Construction of the 3D-world in the brain

ADVANCED SEMINAR

submitted by
Simon Bilgeri

NEUROSCIENTIFIC SYSTEM THEORY
Technische Universität München

Supervisors: M.Sc. Lukas Everding, M.Sc. Mohsen Firouzi
Final Submission: 18.01.2016

TECHNISCHE UNIVERSITÄT MÜNCHEN
NEUROWISSENSCHAFTLICHE SYSTEMTHEORIE
PROF. DR. JÖRG CONRADT

2015-10-01

A D V A N C E D S E M I N A R

Construction of the 3D-world in the brain

Problem description:

Reconstructing the shape of the world from visual information is a highly complex problem. One important subtask, that must be solved, is finding the distances to objects in the environment. Brains of several types of mammals have developed the ability to successfully do this using different visual cues from two eyes with overlapping visual fields. Of these cues, binocular disparity is usually seen as the strongest one for the estimation of distances. During the last decades there have been multiple studies on the neural mechanisms behind the distance computing.[1,2,3] For this project, we would like you to investigate how the brain is able to find depth from disparity and how we can use this knowledge to improve computer vision: give an introduction to the neurological mechanisms, present a computational model, find algorithms that take advantage of these neurophysiological findings and explain how exactly they do it.

- Get familiar with the stereo problem
- Describe how the brain solves this problem
- Present a computational model of how the brain solves it
- Compare cortically inspired algorithms with the biological solution

Supervisor: Lukas Everding, Mohsen Firouzi

(Jörg Conradt)
Professor

Bibliography:

[1] S. Georgieva, R. Peeters, H. Kolster, J. Todd and G. Orban *The Processing of Three-Dimensional Shape from Disparity in the Human Brain* The Journal of Neuroscience, 29(3) p. 727-742, Jan. 2009

[2] A.J. Parker, B.G. Cumming *Cortical mechanisms of binocular stereoscopic vision* Progress in Brain Research, Vol. 134, 2001

[3] Y. Zhu, N. Qian *Binocular Receptive Field Models, Disparity Tuning, and Characteristic Disparity* Neural Computation 8 p. 1611-1641, 1996

Abstract

Reconstructing the 3D-world with only two 2D-projections available is a very challenging task. Corresponding points within the two images have to be found in order to compute their disparity. Homogeneous areas or repeating patterns, however, can pose difficulties for finding these matches. Similar to stereo cameras used in artificial systems, the human visual system largely relies on this disparity cue. Although there are also monocular depth cues, navigation or grasping would be arguable more difficult with only 2D vision available. However, the human brain shows to be remarkably good in inferring the third dimension. Thus, this report deals with the question how disparity is computed in the brain and how these insights can be used in computer vision. To do so, firstly a short introduction on the human visual system is given. Neurophysiological findings on how disparity is represented in both striate and extrastriate cortex are summarized. Moreover, a mathematical model of the disparity tuning found in the striate cortex is introduced. Finally, cortical algorithms using the disparity model are analysed.

Contents

Chapter 1

Introduction

The ability to perceive the world in three dimensions has many advantages in our everyday lives. For example, it is easier to navigate knowing the distance to nearby obstacles, or to interact with objects. With only 2D information at hand, it would be arguably more likely to bump into obstacles or to miss the target in an attempt of grasping. Nevertheless, the projected images on each of our eyes retinas are 2D at first. Many artificial systems like autonomous robots rely on stereo cameras as a depth sensor, which also provide a pair of 2D images. Hence, learning from the processes within the highly accurate human visual system, could help to develop artificial visual systems. The question is, how the brain is able to infer the third dimension using two 2D images?
Monocular cues like motion parallax, occlusion or relative size, can help to perceive depth from each 2D image separately. While being in motion, a closer object appears to move faster than far ones. Hence, this motion parallax can be used to infer how far these objects are away. Another cue is given by an object which occludes another object and thus has to be in front of it. Moreover, while observing two similar objects, the smaller one has to be further away [Hub95, Pal99]. However, these monocular cues only provide a crude impression of 3D vision[1]. To explain the highly accurate depth perception of humans, binocular disparity is generally seen as the most important depth cue [Hub95]. Binocular disparity can be defined as the difference in angle between the two corresponding projections relative to the center of the retina. As illustrated in Figure 1.1, the eyes are fixated on some point F with zero disparity. Near points (A) yield a negative (crossed) disparity ($\alpha_l - \alpha_r < 0$) and far points result in a positive (uncrossed) value ($\beta_l - \beta_r > 0$) [Pal99]. Moving point A closer to the eyes or point B further away than the fixation plane increases their absolute disparity. Hence, binocular disparity is directly related to the depth of the object. Another interpretation is given by the correspondence problem. Considering both 2D images projected onto the retinas, disparity is the relative angular shift between the left and right image points. In order to compute this shift, the corresponding points have to be found, which can lead to confusions if there are sim-

[1] e.g. try to thread a needle with only one eye opened

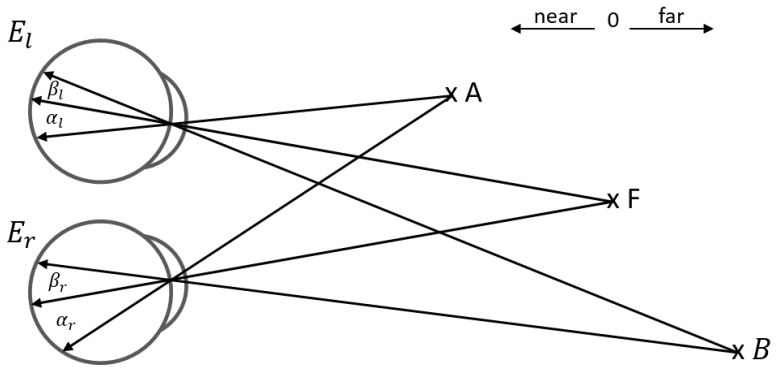

Figure 1.1: Binocular disparity between left and right eye (E_l, E_r). Eyes are fixated on point F, points before (A), and behind (B) show different projection angles relative to the center of each retina.

ilar features in the scene. The brain, however, is remarkably robust in eliminating potential false matches [PC01]. The retinal disparity of the corresponding points then provides an important depth cue to locate the real point in 3D space.
In this report I explain how disparity is represented in the brain and what we can learn from these neurophysiological findings in computer vision. Firstly, a short introduction on the human visual system is given, which summarizes essential concepts. This is followed by chapter 3 focusing on the physiological representation of disparity in the brain. The question how these findings can be mathematically modelled is answered in 4. Finally, algorithms which are based on the model are discussed in 5.

