

Vergence Control and Disparity Estimation with Energy Neurons: Theory and Implementation

Wolfgang Stürzl, Ulrich Hoffmann, and Hanspeter A. Mallot

Universität Tübingen, Zoologisches Institut, Kognitive Neurowissenschaften,
72076 Tübingen, Germany
wolfgang.stuerzl@uni-tuebingen.de
<http://www.uni-tuebingen.de/cog/>

Abstract. The responses of disparity-tuned neurons computed according to the energy model are used for reliable vergence control of a stereo camera head and for disparity estimation. Adjustment of symmetric vergence is driven by minimization of global image disparity resulting in greatly reduced residual disparities. To estimate disparities, cell activities of four frequency channels are pooled and normalized. In contrast to previous active stereo systems based on Gabor filters, our approach uses the responses of simulated neurons which model complex cells in the vertebrate visual cortex.

1 Introduction

In the literature a variety of stereo algorithms has been proposed for estimating depth from disparities, i.e. local image shifts caused by the different positions of the two eyes or cameras, see for example [6], [9]. Findings from psychophysical studies of human stereo vision and vergence adjustment were used in active stereo camera systems. Based on physiological recordings from binocular neurons in the visual cortex of cats and monkeys, so-called energy models for stereo processing have been suggested. In this paper we will describe how the responses of disparity-tuned energy neurons can be used for vergence control of a stereo camera head and for disparity estimation on images taken by two monochrome video cameras.

2 Binocular Energy Model

We give a short overview on energy neurons modeling disparity-tuned cells found in visual cortex of mammals (e.g. [3], [7]).

Linear Stage: In this paper, we consider only vertically oriented receptive fields since they are best suited to deal with horizontal disparities. The receptive field of a quadrature pair of monocular linear neurons is modeled (e.g. [1], [8]) as

$$f_{\nu}(x, y, \varphi) = \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \left(\cos(2\pi\nu x + \varphi) + i \sin(2\pi\nu x + \varphi)\right) \quad . \quad (1)$$

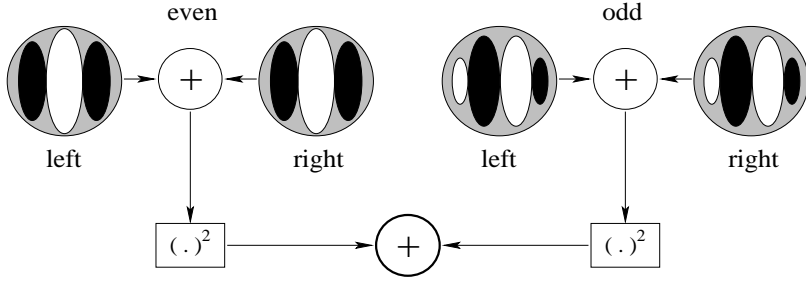


Fig. 1. Illustration of (3): Complex cells combine the output of linear neurons modeled as convolution of left and right images with even/odd (cosine/sine) Gabor kernels including a quadratic nonlinearity.

The output of the linear neurons is formulated mathematically as a convolution of image I with the receptive field function f_ν ,

$$Q_\nu(x, y, \varphi) = |Q_\nu(x, y)|e^{i(\phi(x, y) + \varphi)} = \int f_\nu(x - \xi, y - \eta, \varphi)I(\xi, \eta) d\xi d\eta \quad (2)$$

Complex Cells: Binocular complex cells combine the output of the linear filters applied to both images as shown in Fig. 1,

$$C_\nu = |Q_{l\nu} + Q_{r\nu}|^2 = (\text{Re}[Q_{l\nu}] + \text{Re}[Q_{r\nu}])^2 + (\text{Im}[Q_{l\nu}] + \text{Im}[Q_{r\nu}])^2 \quad (3)$$

Two different models have been proposed for complex cells tuned to disparity D , see for example [1]:

- Phase-shift: corresponding left and right receptive fields have different phases, $C_{D\nu}(x, y, \varphi) = |Q_{l\nu}(x, y, \varphi + 2\pi\nu\frac{D}{2}) + Q_{r\nu}(x, y, \varphi - 2\pi\nu\frac{D}{2})|^2$.
- Position-shift: corresponding left and right receptive fields are centered at shifted positions, $C_{D\nu}(x, y, \varphi) = |Q_{l\nu}(x + \frac{D}{2}, y, \varphi) + Q_{r\nu}(x - \frac{D}{2}, y, \varphi)|^2$.

As discussed e.g. in [8], complex cells of the phase-shift type have limited range of preferred disparities, $D \in [-\frac{\pi}{\nu}, \frac{\pi}{\nu}]$. Consequently, only neurons with low central frequencies (small ν) can code large disparities. Since there is evidence that in humans high frequencies also contribute to perception of large disparities without utilizing a coarse-to-fine strategy [4], we use energy neurons of the pure position-shift type. Phase φ is set to zero.

