VML

MAC(H)INNES CAN SEE

Models of fixation

Sofia Krasovskaya

MODELS OF VISUAL ATTENTION

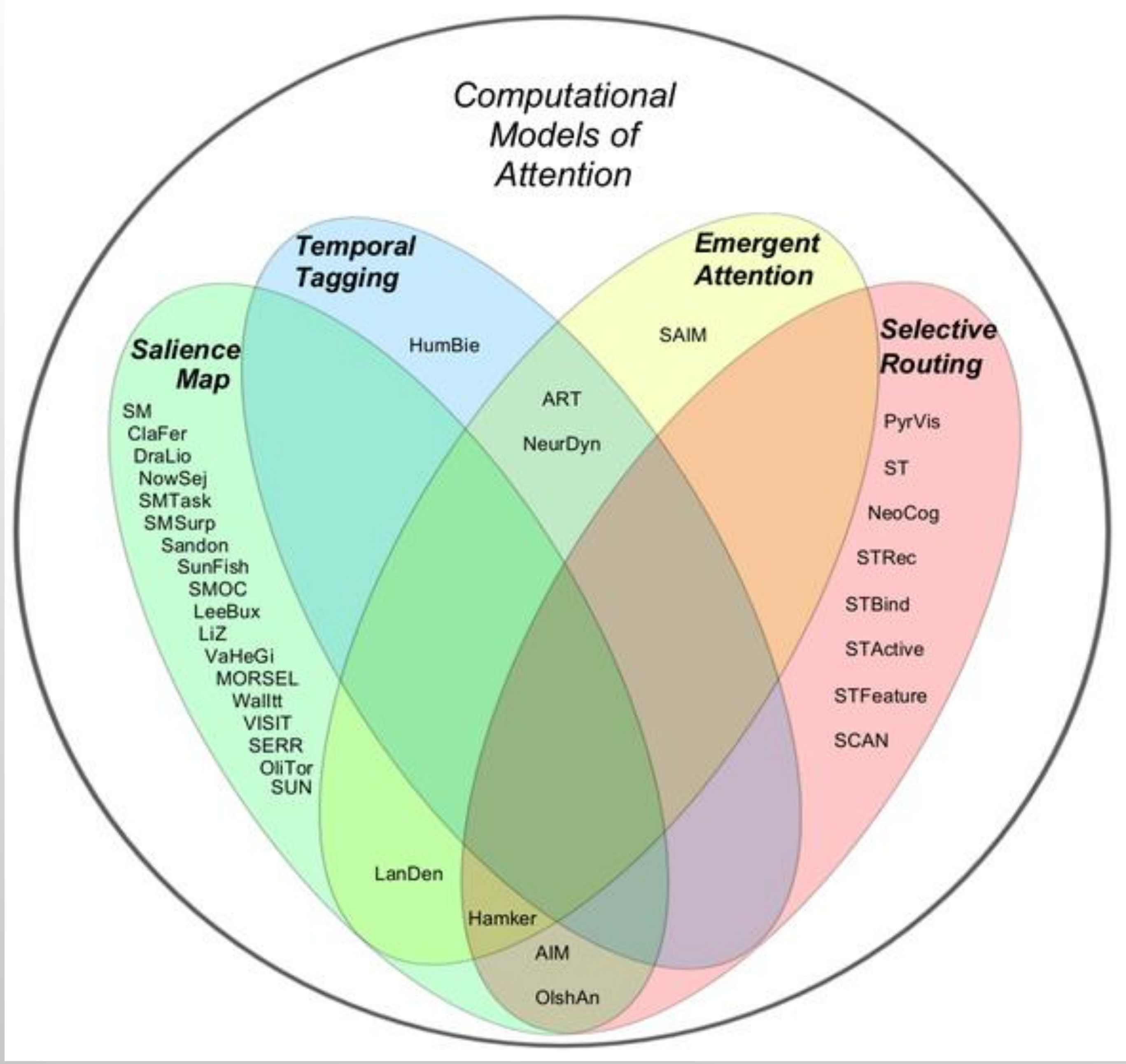
- Describe the observed or predicted behavior of human and primate visual attention.
- Employ natural language, system block diagrams, mathematics, algorithms or computations as their embodiment
- Attempt to mimic, explain or predict visual attentive behavior.
- Must be tested by experiments to assess their predictive validity.

John K. Tsotsos and Albert Rothenstein (2011)

COMPUTATIONAL MODELS OF VISUAL ATTENTION

- An instance of a model of visual attention
- Attempt to explain how attention is computed and can be tested by providing image inputs and by comparing model performance with human performance.
- Mathematical equations solved and/or simulated via computer + Marr's levels of analysis (Marr 1982): the computational level (a formal statement of the problems that must overcome), the algorithmic level (the strategy that may be used), and the implementation level (how the task is actually performed in the brain.
- Computational models attempt to explain complex processes with the help of computers. Allow us to simulate increasingly complex brain functionality and cognitive processes

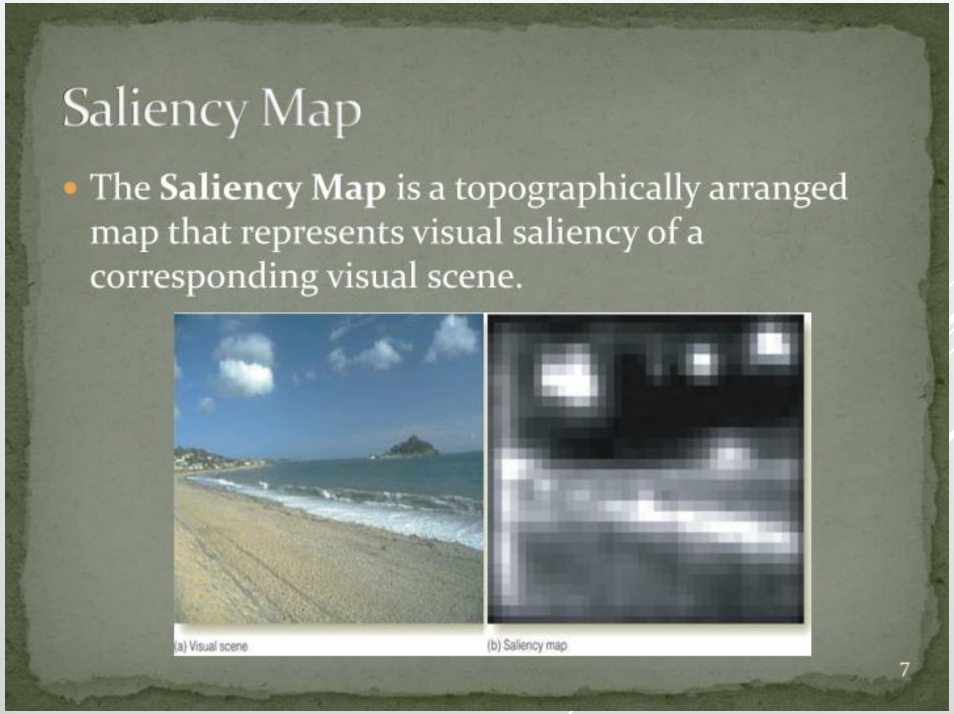
John K. Tsotsos and Albert Rothenstein (2011)



Source: http://www.scholarpedia.org/w/images/7/7d/Cmoa.png

SALIENCY MAP

- It is a topographically arranged map that represents visual saliency of a corresponding visual scene (Niebur, 2007).
- Help overcome information overload.
- highly influenced by phenomena such as ‘*top-down*’ and ‘*bottom-up*’ factors.
- Bottom-up: based solely on visual input. Include characteristics such as size, colour, orientation etc.
- Top-down: gaze is attracted to locations which are task-relevant.