Chapter 2

Human visual system

In order to delve deeper into the question how disparity is represented in the brain, it is essential to have a sound knowledge of the human visual system. Thus, I will summarize most important concepts in the following.
Figure 2.1 illustrates a simplified signal flow of the human visual system. From an

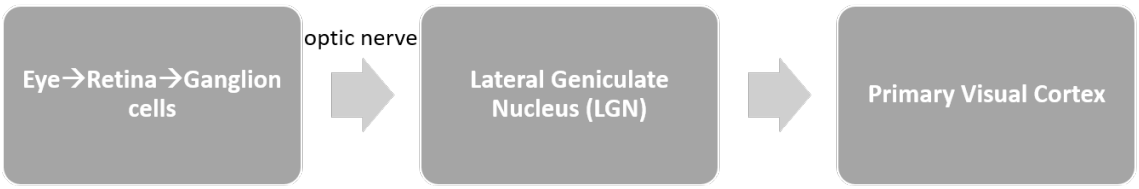

Figure 2.1: Simplified pathway of visual signals from eye to striate cortex.

engineering point of view one can consider the eyes as sensors collecting visual information of the environment and the brain as the processing unit, which interprets the data.
Let us first consider the first block. The eye's adjustable lens and pupil take care that the image is focused on the retina and the right amount of light enters the eye [Hub95]. The retina translates the light into nerve signals with the help of rods and cones. Rods are good for vision in dim light, whereas cones are used for color vision and fine detail [Hub95]. It is important to state that the number of receptors is not uniformly distributed across the retina. The fovea[1] features only cones and is densely packed [Hub95]. Outside the fovea the receptors are more sparsely distributed and rods are more common than cones. This means that the sharpest vision is found within a small region around the center of the retina. The visual signals are processed and passed to retinal ganglion cells. These cells feature center-surround

[1] small area at the center of the retina

receptive fields[2] (RFs). This means that each cell either responds to a white spot on black ground (on-center) or a black spot on white (off-center) as illustrated in figure 2.2a. Covering all receptors within the receptive field with light will give no response

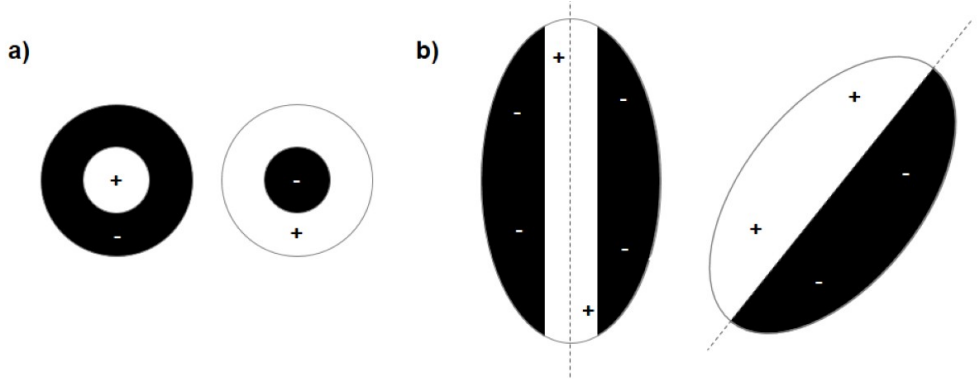

Figure 2.2: Illustration of receptive fields. a) on-center and off-center RFs of ganglion or LGN neurons. b) elongated and oriented RFs of simple cells with distinctive excitatory "+" and inhibitory zones "-".

due to the fact that the inhibitory and excitatory synapses counteract. Thus, at this stage, the neurons are already only sensitive to relative lighting changes within the scene. The output of the ganglion cells is transmitted through the optic nerve and nerve fibers of the left and right eyes are joined at the optic chiasm. The left visual hemifield is transmitted to the right half of the brain and the right visual hemifield to the left one [Kan13].
In each half of the brain, 90 percent of the retinal axons terminate in the lateral geniculate nucleus (LGN) [Kan13]. The LGN neurons also feature center-surround receptive fields and the information of left and right eye is still separated. Furthermore, the projection from ganglion cells to LGN is orderly (retinotopic), which means that nearby neurons in LGN correspond to nearby receptive fields [Kan13].
The output of LGN cells is passed to the primary visual cortex[3]. It consists of six layers of cells, with the inputs of LGN arriving in layer four. Unlike in LGN, the receptive fields of neurons in layers above or below are not center-surround anymore, but respond best to linear, elongated stimuli [Hub95, BR02] as illustrated in figure 2.2b. Moreover, there are many cells which respond to stimuli presented to the left and the right eye [LT99], hence the binocular information is fused. Two major groups of cells are differentiated: simple and complex [Hub95]. Simple cells are tuned to a specific orientation and respond well to bar stimuli. Like center-surround receptive fields, simple cells have excitatory and inhibitory parts in their

[2]A receptive field (RF) of a neuron is defined as the area on the retina which can influence the firing frequency of the cell [Hub95]. This should not be confused with the visual field, which describes the actual scene seen by both eyes [Hub95].

[3]also called V1 or striate cortex, 2mm thick outer layer at the back of the brain [Kan13]

fields and thus the stimulus position and shape is important for the cell's response. One explanation for this behaviour is that simple cells get their input from many LGN neurons with overlapping circular fields. Complex cells, in contrast, do not have specific excitatory or inhibitory zones, and thus are not dependent on the exact position of the stimulus. Nevertheless, complex cells are also tuned to orientation. Besides orientation, simple and complex cells can also be tuned for their preferred spatial frequency, which can be done using sinusoid-gratings [BR02]. A complex cell has a larger receptive field than a simple one and thus it is likely that it receives input from many simple cells with same orientation and frequency but with slightly shifted receptive field positions [Hub95].
Like in LGN, the organisation in V1 is orderly. Columnar structures, perpendicular to the cortical surface, group cells with similar features. Cells within orientation columns prefer the same orientation. Many of such columns next to each other provide sensitivity to all orientation for a certain region in space [Kan13].

