Workshop on "Active Visual Learning and Hierarchical Visual Representations for General-Purpose Robot Vision"

Hong Kong - May 31st, 2014

Deep Representation Hierarchies for 3D Active Vision

Silvio P. Sabatini

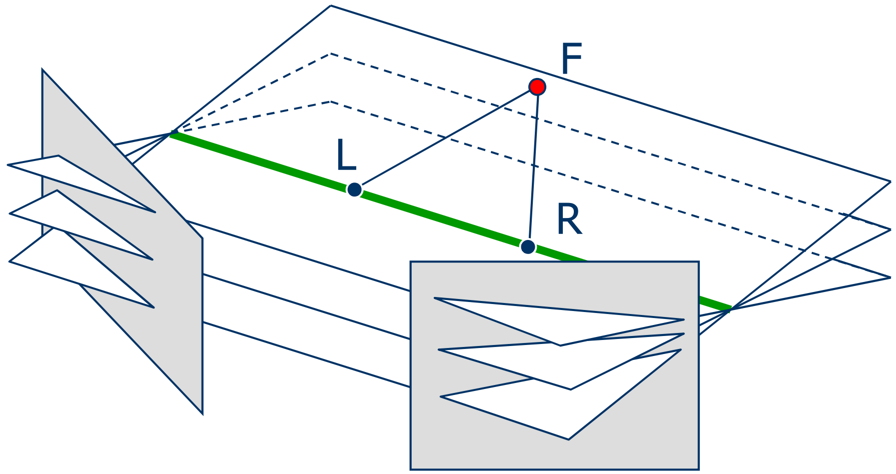
Department of Informatics, Bioengineering, Robotics and Systems
University of Genoa

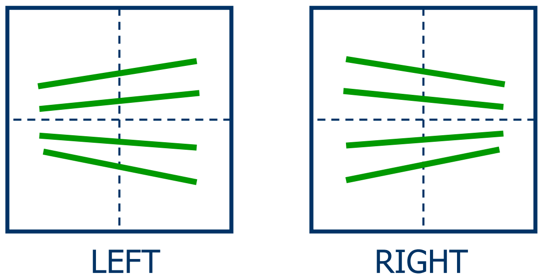
UNIVERSITÀ DEGLI STUDI DI GENOVA dibris PSPC The Physical Structure of Perception and Computation www.pspc.unige.it

Binocular eye movements & stereopsis

Convergent optical axes

- Deviations from primary position rotate the epipolar lines and vertical disparities (VD) become possible
- As the eyes move the epipolar lines move and become more tilted
- Larger search zones to solve the stereo correspondence problem





An active vergent system has to cope with the attendant aperture problem for binocular disparity

Different specializations

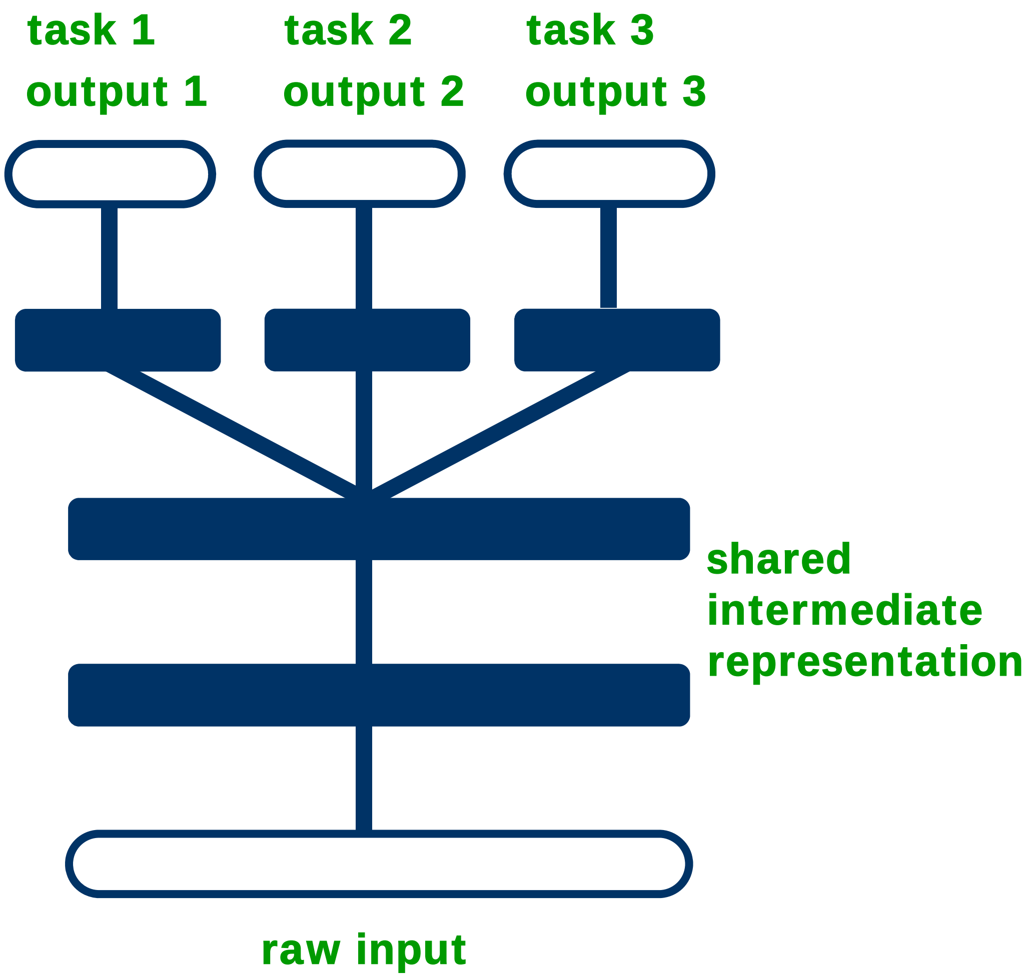
… *for reciprocal improvement of stereopsis and binocular control of eye movements*

→ ***VERGENCE AS A PARADIGMATIC TASK***

- The question arises how to learn *disparity-vergence response curves*, directly (without explicit calculation of the disparity map)
- We will demonstrate that it is possible to gain different specializations according to the paradigm of deep architecture

Deep architectures

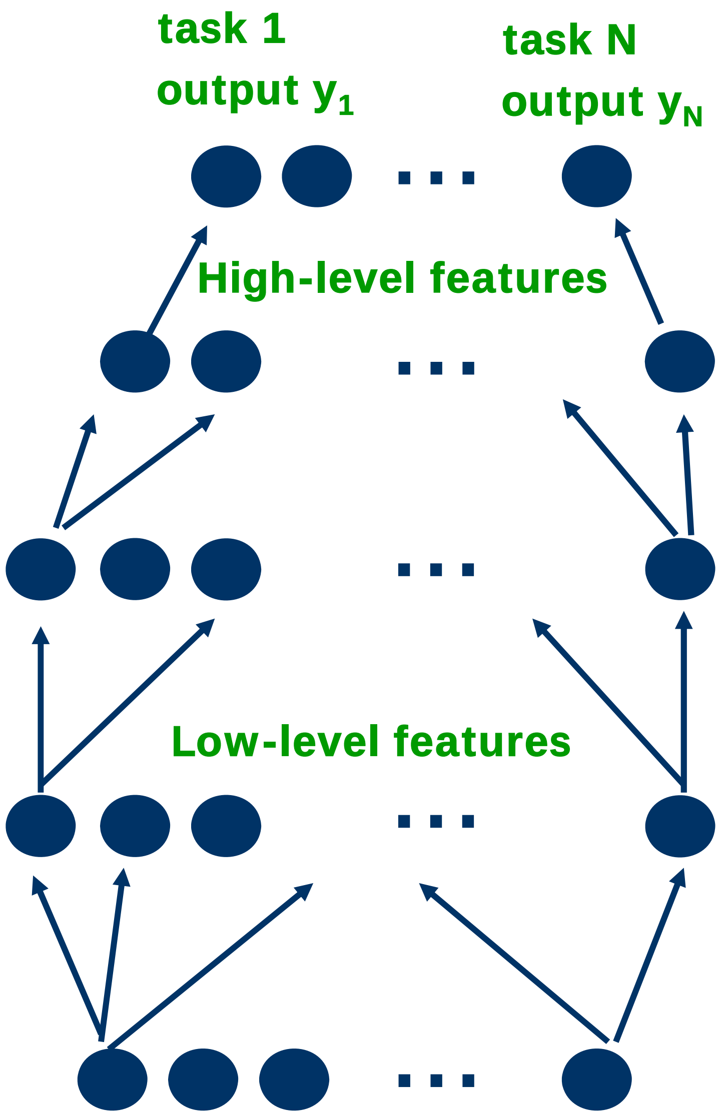
- Deep architectures learn good intermediate representations that can be *shared* across tasks



[Adapted from Bengio, 2009]

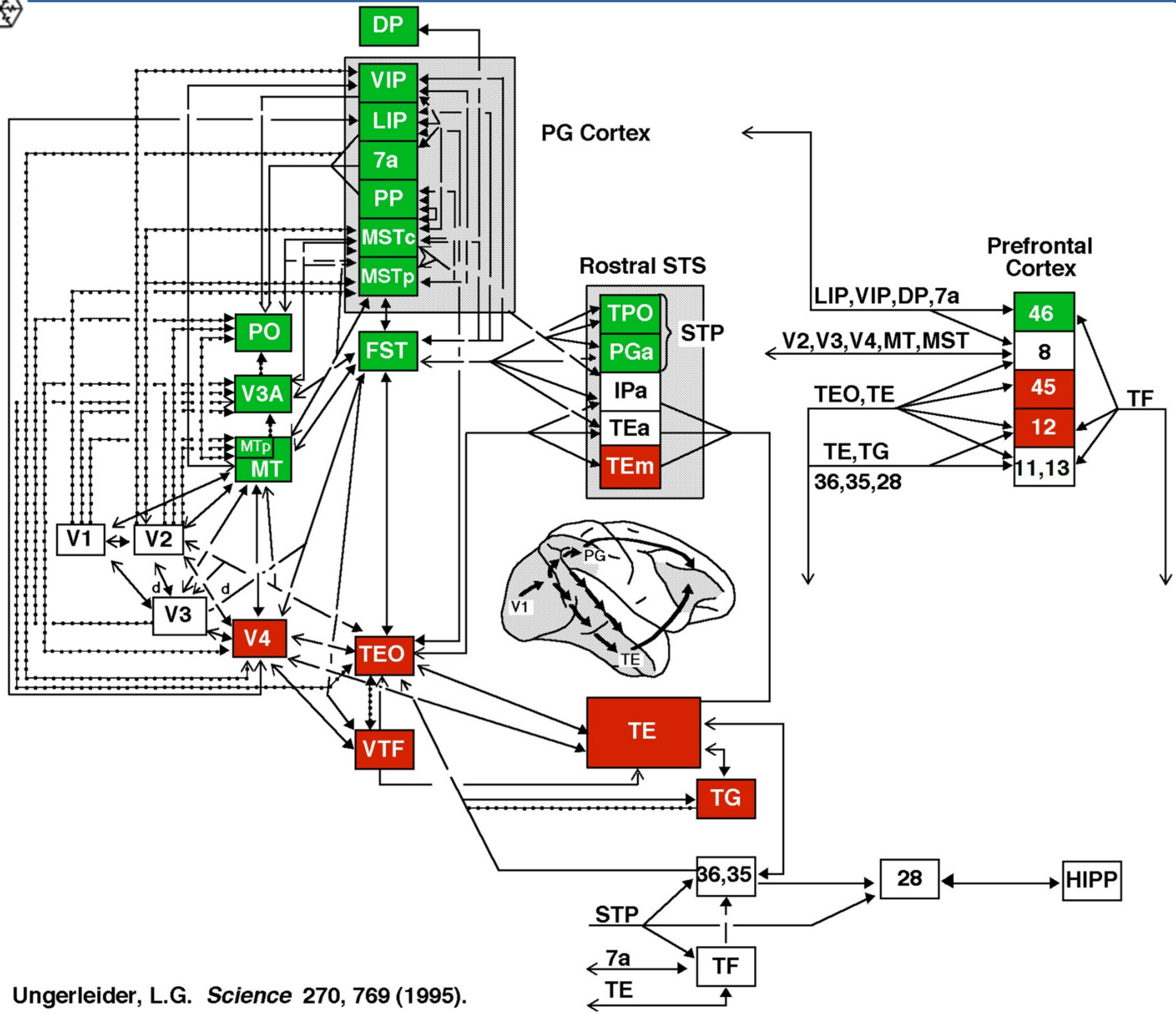
Deep architectures

- Deep architectures learn good intermediate representations that can be *shared* across tasks
- Different tasks can share the same high-level feature
- Different high-level features can be built from the same set of lower-level features



[Adapted from Bengio, 2009]

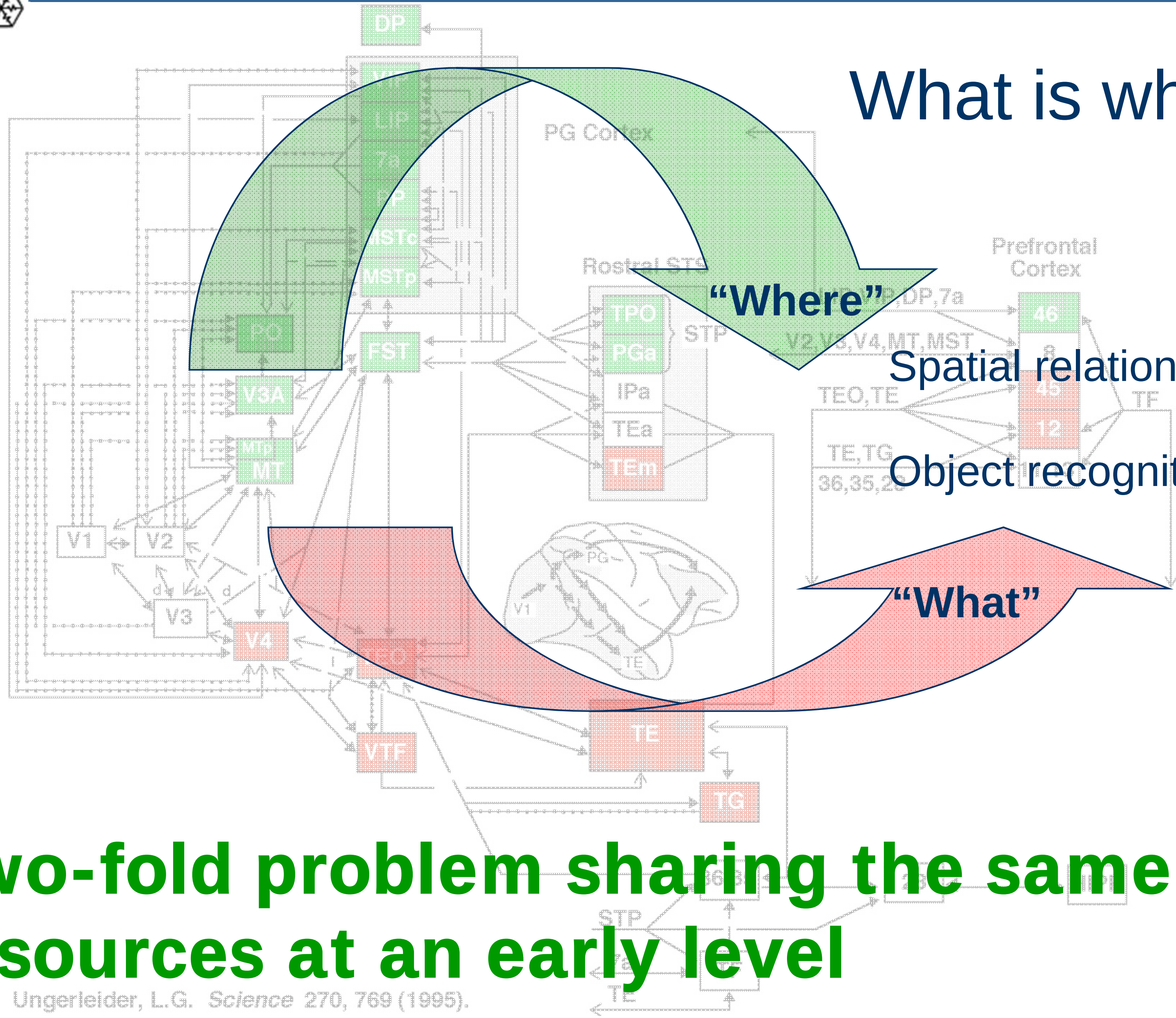
Cortical architecture



Ungerleider, L.G. *Science* 270, 769 (1995).

