Master SIF - REP (Part 10)
Perception

Thomas Maugey (courtsey of Olivier Le Meur)
thomas.maugey@inria.fr

Université de Rennes
Inria

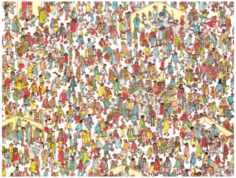
Fall 2023

Outline

1. Visual attention
   - Presentation
   - Overt vs covert
   - Bottom-Up vs Top-Down

Introduction to visual attention (1/5)

Natural visual scenes are cluttered and contain many different objects that cannot all be processed simultaneously.



Where is Waldo, the young boy wearing the red-striped shirt...

Amount of information coming down the optic nerve $10^8 - 10^9$ bits per second

Far exceeds what the brain is capable of processing...

Introduction to visual attention (2/5)

WE DO NOT SEE EVERYTHING AROUND US!!!

Test Your Awareness : Whodunnit?

YouTube link: www.youtube.com/watch?v=ubNF9QNEQLA

Introduction to visual attention (3/5)

Visual attention

Posner proposed the following definition (Posner, 1980). Visual attention is used:

- to select important areas of our visual field (alerting);
- to search for a target in cluttered scenes (searching).

There are several kinds of visual attention:

- Overt visual attention: involving eye movements;
- Covert visual attention: without eye movements (Covert fixations are not observable).

Introduction to visual attention (4/5)

Bottom-Up vs Top-Down

- **Bottom-Up**: some things draw attention reflexively, in a task-independent way (Involuntary; Very quick; Unconscious);


- **Top-Down**: some things draw volitional attention, in a task-dependent way (Voluntary; Very slow; Conscious).

Bottom-Up vs Top-Down

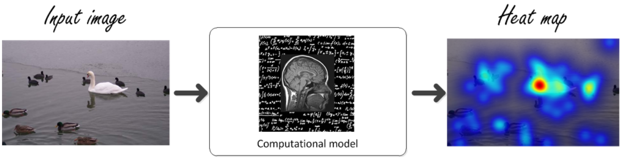
- **Bottom-Up**: some things draw attention reflexively, in a task-independent way (Involuntary; Very quick; Unconscious);

- **Top-Down**: some things draw volitional attention, in a task-dependent way (Voluntary; Very slow; Conscious).

Introduction to visual attention (5/5)

Computational models of visual attention aim at predicting where we look within a scene.

In this presentation, we are focusing on Bottom-Up models of overt attention but we want to go beyond.



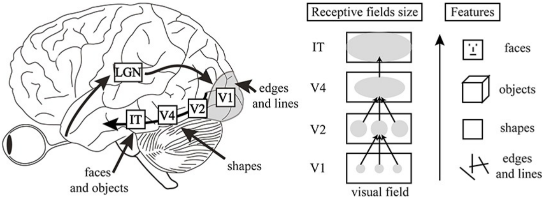
Computational models of visual attention

2 Computational models of visual attention
- Main hypothesis
- Taxonomy
- Information theoretic model
- Cognitive model

Computational models of Bottom-up visual attention (1/5)

Main ingredients

Computer vision models often follow closely the philosophy of neurobiological feedforward hierarchies.



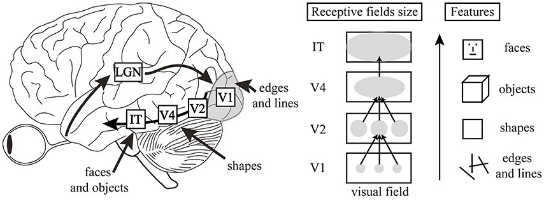
Adapted from (Herzog and Clarke, 2014, Manassi et al., 2013).

- Basic features (e.g. edges and lines) are analyzed by independent filters (V1);
- Higher-level neurons pool information over multiple low-level neurons with smaller receptive fields and code for more complex features.

Computational models of Bottom-up visual attention (2/5)

Main ingredients

Computer vision models often follow closely the philosophy of neurobiological feedforward hierarchies.



Adapted from (Herzog and Clarke, 2014, Manassi et al., 2013).

The deeper we go, the more complex features we extract...

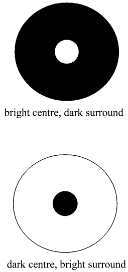
Deep features.

Computational models of Bottom-up visual attention (3/5)

Main ingredients

Computer vision models often follow closely the philosophy of neurobiological feedforward hierarchies.

Receptive Field = region of the retina where the action of light alters the firing of the neuron

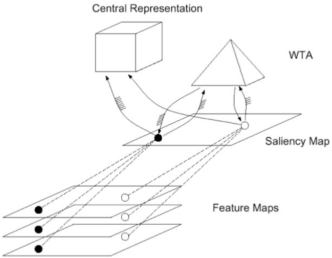


- RF = center + surrround;
- The size of the RF varies: for V1 neurons (0.5-2 degrees near the fovea), inferotemporal cortex neurons (30 degrees).
- Simulated by DoG, Mexican Hat...

Computational models of Bottom-up visual attention (4/5)

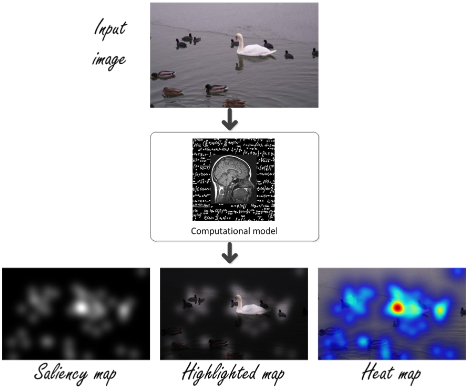
Main ingredients

Most of the computational models of visual attention have been motivated by the seminal work of (Koch and Ullman, 1985).



- a plausible computational architecture to predict our gaze;
- a set of feature maps processed in a massively parallel manner;
- a single topographic saliency map.

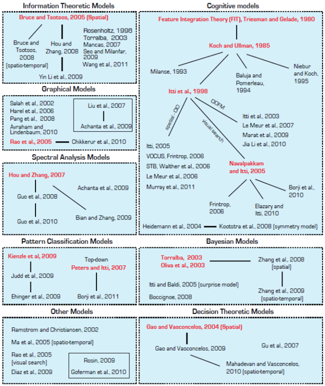
Computational models of Bottom-up visual attention (5/5)



Computational models of Bottom-up visual attention (1/1)

Taxonomy of models:

- Information Theoretic models;
- Cognitive models;
- Graphical models;
- Spectral analysis models;
- Pattern classification models;
- Bayesian models.
- Deep network-based models.

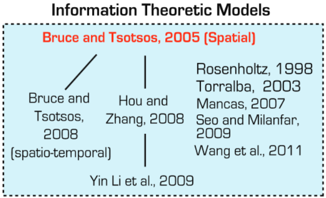


Extracted from (Borji and Itti, 2013).

Information theoretic model (1/3)

Information Theory

- Self-information,
- Mutual information,
- Entropy...



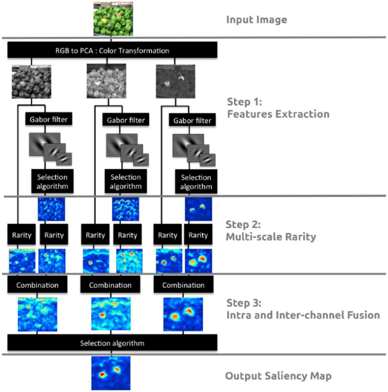
Extracted from (Borji and Itti, 2013).

Self-information is a measure of the amount information provided by an event. For a discrete X r.v defined by $\mathcal{A} = \{x_1, ..., x_N\}$ and by a pdf, the amount of information of the event $X = x_i$ is given by:

$$\boxed{I(X = x_i) = -log_2 p(X = x_i), \text{ bit/symbol}}$$

Information theoretic model (2/3)

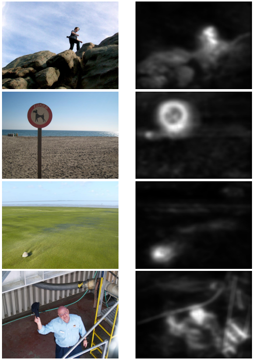
(Riche et al., 2013)'s model (RARE2012)



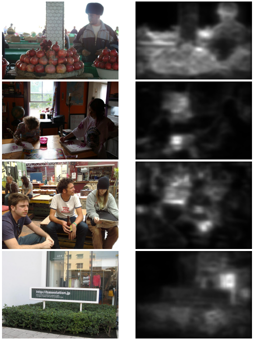
Information theoretic model (3/3)

(Riche et al., 2013)'s model (RARE2012)

- Good prediction:



- Difficult cases:

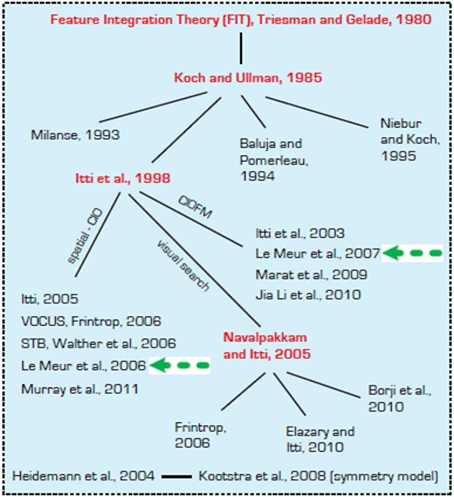


Cognitive model (1/3)

as faithful as possible to the Human Visual System (HVS)

- inspired by cognitive concepts;
- based on the HVS properties.