Chapter 3

Representation of disparity in the cortex

After covering the neurophysiological basis on the human visual system in 2, the present chapter deals with the question how disparity is represented in the brain. Besides frequency and orientation tuned cells, do neurons exist which are sensitive to disparity as well?
In short, yes, disparities sensitive cells are found in primary visual cortex and in many higher cortical regions [Par07]. Figure 3.1 illustrates the visual areas of the

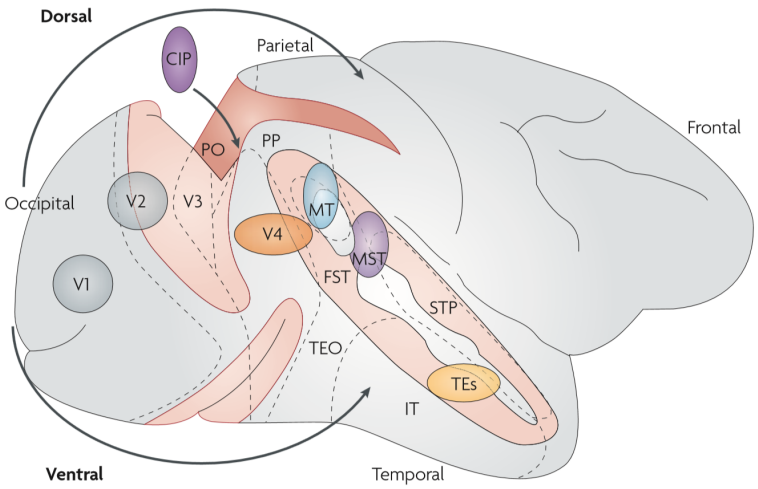

Figure 3.1: Illustration of a human brain with striate cortex V1, extrastriate cortex V2-V5(MT) with dorsal and ventral pathways. (from [Par07])

brain. From early primary visual cortex V1 through V2 the dorsal and ventral pathways lead to the higher areas V3-V5(MT). The first section describes the physiological findings on disparity tuned cells within V1. Insights on the role of extrastriate areas are given in the second part.

3.1 Striate cortex

As explained in chapter 2, most connections of the lateral geniculate nucleus (LGN) terminate at the fourth layer of the primary visual cortex. Moreover, most simple and complex cells are binocular, which means that they respond to stimuli presented to both eyes. The question is, how these binocular cells respond to binocular stimuli which are shifted and thus lead to shifted projections on the retinas. To answer this question, single neurons were recorded from paralyzed animals like cats [FK79] or awake monkeys [LT99], which were trained to fixate their eyes. A stimulus is then presented to each eye and shifted to vary the disparity. The resulting curve is called the disparity tuning curve and if it is not constant, the corresponding cell is said to be disparity selective (see figure 3.2). However, early studies mainly relied on

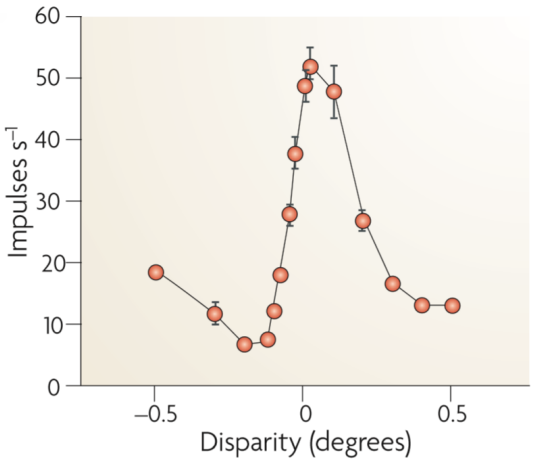

Figure 3.2: Disparity tuning curve of a cell recorded from a monkey using random dot stereograms. (from [Par07])

bar stimuli, which also change the monocular responses while shifting the stimulus [CD01]. As explained in the previous chapter, this is especially true for simple cells, which feature distinct inhibitory or excitatory parts within their RFs. Thus, a change in response does not necessarily point to disparity selectivity. This problem was solved by using random dot stereograms (RDS) as illustrated in figure 3.3. The monocular images are randomly distributed and only the center region is shifted with respect to the surround. In that way the monocular response only changes marginally between stimuli with varying shift. Presented to both eyes separately, the center region is seen in front or behind the background depending on the shift direction.
In general, three different types of tuning curves are distinguished in the literature [FK79, CD01]. Tuned excitatory cells show maximum response to zero or near-zero disparities and a symmetrical profile. This group can be further subdivided in tuned near (peak for crossed disparity), tuned far (peak for uncrossed disparity) and tuned

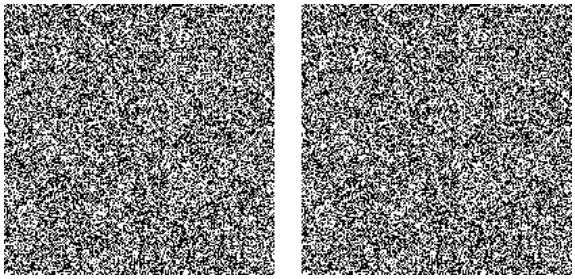

Figure 3.3: 200x200[px] random dot stereogram (RDS). The center 100x100[px] area of the right image is shifted 5[px] to the left. When presented to the left and right eye separately, the central square is seen before the background.

zero (peak for zero disparity). Tuned inhibitory cells show an inverted behaviour with the lowest response near zero disparity. Lastly, asymmetric response profiles have been found for near or far disparities [CD01]. How these tuning characteristics can be modelled is further explained in chapter 4 focusing on the mathematical description.
As mentioned in chapter 2, complex cells do have larger receptive fields than simple cells indicating some sort of pooling mechanisms. At the same time, the cells are tuned to frequency, orientation and disparity, which raises the question how this pooling is done. Spatial pooling could be done combining neighboring cells with equal tuning characteristics, which increases their receptive fields [QZ97]. Additionally, there are indications that something like spatial frequency pooling or a coarse-to-fine disparity detection is performed. A recent study [BSO15] found pooling across different spatial frequencies for complex cells, yielding more broadband RFs. For most cells the preferred disparity does not change across different spatial frequencies, indicating that these pooled cells could play a role in disparity detection. Considering the timing of the response, cells that prefer low frequencies have a short response delay, whereas cells that are sensitive to high spatial frequencies have a larger latency [MF03]. This could mean that the coarse outputs (low frequency) are passed on to the fine cells. As pointed out in chapter 2, many orientation columns provide sensitivity to different oriented stimuli projected onto the retina. Additionally, these cells with varying orientation are also tuned for disparity, which also suggests a pooling across different orientations [CQ04].
Interestingly, complex cells also respond to anti-correlated RDS[1] but with inverted tuning curve [Ohz98, PC01]. Hence, there is a peak where there was no response and a minimum at the former peak. However, humans do not perceive the depth indicated by the inverted disparity peak [Par07]. Therefore, it is unlikely that the disparity tuned complex cells are directly responsible for the perception of depth. The result rather indicates that the cells perform something similar to local correlation within their receptive fields, because an inversion of the stimuli yields an

[1] black pixels in the left image are white in the right one.

inversion of disparity tuning. Moreover, complex cells in V1 are only sensitive to absolute disparity, whereas the human perception of depth is mainly dependent on relative disparity[2] [CD01]. Hence, disparity tuned cells in V1 can be described as local absolute disparity detectors which could contribute to vergence control or perception of depth in further stages within the visual cortex [Par07].