Saliency Map

- The **Saliency Map** is a topographically arranged map that represents visual saliency of a corresponding visual scene.

(a) Visual scene

(b) Saliency map

7

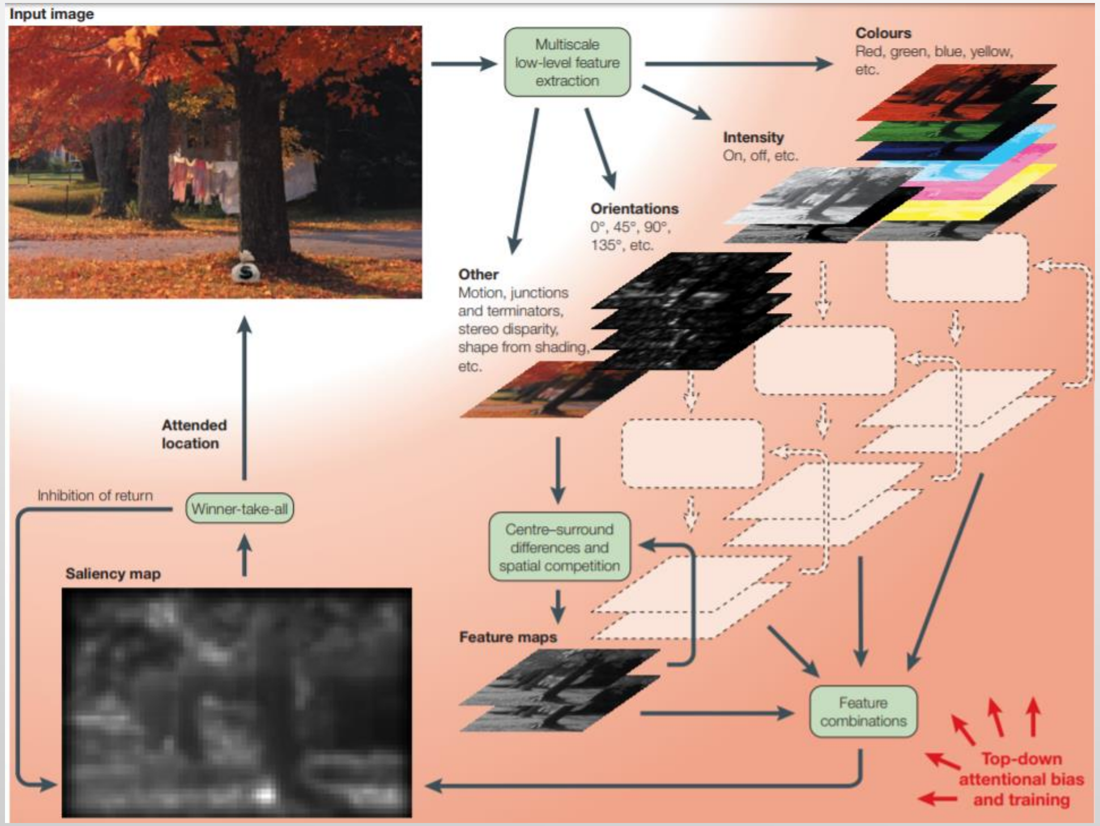
ITTI & KOCH MODEL OF BOTTOM-UP VISUAL ATTENTION

- **Laurent Itti & Christof Koch** (2000), Model of saliency in visual attention: biological aspect (pyramidal cells in centre-surround receptive fields + feature integration theory (Treisman & Gelade, 1980), attention - multiple object feature recognition and integration.
- One of the most influential models of the last 20 years
  - Vision in particular, but well known in all subfields
- Why?
  - Grounded in theory
  - Neurally plausible
  - Testable
  - Controversial? MANY attempts to improve, but not to challenge. Launched many studies that contributed to the understanding of layers of vision and the sphere of visual attention.

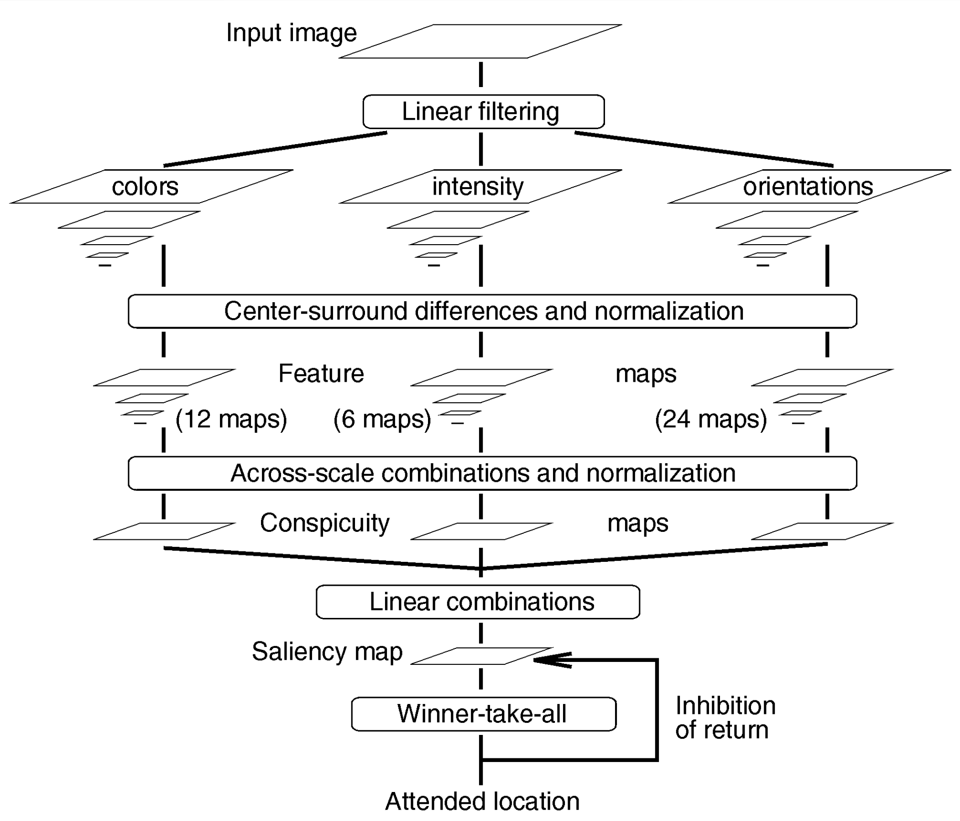
FRAMEWORK FOR A COMPUTATIONAL AND NEUROBIOLOGICAL UNDERSTANDING OF VISUAL ATTENTION, ACCORDING TO ITTI & KOCH (2000):

1) The perceptual saliency of stimuli critically depends on the surrounding context.
2) A unique 'saliency map' is an efficient and plausible bottom-up control strategy.
3) IOR is a crucial element of attentional deployment.
4) Attention and eye movements tightly interplay, posing computational challenges with respect to the coordinate system used to control attention.
5) Scene understanding and object recognition strongly constrain the selection of attended locations.