Cognitive models



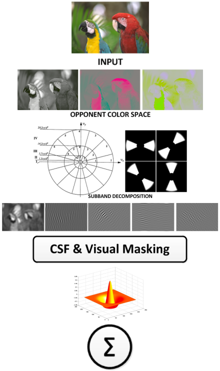
Extracted from (Borji and Itti, 2013).

Cognitive model (2/3)

(Le Meur et al., 2006)'s cognitive model

In (Le Meur et al., 2006), we designed a computational model of bottom-up visual attention.

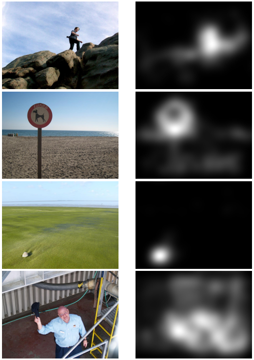
1. Input color image;
2. Projection into a perceptual color space;
3. Subband decomposition in the Fourier domain;
4. CSF and Visual Masking;
5. Difference of Gaussians;
6. Pooling.



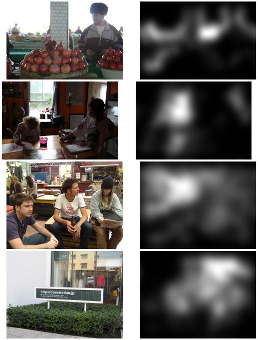
Cognitive model (3/3)

(Le Meur et al., 2006)'s cognitive model

- Good prediction:



- Difficult cases:



Performances

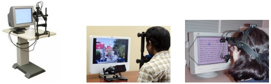
3 Saliency model's performance
- Ground truth
- Similarity metrics
- Benchmark

Ground truth (1/2)

The requirement of a ground truth

- Eye tracker (sampling frequency, accuracy...);
- A panel of observers (age, naive vs expert, men vs women...);
- An appropriate protocol (free-viewing, task...).

Cambridge research system



Tobii

Apple bought SMI.

Ground truth (2/2)

- Discrete fixation map f^i for the i^{th} observer:

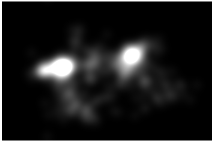
$$f^i(\mathbf{x}) = \sum_{k=1}^{M} \delta(\mathbf{x} - \mathbf{x}_k)$$

where M is the number of fixations and $\mathbf{x}_k$ is the k^{th} fixation.

- Continuous saliency map S:

$$S(\mathbf{x}) = \left(\frac{1}{N} \sum_{i=1}^{N} f^i(\mathbf{x}) \right) * G_\sigma(\mathbf{x})$$

where N is the number of observers.



Similarity metrics

- Comparing two maps:
  - The linear correlation coefficient, $cc \in [-1, 1]$;
  - The similarity metric sim uses the normalized probability distributions of the two maps (Judd et al., 2012). The similarity is the sum of the minimum values at each point in the distributions:

$$sim = \sum_{\mathbf{x}} \min\left(pdf_{map1}(\mathbf{x}), pdf_{map2}(\mathbf{x})\right) \tag{1}$$

    $sim = 1$ means the pdfs are identical, $sim = 0$ means the pdfs are completely opposite.
  - Earth Mover's Distance metric EMD is a measure of the distance between two probability distributions. It computes the minimal cost to transform one probability distribution into another one.

$EMD = 0$ means the distributions are identical, i.e. the cost is null.

  - Receiver Operating Analysis.

Le Meur, O. & Baccino, T., Methods for comparing scanpaths and saliency maps: strengths and weaknesses, Behavior Research Method, 2013.

Similarity metrics

KL-divergence and CC between two maps

- KL-Divergence:

$$KL(p|h) = \sum_{i,j} p(i,j) log_2 \frac{p(i,j)}{h(i,j)}$$

where p and h are the pdf of the predicted and human saliency maps.

$$\begin{aligned} p(i,j) &= \frac{SM_p(i,j)}{\sum_{k,l} SM_p(k,l)} \\ h(i,j) &= \frac{SM_h(i,j)}{\sum_{k,l} SM_h(k,l)} \end{aligned}$$

KL is a divergence: $KL = 0$ when p and h are strictly the same, $KL \geq 0$.

- Linear correlation coefficient:

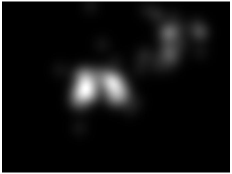
$$CC(p,h) = \frac{cov_{ph}}{\sigma_p \sigma_h}$$

where σ_k is the standard deviation of k and cov_{ph} is the covariance between p and h. CC is between -1 and 1.

Similarity metrics

ROC between two maps

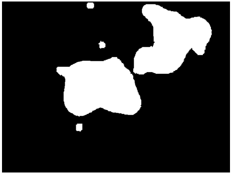
(a) Original



(b) Human

(c) Itti's model

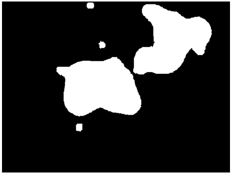
(1) Label the pixels of the human map as fixated (255) or not (0):



The threshold is often arbitrary chosen (to cover around 20% of the picture).

(2) Label the pixels of the predicted map as fixated (255) or not (0) by a given threshold T_i:

(3) Count the good and bad predictions between human and predicted maps:



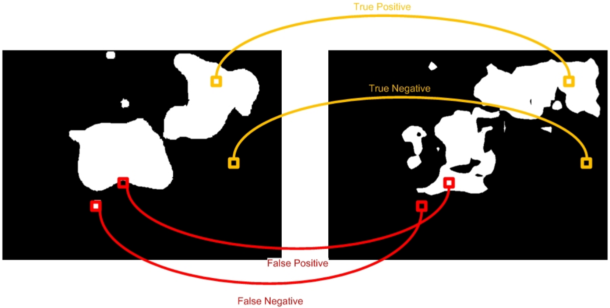
(a) Human Bin.

(b) Predicted Bin.

Similarity metrics

ROC between two maps

(3) Count the good and bad predictions between human and predicted maps:



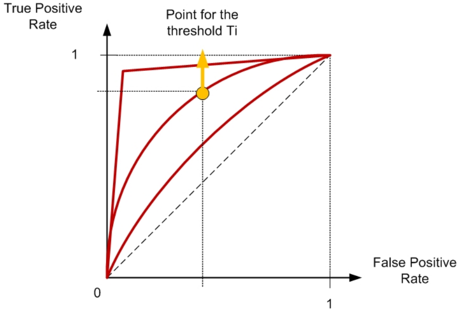
False Positive Rate = True Positive / (True Positive+False Negative)

True Positive Rate = False Positive / (False Positive+True Negative)

Similarity metrics

ROC between two maps

(4) Go back to (2) to use another threshold... Stop the process when all thresholds are tested.



AUC (Area Under Curve)

Similarity metrics

➟ Comparing a map and a set of visual fixations:

- Receiver Operating Analysis;
- Normalized Scanpath Saliency (Parkhurst et al., 2002, Peters et al., 2005);
- The Kullback-Leibler divergence (Itti and Baldi, 2005).

Le Meur, O. & Baccino, T., Methods for comparing scanpaths and saliency maps: strengths and weaknesses, Behavior Research Method, 2013.

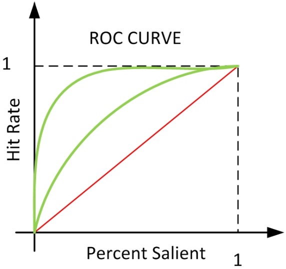
Similarity metrics

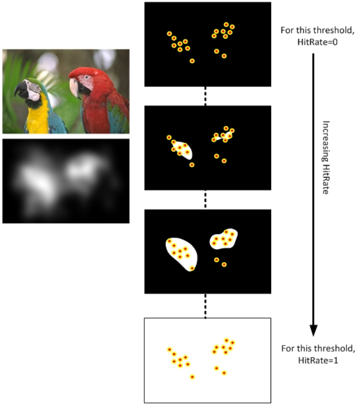
ROC between a map and a set of fixations

ROC analysis is performed between a continuous saliency map and a set of fixations.