3.2 Extrastriate areas

As described in the previous section, disparity tuned cells in V1 are only tuned to absolute disparities and most likely are not directly responsible for depth perception. Thus, the function of other cortical areas within the visual cortex has to be examined. The most important questions are whether there is a region which is responsible for depth perception outside V1 and if there are cells tuned to relative disparity. Furthermore, it would be interesting to know what the roles of the dorsal and ventral pathways are. The characteristics of each region is analysed with single neuron recordings or with fMRI[3] [GPK+09]. In that way, certain specializations of extrastriate regions have been found.
A cortical region responsible for depth perception should show no or a greatly reduced response to anti-correlated RDS compared to V1. In V2 and in the dorsal stream (V5, MST[4]) the amplitude of the response is comparable to V1. In the ventral stream (V4, TEs[5]), however, a decreased or complete suppressed response to anti-correlated RDS was found. Moreover, the corrective vergence eye movements while changing the disparity of the RDS are also inverted [Par07, CD01].
In the search of areas tuned to relative disparity it is important to distinguish the spatial configuration in depth of the stimulus. For instance some cells may respond to the relative disparity of a slanted plane but not to a center-surround[6] configuration. Sensitive neurons to center-surround stimuli are found in V2 [PC01] and V4, surface separation in depth is found in MST, V5 and relative disparity sensitivity to surface slant is found in V5 [Par07]. In areas TEs and V3 also a sensitivity to surface curvature (second-order disparity) has been identified [GPK+09]. Hence, these neurons could represent the 3D-shape of an object.
To come back to the initial questions, there are neurons tuned to a specific type of relative disparity in dorsal and ventral areas, however all of them are sensitive to absolute disparity as well [Par07]. The dorsal pathway seems to rely on something similar to binocular cross-correlation, due to the response to anti-correlated RDS (like V1). Additionally, only rather simple relative disparity computation (slant of

[2]Relative disparity is the difference between the absolute disparities of two features. In contrast to absolute disparity, the value does not change with vergence eye movements [CD01].
[3]Functional magnetic resonance imaging, a method to measure the neural activity of brain regions.
[4]medial superior temporal area
[5]areas in the anterior inferior temporal cortex [Par07]
[6]center of RDS has different absolute disparity than surround

surface, depth segregation) is performed. Thus, vergence eye movements could be controlled by signals from the dorsal areas. In the ventral areas, in contrast, more sophisticated computation are performed. There is a reduced or no response to anti-correlated RDS and sensitivity to 3D-shape of objects. Hence, both pathways perform different kind of computations and may both contribute to the perception of depth [Par07]. Interestingly, regions in V5 have been identified, which either prefer near or far disparities. The preferred disparities change only along the surface of the area but not perpendicular to it, suggesting a columnar organisation [CD01]. During a near-far discrimination task, microstimulation of the near regions in a monkey's brain induced it to decide more often for near depth and vice versa [PC01]. However, more research is needed to fully understand how the extrastriate areas contribute to the perception of depth.

Chapter 4

Mathematical disparity model

This chapter deals with the mathematical description of the neurophysiological findings described in the previous chapters. How can simple and complex cells be modelled and how can the preferred disparity be obtained? Moreover, the biological plausibility of the model is discussed in section 4.2. Finally, in 4.3, the model is compared to standard cross-correlation.

4.1 Derivation

Let us first consider the modelling of a simple cell. As explained before, binocular simple cells are sensitive to RFs in both left and right eyes. Moreover, they have well defined inhibitory and excitatory areas within their elongated RF and are tuned to a preferred frequency. These features can be well described by Gabor functions[1] [FO90]. In order to simplify the derivation, only one dimension of a vertically oriented RF is considered for now.

$$f_l = \exp\left(-\frac{x^2}{2\sigma^2}\right)\cos(\omega_0 x + \Phi_l) \tag{4.1}$$

$$f_r = \exp\left(-\frac{(x-s)^2}{2\sigma^2}\right)\cos(\omega_0 (x-s) + \Phi_r) \tag{4.2}$$

with spatial position x and Gaussian standard deviation σ, spatial frequency ω_0, phase offsets Φ_l, Φ_r and spatial shift s. The size of the RF is determined by the σ of the Gaussian envelope and the preferred frequency by ω_0. Positive values correspond to excitatory and negative to the inhibitory parts of the RF. An example of a Gabor function is illustrated in figure 4.1a. A difference in the profiles, which is essential to relate any disparity, can be achieved in two ways. Either the phase is shifted between the left and right sinusoid with identical Gaussian envelope ($s = 0, \Phi_l \neq \Phi_r$), or the phase is equal ($\Phi_l = \Phi_r = \Phi$) with shifted Gaussian envelope ($s \neq 0$). The position-shift approach assumes two shifted RF profiles with equal shape (see figure 4.1b).

[1] Gaussian function multiplied with sinusoid

The phase-shift model, in contrast, leads to a relative displacement of the sinusoid and a difference in shape (see figure 4.1c). The response of a binocular simple cell

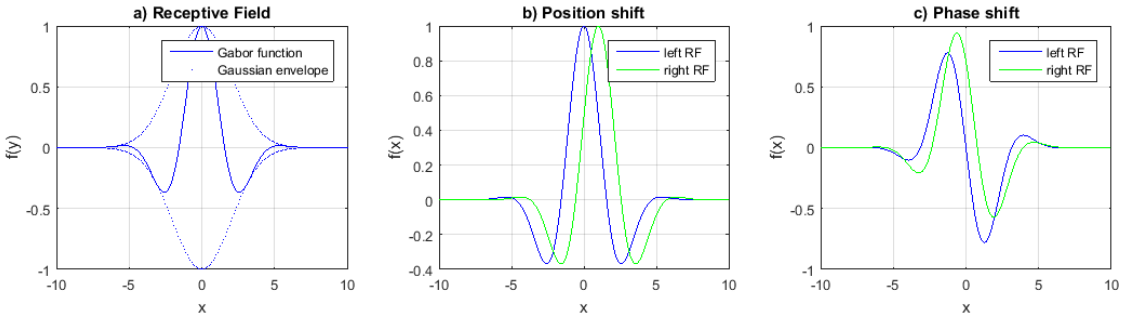

Figure 4.1: a) Gabor function with Gaussian envelope; b) Position-shift model: $s = 1, \Phi_{l,r} = 0$; c) Phase-shift model: $s = 0, \Phi_l = \frac{\pi}{2}, \Phi_r = \frac{\pi}{4}$; Values: $\sigma = 2, w_0 = 1$

can then be modelled as the linear combination of the filtered left and right images [FO90].

