Perception: Psychophysics and Modeling

13 | Visual Saliency

Felix Wichmann

Neural Information Processing Group
Eberhard Karls Universität Tübingen

… and the application of ML techniques to vision science …

Supplementary Reading

Barthelmé, S., Trukenbrod, H. A., Engbert, R., and Wichmann, F. A. (2013). Modeling fixation locations using spatial point processes. *Journal of Vision*, 13(12):1, 1–34.

Laurent Itti (2007), *Scholarpedia*, 2(9):3327.

Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506.

Jäkel, F., Schölkopf, B. & Wichmann, F.A. (2009). Does Cognitive Science need kernels? *Trends in Cognitive Sciences*, 13(9): 381-388.

Kienzle, W., Franz, M.O., Schölkopf, B. & Wichmann, F.A.. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, 9(5):7, 1-15.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., and Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision*, 13:201–214.

What is visual saliency? (1/2)

Visual salience (or visual saliency) is the distinct subjective perceptual quality which makes some items in the world stand out from their neighbours and immediately grab our attention. (Itti, 2007)

It is important for complex biological systems to rapidly detect potential prey, predators, or mates in a cluttered visual world. (Itti, 2007)

Complexity of processing all stimuli in the visual field **simultaneously** is prohibitive.

One solution, adopted by primates and many other animals, is to restrict complex object recognition process to a small area or a few objects at any one time. (Itti, 2007)

Solution: **Serialisation** of visual scene analysis after the gist is understood quickly.

However, this solution produces a problem. If you are only going to process one region or object at a time, how do you select (Itti, 2007) the next target of processing?

Early stages of visual processing give rise to a distinct subjective perceptual quality which makes some stimuli stand out from among other items or locations. Our brain has evolved to rapidly compute salience in an automatic manner and in real-time over the entire visual field. (Itti, 2007)

What is visual saliency? (2/2)

Visual salience (or visual saliency) is the distinct subjective perceptual quality which makes some items in the world stand out from their neighbours and immediately grab our attention. (Itti, 2007)

Our brain has evolved to rapidly compute salience in an automatic manner and in real-time over the entire visual field. (Itti, 2007)

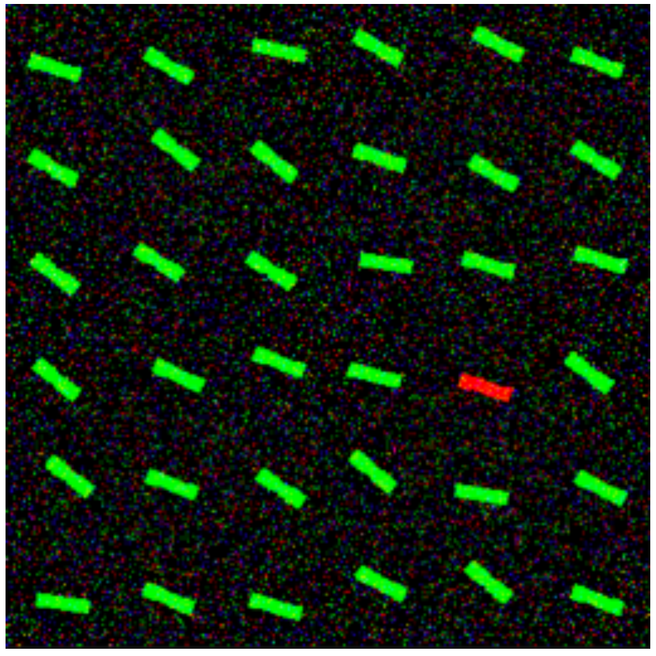
The core of visual salience is a ***bottom-up, stimulus-driven*** *signal that announces "this location is sufficiently different from its surroundings to be worthy of your attention".* (Itti, 2007)

Visual salience is sometimes carelessly described as a physical property of a visual stimulus. It is important to remember that salience is the consequence of an interaction of a stimulus with other stimuli, as well as with a visual system (biological or artificial). (Itti, 2007)

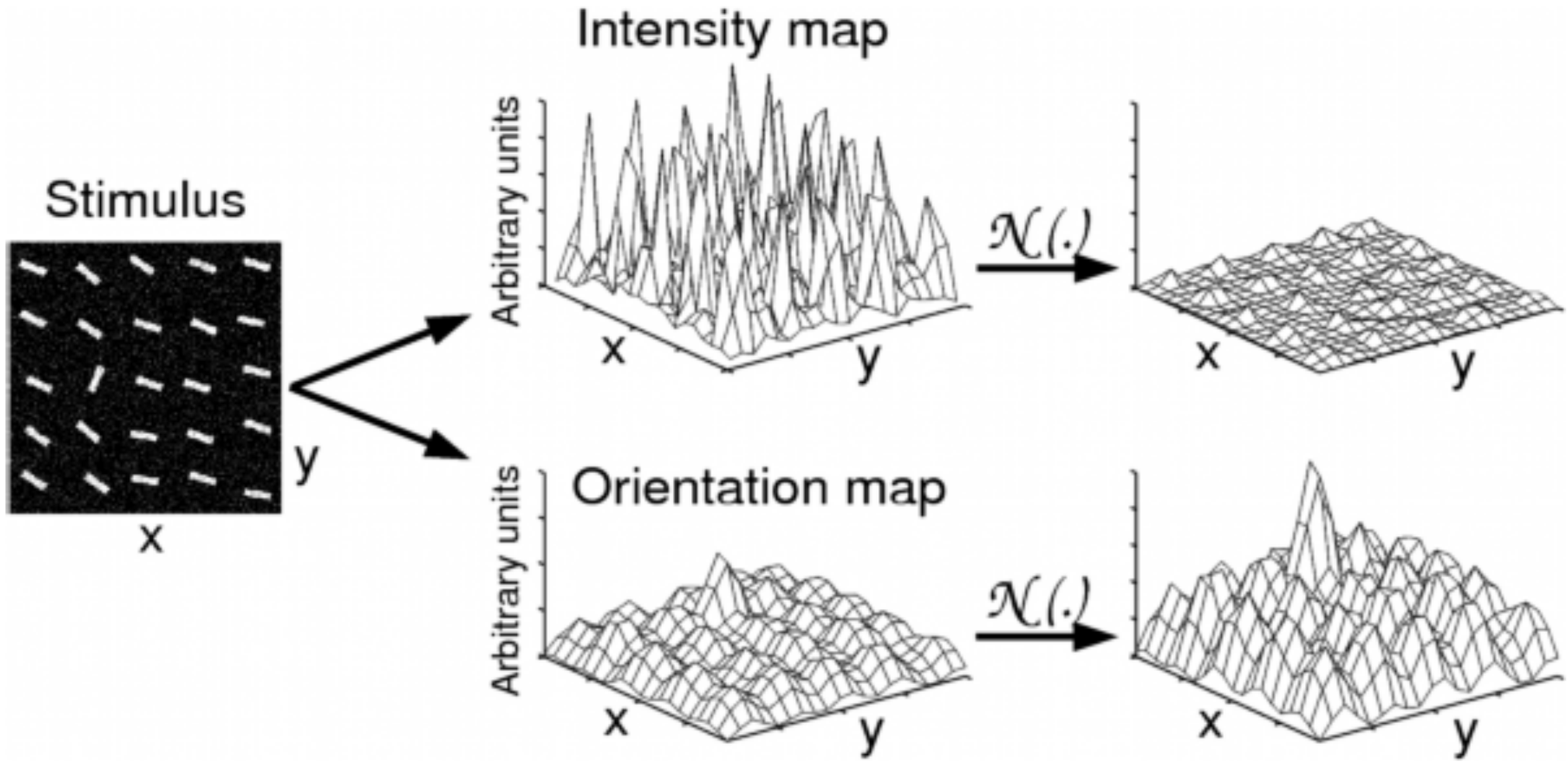
As a straight-forward example, consider that a color-blind person will have a dramatically different experience of visual salience than a person with normal color vision, even when both look at exactly the same physical scene … . (Itti, 2007)

Examples

Perceptual salience is computed automatically, effortlessly, and in real-time. In natural environments, highly salient objects tend to automatically draw attention towards them. Designers have long relied on their own salience system to create objects, such as this emergency triangle, which would also appear highly salient to others in a wide range of viewing conditions. (Itti, 2007)



The essence of salience: competing for representation



http://www.scholarpedia.org/article/Visual_salience

PERGAMON

Vision Research 40 (2000) 1489–1506

Vision Research

www.elsevier.com/locate/visres

A saliency-based search mechanism for overt and covert shifts of visual attention

Laurent Itti, Christof Koch *

Computation and Neural Systems Program, Division of Biology, California Institute of Technology, Mail-Code 139-74, Pasadena, CA 91125, USA

