Neural Mechanisms for Form and Motion Detection and Integration: Biology Meets Machine Vision

Heiko Neumann[1] and Florian Raudies[2]

[1] Institute for Neural Information Processing, Ulm Univ., Germany
[2] Center for Computational Neuroscience and Neural Technology, Boston Univ., USA

Abstract. General-purpose vision systems, either biological or technical, rely on the robust processing of visual data from the sensor array. Such systems need to adapt their processing capabilities to varying conditions, have to deal with noise, and also need to learn task-relevant representations. Here, we describe models of early and mid-level vision. These models are motivated by the layered and hierarchical processing of form and motion information in primate cortex. Core cortical processing principles are: (i) bottom-up processing to build representations of increasing feature specificity and spatial scale, (ii) selective amplification of bottom-up signals by feedback that utilizes spatial, temporal, or task-related context information, and (iii) automatic gain control via center-surround competitive interaction and activity normalization. We use these principles as a framework to design and develop bio-inspired models for form and motion processing. Our models replicate experimental findings and, furthermore, provide a functional explanation for psychophysical and physiological data. In addition, our models successfully process natural images or videos. We show mechanism that group items into boundary representations or estimate visual motions from opaque or transparent surfaces. Our framework suggests a basis for designing bio-inspired models that solve typical computer vision problems and enable the development of neural technology for vision.

1 Introduction - Vision Processes in Man and Machine

The visual system of primates is characterized by its flexibility and robustness to process data under various imaging conditions, such as noise, illumination, and partial occlusions. Also, it adapts its functionality by using learning mechanisms. Form and motion information is mainly processed along two segregated pathways, namely the ventral stream for form representation and the dorsal stream for motion representation. Machine vision mainly focused on developing given task-related constraints. Their aim is the optimization of engineering-like defined objective functions, which are often hard to transfer to other problem domains. Only a few approaches investigate the problem how higher-level knowledge can be used to stabilize and disambiguate sensory signals.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part I, LNCS 7583, pp. 468–473, 2012.

We sketch a core model architecture that has been motivated by knowledge from neuroscience. Several model instances have been derived from this core model to explain a wealth of experimental data from psychophysics and physiology and process natural images and videos besides psychophysics stimuli. Next, we describe our model architecture which is followed by the results and discussion and conclusion.

2 Neural Modeling of Dynamic Vision in Cortex

Biological information processing can be described at various levels of detail. For instance, it can be described at the level of individual neurons, their interconnections and biophysical dynamics, or at the level of neuronal layers and cortical areas and their connectivity as a graph-like structure. To capture the general aspects of visual processing, we employ simple model neurons that are organized into columns at each spatial location. To model the dynamics of such neurons, we use single compartment model neurons with gradual activation dynamics formally denoting the neuronal firing rate. All columns together define an area with a specific functionality, like in visual cortex. We model the interactions in and between such columns. Neurons are laterally interconnected, signaling to their spatially neighboring columns as in cortex. The modeling on a macroscopic level is motivated by the distributed and hierarchical organization of cortical areas and their interconnections [4]. One striking principle is that areas in visual cortex are mostly bi-directionally connected: an area not only sends feed forward (FF) signals to an area higher up in the hierarchy but also receives feedback (FB) signals from higher areas. The role of FB is a topic of active research. The key principle is that FB connections are mainly modulatory. They cannot generate activations alone without driving input.

Our model architecture employs a simplified version of layered processing that is mapped to a three-stage processing cascade consisting of: (i) an initial filtering for e.g. orientation or motion direction, that generates a representation of the driving visual stimulus, (ii) a stage of topographically organized re-entrant signals from areas higher in the hierarchy that amplify initial filtering signals, and (iii) an output stage that normalizes signals to keep the overall energy within bounds. The membrane conductance of model neurons is

$$\dot{v}_i^{(1)}(t) = -v_i^{(1)}(t) + (E_{ex} - v_i^{(1)}(t)) \cdot \{s * F^+\}_i - (E_{in} + v_i^{(1)}(t)) \cdot \{s * F^-\}_i \quad (1)$$

$$\dot{v}_i^{(2)}(t) = -v_i^{(2)}(t) + (E_{ex} - v_i^{(2)}(t)) \cdot [v_i^{(1)}(t))]_+ \cdot (1 + net_i^{FB}) \quad (2)$$

$$\tau \dot{v}_i^{(3)}(t) = -E_L v_i^{(3)}(t) + (E_{ex} - v_i^{(3)}(t)) \cdot v_i^{(2)}(t) - \alpha v_i^{(3)}(t) \cdot f_p(w_i^p(t)) \quad (3)$$

$$\tau_p \dot{w}_i^p(t) = -w_i^p(t) + (E_{ex}^p - w_i^p(t)) \cdot f_s(v^{(3)}(t)) * \Lambda^p. \quad (4)$$