OVERVIEW


Input image
Multiscale low-level feature extraction
Colours
Red, green, blue, yellow, etc.
Intensity
On, off, etc.
Orientations
0°, 45°, 90°, 135°, etc.
Other
Motion, junctions and terminators, stereo disparity, shape from shading, etc.
Attended location
Inhibition of return
Winner-take-all
Saliency map
Centre-surround differences and spatial competition
Feature maps
Feature combinations
Top-down attentional bias and training

Koch & Ullman



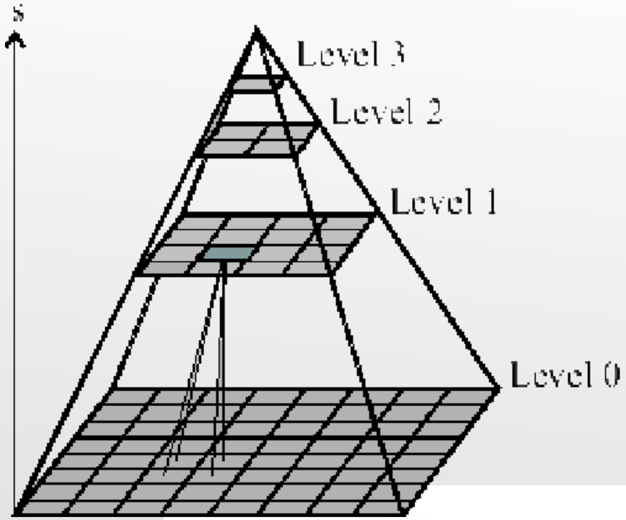
Itti & Koch

1) The input image is decomposed through several pre-attentive feature detection mechanisms (sensitive to colour, intensity and so on), which operate in parallel over the entire visual scene.
2) Neurons feature maps encode for spatial contrast in each of the feature channels. Neurons in each feature map spatially compete for salience.
3) The feature maps are combined into a unique saliency map, which topographically encodes for saliency irrespective of the feature channel in which stimuli appeared salient.
4) The SM is sequentially scanned by attention through the interplay between a winner-take-all algorithm and IOR
5) Top-down attentional bias and training can modulate most stages of this bottom-up model (red shading and arrows).

DYADIC GAUSSIAN PYRAMID

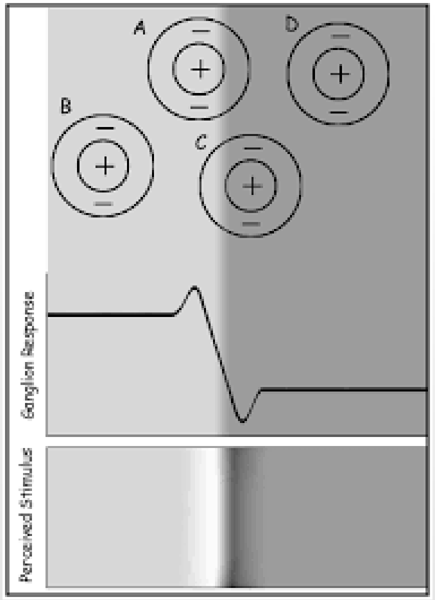
- Input: static color images, 640 x 480
  - 1. low pass filter (blur)
  - 2. Downsample (keep only every nth pixel)

- This is done with nine spatial scales from 1:1 (no downsampling) to 2^8 (1:256) using Dyadic Gaussian Pyramids
- Massively parallel



EXTRACTION OF FEATURES

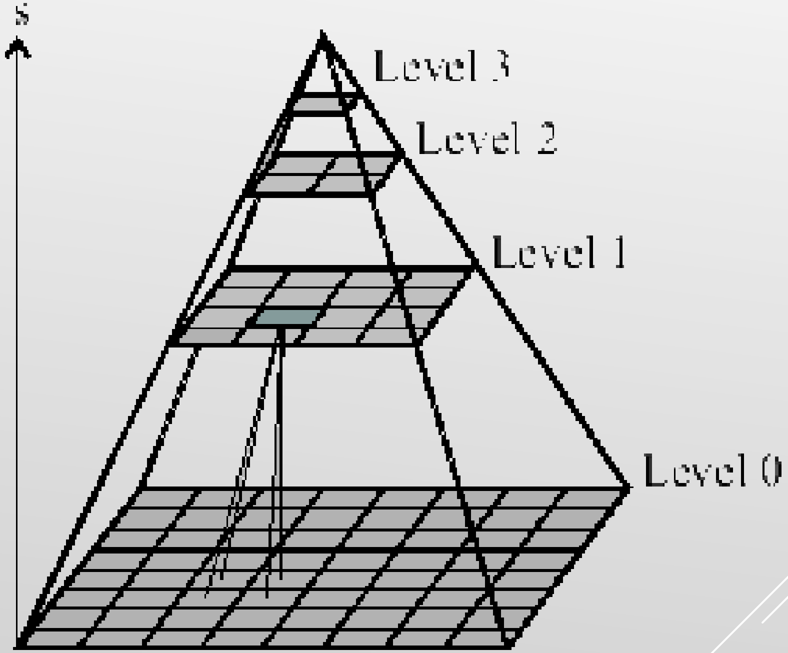
- Each feature is computed by a set of linear “center-surround” operations similar to visual receptive fields
- Typical visual neurons are most sensitive in a small region of the visual space (the center), while stimuli presented in a broader, weaker antagonistic region concentric with the center (the surround) inhibit the neuronal response.
- Center-surround is implemented in the model as the difference between fine and coarse scales.
- Increased smoothing and downsampling reduces edges
- Each pixel on scale n contains a local average that corresponds to an entire pixel **neighbourhood** on scale n + x of the pyramid.



E.g: groups of retinal ganglion cells

FEATURE EXTRACTION (2)

- Centre(C) defined as pixel at scale $c \in \{2, 3, 4\}$
- Surround defined as pixel at scale $s = c + \delta$, *with* $\delta \in \{3, 4\}$
- So that pixel includes averages of many surrounding pixels from the lower scale
- 'Multiscale' feature extraction improved by including different size ratios between the center and surround region (opposed to single scale used previously)

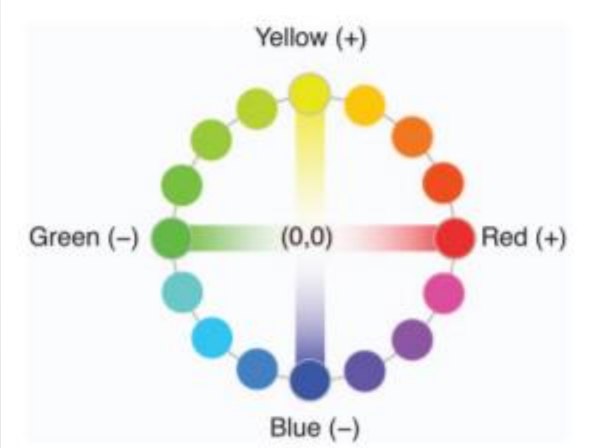


FEATURE MAPS: INTENSITY *(I)*

- Image pixels defined as 0..255 (8 bit) per colour
- Intensity for image in a colour image is *I = (R+G+B)/3*
- 6 maps
  - Centre *(c)* chosen from scale *2,3,4*
  - Surround chosen from *c + 2 or 3*
- Intensity contrast: maps combine light centre with dark surround and dark centre with light surround
  - Both types of sensitivities computed simultaneously in a set of 6 maps then combined (rectified)