Received 27 May 1999; received in revised form 19 July 1999

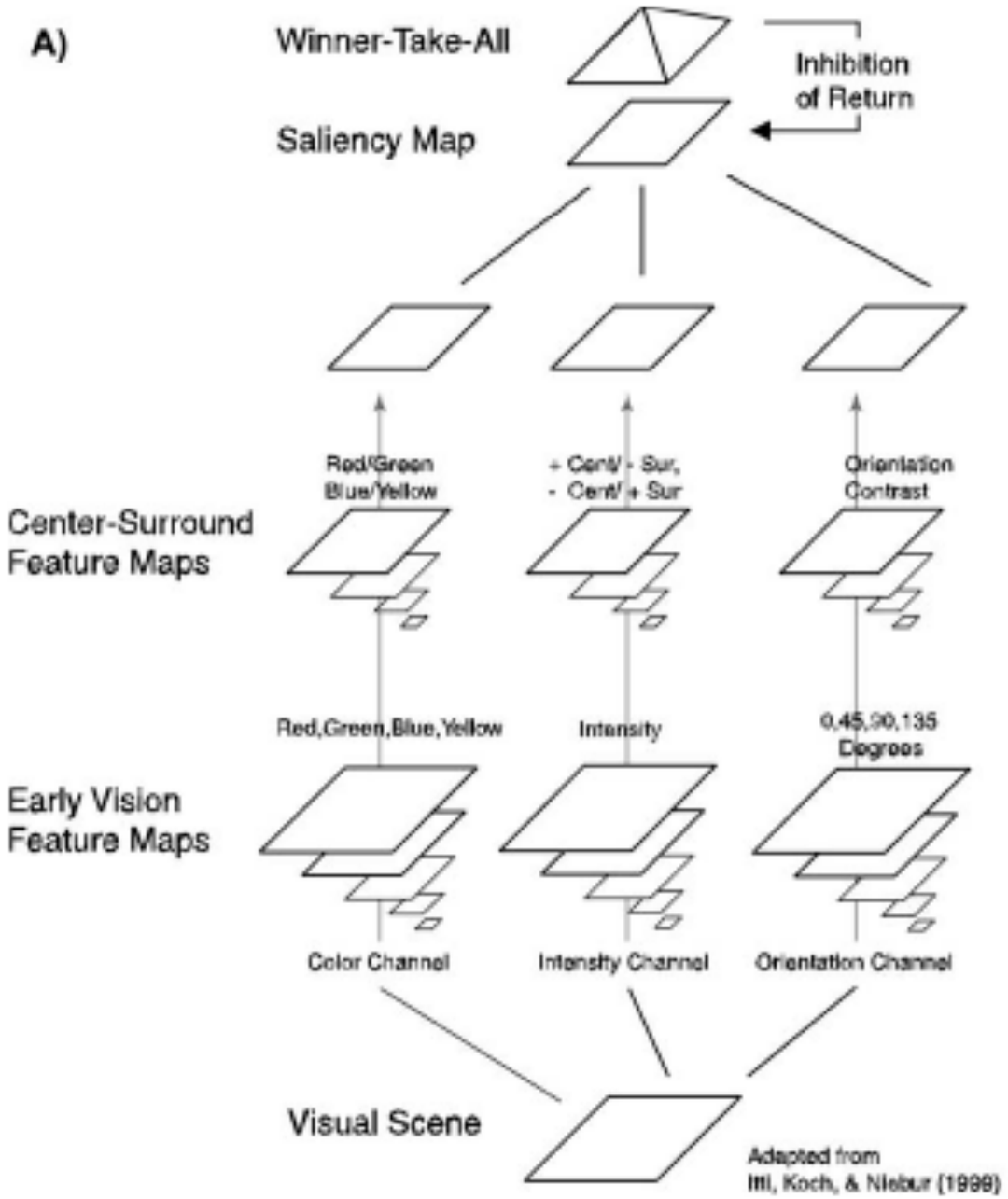
Abstract

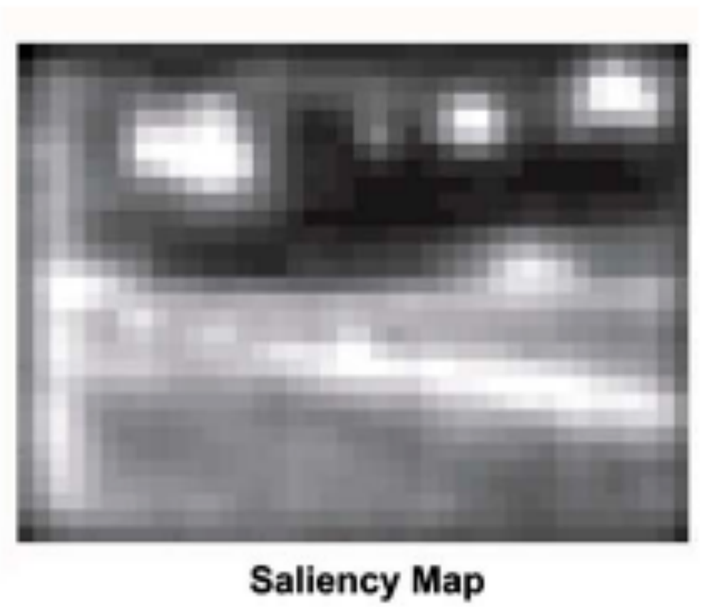
Most models of visual search, whether involving overt eye movements or covert shifts of attention, are based on the concept of a *saliency map*, that is, an explicit two-dimensional map that encodes the saliency or conspicuity of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the next attended target. Inhibiting this location automatically allows the system to attend to the next most salient location. We describe a detailed computer implementation of such a scheme, focusing on the problem of combining information across modalities, here orientation, intensity and color information, in a purely stimulus-driven manner. The model is applied to common psychophysical stimuli as well as to a very demanding visual search task. Its successful performance is used to address the extent to which the primate visual system carries out visual search via one or more such saliency maps and how this can be tested.

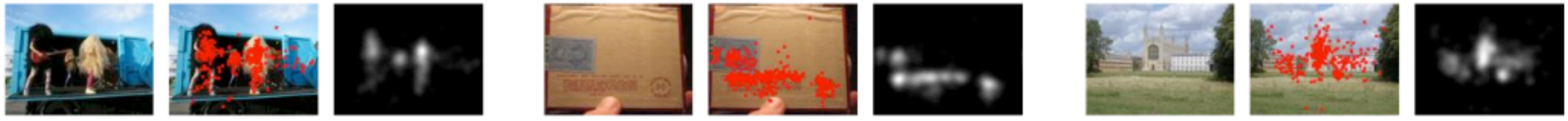
Keywords: Visual attention; Saliency; Vision systems

First well-known and inspirational visual saliency model

The so-called ‘‘Itti & Koch’’ model







about | **results** | datasets | submission | downloads

mit300 | cat2000

mit saliency benchmark results: mit300

The following are results of models evaluated on their ability to predict ground truth human fixations on our **benchmark data set containing 300 natural images with eye tracking data from 39 observers**. We post the results here and provide a way for people to submit new models for evaluation.

citations

If you use any of the results or data on this page, please cite the following:

```
@misc{mit-saliency-benchmark,
  author      = {Zoya Bylinskii and Tilke Judd and Ali Borji and Laurent Itti and Fr{\'e}do Durand and Aude Oliva and Antonio Torralba},
  title       = {MIT Saliency Benchmark},
}
```

These evaluations are released in conjuction with the following papers:

```
@article{salMetrics_Bylinskii,
  title      = {What do different evaluation metrics tell us about saliency models?},
  author     = {Zoya Bylinskii and Tilke Judd and Aude Oliva and Antonio Torralba and Fr{\'e}do Durand},
  journal    = {arXiv preprint arXiv:1604.03605},
  year       = {2016}
}

@InProceedings{Judd_2012,
  title      = {A Benchmark of Computational Models of Saliency to Predict Human Fixations},
  author     = {Tilke Judd and Fr{\'e}do Durand and Antonio Torralba},
  booktitle = {MIT Technical Report},
  year       = {2012}
}
```

images

300 benchmark images (the fixations from 39 viewers per image are not public such that no model can be trained using this data set).

model performances

Model Visualizations

68 models, **5** baselines, **8** metrics, and counting...

Performance numbers prior to September 25, 2014.

Matlab code for the metrics we use.

Sorted by: AUC-Judd metric

Visual saliency as a machine learning problem

What is special about the local image structure at fixation points?

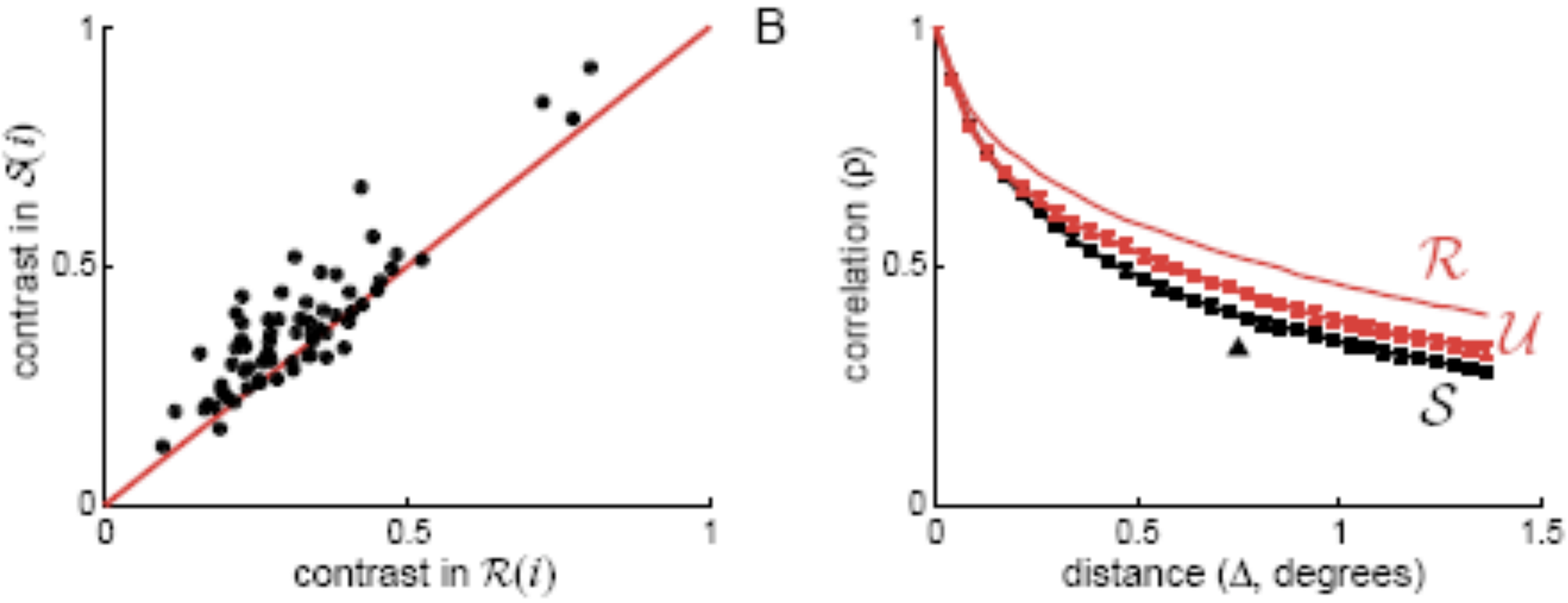
How does *p*(fixation) depend on local image statistics?