Hit rate is measured in function of the threshold used to binarize the saliency map (Judd et al., 2009):

ROC curve goes from 0 to 1!

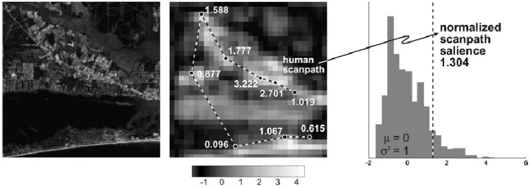




Similarity metrics
NSS

NSS (Normalized Scanpath salience) gives the degree of correspondence between human fixation locations and predicted saliency maps (Parkhurst et al., 2002),(Peters et al., 2005).

1. Each saliency map is normalized to have zero mean and one unit standard deviation.
2. Extraction of the predicted saliency at a given human fixation point.
3. Average of the previous values.



From (Peters et al., 2005)

$NSS = 0$: random performance;
$NSS >> 0$: correspondence between human fixation locations and the predicted salient points:
$NSS << 0$: anti-correspondence.

Benchmark (1/1)

Online benchmarks: http://saliency.mit.edu/

MIT300 and CAT2000

Dataset	Citation	Images	Observers	Tasks	Durations	Extra Notes
MIT300	Tilke Judd, Fredo Durand, Antonio Torralba. A Benchmark of Computational Models of Saliency to Predict Human Fixations [MIT tech report 2012]	**300** natural indoor and outdoor scenes *size:* max dim: 1024px, other dim: 457-1024px 1 dva* ~ 35px	**39** *ages:* 18-50	free viewing	3 sec	This was the first data set with held-out human eye movements, and is used as a benchmark test set. *eyetracker:* ETL 400 ISCAN (240Hz) Download 300 test images.
CAT2000	Ali Borji, Laurent Itti. CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research [CVPR 2015 workshop on "Future of Datasets"]	**4000** images from 20 different categories *size:* 1920x1080px 1 dva* ~ 38px	**24** per image (120 in total) *ages:* 18-27	free viewing	5 sec	This dataset contains two sets of images: train and test. Train images (100 from each category) and fixations of 18 observers are shared but 6 observers are held-out. Test images are available but fixations of all 24 observers are held out. *eyetracker:* EyeLink1000 (1000Hz) Download 2000 test images. Download 2000 train images (with fixations of 18 observers).

For a fair comparison, download the images, run your model and submit your results.

Matlab software is available on the webpage: http://saliency.mit.edu/.

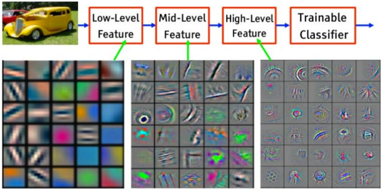
A new breakthrough but...

4. A new breakthrough
   - Convolutional Neural Network
   - CNN-based saliency prediction

A new breakthrough... (1/3)

Convolutional Neural Network in a nutshell

- A neural network model is a series of hierarchically connected functions;
- Each function's output is the input for the next function;
- These functions produce features of higher and higher abstractions;



- End-to-end learning of feature hierarchies.

Image courtesy: http://www.iro.umontreal.ca/~bengioy/talks/DL-Tutorial-NIPS2015.pdf

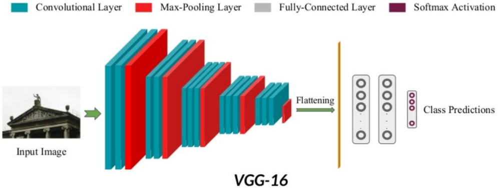
A new breakthrough... (2/3)

- Extremely big annotated datasets...
  - Imagenet, ≈ 16 Million images annotated by humans, 1000 classes (Deng et al., 2009).

- More power (GPU).

A new breakthrough... (3/3)

- One of the best CNN for image classification:



Composed of 16 layers (13 convolutional layers + 3 FC layers) (Simonyan and Zisserman, 2014) trained on Imagenet.
The number of filters of convolutional layer group starts from 64 and increases by a factor of 2 after each max-pooling layer, until it reaches 512.

- One layer = convolution + ReLU (Rectified Linear Unit ≈ truncation / nonlinear function) + Pooling (average, max)

CNN-based saliency prediction (1/9)

- *DeepGaze I: Boosting saliency prediction with feature maps trained on Imagenet*, (Kümmerer et al., 2014):

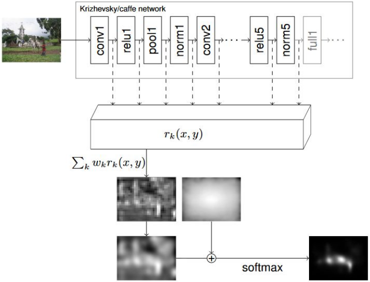
$r_k(x,y)$ represents rescaled neural responses;

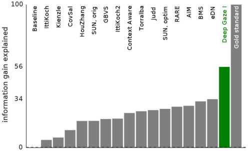
$$s(x,y) = \sum_k w_k r_k(x,y) * G_\sigma;$$

$$o(x,y) = s(x,y) + \alpha \times c(x,y);$$

SoftMax:

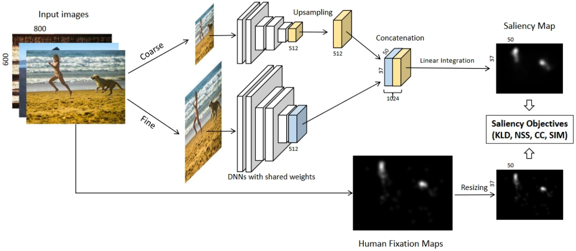
$$p(x,y) = \frac{exp(o(x,y))}{\sum_{x,y} exp(o(x,y))}.$$





CNN-based saliency prediction (2/9)

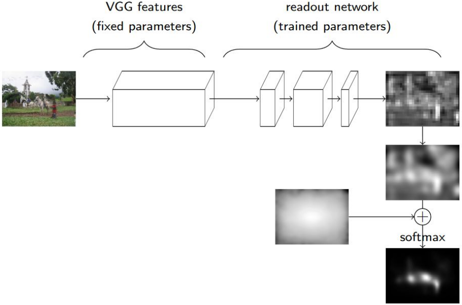
➠ *Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks* (Huang et al., 2015):



- integration of information at different image scales;
- saliency evaluation metrics;
- end-to-end learning.

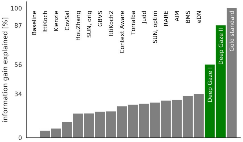
CNN-based saliency prediction (3/9)

- *DeepGaze II: Reading fixations from deep features trained on object recognition* (Kümmerer et al., 2016):



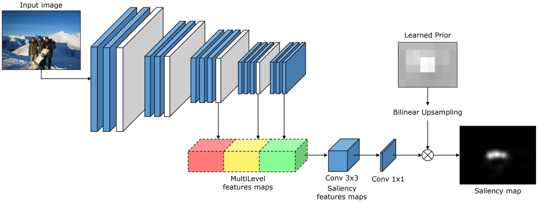
VGG-19 network is now used feature maps from conv5_1, ReLU5_1, ReLU5_2, conv5_3, ReLU5_4;

4 layers of 1×1 convolution + ReLU (second neural network that needs to be trained).



CNN-based saliency prediction (4/9)

- *A Deep Multi-Level Network for Saliency Prediction* (Cornia et al., 2016):

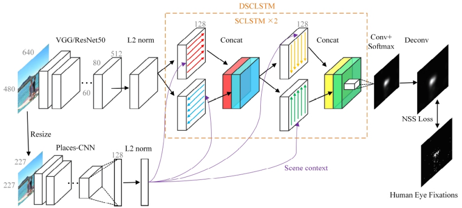


$$\mathcal{L}(S,\hat{S})_{MLNET} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{\alpha - S_j}(S_j - \hat{S}_j)^2, \alpha = 1.1$$

with, $S, \hat{S} \in [0, 1]$

CNN-based saliency prediction (5/9)

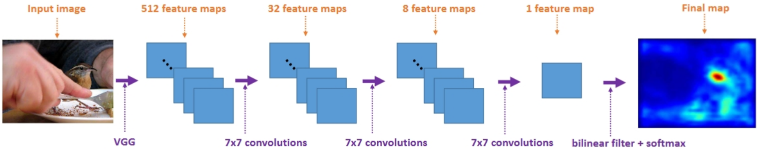
➡ *A Deep Spatial Contextual Long-term Recurrent Convolutional Network for Saliency Detection* (Liu and Han, 2016):



- Local Image Feature Extraction using CNNs (normalize and rescale);
- Scene feature extractor CNN (Places-CNN (Zhou et al., 2014));
- DSCLSTM model incorporates global context information and scene context modulation.