Hierarchical processing of depth

What is where?



Spatial relationships

Object recognition

A two-fold problem sharing the same resources at an early level

Ungerleider, L.G. *Science* 270, 769 (1995).

Building distributed representations of the binocular visual signal

Harmonic featureless representation

Q: What features?

A: Local amplitude, phase and orientation

Through a multi-channel Gabor-like decomposition of the visual signal

Pros:

- **Higher flexibility** having not decided *a priori* what features to be extracted
- We can rely on a **powerful computational theory**

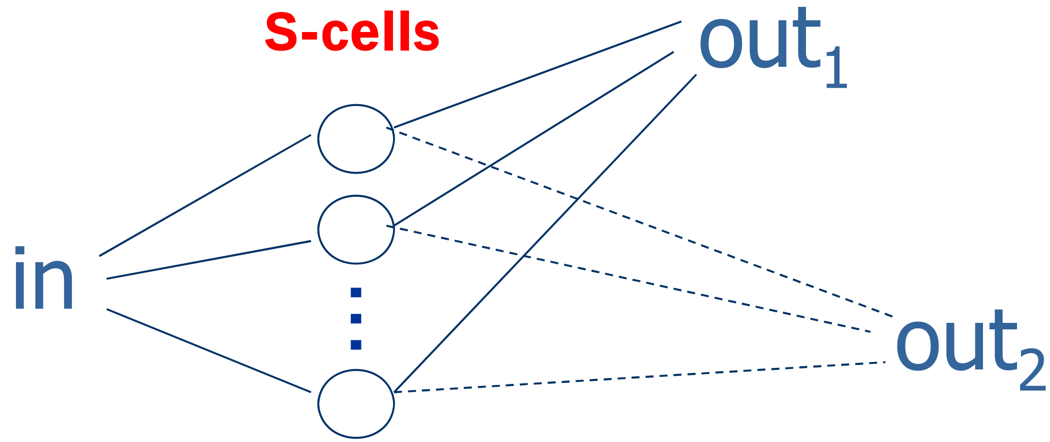
Cons:

- Features are derived qualities based on local phase properties

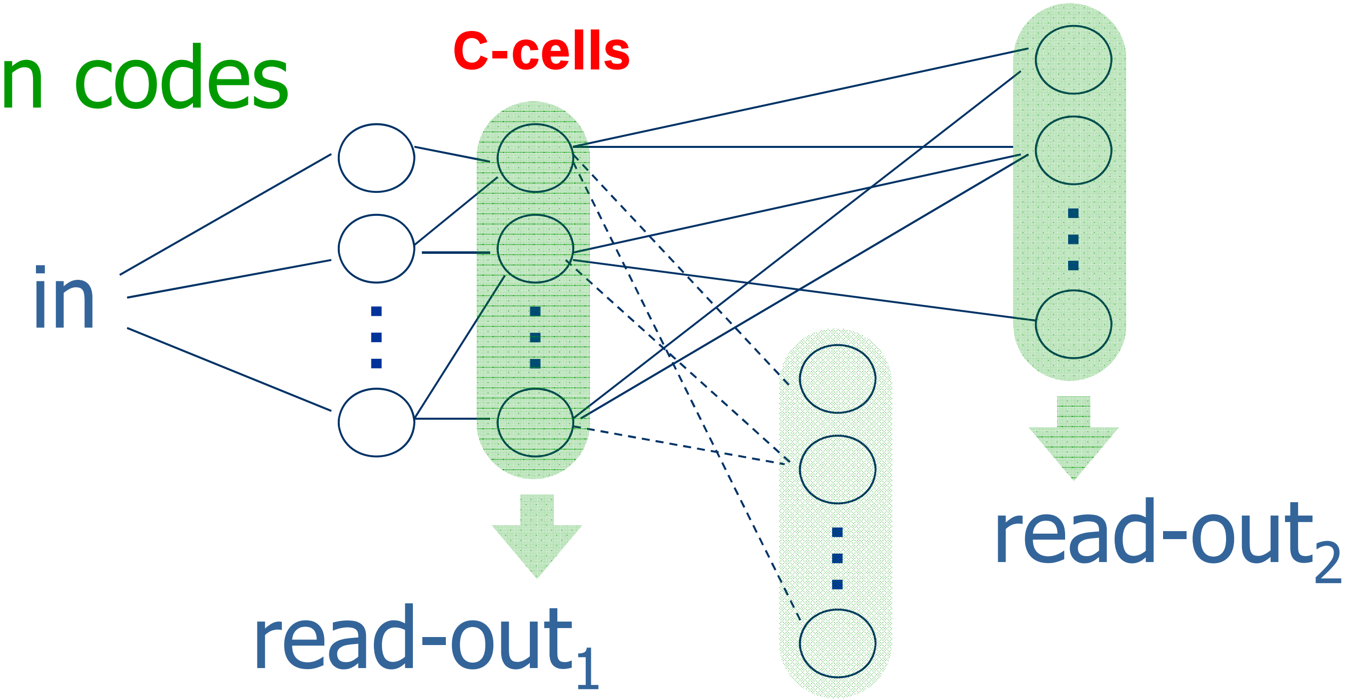
Contrast discontinuities	→	phase *congruency*
Binocular disparity	→	phase *difference*
Visual motion	→	phase *constancy*

Deep representation hierarchies

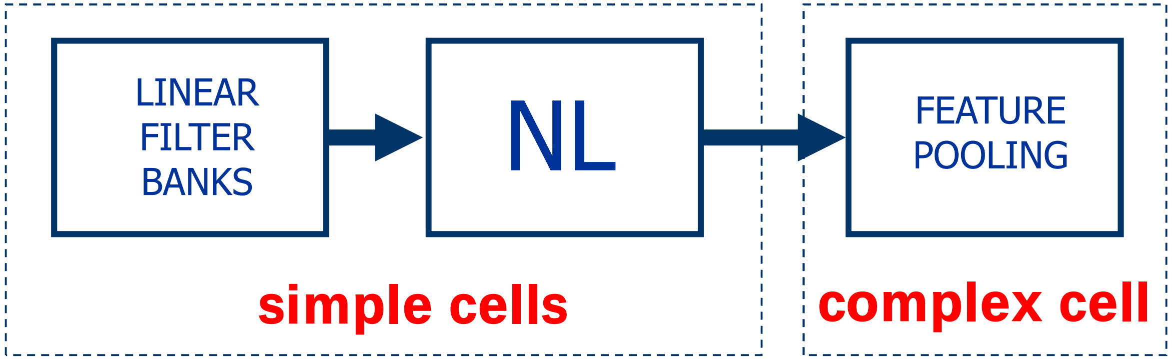
Measures



Population codes

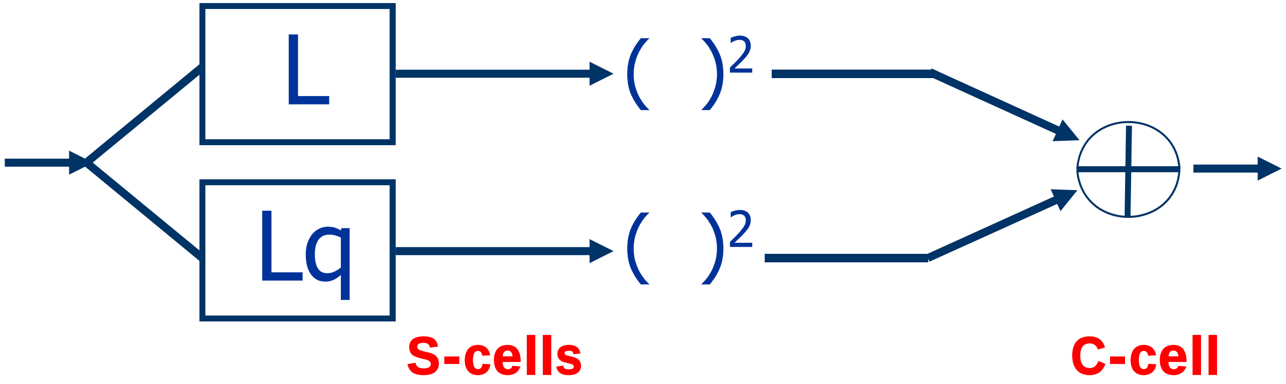


Complex cells



Complex cells “pool” the output of simple cells within a retinotopic neighborhood

Energy models



Linking phase and energy models

Phase-based measures …

Contrast discontinuities → phase *congruency*

Visual motion → phase *constancy*

Binocular disparity → phase *difference*

vs.

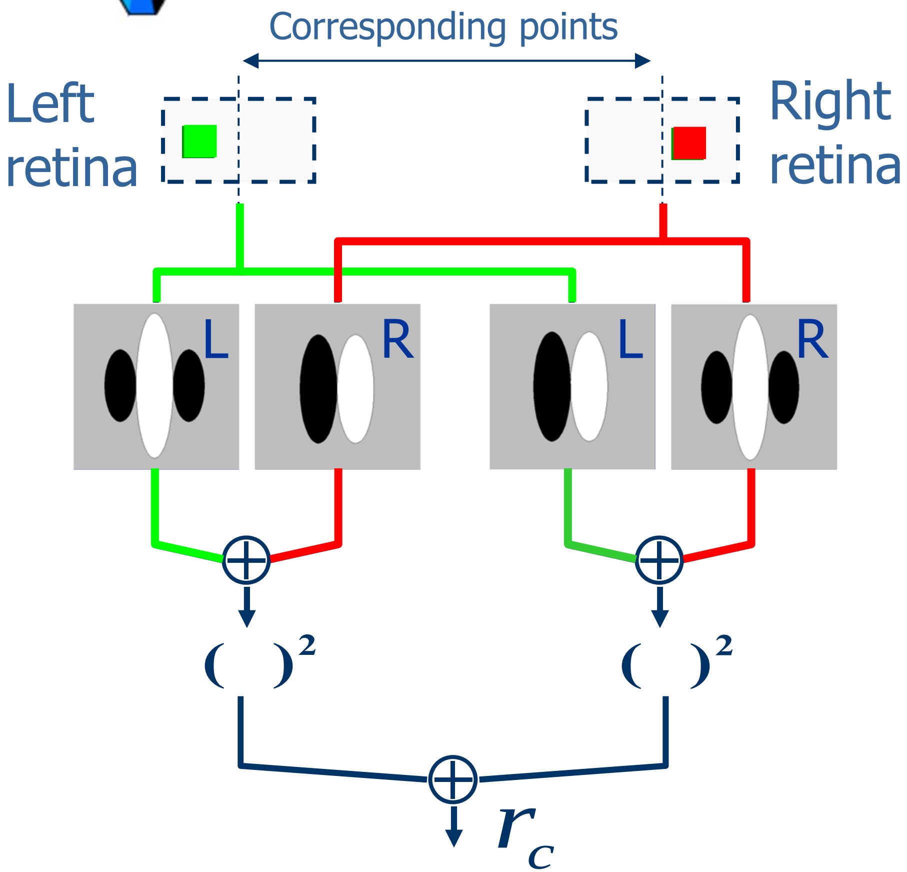
… energy coding

→ Contrast energy [Morrone & Burr, 1982, 1988]

→ Motion energy [Adelson & Bergen, 1985]

→ Binocular energy [Ohzawa et al., 1990]

Binocular energy unit



$$I^L(x), \quad I^R[x+\delta(x)]$$

$$\mathbf{h}^L(x;k_0,\psi_L)=e^{-x^2/\sigma^2}e^{i(k_0x+\psi_L)}$$
$$\mathbf{h}^R(x;k_0,\psi_R)=e^{-x^2/\sigma^2}e^{i(k_0x+\psi_R)}$$

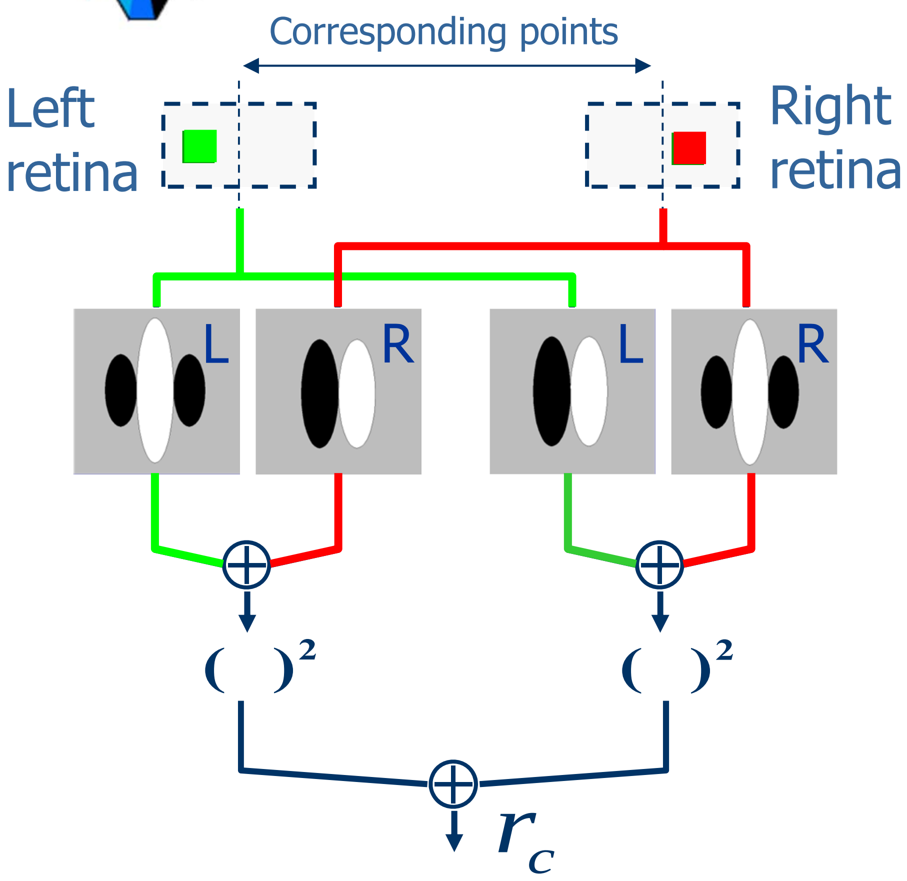
$$Q^{L/R}(x)=\mathbf{h}^{L/R}*I^{L/R}(x)e^{-j\psi_{L/R}}$$

$$r_c(x_0)=\left|Q^L(x_0)+\mathrm{e}^{j\Delta\psi}Q^R(x_0)\right|^2$$

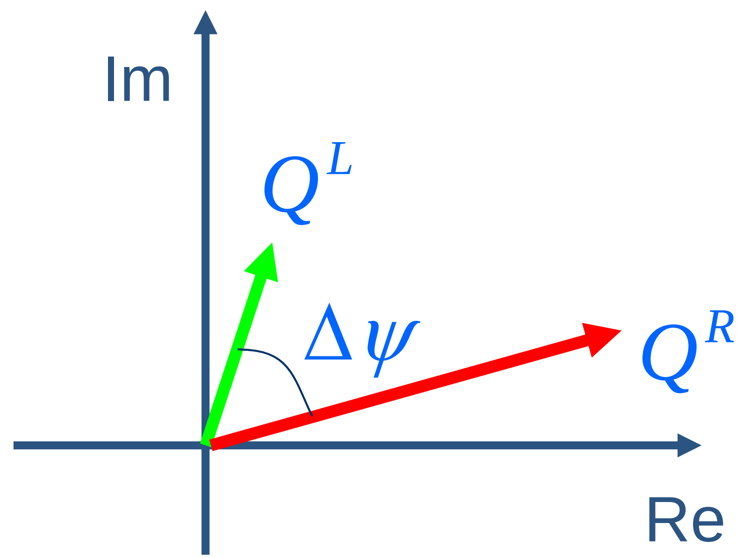
where $\Delta\psi=\psi_R-\psi_L$

[Qian, 1994][Fleet et al., 1996]

