Outline of a Computational Theory of Human Vision

Fridolin Wild

Department of Information Systems and New Media,
Vienna University of Economics and Business Administration (WUW),
Augasse 2-6, A-1090 Vienna, Austria,
fridolin.wild@wu-wien.ac.at

Abstract. Human vision is a powerful yet highly efficient processing system. Drawing on an extensive review of empirical findings and theoretical foundations of human visual-spatial perception originating from various disciplines, results of an in-depth analysis of the human visual perception process will be presented.
Several stages in the visual pathway will be identified that segment into processes in eye and retina, control processes for exploration and sampling, processes in the primal visual cortex, feature detection, and, furthermore, object recognition processes. Based on the review, a computational theory of human vision will be drafted with emphasis on developing an architectural model.
The model introduced will couple enhanced perceptrons over an intermediated layer, which is responsible for controlling exploration and sampling, to primal feature decomposition modules. A feature based router then is responsible for distributing this preprocessed input to concept detector components. The architecture will allow both bottom-up and top-down data flows. Moreover, it will facilitate 'lazy' processing by introducing means for focused, concept driven attention.

1 Introduction

Human visual-spatial perception is an active process of the brain starting with primal processing already on the eye's retina. Researchers from various disciplines (including among others psychology, cognitive science and neurology) up to now have fairly well succeeded in segmenting this process into several different stages. Vision itself, however, remains an eternal conundrum.

Visual-spatial perception (or 'vision') can be defined as the process[1] of building an internal representation of an object, a scene, an event, and simply any concept (or compilation thereof) in the mind of the beholder. This encompasses entities or relations that are beliefed to exist in an external reality and that can be derived by processing reflected light rays (or an absence thereof).

Besides visual charateristica of the objects, this process is shaped by the human and individual anatomy (cf. Pfeiffer's 'embodiment' theses, [23]), prior

[1] Including the result of this process, cf. [27].

experiences, current task and context, expectations, aims, and self-regulatory strategies (for the latter cf. [7, 32]). To rephrase this in a nutshell: already existing activities[2] and the easiness of traversion of neuron connecting axons[3] both drive the spreading of activations in the human brains neural net. The vision process is, so to say, fundamentally 're-constructive' in nature (cf. [25]).

In the remaining sections of this paper, the author will first analyse the stages of the human vision process in more detail to, second, draft the outlines of a model emulating the human approach of perception.

2 Human Vision

Starting with a Section on primal processing in the eye and the retina, a Section on exploration and sampling through the motor movements of the eye will follow. Moreover, an overview over the workflows to and in the primal visual cortex will be given. These findings will result in a Section on (atomic) feature detection which then will be investigated under the premises of (human) object recognition theories.

2.1 Eye and Retina

The human eye is sensitive for lightwaves in the range between 400nm (violet) and 700nm (red). The eye works in a manner similar to a camera. The cornea bends the light beams through the pupil, the opening in the iris, to the retina at the back of the eye. The iris thereby acts just like the apperture of a camera, contracting when exposed to bright light (thus letting less light in) and expanding when experiencing little light. The lense is responsible for focusing light to the retina, thereby reflecting the picture upside down.

The retina consists of two types of photoreceptors: the cones, which respond to colours (either red-, green- or blue-sensitive) and are mostly situated in the centre of the retina, the so called fovea (cf. [19]). The second type of receptors are the rods, which respond to brightness and can only be found outside the fovea. The eye performs detailled distinctions only in the fovea, which covers approximately the size of a thumb nail in an arm-length distance of the field of sight. Outside the fovea, acuity decreases tremendously. However, stimuli in these peripheral areas are processed and sensitivity to peripheral stimuli can even be enhanced by training (cf. [12]).