$$r_s = \int_{-\infty}^{\infty} (f_l(x)I_l(x) + f_r(x)I_r(x))\, dx \tag{4.3}$$

with the previously defined RF profiles $f_l(x), f_r(x)$ and the retinal images $I_l(x), I_r(x)$. However, for different RDS, the resulting disparity tuning curves $r_s(s), r_s(\Delta\Phi)$ do not have the same peak location for neither the position nor the phase-shift model [ZQ96]. This is due to the fact, that the response is strongly dependent on the Fourier phase of the stimulus, which is also the case for real simple cells (see chapter 2). Thus, one simple cell alone cannot be enough for disparity detection [ZQ96]. Therefore, a complex cell is modelled as the sum of two squared simple cell responses, which are related by a phase shift of $\frac{\pi}{2}$ (quadrature pair).

$$r_q = r_{s1}^2 + r_{s2}^2 \tag{4.4}$$

with the simple cell responses r_{s1}, r_{s2}, which are related by $\Phi_{l2} = \Phi_{l1} + \frac{\pi}{2}, \Phi_{r2} = \Phi_{r1} + \frac{\pi}{2}$. In that way, the response is not dependent on the monocular Fourier phase anymore, because either the first or the phase shifted second simple cell contribute to the response. Only a binocular phase difference influences the cell's response. This behaviour is also observable for real complex cells, which do not depend on the spatial position of the stimulus. Due to the fact that a complex cell has a larger receptive field than simple cells, a spatial pooling between neighboring quadrature pairs is introduced [ZQ96].

$$r_c = r_q * w \tag{4.5}$$

with weighting function w and convolution operator $*$. The described response model of a complex cell is also called energy model within the literature [Ohz98, MG10, CQ04].

Based on a general stimulus image shifted horizontally $I_l(x), I_r(x + d)$, it can be

shown [CQ04, ZQ96] that the preferred disparity of a complex cell can be well approximated by the RF parameters of the simple cells.

$$d_{pos} \approx s \tag{4.6}$$

$$d_{phase} \approx \frac{\Delta\Phi}{\omega_0} \in [-\frac{\pi}{\omega_0}, \frac{\pi}{\omega_0}] \tag{4.7}$$

Hence, in case of the position-shift model, the preferred disparity of the complex cell is equal to the shift between the RFs of its simple cells. For the phase-shift model, the disparity depends on the phase difference $\Delta\Phi$ and the preferred frequency ω_0. Furthermore, the range of disparities is restricted to $[\frac{-\pi}{\omega_0}, \frac{\pi}{\omega_0}]$, due to the periodicity of the cosine function. Hence, larger disparities can be covered with a smaller preferred frequency ω_0.
Instead of separating the position and phase shift approach, also a hybrid model can be stated [ZQ96]. In this case, the shift s of the phase-shift model is not zero anymore and thus the preferred disparity is found as the linear combination of both models [ZQ96].

$$d_{hybrid} \approx s + \frac{\Delta\Phi}{\omega_0} \in [s - \frac{\pi}{\omega_0}, s + \frac{\pi}{\omega_0}] \tag{4.8}$$

Hence, the position and phase-shift model can also be interpreted as specializations of the hybrid model for $\Delta\Phi = 0$ or $s = 0$ respectively.
As described in chapter 2, the RF of a real simple cell is not one dimensional but two dimensional with a preferred orientation. However, the above derivations can be extended to the 2D case. The left and right vertically oriented RFs are now modelled as 2D Gabor functions (see Eq. 4.1).

$$f_l = \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right)\cos(\omega_0 x + \Phi_l) \tag{4.9}$$

$$f_r = \exp\left(-\frac{(x-s)^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right)\cos(\omega_0 (x-s) + \Phi_r) \tag{4.10}$$

with horizontal shift s and horizontal phase offset $\Delta\Phi = \Phi_l - \Phi_r$. An illustration of a 2D Gabor function is given in figure 4.2 with black corresponding to -1 and white to 1. The preferred orientation can be set by rotating the above vertical RFs by an angle $(\Theta - 90°)$. A horizontal disparity d in a stimulus can then be split into a component parallel $d_{para} = \cos(\Theta)d$ and perpendicular to the orientation $d_{perp} = \sin(\Theta)d$. As it can be seen in figure 4.2, the RF profile parallel to its orientation changes very slowly, thus this component can be neglected [CQ04]. The preferred disparities (Eq. 4.6) can now be rewritten as [CQ04]

$$d_{pos} \approx \frac{s}{\sin(\Theta)} \tag{4.11}$$

$$d_{phase} \approx \frac{\Delta\Phi}{\sin(\Theta)\omega_0} \in [-\frac{\pi}{\omega_0}, \frac{\pi}{\omega_0}] \tag{4.12}$$

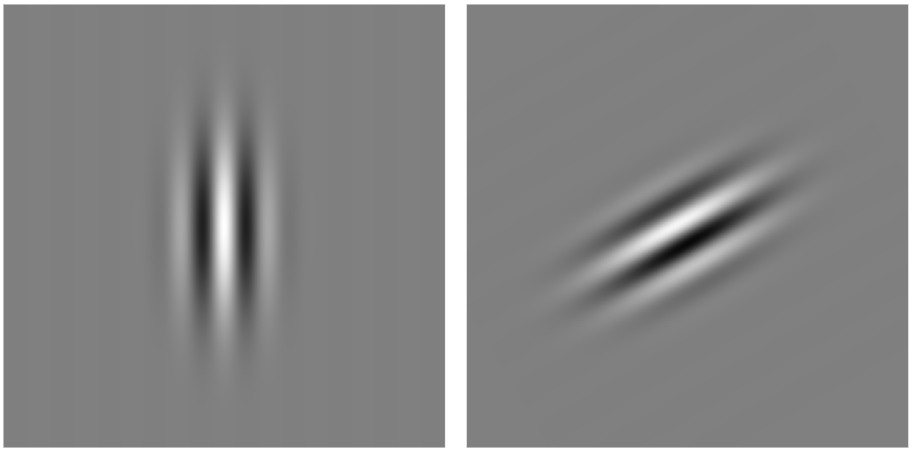

Figure 4.2: Illustration of a vertically oriented (left) and rotated (right) 2D Gabor function. Black values correspond to -1 and white values to 1. Values: $w_0 = 1, \sigma_x = 4, \sigma_y = 8$, left: $\Phi = 0, \Theta = 90$, right: $\Phi = \frac{\pi}{2}, \Theta = 30$.

with orientation Θ defined relative to the positive x-axis ($\Theta = 90°$ yields original vertical RFs).