3 Implementation

We use four frequency channels ($k = 1, 2, 3, 4$) with bandwidth of two octaves, i.e. $\frac{\sigma\nu}{\nu} = \frac{3}{5}$, and center frequencies $\nu_k \in \{5\nu_0, 10\nu_0, 20\nu_0, 40\nu_0\}$, where $\nu_0 = \frac{1}{L_x}$ is the minimal frequency determined by image width L_x (see Fig. 2 b). The Gabor functions were sampled and shifted to yield zero DC component. The resulting convolution kernels for $5\nu_0$ are shown in Fig. 2 a.

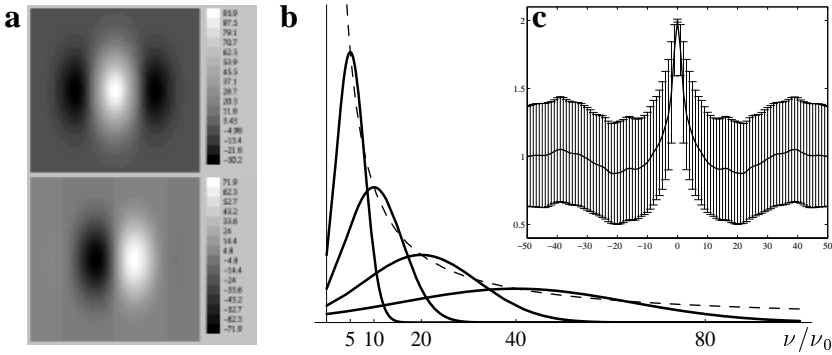


Fig. 2. **a** Gabor filters with bandwidth of two octaves: convolution kernel (size 49×49 pixels) for the lowest frequency $\nu_1 = 5\nu_0$. **b** In the frequency domain, complex Gabor functions are Gaussians of width $\sigma_\nu = \sigma_x^{-1}$ centered at $\nu/\nu_0 = 5, 10, 20, 40$. The inset **(c)** shows a tuning curve of a normalized complex cell (5) tuned to zero disparity (for each stimulus disparity 100,000 responses were evaluated, error bars represent standard deviation).

Monocular normalization: To reduce the influence of “inter-ocular” contrast differences the filter outputs are normalized according to

$$\hat{Q}_k(x_i, y_i) = \frac{Q_k(x_i, y_i)}{(L_x L_y)^{-1} \sum_j |Q_k(x_j, y_j)|} \quad (4)$$

(L_x and L_y are image width and height respectively).

Combination of frequency channels and binocular normalization: Complex cells of the different frequency channels, but tuned to same disparity are combined and normalized to reduce influence of local image contrast, resulting in a normalized complex cell,

$$\hat{C}_D(x_i, y_i) = \frac{\sum_k C_{Dk}(x_i, y_i)}{\epsilon + \sum_k |\hat{Q}_{lk}(x_i + \frac{D}{2}, y_i)|^2 + |\hat{Q}_{rk}(x_i - \frac{D}{2}, y_i)|^2} \quad (5)$$

ϵ is a constant avoiding high complex cell activity in case of very low contrast which causes $\sum_k |\hat{Q}_{lk}|^2 + |\hat{Q}_{rk}|^2 \approx 0$ (in the current implementation ϵ is set to 4.0). We can rewrite (5) using (3) and dropping index k ,

$$\begin{aligned} \hat{C}_D &= \frac{\sum_k |\hat{Q}_{lk} + \hat{Q}_{rk}|^2}{\epsilon + \sum_k |\hat{Q}_{lk}|^2 + |\hat{Q}_{rk}|^2} = \frac{\sum 2|\hat{Q}_l|^2 + 2|\hat{Q}_r|^2 - |\hat{Q}_l - \hat{Q}_r|^2}{\epsilon + \sum |\hat{Q}_l|^2 + |\hat{Q}_r|^2} \\ &= \frac{2 \sum |\hat{Q}_l|^2 + |\hat{Q}_r|^2}{\epsilon + \sum |\hat{Q}_l|^2 + |\hat{Q}_r|^2} \left(1 - \frac{\sum |\hat{Q}_l - \hat{Q}_r|^2}{2 \sum |\hat{Q}_l|^2 + |\hat{Q}_r|^2} \right) \end{aligned} \quad (6)$$

$$\approx 2 \left(1 - \frac{\sum |\hat{Q}_l - \hat{Q}_r|^2}{2 \sum |\hat{Q}_l|^2 + |\hat{Q}_r|^2} \right), \quad \text{if } \epsilon \ll \sum |\hat{Q}_l|^2 + |\hat{Q}_r|^2 \quad (7)$$