FEATURE MAPS: COLOUR

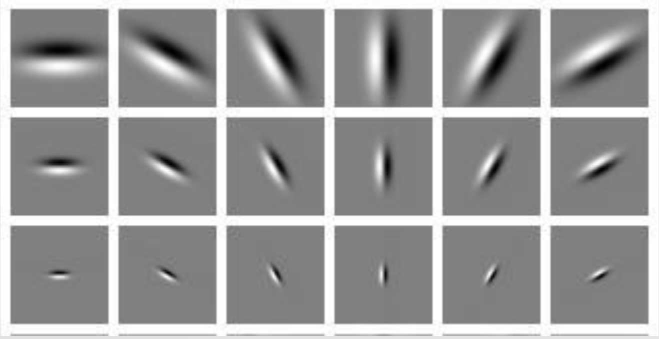
- RGB channels are normalized by the intensity channel to decouple hue from I
    - Otherwise colour maps would hue + I, and we already have separate intensity maps
- Colour values less than 1/10th of max over entire image were dropped
    - Hue differences at low luminance are not perceivable (not salient)
- *Colour double-opponent system*
    - Centre activated by one colour and surround inhibited by a second
    - Implemented G/R, R/G, B/Y and Y/B
- 12 maps in total



Steven K. Shevell and Paul R. Martin, "Color opponency: tutorial," J. Opt. Soc. Am. A 34, 1099-1108 (2017)

FEATURE MAPS: ORIENTATION

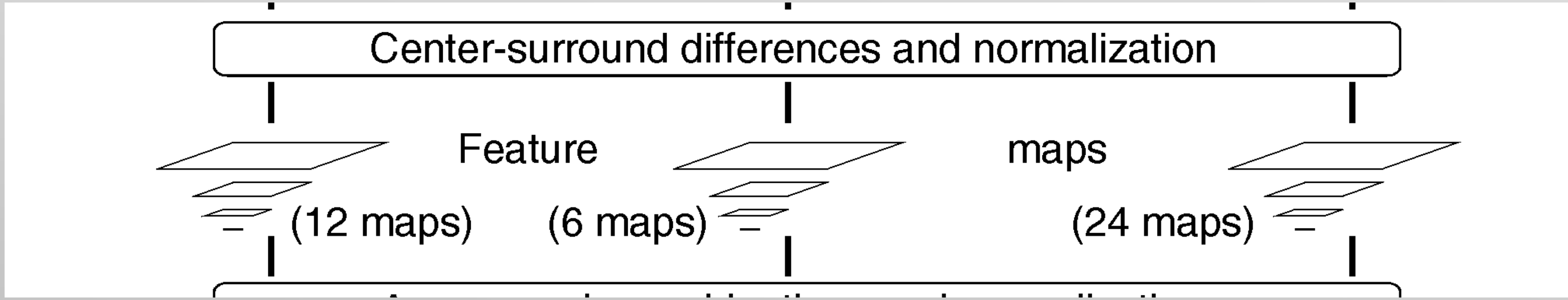
- Location orientation obtained from I using Gabor filters: mimic orientation selective neurons in primary visual cortex
  - (Maybe up to V4 which interprets colour and form)
  - 0, 45, 90 & 135 degrees
  - Orientation feature maps encode, as a group, local orientation contrast between the *c* and *s* scales
  - 24 orientation maps



FEATURE MAP: SUMMARY

- 6 intensity maps
  - 3 centres x 2 surrounds
- 12 normalized colour maps
  - 3 centres x 2 surrounds x 2 colour dyads
- 24 orientation maps
  - 3 centres x 2 surrounds x 4 orientations

42 FM’s in total



NORMALIZATION

- Problem in combining different feature maps: not comparable modalities, with different dynamic ranges and extraction mechanisms.

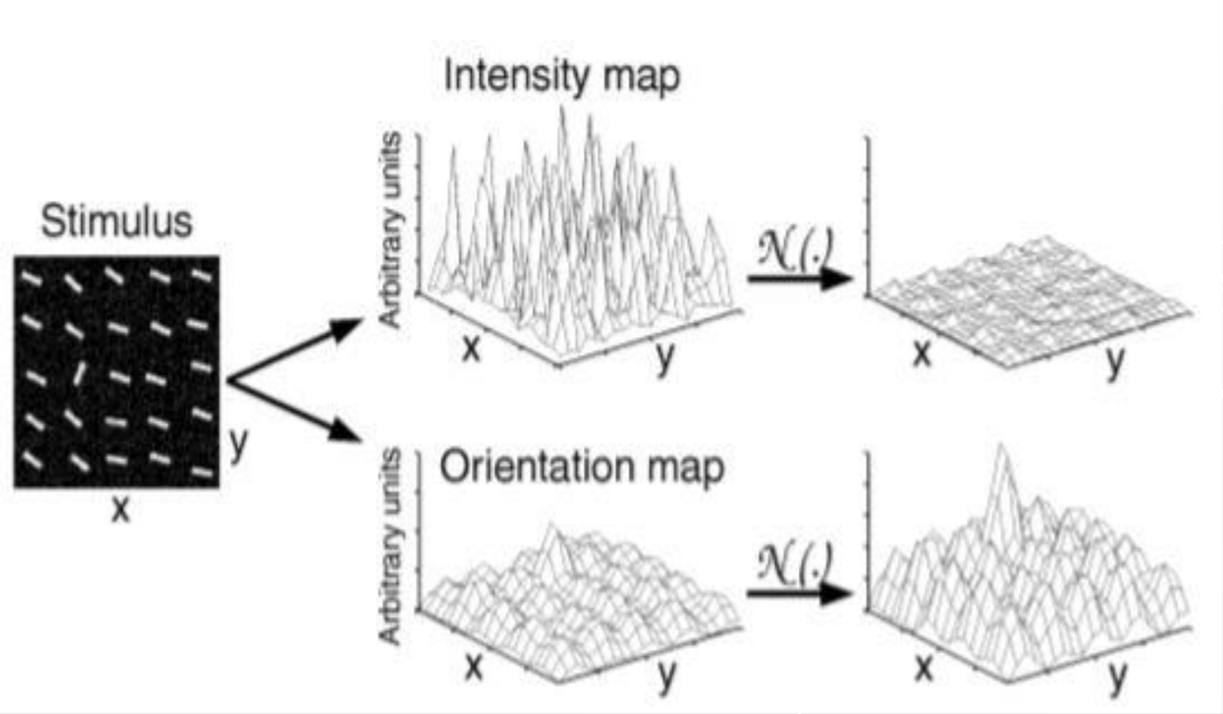
- Due to the combination of 42 maps, salient objects appearing strongly in only a few maps may be masked by noise or by less-salient objects present in a larger number of maps.

- => map normalization operator, *N*(.): globally promotes maps with a small number of strong peaks of activity (conspicuous locations); globally suppresses maps with numerous comparable peak responses.

