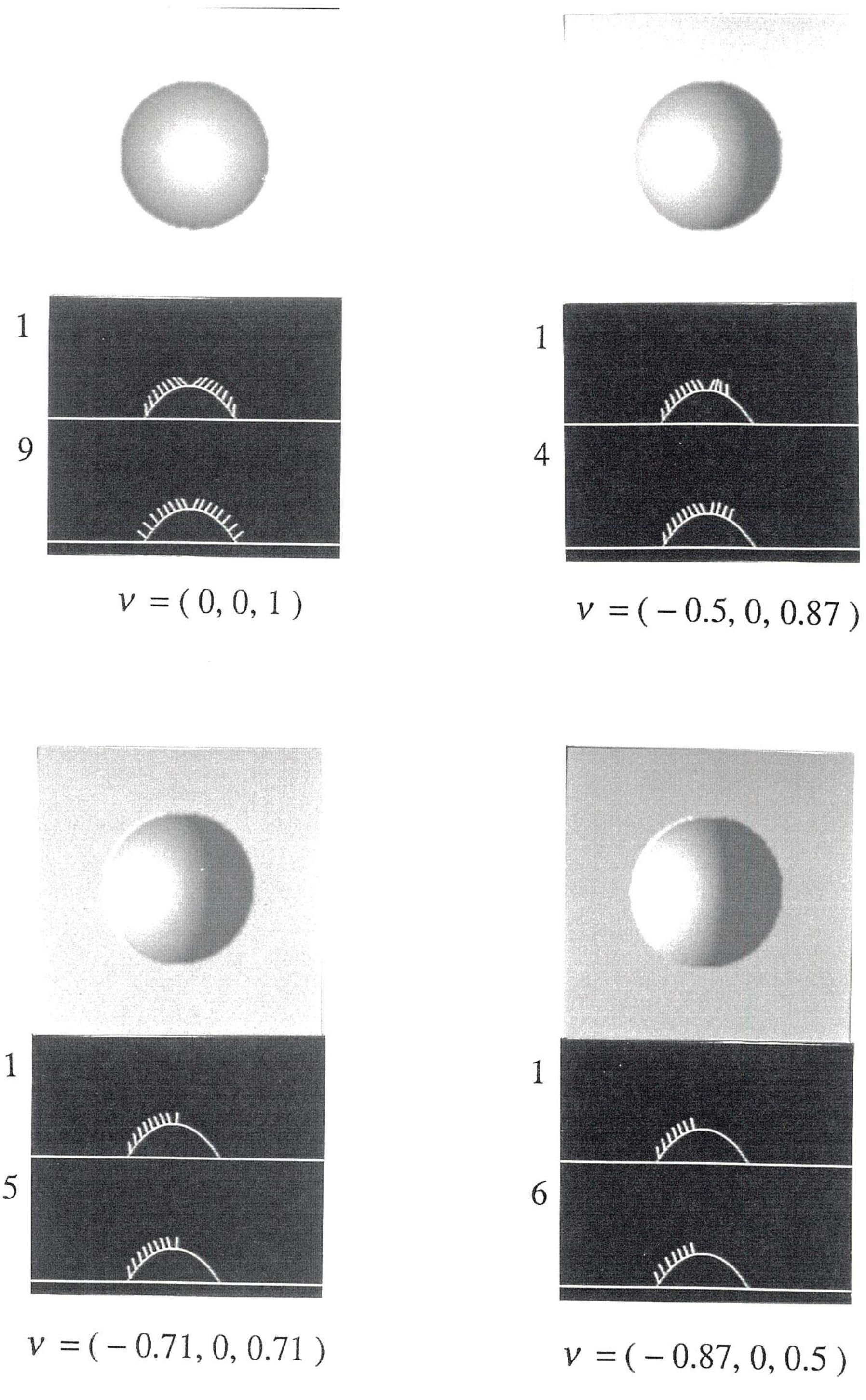
$v = (0, 0, 1)$

$v = (-0.5, 0, 0.87)$

$v = (-0.71, 0, 0.71)$

$v = (-0.87, 0, 0.5)$

Figure 6..7: Result of estimation of surface orientation for the image of a sphere.



$v = (0, 0, 1)$

$v = (-0.5, 0, 0.87)$

$v = (-0.71, 0, 0.71)$

$v = (-0.87, 0, 0.5)$

Figure 6..8: Result of estimation of surface orientation for the image of an ellipsoid.

strongly supports our model.

Some of the simulation results in this section have been reported in preliminary form (Hayakawa, Inui & Kawato, 1991).

7. Computational theory and neural network model for the brightness illusion

In this section, we will develop a specific computational theory and a concrete neural network model for brightness illusion based on our general framework developed in the previous sections. In recent studies of computational vision, it is believed that the main role of vision is to reconstruct a three-dimensional structure from a two-dimensional image. The purpose of this section is twofold: one is to model the early visual processes from the standpoint of visual reconstruction. The other is to model two important functions: detection of intensity discontinuity and filling-in. Some of the results in this chapter have been reported in preliminary articles (Inui, Kawato & Hongo, 1990, Inui, Hongo & Kawato, 1990, Hongo, Inui & Kawato, 1990).

7.1 Mach bands

In this section, we propose a new model for the brightness illusion of Mach bands. We perceive a white band and a dark band clearly near the boundary between a brightly illuminated region and a dimly illuminated region of the visual field (Fiorentini, 1972). Recently, new properties for these phenomena are reported as follows.

1. Mach bands are not visible on ramps whose width is less than 4 minutes of arc (Ross, Morrone & Burr, 1989).

2. Mach bands disappear when a high contrast line is presented in the ramp region (Ratliff, 1984).

3. The dark and light bands show different properties (e.g. Shipley & Wier, 1972).

These properties are very important. If we adopt a simple and popular model of the second derivative of intensity (Hartline & Ratliff, 1957), the clearest Mach band should be seen in the case of a step pattern. However, we cannot see it in reality. In this respect, the Chevreul illusion is different from the Mach band (see also Ross, Holt & Johnstone, 1981). In fact, at least three panels are required for the Chevreul illusion (Bekesy, 1968). These properties imply that we detect the intensity discontinuity by a line process and that Mach bands are inhibited by activation of the line process. In the following section, we propose a two-layered neural network model based on this fundamental idea.

7.1.1 Computational theory of Mach band illusion

We propose a computational theory of the illusion from the standpoint of visual reconstruction. In the theory, it is assumed that an approximate intensity profile is reconstructed from a given piece of sparse intensity data by minimizing an energy function. The energy function which we propose is as follows:

$$\begin{aligned} E &= E_D + \lambda E_I + \gamma E_L \\ E_D &= \int \{f(x) - d(x)\}^2 dx \\ E_I &= \int \{1 - l(x)\}\{\partial^2 f(x)/\partial x^2\}^2 dx \\ E_L &= \int l(x)\{1 - l(x)\}dx + (\epsilon/\gamma)\int l(x)dx, \end{aligned} \tag{7.1}$$

where $f(x)$ represents the reconstructed intensity profile. $d(x)$ represents a given piece of data of the contrast. In this model $d(x)$ is given around the edge. $l(x)$ represents the output of a neuron in the line process.

The first equation is the data fitting term of the energy function. The second equation is the interpolation term. The second derivative term of this equation determines the smoothness of interpolation. In this case, it is a bending energy. In other words, it is the curvature of an interpolated surface. This energy is often called the thin plate potential. Grimson (1981) adopted the thin plate potential energy as a constraint function in the visual surface reconstruction. He pointed out the importance of the discontinuity detection because this smoothness constraint caused the warping and ripping of the reconstructed surface (Gibbs phenomena), which is undesirable (Grimson, 1982). If the bending energy is large enough, the output of the line process neuron $l(x)$ becomes 1, and the network stops interpolating smoothly. As a result, surface discontinuity occurs. The third equation determines the behavior of $l(x)$. The first term allows the value of $l(x)$ to be either 1 or 0. The second term is a penalty term, which prohibits $l(x)$ from being 1 everywhere. λ determines the balance between the data fitting term and the constraint term.

λ is the regularization parameter determined by the signal-to-noise ratio (see Introduction). In related psychophysical experiments, λ becomes smaller if the stimulus contrast is high or the sensitivity is high.

