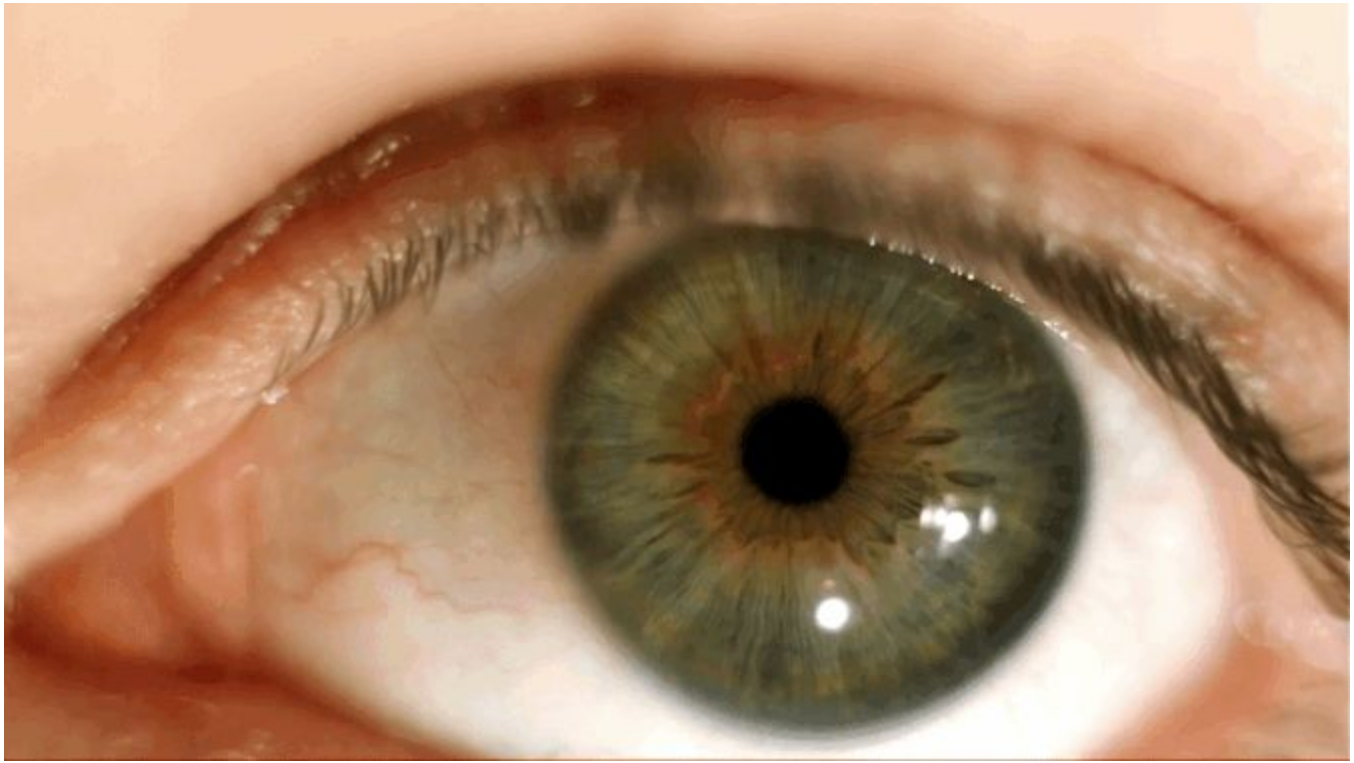
Sailab seminars

Towards Laws of Visual Attention

Dario Zanca :: Marco Gori :: Stefano Melacci

March 20th, 2019



Eymol

Eymol

D. Zanca and M. Gori. “Variational laws of visual attention for dynamic scenes”. NIPS 2017, p. 3823-3832.

The emergence of biological mechanisms of visual attention might obey universal and unifying functional principles.