4.2 Biological plausibility

Both, position and phase shift representations of RFs are biologically plausible mechanisms found in real complex cells. Thus, it is likely that both approaches contribute to disparity coding [CD01].
The question is, how well this model predicts a real disparity tuning curve of a complex cell as described in chapter 3? Studies have shown, that the model is capable of yielding similar tuning curves with respect to random dot stimuli [ZQ96, Ohz98]. Furthermore, the three different types of tuning curves can be described by position- and phase-shifts. Tuned excitatory cells correspond to complex cells with no phase-shift, which leads to symmetric responses. Furthermore, inhibitory cells can be described by a phase shift of π, yielding a zero response at the preferred disparity. Finally, asymmetric cells can be reproduced with a phase-shift near $\frac{\pm\pi}{2}$. With this reasoning, it seems more intuitive to describe the different response profiles as a continuum [CD01]. On top of that, the model also predicts the inversion of disparity tuning for anti-correlated RDS [Ohz98].
However, it is also important to state the limitations of the model. The model predicts, that the inverted disparity tuning curve for anti-correlated RDS should have the same amplitude as in the correlated case with a phase shift of 180°. Real complex cells, in contrast, show a much weaker response (smaller amplitude) for anti-correlated RDS [Ohz98]. This could be fixed by introducing a non-linearity in the monocular stages before combining them to a simple cell. However, the extended

model would lose its simplicity, as more computational steps are required [PC01]. Furthermore, the model only describes how absolute disparity is detected in V1 but does not predict the behaviour of higher cortical areas. Further research is needed to expand the model to extrastriate areas and relative disparity. One possible approach would be to use the response of two V1 complex cells to create relative disparity sensitivity found in higher cortical areas [Par07].

4.3 Relation to cross-correlation

In the previous chapter 3, the neurons in V1 are described as local absolute disparity detectors, performing something similar to cross-correlation. Does this also become apparent in the model?
The simple cell response can be written as

$$r_s = L + R \tag{4.13}$$

with the filtered left and right images L,R. The complex cell is built from two simple cells in quadrature and hence

$$r_q = (L_1 + R_1)^2 + (L_2 + R_2)^2 \tag{4.14}$$

The squaring yields two cross-terms which determine the cells disparity tuning behaviour [QZ97].

$$r_q \approx \text{const} + 2L_1R_1 + 2L_2R_2 \tag{4.15}$$

Hence, the cross-terms in the model relate the filtered left and right images by multiplication, similar to cross-correlation. However, compared to the standard cross-correlation between two images

$$r(d) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I_l(x,y)I_r(x-d,y)dxdy \tag{4.16}$$

the images are filtered with the receptive fields before. Also, there are two cross-terms in the cortical approach compared to one in the standard formulation. Furthermore, in equation 4.16 the whole image patch is integrated, while in the cortical model L and R are scalars [QZ97].

Chapter 5

Cortically inspired algorithms

The disparity model, introduced in chapter 4, is a good approximation for the response of a disparity tuned complex cell. The question is, how this could be used in computer vision?
In order to compute disparity maps of binocular stimuli, not a single disparity tuning curve of a complex cell is of interest but the response of a population of cells [CQ04]. At each spatial location (e.g. pixel) there are many complex cells with different properties (frequency (scale), orientation, phase offset, position offset) leading to various preferred disparities. Given binocular stimuli, the response of each population of cells is used to estimate the disparity. The question is, how the complex cells at each location should be chosen (model, orientation, frequency) and how the disparity can be estimated accurately from the responses.
Hence, cortically inspired algorithms to compute disparity maps using the disparity model are explained in section 5.1. Their biological plausibility is assessed in 5.2.

5.1 Algorithms

A simple algorithm to compute disparity maps of RDS was proposed by Quian et al. [QZ97] (see figure 5.2A). One-dimensional vertically oriented Gabor functions and 8 complex cells in each location were used. The 8 cells vary in phase difference or in position offset for each model respectively. Before the disparity estimation, the spatial pooling described in equation 4.5 is applied. In order to account for the sparse population of tuned complex cells, parabolic interpolation is used for estimation [QZ97]. The peak response then yields the disparity estimate. For this specific approach, the phase and position approach yield similar results. However, the RFs of real complex cells are two dimensional and not always vertically oriented. Additionally, natural scenes have different characteristics than synthetic RDS.
To tackle these limitations Chen et al. suggested another approach [CQ04] (see figure 5.2B). In contrast of using only the phase or position-shift model, both are combined in a coarse-to-fine approach. They found that the population response using the phase shift-model is more reliable and accurate for small stimulus disparity than

the position-shift model. However, the accuracy decreases with increasing disparity, hence an iterative algorithm is proposed. At first, the disparity is estimated using the phase-shift model and no position shift ($s = 0$), yielding an estimate d_0. This value is used as a position shift ($s = d_0$) to decrease the disparity of the stimulus and an estimate d_1 is computed again using the now shifted population of complex cells. This is repeated iteratively with $s = d_i$ and $d_{est} = d_0 + ... + d_i$. On top of that, 2D Gabor functions with different orientations and scales are considered in each location. To do so, there is a population of complex cells with different orientation and phase difference, whereas the shift s and the scale are fixed in each iteration. After each update, the shift is set to the current estimate and the scale is decreased (w_0 gets larger and Gaussian envelope σ smaller, $w_0\sigma = \pi$). Furthermore, spatial and orientation pooling is applied before estimation to gain more robust results. The algorithm is applied to both, synthetic RDS and natural images. Figure 5.1 illustrates how the iterative algorithm refines the resulting disparity map in each step.