CNN-based saliency prediction (6/9)

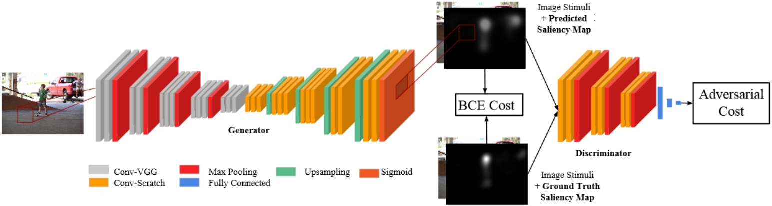
➡ *End-to-End Saliency Mapping via Probability Distribution Prediction* (Jetley et al., 2016):



- VGG Net without the fully-connected layers;
- Three additional convolutional layers + upsampling and softmax.

CNN-based saliency prediction (7/9)

➡ *SalGan: Visual saliency prediction with generative adversarial networks* (Pan et al., 2017):



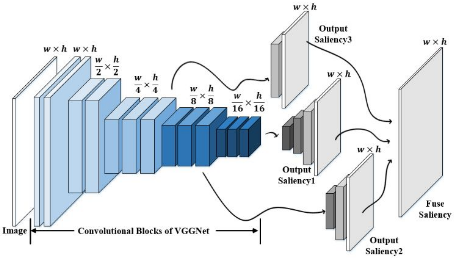
- Training generator (15 epochs), Binary Cross entropy Loss (down-sampled output and ground truth saliency);
- Alternate the training of the saliency prediction network and discriminator network after each iteration (batch).

	sAUC ↑	AUC-B ↑	NSS ↑	CC ↑	IG
MSE	0.728	0.820	1.680	0.708	0.628
BCE	0.753	0.825	2.562	0.772	0.824
BCE/4	0.757	0.833	**2.580**	0.772	1.067
GAN/4	**0.773**	**0.859**	2.560	**0.786**	**1.243**

Table 4. Best results through epochs obtained with non-adversarial (MSE and BCE) and adversarial training. BCE/4 and GAN/4 refer to downsampled saliency maps. Saliency maps assessed on SALICON validation.

CNN-based saliency prediction (8/9)

➠ *Deep visual attention prediction* (Wang and Shen, 2017):



- Encoder - Decoder approach;
- Multi-scale predictions are learned from different layers with different receptive field sizes;
- Fuse saliency thanks to 1×1 convolution layer $(F = \sum_{m=1}^{M} w_f^m S^m)$.

Ablation study:

Aspect	Variant	TORONTO			
		s-AUC ↑	Δs-AUC	CC ↑	ΔCC
	whole model	**0.76**	-	**0.72**	-
submodule	*conv3-3* output	0.68	-0.08	0.57	-0.15
	conv4-3 output	0.69	-0.07	0.65	-0.07
	conv5-3 output	0.69	-0.07	0.69	-0.03
fusion	*avg.* output	0.72	-0.04	0.68	-0.04
supervision	w/o deep supervision	0.71	-0.05	0.68	-0.04
upsampling	bilinear interpolation kernel	0.74	-0.02	0.70	-0.02

CNN-based saliency prediction (9/9)

➡ Snapshot of performance (MIT benchmark, 19^{th} Oct. 2017):

Model Name	Published	Code	AUC-Judd [?]	SIM [?]	EMD [?]	AUC-Borji [?]	sAUC [?]	CC [?]	NSS [?]	KL [?]	Date tested [key]	Sample [img]
Baseline: infinite humans [?]			0.92	1	0	0.88	0.81	1	3.29	0		
Deep Spatial Contextual Long-term Recurrent Convolutional Network (DSCLRCN)	Nian Liu, Junwei Han. A Deep Spatial Contextual Long-term Recurrent Convolutional Network for Saliency Detection [arXiv 2016]		0.87	0.68	2.17	0.79	0.72	0.80	2.35	0.95	first tested: 16/06/2016 last tested: 27/07/2016 maps from authors	
Saliency Attentive Model (SAM-ResNet)	Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model [arXiv 2016]	python	0.87	0.68	2.15	0.78	0.70	0.78	2.34	1.27	first tested: 10/30/2016 last tested: 03/03/2017 maps from authors	
Saliency Attentive Model (SAM-VGG)	Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model [arXiv 2016]	python	0.87	0.67	2.14	0.78	0.71	0.77	2.30	1.13	first tested: 10/30/2016 last tested: 03/03/2017 maps from authors	
DeepFix	Srinivas S S Kruthiventi, Kumar Ayush, R. Venkatesh Babu DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations [arXiv 2015]		0.87	0.67	2.04	0.80	0.71	0.78	2.26	0.63	first tested: 02/10/2015 last tested: 02/10/2015 maps from authors	
DenseSal	Taiki Oyama, Takao Yamanaka		0.87	0.67	1.99	0.81	0.72	0.79	2.25	0.48	first tested: 14/06/2017 last tested: 14/06/2017 maps from authors	
SALICON	Xun Huang, Chengyao Shen, Xavier Boix, Qi Zhao		0.87	0.60	2.62	0.85	0.74	0.74	2.12	0.54	first tested: 19/11/2014 last tested: 15/11/2015 maps from authors	
Probability Distribution Prediction (PDP)	Saumya Jetley, Naila Murray, Eleonora Vig. End-to-End Saliency Mapping via Probability Distribution Prediction [CVPR 2016]		0.85	0.60	2.58	0.80	0.73	0.70	2.05	0.92	first tested: 05/11/2015 last tested: 05/11/2015 maps from authors	
ML-Net	Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, Rita Cucchiara. A Deep Multi-Level Network for Saliency Prediction [ICPR 2016]	Python	0.85	0.59	2.63	0.75	0.70	0.67	2.05	1.10	first tested: 25/01/2016 last tested: 01/09/2016 maps from authors	
SalGAN	Junting Pan, Cristian Canton, Kevin McGuinness, Noel E. Oâ€™Connor, Jordi Torres, Elisa Sayrol and Xavier Giro-i-Nieto. SalGAN: Visual Saliency Prediction with Generative Adversarial Networks [arXiv 2017]	python	0.86	0.63	2.29	0.81	0.72	0.73	2.04	1.07	first tested: 10/30/2016 last tested: 10/30/2016 maps from authors	
Learning Human												

Limitations (1/1)

The picture is much clearer than 10 years ago!
BUT...

Important aspects of our visual system are clearly overlooked

- Current models implicitly assume that eyes are equally likely to move in any direction;
- Viewing biases are not taken into account;
- The temporal dimension is not considered (static saliency map).

Limitations (1/1)

The picture is much clearer than 10 years ago!
BUT...

Important aspects of our visual system are clearly overlooked

- Current models implicitly assume that eyes are equally likely to move in any direction;
- Viewing biases are not taken into account;
- The temporal dimension is not considered (static saliency map).

Limitations (1/1)

The picture is much clearer than 10 years ago!
BUT...

Important aspects of our visual system are clearly overlooked

- Current models implicitly assume that eyes are equally likely to move in any direction;
- Viewing biases are not taken into account;
- The temporal dimension is not considered (static saliency map).

Limitations (1/1)

The picture is much clearer than 10 years ago!
BUT...

Important aspects of our visual system are clearly overlooked

- Current models implicitly assume that eyes are equally likely to move in any direction;
- Viewing biases are not taken into account;
- The temporal dimension is not considered (static saliency map).

Saccadic model

Presentation (1/1)

- Eye movements are composed of fixations and saccades. A sequence of fixations is called a visual scanpath.
- When looking at visual scenes, we perform in average 4 visual fixations per second.

Saccadic models are used:

1. to compute plausible visual scanpaths (stochastic, saccade amplitudes / orientations...);
2. to infer the scanpath-based saliency map ⇔ to predict salient areas!!

Proposed model (1/8)

So, what are the key ingredients to design a saccadic model?

- The model has to be stochastic: the subsequent fixation cannot be completely specified (given a set of data).
- The model has to generate plausible scanpaths that are similar to those generated by humans in similar conditions: distribution of saccade amplitudes and orientations, center bias...
- Inhibition of return has to be considered: time-course, spatial decay...
- Fixations should be mainly located on salient areas.

O. Le Meur & Z. Liu, Saccadic model of eye movements for free-viewing condition, Vision Research, 2015.

O. Le Meur & A. Coutrot, Introducing context-dependent and spatially-variant viewing biases in saccadic models, Vision Research, 2016.