The photoreceptors are connected through bipolar cells to ganglion cells which communicate the sensory excitation to the brain (see figure 2.1, cf. [18], [17],[21]). Usually several rods and cones together with horizontal cells converge to one bipolar cell, which again converge with other bipolars into one ganglion. Figure 2.1 shows (simplified) how on-bipolar cells and off-bipolar cells inhibit respectively strengthen activations of surrounding cells (cf. [24]).

Horizontal cells and amacrine cells transmit signals laterally. Depending on the input receptors, the signals are merged and converted into different (colour)

[2] Which nodes are already active and are emitting/relaying action potentials?
[3] Which trajectory is inhibatory, which is facilitating?

contrast signals. Amacrinal cells similarily show an antagonistic behavior. In some cases, however, they react only to stimuli changes or show phasic behavior.

All receptors of a retinal ganglion cell form a receptive field of this cell. The larger this field is, the fuzzier the perceived picture will be – as the origins of impulses cannot be exactly identified. In the fovea these receptive fields are very small and, accordingly, the resolution is very high.

Ganglions bundle input from the above mentioned layers into receptive fields, either into an on-centre, into an off-centre or into an on-off-centre field type. Moreover, they can be distinguised into transient (sensitive only to changes) and sustained (i.e. sensitive constantly throughout the stimulation) cells. Taking into account these different behaviors, ganglion cells can be functionally distinguished — for example according to their sensitivity to colour antagonisms, luminance, movement, directions, specific spatial frequencies, and others (cf. [18]).

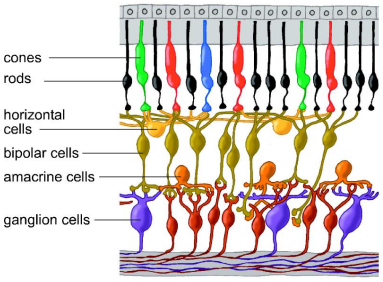

Fig. 1. Cross-Section of the Retina (Kolb)

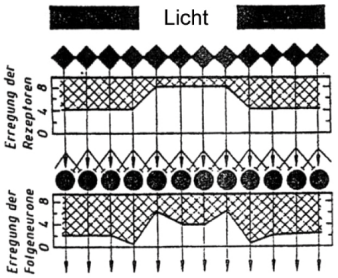

Fig. 2. Inhibition and Excitation on the Retina (Pichler).

2.2 Exploration and Sampling

The sampling process with which the eye explores the field of sight consists of fixations and saccades. Saccades are the rapid eye movements (approx. 25ms duration) with which the high-resolution central field is pointed to the area of desire. Fixations (and slow eye movements) of relatively long duration (300ms) follow these extremly quick redirections.

Movement and focus are controlled by 'motor maps' (residing in the Colliculi Superiores), a kind of neural sketch-pad activated from the retina itself and from regions of higher processing and cognition. Retina activation in these motor maps for example aligns the eye's orientations towards hard contrast changes. Activation from higher processing stages is top-down responsible for e.g. integrating expectations or tasks.

2.3 The Primal Visual Cortex

The primal visual cortex (V1) gets input from the Chiasma Opticus, the crossing of the nerve cells attached to the ganglions in the middle between retina and primal visual cortex at the back of the brain.

The input is organised in two times three neuron layers responsible for red-green antagonism, yellow-blue antagonism and luminance-antagonism, all for each eye separate. Only about 10-20% of the stimuli from the retinal ganglion cells reach the cortex.

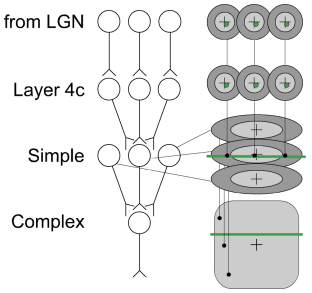

Fig. 3. Celltype connectivity.

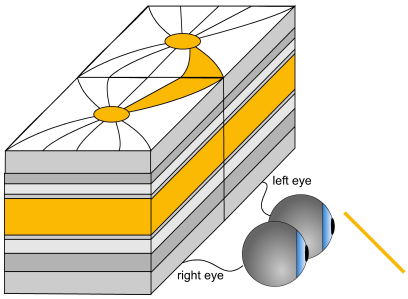

Fig. 4. Hypercolumns.