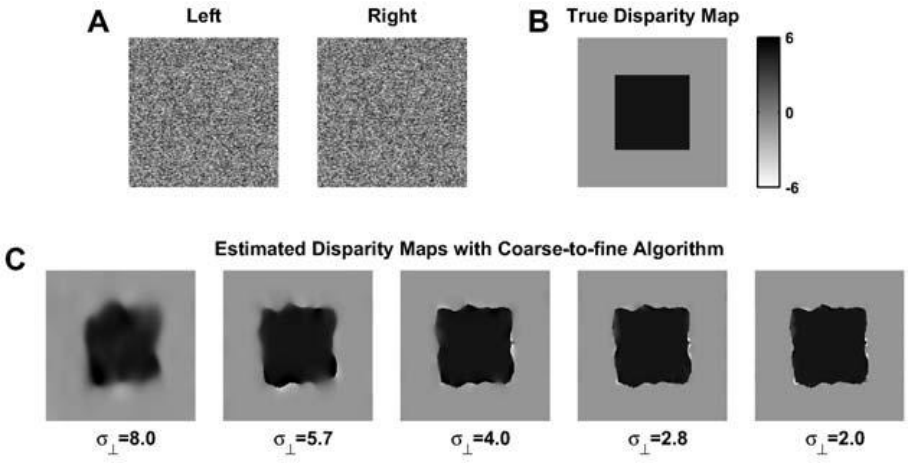

Figure 5.1: A) RDS used; B) true disparity map C) iterative result using coarse-to-fine approach (rearranged from [CQ04])

In [TS09] a Bayesian approach for estimating the disparity is introduced (see figure 5.2C). Instead of using the peak (P) or average (M) of a population response as an estimate, they propose a normalized response $R = \frac{P-M}{M}$, which is approximately linear dependent on the log Bayes factor[1]. In that way they use the value of R as a scaled measure of the evidence. Moreover, compared to the coarse to fine method, they only use one preferred frequency (scale), five different orientations and the hybrid energy model. Hence, in each location there are complex cells with varying position- and phase-shift, orientation but equal standard deviation and frequency. Like in the other algorithms, spatial pooling is applied between complex cells with

[1]The factor indicates whether the hypotheses that the population disparity is close to the true value or not is valid.

equal position, phase-shift and orientation. The normalized response R is then computed for each phase-shifted population at a fixed position-shift $s \in \{0, 1, ..., 127\}$, and fixed orientation. To accumulate the evidence across the orientations, the normalized responses are pooled. The position-shifted population with maximum normalized response R is chosen as disparity estimate. As R is normalized, it can also be used to detect occluded areas, which yield small responses below a threshold. This approach is extended in [MG10], which also includes pooling across two different scales of normalized responses R after orientation pooling. For an overview of all presented algorithms see figure 5.2.

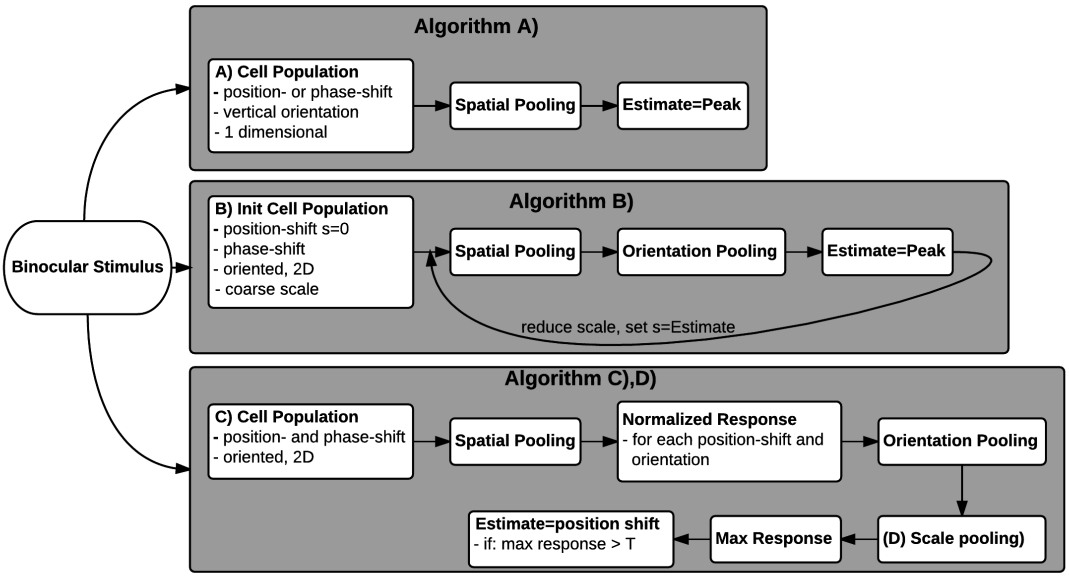

Figure 5.2: A) [QZ97] Peak of spatially pooled response of 8 complex cells with position or phase-shift yields estimate; B) [CQ04] Coarse-to-fine, iterative computation with reduced scale; C,D) [TS09],[MG10] Hybrid population, estimate found choosing the maximum normalized response, scale pooling only in algorithm D).

In order to compare the performance of these cortical algorithms it is interesting to discuss their performance on the well-known Middlebury College data sets. Figure 5.3 shows the resulting disparity maps of the "Cones" and "Teddy" test images as well as the ground truth. Disparities are horizontal and in the range of $\{0, 1, ..., 127\}$. The percentage of false disparity values within the non-occluded areas is lowest for the normalized response algorithm with scale pooling [MG10] (5.3D), followed by the similar approach without pooling [TS09] (5.3C). The coarse to fine approach yields a higher error rate [TS09], which likely is due to the rather large range of disparities within the test-sets (5.3B).

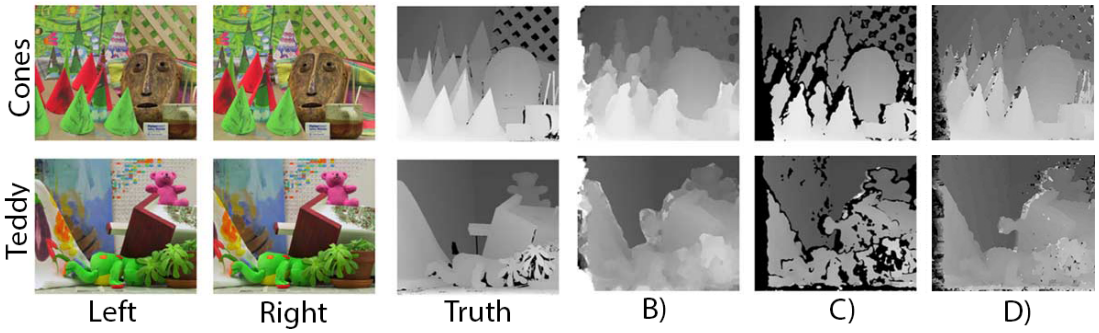

Figure 5.3: B) Coarse-to-fine algorithm [CQ04]; C) Normalized response algorithm [TS09]; D) with scale pooling [MG10] (rearranged from [TS09], [MG10])

5.2 Biological plausibility

Although it is not yet known how the brain decides for a specific disparity value given a population response, all described algorithms are biologically plausible. Pooling of cells with nearby spatial locations, different orientation and frequency are possible computations performed by real cells. Additionally, there is also evidence for the coarse to fine approach as described in chapter 3. However, it is clear that the disparity tuned cells in V1 are not the final processing stage for depth perception and hence the cortical algorithms also have to be evaluated this way. They provide a proof of concept for the local detection of absolute disparity in a biologically plausible manner. A more detailed understanding of the further cortical processing stages is needed to derive improved algorithms more closely related to the actual perception of depth.