Statistical properties of fixation locations (1)

Correlation coefficient of RMS and model output: 0.69

Center pixel ‘‘more different’’ to surrounding pixels in fixation patches

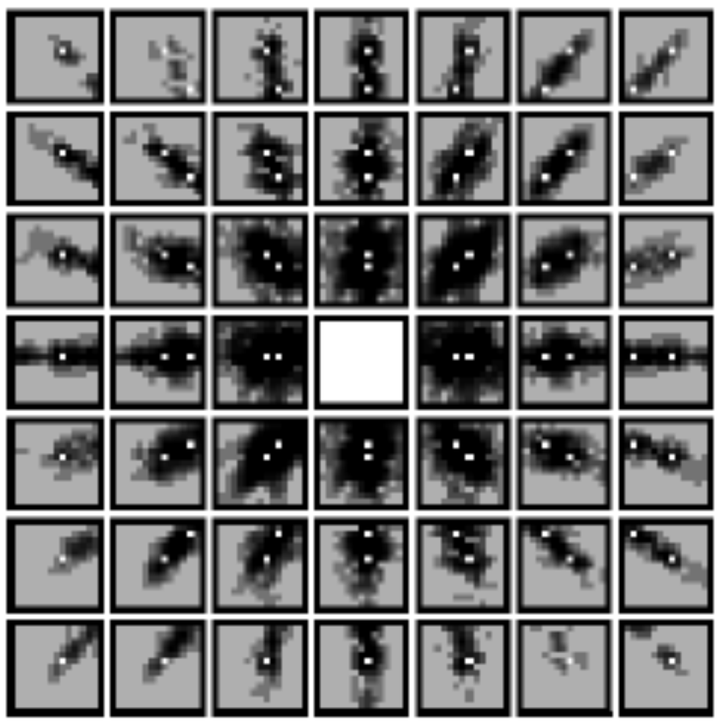
(Reinagel & Zador, 1998)



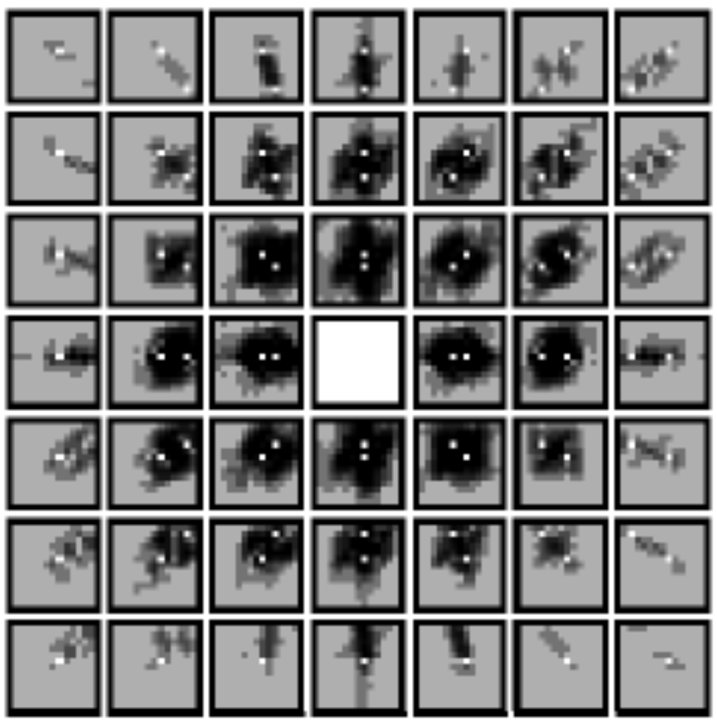
Statistical properties of fixation locations (2)

For third order statistics, the ‘‘energy distribution is more circular’’:
‘‘The saccadic selection system avoids image regions which are dominated by a single oriented structure. Instead, it selects regions containing different orientations, like occlusions, corners, etc.’’ (Krieger et al., 2000)

$C_3^{U_{rand}}(f_{x1}, f_{y1}, f_{x2}, f_{y2})$



$C_3^{U_{eye}}(f_{x1}, f_{y1}, f_{x2}, f_{y2})$



What is Machine Learning?

Algorithmic approach to the science of learning from data, initially mainly developed in computer science.

Statistics is the science of learning from data, too ...

These fields are identical in intent although they differ in their history, conventions, emphasis and culture.

Larry Wasserman, *Rise of the Machines*, 2014

Machine learning algorithms (at their best) excel at discovering hidden structure in existing data in order to predict novel data—exploratory methods, strong emphasis on prediction.

Machine learning (at its worst) designs algorithms nobody would ever want to use, and packages together ad-hoc heuristics for ‘‘data mining.’’

Machine learning driven to a non-trivial extent by companies—Google, Facebook, Apple—and, presumably, government agencies—NSA—interested in solving very large-scale inference problems, where ‘‘data’’ might mean gene sequences, graphs, images, videos, twitter feeds, and web pages, not ‘‘just’’ numbers.

ML places great emphasis on algorithms ...

When is an algorithm considered ‘‘good’’ in computer science?

- Low time complexity (‘‘fast’’)
- Low space complexity (‘‘small memory’’)
- Approximation guarantees (‘‘close to global optimum’’)

When is an algorithm considered ‘‘good’’ from a statistical point of view?

- Consistency: For more and more input data, the algorithm should converge to the true solution
- Confidence: The algorithm should know about its own reliability

Statistical goals do not necessarily align well with those aspired to by computer science.

ML at its finest: Find a good trade-off between all of these requirements.

Given that ML is (still) largely computer science based, it has perhaps a stronger emphasis on computation ... (computability in polynomial time, convex optimization): a very practical, pragmatic approach.

ML approach to visual saliency

Previously: Classic top-down, mechanistic modeling approach developing ‘‘biologically inspired’’ models built using ‘‘neurophysiological-hardware’’ like Gabor filters, ...

Style of modelling well-suited and time-proven if abundant domain knowledge is available—often not the case in sensory psychology: Many more or less ad-hoc choices have to be made, e.g. exact filter types, sizes, numbers, combination strategies, ...

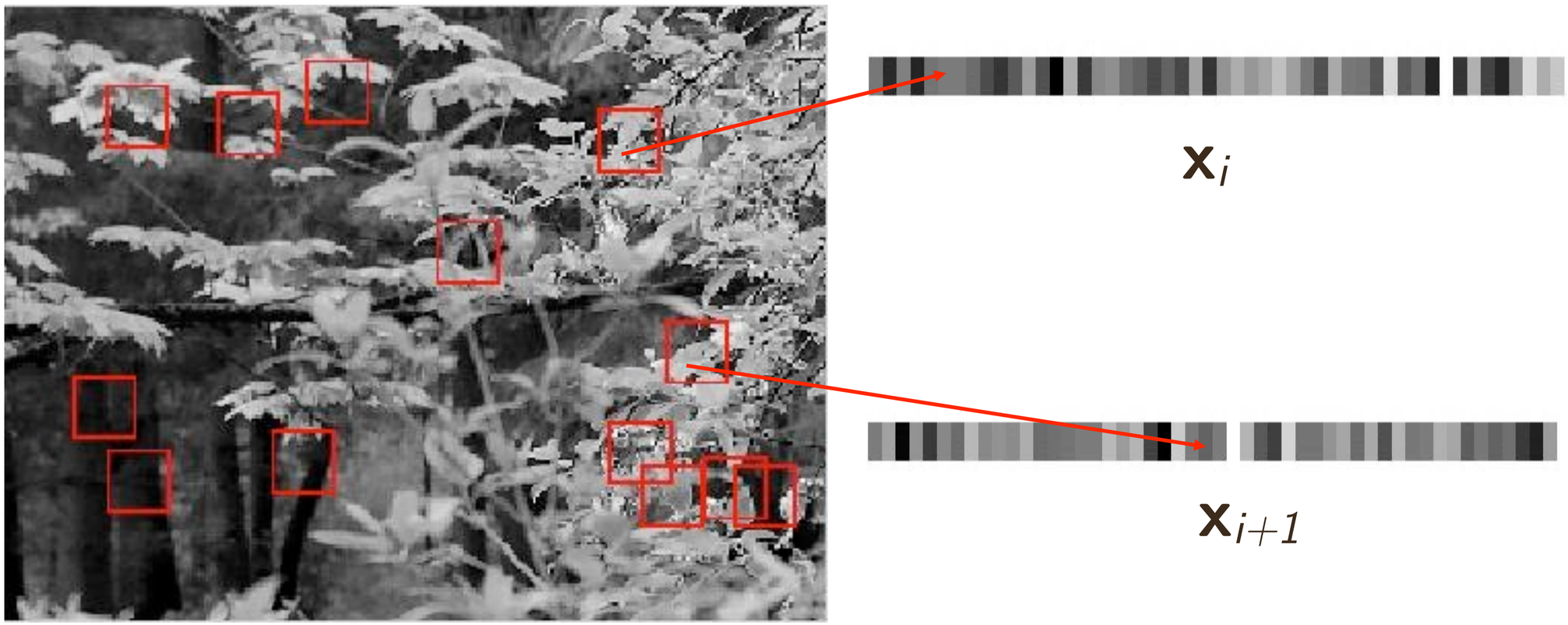
Machine Learning approach: construct a model from the data, i.e. ...