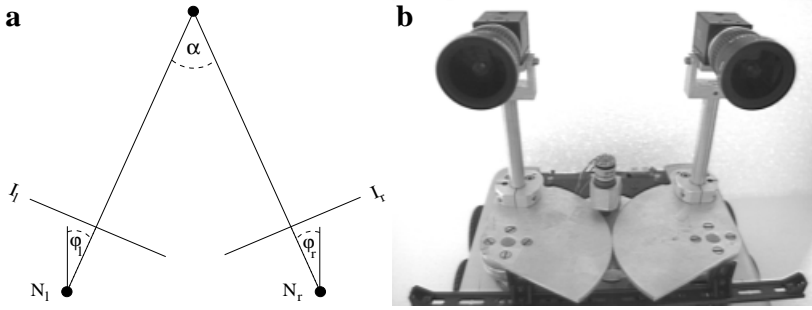


Fig. 3. **a** Camera vergence angle α is defined as the angle between the two camera axes ($N_{l/r}$: left/right camera nodal point, $I_{l/r}$: left/right image). **b** Stereo camera head with symmetrical vergence adjustment ($\varphi_l = \varphi_r = \frac{1}{2}\alpha$) ensured by two cogwheels driven by a single stepper motor. Each of the two monochrome cameras has a field of view of approx. 80° . Nodal point separation is 14.5 cm.

From (6) and (7) we see that \hat{C}_D reaches its maximum value if \hat{Q}_l equals \hat{Q}_r , i.e. if the left and right receptive field “look” at corresponding image parts. Using the triangle inequality it can be shown that $\hat{C}_D \in [0, 2]$:

Verging Camera Head: Taking the mean of \hat{C}_D for each disparity over all positions, we compute responses of vergence controller cells C_D^V . From their activity global image disparity is estimated according to

$$D_{\text{glob}}^{\text{est}} = \arg \max_D C_D^V \quad , \quad (8)$$

$$C_D^V = (L_x L_y)^{-1} \sum_i \hat{C}_D(x_i, y_i) \quad . \quad (9)$$

We use a large range of (global) disparities with inhomogeneous resolution (highest at zero disparity), i.e. $D \in \{0, \pm 1, \pm 2, \pm 3, \pm 5, \pm 7, \pm 10, \pm 14, \pm 19, \pm 25, \pm 32, \pm 40, \pm 49, \pm 59 \text{ pixels}\}$.

As long as $D_{\text{glob}}^{\text{est}} \neq 0$, camera vergence is changed symmetrically ($\varphi_l = \varphi_r = \frac{1}{2}\alpha$) using a stepper motor which drives two cogwheels (see Fig. 3). The vergence angle α is approximately proportional to $D_{\text{glob}}^{\text{est}}$ (for our stereo system we use $\alpha = 0.5^\circ D_{\text{glob}}^{\text{est}}$).

Disparity Estimation: After adjustment of vergence angle, the range of residual space dependent disparities is usually greatly reduced. Residual disparities are analysed with neurons tuned to disparities $D \in \{0, \pm 1, \pm 2, \dots, \pm 10\}$. The population activity of these neurons is a rich code of stereo information that will be useful for many tasks. If disparity maps are sought, they can be obtained as the preferred disparity of the locally most active neuron,

$$D_{\text{loc}}^{\text{est}}(x_i, y_i) = \arg \max_D \hat{C}_D(x_i, y_i) \quad . \quad (10)$$

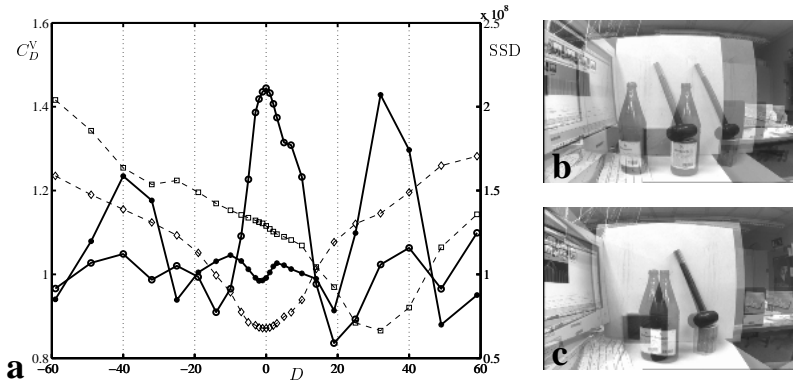


Fig. 4. **a** Responses of vergence controller neurons (small dots and circles) are compared to SSD (rectangles and diamonds, dashed curves) of shifted left and right images (shifts were applied according to the preferred disparity of the neurons, note different scaling on y-axes). Initially when camera axes are aligned corresponding to vergence angle 0° (see upper superimposed image, **b**) peak response is at $D = 32$ (small dots). After vergence adjustment global disparity is minimized corresponding to maximum activity of C_0^V (circles). At the resulting vergence angle of approx. 16° the residual disparities are much smaller and better fitted to the range of the local disparity detectors (**c**).