Chapter 6

Conclusion

In this report, the question how the 3D-world is constructed in the brain is answered with focus on binocular disparity. There are neurons tuned to binocular disparity in V1 and most other areas of the visual cortex. Cells in the striate cortex can be described as absolute local disparity detectors, whereas neurons in extrastriate areas are also sensitive to relative disparity. The disparity model has shown to be a good approximation for disparity tuned complex cells in V1, which is used by cortically inspired algorithms for disparity estimation. However, it is still not clear how the binocular disparity information is transformed into actual depth perception.
The human eyes, converge on a point of fixation, which defines zero disparity. Thus, without knowing the distance to the fixation plane, the perceived depth of an object would change with viewing distance. The vergence signal of the eyes could provide crucial information for depth perception and constancy [MWTR00]. However, as pointed out before, there is evidence that corrective vergence movements are driven by absolute disparity computations before depth perception. Anti-correlated RDS do not result in depth perception but yield inverted vergence movements [MBM97]. Moreover, a recent study showed that depth can even be recovered under diplopia[1] [LWAH14], which also argues against a strong involvement in depth perception. Thus, vergence may only be seen as mechanism to fuse binocular images and to limit the range of possible disparities.
Further research is needed to fully understand how the brain provides a 3D-world from two 2D images. The disparity sensitive cells in V1, which respond to anti-correlated RDS, cannot fully describe our perception of depth. Hence, the role of extrastriate areas has to be better understood. A detailed knowledge of the different cortical processing stages beyond V1 may then be useful to derive enhanced models and cortical algorithms for artificial vision systems.

[1] left and right images are not fused, double vision

List of Figures

Bibliography

[BR02] Christine E Bredfeldt and DL Ringach. Dynamics of spatial frequency tuning in macaque v1. *The Journal of Neuroscience*, 22(5):1976–1984, 2002.

[BSO15] Mika Baba, Kota S Sasaki, and Izumi Ohzawa. Integration of multiple spatial frequency channels in disparity-sensitive neurons in the primary visual cortex. *The Journal of Neuroscience*, 35(27):10025–10038, 2015.

[CD01] B. G. Cumming and G. C. Deangelis. The Physiology of Stereopsis. *Annual Review of Neuroscience*, 24:203–238, 2001.

[CS02] Jorg Conradt, Pascal Simon, Michael Pescatore and Paul FMJ Verschure. Saliency map operating on stereo images detect landmarks and their distance, ICANN 2002, 795-800.

[FK79] B Fischer and J Krüger. Disparity tuning and binocularity of single neurons in cat visual cortex. *Experimental brain research*, 35(1):18, March 1979.

[FO90] Ralph D Freeman and Izumi Ohzawa. On the neurophysiological organization of binocular vision. *Vision research*, 30(11):1661–1676, 1990.

[GPK+09] Svetlana Georgieva, Ronald Peeters, Hauke Kolster, James T Todd, and Guy A Orban. The processing of three-dimensional shape from disparity in the human brain. *The Journal of Neuroscience*, 29(3):727–742, 2009.

[Hub95] D.H. Hubel. *Eye, Brain, and Vision*. Scientific American Library Series. Henry Holt and Company, 1995.

[Kan13] E. Kandel. *Principles of Neural Science, Fifth Edition*. Principles of Neural Science. McGraw-Hill Education, 2013.

[LT99] M. S. Livingstone and D. Y. Tsao. Receptive fields of disparity-selective neurons in macaque striate cortex. *Nature Neuroscience*, 2(9):825–832, 1999.

[LWAH14] Arthur J Lugtigheid, Laurie M Wilcox, Robert S Allison, and Ian P Howard. Vergence eye movements are not essential for stereoscopic depth. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1776):20132118, 2014.

[MBM97] GS Masson, C Busettini, and FA Miles. Vergence eye movements in response to binocular disparity without depth perception. *Nature*, 389(6648):283–286, 1997.

[MF03] M. D. Menz and R. D. Freeman. Stereoscopic depth processing in the visual cortex: a coarse-to-fine mechanism. *Nature Neuroscience*, 6:59–65, 2003.

[MG10] Flavio Mutti and Giuseppina Gini. Bio-inspired disparity estimation system from energy neurons. In *Proc. IEEE Int. Conf. on Appl. Bionics and Biomechanics*, pages 1–6, 2010.

[MWTR00] Mark Mon-Williams, James R Tresilian, and Andrew Roberts. Vergence provides veridical depth perception from horizontal retinal image disparities. *Experimental brain research*, 133(3):407–413, 2000.

[Ohz98] Izumi Ohzawa. Mechanisms of stereoscopic vision: the disparity energy model. *Current opinion in neurobiology*, 8(4):509–515, 1998.

[Pal99] S.E. Palmer. *Vision Science: Photons to Phenomenology*. A Bradford book. Bradford Bokk, 1999.

[Par07] Andrew J. Parker. Binocular depth perception and the cerebral cortex. *Nature reviews. Neuroscience*, 8(5):379–391, May 2007.

[PC01] AJ Parker and BG Cumming. Cortical mechanisms of binocular stereoscopic vision. *Progress in brain research*, 134:205–216, 2001.

[QZ97] Ning Qian and Yudong Zhu. Physiological computation of binocular disparity. *Vision research*, 37(13):1811–1827, 1997.

[TS09] Eric KC Tsang and Bertram E Shi. Disparity estimation by pooling evidence from energy neurons. *Neural Networks, IEEE Transactions on*, 20(11):1772–1782, 2009.

[ZQ96] Yu-Dong Zhu and Ning Qian. Binocular receptive field models, disparity tuning, and characteristic disparity. *Neural Computation*, 8(8):1611–1641, 1996.

License