In this set of equations, s denotes the bottom-up driving input signals, $F^{+,-}$ denotes filter kernels, $*$ denotes the convolution in space and/or feature domain, and $[\bullet]_+$ denotes a half-wave rectification. The constants E, τ, and α define levels of activity saturation, membrane efficacy, and modulation strength, respectively.

The normalization integrates the activity in a pool of neurons employing an integration kernel Λ^p. This pool activity, in turn, inhibits the output activation (Eq. 3) via shunting surround inhibition. Output activities are passed through a non-linear transfer function of sigmoidal shape $f(\bullet)$ to account for the firing rate function of neurons. The re-entrant modulating signal net^{FB} amplifies the respective driving FF activation by FB, but leaves FF unchanged when FB is absent. On the other hand, if no FF signal is present, FB cannot generate a response. This asymmetry between FF and FB signals is captured in the model by incorporating a tonic level of FB at each location in the space-feature domain.

3 Simulation Results for Form and Motion Processing

Form Processing and Grouping. The goal of form and shape processing is to detect and interpolate arrangements of local oriented contrasts, to generate representations of (surface) boundaries, to infer 3D ordinary layout of surfaces from junctions of different types, or to infer the geometric properties of 3D object surfaces irrespective of their material properties. Based on early modeling investigations of boundary grouping [5], we developed a taxonomy of mechanisms for feature cooperation, which employs a measure of feature compatibility (for details, see [11]). Based on this taxonomy, we suggest that the outputs from collinear sub-field filtering are combined conjunctively in order to generate contour completions as observed for illusory boundary formation. Oriented contour integration operates at the spatial scale of area V1 and V2 (see [16]). Processing mechanisms of this model V1-V2 for contour integration are formulated as a variant of the generic Equations 1-4. Here, we only highlight the most important aspects. The input stage for long-range grouping at model V2 is defined by the instance of Eqn. 1

$$\dot{v}^{V2,1}_{i,\theta}(t) = -v^{V2,1}_{i,\theta}(t) + (E_{ex} - v^{V2,1}_{i,\theta}(t)) \cdot \{s^{V1}_{\theta} * \Gamma^{V2,L}\}_i \cdot \{s^{V1}_{\theta} * \Gamma^{V2,R}\}_i. \quad (5)$$

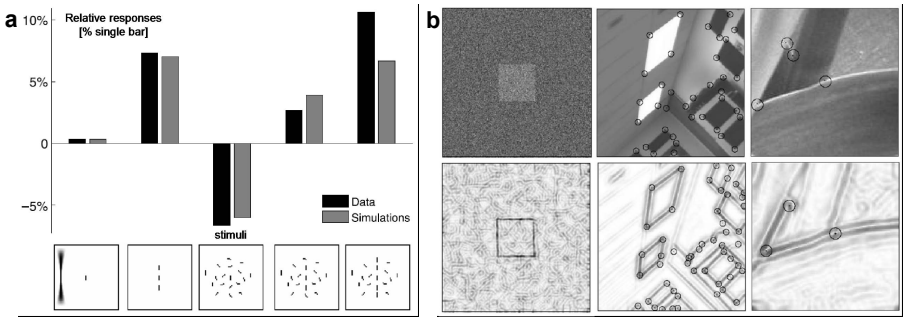

Fig. 1. Simulation results for form processing. a) Oriented contour representations are modulated by local context (black bars: physiological data from [9], grey bars: model simulations from [7]). b) Contour enhancement and junctions detected from local activity distributions for natural images (from [6]).