The cortex combines both parts again, both hemispheres are connected by the Corpus Callosum (see figure, cf. [14]). The primal visual cortex (V1) is partitioned into six layers. The fourth layer receives input of the concentric receptive fields originating from the retina. From there, several cells converge onto simple cells. Simple cells are thus sensitive to bar-shaped (bars, lines, edges) light stimuli of a specific orientation. Simple cells converge onto complex cortex cells with receptive fields similar to those of simple cells (see figure 3). However, they cover a larger field of the retina (thus generating positional invariance) and they fire most to moving lines (cf. [31]). Furthermore, other more complex types of cells have been identified in the cortex, for example hypercomplex cells that respond to lines of specific length or to combinations of orientations. Layer 4c input cell can be wired with many different simple cells.

The cortex is organised in hypercolumns: position columns are organised retinotrop, i.e. their spatial distribution resembles the distribution in the retina. Ocular dominance columns have a pinwheel structure (for the orientation sensitivity, the so called orientation columns) and – at their centres – a colour responsive blob (see figure 4).

2.4 Feature Detection

Starting with the processing in the hypercolumns, channels can be assumed which split visual input data along their spatial frequencies into nine different channels (see figure 5, cf. [11], [15],[16]). This acts just like the biological realisation of a coarse fourier analysis. Figure 6 shows on the left the original picture which is separated as described into its frequency channels. A revisualisation of the first four low-frequency channels to the right shows that the human vision system perceives the dotted line on some of the channels as if they were connected[4]. Not all features are processed simultaneously, e.g. usually colour is processed quicker than shape and shape again is quicker than movement ([34]).

[4] Wertheimer called this in his investigations in Gestalt laws the law of proximity.

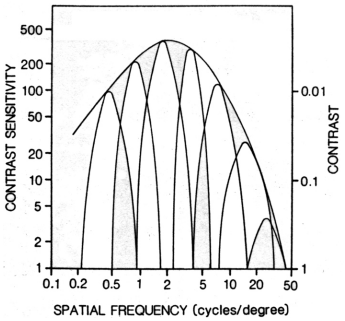

Fig. 5. Frequency Channels (Ginsburg).

Fig. 6. Original and revisualisation of four low frequency channels (Ginsburg).

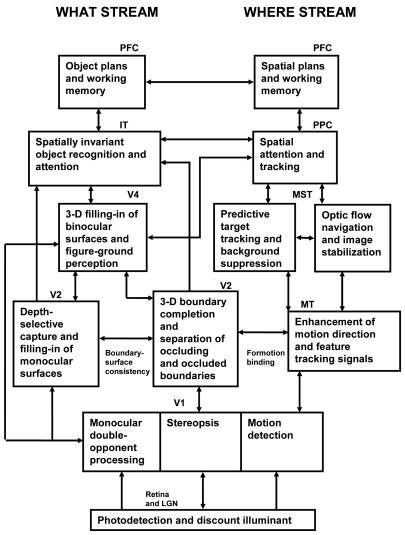

Fig. 7. 'What'- and 'where'-stream (Grossberg).

From the hypercolumns in the primal visual cortex, activations are spread to the other areas of the brain (including the other areas of the visual cortex).

Further down, a 'what'- and a 'where'-stream can be differentiated (see figure 7, cf. [13]). The 'what'-stream is responsible for object recognition, the 'where'-stream localises where these objects and events are. The 'what'-stream is assumed to be allocentric, i.e. object centered, with basically retinotrop organisation or at least resemblance, whereas the 'where'-stream is organised egocentric, i.e. observer centered.