- Coarse replication of biological cortical lateral inhibition mechanisms, in which neighboring similar features inhibit each other

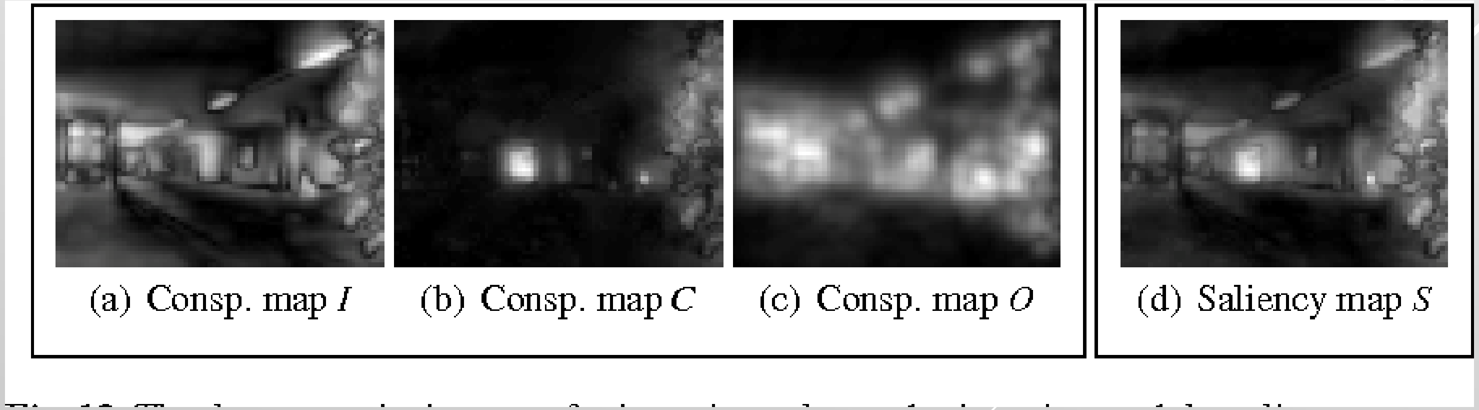
NORMALIZATION *N(.)*

- 1) normalizing the values in the map to a fixed range [0..M], in order to eliminate modality-dependent amplitude differences;
- 2) finding the location of the map's global maximum M and computing the average m of all its other local maxima;
- 3) globally multiplying the map by $(M – m)^2$



CONSPICUITY MAPS

- **FM's are combined into 3 CM's: $\bar{I}$, $\bar{O}$ & $\hat{C}$**
  - Saliency competition is strong within features, while different modalities contribute independently to the saliency map
- **For each feature**
  - Downsample all maps to scale 4
  - Add all points to get total map
  - Orientation combines within orientation first, then between orientation
- **Final step, normalize each of the three intermediate maps again, then combine**
  - Equal weight for each feature
- ***$S = 1/3\,(N(\bar{I}) + N(\hat{C}) + N(\bar{O}))$***

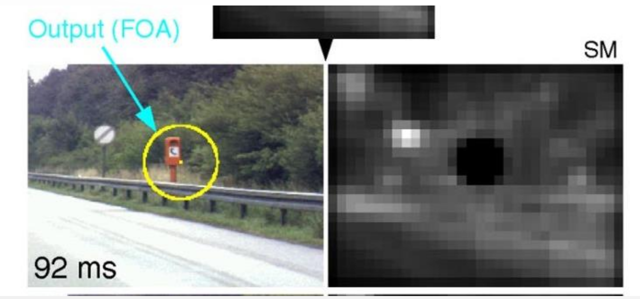


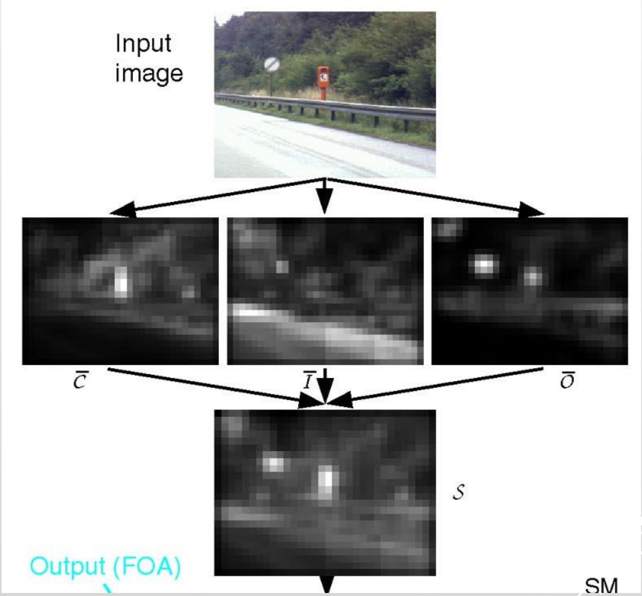
(a) Consp. map I (b) Consp. map C (c) Consp. map O (d) Saliency map S

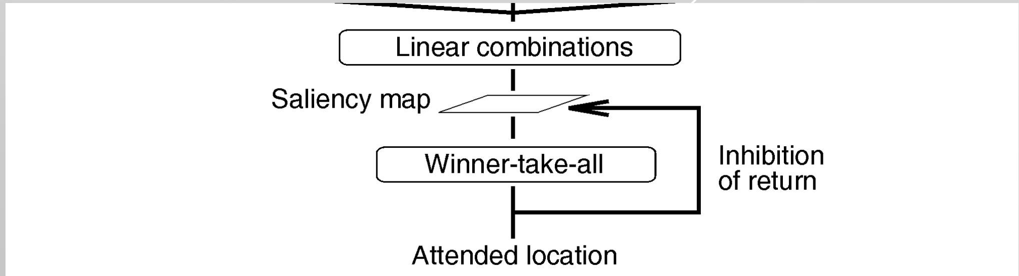
Source: http://portaldocandomble.com.br/papers.php?bmm3q.html

FINAL MAP AND SELECTION

- Area of highest peak now suggests the salient feature where attention should focus
  - But they wanted to model attention selection as well
- This final map is a 2D layer of leaky *integrate-and-fire neurons* at scale four
  - Scale 4 means attention doesn't focus at 'single pixel'
  - Capacity builds until a threshold is reached, then reset
  - Dynamic neural net since input/output change over time
- 'Winner take all' network
  - Only one possible output from a noisy system with many potential options
  - Saccades, attention, forced choice RT, decision making,...
  - In this network, connections suppress all but the most active neurons