Proposed model (1/8)

So, what are the key ingredients to design a saccadic model?

- The model has to be stochastic: the subsequent fixation cannot be completely specified (given a set of data).
- The model has to generate plausible scanpaths that are similar to those generated by humans in similar conditions: distribution of saccade amplitudes and orientations, center bias...
- Inhibition of return has to be considered: time-course, spatial decay...
- Fixations should be mainly located on salient areas.

O. Le Meur & Z. Liu, Saccadic model of eye movements for free-viewing condition, Vision Research, 2015.
O. Le Meur & A. Coutrot, Introducing context-dependent and spatially-variant viewing biases in saccadic models, Vision Research, 2016.

Proposed model (1/8)

So, what are the key ingredients to design a saccadic model?

- The model has to be stochastic: the subsequent fixation cannot be completely specified (given a set of data).
- The model has to generate plausible scanpaths that are similar to those generated by humans in similar conditions: distribution of saccade amplitudes and orientations, center bias...
- Inhibition of return has to be considered: time-course, spatial decay...
- Fixations should be mainly located on salient areas.

O. Le Meur & Z. Liu, Saccadic model of eye movements for free-viewing condition, Vision Research, 2015.
O. Le Meur & A. Coutrot, Introducing context-dependent and spatially-variant viewing biases in saccadic models, Vision Research, 2016.

Proposed model (1/8)

So, what are the key ingredients to design a saccadic model?

- The model has to be stochastic: the subsequent fixation cannot be completely specified (given a set of data).
- The model has to generate plausible scanpaths that are similar to those generated by humans in similar conditions: distribution of saccade amplitudes and orientations, center bias...
- Inhibition of return has to be considered: time-course, spatial decay...
- Fixations should be mainly located on salient areas.

O. Le Meur & Z. Liu, Saccadic model of eye movements for free-viewing condition, Vision Research, 2015.
O. Le Meur & A. Coutrot, Introducing context-dependent and spatially-variant viewing biases in saccadic models, Vision Research, 2016.

Proposed model (1/8)

So, what are the key ingredients to design a saccadic model?

- The model has to be stochastic: the subsequent fixation cannot be completely specified (given a set of data).
- The model has to generate plausible scanpaths that are similar to those generated by humans in similar conditions: distribution of saccade amplitudes and orientations, center bias...
- Inhibition of return has to be considered: time-course, spatial decay...
- Fixations should be mainly located on salient areas.

O. Le Meur & Z. Liu, Saccadic model of eye movements for free-viewing condition, Vision Research, 2015.
O. Le Meur & A. Coutrot, Introducing context-dependent and spatially-variant viewing biases in saccadic models, Vision Research, 2016.

Proposed model (2/8)

Let $\mathcal{I} : \Omega \subset \mathcal{R}^2 \mapsto \mathcal{R}^3$ an image and $\mathbf{x}_t$ a fixation point at time t.

We consider the 2D discrete conditional probability:

$$\boxed{p\left(\mathbf{x}|\mathbf{x}_{t-1}, S\right) \propto p_{BU}(\mathbf{x}) p_B(d, \phi|F, S) p_M(\mathbf{x}|\mathbf{x}_{t-1})}$$

- $p_{BU} : \Omega \mapsto [0, 1]$ is the grayscale saliency map;
- $p_B(d, \phi|F, S)$ represents the joint probability distribution of saccade amplitudes and orientations.
  - d is the saccade amplitude between two fixation points $\mathbf{x}$ and $\mathbf{x}_{t-1}$ (expressed in degree of visual angle);
  - ϕ is the angle (expressed in degree between these two points);
  - F and S correspond to the frame index and the scene type, respectively.
- $p_M(\mathbf{x}|\mathbf{x}_{t-1})$ represents the memory state of the location $\mathbf{x}$ at time t. This time-dependent term simulates the inhibition of return.

Proposed model (2/8)

Let $\mathcal{I} : \Omega \subset \mathcal{R}^2 \mapsto \mathcal{R}^3$ an image and $\mathbf{x}_t$ a fixation point at time t.

We consider the 2D discrete conditional probability:

$$\boxed{p\left(\mathbf{x}|\mathbf{x}_{t-1}, S\right) \propto p_{BU}(\mathbf{x}) p_B(d, \phi|F, S) p_M(\mathbf{x}|\mathbf{x}_{t-1})}$$

- $p_{BU} : \Omega \mapsto [0, 1]$ is the grayscale saliency map;
- $p_B(d, \phi|F, S)$ represents the joint probability distribution of saccade amplitudes and orientations.
  - d is the saccade amplitude between two fixation points $\mathbf{x}$ and $\mathbf{x}_{t-1}$ (expressed in degree of visual angle);
  - ϕ is the angle (expressed in degree between these two points);
  - F and S correspond to the frame index and the scene type, respectively.
- $p_M(\mathbf{x}|\mathbf{x}_{t-1})$ represents the memory state of the location $\mathbf{x}$ at time t. This time-dependent term simulates the inhibition of return.

Proposed model (2/8)

Let $\mathcal{I} : \Omega \subset \mathcal{R}^2 \mapsto \mathcal{R}^3$ an image and $\mathbf{x}_t$ a fixation point at time t.

We consider the 2D discrete conditional probability:

$$\boxed{p\left(\mathbf{x}|\mathbf{x}_{t-1}, S\right) \propto p_{BU}(\mathbf{x}) p_B(d, \phi|F, S) p_M(\mathbf{x}|\mathbf{x}_{t-1})}$$

- $p_{BU} : \Omega \mapsto [0, 1]$ is the grayscale saliency map;
- $p_B(d, \phi|F, S)$ represents the joint probability distribution of saccade amplitudes and orientations.
  - d is the saccade amplitude between two fixation points $\mathbf{x}$ and $\mathbf{x}_{t-1}$ (expressed in degree of visual angle);
  - ϕ is the angle (expressed in degree between these two points);
  - F and S correspond to the frame index and the scene type, respectively.
- $p_M(\mathbf{x}|\mathbf{x}_{t-1})$ represents the memory state of the location $\mathbf{x}$ at time t. This time-dependent term simulates the inhibition of return.

Proposed model (2/8)

Let $\mathcal{I} : \Omega \subset \mathcal{R}^2 \mapsto \mathcal{R}^3$ an image and $\mathbf{x}_t$ a fixation point at time t.

We consider the 2D discrete conditional probability:

$$\boxed{p\left(\mathbf{x}|\mathbf{x}_{t-1}, S\right) \propto p_{BU}(\mathbf{x}) p_B(d, \phi|F, S) p_M(\mathbf{x}|\mathbf{x}_{t-1})}$$

- $p_{BU} : \Omega \mapsto [0, 1]$ is the grayscale saliency map;
- $p_B(d, \phi|F, S)$ represents the joint probability distribution of saccade amplitudes and orientations.
  - d is the saccade amplitude between two fixation points $\mathbf{x}$ and $\mathbf{x}_{t-1}$ (expressed in degree of visual angle);
  - ϕ is the angle (expressed in degree between these two points);
  - F and S correspond to the frame index and the scene type, respectively.
- $p_M(\mathbf{x}|\mathbf{x}_{t-1})$ represents the memory state of the location $\mathbf{x}$ at time t. This time-dependent term simulates the inhibition of return.

Proposed model (3/8)

Bottom-up saliency map

$$\boxed{p\left(\mathbf{x}|\mathbf{x}_{t-1}, S\right) \propto p_{BU}(\mathbf{x})p_B(d, \phi|F, S)p_M(\mathbf{x}|\mathbf{x}_{t-1})}$$

➡ p_{BU} is the bottom-up saliency map.

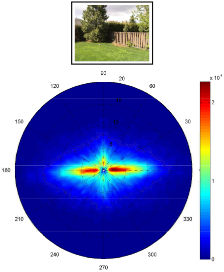
- Computed by GBVS model (Harel et al., 2006). According to (Borji et al., 2012)'s benchmark, this model is among the best ones and presents a good trade-off between quality and complexity.
- $p_{BU}(\mathbf{x})$ is constant over time. (Tatler et al., 2005) indeed demonstrated that bottom-up influences do not vanish over time.

Proposed model (4/8)

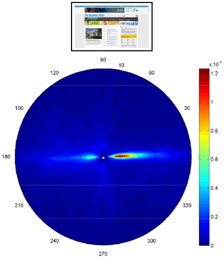
Viewing biases

$$\boxed{p\left(\mathbf{x}|\mathbf{x}_{t-1}, S\right) \propto p_{BU}(\mathbf{x})p_B(d,\phi|F,S)p_M\left(\mathbf{x}|\mathbf{x}_{t-1}\right)}$$

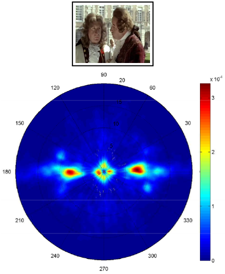
- $p_B(d,\phi|F,S)$ represents the joint probability distribution of saccade amplitudes and orientations $\Rightarrow$ learning from eye-tracking data.
  d and ϕ represent the distance and the angle between successive fixations.