Our intent is to determine laws of attention which obey to by certain Principles of Visual Attention:

1. Boundedness of the trajectory
2. Curiosity (local+peripheral)
3. Brightness invariance

Eymol

D. Zanca and M. Gori. “Variational laws of visual attention for dynamic scenes”. NIPS 2017, p. 3823-3832.

The brightness that hits the retina at each time instant is defined, on each position by

$$b : \mathbb{R}^+ \times \mathbb{R}^2 \to \mathbb{R}$$

INPUT

The trajectory that draws an attentive scanpath over the image is defined by

$$x : \mathbb{R}^+ \to \mathbb{R}^2$$

OUTPUT

Eymol

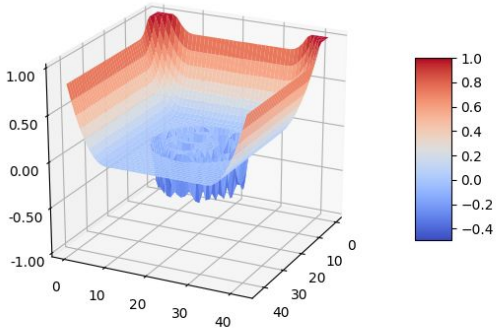
D. Zanca and M. Gori. “Variational laws of visual attention for dynamic scenes”. NIPS 2017, p. 3823-3832.

$$\int_{\mathcal{T}} L(t,x,\dot{x})\,dt = \int_{\mathcal{T}} \frac{1}{2}m\dot{x}^2 - [\underbrace{V(x)}_{retina} - \underbrace{C(t,x)}_{curiosity} + \underbrace{B(t,x,\dot{x})}_{bright.inv.}]\,dt$$

$$V(x) = k \sum_{i=1,2} \left((l_i - x_i)^2 \cdot [x_i > l_i] + (x_i)^2 \cdot [x_i < 0] \right)$$

$$C(t,x) = b_x^2$$

$$B(t,x,\dot{x}) = \left(\frac{db}{dt}\right)^2 = (b_t + b_x\dot{x})^2$$



Eymol

D. Zanca and M. Gori. “Variational laws of visual attention for dynamic scenes”. NIPS 2017, p. 3823-3832.

We can determine the motion trajectory by searching for the stationary point of the action functional, which is given by the correspondent Euler-Lagrange equations.

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = \frac{\partial L}{\partial x}$$

That, in our case, are:

$$m\ddot{x} - \lambda \frac{d}{dt} B_{\dot{x}} = -V_x + \eta C_x - B_x$$

Eymol

D. Zanca and M. Gori. “Variational laws of visual attention for dynamic scenes”. NIPS 2017, p. 3823-3832.

Top 10 scores on CAT2000, the biggest dataset of eye-fixations data, sorted by NSS metric.

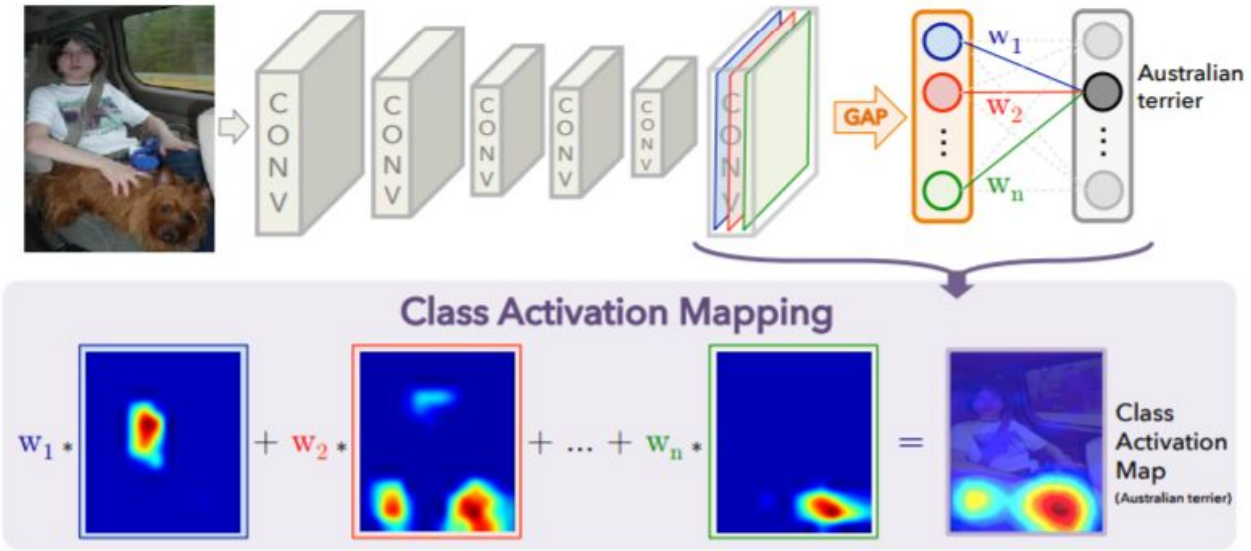
Our model is indicated with EYMOL (EYe MOvement Laws)

