Deep Representation Hierarchies for 3D Active Vision

Silvio P. Sabatini
Department of Informatics, Bioengineering, Robotics and Systems
University of Genoa

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 and 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
... 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 learn good intermediate representations that can be *shared* across tasks.

[Adapted from Bengio, 2009]
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
Cortical architecture

Hierarchical processing of depth

What is where?

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

“Where”

Spatial relationships

Object recognition

“What”
Building distributed representations of the binocular visual signal
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

S-cells

C-cells

read-out$_1$

read-out$_2$
Complex cells “pool” the output of simple cells within a retinotopic neighborhood.

Energy models:

\[ L \rightarrow (\cdot)^2 \rightarrow \bigoplus \]

\[ Lq \rightarrow (\cdot)^2 \rightarrow \bigoplus \]

S-cells

C-cell

Simple cells

Complex cell

Linear filter banks

Non-linear (NL)

Feature pooling
Phase-based measures ...

- Contrast discontinuities $\rightarrow$ phase *congruency*
- Visual motion $\rightarrow$ phase *constancy*
- Binocular disparity $\rightarrow$ phase *difference*

*vs.*

... energy coding

- Contrast energy $\rightarrow$ [Morrone & Burr, 1982, 1988]
- Motion energy $\rightarrow$ [Adelson & Bergen, 1985]
- Binocular energy $\rightarrow$ [Ohzawa et al., 1990]
Binocular energy unit

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

\[ h^L(x; k_0, \psi_L) = e^{-\frac{x^2}{\sigma^2}} e^{i(k_0 x + \psi_L)} \]

\[ h^R(x; k_0, \psi_R) = e^{-\frac{x^2}{\sigma^2}} e^{i(k_0 x + \psi_R)} \]

\[ Q^{L/R}(x) = h^{L/R} \ast I^{L/R}(x)e^{-j\psi_{L/R}} \]

where \[ \Delta \psi = \psi_R - \psi_L \]

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

[Qian, 1994][Fleet et al., 1996]
The binocular energy unit maximally responds when $\Delta \psi$ matches the image phase disparity $\Delta \phi$.
The binocular energy unit maximally responds when $\Delta \psi$ matches the image phase disparity $\Delta \phi$. 

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

Disparity tuning curve

Stimulus disparity $[D_{\text{max}}]$
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

\[ \delta_{\text{pref}} \propto \frac{\Delta \psi}{k_0} \]

Large scale cortical architectures

2×56 binocular receptive fields for each pixel

\[ h^L \quad h^R \]

\[ \begin{array}{cccccccc}
0 & \pi/8 & \pi/4 & 3\pi/8 & \pi/2 & 5\pi/8 & 3\pi/4 & 7\pi/8 \\
-3\pi/4 & -\pi/2 & -\pi/4 & 0 & \pi/4 & \pi/2 & 3\pi/4 & -3\pi/4 \\
\end{array} \]

Cell response

Disparity tuning surface

\[ \delta_{ij_{\text{pref}}} \propto k^j_0 \frac{\Delta \psi^i}{\|k^j_0\|^2} \]

Enabling disparity-vergence responses in stereo-heads
**Disparity estimation**

\[ \Delta = \text{maximum detectable disparity along the direction orthogonal to the cell’s orientation, equals one half cycle of the peak spatial frequency of the RF} \]

**Direct vergence control**

Assuming \( VD \approx 0 \), the orientation is used to extend the sensitivity range of the cells’ population to HD stimuli.
3D active vision requires ≠ specializations

Disparity estimation

\[ \delta_{est}(x) = \arg \min_{\delta(x)} \sum_{\theta} \left( \delta_{\theta}(x) - \frac{k_0^T}{k_0} \delta(x) \right) \]

Direct vergence control

\[ v_S(x) = \sum_{x \in \Omega} G(x) \sum_i w_i r_c^i(x) \]
**Learning Algorithms**

**Reinforcement learning**, based on particle swarm optimization algorithm


**Supervised learning**, based on LeNet non-linear convolutional network


**Differential Hebbian Rule:**

\[
\begin{align*}
\Delta w_i^t &= (1 - \eta) w_i^{t-1} + \eta V_S(r_c^{t-1}) \Delta r_c^i \\
\Delta r_c^i : \text{Differential population response} \\
V_S(r_c^{t-1}) : \text{Vergence signal at instant } t - 1 \\
\eta : \text{Learning rate} = \Delta \text{STD of the population response}
\end{align*}
\]

\( \Rightarrow \text{INTRINSIC REWARD!} \)
Results

@trial 50

weight distribution

$\delta_V$

$\delta_H$

Time steps

$\theta$

$\Delta \psi$

$\Delta \theta$
Results

@trial 200

weight distribution

$\theta$

$\Delta \psi$

$\delta_V$

$\delta_H$

Time steps

@trial 200

$\delta_V$

$\delta_H$
Simulation Results

moving stimuli

Model

Real data

[Hung, 1997]

Simulation Results

moving stimuli

Model

Real data

[Hung, 1997]

Experimental Results

moving stimuli

iCub

Real data


[Hung, 1997]
Experimental results

Switching fixations among static visual targets

Stepping and waving objects

Videos: http://www.eyeshots.it/res_news.php
Combined control of horizontal and vertical vergence

VC_H

Vertical Disparity $\delta_V$

VC_V

Horizontal Disparity $\delta_H$
Measured vergence working ranges

Helmholtz (=Tilt-Pan) system

[Gibaldi et al., submitted]
Measured vergence working ranges

Fick (=Pan-Tilt) system

[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


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

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

\[ \Delta \psi \]

\[ j \theta \]

\[ h^L \]

\[ h^R \]

\[ i \]

\[ h \]

\[ R \]

for \( i = 0, \ldots, N_\psi - 1 \)

\[ \text{Re}\{Q_{0,0}\} \]

\[ \text{Im}\{Q_{0,0}\} \]

\[ \text{Re}\{Q_{0,1}\} \]

\[ \text{Im}\{Q_{0,1}\} \]

\[ \text{Re}\{Q_{0,2}\} \]

\[ \text{Im}\{Q_{0,2}\} \]

\[ \text{Re}\{Q_{0,3}\} \]

\[ \text{Im}\{Q_{0,3}\} \]

\[ \text{Re}\{Q_{0,4}\} \]

\[ \text{Im}\{Q_{0,4}\} \]

\[ \text{Re}\{Q_{0,5}\} \]

\[ \text{Im}\{Q_{0,5}\} \]
Progress toward a Neuroware

Examples

calcEnergy_noGroup

Left retina

Right retina

Corresponding points

\( r_c \)

for \( i = 0, \ldots, N_v - 1 \)

\( \text{EPS}_{\text{stat}} = 10^{-6} \)

\( \text{Re}(Q_{\text{corr}}) \)

\( \text{Re}(Q_{\text{corr}}) \)

\( \text{Im}(Q_{\text{corr}}) \)

\( \text{Im}(Q_{\text{corr}}) \)

set 0 if \( \text{EPS}_{\text{stat}} \)

set 0 if \( \text{EPS}_{\text{stat}} \)

\( \text{Re}(Q_{\text{corr}}) \rightarrow E_{\text{corr}} \)
Progress toward a Neuroware

Performance evaluation for full disparity estim.

\[
\begin{align*}
\frac{62}{17} &\approx 3.7 \times \\
\frac{187}{20} &\approx 9.4 \times \\
\frac{873}{36} &\approx 24.3 \times \\
\frac{3339}{96} &\approx 34.8 \times
\end{align*}
\]
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