Use a very general model class that does not ‘‘know’’ about the problem, but can adapt very well to a large class of problems.

Numerically learn—**optimise**—its parameters such that new data is predicted best.

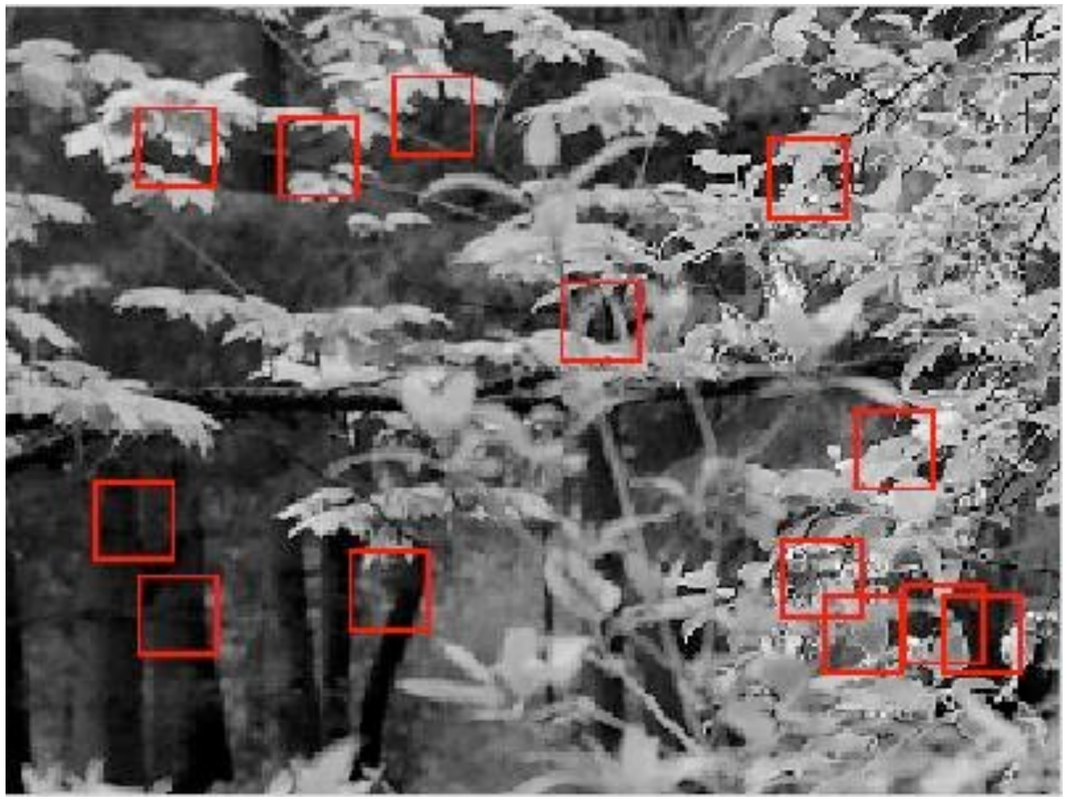
Data representation

For each fixation location (data point, i = 1 … 36,000), store local pixel values in a feature vector $\boldsymbol{x}_i$ and associate a label $\mathbf{y}_i$ = 1/-1 (fixation/

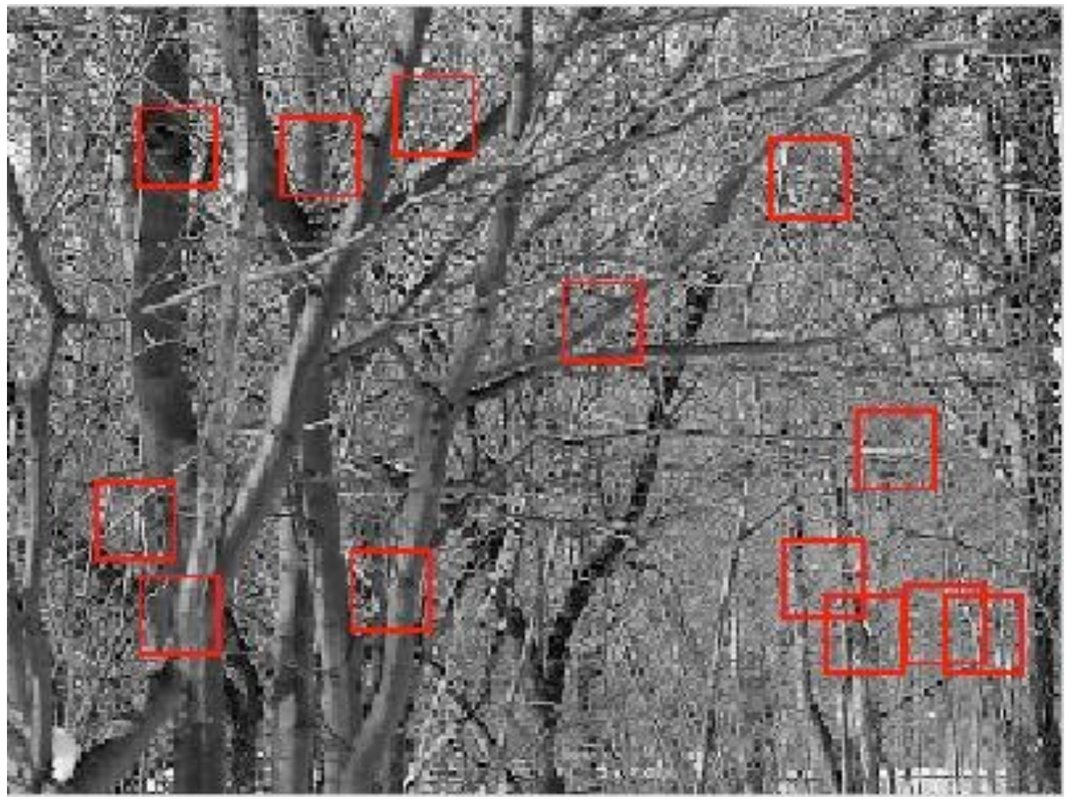


What is a “non-fixated” patch?

Generate background examples with same spatial distribution as fixations, the recipe suggested by Reinagel & Zador (1998).



Fixations



Background

ML method

Overall strategy: make the model class as general as possible

The model is a radial basis function (RBF) network with one basis function centred on each training example. (“Nonparametric” as its complexity grows with the number of data points.)

General? Universal approximation property, no preference for any image structure, no knowledge about shape or size of receptive fields in the human visual system.

We compute the weights (α_i) using hinge loss + L2-regularizer (= SVM)—finding α_i is convex, i.e. efficient and guaranteed to find the global optimum.

We find the *design parameters* λ, γ, and patch size d via exhaustive grid-search, using cross-validation estimates of accuracy—feasible, as problem only 3D.

Radial-basis-function support vector machine (RBF-SVM)

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2\right)$$

Kernel bandwidth

Patch size: d

Weights

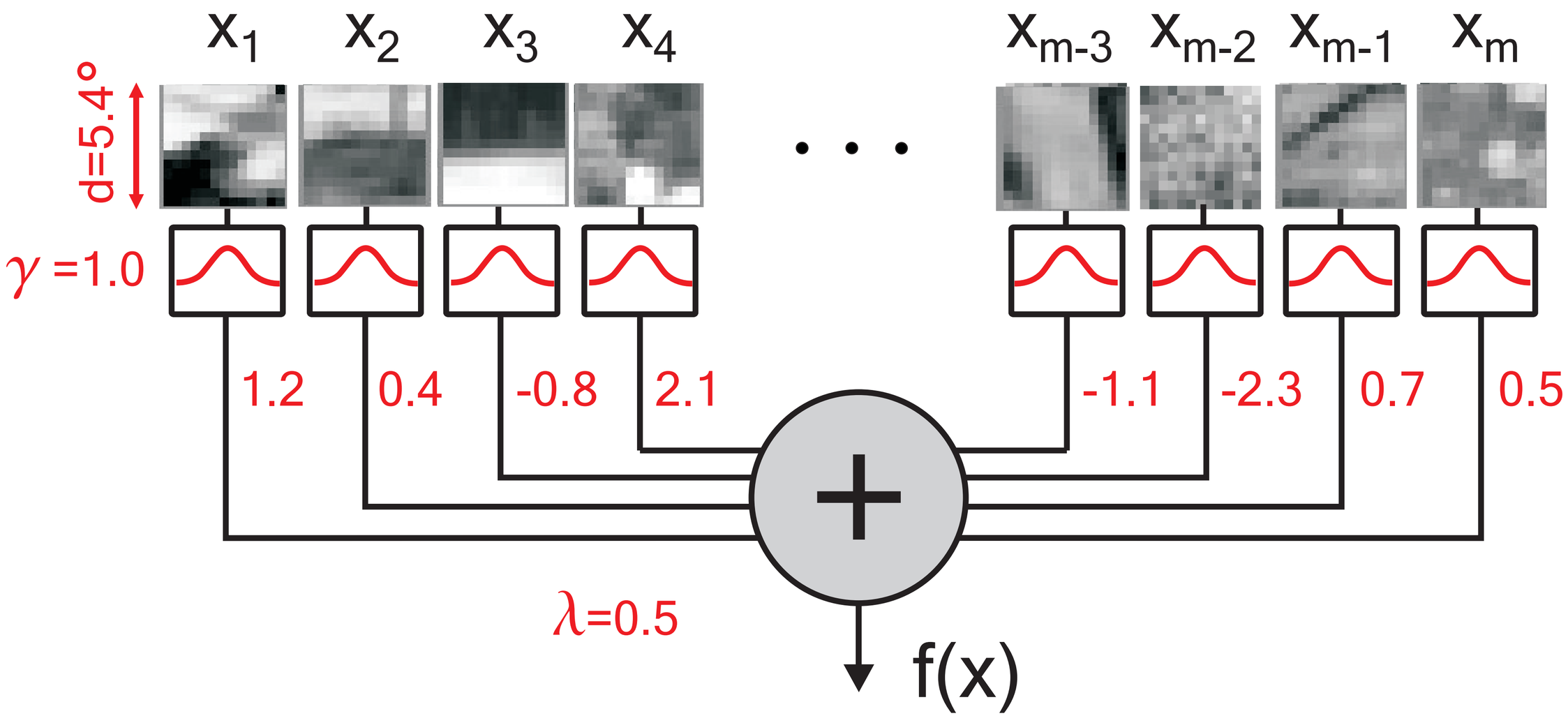
$$\lambda \|f\|^2 + \sum_{i=1}^{m} \max(0, 1 - y_i f(\mathbf{x}_i))$$