Strong horizontal bias

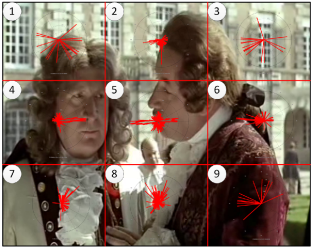


Strong horizontal bias but mainly in the rightward direction.



Three modes in the distribution

Spatially-invariant to spatially-variant and scene-dependent distribution $p_B(d, \phi | F, S)$:
rather than computing a unique joint distribution per image, we evenly divide the image into a $N \times N$ equal base frames.

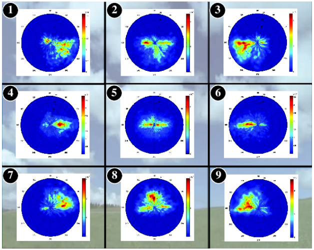


$$N = 3$$

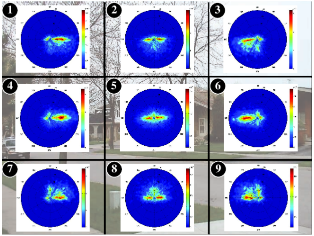
Proposed model (6/8)

Viewing biases

Estimation of the joint distribution $p_B(d, \phi | F, S)$, given the frame index F ($F \in \{1, ..., 9\}$) and the scene category S (Natural scenes, webpages, conversational...):



Dynamic landscape.



Natural scenes.

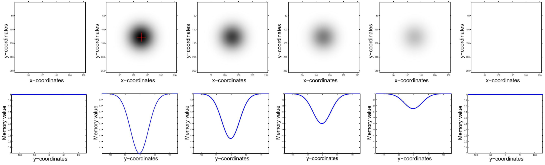
➡ Re-positioning saccades allowing us to go back to the screen's center. Interesting to reproduce the center bias!

Proposed model (7/8)

Memory effect and inhibition of return (IoR)

$$\boxed{p\left(\mathbf{x}|\mathbf{x}_{t-1}, S\right) \propto p_{BU}(\mathbf{x})p_B(d,\phi|F,S)p_M(\mathbf{x}|\mathbf{x}_{t-1})}$$

➡ $p_M(\mathbf{x}|\mathbf{x}_{t-1})$ represents the memory effect and IoR of the location $\mathbf{x}$ at time t. It is composed of two terms: Inhibition and Recovery.



(a) Initialization (b) Inhibition (c) Recovering $\frac{1}{4}$ (d) Recovering $\frac{2}{4}$ (e) Recovering $\frac{3}{4}$ (f) Recovering $\frac{4}{4}$

- The spatial IoR effect declines as a Gaussian function $\Phi_{\sigma_i}(d)$ with the Euclidean distance d from the attended location (Bennett and Pratt, 2001);
- The temporal decline of the IoR effect is simulated by a simple linear model.

Proposed model (8/8)

Selecting the next fixation point

$$\boxed{p\left(\mathbf{x}|\mathbf{x}_{t-1},S\right) \propto p_{BU}(\mathbf{x})p_B(d,\phi|F,S)p_M\left(\mathbf{x}|\mathbf{x}_{t-1}\right)}$$

- Optimal next fixation point (*Bayesian ideal searcher* proposed by (Najemnik and Geisler, 2009)):

$$\mathbf{x}_t^* = \arg\max_{\mathbf{x}\in\Omega} p\left(\mathbf{x}|\mathbf{x}_{t-1}\right) \quad (2)$$

Problem: this approach does not reflect the stochastic behavior of our visual system and may fail to provide plausible scanpaths (Najemnik and Geisler, 2008).

- Rather than selecting the best candidate, we generate $N_c = 5$ random locations according to the 2D discrete conditional probability $p\left(\mathbf{x}|\mathbf{x}_{t-1}\right)$.
The location with the highest saliency is chosen as the next fixation point $\mathbf{x}_t^*$.

Proposed model (8/8)

Selecting the next fixation point

$$\boxed{p\left(\mathbf{x}|\mathbf{x}_{t-1}, S\right) \propto p_{BU}(\mathbf{x}) p_B(d, \phi|F, S) p_M\left(\mathbf{x}|\mathbf{x}_{t-1}\right)}$$

- Optimal next fixation point (*Bayesian ideal searcher* proposed by (Najemnik and Geisler, 2009)):

$$\mathbf{x}_t^* = \arg\max_{\mathbf{x}\in\Omega} p\left(\mathbf{x}|\mathbf{x}_{t-1}\right) \quad (2)$$

Problem: this approach does not reflect the stochastic behavior of our visual system and may fail to provide plausible scanpaths (Najemnik and Geisler, 2008).

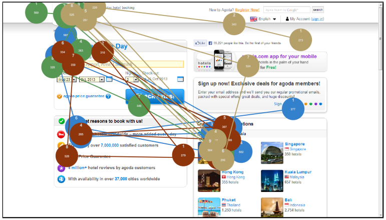
- Rather than selecting the best candidate, we generate $N_c = 5$ random locations according to the 2D discrete conditional probability $p\left(\mathbf{x}|\mathbf{x}_{t-1}\right)$.
The location with the highest saliency is chosen as the next fixation point $\mathbf{x}_t^*$.

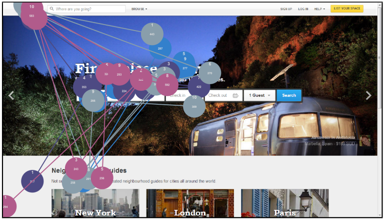
Results (1/5)

The relevance of the proposed approach is assessed with regard to **the plausibility**, **the spatial precision** of the simulated scanpath and ability **to predict saliency areas**.

- Do the generated scanpaths present the same oculomotor biases as human scanpaths?
- What is the similarity degree between predicted and human scanpaths?
- Could the predicted scanpaths be used to form relevant saliency maps?

Results (2/5)


Sign up now! Exclusive deals for agoda members!
Singapore
Hong Kong
Kuala Lumpur
Phuket
Bali

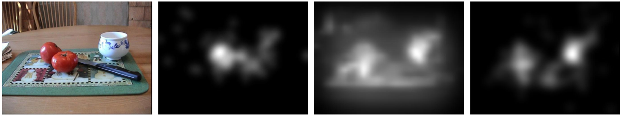

Search
New York
London
Paris

Results (3/5)

Scanpath-based saliency map

- We compute, for each image, 20 scanpaths, each composed of 10 fixations.

- For each image, we created a saliency map by convolving a Gaussian function over the fixation locations.

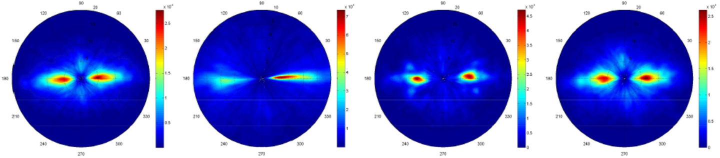


(a) (b) (c) (d)

(a) original image; (b) human saliency map; (c) GBVS saliency map; (d) GBVS-SM saliency maps computed from the simulated scanpaths.

Results (4/5)

Are the predicted scanpaths similar to human ones?



(a) Natural scenes (b) Static webpages (c) Conversational video (d) Dynamic landscapes

Figure 11: Joint distribution of predicted scanpaths shown on polar plot for (a) Natural scenes, (b) Webpages, (c) conversational video and (d) dynamic landscapes. Scanpaths are generated by the context-dependent saccadic saliency model (Top2(R+H), $N = 3$).

Yes, predicted scanpaths show similar patterns as the human scanpaths!

Results (5/5)

Mixing together bottom-up saliency and viewing biases.