We also compute a confidence value for each pixel using

$$c(x_i, y_i) = \frac{\max_D \hat{C}_D(x_i, y_i)}{N_D^{-1} \sum_D \hat{C}_D(x_i, y_i) + \epsilon_D} \quad , \quad (11)$$

where ϵ_D was set to 1/4 of maximum cell response, i.e. 0.5 and $N_D = 21$ is the number of preferred neuron disparities used. Locations where true disparity is out of the considered range or where disparity estimation is simply impossible, e.g. due to occlusions, will receive low confidence.

4 Results

In Fig. 4 the response of the vergence controller neurons is compared to the sum of squared differences (SSD) between left and right image. At the optimal vergence angle (approx. 16° in this example), i.e. highest response of C_0^V , the SSD has its minimum at zero image shift corresponding to maximal image correlation. This is in agreement with studies on vergence adjustment in humans, e.g. [5].

In order to check the estimated disparity map calculated after vergence adjustment according to (10) we compute a “cyclopean” view [2], i.e. we fuse left and right image using the disparity map,

$$I_{\text{fused}}(x, y) = \frac{1}{2} \left(I_l \left(x + \frac{1}{2} D_{\text{loc}}^{\text{est}}(x, y), y \right) + I_r \left(x - \frac{1}{2} D_{\text{loc}}^{\text{est}}(x, y), y \right) \right) \quad . \quad (12)$$

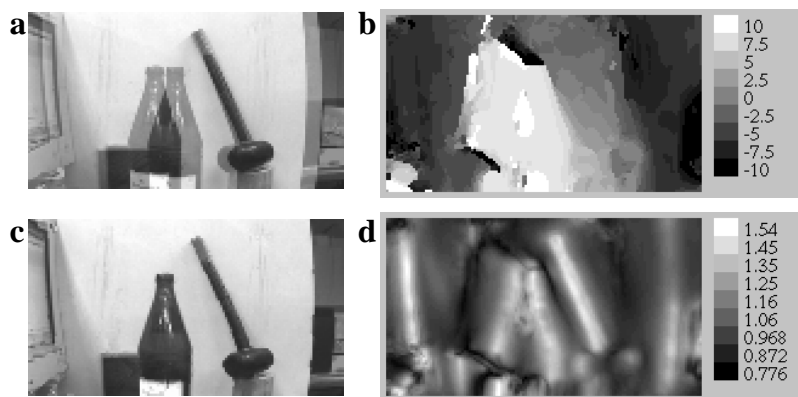


Fig. 5. Disparity estimation on stereo image: Superimposed image after vergence adjustment (a). The left and right images can be fused using the disparity map (b) and (12) eliminating all double vision (c). The confidence map (d) calculated according (11) shows high values at contours of high contrast.

By comparing the resulting fused image (Fig. 5c) with the superposition of left and right image (Fig. 5b) one can see that double vision has almost completely vanished.

Further evaluation of the proposed stereo algorithm will be done on a mobile robot using the activities of the disparity-tuned cells as representation of places.

References

1. Fleet, D., Heeger, D., Wagner, H.: Neural encoding of binocular disparity: Energy model, position shifts and phase shifts. *Vision Research* **36** (1996) 1839–1857
2. Henkel, R.D.: Fast stereovision by coherence detection. In G. Sommer, K.D., Pauli, J., eds.: *Computer Analysis of Images and Patterns, LCNS 1296* (1997) 297–30
3. Hubel, D., Wiesel, T.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* **160** (1962) 106–154
4. Mallot, H., Gillner, S., Arndt, P.: Is correspondence search in human stereo vision a coarse-to-fine process? *Biological Cybernetics* **74** (1996) 95–106
5. Mallot, H., Roll, A., Arndt, P.: Disparity-evoked vergence is driven by inter-ocular correlation. *Vision Research* **36** (1996) 2925–2937
6. Marr, D., Poggio, T.: A cooperative computation of stereo disparity. *Science* **199** (1976) 283–287
7. Ohzawa, I., DeAngelis, G., Freeman, R.: Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science* **249** (1990) 1037–1041
8. Qian, N.: Computing stereo disparity and motion with known binocular cell properties. *Neural Computation* **6** (1994) 390–404
9. Sanger, T.: Stereo disparity computation using gabor filters. *Biological Cybernetics* **59** (1988) 405–418