Smoothness

>24,000 weights

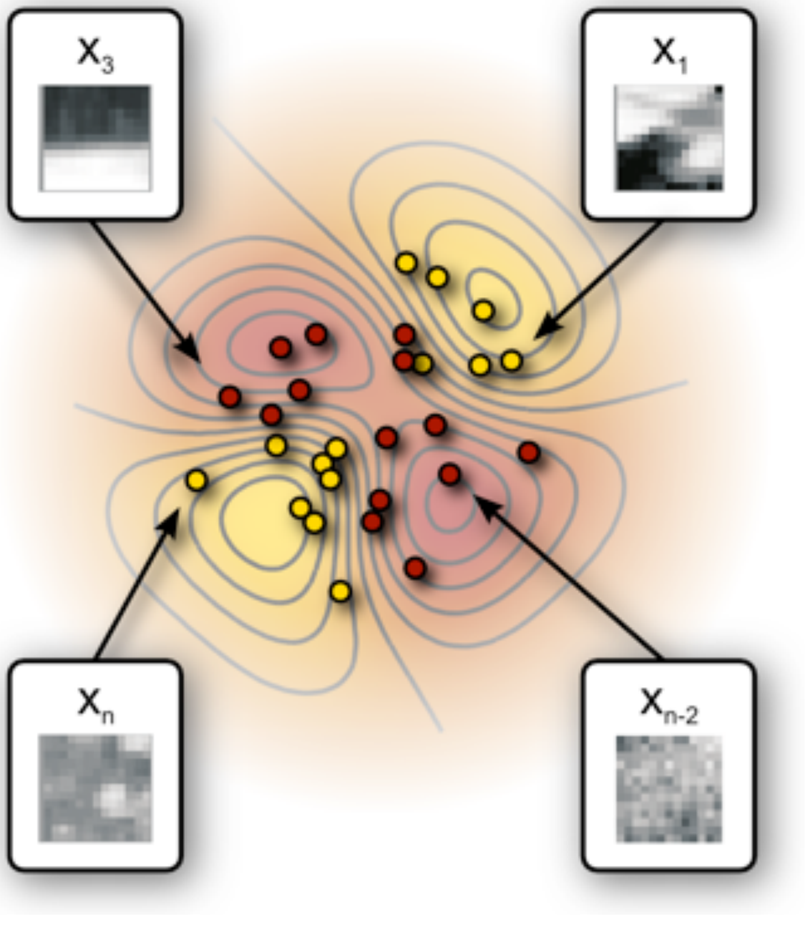
3 design parameters

RBF-SVM after optimization ("learning")



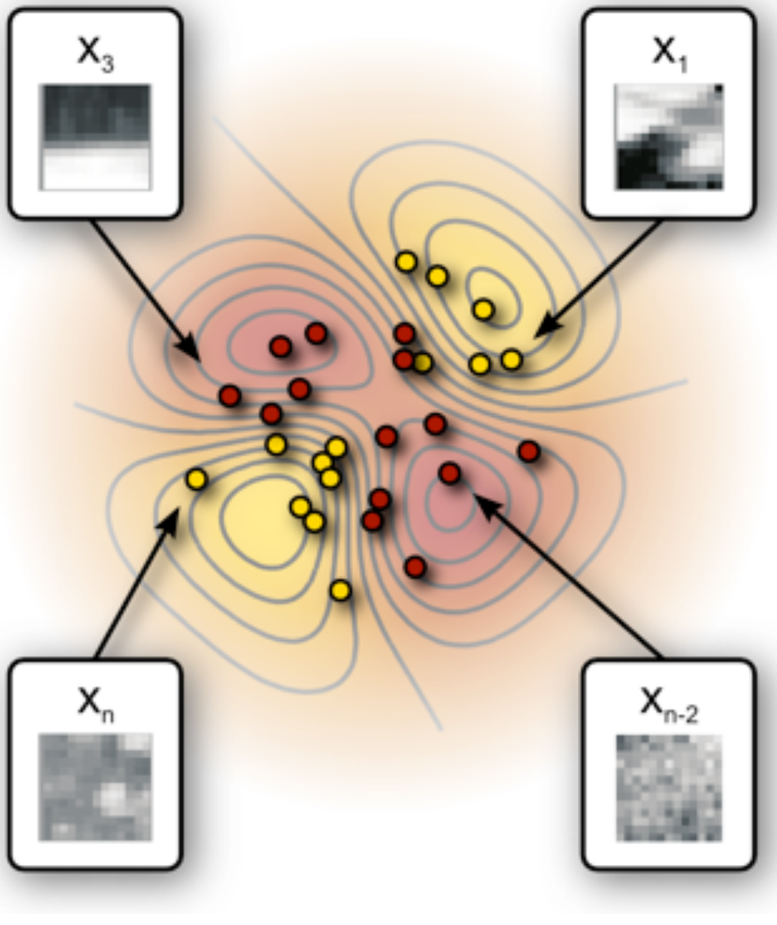
Predicitivity (area under ROC): 0.64±0.010

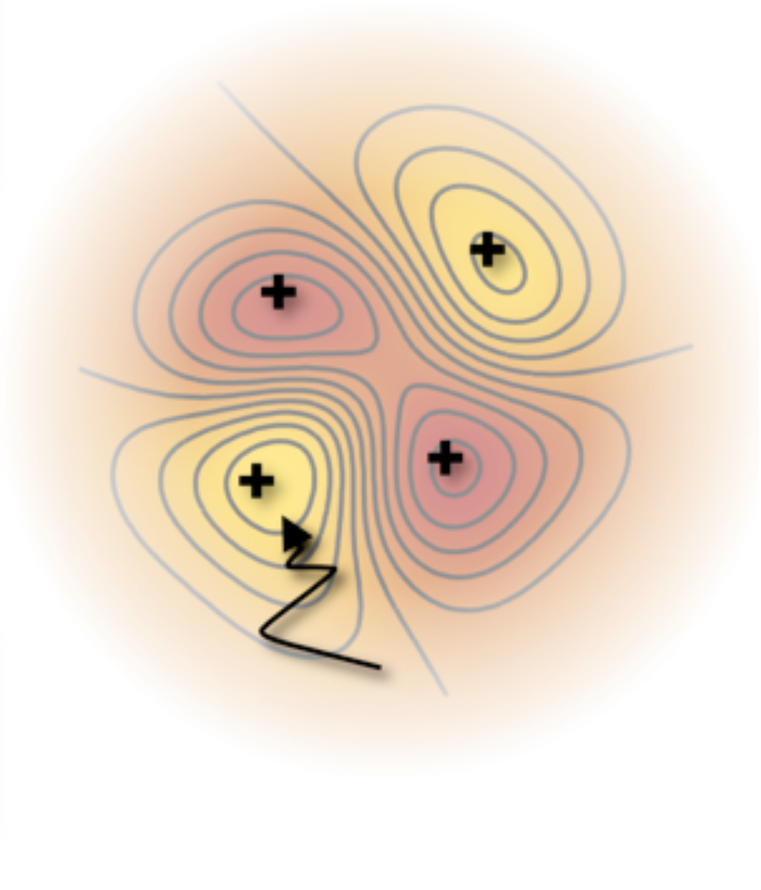
(a)



(a)

(b)





(a)

(b)