7.1.2 Structure of the neural network

In this section, we propose the neural network model which minimizes the energy function described above. Figure 7..1 shows a schematic diagram of the network. The circles show luminance units which have a nonlinear input-output function. The bars show the line detectors which represent discontinuities of the luminance units' states. The arrows show the interaction between a luminance unit and other luminance units or between a line detector and the luminance units. An excited line detector inhibits all interactions across it. Time evolution of the membrane potential of an intensity neuron is described by the equation:

$$dC(x)/dt = -\partial E/\partial f(x) \tag{7.2}$$

Here, $f(x)$ represents the output of a neuron, and $C(x)$ represents the membrane potential of a neuron. The energy function defined above is substituted into this equation. Then we obtain:

$$\begin{aligned} dC(x)/dt &= -2\{f(x) - d(x)\} - 2\lambda\{1 - l(x)\}\{\partial^4 f(x)/\partial x^4\} \\ f(x) &= f_{max}\frac{C(x)^n}{C(x)^n + C_{50}^n} \end{aligned} \tag{7.3}$$

For the minimization, the fourth derivative of $f(x)$ must be calculated. The second equation, called the "Naka-Rushton equation", determines the input-output relationship of the neuron. In this case, $C(x)$ is identified with contrast of the visual stimulus. This function has been shown to provide a good fit to the contrast response functions of the visual cortical neurons (Albrecht & Hamilton, 1982, see also Sclar, Maunsell & Lennie, 1990). $d(x)$ represents the given data of the contrast. In this model, $d(x)$ is given around the edge, while $l(x)$ represents the output of a neuron in the line process.

$$\begin{aligned} du(x)/dt &= -\partial E/\partial l(x) \\ &= \lambda\{\partial^2 f(x)/\partial x^2\}^2 - \gamma\{1 - 2l(x)\} - \epsilon \\ l(x) &= 1/\{1 + e^{-2\lambda u(x)}\} \end{aligned} \tag{7.4}$$

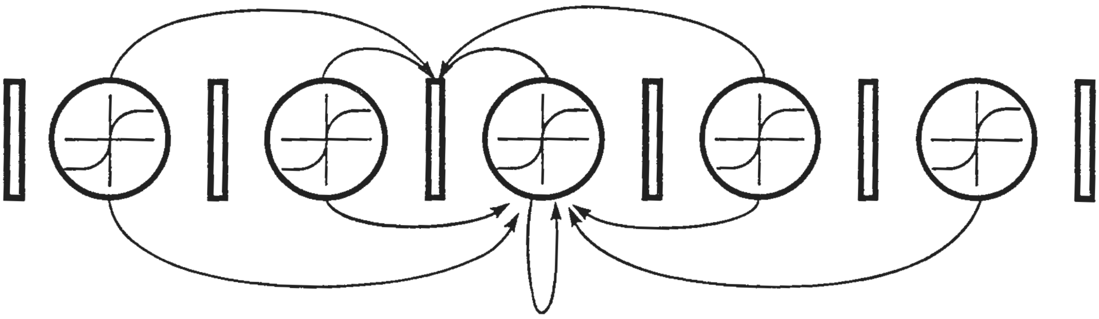


Figure 7..1: Schematic diagram of the network for brightness illusion.

This equation describes the change in the membrane potential of a line process neuron. $u(x)$ is the membrane potential. The second equation determines the input-output relationship.

From the coefficient of the differential equations 7.4, the synaptic weight coefficients between neurons are determined. Figure 7..2 shows their values. Although Eqs. 7.3 and 7.4 are written with a continuous expression of space x, Fig. 7..2 is displayed with a spatially discrete representation of these equations with a number of neurons. In this procedure, we make x discrete and use the central difference scheme to approximate the fourth and second spatial derivatives. Then the connection weights between neurons are derived from these schemes to make x discrete.

(a) shows the interaction between luminance units. The output of the luminance unit indexed by n excites itself. The nearest-neighboring luminance units indexed by $n-1$ and $n+1$ have inhibitory connections, and the next-nearest neighbors indexed by $n-2$ and $n+2$ have excitatory connections on the unit n. (b) shows a functional form of the Gabor filter. (c) shows the interactions between a line detector unit and the luminance units. Two hidden units indexed by h and $h+1$ are shown in this figure. The luminance units $n-1$ and $n+1$ have excitatory connections, and n has an inhibitory connection with h. The hidden unit $h+1$ also has the same connections from n, $n+1$ and $n+2$. The summation of outputs from h and $h+1$ excites the line detector unit m. (d) shows a functional form of the DOG (difference of Gaussian) filter. It appears that the connections between the intensity process are similar to the Gabor function (compare (a) and (b) of Fig. 7..2). The interaction between the line detector unit and the luminance unit is similar to DOG or the Laplacian Gaussian.

In general, the Laplacian operation can be regarded as one example of the inverse optics, because it is useful for detecting edges from a two-dimensional image. On the other hand, inhibition from the line process to the intensity processes can, in a broad sense, be considered one example of the optics, because two sides of an edge are different regions if the edge exists. Therefore, this two-layered network simultaneously realizes the optical and inverse optical operations for visual reconstruction.

7.1.3 Simulation results

Figure 7..3 shows the experimental data given by Lowry and De Palma (1961). The dashed line shows the stimulus intensity profile. Open circles show the

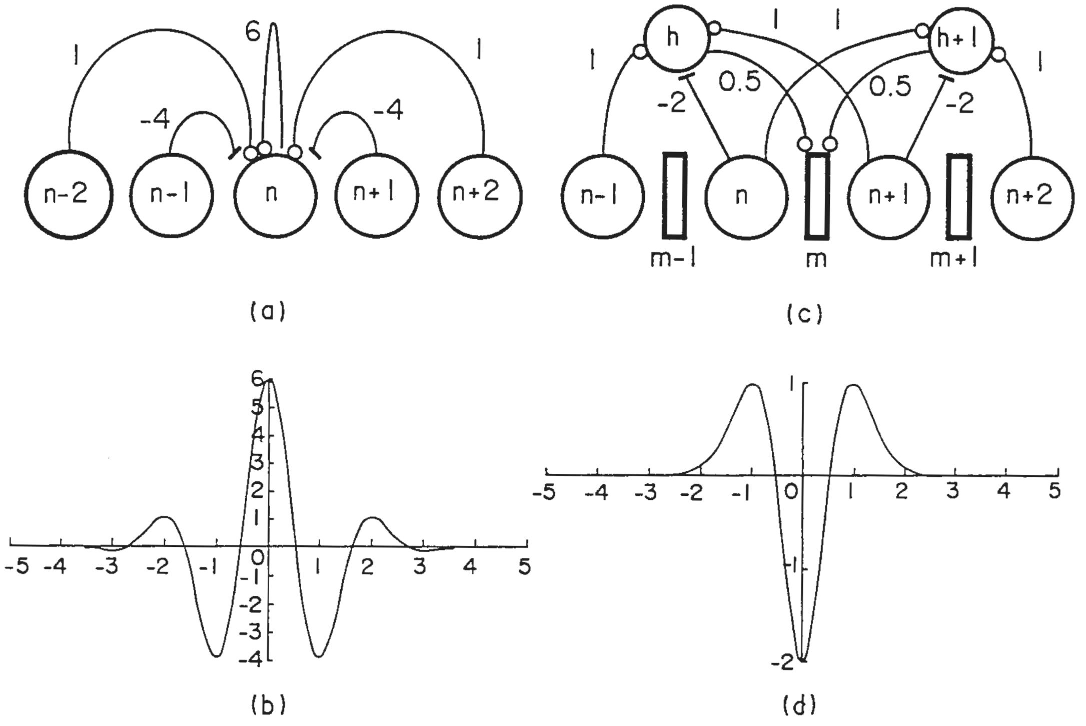


Figure 7..2: (a) Synaptic interactions among intensity units. The numbers in the figure indicate the synaptic weights. (b) Shape of the Gabor filter. (c) Synaptic interactions between intensity, hidden and line-process units. (d) Shape of the difference of Gaussian.

subjective brightness.

Figure 7..4 shows the result of the simulation. The values of the parameters are shown in the figure caption. The output of the network is very close to the experimental data by Lowrey & De Palma (1961).

Figure 7..5 shows the experimental data given by Fiorentini and Radici (1957). The dashed line shows the stimulus intensity profile. The solid line shows the subjective brightness.

Figure 7..6 shows the result of the simulation. The values of the parameters of the network are also described in the figure caption. They are the same as those used in the simulation of the experiment by Lowrey and De Palma (1961), except for λ. As λ determines the signal-to-noise-ratio, it is reasonable that λ is changed for different experimental conditions and subjects. The output of the network is also very close to the experimental data by Fiorentini and Radici (1957).

The performance of our two-layered neural network closely approximates the psychophysical data. In this simulation, the data given to the network are contrast. If the data are binocular disparity, the network reconstructs the depth of the surface as in the case of brightness. However, it has been reported that we cannot see Mach bands in the depth domain for random-dot stereograms (Brookes & Stevens, 1989). If this is the case, we should adopt the membrane-type potential (second derivative) as the smoothness term of the energy function for computation of disparity.

Another line of evidence also suggests a membrane-type potential, even for luminance reconstruction. It is well known that the second derivatives of the intensity profile are calculated in the early process, probably in the retina. Therefore, we are now examining a new neural network model. In this new model, it is assumed that there are two kinds of visual channels: one transmits the second derivative of the intensity, and the other transmits the absolute level of the intensity. We call the latter channel a "luminance unit". We adopt the membrane-type potential as the smoothness constraint for the network. For energy minimization, the second derivative of the data must be calculated. This second derivative is different from the spatial derivative executed in the retina. The network executes filling-in and detection of intensity discontinuity, as in the previous network. The difference between the old and new model is the interpolation function. We are currently examining the performance of the new network. The brightness computation which we mentioned before will be reproduced by the new model.