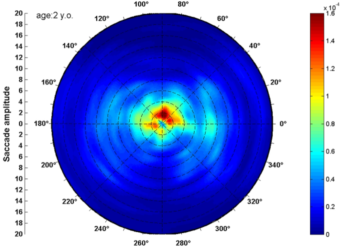
	Metric	CC	SIM	EMD
(B) Bottom-up features alone	State-of-the-art saliency models			
	(Itti et al., 1998)	0.27±0.18	0.37±0.05	3.41±0.65
	(Le Meur et al., 2006)	0.38±0.20	0.43±0.09	3.03±1.06
	(Harel et al., 2006)	0.56±0.14	0.48±0.05	2.49±0.53
	(Bruce & Tsotsos, 2009)	0.31±0.10	0.37±0.04	3.44±0.56
	(Judd et al., 2009)	0.42±0.13	0.40±0.04	3.25±0.57
	(Garcia-Diaz et al., 2012)	0.42±0.18	0.43±0.06	3.30±0.76
	(Riche et al., 2013)	0.54±0.18	0.48±0.06	2.61±0.71
	Top 2 models combined: (Riche et al., 2013) + (Harel et al., 2006)			
	Top2(R+H)	0.62±0.13	0.514±0.05	2.282±0.56
Combining (V) and (B)	Saccadic saliency model (Top2(R+H)) context-independent, $N = 1$			
	(Le Meur & Liu, 2015)	0.641±0.18	0.568±0.09	2.03±0.85
	Saccadic saliency model (Top2(R+H)) context-dependent, $N = 3$			
	Natural scenes	0.649±0.18	0.566±0.09	2.068±0.84
	Webpages	0.641±0.18	0.561±0.09	2.177±0.88
	Conversational	0.628±0.17	0.561±0.09	2.061±0.84
	Landscapes	0.653±0.17	0.571±0.08	2.034±0.85

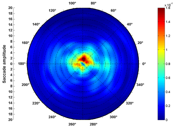
Table 2: Performance (average ± standard deviation) of saliency models over Bruce's dataset. In pink cells, we compare state-of-the-art saliency maps with human saliency maps. We add the top 2 models ((Riche et al., 2013) + (Harel et al., 2006)) into a single bottom-up model: Top2(R+H). In green cells, we compare the performances when low-level visual features from Top2(R+H) and viewing biases are combined. First, we assess the context-independent saccadic model based on a single distribution (N=1) from (Le Meur & Liu, 2015). Second, we assess our context-dependent saccadic model based on 9 distributions (N=3), with viewing biases estimated from 4 categories (Natural Scenes, Webpages, Conversational videos and Landscape videos). Three metrics are used: CC (linear correlation), SIM (histogram similarity) and EMD (Earth Mover's Distance). For more details please refer to the text.

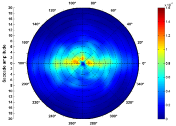
(i) When the quality of the input saliency map increases, performance of saccadic model increases;
(ii) The gain brought by spatially-variant and context-dependent distributions is not significant;
(iii) Spatially-variant and context-dependent distributions are required to generate plausible visual scanpaths (see previous slides).

Tailoring the model for different contexts!

- Task-dependent saccadic model (free-viewing vs quality task...)
- Age-dependent saccadic model.... (2 y.o., 4-6 y.o., 6-10 y.o, adults) (Helo et al., 2014)







Le Meur et al., Visual attention saccadic models learn to emulate gaze patterns from childhood to adulthood, IEEE Trans. Image Processing, 2017.

Limitations

Still far from the reality...

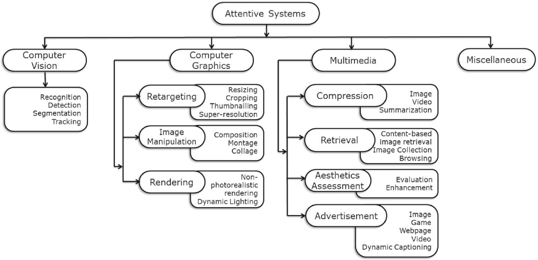
- We do not predict the fixation durations. Some models could be used for this purpose (Nuthmann et al., 2010, Trukenbrod and Engbert, 2014).
- Second-order effect. We assume that the memory effect occurs only in the fixation location. However, are saccades independent events? No, see (Tatler and Vincent, 2008).
- High-level aspects such as the scene context are not included in our model.
- Should we recompute the saliency map after every fixations? Probably yes...
- Randomness (N_c) should be adapted to the input image. By default, $N_c = 5$.
- Is the time course of IoR relevant? Is the recovery linear?
- Foveal vs peripheral vision? Cortical magnification...

Attentive applications

6. Attentive applications
   - Taxonomy
   - Saliency-based applications
   - Eye Movements-based applications

Taxonomy

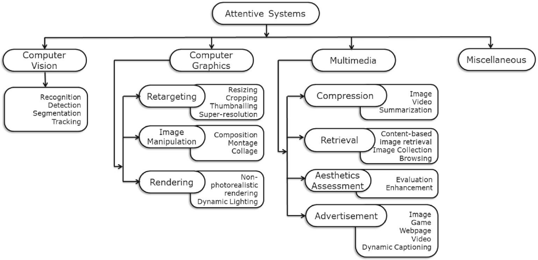
- A sheer number of saliency-based applications....



Extracted from (Nguyen et al., 2017). See also (Mancas et al., 2016).

Taxonomy

- A sheer number of saliency-based applications....

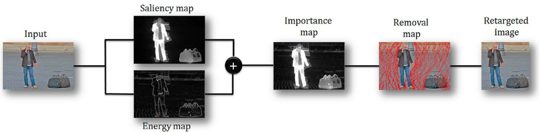


Extracted from (Nguyen et al., 2017). See also (Mancas et al., 2016).

- More and more eye-movements-based applications...

Saliency-based applications (1/2)

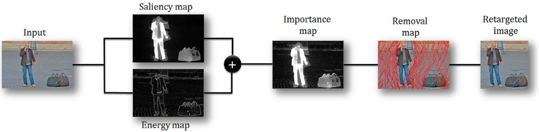
- Saliency-based seam carving (Avidan and Shamir, 2007):



Extracted from (Nguyen et al., 2017).

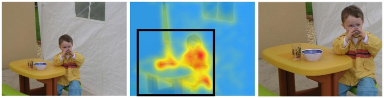
Saliency-based applications (1/2)

- Saliency-based seam carving (Avidan and Shamir, 2007):



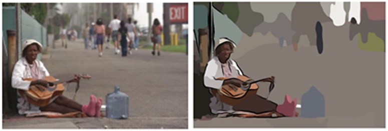
Extracted from (Nguyen et al., 2017).

- Retargeting (Le Meur et al., 2006):



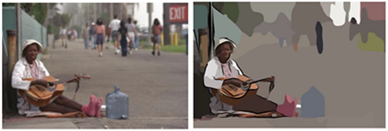
Saliency-based applications (2/2)

- Non photorealistic rendering (DeCarlo and Santella, 2002):

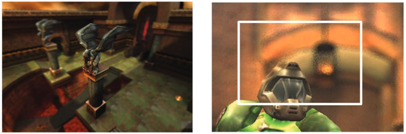


Saliency-based applications (2/2)

- Non photorealistic rendering (DeCarlo and Santella, 2002):



- First-Person Navigation in Virtual Environments (Hillaire et al., 2008):



Eye Movements-based applications (1/3)

- Predicting Moves-on-Stills for Comic Art using Viewer Gaze Data (Jain et al., 2016)

The Ken Burns effect is a type of panning and zooming effect used in video production from still imagery.

More results on http://jainlab.cise.ufl.edu/comics.html

Eye Movements-based applications (2/3)

- Gaze-driven Video Re-editing (Jain et al., 2015)

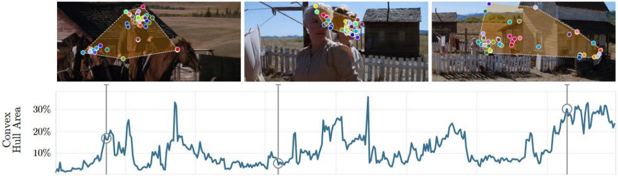
We record gaze data from viewers on the original widescreen video. Each viewer is marked in a different color.

A cut from the woman's face to the man's face.

The cropping window pans to the left while zooming in.

Eye Movements-based applications (3/3)

- Gaze Data for the Analysis of Attention in Feature Films (Breeden and Hanrahan, 2017)



Smaller values indicate increased attentional synchrony.

Conclusion

7 Conclusion

Conclusion (1/2)

Take Home message:

- Saliency model $\Rightarrow$ 2D saliency map;
- Saccadic model $\Rightarrow$
  - to produce plausible visual scanpaths;
  - to detect the most salient regions of visual scenes.
  - can be tailored to specific visual context.
- A number of saliency-based / eye-movements-based applications.

Conclusion (1/2)

Take Home message:

- Saliency model $\Rightarrow$ 2D saliency map;
- Saccadic model $\Rightarrow$
  - to produce plausible visual scanpaths;
  - to detect the most salient regions of visual scenes.
  - can be tailored to specific visual context.
- A number of saliency-based / eye-movements-based applications.