	AUC	SIM	EMD	CC	NSS
SAM	0.88	0.77	1.06	0.90	2.40
DeepFix	0.87	0.74	1.15	0.87	2.28
MixNet	0.86	0.66	1.63	0.76	1.92
EYMOL	0.83	0.61	1.91	0.72	1.78
BMS	0.85	0.61	1.95	0.67	1.67
iSEEL	0.84	0.62	1.78	0.66	1.67
{Perm. control}	0.80	0.55	2.25	0.63	1.63
FES	0.82	0.57	2.24	0.64	1.61
Aboudib	0.81	0.58	2.10	0.64	1.57
{One human}	0.76	0.43	2.51	0.56	1.54

CF-Eymol

CF-Eymol

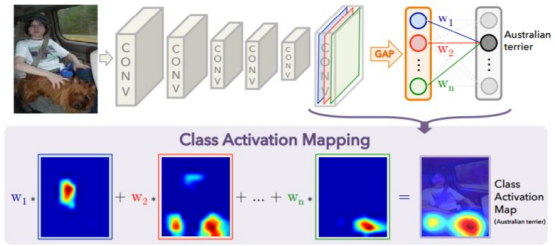
D. Zanca, M. Gori and A. Rufa,"A Unified Computational Framework for Visual Attention Dynamics", Progress in Brain Research, vol. 248, 2018.

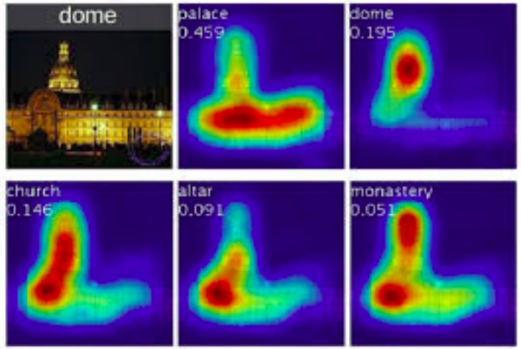


http://cnnlocalization.csail.mit.edu/

CF-Eymol

D. Zanca, M. Gori and A. Rufa,"A Unified Computational Framework for Visual Attention Dynamics", Progress in Brain Research, vol. 248, 2018.





For a given input image, let $f_k(x)$ represents the activation of unit k in the last convolutional layer "pool" at spatial location $x = (x_1, x_2)$. Then, we can indicate the result of global average pooling as