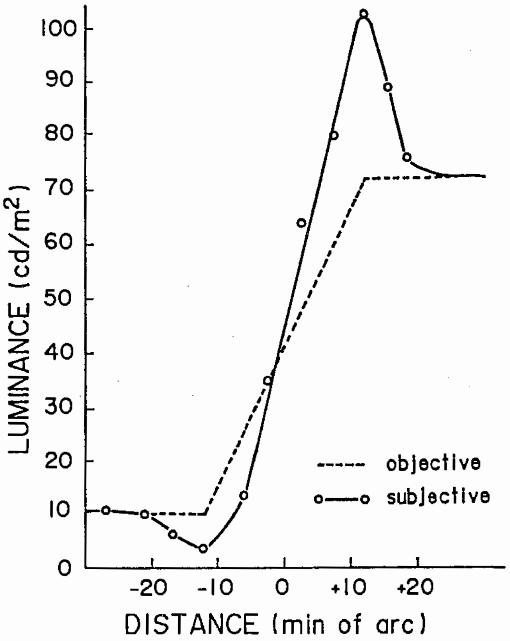


Figure 7..3: Subjective brightness distribution (open circle) and luminance profile (dashed line). (From Lowry and De Palma, 1961).

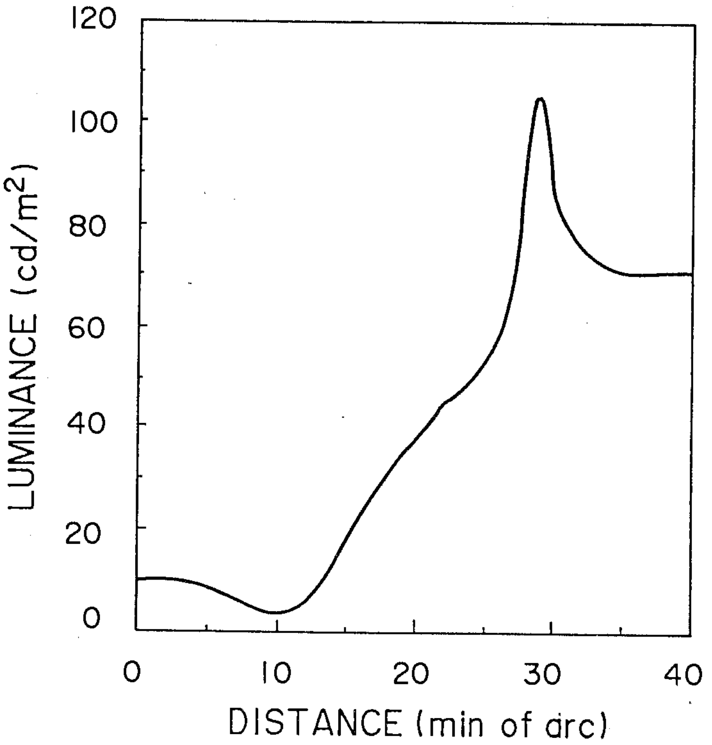


Figure 7..4: Simulation results for the data by Lowry and De Palma (1961). Parameter values used in this simulation are: λ=1.4, C_{50}=0.1, $n(+)$=3.4, $n(-)$=1.7.

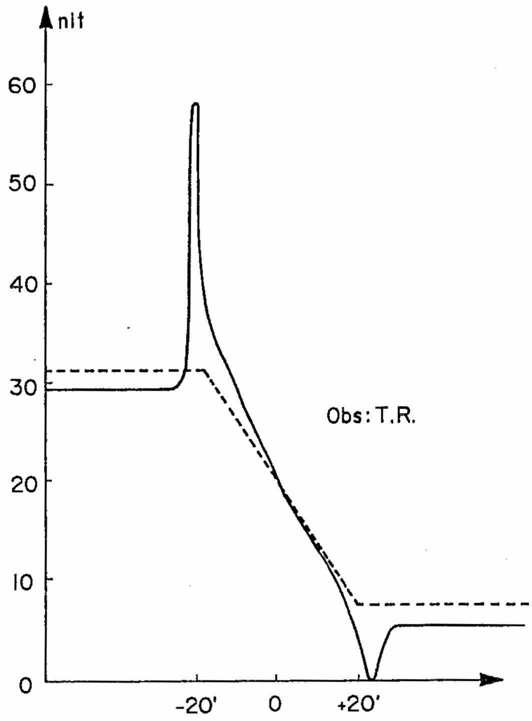


Figure 7..5: Subjective brightness distribution (solid line) and luminance profile (dashed line). (From Fiorentini and Radici, 1957).

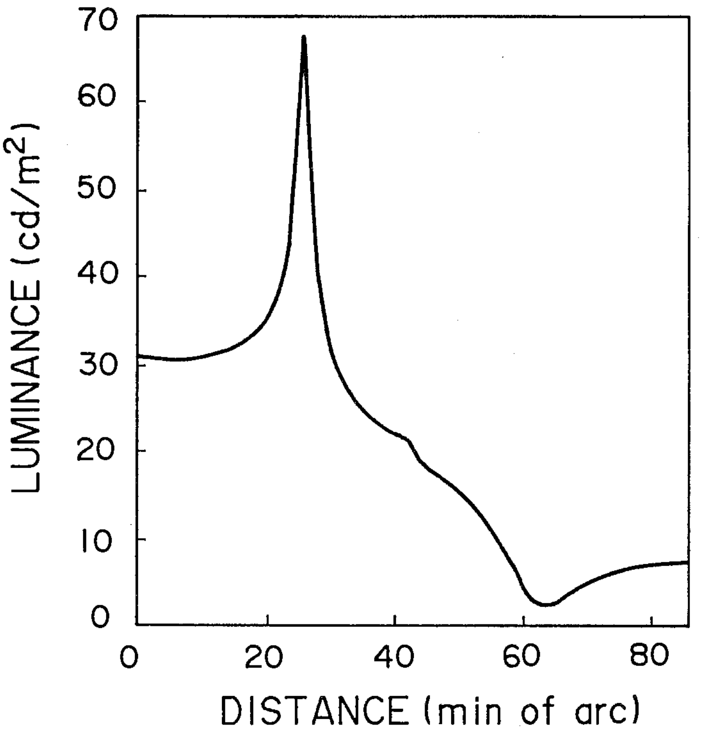


Figure 7..6: Simulation results for the data by Fiorentini and Radici (1957). Parameter values used in this simulation are the same as that of Lowry and De Palma's data except for λ; λ= 2.8.

7.2 Craik-O'Brien-Cornsweet effect

7.2.1 Phenomenon

A two-dimensional luminance pattern which is uniform except for a narrow differential region of the luminance is perceived as a step luminance pattern (see Fig. 7..7). This illusion is called the Craik-O'Brien-Cornsweet effect (COCE). Fig. 7..7 shows a one-dimensional slice across a two-dimensional luminance profile which induces COCE (a). The left side luminances are equal to those on the right. This profile looks like a luminance step (b). Burr (1987) reported that the COCE occurs only when the contrast is very low.

In his experiment, a pattern which has two narrow differential regions was used (see Fig. 7..8). When the illusion typically occurs, one cycle of a square wave with period T is perceived. The patterns were generated by microcomputer and displayed on the face of an oscilloscope at 100 frames/sec, 1000 lines/frame. The oscilloscope face was surrounded by a 1 m square screen uniformly floodlighted to 400 cd/m^2, the same mean luminance as the oscilloscope. One cycle of the grating was displayed for all conditions. Spatial frequency was varied by varying viewing distance, from 0.22 to 200 m, using mirrors and reverse binoculars where necessary.

Observers were required to match the brightness of the central bar of the waveform with that of the square wave. Either the COCE wave form or the square wave could be displayed on the screen at the observer's request. Observers looked alternately at the two stimuli, and adjusted the contrast of the square wave until the brightness of its central bar was comparable to that of the COCE stimulus. Judgements were made on the overall brightness of the bar, ignoring the regions adjacent to the border: the rest of the waveform appeared uniformly bright. Five measurements were made for each condition, and the average reported (Burr, 1987).

Fig. 7..9 shows the apparent brightness produced by the COCE, as a function of stimulus contrast (Burr, 1987). The fundamental spatial frequency $2\pi/T$ of the waveform is 0.25 (solid circles) and 2 (open circles) c/deg. The broken line indicates the equivalent of the two stimuli. The effect of contrast is brought out in this figure, which reports brightness matches as a function of the peak-to-peak contrast of the COCE stimulus.