2.5 Object Recognition

Object recognition theories have a long tradition in research on the human vision process (cf. [20], [4],[29],[28], [33],[14],[10], [13]). Object recognition theories, according to Ullman ([30]), must be able to cope with four variability effects: photometric effects, context effects[5], effects of changing perspectives, and effects of shape morphs[6]. By combining the above mentioned stream processing method with Guided Search (an enhanced Feature Integration Theory, cf. [29], [28] and [33]), these obstacles can be overcome. The coupling processes base on feature bundles and, thereby, guide attention bottom-up. The other way round, expectations lead to higher pre-activation of expected features (in the expected locations). Retinotrope feature maps split features according to their types, which are then processed by the functional modules (as described in [13]). From there stimuli associate to the brain areas where memories reside. The short term memory

[5] For example, whenever an object is partly hidden by another object.
[6] For example, a sitting vs. a standing person.

is responsible for keeping relevant areas active (for more information on memory see [1], [2], [3], [22], [5], [6], [8]).

3 A Computational Model of Human Vision

3.1 Enhanced Perceptrons: e-Ganglions

The first layer of an artificial vision system emulating human vision consists of perceptron-like (cf. [26]) electronic ganglions (e-ganglions) that split the concentric field of sight of a retina-equivalent into smaller, overlapping fields of sight of ganglion-equivalents. All major types of circular receptive fields of the human ganglions[7] in varying sizes have to be modeled. The various types of behavior can be imitated with pattern matching algorithms. As a side-effect of this, some photometric effects can be already eliminated in this layer. In figure 8, the overlapping receptive fields are represented as petals of the blossom unfolding around the fixation points.

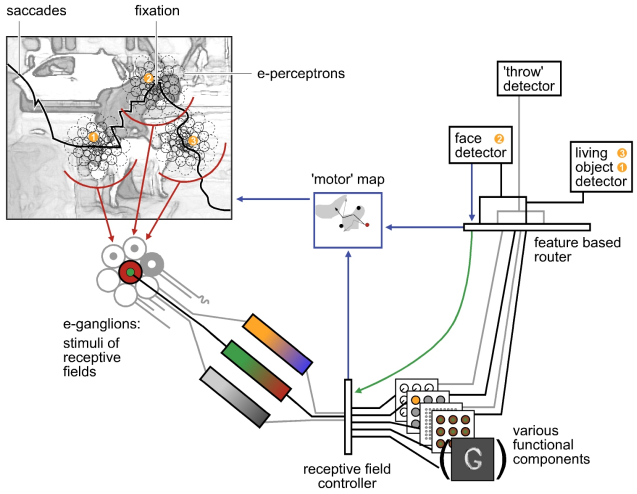

Fig. 8. Model: Architecture.

3.2 Intermediate Layer I: Receptive Field Controller

The functional feature decomposition components register at the receptive field controller in order to limit their input to a specified selection of all available data fired by the e-ganglions. They can send down impulses (coming e.g. from higher areas of cognition or from their own processing results) that impede or enhance the activitations from the e-ganglions.

[7] Luminance-, red-green-, and blue-yellow antagonism with on-centre, off-centre, and on-off-centre behavior, both in sustained and transient activation mode.

The receptive field controller acts as a router organising and forwarding the e-ganglion output to the feature decomposition components (and between them). Moreover, the controller sends data to the 'motor' maps component to influence exploration and sampling. Microtremors (very small movements of the fixation point) ensure that pattern matching is not too sensitive to absolute positioning.

3.3 Primal Feature Decomposition

Fed from the receptive field controller, primal feature separation takes place in the feature decomposition components. These, for example, imitate the frequency band filtering mechanisms or the orientation splitting in the hypercolumns of the visual cortex. Here again it is important to support not only bottom up processing, but to facilitate top-down communication of activations in order to drive exploration and sampling. This is especially neccessary, as in a fixation period only a small (thumb-nail sized) clipping of the potential stimulus material of a complete image will be processed.