$$F_k = \sum_x f_k(x).$$

For each class c of the dataset, the input of the softmax is

$$\sum_k w_k^c F_k,$$

Class-specific activation map

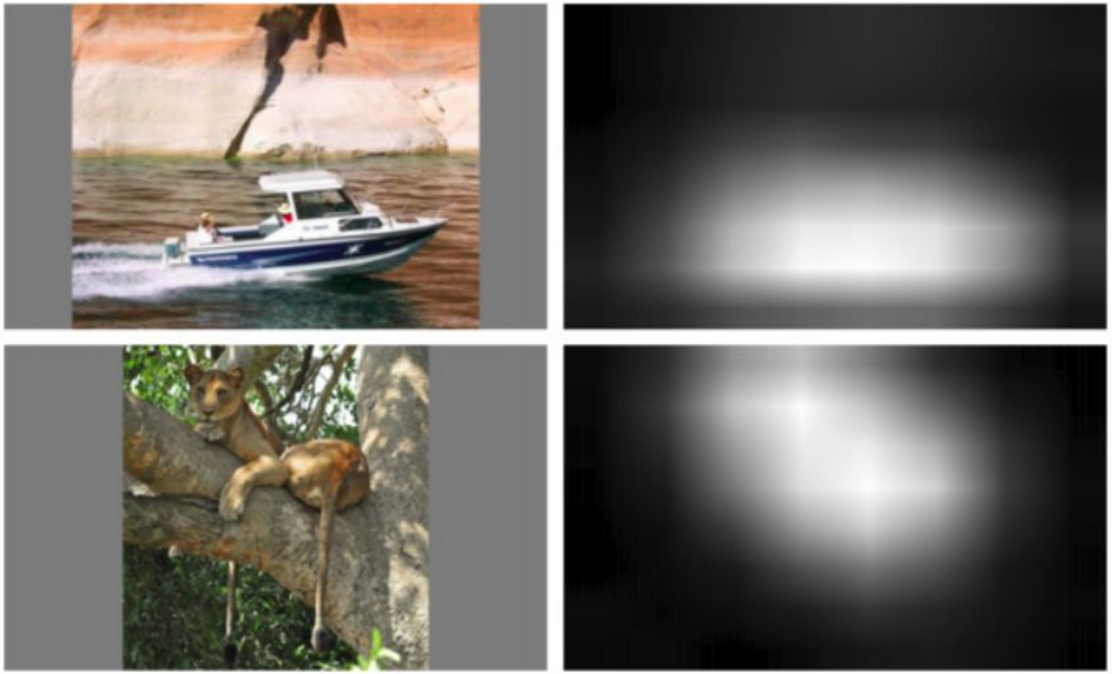
$$M_c(x) = \sum_k w_k^c f_k(x).$$

Convolutional Features (CF) activation map

$$M(x) = \frac{1}{k} \sum_k f_k(x)$$

CF-Eymol

D. Zanca, M. Gori and A. Rufa,"A Unified Computational Framework for Visual Attention Dynamics", Progress in Brain Research, vol. 248, 2018.



(a) Stimulus (b) M

Examples of CF activation map

CF-Eymol

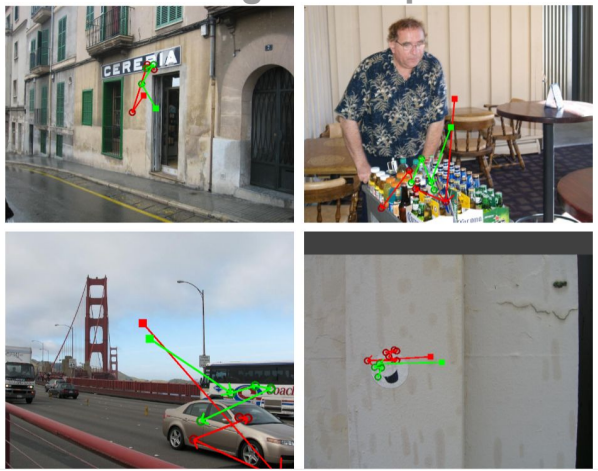
D. Zanca, M. Gori and A. Rufa,"A Unified Computational Framework for Visual Attention Dynamics", Progress in Brain Research, vol. 248, 2018.