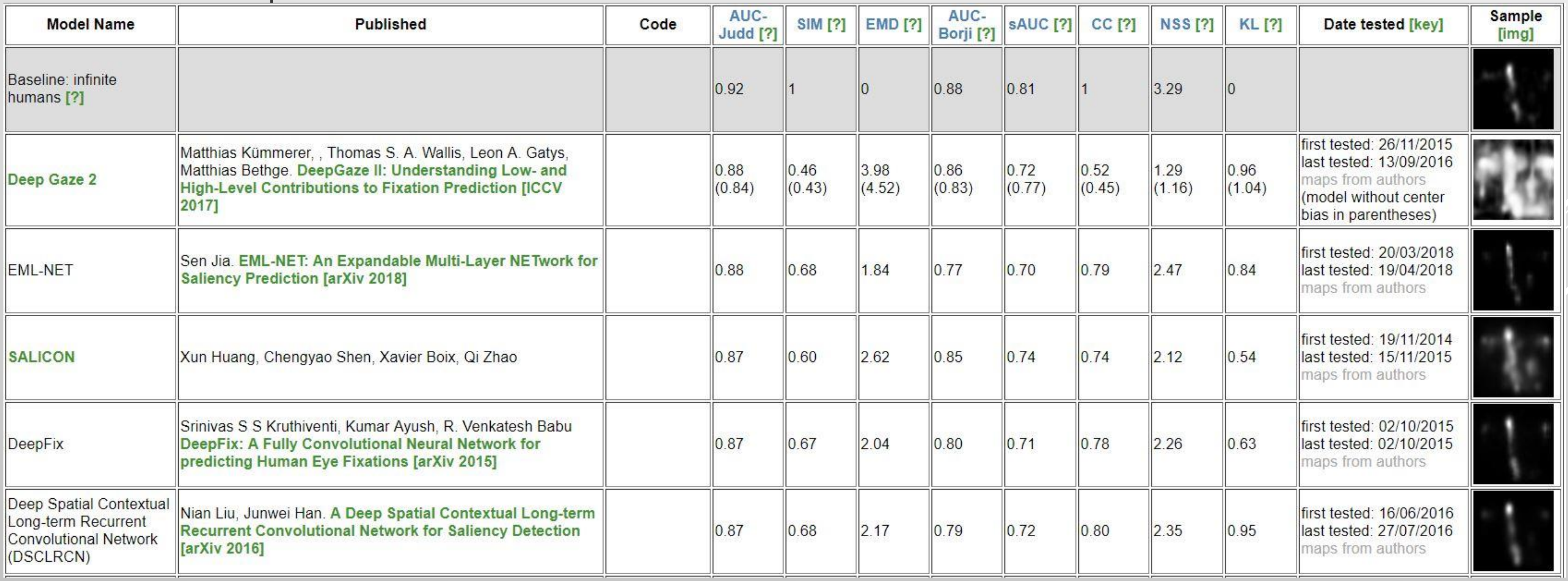


SUMMARIZING

- Model similar to aspects of primate visual cortex in neurophysiology
- Massively parallel and feedforward
  - Faster than previous iterative algorithms
- Covers early feature extraction and attention selection
- Normalization allows for realistic conjunctions of features
- **Limitations:**
  - Only three feature types
  - Feedforward does not include feedback mechanisms
  - No top down attention
  - Fails for non-implemented feature types (corners)
  - Good spatial predictions, but how about temporal performance?
- **Demo**: https://www.youtube.com/watch?v=zeFCYvwbIGU

WHAT ELSE IS THERE…?

- Efficient recreation of the visual system = biological accuracy of older models with fewer parameters or complex newer approaches?
- Interdisciplinary approach - 'reverse engineering'. Computational modelling + human vision research
- Complex hierarchical state-of-the-art DLNN's
- MIT Benchmark Top 5:



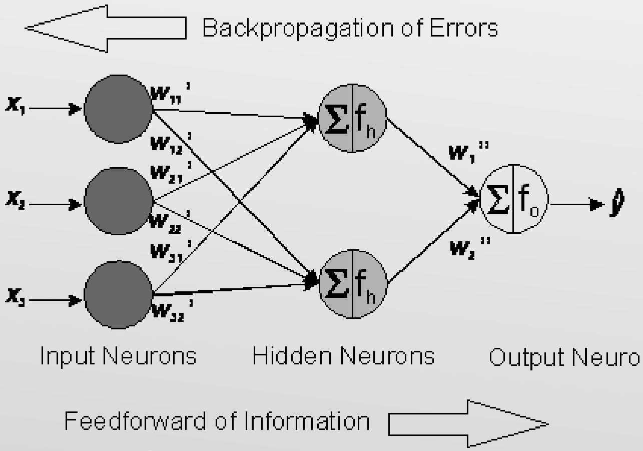
Model Name	Published	Code	AUC-Judd [?]	SIM [?]	EMD [?]	AUC-Borji [?]	sAUC [?]	CC [?]	NSS [?]	KL [?]	Date tested [key]	Sample [img]
Baseline: infinite humans [?]			0.92	1	0	0.88	0.81	1	3.29	0		
Deep Gaze 2	Matthias Kümmerer, , Thomas S. A. Wallis, Leon A. Gatys, Matthias Bethge. DeepGaze II: Understanding Low- and High-Level Contributions to Fixation Prediction [ICCV 2017]		0.88 (0.84)	0.46 (0.43)	3.98 (4.52)	0.86 (0.83)	0.72 (0.77)	0.52 (0.45)	1.29 (1.16)	0.96 (1.04)	first tested: 26/11/2015 last tested: 13/09/2016 maps from authors (model without center bias in parentheses)	
EML-NET	Sen Jia. EML-NET: An Expandable Multi-Layer NETwork for Saliency Prediction [arXiv 2018]		0.88	0.68	1.84	0.77	0.70	0.79	2.47	0.84	first tested: 20/03/2018 last tested: 19/04/2018 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	
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	
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	

Source: http://saliency.mit.edu/results_mit300.html

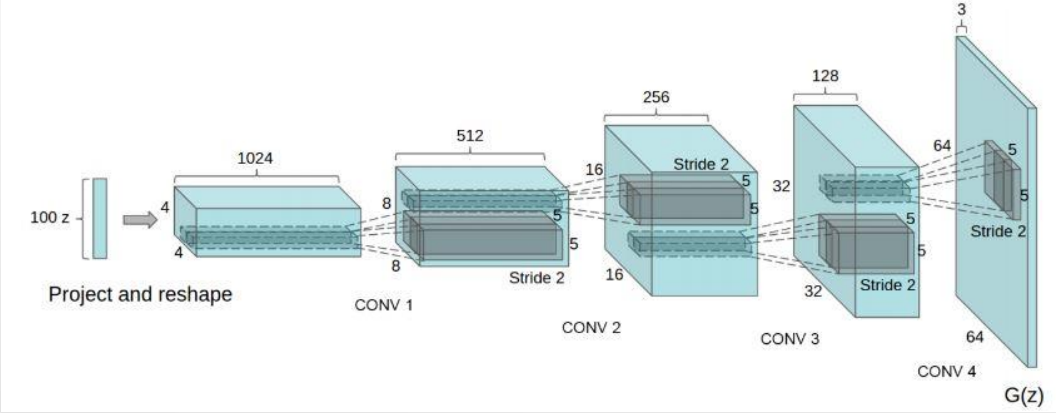
WHY GO DEEP?

- *Deep learning* - a type of machine learning algorithm that uses a **non-linear function for parallel information processing** (Deng & Yu, 2014).
- It focuses on the creation of a **multi-layered neural network** and management of the neural weights of this network in order to solve complex tasks or to replicate natural phenomena.
- are a versatile, accurate and powerful tool in domains linked to **complex data analysis** (LeCun, Bengio & Hinton, 2015).
- effective tools for modelling **high levels of abstraction** (vision, speech)