The activations originating the receptive fields of the e-ganglions, which are agglomerated by the mechanisms of the band width filtering process (and others), are used to stimulate parts of models of concepts through the next intermediate layer, the intelligent feature router. When reaching a certain threshold level in one of these components, they lead to a complete activation of the matching model and to the recognition of a corresponding label, if the concept has already been learnt. The compartments of the 'what'- and 'where'-stream described above in figure 7 rest mainly in this layer. Some of them, however, have to be interconnected or even serialized. For example the output of the four low frequency channels (as depicted in the 'G'-component in figure 8 resp. 6) converge onto a primal feature detector that serves as preprocessor for a figure ground separator component.

3.4 Intermediate Layer II: An Intelligent Feature Router

By introducing an intermediate bidirectional permissive layer between the crude feature detection mechanisms and the components which emulate a specific partial concept activation mechanism of the human long term memory, the way for a distributed architecture is paved: without understanding, what actually is processed, a feature router can be installed that forwards potentially interesting material to specific detector components (in parallel or consecutively). The other way round, detector components can send feature requests: they can emit (spatially bound) mark-up of interesting features or regions within the field of sight – top-down to the controller and the 'motor' map component. The router may be rule-based, pattern-based or based on a trainable neural net.

A message channel ensures that information gained in one primal feature or concept detector component can be accessed by the others.

3.5 Concept Detector Components

Concept detector components receive input from various primal feature detectors (similarily to the components of the 'what'- and 'where'-stream). By driving attention and focus top-down, first assumptions can be discarded or asserted later on in the perception process. As the retinotrop perception favors lazy, spot-oriented processing, many object recognition problems (see above) can be

avoided, especially context effects and shape morph effects. Several alternatives arise considering the basic working process of a concept detector component. Static and dynamic (=learning via a neural net) pattern matching mechanisms compete with graph-based feature segmentation methods (cf. [9], [10]).

3.6 Controlling Exploration: Motor Maps

Motor maps are fed from both intermediate layers. They are not retinotrop but have an egocentric field of sight resolution. They control where the next saccade is pointing to and which location will be fixated. The motor map component especially has to be coupled to components of a 'where'-stream equivalent.

3.7 Learning an Ontology

To develop new concept detector components in a system implementing this architectural model, the resulting (semantics-free) output of the feature decomposition layer needs to be analysed. Especially, revisualising low-frequency filter band information helps in the identification of course shapes. Combined with information from the 'motor' map component that enables spatial localisation, this can be used to find rules, build pattern algorithms or train a neural net capable of matching the visual representations of the concept of desire.

4 Conclusion and Future Work

The human vision process has been described in detail. An model imitating this vision process has been outlined. However, an important aspect for human vision, the clocking, i.e. timing constraints in processing, remains uncovered. Furthermore, the 'grammar' (e.g. graphs) and 'vocabulary' (e.g. visual variables) of human vision at the interface of primal feature detection and concept detection needs further investigation. Ways of automatic visual learning (e.g. based on MPEG7 shape mark-up) will additionally guide future research.

References

1. A. Baddeley. The fractionation of working memory. *Proceedings of the National Academy of Science of the USA*, 93(24):13468–13472, 1996.
2. A. Baddeley. *Episodic memory. New directions in research.* Oxford University Press, Oxford, UK, 2002.
3. L. Barsalou, K. Simmons, A. Barbey, and C. Wilson. Grounding conceptual knowledge in modality-specific systems. *TRENDS in Cognitive Science*, 7(2):84–91, 2003.
4. I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.
5. T. Braver and J. Cohen. Working memory, cognitive control and the prefrontal cortex. *Cognitive Processing*, 7(1):25–55, 2001.
6. C. Curtis and M. D'Esposito. Persistent activity in the prefrontal cortex during working memory. *TRENDS in Cognitive Science*, 7(9):415–423, 2003.
7. U. Drewniak. *Lernen mit Bildern in Texten.* Waxman, Münster, Germany, 1992.
8. J. Engelkamp. Gedächtnis für Bilder. In K. Sachs-Hombach and K. Rehkämper, editors, *Bild – Bildwahrnehmung – Bildverarbeitung*, Wiesbaden, 1998. Universitäts-Verlag.
9. J. Feldman. What is a visual object? *TRENDS in Cognitive Science*, 7(6):252–256, 2003.