The measurements were made with the fundamental spatial frequency of 0.25 c/deg and 2 c/deg. At low contrasts, (up to about 0.06) the matched contrast for brightness is similar to that of the stimulus. At higher contrasts, the matched contrast continues to increase at a lower rate up to a contrast of

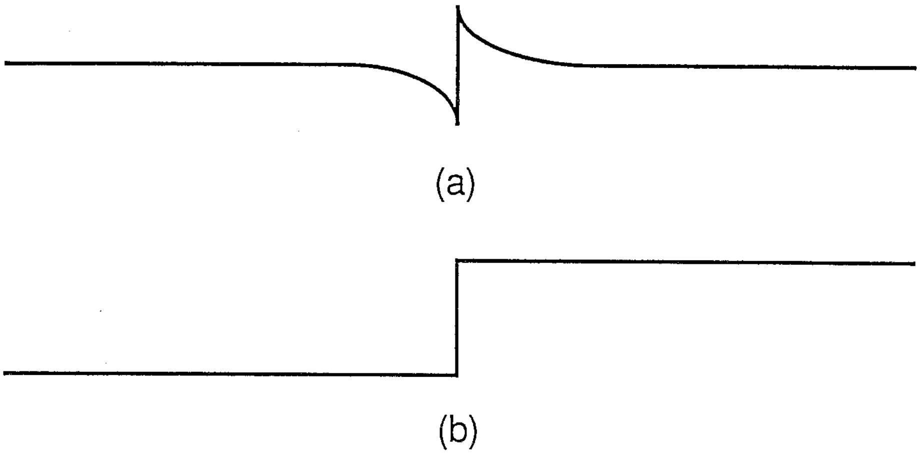


Figure 7..7: (a) A one-dimensional slice across a two-dimensional COCE luminance profile. The left side luminances are equal to those on the right. (b) This profile looks like a luminance step.

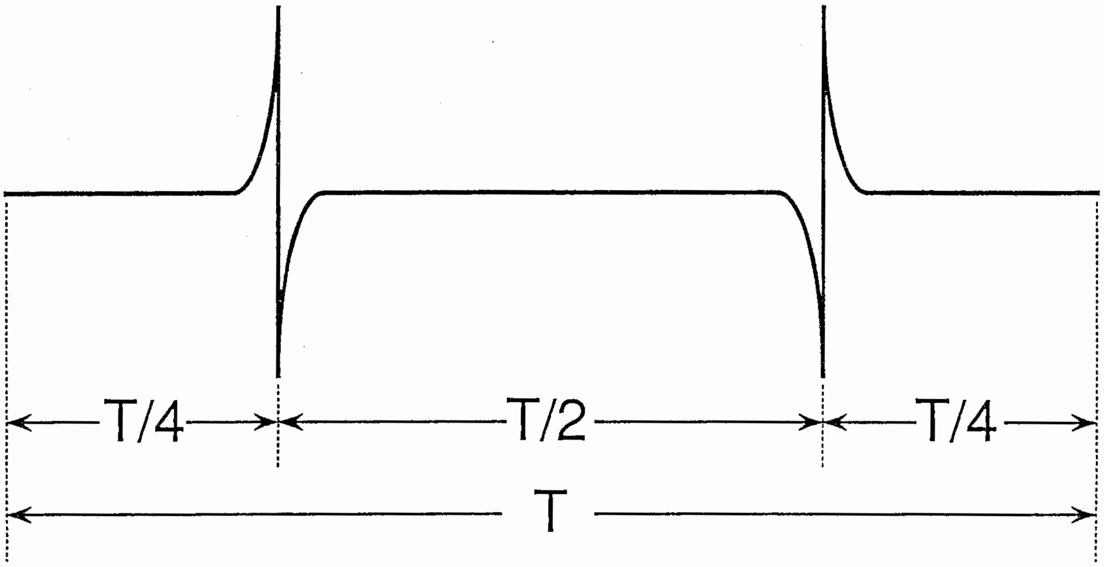


Figure 7..8: A one-dimensional slice of the COCE waveform which is used in the Burr measurement. Typically, when the illusion occurs, one cycle of a square wave whose period is T is perceived.

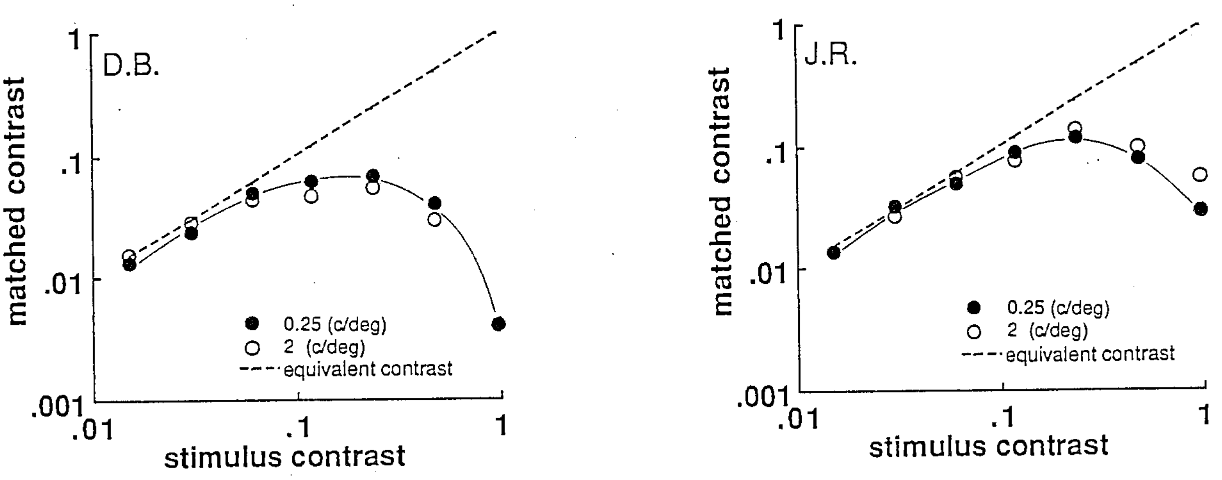


Figure 7..9: The apparent brightness produced by the COCE, as a function of stimulus contrast (Burr, 1987).

about 0.2, whereupon it decreases.

7.2.2 Simulation results

In this section we show the simulation results of COCE illusion. Here, we emphasize that the simulation was conducted using essentially the same neural network model used in the simulation of the Mach bands. Thus, our model is coherent in reproducing both the Mach bands and the COCE illusion.

We used a luminance pattern which has two differential regions according to the experiment by Burr (1987). If the illusion does occur, the central region will be perceived to be darker than the left and right regions. The thirty-two intensity units and the thirty-one line detectors placed between the intensity units are used in this simulation. The following parameter values are used: λ=2.8, γ=104 and ϵ=7.2×10^{-5}.

The contrast under which this illusion is perceived is less than 10% of that under which the Mach band is perceived. When the contrast is low, the input-output characteristic function of the neuron can be approximated by the linear function. Thus, the nonlinear input-output function is not used in this simulation.

The initial values of the intensity units are represented by the solid line in Fig. 7..10. The data around the mean luminance are used only as the initial values, but the maximum and minimum values which are shown by open circles are also used as the continuous input data.

The results of the simulation are shown in Fig. 7..11. The dashed line shows the initial value of each luminance unit, and the circles show the continuous input. The solid line shows the simulated brightness perception. When the input contrast is small, the COCE occurs (a). On the other hand, when the contrast exceeds 6%, the discontinuities on the gentle slope are detected and the illusion does not occur (b, c, d). These results agree with Burr's measurement.

Some of the results shown in this subsection have been previously reported in preliminary form (Inui, Kawato & Hongo, 1990; Hongo, Inui & Kawato, 1990).

7.3 Comparison with other models

Todorovic (1987) discussed the advantages and disadvantages of different types of theoretical attempts to explain the structure of the brightness percept in

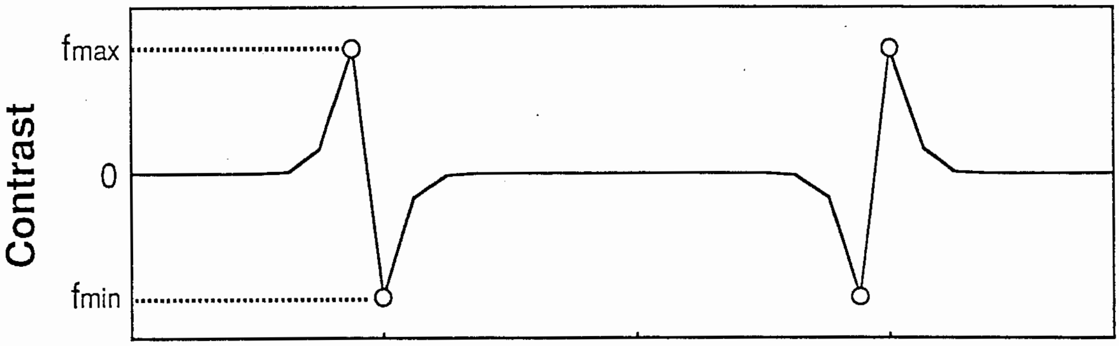


Figure 7..10: A one-dimensional COCE luminance profile which is used in the simulation. The initial value of each luminance unit is shown by the solid line. The maximum and minimum contrasts which are shown by open circles are given continuously.