(c)

x_3

x_1

x_n

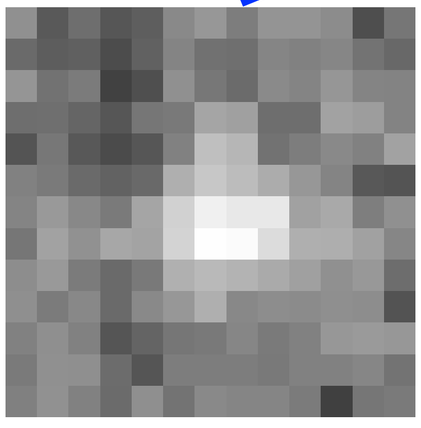
x_{n-2}

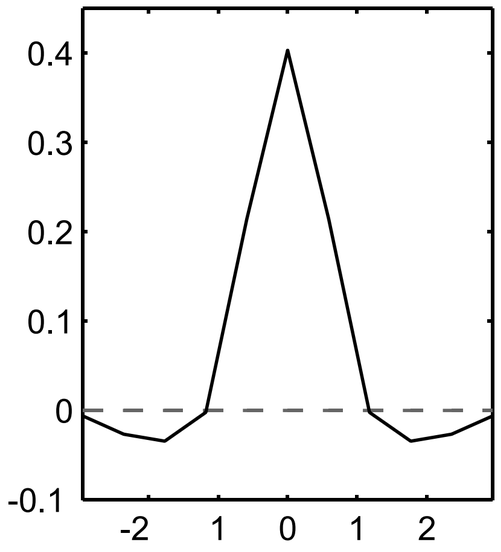
z_1

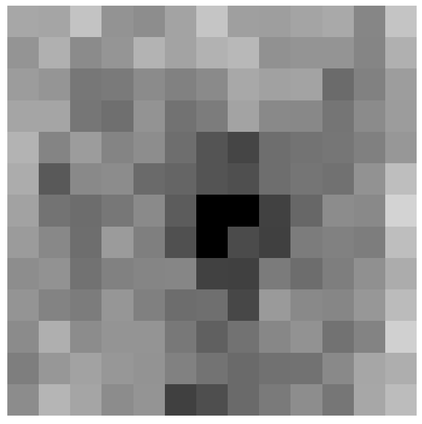
z_3

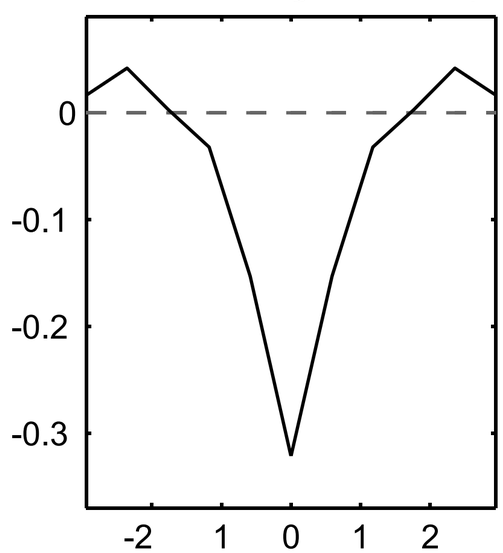
z_4

z_2





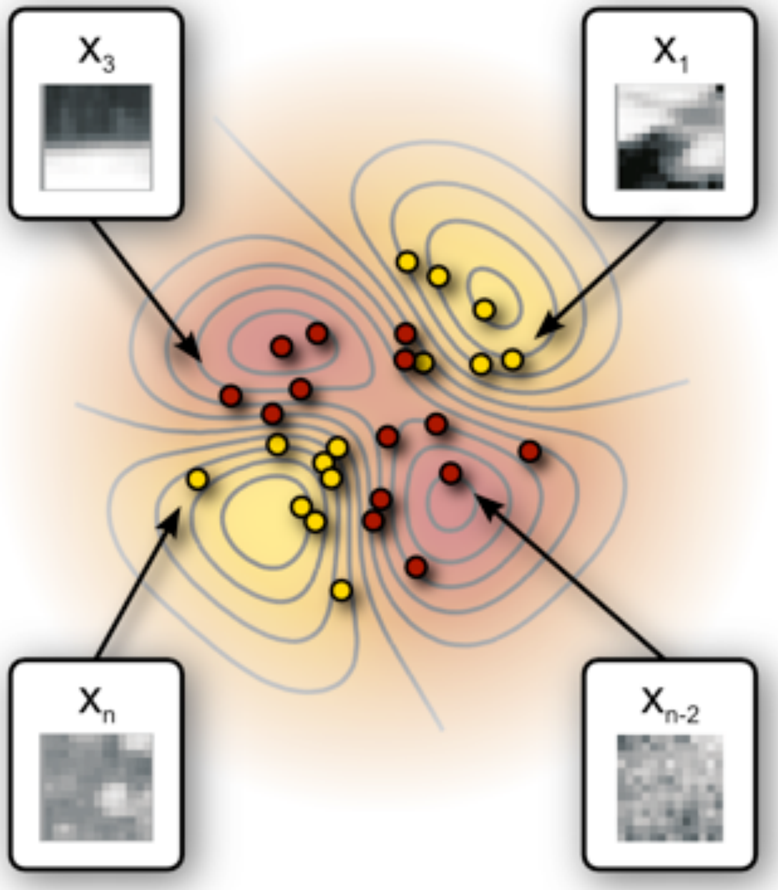


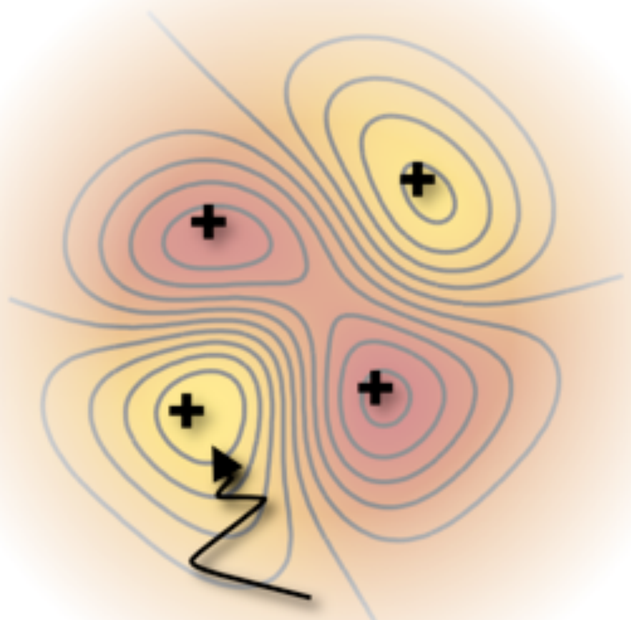


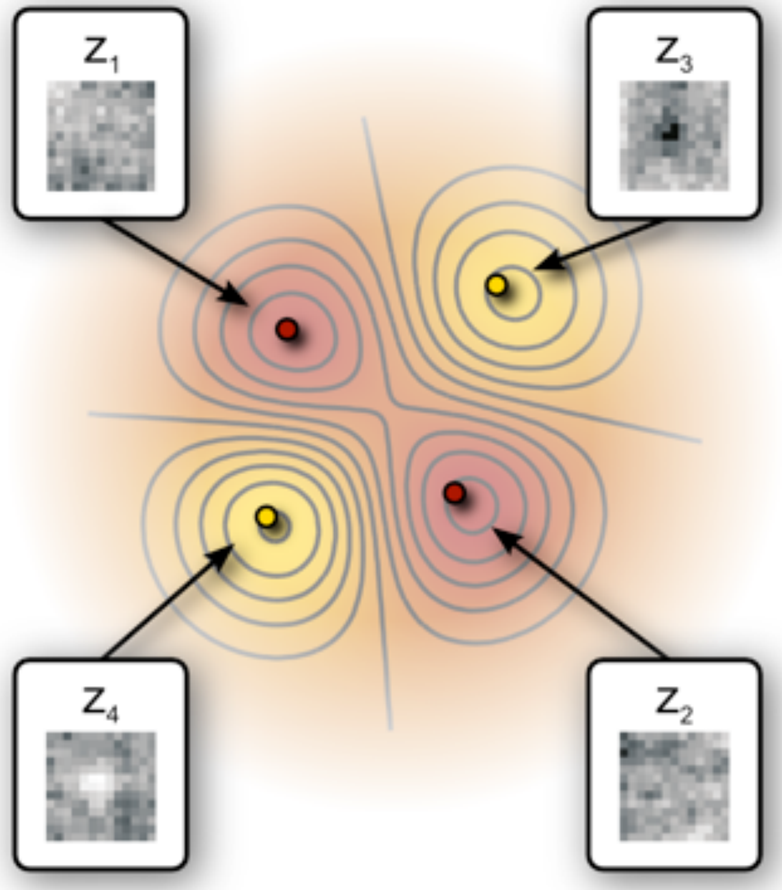
(a)

(b)