The input filtering uses pairs of elongated sub-field kernels $\Gamma^{V2,L/R}$ that integrate oriented bottom-up inputs. The output at stage three of model V2 is fed back to enhance oriented filter activities in model V1, formally, defined by the variant of Eqn. 2

$$\dot{v}_{i,\theta}^{V1,2}(t) = -v_{i,\theta}^{V1,2}(t) + (E_{ex} - v_{i,\theta}^{V1,2}(t)) \cdot v_{i\theta}(t) \cdot (1 + \Psi^{V2}(\{v_{\theta}^{V2,3} * \Lambda^{FB}\}_i)). \quad (6)$$

Model cells show the same behavior as recorded cells when probed by oriented input contrast items with surround context (see Fig. 1 a). The grouping mechanism successfully processes real-world stimuli as well. For example, the network is capable of enhancing junction configurations (Fig. 1 b) so that different configurations of them can be reliably read out from V1 model cells. Simulations outperform 2D landmark detection using the structure tensor (see [6] for details).

Motion Detection and Integration. Our second model instance and related simulations focuss on the detection and integration of visual motion. Our goal is the estimation of spatio-temporal responses and their integration to resolve motion ambiguities, e.g. caused by the aperture problem or noisy input. Our model consists of areas V1, MT, and MST, all part of the dorsal pathway. Initial motions are detected by direction selective cells in model area V1 which feed their responses to motion integrative cells in model area MT. Model area MT output signals subsequently feed into area MST that contains cells selective for motion patterns that are generated on the image plane, e.g. by self-motion. Thus, model MST cells are selective for translational, rotational, or expansion/contraction optic flow or a combination thereof. A major challenge is allowing for motion perception of opaque as well as transparent motion. A key mechanism for the processing of transparent motion is a soft competition between model MT cells formulated as center-surround interaction in the velocity space. This soft competition is a simplified variant of Eqn. 3 and 4

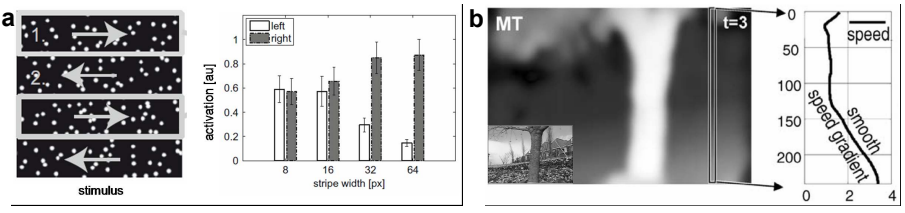

Fig. 2. Simulation results for motion analysis. a) Patterns arranged in lanes of opposite motions are integrated by motion-sensitive cells. Large widths mainly drive cells with matching velocity selectivity (grey bars) and for small widths, cells selective to the opposite motion direction become increasingly active (white bars; FB from MST stabilize these activities; from [14]). b) Motion detection and integration for a computer vision benchmark sequence (inset). Depth variations are represented by smooth motion gradients (from [2]).

$$\dot{v}^{MT,3}_{i,vel}(t) = -v^{MT,3}_{i,vel}(t) + \{v^{MT,2} * \Lambda^{+}\}_{i,vel} - \alpha v^{MT,3}_{i,vel}(t) \cdot w_{i,vel}(t) \quad (7)$$

$$\dot{w}_{i,vel}(t) = -w_{i,vel}(t) + \{v^{MT,2} * \Lambda^{-}\}_{i,vel} \quad (8)$$

with $vel = (\phi, ||\mathbf{u}||)$. Similar motions that fall into the excitatory zone are combined, while motions are suppressed and repelled when they fall in the inhibitory zone [15]. Motion (semi-) transparency occurs when spatially separate motions with significant velocity difference occur which cannot be resolved by the visual sensing mechanisms. This case generates multiple activations at the same image location (Fig. 2 a). Our model implicitly detects different transparent motion layers at MT level, but it does not require a pre-defined number of motions. This is unlike computational vision approaches which detect multiple layers of motion with an iterative routine treating motion from other layers as outliers (e.g., [3]).

Real-world video input has been successfully processed as well. The flower garden sequence shows a tree moving in front of a background due to depth variations relative to a translating observer. The result demonstrates that the model successfully generates discontinuous motion at occlusions based on texture-defined motion discontinuities. In addition, it illustrates the ability to represent speed gradients (see the marked column in the image where the speed gradient is induced by the slanted ground plane, Fig. 2 b). Such gradients are prerequisite of inferring spatial layout from motion or help to solve the structure-from-motion problem ([12]).

