Toward predictive machine learning for active vision

Emmanuel Daucé

Institut de Neurosciences des Systèmes (UMR S1106), Marseille, France.

emmanuel.dauce@univ-amu.fr

Inserm

Institut de Neurosciences des Systèmes

Aix*Marseille université

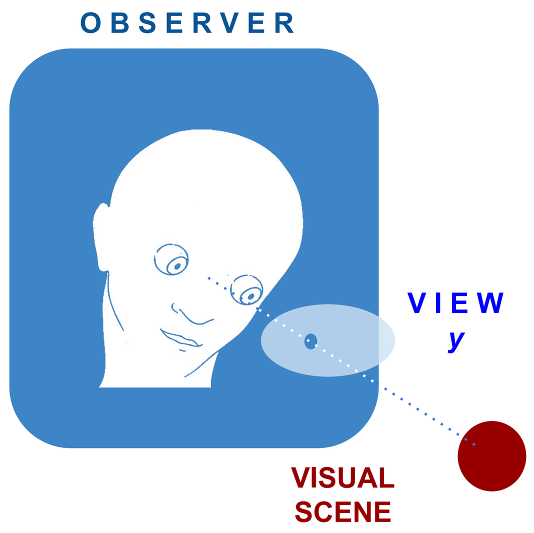
Theoretical Neuroscience Group

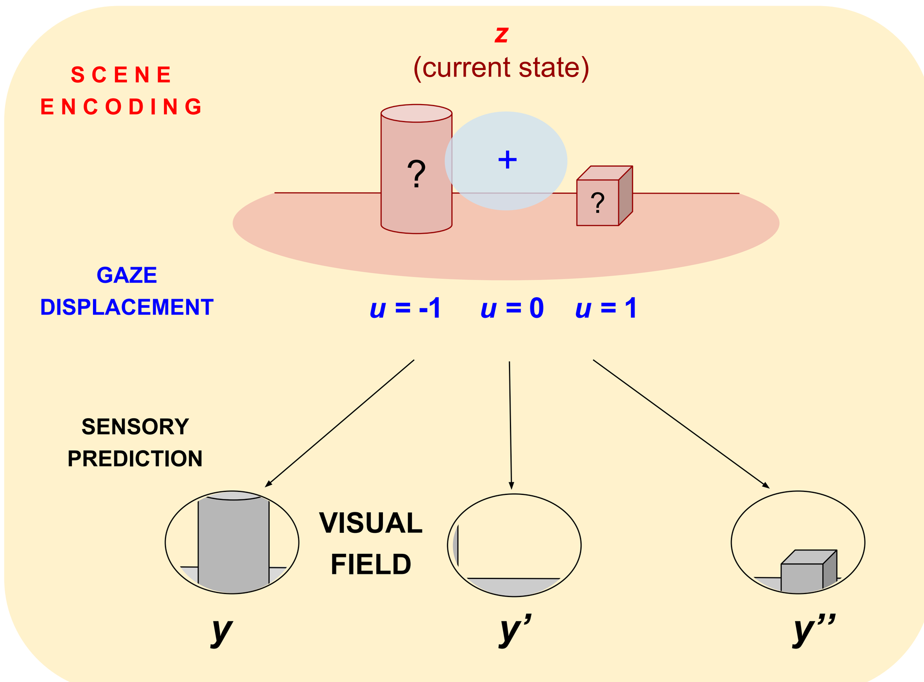
Active perception was proposed to rely on sensory prediction through a generative model (Friston et al, 2012). A "three-party" generative framework based on three mutually independent domains (i.e object-in-space, gaze orientation and visual field) is proposed here. An control policy devoted to the recognition of objects in a scene through a foveated sensor is developed. It is shown efficient on a digit recognition database, providing biologically-realistic sequence of saccades and state-of-the-art recognition rates.