Binocular energy unit



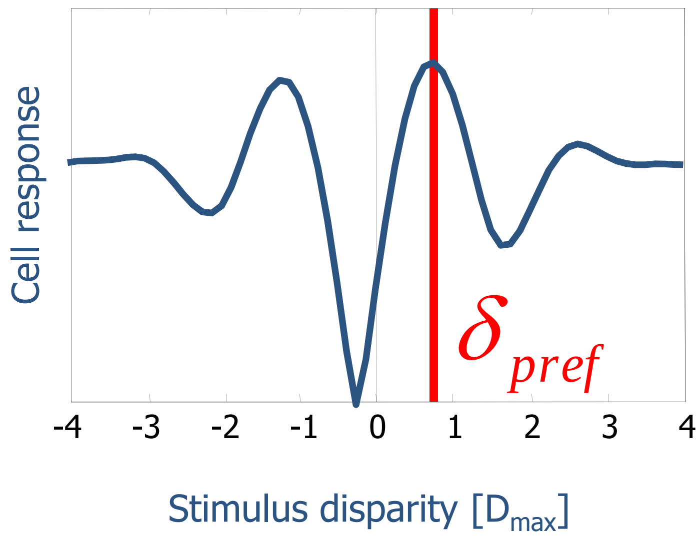
$$r_c(x_0) = \left| Q^L(x_0) + e^{j\Delta\psi} Q^R(x_0) \right|^2$$



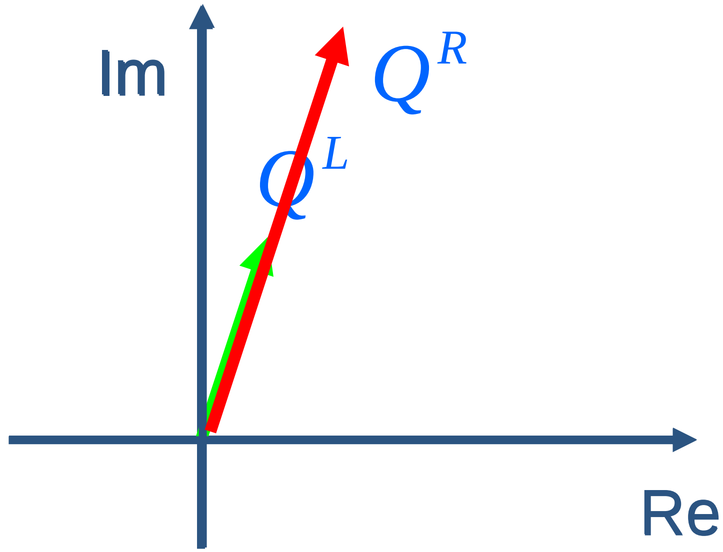
The binocular energy unit maximally responds when $\Delta\psi$ matches the image phase disparity $\Delta\phi$.

Binocular energy unit

Disparity tuning curve



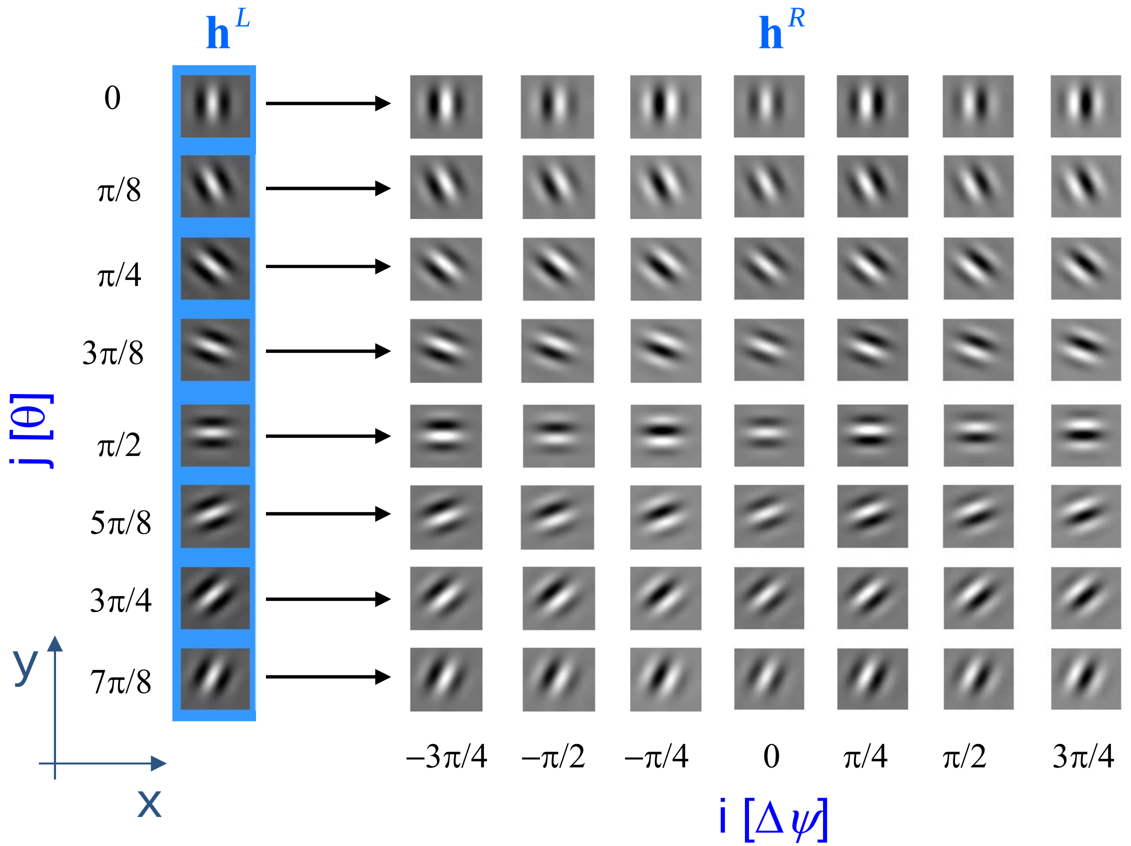
$$\boxed{\delta_{pref} \propto \frac{\Delta\psi}{k_0}}$$



The binocular energy unit maximally responds when $\Delta\psi$ matches the image phase disparity $\Delta\phi$.

Large scale cortical architectures

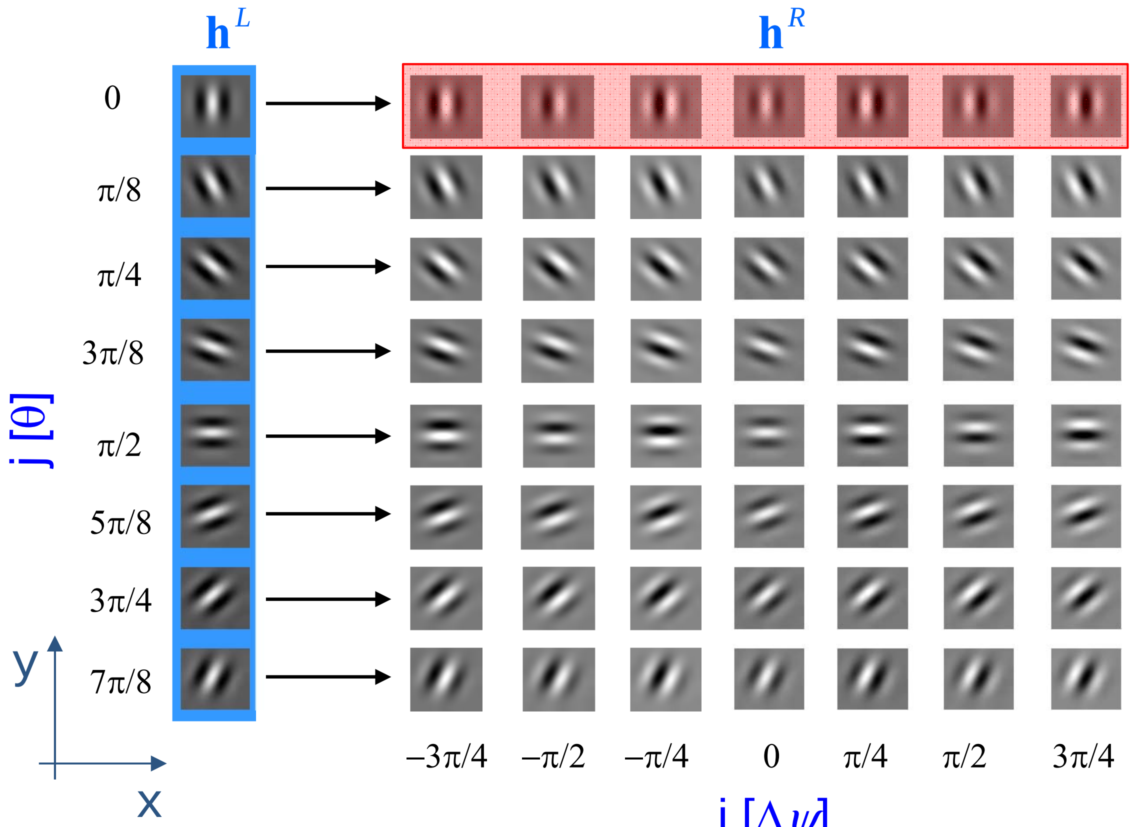
2×56 binocular receptive fields for each pixel

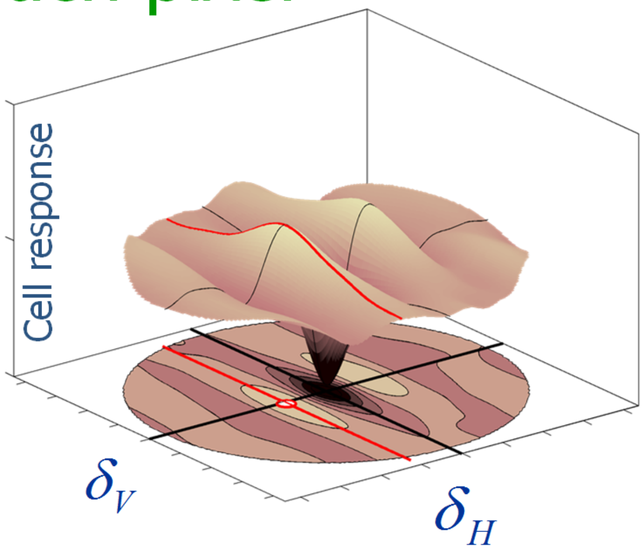


A set of oriented Gabor receptive fields with different phase shifts but centered at the same retinal position.

Large scale cortical architectures

2×56 binocular receptive fields for each pixel





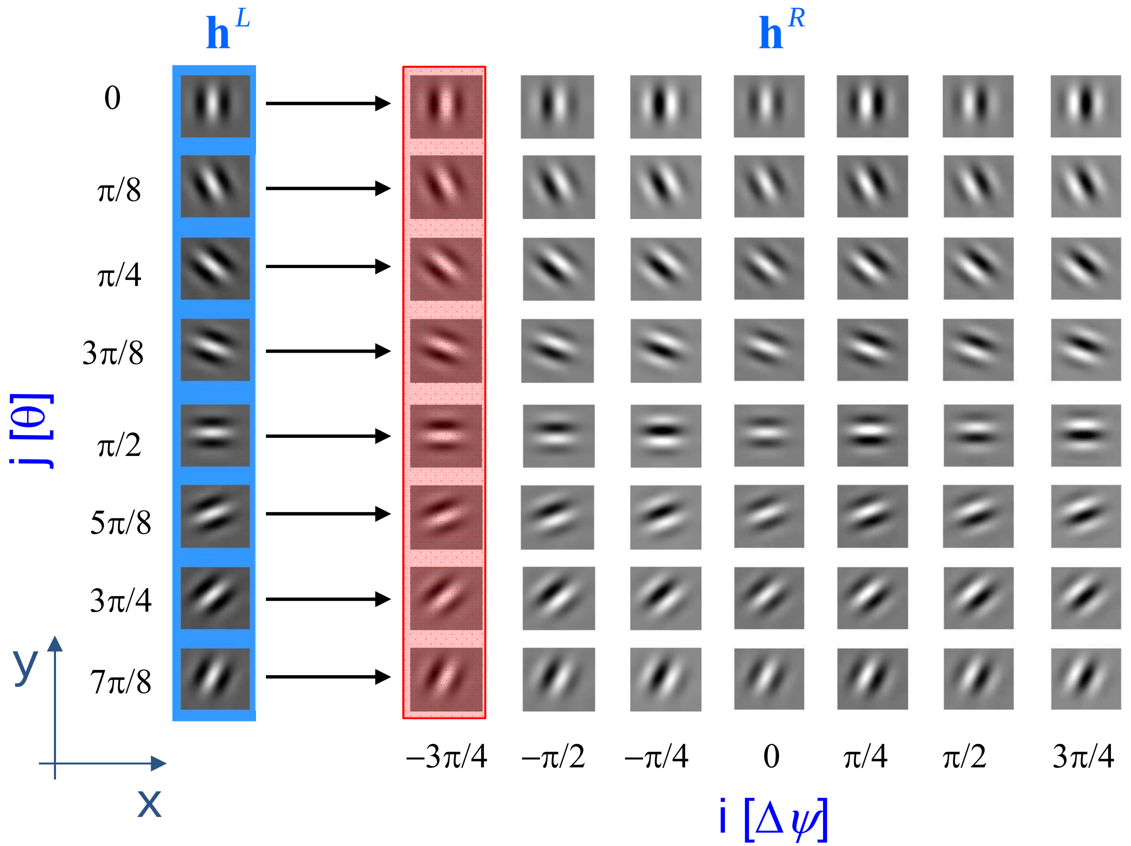
Disparity tuning surface

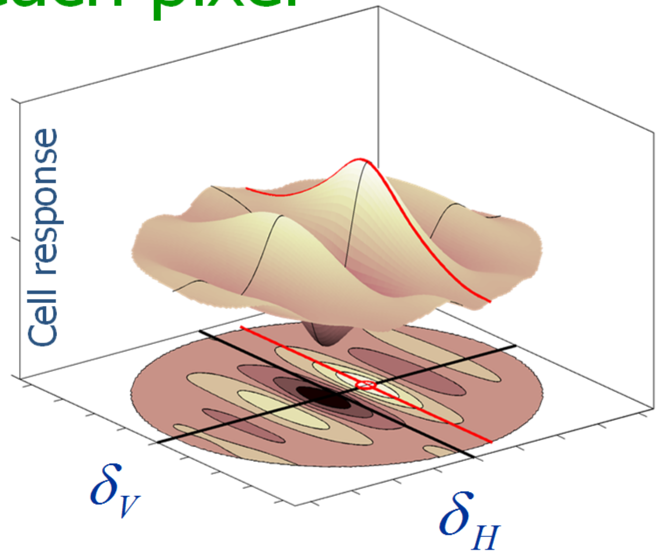
$$\boxed{\delta_{pref} \propto \frac{\Delta\psi}{k_0}}$$

[M. Chessa, S.P. Sabatini and F. Solari *A fast joint bioinspired algorithm for optic flow and two-dimensional disparity estimation*. 7th Int. Conference on Computer Vision Systems (ICVS'09), 13-15 October 2009, Liege, Belgium.]