4 Discussion and Conclusion

We presented the outline of a biologically inspired architecture for visual processing. The components have been motivated by functional principles of cortical architecture resulting in a cascade of three processing steps for bottom-up driving FF signal flow and modulating re-entrant FB that delivers context information from higher areas. Together with the activity normalization this leads to a biased competition of activity distributions. Our models explain various experimental data from psychophysics and physiology. They also robustly process real-world data at a performance level that is comparable to computer vision algorithms. Our framework provides an approach to the design of general purpose vision systems utilizing core principles of cortical processing. The building blocks of our framework are flexible and modular. Additional functionality can be easily added without causing interface problems for input/output representations and their processing. For example, we demonstrated the interaction between motion and form processing for figure-ground segregation where the figures are only defined by kinetic contours [13]. Unsupervised learning mechanisms have been incorporated into the framework to build feature detectors and representations for higher level processing in animated motion sequence analysis for action recognition [10].

In summary, our simulations from various domains of vision demonstrate the generality of our approach to develop bio-inspired vision models. Together with

newly available hardware that accelerates the execution of distributed processing at various stages, our approach allows for building robust and adaptive vision technologies capable to function in various, unconstrained applications.

Acknowledgements. HN was supported by the Transregional Collaborative Research Centre "A Companion Technology for Cognitive Technical Systems" funded by DFG. FR acknowledges support from the Office of Naval Research (ONR N00014-11-1-0535 and ONR MURI N00014-10-1-0936).

References

1. Bayerl, P., Neumann, H.: Disambiguating visual motion through contextual feedback modulation. Neural Computation 16, 2041–2066 (2004)
2. Bayerl, P., Neumann, H.: Disambiguating visual motion by form-motion interactional computational model. Int'l J. of Comp. Vis. 72(1), 27–45 (2007)
3. Black, M.J., Anandan, P.: The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. Comp. Visi. and Image Understanding 63(1), 75–104 (1996)
4. Felleman, D.J., van Essen, D.C.: Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex 1, 1–47 (1991)
5. Grossberg, S., Mingolla, E.: Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentation. Percept. & Psychophys. 38, 141–171 (1985)
6. Hansen, T., Neumann, H.: Neural mechanisms for the robust representation of junctions. Neural Computation 16, 1013–1037 (2004)
7. Hansen, T., Neumann, H.: A recurrent model of contour integration in primary visual cortex. J. of Vision 8(8), 1–25 (2008)
8. Hirsch, J.A., Gilbert, C.D.: Synaptic physiology of horizontal connections in the cats visual cortex. J. of Neuroscience 11, 1800–1809 (1991)
9. Kapadia, K.M., Ito, M., Gilbert, C.D., Westheimer, G.: Improvement in visual sensitivity by changes in local context: Parallel studies in human observers and in V1 of alert monkeys. Neuron. 15, 843–856 (1995)
10. Layher, G., Giese, M.A., Neumann, H.: Learning Representations for Animated Motion Sequence and Implied Motion Recognition. In: Villa, A.E.P., Duch, W., Érdi, P., Masulli, F., Palm, G. (eds.) ICANN 2012, Part I. LNCS, vol. 7552, pp. 288–295. Springer, Heidelberg (2012)
11. Neumann, H., Yazdanbakhsh, A., Mingolla, E.: Seeing surfaces: The brain's vision of the world. Physics of Life Reviews 4, 189–222 (2007)
12. Orban, G.: Higher order visual processing in macaque extrastriate cortex. Physiol. Rev. 88, 69–89 (2008)
13. Raudies, F., Neumann, H.: A neural model of the temporal dynamics of figure-ground segregation in motion perception. Neural Networks 23, 160–176 (2010)
14. Raudies, F., Neumann, H.: A model of neural mechanisms in monocular transparent motion perception. J. of Physiology 104, 71–83 (2010)
15. Raudies, F., Mingolla, E., Neumann, H.: A model of motion transparency processing with local center-surround interactions and feedback. Neural Computation 23, 2868–2914 (2011)
16. Weidenbacher, U., Neumann, H.: Extraction of surface-related features in a recurrent model of V1-V2 interactions. PLoS One 4(6), e5909 (2009)