Variational Laws of Visual Attention for Dynamic Scenes

dariozanca
dario.zanca@unifi.it
marcogori
marco@diism.unisi.it

Abstract

Computational models of visual attention are at the crossroad of many disciplines. In computer vision, there is an important need for mechanisms that allow to disentangle the huge amount of information carried by a video stream. In this work we aim at developing a real-time visual attention model, which tries to capture saliency as well as dynamic properties of a scene. The whole work is inspired by the human mechanism of eye movement together with more general functional principles.

Principles of Visual Attention

Three fundamental principles drive the attention. They are given a mathematical formulation that lead to the introduction of the correspondent terms of the Lagrangian of the action functional.

- **Retina**

  Scanpaths are limited inside the input frame, where the totality of the information resides

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

- **Curiosity**

  Attention is directed toward details, more than uniform parts of the visual input

$$C(t,x) = b_x^2 \cos^2(\omega t) + p_x^2 \sin^2(\omega t)$$

- **Brightness invariance**

  Trajectories which exhibit brightness invariance are to be preferred, because this brings two interesting behaviors: fixations and tracking

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

EYMOL: Eye Movement Laws

Scanpaths of visual attention are modeled as the motion of a particle of mass *m* within the potential field defined by the terms described above. Given the action

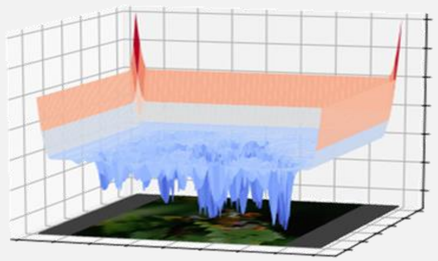
$$\int_{\mathcal{T}} L(t,x,\dot{x})\,dt = \int_{\mathcal{T}} \frac{1}{2} m\dot{x}^2 - [\underbrace{V(x)}_{retina} - \underbrace{C(t,x)}_{curiosity} + \underbrace{B(t,x,\dot{x})}_{bright.inv.}]\,dt$$

the scanpath turns out to be a stationary point of this functional, which is given by the correspondent Euler-Lagrange equations. In our case,

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

Potential field

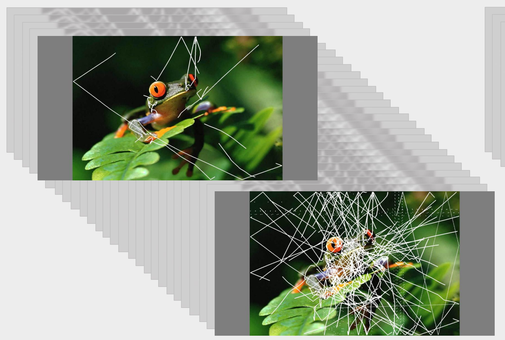
Stimulus



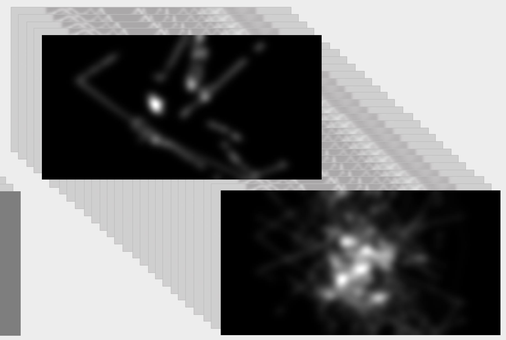
Potential field

Given a stimulus (left), it is converted to grayscale. The principles of visual attention define a potential field (rigth) that guide attention. The potential field is a continuous function of position and time so that this approach is naturally applicable in case of dynamic scenes. EYMOL differential equations describe the movement of a mass m within this potential field.

Simulation and saliency map generation



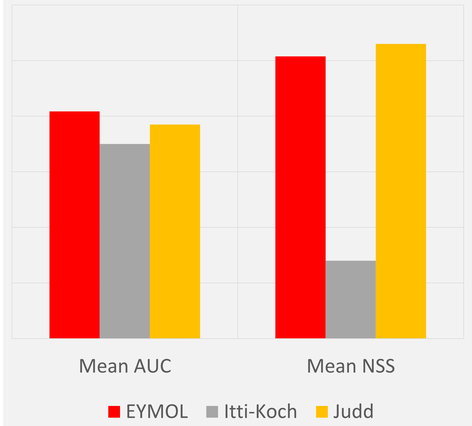
Simulated scanpaths



Saliency map

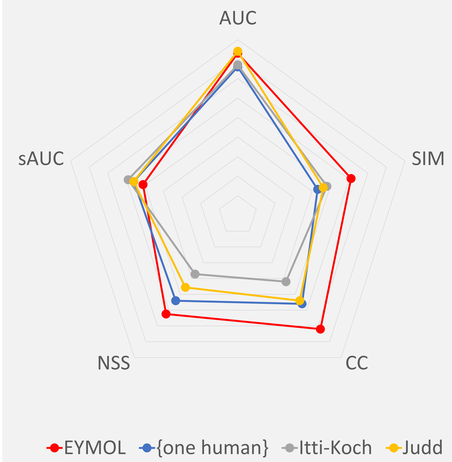
In order to produce saliency maps, 199 different scanpaths (left) of visual attention have been calculated. Different scanpaths are obtained by small changes on initial conditions of the system, i.e. initial position and velocity. Saliency maps (right) are obtained by integrating all the scanpaths. Maps are convolved with a gaussian filter, because this improves performance in the task of saliency maps estimation.

Results on SFU (dynamic)



The SFU dataset contains 12 clips and fixations of 15 observers, each of them have watched twice every video. A saliency map is computed for each frame, then the mean values of AUC and NSS among all frames is used to compare the models.

Results on CAT2000 (static)



The CAT2000 dataset contains 2000 images for train and 2000 images for test. The 2000 images for test are kept private and scores are officially provided by MIT Saliency Benchmark Team.
The baseline {one human} indicates how well a fixation map of one observer of the dataset (taken as a saliency map) predicts the fixations of the other N-1 observers.

Lesson learned

Current computational models of visual attention postulate a centralized rule of the saliency map. This implies that saliency, and as a consequence attention, is the product of global computation over an entire stack of feature maps.

Instead, we show that also a local approach is able to capture important aspect of this phenomenon. Moreover, in terms of performance in a task of saliency detection, it is even better of many other model with respect of different metrics.

Our approach has the important advantage of avoiding global computation and to be, by its mathematical definition, online and adapt for dynamic scenes.

Future directions

In the present work we investigated the human attention mechanisms in their early stage of vision, which we assume completely data-driven. For the future, we aim at:

- investigating behavioral data, not only in terms of saliency maps, but also by comparing actual generated scanpaths with human data in order to discover temporal correlations;
- providing the integration of the presented model with a theory of feature extraction;

Links and downloads

A library for an easy use of the model is available on GitHub. It contains many additional tools for scanpath simulation and saliency map estimation.

We are very open for collaboration and invite everyone who is interested to contact us.

https://github.com/dariozanca/eymol