Large scale cortical architectures

2×56 binocular receptive fields for each pixel





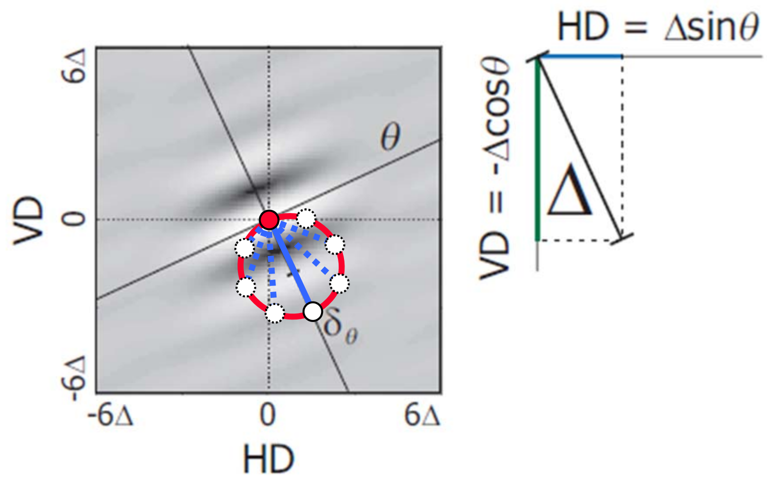
Disparity tuning surface

$$\delta^{ij}_{pref} \propto \mathbf{k}_0^j \frac{\Delta\psi^i}{\left\|\mathbf{k}_0^j\right\|^2}$$

[M. Chessa, S.P. Sabatini and F. Solari *A fast joint bioinspired algorithm for optic flow and two-dimensional disparity estimation*. 7th Int. Conference on Computer Vision Systems (ICVS'09), 13-15 October 2009, Liege, Belgium.]

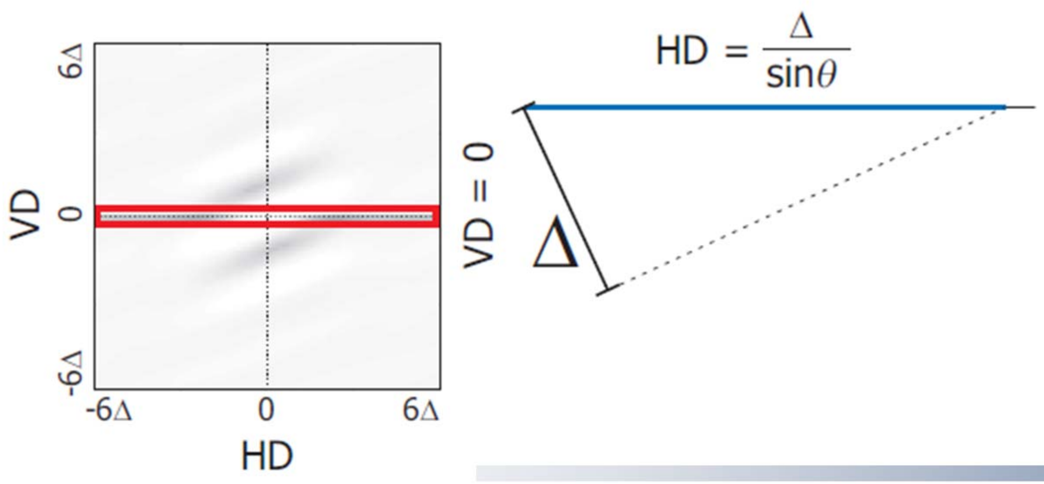
Enabling disparity-vergence responses in stereo-heads

Disparity estimation



Δ = maximum detectable disparity along the direction orthogonal to the cell's orientation, equals one half cycle of the peak spatial frequency of the RF

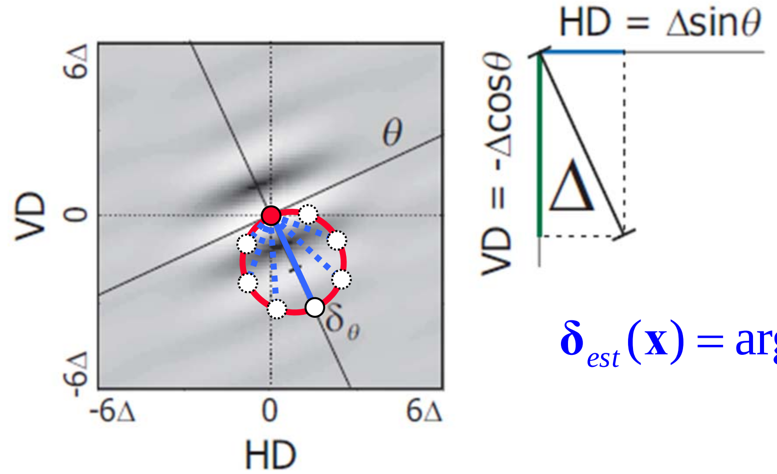
Direct vergence control



Assuming VD ≅ 0, the orientation is used to extend the sensitivity range of the cells' population to HD stimuli.

3D active vision requires ≠ specializations

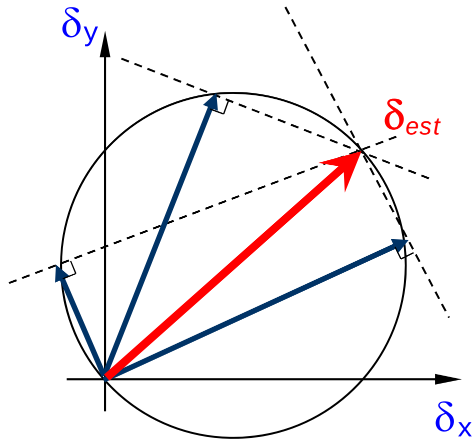
Disparity estimation



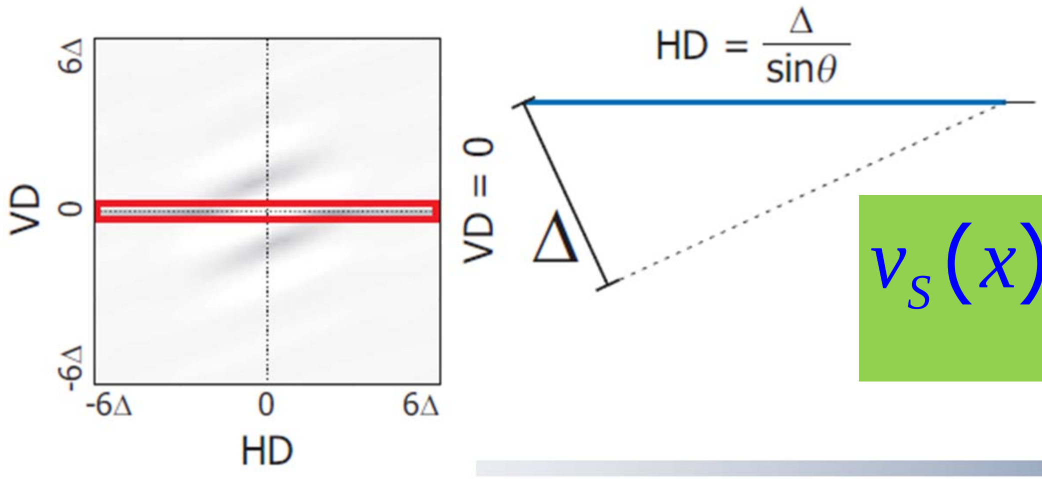
$$\delta^{\theta}_{est} = \frac{\sum_i \delta^i_{pref}\, r^i_c(x)}{\sum_i r^i_c(x)}$$

$$\boldsymbol{\delta}_{est}(\mathbf{x}) = \arg\min_{\boldsymbol{\delta}(\mathbf{x})} \sum_\theta \left(\delta_\theta(\mathbf{x}) - \frac{\mathbf{k}_0^T}{k_0} \boldsymbol{\delta}(\mathbf{x}) \right)$$

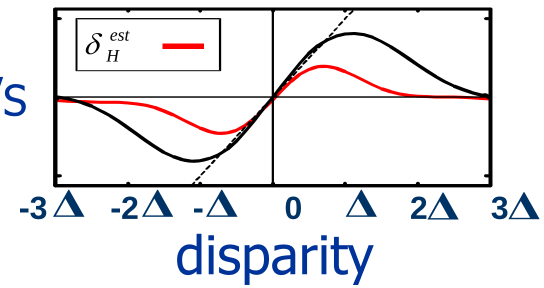
Intersection of constraints



Direct vergence control



Vs



$$v_S(x) = \sum_{x\in\Omega} G(x) \sum_i w_i\, r^i_c(x)$$

Learning Algorithms

Reinforcement learning, based on particle swarm optimization algorithm

[A. Gibaldi, A. Canessa, M. Chessa, F. Solari, S.P. Sabatini. *How a population-based representation of binocular visual signal can intrinsically mediate autonomous learning of vergence control*. Procedia Computer Science 13: 212–221, 2012]

Supervised learning, based on LeNet non-linear convolutional network [N. Chumerin, A. Gibaldi, S.P. Sabatini and M.M. Van Hulle *Learning Eye Vergence Control from a Distributed Disparity Representation*. International Journal of Neural Systems, Vol. 20, p 267-278, 2010]

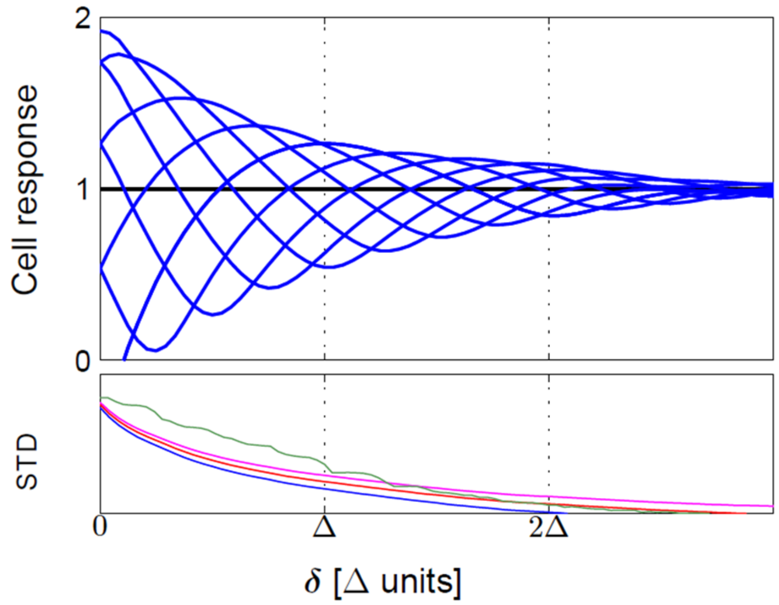
Differential Hebbian Rule:

$$\boxed{w_i\big|_t = (1-\eta)\, w_i\big|_{t-1} + \eta\, \mathrm{V}_S\left(\mathbf{r}_c\big|_{t-1}\right) \Delta r_c^i}$$

Δr_c^i : Differential population response

$\mathrm{V}_S\left(\mathbf{r}_c\big|_{t-1}\right)$: Vergence signal at instant t-1

η : Learning rate = **ΔSTD of the population response**

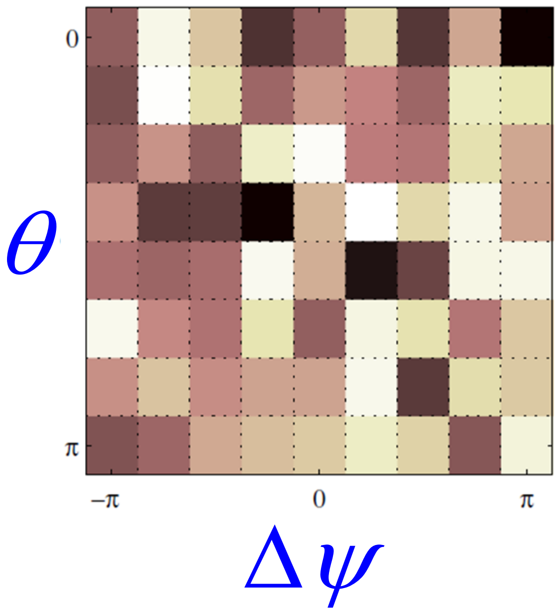


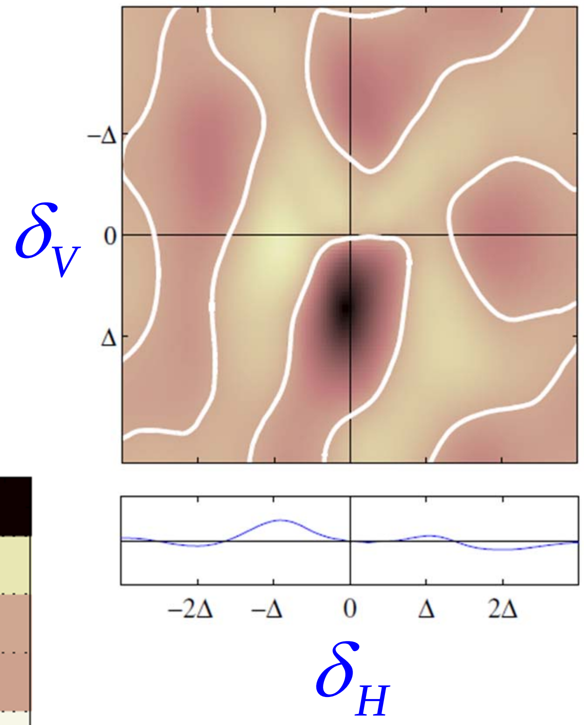
→ INTRINSIC REWARD!