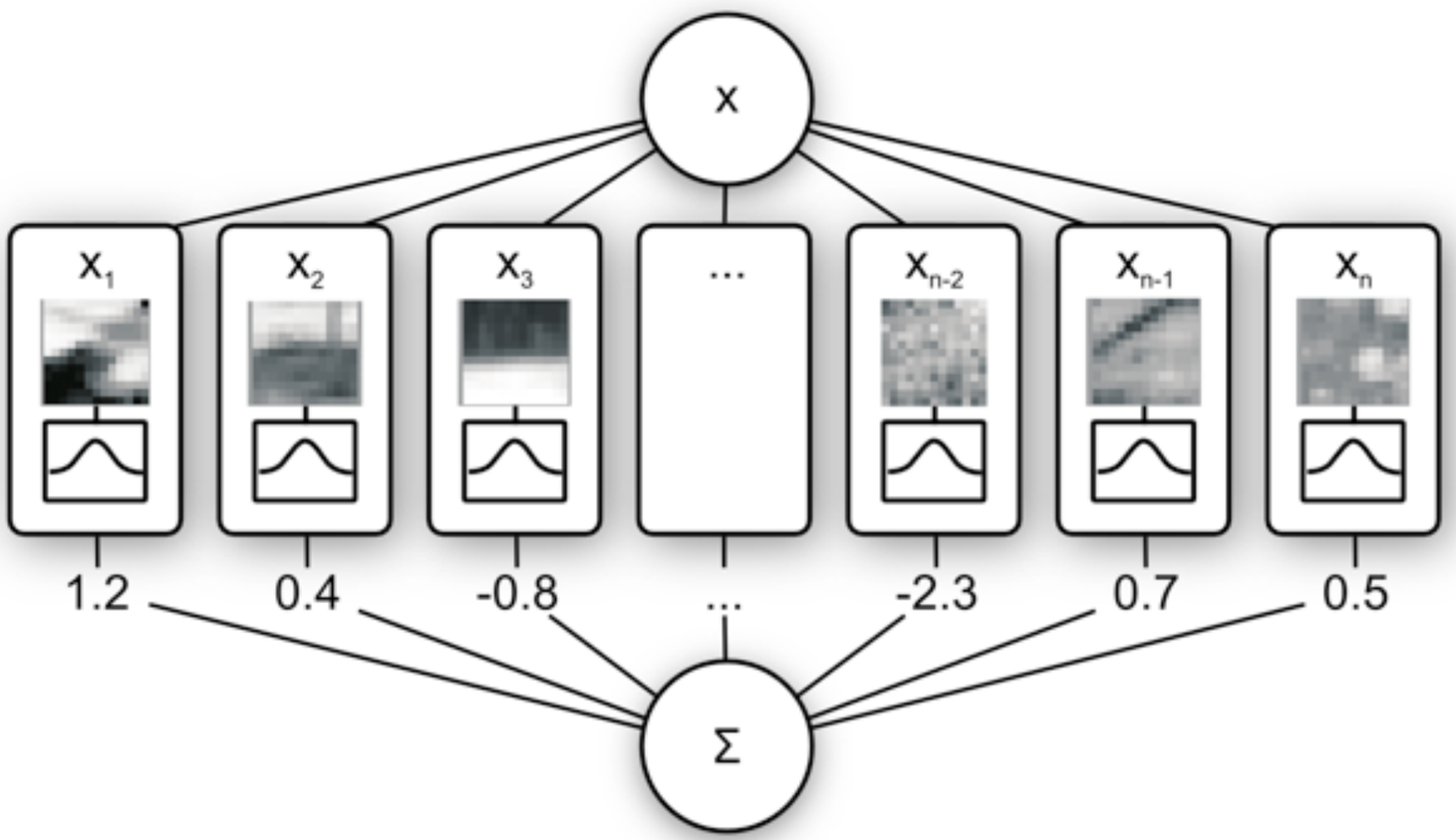
(c)


x_3
x_1
x_n
x_{n-2}

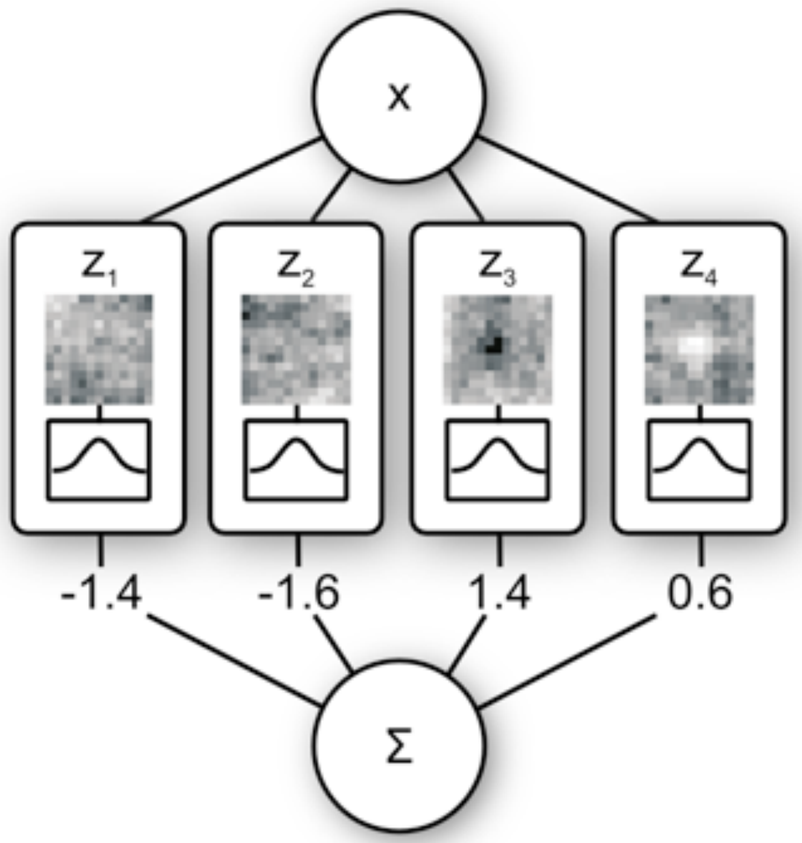



z_1
z_3
z_4
z_2

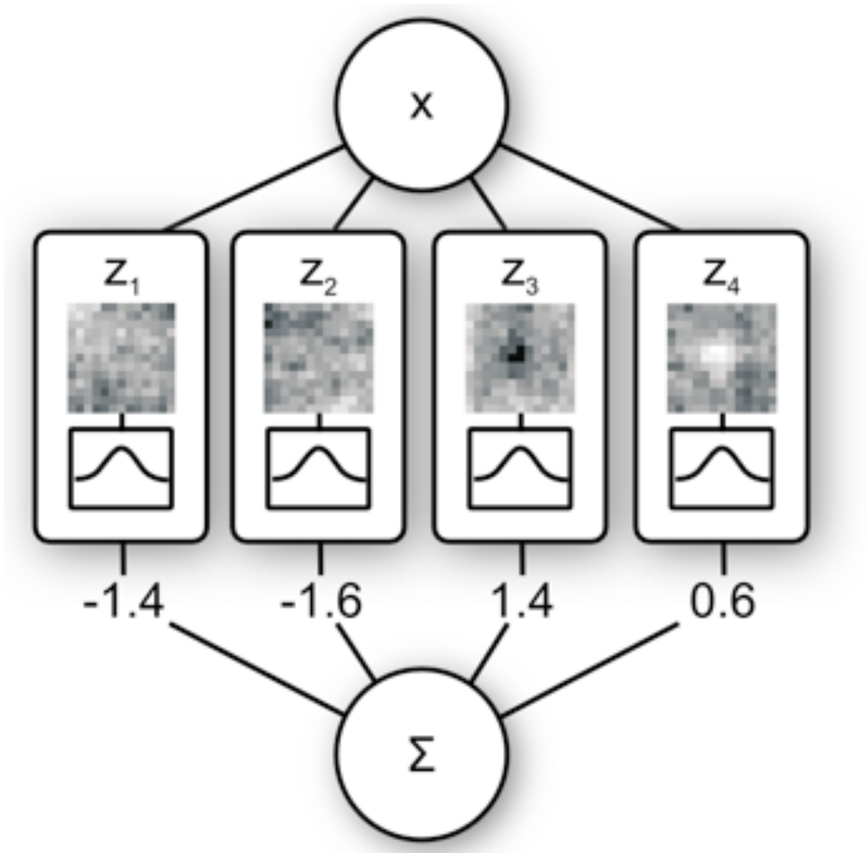
(d)

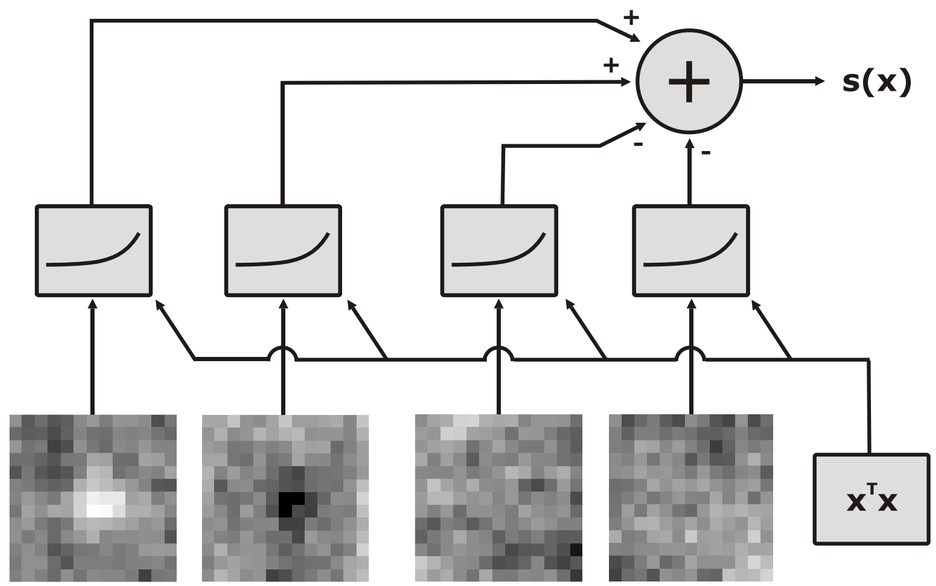

x
x_1
x_2
x_3
...
x_{n-2}
x_{n-1}
x_n
1.2
0.4
-0.8
...
-2.3
0.7
0.5
Σ

≈


x
z_1
z_2
z_3
z_4
-1.4
-1.6
1.4
0.6
Σ

Non-linear decision-image network for visual saliency





Critical controls

1. Ground-Truth Test ✓

2. Generalization to novel data set:

ML-model: 0.64 ± 0.010 s.e.m.
Itti-Koch: 0.62 ± 0.020 s.e.m.

ML-model: 0.62 ± 0.012 s.e.m.
Itti-Koch: 0.57 ± 0.020 s.e.m.

Interim Conclusions

Bottom-up saliency can be inferred from data, without prior assumptions regarding the computational architecture.