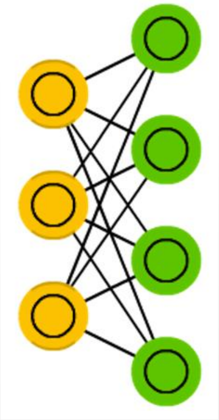
TYPES OF DLNN'S



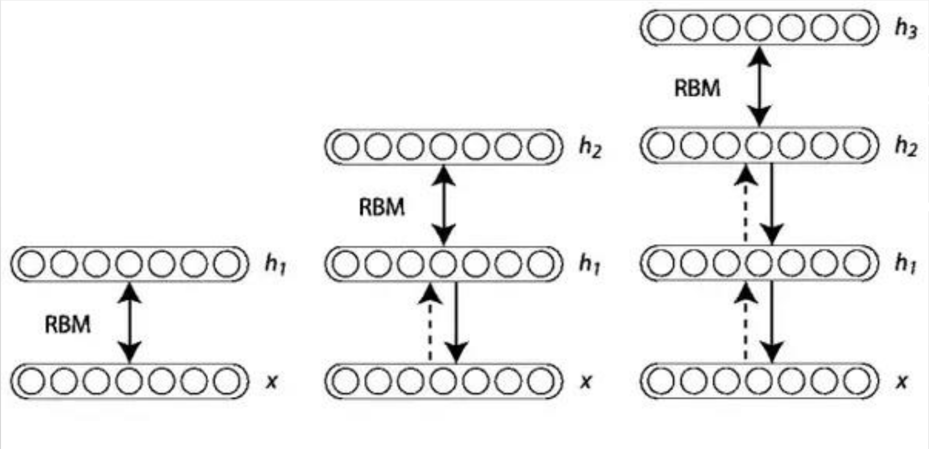
Feedforward backprop NN



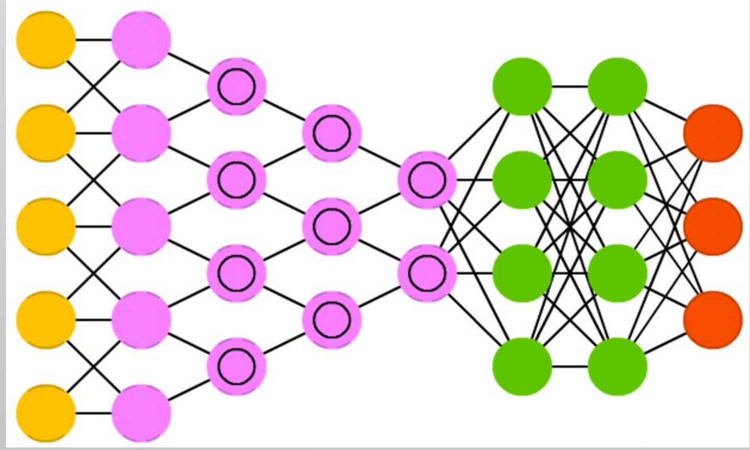
DCGAN



RBM

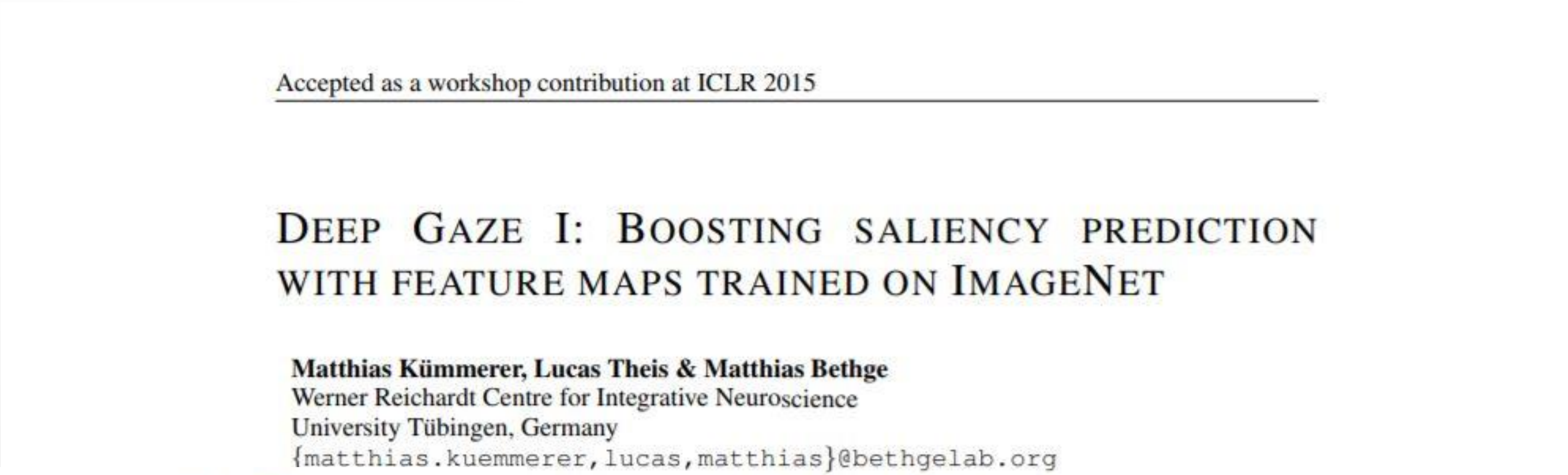


DBN



CNN

and many, many more…



Accepted as a workshop contribution at ICLR 2015

DEEP GAZE I: BOOSTING SALIENCY PREDICTION WITH FEATURE MAPS TRAINED ON IMAGENET

Matthias Kümmerer, Lucas Theis & Matthias Bethge
Werner Reichardt Centre for Integrative Neuroscience
University Tübingen, Germany
{matthias.kuemmerer,lucas,matthias}@bethgelab.org

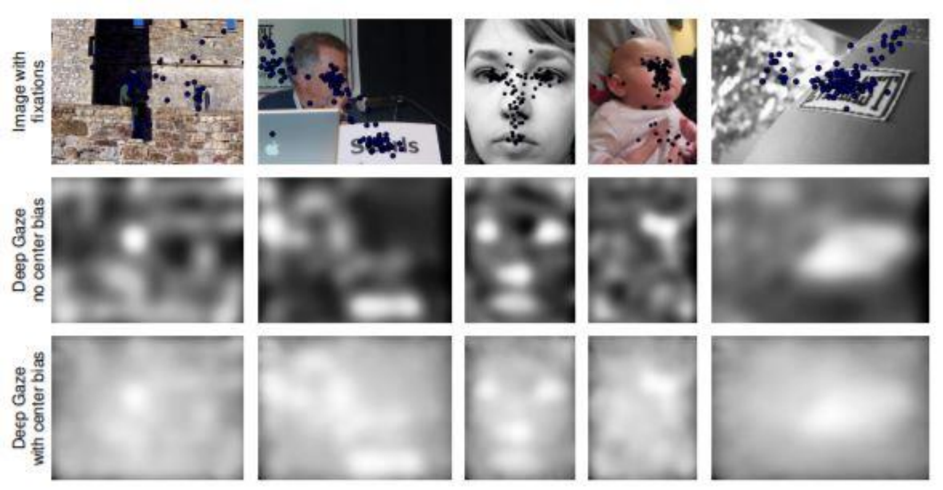


Figure 1: Example saliency maps: The top row shows example images from the dataset by Judd et al. (2009). The fixations of the subjects are indicated by dots. The middle row shows the logdensities produced by Deep Gaze I for these images when assuming a uniform prior distribution instead of a center bias. The bottom row shows the log-densities for the same images when using the center bias of the full dataset. Note that only the first two images were included in the set of images used to train Deep Gaze I.