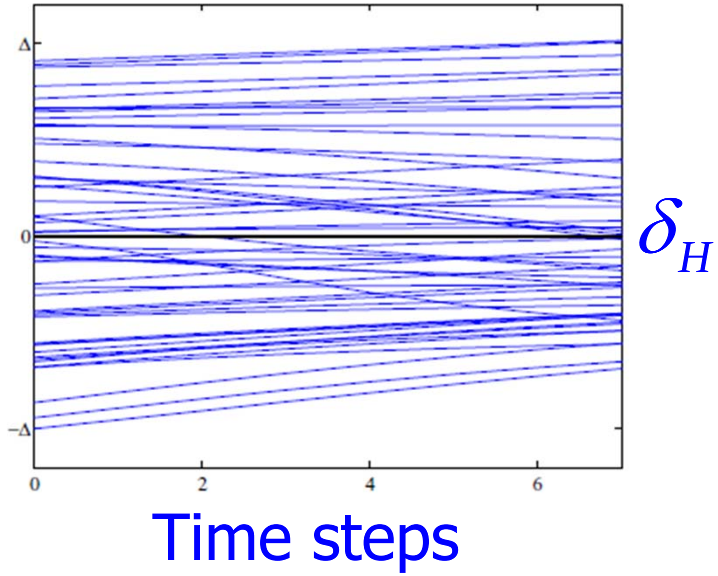
Results

@trial 50

weight distribution



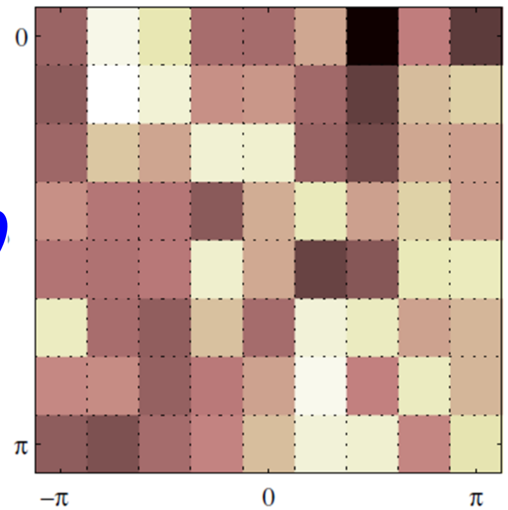


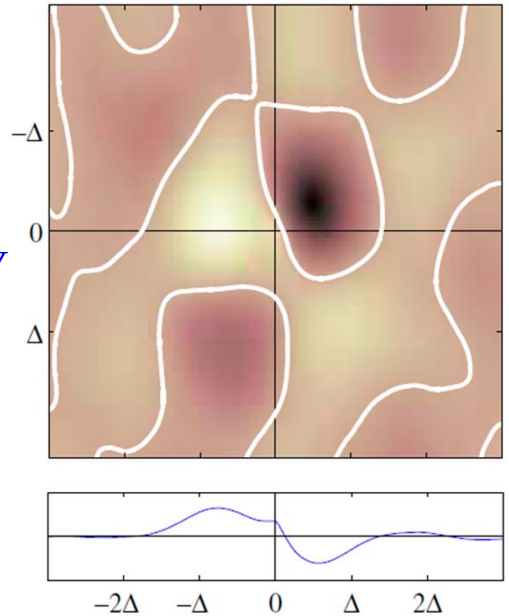


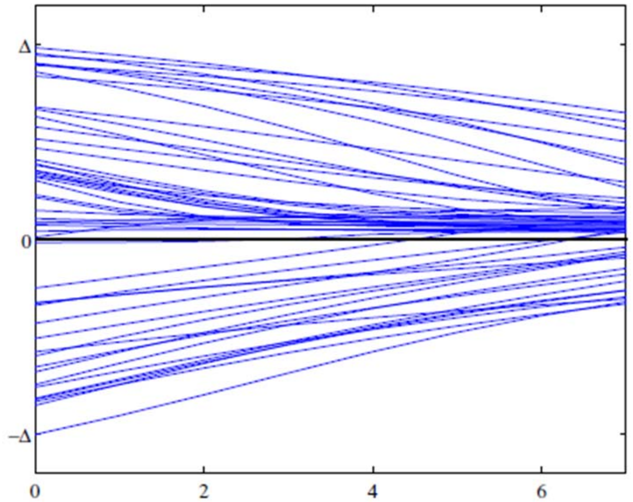
Results

@trial 200

weight distribution



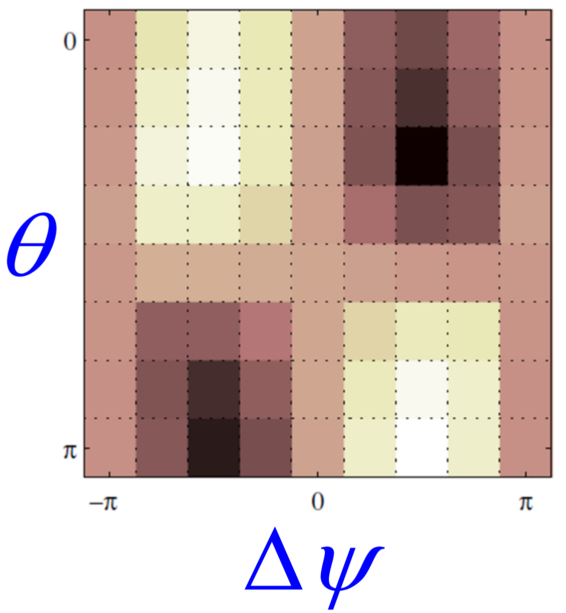


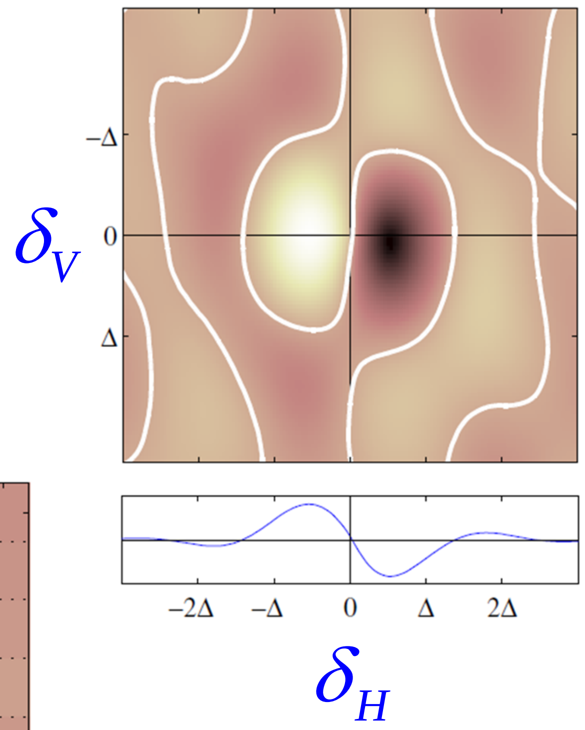


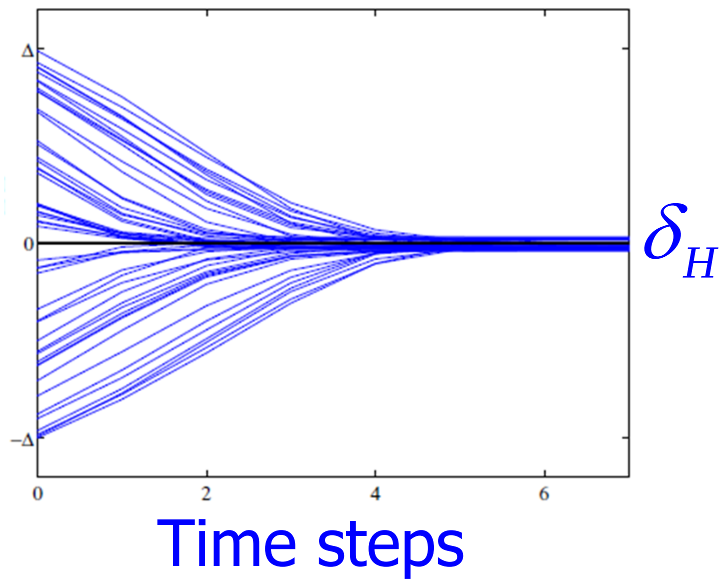
Results

@trial 1500

weight distribution





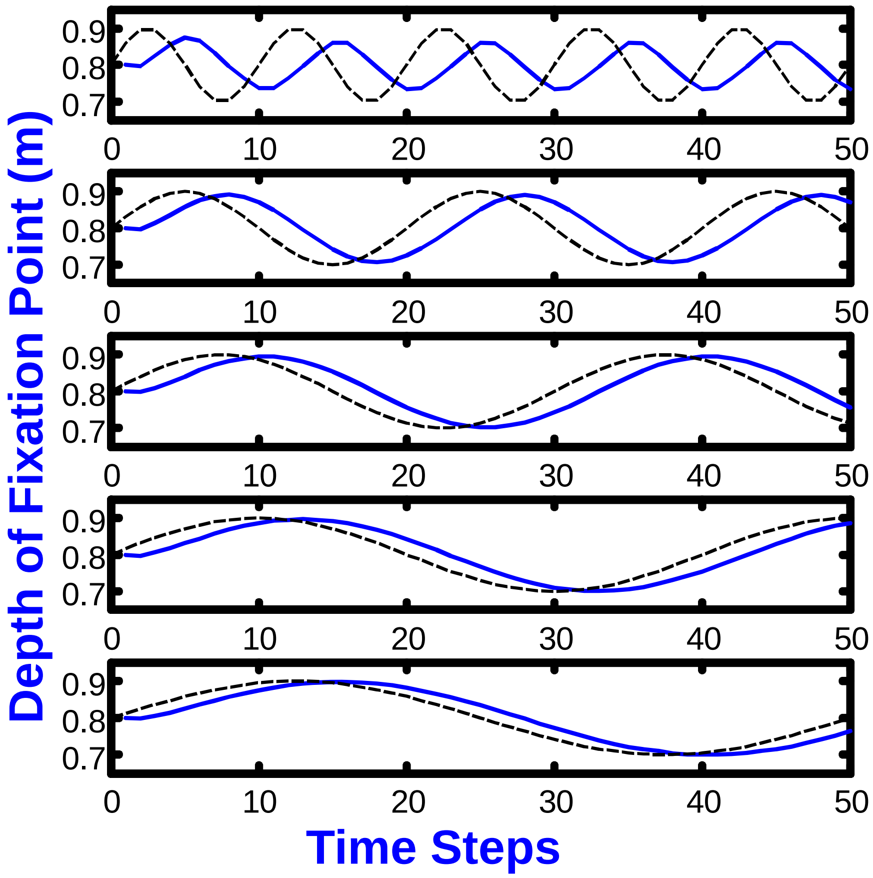


[Gibaldi et al., *Autonomous Learning of Disparity-Vergence Behaviour through distributed coding and population reward: basic mechanisms and real-world conditioning on a robot stereo head*. Robotics & Automation Systems Journal , submitted]

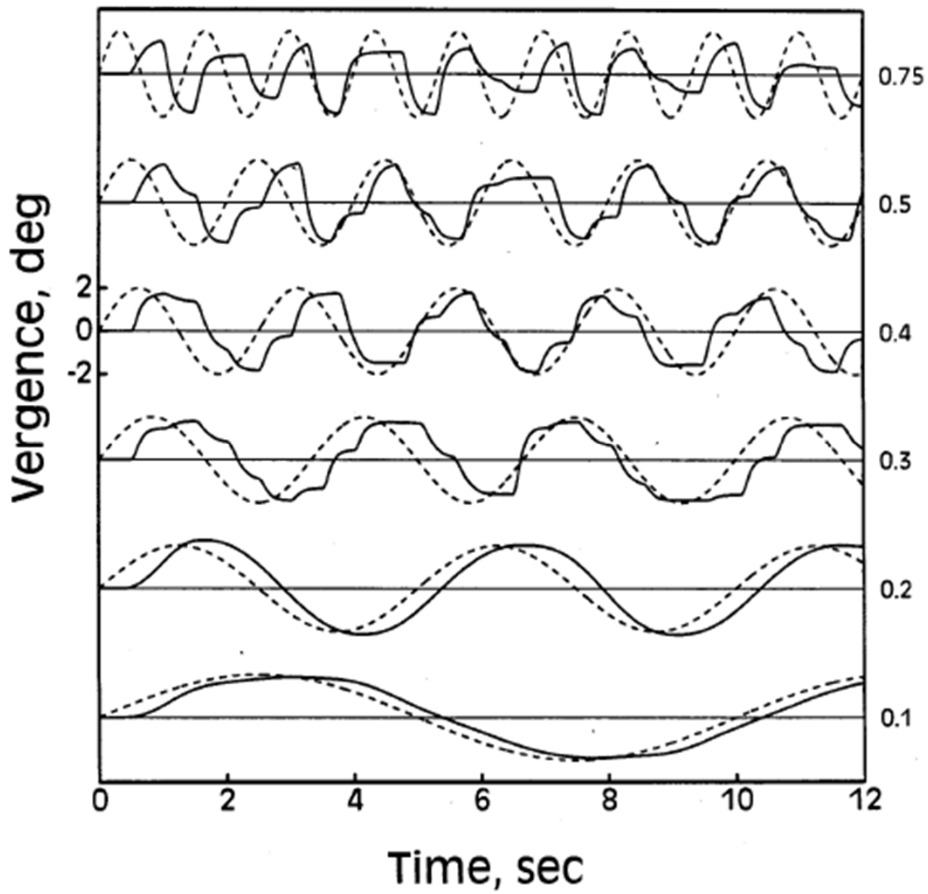
Simulation Results

moving stimuli

Model



Real data



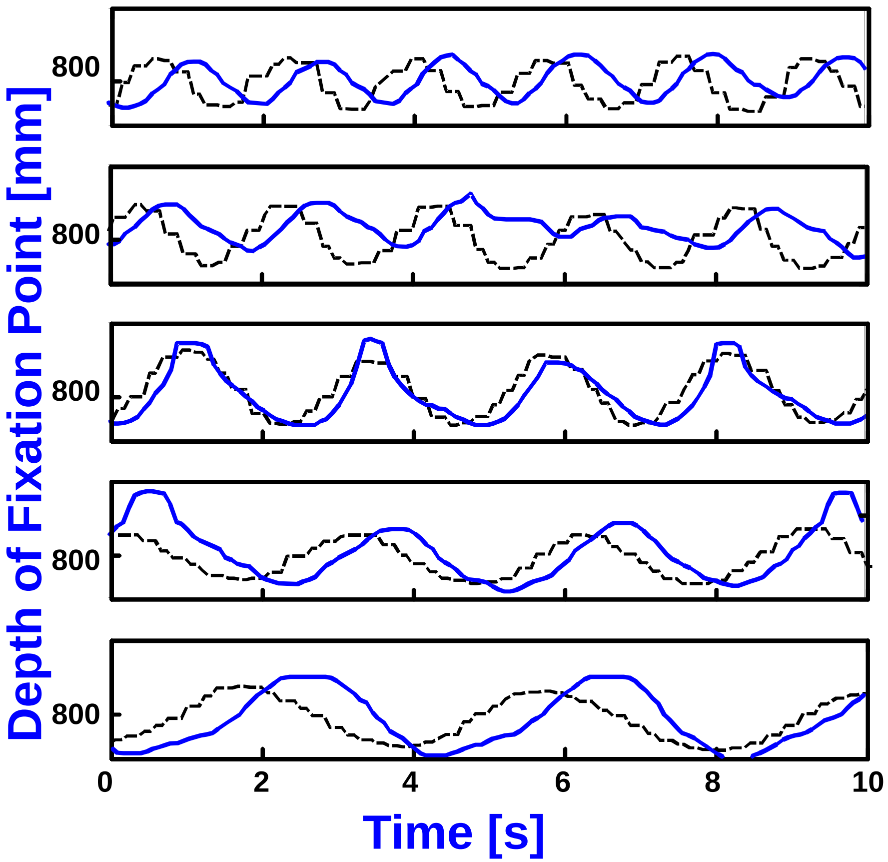
[A. Gibaldi, M. Chessa, A. Canessa, S.P. Sabatini, F. Solari A cortical model for binocular vergence control without explicit calculation of disparity. Neurocomputing, Vol. 73, p 1065-1073, 2010.]

