A probabilistic tour of visual attention and gaze shift computational models

Giuseppe Boccignone
Dipartimento di Informatica
Università di Milano
Giuseppe.Boccignone@unimi.it
http://phuselab.di.unimi.it

Active Vision and perception in Human (-Robot) Collaboration (AVHRC 2020) September, 2020
Some key points of this talk

- Stochastic processes
- Lévy flights
- Bayesian modelling
- Foraging theory
- Bayesian decision-making theory
- Emotions
- Eye guidance modelling
Yet, we have this wandering eye…
//at the heart of active sensing

The world $W(t)$ as we perceive it at time $t$.
Computational models of eye guidance

\textbf{\textit{The bare essence}}

1. Where do people look?

\[ I \xrightarrow{T} \{ r_F(1), r_F(2), \cdots \} \]

2. How do people look there?
Computational models of eye guidance
\textbackslash the bare essence

2. How do people look there?

I. Where do people look?

Goal, Task \( G \) \( \rightarrow \) \( r_F(t) \) \( \rightarrow \) \( r_F(t + 1) \)

\( W(t) \) \( \rightarrow \) \( \mathcal{W}(t) \) \( \rightarrow \) \( \mathcal{W}(t + 1) \)

Stimulus

Perceptual evaluation

Gaze shift
Computational models of eye guidance
//The historical baseline: Itti, Koch & Niebur model

1. Where do people look?

2. How do people look there?
Computational models of eye guidance

\[ \text{the bare essence} \]

1. Where do people look?

\[ I \mapsto \mathcal{R} \]  
(e.g., saliency map)

\[ \mathcal{R} \mapsto \{ r_F(1), r_F(2), \cdots \} \]

2. How do people look there?
Computational models of eye guidance
//anatomy & misery of saliency maps
Computational models of eye guidance
// anatomy & misery of saliency maps

1. Where do people look?
   \[ I \mapsto \mathcal{R} \] (e.g., saliency map)

2. How do people look there?
   \[ \mathcal{R} \mapsto \{ r_{F}(1), r_{F}(2), \ldots \} \]
Computational models of eye guidance //anatomy & misery of saliency maps

posterior prob. of gaze shift

\[ P(\mathbf{r} | \mathcal{W}) \]

data likelihood under the shift

\[ \frac{P(\mathcal{W} | \mathbf{r})}{P(\mathcal{W})} \]

gaze shift prior

\[ \hat{P}(\mathbf{r}) \]

ts this is a shift
Computational models of eye guidance // anatomy & misery of saliency maps

posterior prob. of gaze shift
\[
P(r | W) = \frac{P(W | r)}{P(W)} \cdot \hat{P}(r).
\]
this is a shift
\[
P(r) = P(r_F(t) - r_F(t-1)) \approx P(r_F(t) | r_F(t-1)) = P(r_F(t)).
\]

posterior prob. of gazing at
\[
P(r_F | W) = \frac{P(W | r_F)}{P(W)} \cdot \hat{P}(r_F)
\]
this is a point
\[
r_F(t) \to r_F(t + 1)
\]
Computational models of eye guidance // anatomy & misery of saliency maps

- posterior prob. of gaze shift
  \[ P(\mathbf{r} \mid \mathbf{W}) = \frac{P(\mathbf{W} \mid \mathbf{r})}{P(\mathbf{W})} \]

  data likelihood under the shift
  \[ \hat{P}(\mathbf{r}) \]

  gaze shift prior

  this is a shift

- posterior prob. of gazing at
  \[ P(\mathbf{r}_F \mid \mathbf{W}) = \frac{P(\mathbf{W} \mid \mathbf{r}_F)}{P(\mathbf{W})} \]

  data likelihood under gaze at
  \[ \hat{P}(\mathbf{r}_F) \]

  prior prob. of gazing at

  this is a point

- posterior prob. of selecting location L
  \[ P(\mathbf{L} \mid \mathbf{F}) = \frac{P(\mathbf{F} \mid \mathbf{L})}{P(\mathbf{F})} \]

  feature likelihood under location L
  \[ \hat{P}(\mathbf{L}) \]

  prior prob. of location L

  this is a map

\[ \mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t + 1) \]
Computational models of eye guidance //anatomy & misery of saliency maps

\[
P(r | \mathcal{W}) = \frac{P(\mathcal{W} | r)}{P(\mathcal{W})} P(r).
\]

\[
P(r_F | \mathcal{W}) = \frac{P(\mathcal{W} | r_F)}{P(\mathcal{W})} P(r_F).
\]

\[
P(L | F) = \frac{P(F | L)}{P(F)} P(L).
\]

\[
P(L | F) \propto \frac{1}{P(F)}.
\]