CONTEXT

Active Vision

- **The active sensing** framework primarily relies on emitting a signal to sense the environment, as is typically done by a radar or echolocation.
- **Active vision** (or active perception in general)
  - rests on a multi-view processing of a scene
  - generalizes to the concept of ***action-for-perception*** where a sensing device is moved around to increase its range and/or its resolution.
- **Bayesian estimation:** ``*The problem of Active Sensing can be stated as a problem of controlling strategies applied to the data acquisition process which will depend on the current state of the data interpretation and the goal or the task of the process.*" **(Bajcsy, 1988)**
- **Predictive framework** :
  - A general setup proposed by Friston and colleagues **(Friston & Kiebel, 2009)**
  - Based on signal filtering theory **(Kalman, 1960)**
  - Interpreted as a general tendency of the brain to **counteract surprising sensory events through action.**





Gaze orientation

- **Visuo-motor control**
- Superior vertebrates visual apparatus relies on:
  - a **foveated** retina
  - that concentrates light photoreceptors over a small central portion of the visual field
- Scene scanning through **saccades** (Yarbus, 1967)
  - High-speed targeted eye movements
  - Sequential scene exploration

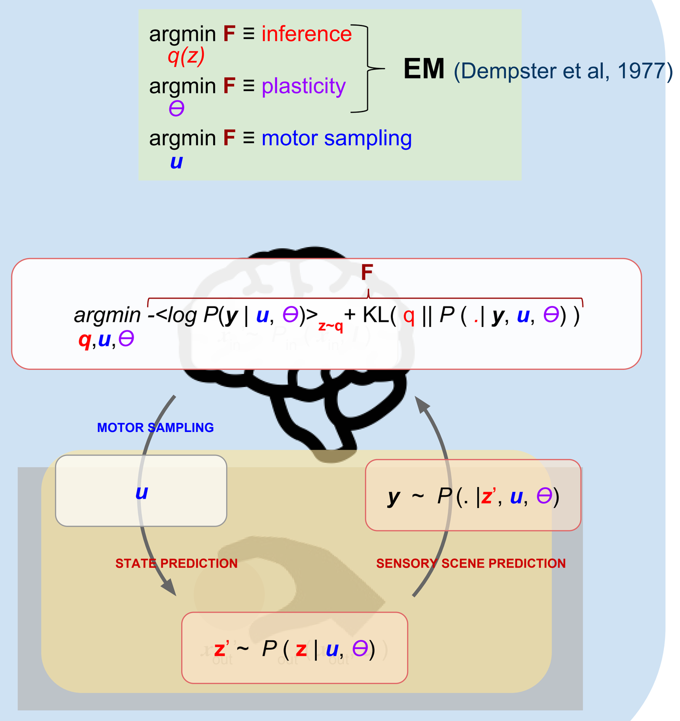
MODELS

Active inference

(Friston & Kiebel, 2009)

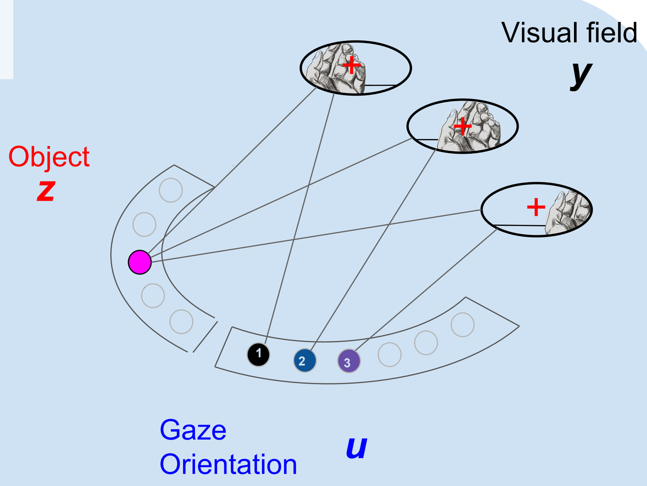
- The brain builds a generative model **P**
- so as to improve its predictions over time.
- This improvement is done :

- through sampling the environment (u)
- and extracting statistical invariants (z)
- that are used in return to predict upcoming events (y).