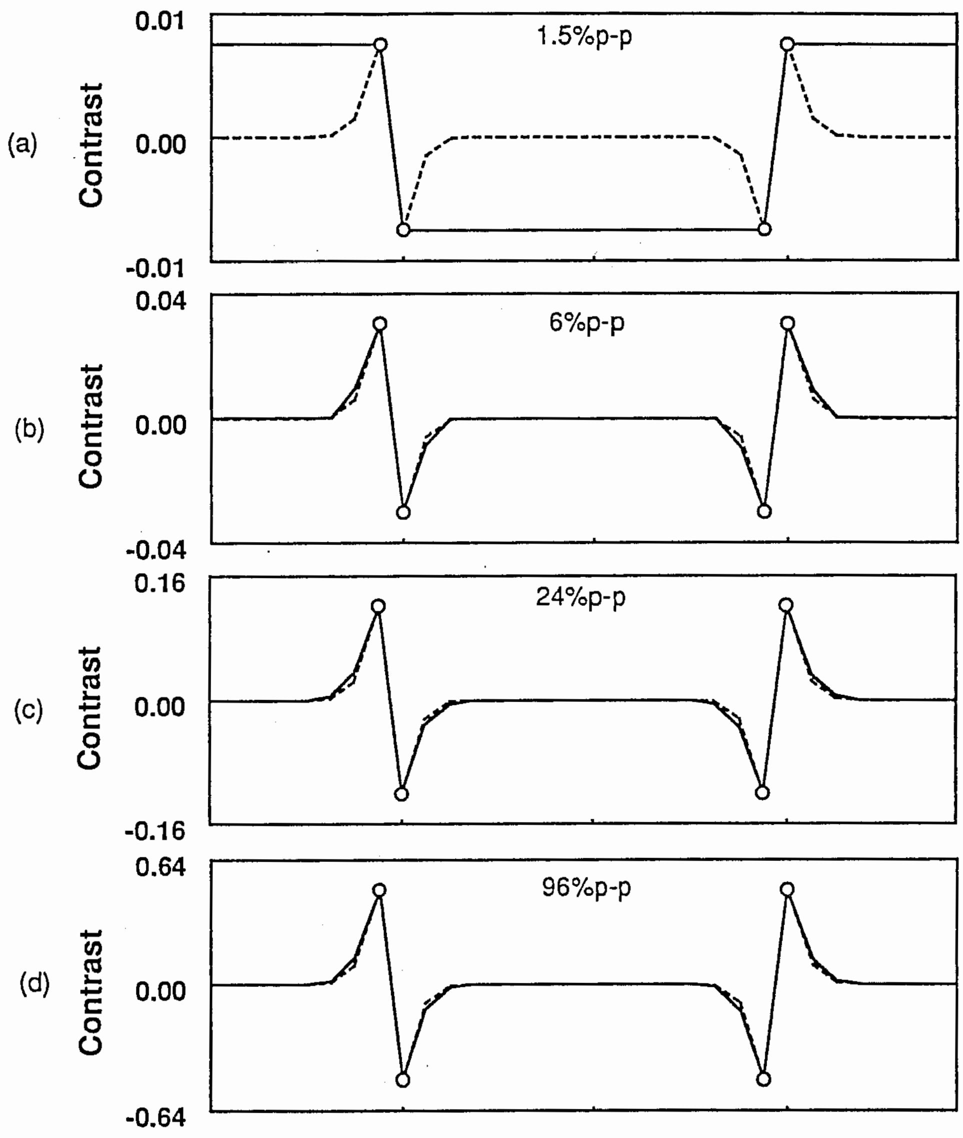


Figure 7..11: Simulation of the COCE for the input contrast from 1.5% to 96% peak to peak.

the COCE. He developed new visual displays related to the COCE and estimated their relevance for theories. He classified theories of the effect into the nonisomorphistic, the cognitive, and the mechanistic approaches.

The guiding idea of nonisomorphistic approach is that the similarity of the profiles of neural activities induced by the steps and cusps is sufficient to explain the similarity of percepts; it is not necessary to require isomorphism, that is, the identity of shapes of spatial distributions of percepts and the underlying neural activities. This approach is criticized as being incomplete. For example, it can not explain the COCE disappearance, when a luminance cusp is embedded into a uniform background (Todorovic, 1987).

The cognitively oriented approach is founded on the idea: "the source of this illusion may be situated at a much higher level, where neural excitation patterns are recognized and interpreted". This approach is criticized as inadequate based on the idea which we cognitively know that apart from the gradients, each region is equiluminant.

Mechanistic theories are divided into integration theories, which model the visual processing stage with standard mathematical operations such as differentiation and integration, and filling-in theories, which stress the processing dynamics of interconnected neural networks. The reported variations of the COCE support the latter class of theories.

Cohen and Grossberg (1984) proposed a filling-in theory which uses a physiological framework, invoking the dynamics of neural processing at different levels of the visual system. Our model is similar to their idea in its basic concept. However, they did not incorporate any mechanism which explains the disappearance of the two kinds of illusions under high-contrast condition. Our model is coherent in that it explains the Mach bands and COCE with a single mechanism.

Morrone and Burr (1988) proposed the local energy model of feature detection. In the model, images are first filtered by pairs of matched filters, one with even-symmetric line-spread functions, the other with odd-symmetric line-spread functions. Then, the outputs of the two filters are squared and summed to give the square of local energy. Features are signaled by peaks in local energy functions. The nature of the features is determined by evaluating the linear responses of the two types of operators at the peaks in local energy. At the positions where local energy peaks on the step-edge, the response of the odd symmetric operator is maximal and the even-symmetric operator zero. This is the signal for an edge. On the other hand, when the local energy is maximal and the odd-symmetric response is zero, it is the signal for a line.

According to the model, we see the Mach bands because at the knee points of the trapezoidal pattern, there are peaks of energy where there is a strong even-symmetric response (Ross, Morrone & Burr, 1989). Similarly, the COCE occurs because at the non-uniform region of the luminance pattern, peaks of local energy occur signaling an edge (Burr & Morrone, 1990). In the local energy model, the peak values will change linearly with the contrast of the pattern, since the model is fundamentally linear. However, this is not the case: these two types of illusion appeared nonlinear. As already mentioned in the previous sections, the COCE disappears for high contrast with exactly the same spatial pattern. Similarly, we found that the Mach band disappears for high contrast with exactly the same spatial pattern (Hongo, Kawato & Inui, unpublished observation). These disappearance phenomena in the COCE and the Mach band can not be explained by the local energy model.

8. Learning of local, parallel and hierarchical model

V1 has marked synaptic plasticity as revealed by studies of orientation selective cells and the ocular dominance column. HVC seems to have even more synaptic plasticity.

In the model shown in Fig. 4..2, the intrinsic connection, the feedforward connection and the feedback connection in this order for the higher visual cortex are more acquired by learning rather than inherently wired. In other words, the feedback connection to V1 is essentially genetically determined. This is because the image generation procedure has common physical properties, while a higher-order abstract model of the visual world, such as O in IT, depends heavily on the environment of individual organisms.

Authors and colleagues have been studying learning acquisition of the coupled MRF and recurrent neural networks from examples of images (Kawato, Ikeda & Miyake, 1988, Kawato, Ikeda, Sonehara, Inui & Miyake, 1989). From these studies, we found that if we provide teaching signals both for $\mathbf{S}$ and I, the image model can be acquired both for synthetic images and natural images (Hongo, Kawato, Inui & Miyake, 1989) irrespective of MRF or recurrent networks (Kawato, Ikeda, Sonehara, Inui & Miyake, 1989).

However, while providing teaching signals about $\mathbf{S}$ might be important for some engineering applications, it is completely inappropriate as a model of learning in the brain. The reason is that the assumption of the existence of the teaching signal for $\mathbf{S}$ is equivalent to the assumption of some other brain area which can obtain a correct estimation of $\mathbf{S}$ without the "visual" cortical areas.

It is well conceivable that sensory modalities other than vision, navigation or object manipulation help the visual areas to acquire a good estimate of $\mathbf{S}$. However, if these other means were able to obtain a rigorous estimate of $\mathbf{S}$, there is no reason for existence of the visual system. Consequently, it is very important to examine whether or not the image model can be acquired when only information about I is provided.

We experimentally found that it is indeed possible (Okamoto & Kawato, 1989 unpublished observation; Ohtsuki & Kawato, 1991). First, we generated an image $(I, \mathbf{S})$ from the coupled MRF with specific energy parameters. Then we gave only I to another MRF with different energy parameters. We used the learning method which we developed for the coupled MRF based on the maximal likelihood criterion (Kawato, Ikeda & Miyake, 1988). As a natural modification of our learning rule for the stochastic model to a deterministic model, we then propose a cross-covariance learning rule for visual cortical areas.

The neural networks shown by broken curves in Fig. 5..2 which realize $-\partial U(\mathrm{S})/\partial \mathrm{S}$ and $-\partial V/\partial \hat{\mathrm{S}}$ contain multiple synapses and interneurons, although they are not shown. By changing the strength of synaptic weights included in this circuit, we can modify $-\partial U(\mathrm{S})/\partial \mathrm{S}$ and $-\partial V/\partial \hat{\mathrm{S}}$, and consequently modify probabilities P associated with internal models of the visual world.