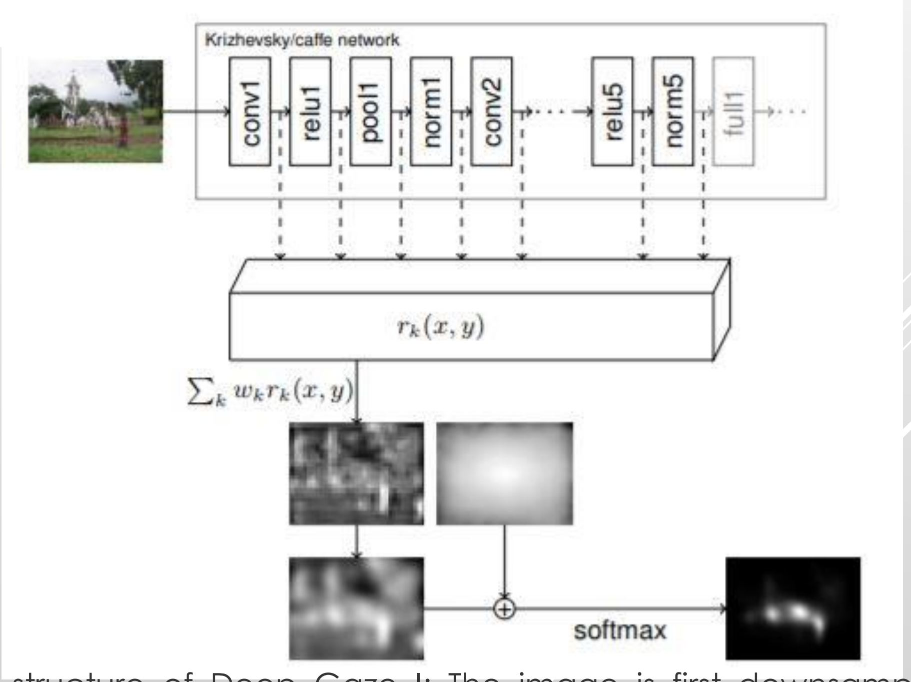


Figure 2: The model structure of Deep Gaze I: The image is first downsampled and preprocessed with the Krizhevsky network. The responses of the layers that are included in the model are then scaled up to the size of the largest network layer and normalized to have unit standard deviation. This list of maps is then linearly combined and blured with a Gaussian kernel. To compensate for the central fixation bias, an estimate of the prior distribution is added. Finally, the model output is fed through a softmax rectification, yielding a two dimensional probability distribution.

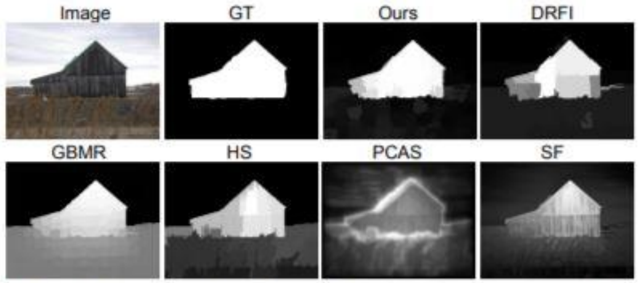


Figure 1. Examples to show problems in conventional approaches. From top left to bottom right: image, groundtruth mask, our saliency maps, and saliency maps of other five latest approaches, including DRFI [25], HS [56], GBMR [57], PCAS [41], and SF[44].

Saliency Detection by Multi-Context Deep Learning

Rui Zhao[1,2] Wanli Ouyang [2] Hongsheng Li [2,3] Xiaogang Wang[1,2]
[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]Department of Electronic Engineering, The Chinese University of Hong Kong
[3]School of Electronic Scicence, University of Electronic Science and Technology of China
{rzhao, wlouyang, hsli, xgwang}@ee.cuhk.edu.hk

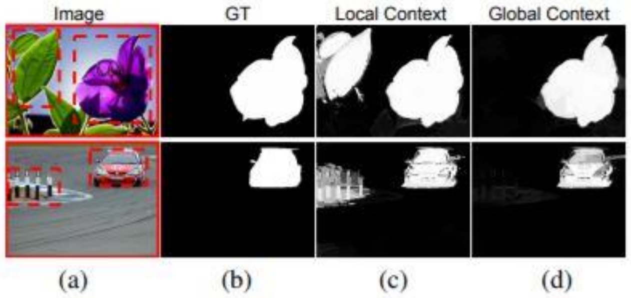


Figure 2. Examples to show importance of global context. From left to right: image, groundtruth saliency mask, our saliency map predicted with local context, and our saliency map predicted with global context.

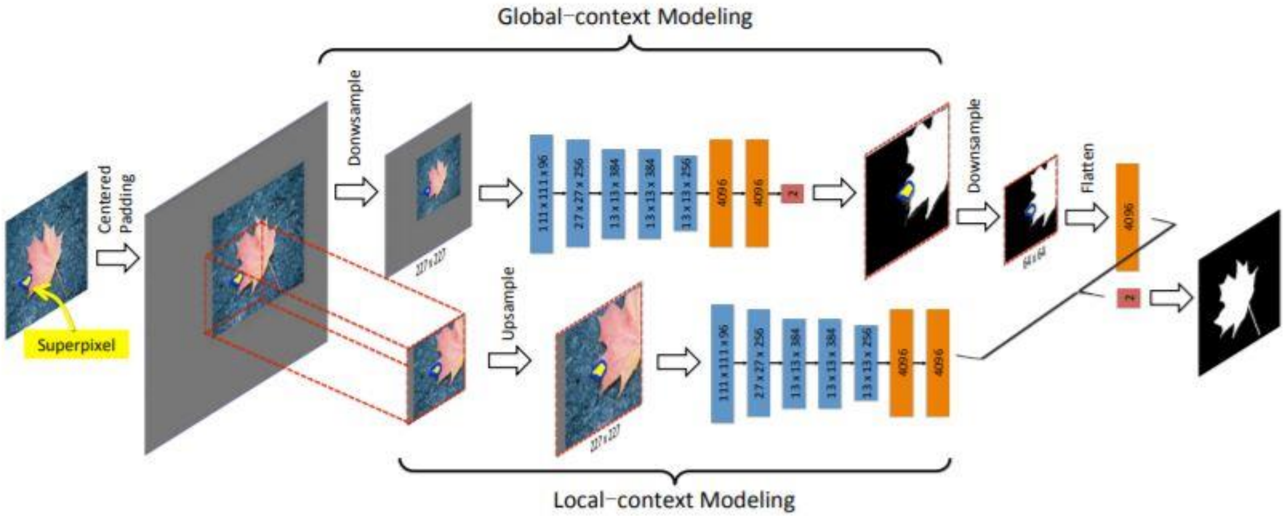


Figure 3. Upper branch: Deep CNN-based global-context modeling for saliency detection with a superpixel-centered window padded with mean pixel value. Lower branch: local-context modeling with a closer-focused superpixel-centered window, and global-context saliency detection results are combined into finally fully-connected layer in the local-context model. We visualize the network layers with their corresponding dimensions, where convolutional layers are in blue, fully connected layers (with parameters initialized using pre-trained model parameters) in orange, and fully connected layers (with parameters randomly initialized) in red. Layers without parameters are omitted in this figure.

IN SUMMARY:

- Older models, such as I&K, are a good basis for research, but may be improved by focusing on both spatial and temporal aspects;
- New approaches, such as DL, provide powerful and complex computational tools, but they are mostly used for computer vision and classification tasks;
- Highly plausible computational models may be possible due to a combination of older theoretical foundations and new state-of-the-art machine learning techniques, together with temporal algorithms for better biological precision.
- Overall, models allow to test various hypotheses and investigate relationships between parameters to predict outcomes in the entire system, which may be applicable in real-life situations.