Three-party generative model

- Many **views** y_u's on the same **scene** :
  - scene $\mathbf{Y} = \{\mathbf{y}_u\}_{u \in U}$
  - independance assumption : $P(\mathbf{Y}) = \prod_u P(\mathbf{y}_u)$
- Latent space = scene encoding :
  $\mathbf{z} = (\mathbf{o}, \mathbf{x})$
  - $\mathbf{o}$ is an object
  - $\mathbf{x}$ is the object coordinates in the peripersonal space
- End-effector control :
  - $\mathbf{u}$ (motor command) is the absolute orientation of the visual sensor



- Steady state assumption : $\dot{\mathbf{z}} = 0$ (static scene)
- Model-based approach :
  - Generative model : $P(\mathbf{y}, \mathbf{z}, \mathbf{u})$
  - object-effector independence assumption : $P(\mathbf{z}|\mathbf{u}) = P(\mathbf{z})$

Sequential scene exploration

Sequential estimate of $q(\mathbf{z}) \approxeq P(\mathbf{z}|\mathbf{Y})$

Objective function : $H = -\langle \log q(\mathbf{z}) \rangle_{z \sim q}$

(Friston et al, 2012)

<u>data</u> : $\mathbf{Y} = \{\mathbf{y}_u\}_{u \in U}$

<u>initiate</u> :

$\forall \mathbf{z}, q(\mathbf{z}) = P(\mathbf{z})$ -- prior

<u>while</u> $H(q) > H_{ref}$:

predict $\mathbf{z} \sim q$

$\forall \mathbf{u} \in U$, predict $\mathbf{y}_u \sim P(\mathbf{y}|\mathbf{z},\mathbf{u})$

choose $\hat{\mathbf{u}} = \mathrm{argmax}_{u \in U} P(\mathbf{z}|\mathbf{y}_u,\mathbf{u})$

read $\mathbf{y}_{\hat{u}}$

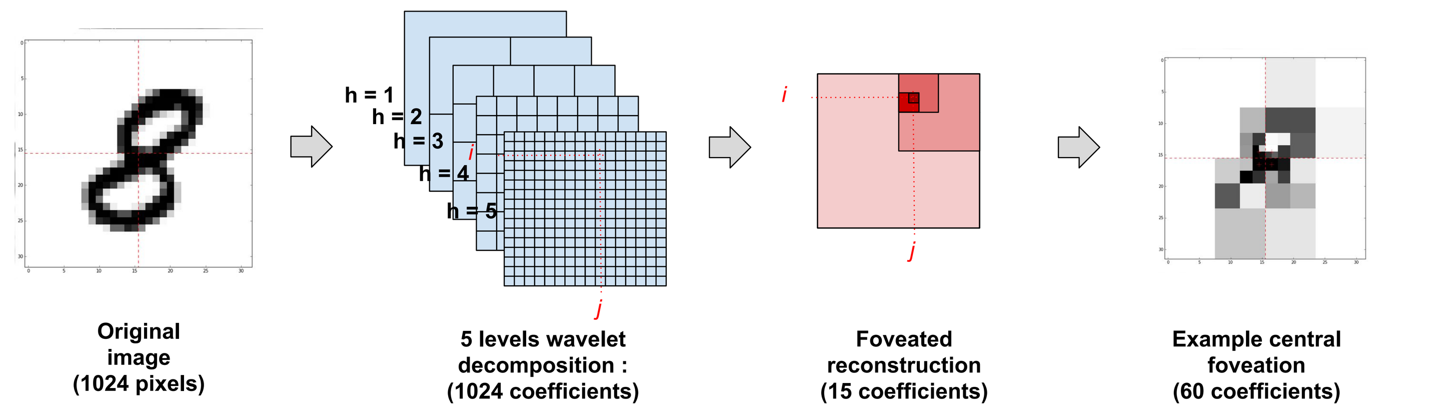
update $\forall \mathbf{z}, q(\mathbf{z}) \leftarrow P(\mathbf{z}|\mathbf{y}_{\hat{u}},\hat{\mathbf{u}})$ -- posterior

-- becomes prior

$U \leftarrow U \setminus \{\hat{\mathbf{u}}\}$

<u>return</u> q

EXAMPLE : MNIST DATASET



- Handwritten digit recognition
- $z \in \{0,...,9\}$
- 28×28 b/w images
- 1 image = 1 visual scene
- Foveated vision from 2D Haar wavelet decomposition :
  - 5 levels
  - image patches :