We now consider one synapse included in this circuit. Let us denote short-time average of the firing frequency of the presynaptic fiber by $f(t)$. $v(t)$ denotes local postsynaptic potential. w represents the synaptic weight. The change of w is proportional to the difference between the covariance of f and v while the visual cortex changes its state according to eqs. 5.4, 5.5, 5.6 and 5.7 and the covariance when the visual input faded out, and the first and the second terms of eq. 5.5 do not exist.

$$\Delta w = \overline{(f - \overline{f})(v - \overline{v})}^{input} - \overline{(f - \overline{f})(v - \overline{v})}^{reconst}, \tag{8.1}$$

here the overline is time average. The first term is calculated by time average while the calculation is done as shown in Fig. 5..2 for 100 or 200 msec after a new image I impinges on the retina because of, for example, saccade. Then, the visual input system adapts and the calculation of the first term and the second term of eq. 5.5 is not carried out. Accordingly, independent of the image data impinging on the retina (although the initial condition is affected), the visual cortical areas start to reconstruct the visual world based only on forward and inverse models of optics, and internal models of the visual world.

While the latter state lasts for 200 to 300 msec, the second term of eq. 8.1 is calculated. This learning equation can be regarded as a natural extension of our previous learning rule for MRF to the case of a recurrent neural network model. This is partly inspired by experimental findings of covariance learning in the hippocampus (Stanton & Sejnowski, 1989). Mitchison (1989) has independently proposed a closely related learning algorithm.

The most important feature of the cross-covariance learning rule proposed above is that the period when the visual input is really conveyed to the visual cortices alternates quite rapidly (100 to 300 msec) with the period when the input is shut down and information processing is done closed within the cortex. Several psychological experimental data support this hypothesis (Inui & Miyamoto, 1981, Sperling, Budiansky, Spirak & Johnson, 1971). In typical eye movement, fixation and saccade alternate. Each fixation period is typically 250 to 300 msec and the saccade lasts for several tens of milliseconds. It is well known that visual input is suppressed during the saccade. Furthermore, it recently became clear that the visual information is acquired only for 100 msec or so after the saccade, and the rest of the duration is not used simply for acquiring visual input. That is, from the psychological measurement, the visual information about the outer world does not seem to be processed after the first 100 msec after the saccade. While the visual input is shut down, the visual cortical areas can reconstruct or generate the representation of the visual world based on the feedback and feedforward connections between areas and the intrinsic connections within areas, as shown in Fig. 5..2. Synaptic weights change so that this reconstructed visual world is in good agreement with the real visual world according to eq. 8.1.

9. Discussion

In this paper we proposed computational and neural network models which coherently explain early, middle and high-level visions based on physiological and anatomical knowledge of the visual cerebral cortices. Feedforward neural connections, feedback neural connections and the intrinsic neural connections are assumed to provide models of the approximated inverse of image generation, the forward models of image generation, and internal models of the visual world, respectively. All these connections are topographical and local, so a single layer can provide a very local model of the visual world (Markov random field model in this sense). However, the brain makes up a global and abstract model of the visual world by folding many layers. In the HVC, which is higher than V1 by several levels, the receptive fields of the neurons are much larger than those in V1 because of convergence of the feedforward neural connections. Consequently, although the intrinsic neural connection in the HVC is still local physically on the cortex surface, it provides a global interaction in the visual field. This design principle is appealing also from the engineering point of view (Hongo, Kawato & Inui, 1991).

Finally, we discuss visual memory and pattern recognition within the framework given in Fig. 4..2. Even if there is no visual input I, if neurons in some HVC are excited, then all areas in Fig. 4..2 are activated because of the feedback and feedforward connections between areas. This corresponds to recall of imagery. In this sense, the internal model of O in IT can be said to possess memorized visual imagery. Rolls (1989a, 1989b) proposed a computational and conceptual model of the hippocampus and the visual cortical areas. In this model he emphasized the functional significance of feedback connections (backprojection in his terminology) in visual pattern recognition. Rolls proposed 7 possible functional roles of this backprojection. We believe that our mathematical interpretation of the feedback connections as a model of the image generation process is fairly compatible with his interpretations.

The pattern recognition is equivalent to convergence of all the states of all areas in Fig. 4..2 to stable equilibria by interaction with feedforward, feedback and intrinsic connections. That is, when all three conditions, feature extraction from the image data, reconstruction of the image data from memorized mental imagery, and prediction by internal model of the visual world, are satisfied for all the description levels from s_1 to s_{14}, the total system is said to recognize something. The fundamental design principles of the visual cortices are parallelism and hierarchy with local connections. The three main parallel flows of information (shape, color, motion) avoid a combinatorial explosion of different attributes. Feedforward connections make relatively fast relaxation possible. Interactions between different areas through feedforward and feedback connections guarantee convergence to a consistent solution all over the visual cortices.

It is known that the hippocampus has an important role in recognition memory, and several models have already been proposed (McNaughton & Morris, 1987, Rolls, 1989a). Recently, it was found that the hippocampus has several reciprocal connections between V4, IT, MT, area 7, area 22 (Van Hosen, 1982). Area 7 is related to spatial perception, and area 22 is related to auditory perception. We assume that hippocampus integrates the various kinds of information into a single episode; visual information for an object is combined with auditory and spatial information which can be cues when it is recalled. Integration of this type is called 'vertical association' (Wickelgrem, 1979). When the auditory or spatial information is given, the reciprocal connections work and the signal will go down from the hippocampus to the higher visual areas, and in turn, the neurons in the lower level (e.g., V2 and V3) will be activated through the reciprocal connections among the visual areas. The state of activation of the visual areas corresponds to the image recall.

In general, feedforward connections calculate the inverse optics (roughly speaking, the pseudo-inverse matrix from lower data to the higher-order solution space), and backward connections calculate optics (higher-order estimation to lower-order data). In other words, they perform hypothesis generation and verification respectively. For example, for shape from shading, feedforward connections compute the surface orientation roughly from the 2D image by a one-shot algorithm (using natural law as a constraint). On the hand hand, feedback connections calculate the 2D image from the estimated surface orientation. Through these loops the network estimates the orientation rapidly and exactly.

References

[1] Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann Machines. *Cognitive Science, 9*, 147-169.

[2] Albrecht, D. G., & Hamilton D. B. (1982). Striate cortex of monkey and cat: contrast response function. *Journal of Neurophysiology, 48*, 217-237.

[3] Ballard, D. H., Hinton, G. E., & Sejnowski, T. J. (1983). Parallel visual computation. *Nature, 306*, 21-26.

[4] Bekesy, G. von (1968). Mach- and Hering-type lateral inhibition in vision. *Vision Research, 8*, 1483-1499.

[5] Brookes, A., & Stevens, K. A. (1989). The analogy between stereo depth and brightness. *Perception, 18*, 601-614.

[6] Buckley, D., Frisby, J. P., & Mayhew, E. W. (1989). Integration of stereo and texture cues in the formation of discontinuities during three-dimensional surface interpolation. *Perception, 18*, 563-588.

[7] Bülthoff, H. H. and Mallot, H. H. (1988). Integration of depth modules: stereo and shading. *Journal of the Optical Society of America, A, 5*, 1749-1758.

[8] Burr, D. C. (1987). Implications of the Craik-O'Brien illusion for brightness perception. *Vision Research, 27*, 1903-1913.

[9] Burr, D. C. & Morrone, M. C. (1990). Feature detection in biological and artificial visual systems. In C. Blakemore (Ed.), *Vision: coding and efficiency*, (pp.185-194) Cambridge: Cambridge University Press.

[10] Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 8*, 679-698.

[11] Cavanagh, P. (1987). Reconstructing the third dimension: Interactions between color, texture, motion, binocular disparity, and shape. *Computer Vision, Graphics, and Image Processing, 37*, 171-195.

[12] Cohen, M. A. & Grossberg, S. (1984). Neural dynamics of brightness perception: Features, boundaries, diffusion, and resonance. *Perception & Psychophysics, 36*, 428-456.

[13] Coren, S. (1972). Subjective contours and apparent depth. *Psychological Review, 79*, 356-367.

[14] Daugman, J. G. (1988). Pattern and motion vision without Laplacian zero crossings. *Journal of the Optical Society of America, A, 5*, 1142-1148.

[15] Desimone, R., Schein, S. J., Moran, J., & Ungerleider, L. G. (1985). Contour, color and shape analysis beyond the striate cortex. *Vision Research, 25*, 441-452.

[16] Desimone, R., & Schein, S. J. (1987). Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. *Journal of Neurophysiology, 57*, 835-868.

[17] Felleman, D. J., & Van Essen, D. C. (1987). Receptive field properties of neurons in area V3 of macaque monkey extrastriate cortex. *Journal of Neurophysiology, 57*, 889-920.

[18] Fiorentini, A., & Radici, T. (1957). Binocular measurements of brightness on a field presenting a luminance gradient. *Atti della fondazinne Giorgio Ronchi anno XII*, 453-461.