Eq. 15. Torralba
Computational models of eye guidance // anatomy & misery of saliency maps

where \( x = x_F(t) - x_F(t-1) \) is the random vector representing the gaze shift (in [101], saccades), and \( D \) generically stands for the input data. As Tatler after Dorr

tures (of an assumption of independence: dynamics. In probabilistic terms we may re-phrase this result as the outcome between Eq. 7 and Eq. 14 is subtle. But, as a matter of fact, Eq. 14 bears no

ject (they may be); then Eq. 14 boils down to the following

terior prob. of selecting location \( L \)

feature likelihood under location \( L \)

prior prob. of location \( L \)

\[
\frac{P(L \mid F)}{P(F)} = \frac{P(F \mid L) P(L)}{P(F)}
\]

By careful inspection, it can be noted that the posterior

posterior prob. of gazing at

by a simplified version of Eq. 7:

\[
P(L) = P(L 
\mid \left[\begin{array}{c}
F_L\
F_L
\end{array}\right])
\]

Further, the prior

by a simplified version of Eq. 7:

\[
P(L) = P(L \mid \left[\begin{array}{c}
F_L\
F_L
\end{array}\right])
\]

To make things even clearer, let us explicitly substitute

for a large number of visual attention models that have been proposed in

the object-based model PGM (Figure 11, center), which is the one previously

represented on the left of Figure 11. As it can be seen, it is a subgraph of

the PGM structure to brain areas underpinning visual attention: early visual areas V1 and

cortex (PFC).

V2, V4, lateral intraparietal (LIP), frontal eye fields (FEF), inferotemporal (IT), prefrontal

salience-based model in the literature proposed by Itti

either luminance, color, texture or motion) and clearly relates to entropy

of such a pure bottom-up approach has b

di

proposed are much or less variations of this leitmotif (experimenting with

to gain the understanding that a great deal computational models so far

they may be); then Eq. 14 boils down to the the following

Eq. 16 tells that the probability of fixating a spatial location

as discussed by Foulsham and Underwood [40].

may be still more predictive than chance while preparing for a memory test

correlational e

of such a pure bottom-up approach has b

di

proposed are much or less variations of this leitmotif (experimenting with

to gain the understanding that a great deal computational models so far

they may be); then Eq. 14 boils down to the the following

Eq. 16 tells that the probability of fixating a spatial location

as discussed by Foulsham and Underwood [40].

may be still more predictive than chance while preparing for a memory test

correlational e

Computational models of eye guidance
//anatomy & misery of saliency maps

\[
\frac{P(L \mid F)}{P(L)} = \frac{P(F \mid L)}{P(F)} = \frac{P(L \mid F)}{P(L)}
\]

posterior prob. of selecting location L

feature likelihood under location L

prior prob. of location L
To sum up...

Eye movements and perception: A selective review

Alexander C. Schütz
Department of Psychology, Gießen University, Gießen, Germany

Doris I. Braun
Department of Psychology, Gießen University, Gießen, Germany

Karl R. Gegenfurtner
Department of Psychology, Gießen University, Gießen, Germany
To sum up...

State-of-the-Art in Visual Attention Modeling

65 models: variations, variations of variations, variations of variations of variations,...... on base schemes (Itti & Koch)
Computational models of eye guidance
//anatomy & misery of saliency maps

1. Where do people look?
   \[ I \mapsto \mathcal{R} \] (e.g., saliency map)
   \[ \mathcal{R} \mapsto \{ r_F(1), r_F(2), \ldots \} \]

2. How do people look there?
   \[ \text{arg max } \mathcal{R} \]

Deterministic gaze shift, no variability!
The problem of variability
// How random are gaze shifts?
The problem of variability
// How random are gaze shifts?

(Itti & Koch)
The problem of variability
//Oculomotor tendencies
The problem of variability
//Oculomotor tendencies

(A) Human

(B) Mean = 13.97
Median = 11.25
Mode = 2.50

(C) (Itti & Koch)

(D) Mean = 17.25
Median = 16.87
Mode = 14.50
The problem of variability

// Oculomotor tendencies

- Oculomotor tendencies:
  - regularities that are common across all instances of and manipulations to the behavior

- Tatler & Vincent:
  - a model based on oculomotor biases alone performs better than the standard salience model

If one samples from prior only

\[ r(t) \sim P(r(t)), \quad t = 1, 2, \ldots \]

blind to visual information, out-performs feature-based accounts of eye guidance:

0. 648 area under the receiver operator curve (AUC) as opposed to 0. 593 for edge information and 0. 565 for salience information!
Computational models of eye guidance
//bringing variability into the game

\[
\begin{align*}
\text{posterior prob. of gaze shift} & \quad = \quad \frac{P(W | r)}{P(W)} \\
\end{align*}
\]
Computational models of eye guidance
\the bare essence

1. Where do people look?

2. How do people look there?

Goal, Task $G \rightarrow r_F(t) \rightarrow r_F(t + 1)$

Stimulus $W(t) \rightarrow W(t + 1)$

Perceptual evaluation

Gaze shift
Computational models of eye guidance
// bringing variability into the game

Eye guidance modelling

Bayesian modelling

Stochastic processes & Lévy flights

Bayesian Decision-making theory
Emotions
Foraging theory
Computational models of eye guidance
//bringing variability into the game


The ecology of gaze shifts

Dirk Brockmann*, Theo Geisel

Fig. 1. Left, Center: Two typical scanpaths on different trials by the same subject. Each scanpath consists of approximately 350 saccades. Right: Saccadic magnitude histogram calculated from the scanpaths depicted. $\theta$ denotes saccadic magnitude in degrees of visual angle.
Bringing variability into the game
//anomalous walks

Brownian (Gaussian) walk

Cauchy walk

Fig. 1. Left, Center: Two typical scanpaths on different trials by the same subject. Each scanpath consists of approximately 350 saccades. Right: Saccadic magnitude histogram calculated from the scanpaths depicted. \( \theta \) denotes saccadic magnitude in degrees of visual angle.
Bringing variability into the game // anomalous walks

Lèvy alpha-stable distributions

\[ \alpha = 1 \quad \alpha = 1.5 \quad \alpha = 1.8 \quad \alpha = 2 \]

Cauchy walk

Gaussian walk
Gaze-shift as a constrained random walk

\[ \mathbf{r}_{\text{new}}(t) = \mathbf{r}(t) - \nabla V + \eta \]
Successful Applications: Robot Action Learning for the iCub (Nagai. 2009)


Fig. 7. Transition of attention of proposed model (a) and two comparative models, (b) and (c). The line color corresponds to the cup color.
Computational models of eye guidance
//gaze shifts as actions
Computational models of eye guidance
//gaze shifts as actions

Ecological Sampling of Gaze Shifts
Giuseppe Boccignone and Mario Ferraro

Matlab simulation code:
https://github.com/phuselab/EcoSampling
Ecological sampling of gaze shifts
//sampling the landscape

- Sampling the natural habitat
  \[ W^*(t) \sim P(W(t)|r_F(t), F(t), I(t)) \]
Ecological sampling of gaze shifts
//sampling the oculomotor action

- Sampling the appropriate motor behavior

\[ A(t)^* \sim P(A(t)|A(t-1), W^*(t)) \]

\[ \pi^*(t) \sim \text{Dir}(\pi|\nu(O(t))) \]

\[ z^*(t) \sim \text{Mult}(z(t)|\pi^*(t)) \]
Ecological sampling of gaze shifts
// oculomotor actions: fixate, pursuit, saccade

Variational Bayesian Student-t mixture model

Lèvy flight
Ecological sampling of gaze shifts //sampling the oculomotor action

- Sampling where to look next

\[ r_F(t+1) \sim P(r_F(t+1)|A(t)^*, W^*(t), r_F(t)) \]

**Action control**

\[ f(\xi; \alpha, \beta, \gamma, \delta) \rightarrow A(t+1) \]

**Perceptual evaluation**

\[ r_F(t) \rightarrow r_F(t+1) \]

\[ W(t) \rightarrow W(t+1) \]

\[ \xi \sim f(\xi; \alpha, \beta, \gamma, \delta) \]

**Alpha-stable (Lévy flight)**

\[ r_F(t_{n+1}) \approx r_F(t_n) - \sum_{p=1}^{N_v} (r_F(t_n) - r_p(t_n)) \tau + \gamma_k \tau^{1/\alpha_k} \xi_k. \]
Ecological sampling of gaze shifts
//some results...