[Hung, 1997]

Experimental Results

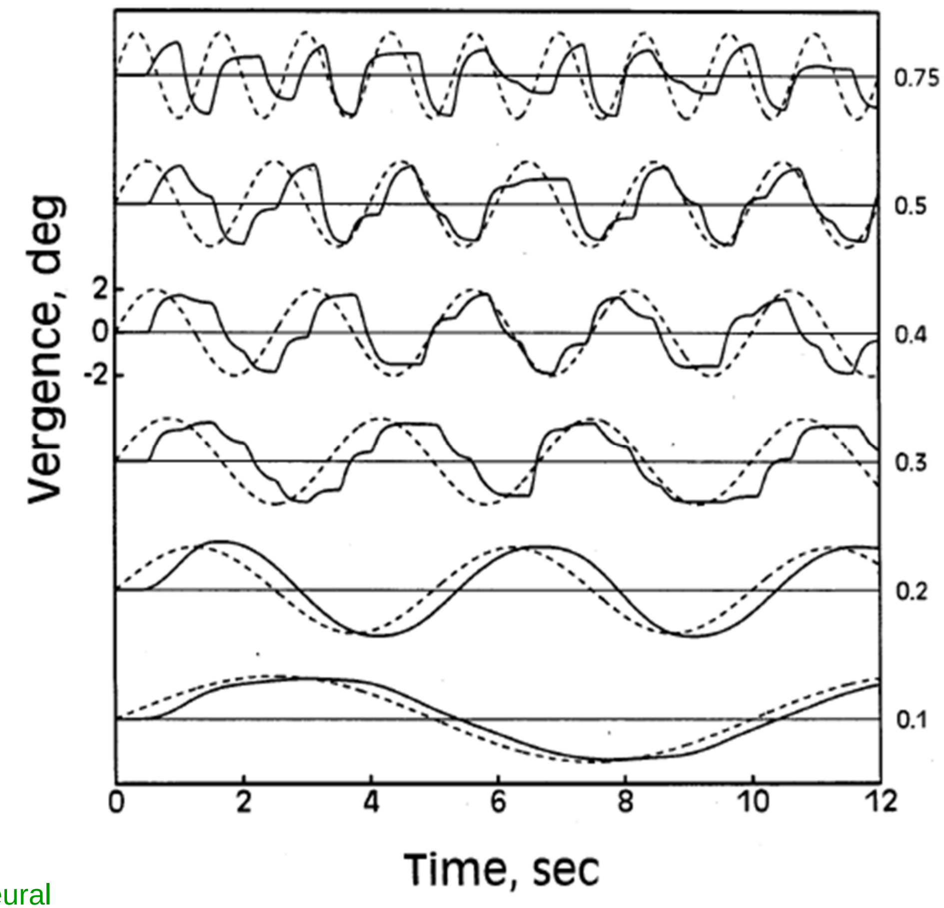
moving stimuli

iCub



[A. Gibaldi, A. Canessa, M. Chessa, F. Solari, S.P. Sabatini. A neural model for coordinated control of horizontal and vertical alignment of the eyes in three-dimensional space. Proc. 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), 24-27 June 2012.]

Real data



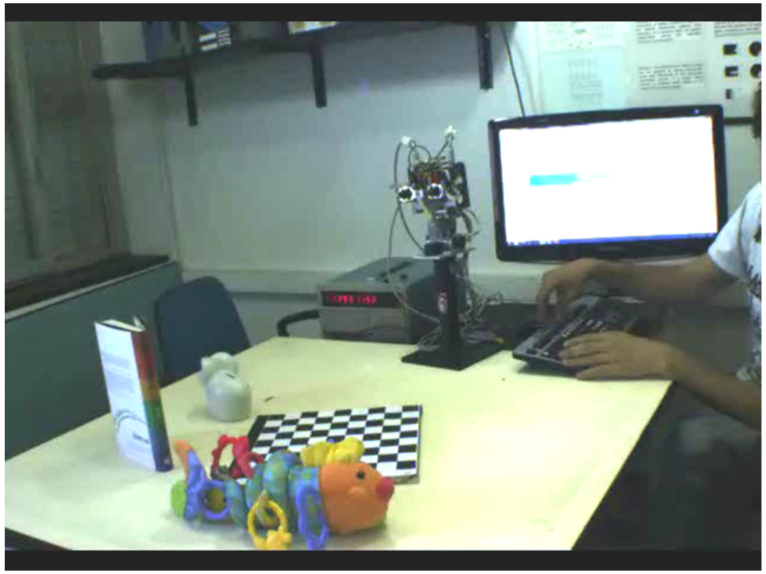
[Hung, 1997]

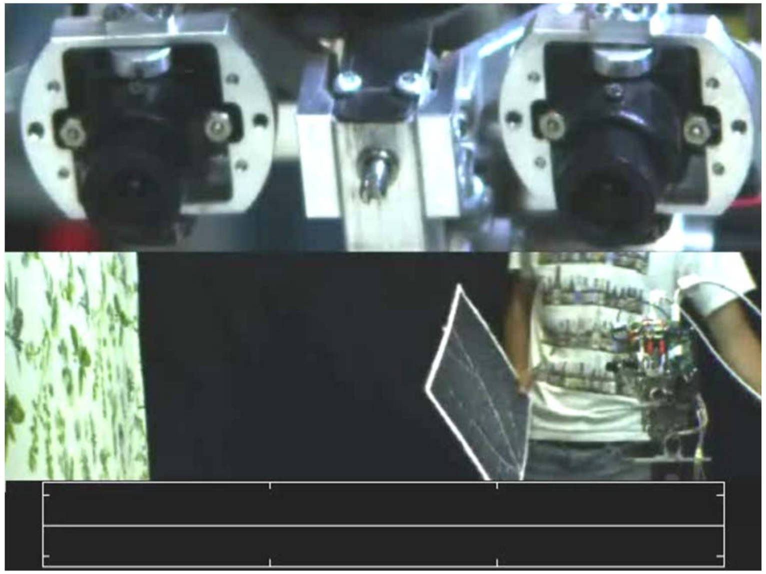
Experimental results

Videos: http://www.eyeshots.it/res_news.php

Switching fixations among static visual targets

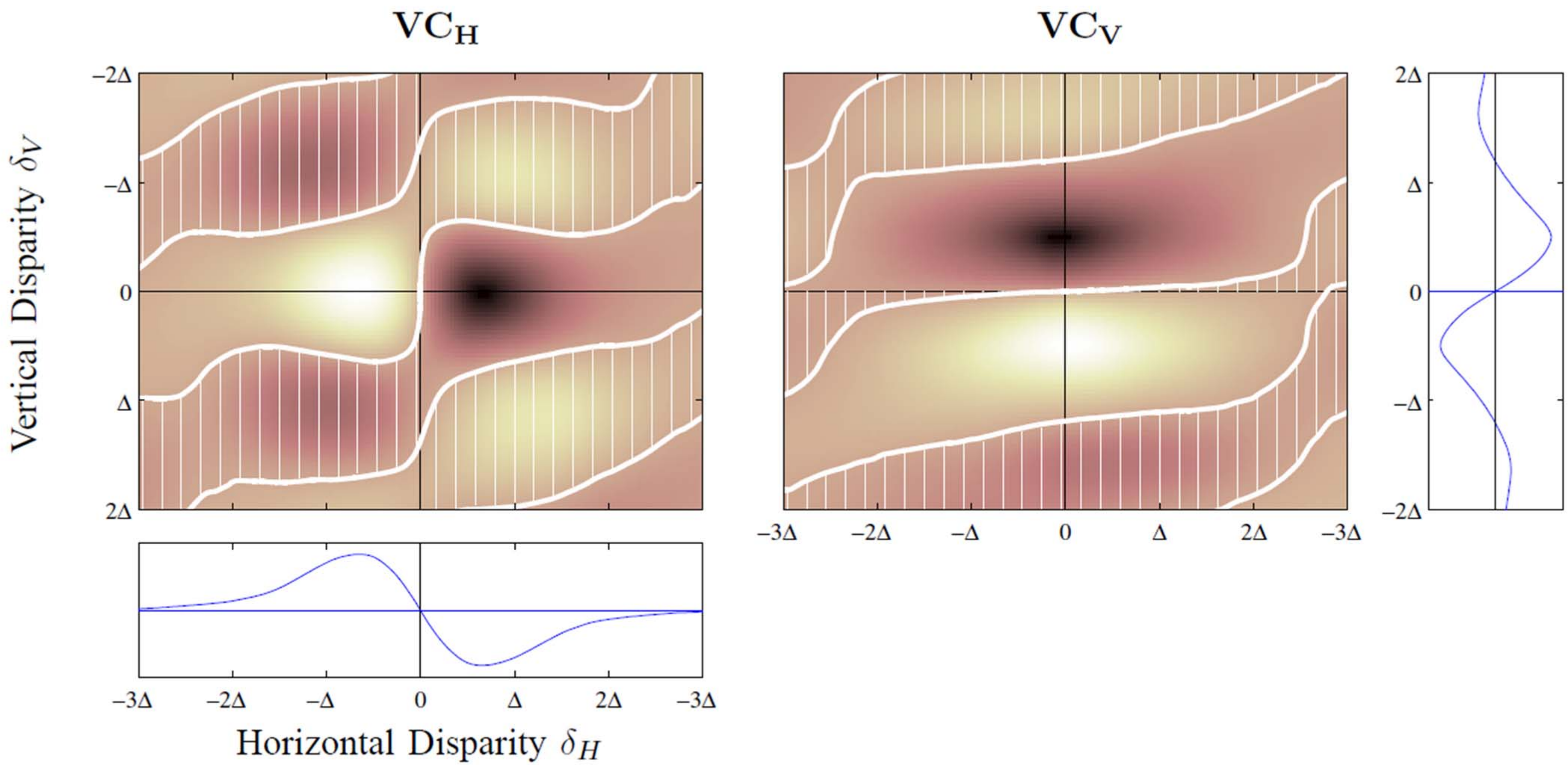
Stepping and waving objects





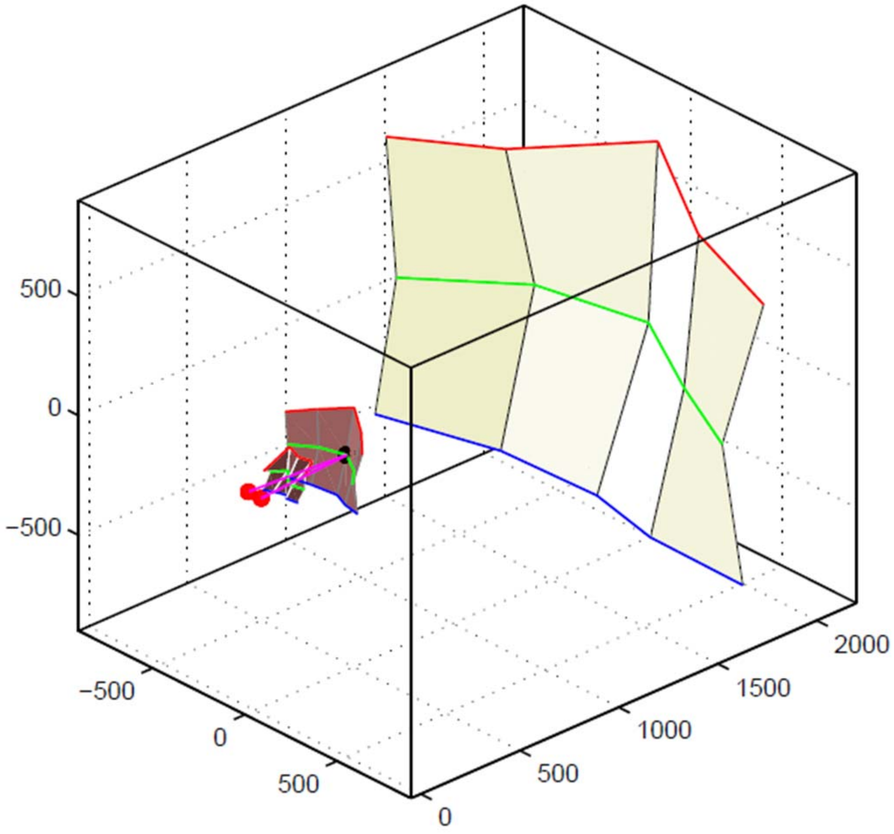
Videos

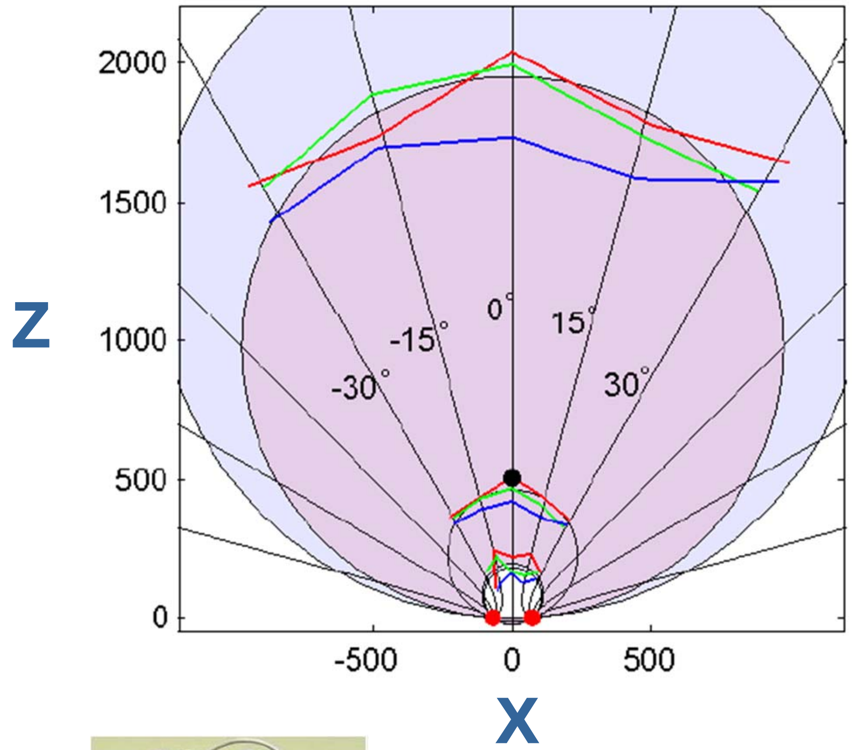
Combined control of horizontal and vertical vergence