Conclusion (1/2)

Take Home message:

- Saliency model $\Rightarrow$ 2D saliency map;
- Saccadic model $\Rightarrow$
  - to produce plausible visual scanpaths;
  - to detect the most salient regions of visual scenes.
  - can be tailored to specific visual context.
- A number of saliency-based / eye-movements-based applications.

Conclusion (2/2)

Advanced DIP

T. Maugey

➡ Eye-movements revolution...

- Diagnosis of neurodevelopmental disorders (see Itti, L. (2015). *New Eye-Tracking Techniques May Revolutionize Mental Health Screening*. Neuron, 88(3), 442-444.);
- Learning Visual Attention to Identify People With Autism Spectrum Disorder (Jiang and Zhao, 2017);
- Alzheimer's disease (Crawford et al., 2015);
- US startup proposes a device for tracking your eyes to see if you're lying...;
- Emotion, gender (Coutrot et al., 2016), age (Le Meur et al., 2017)....

Conclusion (2/2)

➡ Eye-movements revolution...

- Diagnosis of neurodevelopmental disorders (see Itti, L. (2015). *New Eye-Tracking Techniques May Revolutionize Mental Health Screening*. Neuron, 88(3), 442-444.);
- Learning Visual Attention to Identify People With Autism Spectrum Disorder (Jiang and Zhao, 2017);
- Alzheimer's disease (Crawford et al., 2015);
- US startup proposes a device for tracking your eyes to see if you're lying...;
- Emotion, gender (Coutrot et al., 2016), age (Le Meur et al., 2017)....

Conclusion (2/2)

➡ Eye-movements revolution...

- Diagnosis of neurodevelopmental disorders (see Itti, L. (2015). *New Eye-Tracking Techniques May Revolutionize Mental Health Screening*. Neuron, 88(3), 442-444.);
- Learning Visual Attention to Identify People With Autism Spectrum Disorder (Jiang and Zhao, 2017);
- Alzheimer's disease (Crawford et al., 2015);
- US startup proposes a device for tracking your eyes to see if you're lying...;
- Emotion, gender (Coutrot et al., 2016), age (Le Meur et al., 2017)....

Conclusion (2/2)

➡ Eye-movements revolution...

- Diagnosis of neurodevelopmental disorders (see Itti, L. (2015). *New Eye-Tracking Techniques May Revolutionize Mental Health Screening*. Neuron, 88(3), 442-444.);
- Learning Visual Attention to Identify People With Autism Spectrum Disorder (Jiang and Zhao, 2017);
- Alzheimer's disease (Crawford et al., 2015);
- US startup proposes a device for tracking your eyes to see if you're lying...;
- Emotion, gender (Coutrot et al., 2016), age (Le Meur et al., 2017)....

Conclusion (2/2)

➡ Eye-movements revolution...

- Diagnosis of neurodevelopmental disorders (see Itti, L. (2015). *New Eye-Tracking Techniques May Revolutionize Mental Health Screening*. Neuron, 88(3), 442-444.);
- Learning Visual Attention to Identify People With Autism Spectrum Disorder (Jiang and Zhao, 2017);
- Alzheimer's disease (Crawford et al., 2015);
- US startup proposes a device for tracking your eyes to see if you're lying...;
- Emotion, gender (Coutrot et al., 2016), age (Le Meur et al., 2017)....

References

S. Avidan and A. Shamir. Seam carving for content-aware image resizing. In *ACM SIGGRAPH*, volume 26, 2007. 81, 82

P. J. Bennett and J. Pratt. The spatial distribution of inhibition of return:. *Psychological Science*, 12:76–80, 2001. 68

A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35: 185–207, 2013. 16, 17, 20

A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, 2012. 64

Katherine Breeden and Pat Hanrahan. Gaze data for the analysis of attention in feature films. *ACM Transactions on Applied Perception*, 1:1–14, 2017. 87

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A Deep Multi-Level Network for Saliency Prediction. In *International Conference on Pattern Recognition (ICPR)*, 2016. 43

Antoine Coutrot, Nicola Binetti, Charlotte Harrison, Isabelle Mareschal, and Alan Johnston. Face exploration dynamics differentiate men and women. *Journal of vision*, 16(14):16–16, 2016. 92, 93, 94, 95, 96

Trevor J Crawford, Alex Devereaux, Steve Higham, and Claire Kelly. The disengagement of visual attention in alzheimer's disease: a longitudinal eye-tracking study. *Frontiers in aging neuroscience*, 7, 2015. 92, 93, 94, 95, 96

Doug DeCarlo and Anthony Santella. Stylization and abstraction of photographs. In *ACM transactions on graphics (TOG)*, volume 21, pages 769–776. ACM, 2002. 83, 84

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 38

J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2006. 64

A. Helo, S. Pannasch, L. Sirri, and P. Rama. The maturation of eye movement behavior: scene viewing characteristics in children and adults. *Vision Research*, 103:83–91, 2014. 76

Michael H Herzog and Aaron M Clarke. Why vision is not both hierarchical and feedforward. *Frontiers in computational neuroscience*, 8, 2014. 11, 12

Sébastien Hillaire, Anatole Lécuyer, Rémi Cozot, and Géry Casiez. Depth-of-field blur effects for first-person navigation in virtual environments. *IEEE computer graphics and applications*, 28(6), 2008. 84

Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 262–270, 2015. 41

Laurent Itti and Pierre F Baldi. Bayesian surprise attracts human attention. In *Advances in neural information processing systems*, pages 547–554, 2005. 32

Eakta Jain, Yaser Sheikh, Ariel Shamir, and Jessica Hodgins. Gaze-driven video re-editing. *ACM Transactions on Graphics (TOG)*, 34(2):21, 2015. 86

Eakta Jain, Yaser Sheikh, and Jessica Hodgins. Predicting moves-on-stills for comic art using viewer gaze data. *IEEE computer graphics and applications*, 36(4):34–45, 2016. 85

Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-end saliency mapping via probability distribution prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5753–5761, 2016. 45

Ming Jiang and Qi Zhao. Learning visual attention to identify people with autism spectrum disorder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3267–3276, 2017. 92, 93, 94, 95, 96

T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where people look. In *ICCV*, 2009. 33

T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixation. Technical report, MIT, 2012. 26

C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4: 219–227, 1985. 14

Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014. 40

Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016. 42

O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model the bottom-up visual attention. *IEEE Trans. On PAMI*, 28(5):802–817, May 2006. 21, 22, 82

Olivier Le Meur, Antoine Coutrot, Zhi Liu, Pia Rämä, Adrien Le Roch, and Andrea Helo. Your gaze betrays your age. In *EUSIPCO*, 2017. 92, 93, 94, 95, 96

Nian Liu and Junwei Han. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *arXiv preprint arXiv:1610.01708*, 2016. 44

Mauro Manassi, Bilge Sayim, and Michael H Herzog. When crowding of crowding leads to uncrowdingshort title?? *Journal of Vision*, 13(13):10–10, 2013. 11, 12

Matei Mancas, Vincent P Ferrera, Nicolas Riche, and John G Taylor. *From Human Attention to Computational Attention: A Multidisciplinary Approach*, volume 10. Springer, 2016. 79, 80

J. Najemnik and W.S. Geisler. Eye movement statistics in humans are consistent with an optimal strategy. *Journal of Vision*, 8(3): 1–14, 2008. 69, 70

J. Najemnik and W.S. Geisler. Simple summation rule for optimal fixation selection in visual search. *Vision Research*, 42: 1286–1294, 2009. 69, 70

Tam V Nguyen, Qi Zhao, and Shuicheng Yan. Attentive systems: A survey. *International Journal of Computer Vision*, pages 1–25, 2017. 79, 80, 81, 82

A. Nuthmann, T. J. Smith, R. Engbert, and J. M. Henderson. CRISP: A Computational Model of Fixation Durations in Scene Viewing. *Psychological Review*, 117(2):382–405, April 2010. URL http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ884784. 77

Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017. 46

D. Parkhurst, K. Law, and E. Niebur. Modelling the role of salience in the allocation of overt visual attention. *Vision Research*, 42: 107–123, 2002. 32, 34

R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18): 2397–2416, 2005. 32, 34

M. I. Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32:3–25, 1980. 6

N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit. Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, 28(6):642 – 658, 2013. ISSN 0923-5965. doi: http://dx.doi.org/10.1016/j.image.2013.03.009. 18, 19

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2014. 39

B.W. Tatler and B.T. Vincent. Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, 2:1–18, 2008. 77

B.W. Tatler, R. J. Baddeley, and I.D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45:643–659, 2005. 64

Hans A Trukenbrod and Ralf Engbert. Icat: A computational model for the adaptive control of fixation durations. *Psychonomic bulletin & review*, 21(4):907–934, 2014. 77

Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *arXiv preprint arXiv:1705.02544*, 2017. 47

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 44