Human

CCDF: comp.#1(all subj.)

log P(X > x)

empirical
estimated

log x (pixels)

CCDF: comp.#2(all subj.)

log P(X > x)

empirical
estimated

log x (pixels)

CCDF: comp.#3(all subj.)

log P(X > x)

empirical
estimated

log x (pixels)

Model

CCDF: comp.#1(Eco Sampl.)

log P(X > x)

empirical
estimated

log x (pixels)

CCDF: comp.#2(Eco Sampl.)

log P(X > x)

empirical
estimated

log x (pixels)

CCDF: comp.#3(Eco Sampl.)

log P(X > x)

empirical
estimated

log x (pixels)

Lèvy flight
Computational models of purposive eye guidance // decisions on actions: considering task / goals
Some key points of this talk

- Bayesian modelling
- Stochastic processes & Lévy flights
- Eye guidance modelling
- Bayesian decision-making theory
- Foraging theory
- Emotions
Computational models of purposive eye guidance
// Considering task / goals

mobile eye-tracking

homography derived panoramic image
Computational models of purposive eye guidance // Bayesian Decision Theory

\[ P(x|z_{0:t}, M) \propto P(z_{0:t}|x, M)P(x|M) \]

\[ Q_{\text{now}}(a, P(x|z_{0:t}, M)) = \sum_x R(a, x)P(x|z_{0:t}, M) \]

**Theoretical perspectives on active sensing**
Scott Cheng-Hsin Yang\(^1\), Daniel M Wolpert\(^1,3\) and Máté Lengyel\(^1,2,3\)

*Current Opinion in Behavioral Sciences* 2016, 11:100–108
Computational models of purposive eye guidance // Considering task / goals

Perceptual evaluation

Action control

Goal, Task

T

Action

Value representation

Decision = Max Expected Reward

r_F(t + 1) = arg max \[E \left[ R^{(k)} \right] \]

Reward

r_F(t)

Object representation

O

Priority representation

L(t)

Perception

I
Computational models of purposive eye guidance // Considering task / goals

High level of representation

Value for text (search task)

Object-based perception (free viewing task)

Low level of representation

Value for people (search task)

Early salience based perception (free viewing task)
The foraging perspective

- Bayesian modelling
- Stochastic processes & Lévy flights
- Bayesian decision-making theory
- Emotions
- Foraging theory

Eye guidance modelling
Back to the random walks....
The foraging perspective
The foraging perspective

Foraging pattern of spider monkeys in the Yucatan Peninsula

“Foraging pattern” of the eye

Long-tail distributions beyond the Central Limit Theorem

Anomalous diffusion (Cauchy walk)
The foraging perspective

Anomalous diffusion
Lèvy Flights / Walks
The foraging perspective
What was once foraging in a physical space for tangible resources became, over evolutionary time, foraging in cognitive space for information related to those resources.
The foraging perspective
//optimal foraging theory

- What prey to take (optimal diet choice)
- What patch type to search (optimal patch choice)
- When to leave a patch (optimal giving up or departure times, GUT)
- How to move between patches (optimal movements)
The foraging perspective // Charnov’s Marginal Value Theorem

- When to leave a patch (optimal giving up or departure times, GUT)
How do we look at social scenes?

On gaze deployment to audio-visual cues of social interactions

GIUSEPPE BOCCIGNONE, VITTORIO CUCULO, ALESSANDRO D’AMELIO, GIULIANO GROSSI AND RAFFAELLA LANZAROTTI

PPhuSe Lab - Dipartimento di Informatica, Università degli Studi di Milano,
via Celoria 18, Milano 20133, Italy (e-mail: {giuseppe.boccignone, vittorio.cuculo, alessandro.damelio, giuliano.grossi, raffaella.lanzarotti}@unimi.it)

Corresponding author: Alessandro D’Amelio (e-mail: alessandro.damelio@unimi.it).

This work was supported in part by University of Milano under Grant PSR 2019

ABSTRACT

Attention supports our urge to forage on social cues. Under certain circumstances, we spend the majority of time scrutinising people, markedly their eyes and faces, and spotting persons that are talking. To account for such behaviour, this paper develops a computational model for the deployment of gaze within a multimodal landscape, namely a conversational scene. Gaze dynamics is derived in a principled way by reformulating attention deployment as a stochastic foraging problem. Model simulation experiments on a publicly available dataset of eye-tracked subjects are presented. Results show that the simulated scan paths exhibit similar trends of eye movements of human observers watching and listening to conversational clips in a free-viewing condition.

INDEX TERMS
audio-visual attention, gaze models, social interaction, multimodal perception

I. INTRODUCTION

Consider a clip displaying social interactions, in particular a conversational clip (audio and video): the chief concern of this paper is to model the deployment of attention through gaze by a human subject who is viewing and listening to the clip.

Why should this research problem be relevant beyond its merits?