Measured vergence working ranges

Helmholtz (=Tilt-Pan) system



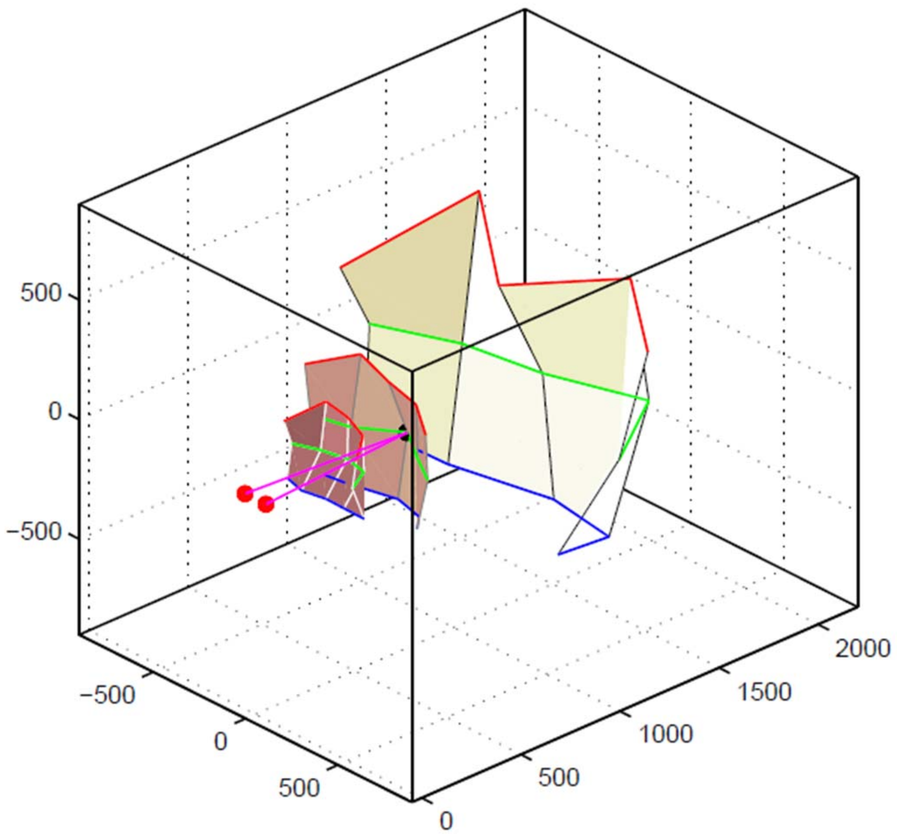


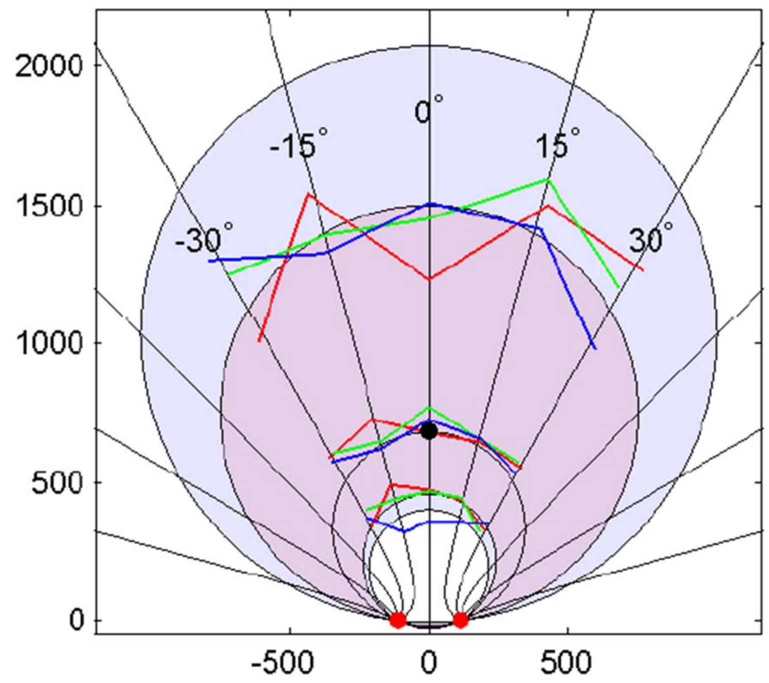
iCub

[Gibaldi et al., submitted]

Measured vergence working ranges

Fick (=Pan-Tilt) system





Koala

[Gibaldi et al., submitted]

Progress toward a *Neuroware* for humanoid robots

Progress toward a *Neuroware*

for humanoid robots

- **Goal:** development of a neural library on GPU to enable real-time perceptual processing through neuromorphic paradigms

  - perceptual engines accessible through SW developed in standard programming languages extended with specific keywords and syntaxes → CUDA C/C++
  - flexible functions to be called in different contexts for enabling basic sensorimotor skills

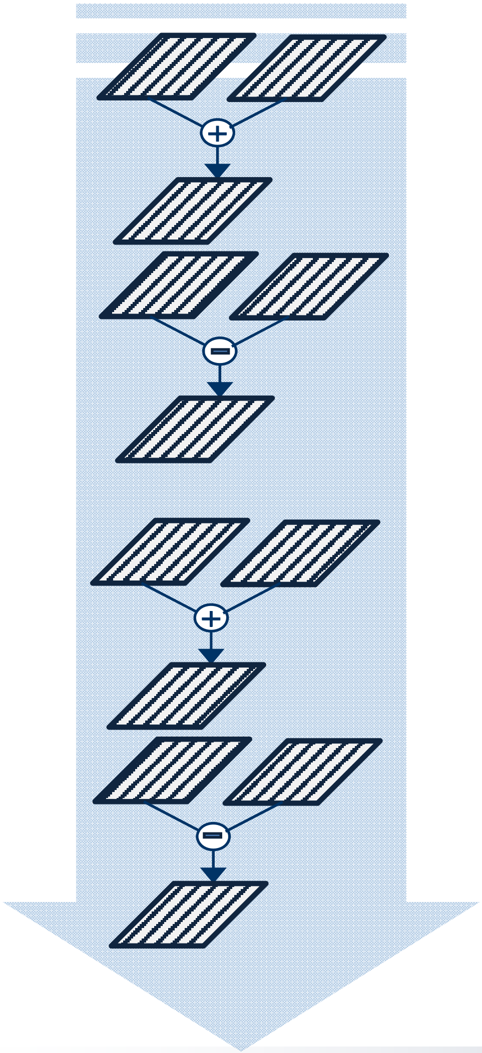
[M. Chessa, V. Bianchi, M. Zampetti, S. P. Sabatini, F. Solari (2012) *Real-time simulation of large-scale neural architectures for visual features computation based on GPU.* Network: Computation in Neural Systems 23(4), pp. 272-291.]

[M. Chessa and G. Pasquale (2013) *Graphics processing unit-accelerated techniques for bio-inspired computation in the primary visual cortex.* Concurrency and Computation: Practice and Experience, DOI: 10.1002/cpe.3118]

Progress toward a *Neuroware*

Comparing different implementation strategies

1. Data **parallelism**…

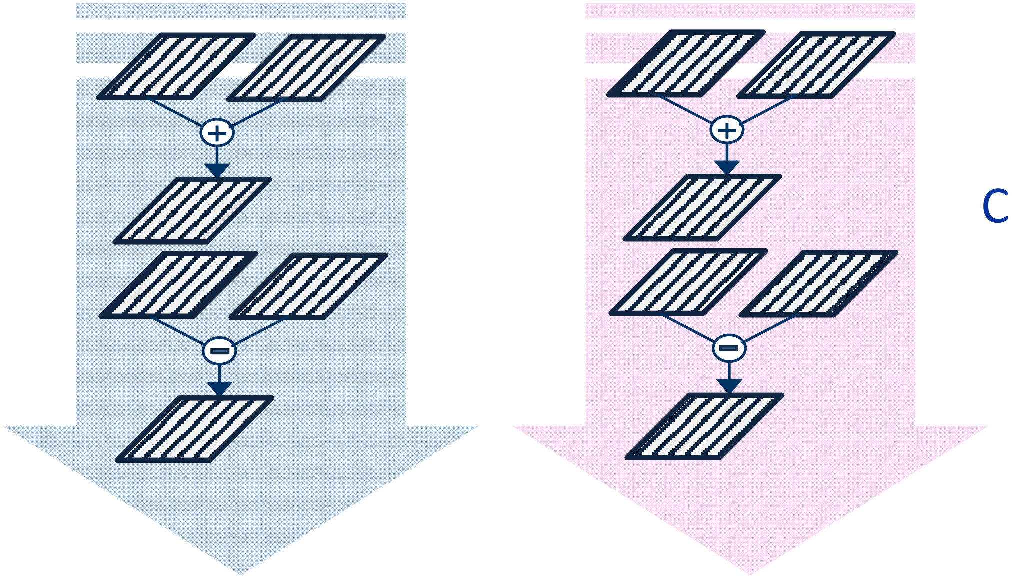


CUDA **kernels**…

Progress toward a *Neuroware*

Comparing different implementation strategies

1. **Data parallelism**… and **task parallelism**

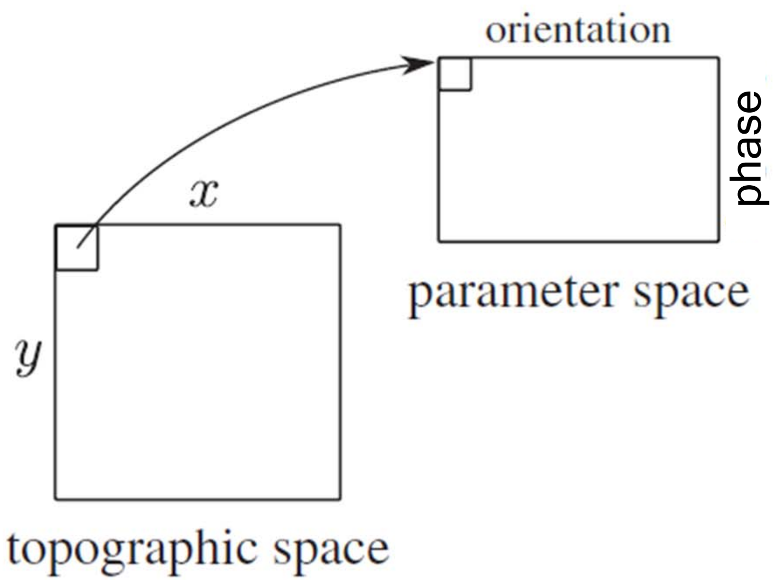


CUDA **kernels**… and CUDA **streams**

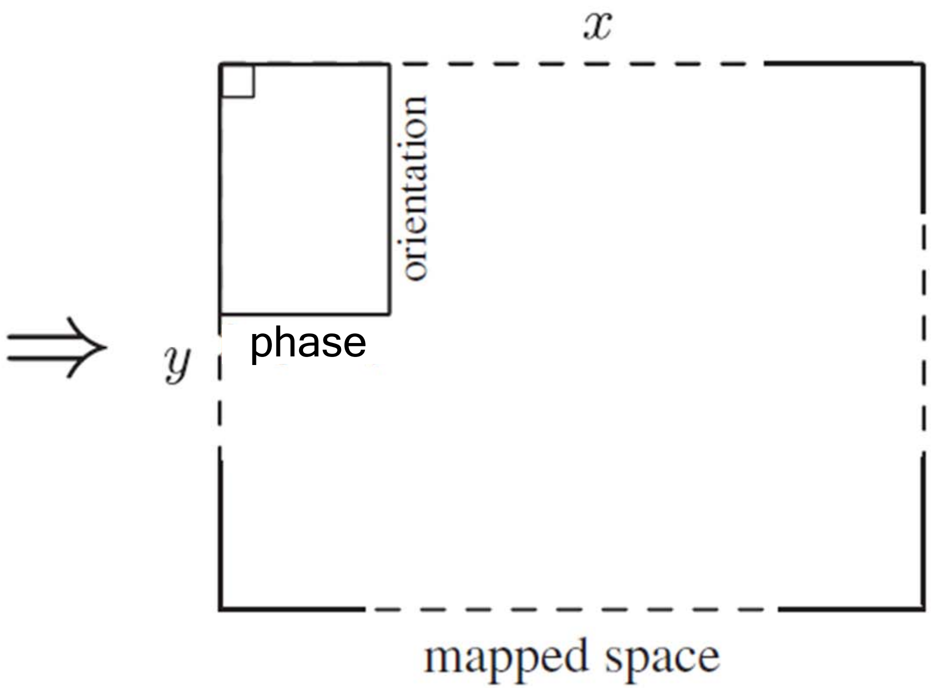
2. **OpenCV / CUDA**
   - C++ using OpenCV's interface to CUDA or OpenCV's processing primitives
   - CUDA C/C++ using CUDA runtime APIs or NVDIA performance primitives
3. "grouped" *vs.* "ungrouped" **data structures**
   - All response matrices in separate memory locations
   - A unique matrix for responses of cells with equal phase-shift
   - Left and right responses replicated in equal matrices