[19] Fiorentini, A. (1972). Mach band phenomena. In D. Jameson & L. M. Hurvich (Eds.), *Handbook of sensory physiology, VII/4, Visual Psychophysics* (pp.188-201) New York: Springer-Verlag.

[20] Fuster, J. M. (1990). Inferotemporal units in selective visual attention and short-term memory. *Journal of Neurophysiology, 64*, 681-697.

[21] Galletti, C., & Battaglini, P. P. (1989). Gaze-dependent visual neurons in area V3A of monkey prestriate cortex. *The Journal of Neuroscience, 9*, 1112-1125.

[22] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions Pattern Analysis Machine Intelligence, 6*, 721-741.

[23] Gerrits, H. J. M. & Vendrik, A. J. H. (1970). Simultaneous contrast, filling-in process and information processing in man's visual system. *Experimental Brain Research, 11* 411-430.

[24] Gilbert, C. D. & Wiesel, T. N. (1983). Clustered intrinsic connections in cat visual cortex. *Journal of Neuroscience, 3*, 1116-1133.

[25] Grimson, W. E. L. (1981). *From Images to surfaces: A computational study of the early visual system.* Cambridge, MA: MIT Press.

[26] Grimson, W. E. L. (1982). A computational theory of visual surface interpolation. *Philosophical Transactions of Royal Society of London, B298*, 395-427.

[27] Hartline, H. K. & Ratliff, F. (1957). Inhibitory interaction of receptor units in the eye of Limulus. *Journal of General Physiology, 40*, 357-376

[28] Hayakawa, H., Inui, T., & Kawato, M. (1991). Computational theory and neural network model of perceiving shape from shading in monocular depth perception. *Proceedings of International Joint Conference on Neural Networks.*

[29] Hinton, G. E. (1989). Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation, 1*, 143-150.

[30] Hochberg, J. E., Triebel, W. & Seaman, G. (1951). Color adaptation under conditions of homogeneous stimulation (ganzfeld). *Journal of Experimental Psychology, 41*, 153- 159.

[31] Hongo, S., Kawato, M., Inui, T. & Miyake, S. (1989). Contour extraction of images - Local, parallel and stochastic algorithm which learns energy parameters. *Proceedings of International Joint Conference on Neural Networks, 1*, 161-168.

[32] Hongo, S., Inui, T., & Kawato, M. (1990). A computational theory and a neural network model on the brightness perception: A Markov random field model for the filling-in process. *Japanese ITEJ Technical Report, 14*, 1-6.

[33] Hongo, S., Kawato, M. & Inui, T. (1991). Contour extraction of natural images based on a multi-layered MRF model: A two-resolution model. *Proceedings of Australian Conference on Neural Networks*, 102-106.

[34] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA, 79*, 2254-2258.

[35] Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences USA, 81*, 3088-3092.

[36] Hopfield, J. J., & Tank, D. W. (1985). Neural computation of decisions in optimization problems. *Biological Cybernetics, 52*, 141-152.

[37] Horn, B. K. P. (1975). Obtaining shape from shading information. In P. H. Winston (Ed.), *The Psychology of Computer Vision.* (pp.115-155). New York: McGraw-Hill.

[38] Horn, B. K. P. (1977). Understanding image intensities. *Artificial Intelligence, 8*, 201-231.

[39] Horn, B. K. P., & Brooks, M. J. (1989). *Shape from Shading.* Cambridge, MA: The MIT Press.

[40] Houser, C. R., Vaughn, J. E., Hendry, S. H. C., Jones, E. G., & Peters, A. (1984). GABA neurons in the cerebral cortex. In E. G. Jones & A. Peters (Eds.), *Cerebral cortex, 2, Functional properties of cortical cells* (pp.63-89). New York:Plenum Press.

[41] Hubel, D. H., & Livingstone, M. S. (1987). Segregation of form, color and stereopsis in primate area 18. *The Journal of Neuroscience, 7*, 3378-3415.

[42] Hubel, D. H., & Livingstone, M. S. (1990). Color and contrast sensitivity in the lateral geniculate body and primary visual cortex of the macaque monkey. *The Journal of Neuroscience, 10*, 2223-2237.

[43] Hurlbert, A. & Poggio, T. (1988). Synthesizing a color algorithm from examples. *Science, 239*, 482-485.

[44] Iba, Y. (1989). Bayesian statistics and statistical mechanics. In H. Takayama (Ed.), *Cooperative dynamics in complex physical systems* (pp. 235-236). Berlin, Heidelberg; Springer-Verlag.

[45] Inui, T. & Miyamoto, K. (1981). The time needed to judge the order of a meaningful string of pictures. *J. Experimental Psychology: Human Leaning and Memory, 7*, 393-396.

[46] Inui, T., Kawato, M. & Hongo, S. (1990). Computational theory and its network implementation on the visual reconstruction process. *Japan IEICE Technical Report, NC90-21*, 61-68.

[47] Inui, T., Hongo, S. & Kawato, M. (1990). A computational model of brightness illusion and its implementation. *Proceedings of ECVP90*, 401.

[48] Julesz, B. (1971). *Foundations of cyclopean perception.* Chicago: University of Chicago Press.

[49] Julesz, B. & Chang, J. J. (1976). Interaction between pools of binocular disparity detectors tuned to different disparities. *Biological Cybernetics, 22*, 107-120.

[50] Kaas, J. H. (1986). The structural basis for information processing in the primate visual system. In J. D. Pettigrew, K. J. Sanderson, & W. R. Levick (Eds.), *Visual Neuroscience* (pp.315-340). Cambridge University Press.

[51] Kawato, M., Ikeda, T. & Miyake, S. (1988). Learning in neural networks for visual information processing. *Journal of Japanese Television Society, 42* 918-924.

[52] Kawato, M., Ikeda, T., Sonehara, N., Inui, T. & Miyake, S. (1989). Information processing of image and neural network models. *Journal of Japanese Society for Artificial intelligence, 4*, 143-150.

[53] Kersten, D., O'Toole, A., Sereno, E., Knil, D. & Anderson, J. (1987) Associative learning of scene parameters from images. *Applied Optics, 26*, 4999-5006.

[54] Koch, C., Marroquin, J. & Yuille, A. (1986). Analog "neural" networks in early vision. *Proc. Natl. Acad. Sci. USA, 83*, 4263-4267.

[55] Legge, G. E., & Gu, Y. (1989). Stereopsis and contrast. *Vision Research, 29*, 989-1004

[56] Lehky, S. R., & Sejnowski, T. J. (1989). Simplifying network models of binocular rivalry and shape-from-shading. In C. Koch & I. Segev (Eds.), Methods in neuronal modeling. From synapses to networks (pp.361-396). Cambridge, MA: The MIT Press.

[57] Livingstone, M. S., & Hubel, D. H. (1983). Specificity of cortico-cortical connections in monkey visual system. *Nature, 304*, 531- 534.

[58] Livingstone, M. S., & Hubel, D. H. (1984). Anatomy and physiology of a color system in the primate visual cortex. *The Journal of Neuroscience, 4*, 309-356.

[59] Livingstone, M. S., & Hubel, D. H. (1987a). Connections between layer 4B of area 17 and the thick cytochrome oxidase stripes of area 18 in the squirrel monkey. *The Journal of Neuroscience, 7*, 3371-3377.

[60] Livingstone, M. S., & Hubel, D. H. (1987b). Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *The Journal of Neuroscience, 7*, 3416-3468.

[61] Lowry, E. M., & De Palma, J. J. (1961). Sine wave response of the visual system I. The Mach phenomenon. *Journal of the Optical Society of America, 51*, 740-746.

[62] Lueck, C, J., Zeki, S., Friston, K. J., Deiber, M. P., Cope, P., Cunningham, V. J., Lammertsma, A. A., Kennard, C., & Frackowiak, R. S. J. (1989). The colour centre in the cerebral cortex of man. *Nature, 340*, 386-389.

[63] Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information.* New York:W. H. Freeman and Company.

[64] Marr, D. & Hildreth, E. (1980) Theory of edge detection. *Proceedings of Royal Society of London, B207*, 187-217.

[65] Marr, D. & Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of Royal Society of London, B204*, 301-328.

[66] McNaughton, B. L. & Morris, R. G. M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neuroscience, 10*, 408-415.

[67] Michael, C. R. (1988). Retinal afferent arborization patterns, dendritic field organizations, and the segregation of functions in the lateral geniculate nucleus of the monkey. *Proceedings of the National Academy of Sciences USA, 85*, 4918-4924.

[68] Mitchison, G. (1988). Planarity and segmentation in stereoscopic matching. *Perception, 17*, 753-782.

[69] Mitchison, G. (1989). Learning algorithms and networks of neurons. In R. Durbin, C. Miall & G. Mitchison (Eds.), *The computing neuron* (pp.35-53) Wokingham: Addison-Wesley.

[70] Morrone, M. C., & Burr, D. C. (1988). Feature detection in human vision: a phase-dependent energy model. *Proceedings of the Royal Society of London, B235*, 221-245.