10. P. Gaerdenfors. *Conceptual Spaces. The Geometry of Thought*. MIT Press, Cambridge, MA, 2000.
11. A. Ginsburg. Spatial filtering and visual form perception. In Boff, Kaufman, and Thomas, editors, *Handbook of Perception and Human Performance, Vol. II: Cognitive Processes and Performance*, pages 1–41, 1986.
12. A. Gramopadhye and K. Madhani. Visual search and visual lobe size. In *IWVF 4, LNCS 2059*, Berlin, 2001. Springer.
13. S. Grossberg. How does the cerebral cortex work? development, learning, attention, and 3d vision by laminar circuits of visual cortex. In S. Grossberg, editor, *Behavioral and Cognitive Neuroscience Reviews*, 2003.
14. D. Hoffman. *Visuelle Intelligenz*. Klett-Cotta, Stuttgart, 2001.
15. R. Höger. Speed of processing and stimulus complexity in low-frequency and high-frequency channels. *Perception*, 26:1039–1045, 1997.
16. R. Höger. *Raumzeitliche Prozesse der visuellen Informationsverarbeitung*. Scriptum Verlag, Magdeburg, 2001.
17. H. Kolb. How the retina works. *American Scientist*, 91(Jan/Feb):28–35, 2003.
18. H. Kolb, E. Fernandez, and R. Nelson. *Webvision. The Organisation of the Retina and Visual System*. John Moran Eye Center of the University of Utah, Salt Lake City, 2005. http://webvision.med.utah.edu/.
19. H. Mallot. *Sehen und die Verarbeitung visueller Information*. Vieweg, Braunschweig, 2000.
20. D. Marr. *Vision*. Freeman, San Francisco, 1982.
21. G. Murch. Human factors of color displays. In *Advances in Computer Graphics*, Berlin, 1986. Eurographics.
22. J. Nairne. The myth of the encoding-retrieval match. *Memory*, 10(5-6):389–395, 2002.
23. R. Pfeifer and C. Scheier. *Understanding Intelligence*. MIT Press, Cambridge, MA, 2001.
24. P. Pichler. Prinzipien der Bildverarbeitung im visuellen System des Menschen, 2002. http://www.informatik.uni-ulm.de/ni/Lehre/SS02/Proseminar_CV/ausarbeitungen2/ppichler.pdf.
25. E. Pöppl. Informationsverarbeitung im menschlichen Gehirn. *Informatik Spektrum*, 16(December):427–437, 2002.
26. F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. In Cummins and Cummins, editors, *Minds, Brains, and Computers*, Oxford, 2000(1958). Blackwell.
27. M. Scaife and Y. Rogers. External cognition: how do graphical representations work? *International Journal of Human-Computer Studies*, 45(2):185–213, 1996.
28. A. Treisman. Features and objects. *Quarterly Journal of Experimental Psychology*, 40A(2):201–237, 1988.
29. A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.
30. S. Ullman. The visual recognition of three-dimensional objects. In Meyer and Kornblum, editors, *Attention and Performance XIV*, Cambridge, MA, 1993. MIT Press.
31. T. Vilis. The physiology of the senses. Transformations for perception and action, 2005. http://www.med.uwo.ca/physiology/courses/sensesweb/.
32. B. Weidenmann. *Psychische Prozesse beim Verstehen von Bildern*. Verlag Hans Huber, Bern, 1988.
33. J. Wolfe. Moving towards solutions to some enduring controversies in visual search. *TRENDS in Cognitive Science*, 7(2):70–76, 2003.
34. S. Zeki. *Inner Vision*. Oxford University Press, London, 1999.