$u = (h,i,j) \rightarrow \mathbf{y}_u = (y_u^{1}, y_u^{2}, y_u^{3})$

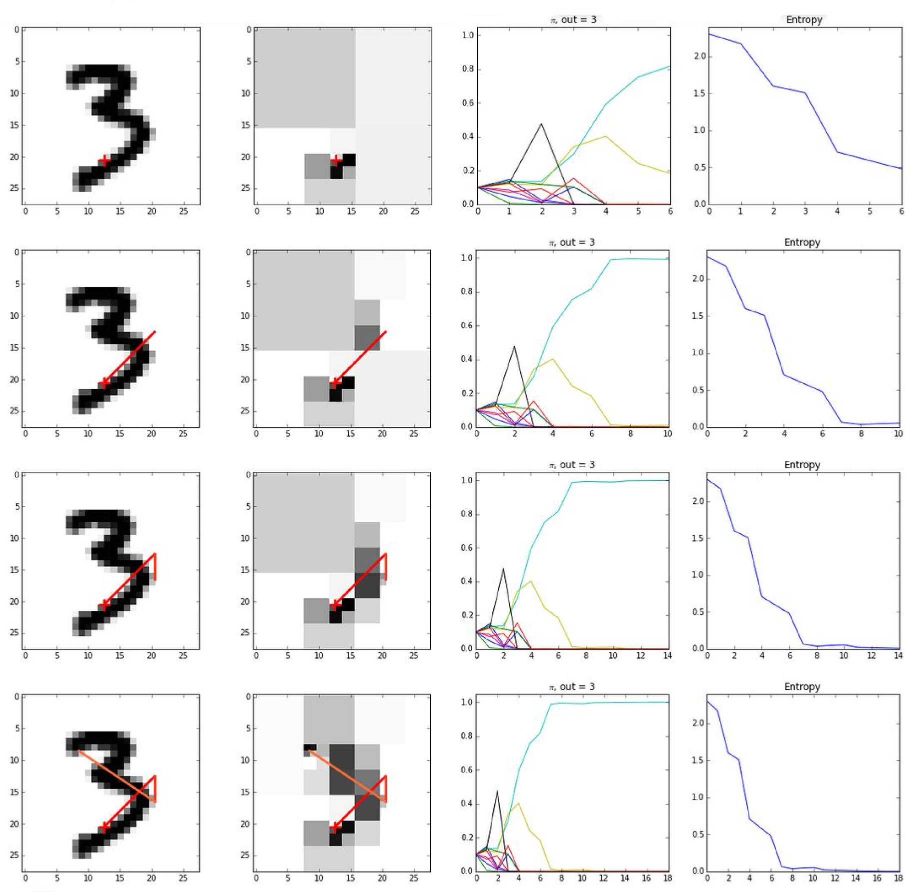
- Generative model :

$\mathbf{y}|z, u \sim \mathcal{B}(\rho_{z,u}) \times \mathcal{N}(\mu_{z,u}, \Sigma_{z,u})$

$\{\rho_{z,u}, \mu_{z,u}, \Sigma_{z,u}\}_{z,u}$ learned on 55.000 examples

- Inference :

$\log q(z) = \log p(z|\mathbf{y}_{1:T},\mathbf{u}_{1:T}) \propto \sum_t \log p(z|\mathbf{y}_t,\mathbf{u}_t)$

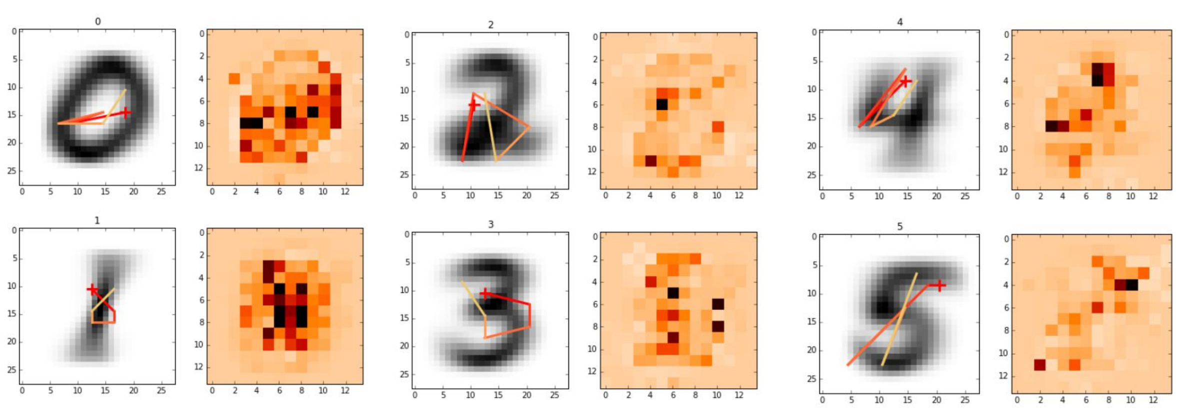


Salience maps

- $\forall z : \forall (i,j)$:

$$\log q(z|i,j) \propto \sum_h \log P(\mu_{z,(h,i,j)}|z,(h,i,j)) \underbrace{- \log \sum_{z'} \prod_{h'} P(\mu_{z',(h',i,j)}|z',(h,i,j))}_{K_{i,j}}$$

- Saccade prototypes :
  - $(i_1,j_1) \rightarrow (i_2,j_2) \rightarrow \ldots \rightarrow (i_T,j_T)$
  - with : $q(z|i_1,j_1) > q(z|i_2,j_2) > \ldots > q(z|i_T,j_T)$
- pre-processed saccades ≃ Table look-up



OUTLOOK & FUTURE WORK

Results

- Generative Model
- End-effector control
- Foveated vision
- Rests on **predictive posterior entropy minimization** ("salience" maximization)
- Visual scan concentrates on class-critical regions
- Reduced visual bitrate (with variable #saccades depending on task difficulty)
- Admits fast table look-up implementation
- The recognition rate depends on a recognition threshold H_{ref}
- Up to 92 % correct recognition ($H_{ref} = 10^{-4}$)

Perspectives

- Controller learning :
  - Salience maps : $M(\mathbf{z},i,j) = -\log q(\mathbf{z}|i,j)$
    1. can be learned by exploration
    2. through an existing generative model P
  - Bandit exploration (Exp4, UCB, ..)?
- Combinatorial predictive models :
  - Scene encoding $\mathbf{z}$ = (object $\mathbf{o}$, position $\mathbf{x}$)
  - What/where pathways
  - Build-up visual field prediction?
    1. Estimate $P(\mathbf{x}|\mathbf{Y}_0)$ -- object position prediction
    2. Orientate sight toward $\mathbf{x}$
    3. Read $\mathbf{Y}_1$
    4. Estimate $P(\mathbf{o}|\mathbf{Y}_1)$ -- object identity prediction
    5. Explore locally through saccades if needed

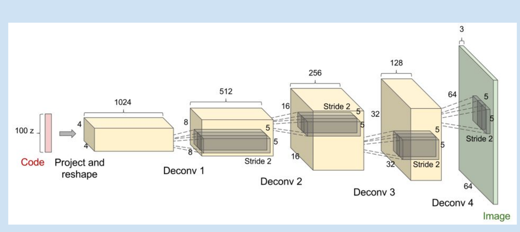
- Generative model learning : $\Theta = \{\rho_{z,u}, \mu_{z,u}, \Sigma_{z,u}\}_{z,u}$
  - Gradient descent over generative model parameters :

$\Delta\Theta \propto -\nabla_\Theta H(q_T) \propto \langle \sum_t \nabla_\Theta \log P(\mathbf{y}_t|\mathbf{u}_t; \Theta) \rangle_{z \sim q}$

  - Online and Reinforcement learning :
    - Final object estimate : $z_T \sim q_T$
    - Category learning $r(z_T, z^*) = \delta(z_T, z^*) - b$

$\Delta\Theta \propto \langle - r(z, z^*) \sum_t \nabla_\Theta \log P(\mathbf{y}_t|\mathbf{u}_t; \Theta) \rangle_{z \sim q}$

- Possible extension to convolutional architectures :



from Radford et al (2015)

Bibliography

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE*, *76*(8), 966-1005.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1521), 1211-1221.

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in psychology*, *3*.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, *82*(1), 35-45.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In *Eye movements and vision* (pp. 171-211). Springer US.

INTRINSICALLY MOTIVATED OPEN-ENDED LEARNING_

IMOL 2017 The third international workshop on Intrinsically Motivated Open-ended Learning
Rome, Italy, October 04-06, 2017