One straightforward reason lies in the classic data mining hurdle. YouTube, Twitch, Facebook Live contain myriads of such clips [1], [2]. Also, large-scale multimodal data conveying social interactions from non-laboratory settings are being increasingly employed to analyse behaviours, emotions, and interactions in real-life situations [3]. It goes without saying, the processing of large spatio-temporal data from multiple media in different contexts is a mind-blowing engineering challenge: spotting sharable highlights, capturing socially relevant events, generate value-based summaries to facilitate browsing and skimming. All such problems call for an ability that is germane to the successful performance of any cognitive task: the ability to predict and to select where the most meaningful and task-relevant information is to be found in the sensory input.

A less evident, albeit earnest need takes root in the challenge of “subject’s mining”: the computational inference of subject’s traits, or expertise, or even expectations from attentive behaviour. Much can be gained indeed by analysing the “mind’s eye” conduct of a subject who scrutinises and forages on the behaviour of other subjects involved in social interactions [4]–[7].

In a nutshell, the research problem addressed here is relevant beyond its peculiar interest because it complies with a quest for parsimony. Under a variety of circumstances, what prima facie might come across as a conundrum of diverse engineering problems, boils down to the modelling of one and only skill: the effective deployment of attention that organisms have evolved to promote survival and well-being. Surprisingly, the dynamics of deployment has been hitherto overlooked in computational approaches.

Problems and challenges.

Throughout our lives, we are bond to unfalteringly sample the environment. Moment-by-moment we strive to answer the question: Where to look next? Attention guides our gaze to the appropriate location of the scene and holds it in that location for the deserved amount of time given current processing demands [8]. In doing so, like other animals with as diverse evolutionary backgrounds, we exhibit a consistent pattern of eye movements. To illustrate at the finest “resolution scale” the signature of gaze dynamics, Fig. 1 plots the raw data recording of one subject’s gaze. The trajectory of gaze is shown as unfolding in time on an excerpt of subsequent frames: large relocations are followed by local clustering of gaze points.

multimodal stimulus (video + audio)
How do we look at social scenes?

Action control

Gaze shift

Perceptual evaluation

Stimulus

Endogenous
Goal, Task

$G$

$A(t) \rightarrow A(t+1)$

$r_F(t) \rightarrow r_F(t+1)$

$W(t) \rightarrow W(t+1)$
How do we look at social scenes?

- Social context and multimodal stimuli
- Foraging in the multimodal landscape

- Salient objects / events in stimuli
  - Visual priority map
  - Audiovisual priority map
  - Patches / attractors for multimodal attention

- Selection history
- Current goals and value

- Patch exploitation
- Patch exploration
- State sampling
- Patch sampling
How do we look at social scenes? //value-based patches

Patches = speaker, non speaker, hand (gestures), body (gestures), ....

social value

inferred from behaviour
How do we look at social scenes?
//giving-up time of a patch (stochastic Charnov)
How do we look at social scenes? //exploitation vs. exploration random walks

Exploitation within patch

Exploration between patches
How do we look at social scenes?
//exploitation vs. exploration random walks
How do we look at social scenes?//exploitation vs. exploration random walks

**Shift amplitude**

**Fixation duration**

**Gaze direction**

- **Real**
- **Generated**
Computational models of purposive eye guidance // the value of “value” (and reward)
The evolutionary role of dopamine in the modulation of goal-directed behavior and cognition is further supported by pathologies of human goal-directed cognition, which have motor and cognitive dysfunction and organize themselves, with respect to dopaminergic activity, perseverative to unfocused.

Dopamine is a key component in foraging behaviors in invertebrates and vertebrates, in vertebrates dopamine is also associated with goal-directed cognition.
At the heart of purposive eye guidance //the dopamine hypothesis (Hills)
Computational models of eye guidance
//bringing emotions into the game

Eye guidance modelling

Bayesian modelling

Stochastic processes & Lévy flights

Bayesian Decision-making theory
Emotions
Foraging theory
At the heart of purposive eye guidance //Value & reward: a doorway to emotions

On the relationship between emotion and cognition

Luiz Pessoa

Nature Reviews Neuroscience Volume 9 | February 2008 | 180

Diagram showing neural networks related to emotion and cognition, including areas such as the nucleus accumbens, anterior cingulate cortex, orbito-frontal cortex, lateral prefrontal cortex, and early and late visual areas. Arrows indicate the direction of information flow, with labels for dopaminergic modulation (reward) and cholinergic modulation (attention).
Future work
//bridging active sensing and emotions
Future work
//bridging active sensing and emotions
Future work
//bridging active sensing and emotions
Future work
//bridging active sensing and emotions

Future work
//bridging active sensing and emotions
Thank you!