Progress toward a *Neuroware*

Different data structures



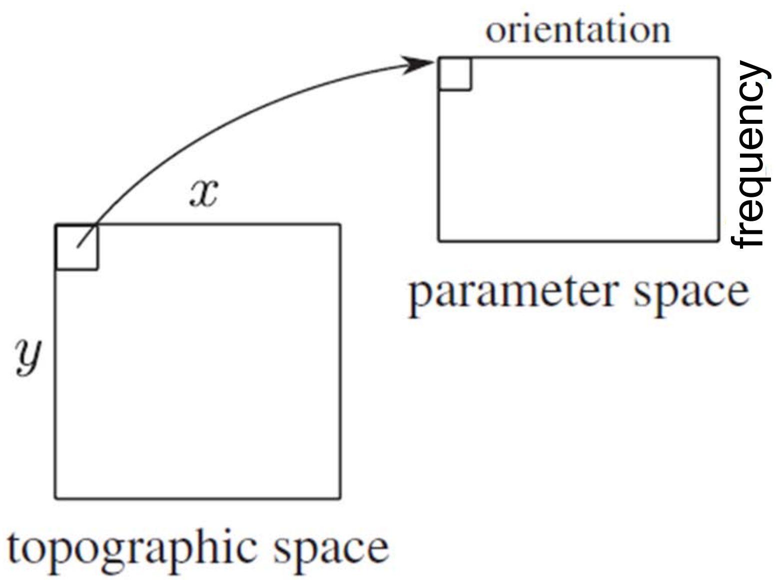
Array of pointers



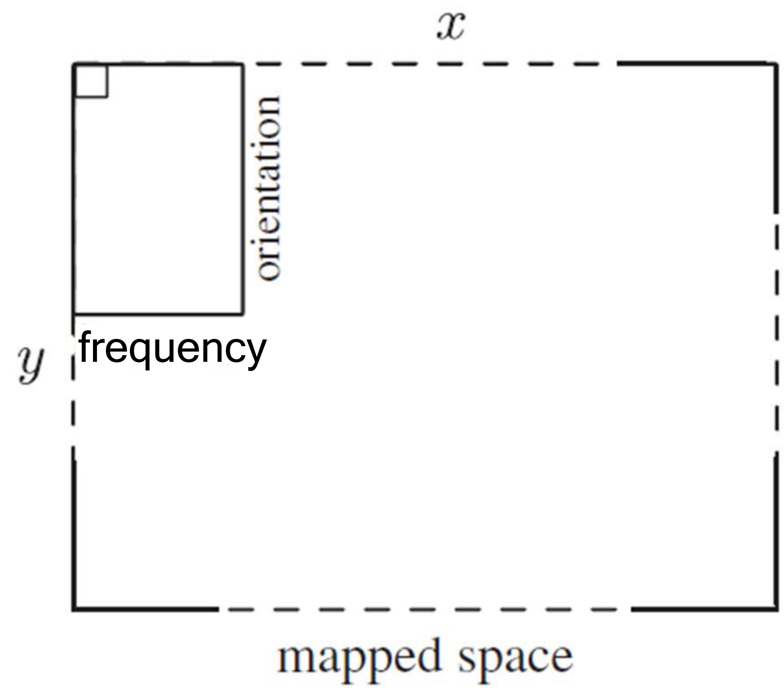
2D array

Progress toward a *Neuroware*

Different data structures



Array of pointers

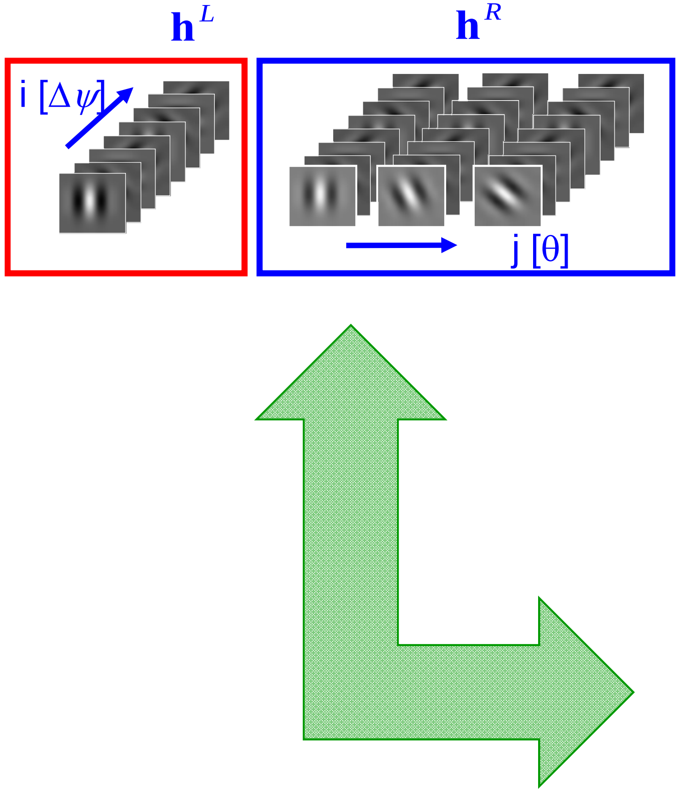


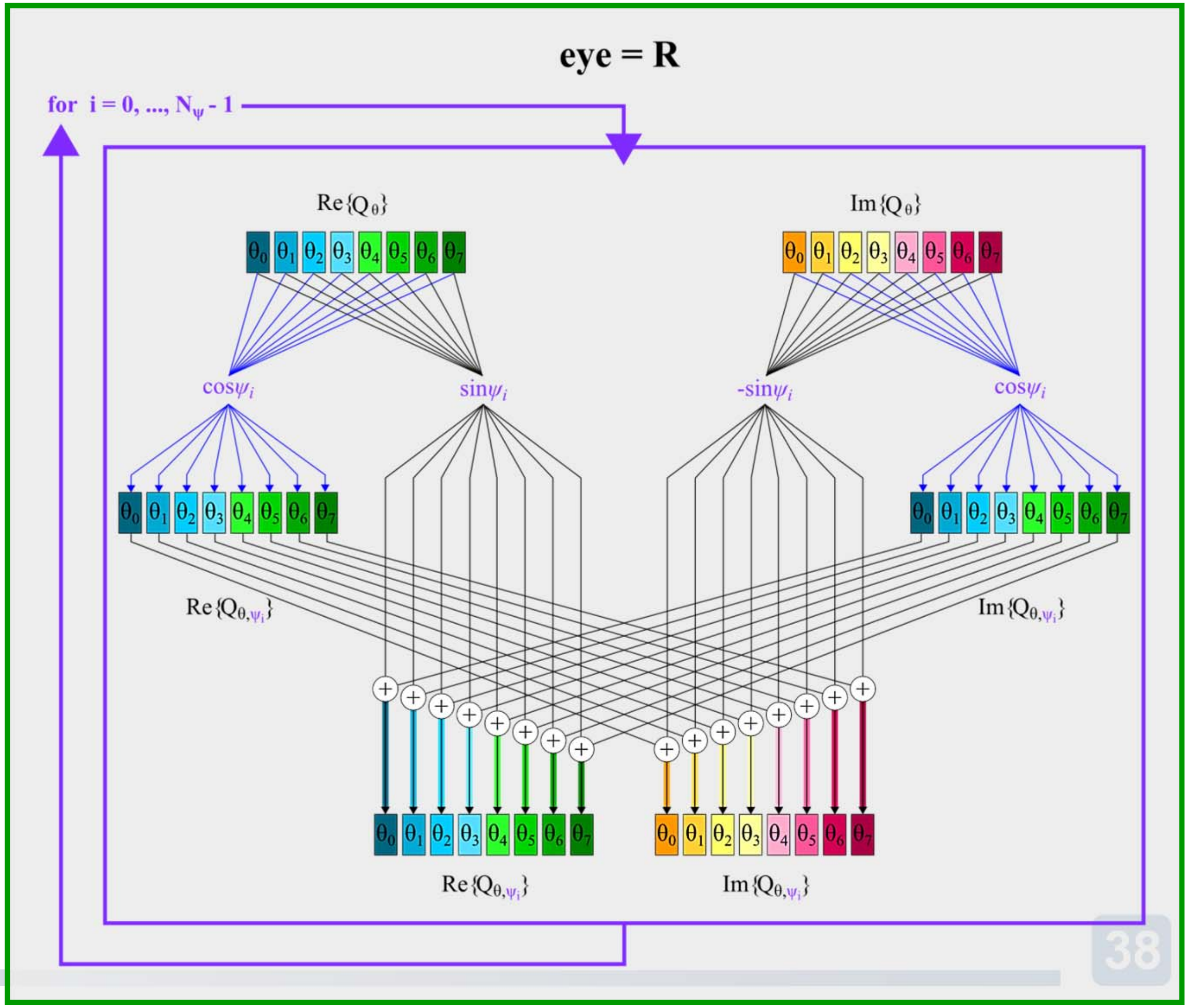
2D array

Progress toward a *Neuroware*

Examples

shiftSimpleResponses_noGroup

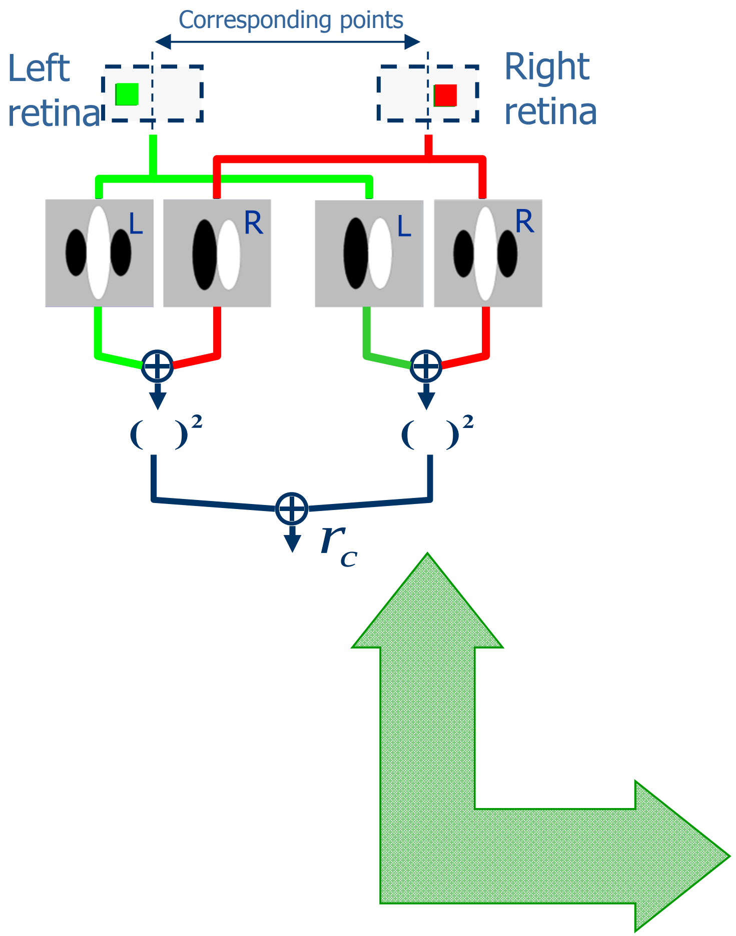


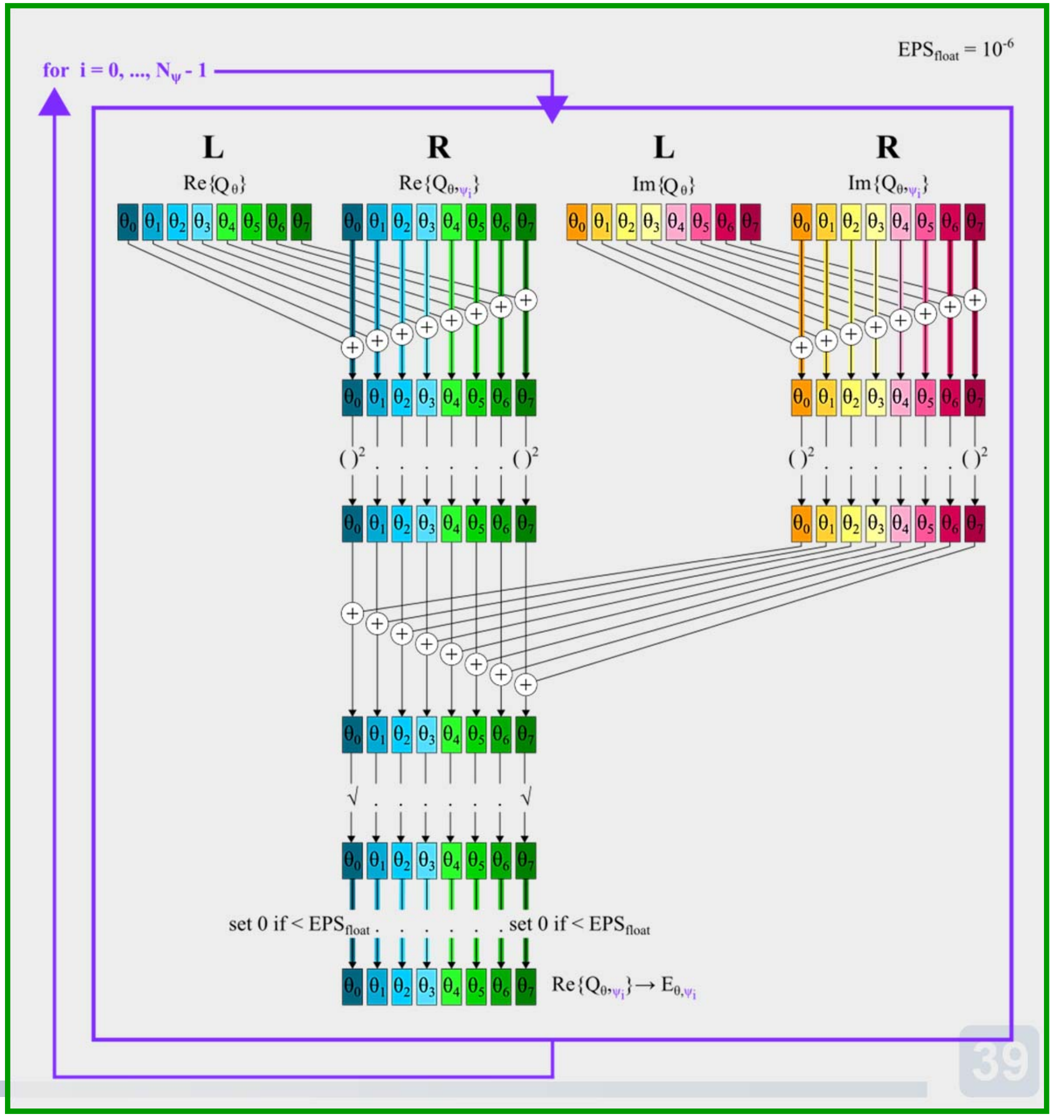


Progress toward a *Neuroware*

Examples

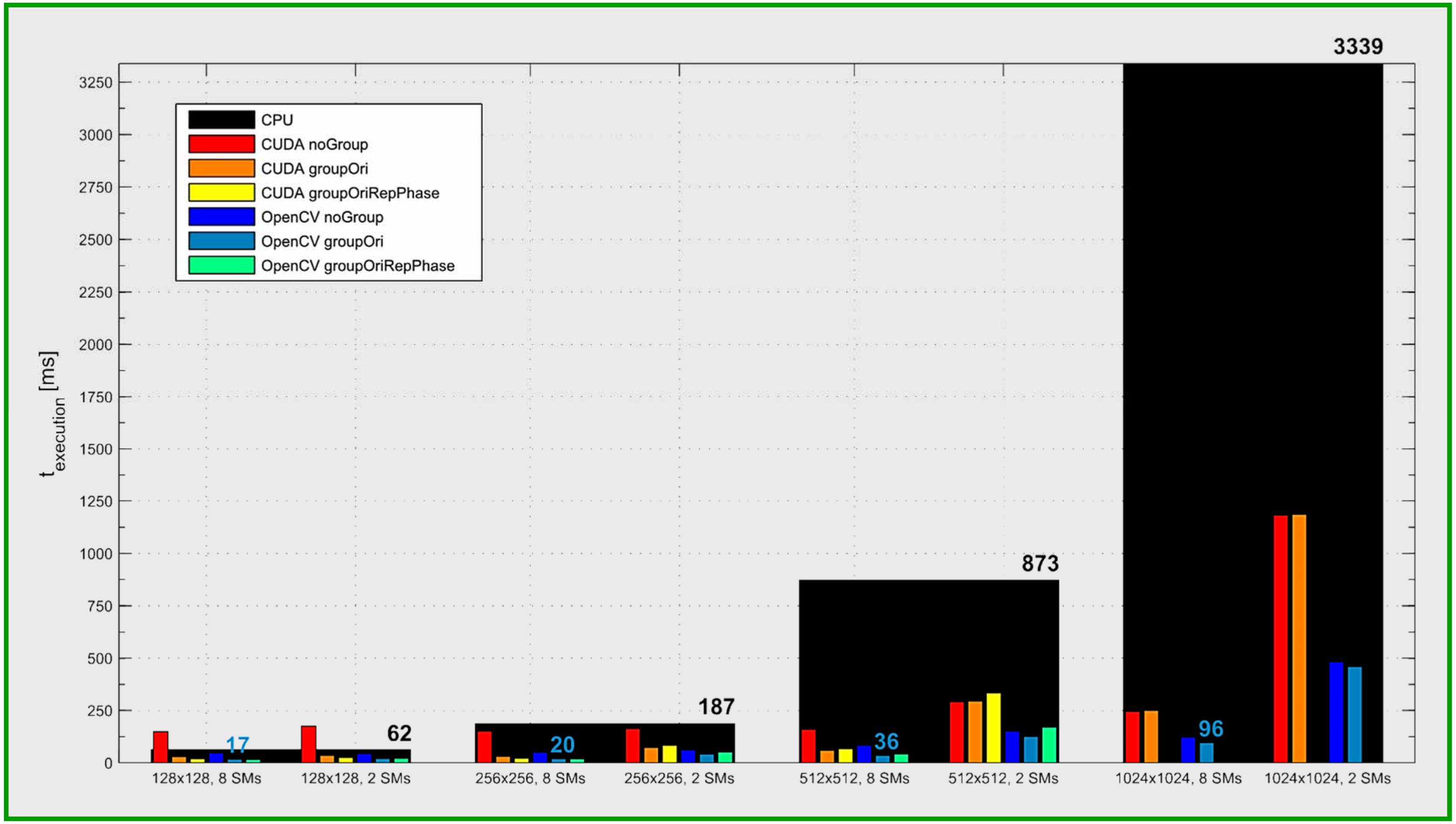
calcEnergy_noGroup





Progress toward a *Neuroware*

Performance evaluation for full disparity estim.



$\frac{62}{17} \approx 3.7\times$ $\frac{187}{20} \approx 9.4\times$ $\frac{873}{36} \approx 24.3\times$ $\frac{3339}{96} \approx 34.8\times$

Conclusions

Take-home messages

- Alternative to feature extraction
  - Deriving features from spatio-temporal properties of the visual *signal* in the harmonic domain
- Alternative to measures
  - Distributed coding through populations of cells tuned to space-time phase relationships
    - Increased flexibility
    - Improved resistance to noise
    - Crucial to avoid sequentialization of sensor and motor processes
- Different modes of specializations through parallel hierarchies
- Efficient implementation on modern graphic cards

The Group

Fabio Solari

Manuela Chessa

Andrea Canessa

Agostino Gibaldi

Contact: **silvio.sabatini@unige.it**

Acknowledgements:

European FP7 Project EYESHOTS – “Heterogeneous 3D perception across visual fragments” – **www.eyeshots.it**

EYESHOTS
Heterogeneous 3D Perception across Visual Fragments