[71] Movshon, J. A., Adelson, E. H., Gizzi, M. S., & Newsome, W. T. (1986). The analysis of moving visual patterns. In C. Chagas, R. Gattass & C. Gross (Eds.) *Pattern recognition mechanisms* (pp.117-151. New York. Springer-Verlag.

[72] Nishihara, H. K. (1988). Hidden information in transparent stereograms. *Proceedings of the Twenty-first Asilomar Conference on Signals Systems and computers*, Pacific Grove, CA, November 2-4, 1987, The Computer Society of the IEEE, Washington, 695-700.

[73] Ohtsuki, H. & Kawato, M. (1991). Training a hierarchical image model without a high-level teacher - Estimation of MRF line-process energy provided only with image intensity data. *Proceedings of International Joint Conference on Neural Networks*, submitted.

[74] Pandya, D. N., & Yeterian, E. H. (1988). Architecture and connections of cortical association areas. In A. Peters & E. G. Jones (Eds.), *Cerebral cortex, 4, Association and auditory cortices* (pp.3-61). New York:Plenum Press.

[75] Parnavelas, J. G. (1984). Physiological properties of identified neurons. In E. G. Jones & A. Peters (Eds.), *Cerebral cortex, 2, Functional properties of cortical cells* (pp.205-239). New York:Plenum Press.

[76] Pentland, A. P. (1982). Finding the illuminant direction. *Journal of the Optical Society of America, 72*, 448-455.

[77] Pentland, A. P. (1984). Local shading analysis. *IEEE Trans. Pattern Analysis. Machine Intelligence, 6*, 170-187.

[78] Pentland, A. P. (1986). Local Shading Analysis. In A. P. Pentland (Ed.), *From pixels to predicates* (pp.40-77). Norwood, NJ: Ablex Publishing Corporation.

[79] Peterhans, E., & von der Heydt, R. (1989). Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps. *The Journal of Neuroscience, 9*, 1749-1763.

[80] Peterson, C. & Anderson, J. R. (1987). A mean field theory learning algorithm for neural networks. *Complex Systems, 1*, 995-1019.

[81] Poggio, G. F., Gonzalez, F., & Krause, F. (1988). Stereoscopic mechanisms in monkey visual cortex: binocular correlation and disparity selectivity. *The Journal of Neuroscience, 8*, 4531-4550.

[82] Poggio, T., Torre, V. & Koch, C. (1985). Computational vision and regularization theory. *Nature, 317*, 314-319.

[83] Poggio, T., Gamble, E.B. & Little, J. J. (1988). Parallel integration of vision modules. *Science, 242*, 436-440.

[84] Ramachandran, V. S. (1988). Perceiving shape from shading. *Scientific American, 259*, 76-83.

[85] Ratliff, F. (1984). Why Mach bands are not seen at the edges of a step. *Vision Research, 2*, 163-165.

[86] Rolls, E. T. (1989a). Functions of neuronal networks in the hippocampus and neocortex in memory. In J. H. Byrne & W. O. Berry (Eds.), *Neural model of plasticity: theoretical and empirical approaches* , (pp. 240-265) New York: Academic Press.

[87] Rolls, E. T. (1989b). The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampus. In R. Durbin, C. Miall & G. Mitchison (Eds.),*The computing neuron* (pp. 125-159) Wokingham: Addison-Wesley.

[88] Ross, J., Holt, J. J., & Johnstone, J. R. (1981). High frequency limitations of Mach bands. *Vision Research, 21*, 1165-1167.

[89] Ross, J. M., Morrone, M. C. & Burr, D. C. (1989). The conditions under which Mach bands are visible. *Vision Research, 2*, 699-715.

[90] Roy, J., & Wurtz, R. H. (1990). The role of disparity-sensitive cortical neurons in signalling the direction of self-motion. *Nature, 348*, 160-162.

[91] Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.) (1986), *Parallel distributed processing: Explorations in the microstructure of cognition Vol. 1, 2.* Cambridge, MA: MIT Press.

[92] Schein, S. J., & Desimone, R. (1990). Spectral properties of V4 neurons in the macaque. *The Journal of Neuroscience, 10*, 3369-3389.

[93] Schiller, P. H., & Logothetis, N. K. (1990). The color-opponent and broad-band channels of the primate visual system. *Trends in Neuroscience, 13*, 392-398.

[94] Sclar, G., Maunsell, J. H. R. & Lennie, P. (1990). Coding of image contrast in central visual pathways of the macaque monkey. *Vision Research, 30*, 1-10.

[95] Shipley, T., & Wier, C. (1972). Asymmetries in the Mach band phenomena. *Kybernetik, 10*, 181-189.

[96] Sperling, G., Budiansky, J., Spivak, J. G. & Johnson, M. C. (1971). Extremely rapid visual search: The maximum rate of scanning letters for the presence of a numeral. *Science, 174*, 307-311.

[97] Stanton, P. K. & Sejnowski, T. J. (1989). Associative long-term depression in the hippocampus induced by hebbian covariance. *Nature, 339*, 215-218.

[98] Stevens, K. A., & Brookes, A. (1988). Integrating stereopsis with monocular interpretations of planar surfaces. *Vision Research, 28*, 371-386.

[99] Tanaka, K., & Saito, H. (1989). Analysis of motion of the visual field by direction, expansion/contraction, and rotation cells clustered in the dorsal part of the medial superior temporal area of the macaque monkey. *Journal of Neurophysiology, 62*, 626-641.

[100] Tanaka, K., Fukada, Y., & Saito,H. (1989). Underlying mechanisms of the response specificity of expansion / Contraction and rotation cells in the dorsal part of the medial superior temporal area of the macaque monkey. *Journal of Neurophysiology, 62*, 642-656.

[101] Todd, J. T., & Reichel, F. D. (1989). Ordinal structure in the visual perception and cognition of smoothly curved surfaces. *Psychological Review, 96*, 643-657.

[102] Todorovic, D. (1987). The Craik-O'Brien-Cornsweet effect: New varieties and their theoretical implications, *Perception & Psychophysics, 42*, 545-560.

[103] Tootel, R. B. H., Hamilton, S. L., & Switkes, E. (1988). Functional anatomy of macaque striate cortex. IV. Contrast and magno-parvo streams. *The Journal of Neuroscience, 8*, 1594-1609.

[104] Tootell, R. B. H., Silverman, M. S., Hamilton, S. L., Switkes, E., & De Valois, R. L. (1988). Functional anatomy of macaque striate cortex. V.Spatial frequency. *The Journal of Neuroscience, 8*, 1610-1624.

[105] Tootell, R. B. H., & Hamilton, S. L. (1989). Functional anatomy of the second visual area (V2) in the macaque. *The Journal of Neuroscience, 9*, 2620-2644.

[106] Torre, V., & Poggio, T. A. (1986). On Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 8*, 147-163.

[107] Toyama, K. (1988). Functional connections of the visual cortex studied by cross-correlation techniques. In P. Rakic & W. Singer (Eds.), *Neurobiology of neocortex* (pp. 203-217). Chichester: John Wiley & Sons.

[108] Ts'o, D. Y., Gilbert, C. D. & Wiesel, T. N. (1986). Relationships between horizontal interactions and functional architecture in cat striate cortexes revealed by crosscorrelation analysis. *Journal of Neuroscience, 6*, 1160-1170.

[109] Tyler, C. W. (1990). A stereoscopic view of visual processing streams. *Vision Research, 30*, 1877-1895.

[110] Ullman, S., & Basri, R. (1989). Recognition by linear combinations of models. *A.I. Memo, 1152*

[111] Van Hoesen, G. W. (1982). The parahippocampal gyrus, New observations regarding its cortical connections in the monkey. *Trends in Neuroscience, 5*, 345-350.

[112] von der Heydt, R., & Peterhans, E. (1989). Mechanism of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *The Journal of Neuroscience, 9*, 1731-1748

[113] Wang, H. T., Mathur, B. & Koch, C. (1989). Computing optical flow in the primate visual system. *Neural Computation, 1*, 92-103.

[114] Wickelgren, W. A. (1979). Chunking and consolidation: a theoretical synthesis of semantic networks, configuring in conditioning, S-R versus cognitive learning, normal forgetting, the amnesic syndrome and the hippocampal arousal system. *Psychological Review, 86*, 44-60.

[115] Wurger, S. M., & Landy, M. S. (1988). Depth interpolation with sparse disparity cues. *Perception, 8*, 39-54.

[116] Yarbus, A. L. (1967). *Eye movements and vision.* New York: Plenum Press.

[117] Yuille, A. L. (1990). Generalized deformable models, statistical physics, and matching problems. *Neural Computation, 2*, 1-24.

[118] Zeki, S. (1983). Colour coding in the cerebral cortex : The reaction of cells in monkey visual cortex to wavelengths and colours. *Neuroscience, 9*, 741-765.

[119] Zeki, S. & Shipp, S. (1988). The functional logic of cortical connections. *Nature, 335*, 311-317.