$$\int_\tau L(t,x,\dot{x}) = \int_\tau \frac{1}{2}m\dot{x}^2 - [V(x) - C(t,x) + B(t,x,\dot{x}) - M(x)]\ dt$$

Incremental results on saliency prediction

Model version	CAT2000 AUC	CAT2000 NSS
EYMOL	0.838 (0.001)	1.810 (0.014)
CF-EYMOL	**0.843 (0.001)**	**1.822 (0.064)**
Itti-Koch	0.77	1.06
AIM	0.76	0.89
Judd Model	0.84	1.30
AWS	0.76	1.09
eDN	0.85	1.30
DeepFix	0.87	2.28
SAM	**0.88**	**2.38**

More meaningful scanpaths



G-Eymol

G-Eymol

- Features act as masses attracting the focus of attention
- Features are defined outside the motion model and in principle, they can also derive from a convolutional neural network
- The model also includes a dynamic process of inhibition to return.

G-Eymol

Two basic features:

1. Spatial gradient of the brightness

$$\mu_1 = \alpha_1 \|\nabla_x b\|$$

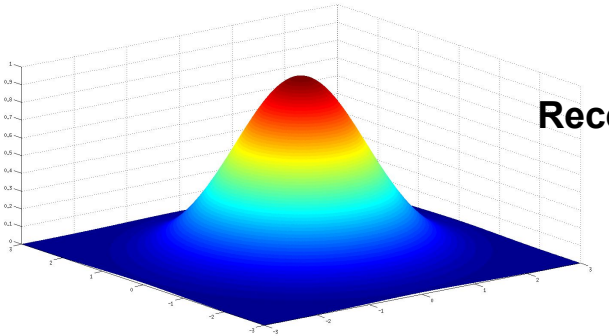
2. Optical flow

$$\mu_2 = \alpha_2 \|v\|$$

G-Eymol

Given any virtual mass μ, that comes from visual features as previously explained, we can construct the overall field by

$$E(a(t)) = -\frac{1}{2\pi}\int_{\mathcal{R}} dx \frac{a(t)-x}{\|a(t)-x\|^2}\mu(x,t).$$



Receptive field

Feature maps
(velocity tag,
spatial gradient)

G-Eymol

We can model the inhibitory function in the same framework as

$$\frac{\partial I(x,t)}{\partial t} + \beta I(x,t) = \beta g(x - a(t))$$

where g(.) is the Gauss function and $0 < \beta < 1$.

G-Eymol

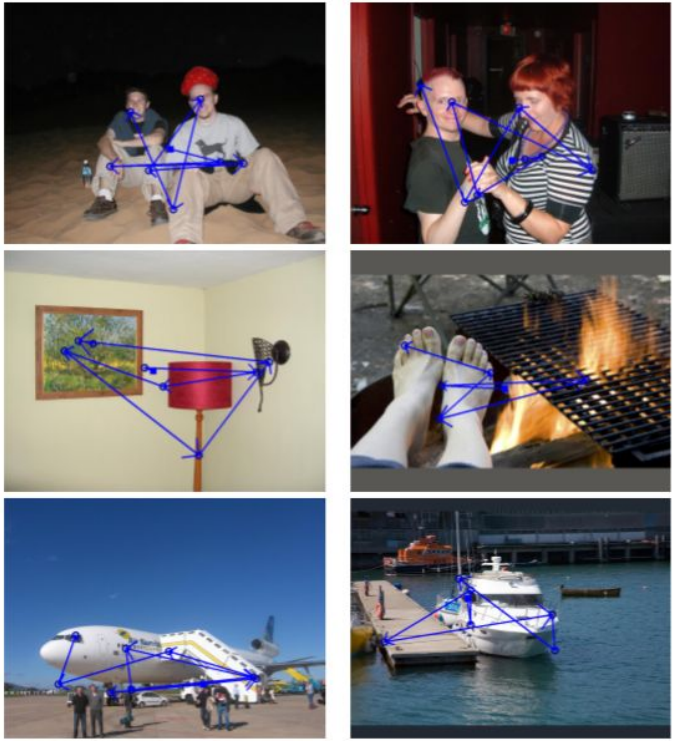
Inhibition of return function is used to suppress contribution of the virtual mass associated with the spatial gradient of the brightness, whereas it is reasonable to not apply it to the optical flow to favour the tracking behaviour

$$\mu(x,t) = \mu_1(x,t)(1 - I(x,t)) + \mu_2(x,t).$$

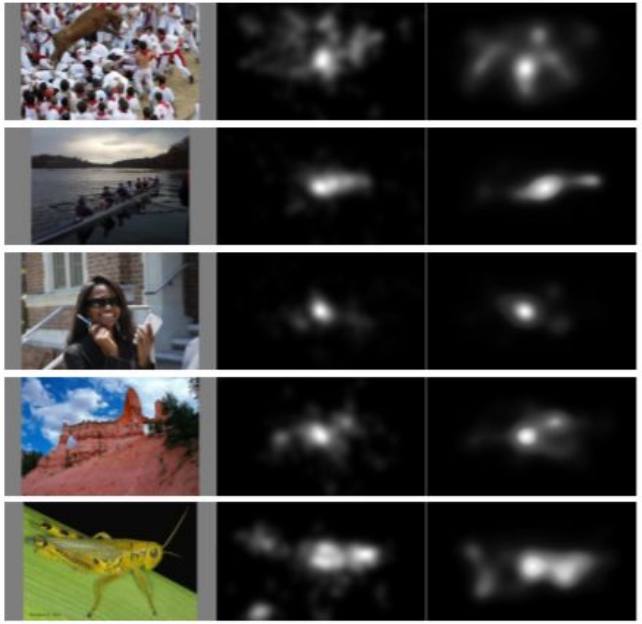
We are now ready to write the Newtonian laws of motion for the system

$$\ddot{a}(t) + \lambda\dot{a}(t) + (e * \mu)(t, a(t)) = 0.$$

G-Eymol



G-Eymol



G-Eymol

Model	FixaTons			
	String-Edit (distance)		Scaled Time-delay embedding (similarity)	
	Mean	Best	Mean	Best
G-Eymol	**7.34 (2.42)**	**3.72 (1.92)**	**0.81 (0.03)**	**0.85 (0.03)**
Eymol [55]	7.94 (2.46)	4.10 (1.95)	0.74 (0.07)	0.81 (0.07)
SAM [13]	8.02 (2.53)	4.25 (1.95)	0.77 (0.08)	0.83 (0.08)
Deep Gaze II [38]	8.17 (2.52)	4.34 (1.95)	0.72 (0.10)	0.79 (0.10)
Itti [21]	8.15 (2.48)	4.36 (1.94)	0.70 (0.09)	0.76 (0.09)
Center	8.13 (2.42)	4.35 (1.90)	0.72 (0.04)	0.77 (0.04)
Random	8.21 (2.40)	4.43 (1.87)	0.70 (0.04)	0.75 (0.04)

Model	COUTROT DATASET			
	String-Edit (distance)		Scaled Time-delay embedding (similarity)	
	Mean	Best	Mean	Best
G-Eymol	**35.68** (13.97)	**23.83** (13.07)	**0.79** (0.05)	**0.86** (0.04)
Eymol	39.90 (11.29)	30.48 (10.76)	0.77 (0.03)	0.84 (0.03)
Center	44.24 (2.24)	36.68 (1.41)	0.74 (0.01)	0.79 (0.001)
Random	45.51 (2.97)	38.45 (1.33)	0.70 (0.01)	0.76 (0.01)

G-Eymol

ONLINE DEMO AT: https://sailab.diism.unisi.it/attention/

Future works

- Image captioning
- Integration of color and sound
- Integration with high level features
- Use Eymol for neurodegenerative diseases diagnosis
- Modeling of task dependence
- Evaluate in a curiosity learning scenario
- ...

Thank you!

Questions?