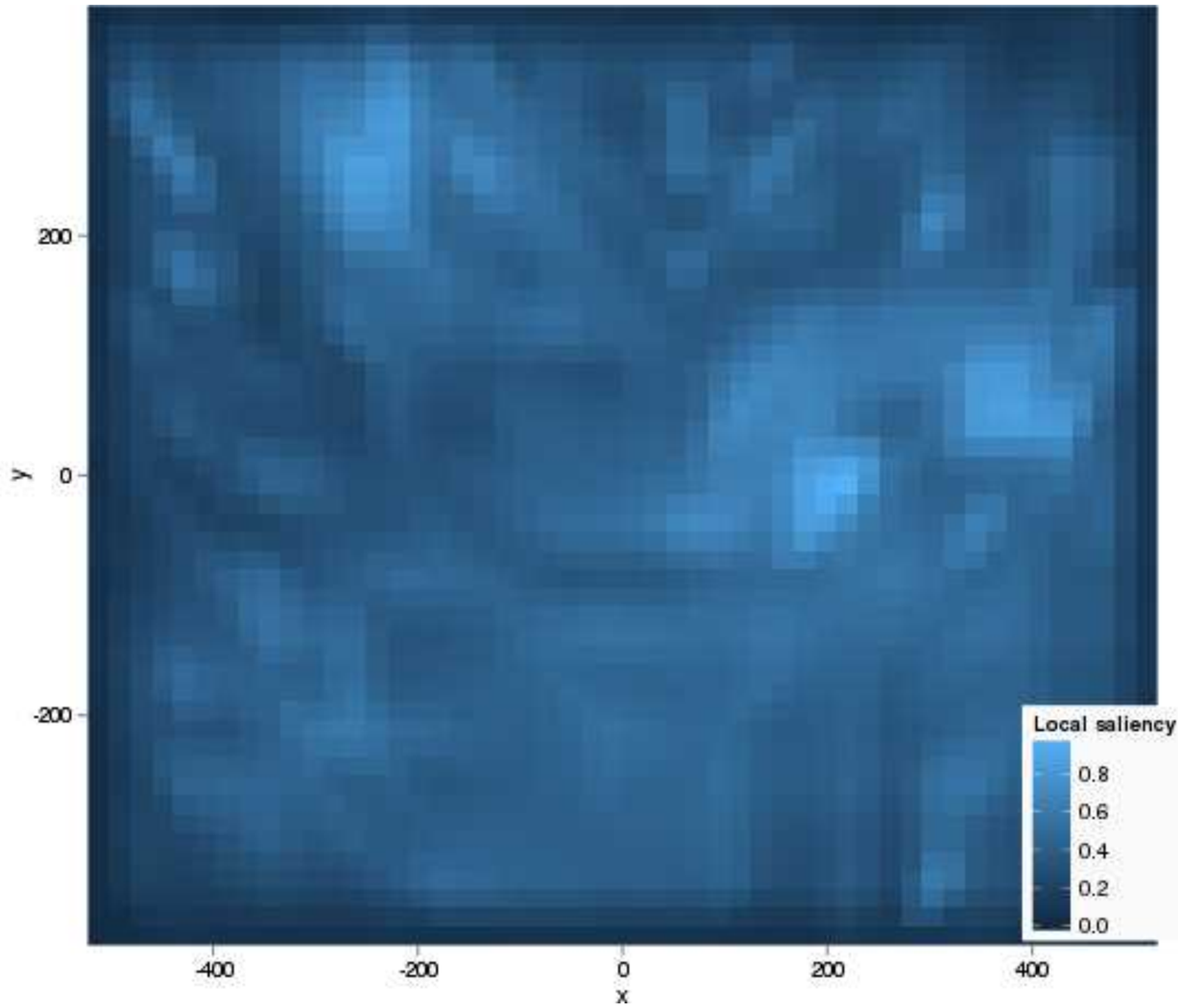
The most relevant regularity in local image structure at fixation is a simple center-surround configuration. (Biologically plausible but learned from the data, not assumed!)

Assembled into a small network with only four linear receptive fields followed by a static nonlinearity and contrast gain-control, the prediction performance of the full RBF-SVM is obtained—this model is very simple compared to previously suggested ones.

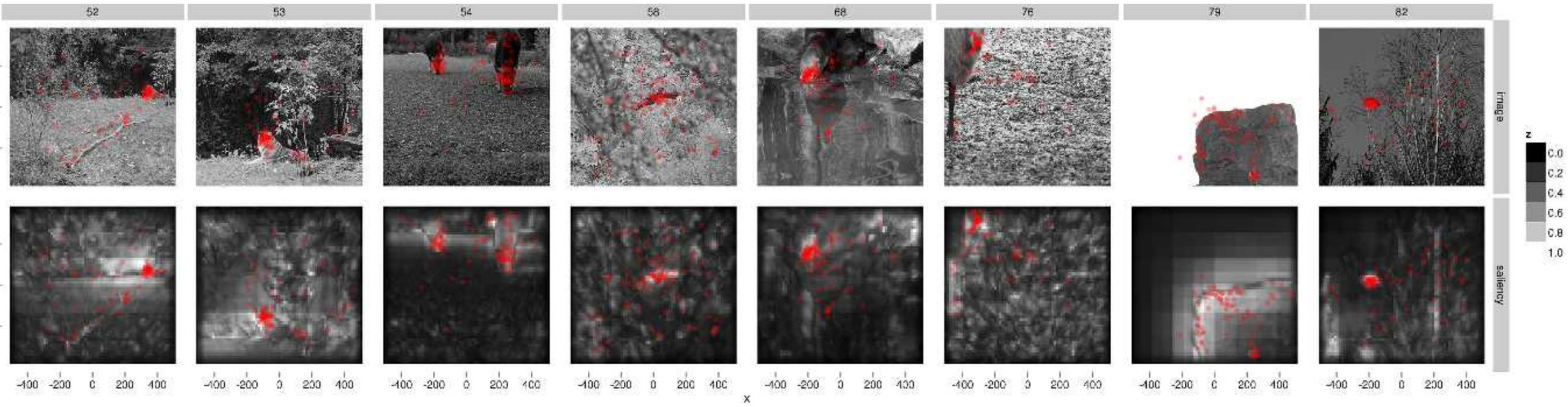
This analysis can be seen as an extended psychophysical receptive or perceptive field analysis, recovering perceptive field—or decision-image—networks.

System identification via reverse-engineering a non-linear kernel machine!

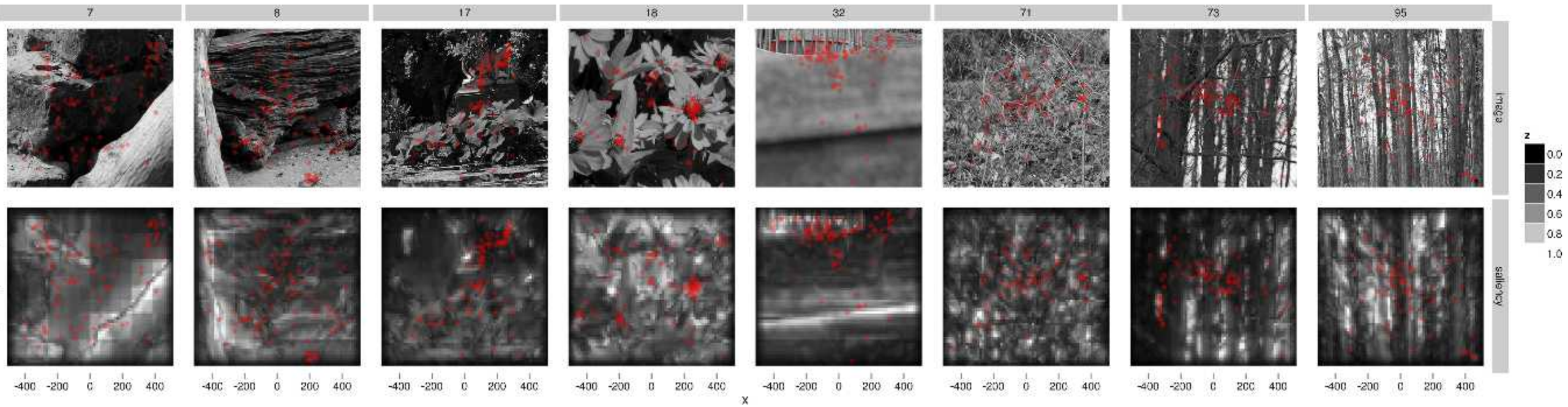
Fixation patterns are stochastic—Itti & Koch and an example image



Itti & Koch: good predictivity on some images ...



Itti & Koch: poor predictivity on other images ...



How to measure the predictivity of visual saliency models?

Use of so-called
spatial point processes
(Barthelmé et al., 2013)

Journal of Vision (2013) 13(12):1, 1–34 http://www.journalofvision.org/content/13/12/1 1

Modeling fixation locations using spatial point processes

Simon Barthelmé Psychology, University of Geneva, Genève, Switzerland ✉

Hans Trukenbrod Psychology, University of Potsdam, Golm, Germany

Ralf Engbert Psychology, University of Potsdam, Golm, Germany

Neural Information Processing Group, Faculty of Science, University of Tübingen, Tübingen, Germany
Bernstein Center for Computational Neuroscience Tübingen, Tübingen, Germany

Felix Wichmann

Max Planck Institute for Intelligent Systems, Empirical Inference Department, Tübingen, Germany

Whenever eye movements are measured, a central part of the analysis has to do with where subjects fixate and why they fixated where they fixated. To a first approximation, a set of fixations can be viewed as a set of points in space; this implies that fixations are spatial data and that the analysis of fixation locations can be beneficially thought of as a spatial statistics problem. We argue that thinking of fixation locations as arising from *point processes* is a very fruitful framework for eye-movement data, helping turn qualitative questions into quantitative ones. We provide a tutorial introduction to some of the main ideas of the field of spatial statistics, focusing especially on spatial Poisson processes. We show how point processes help relate image properties to fixation locations. In particular we show how point processes naturally express the idea that image features' predictability for fixations may vary from one image to another. We review other methods of analysis used in the literature, show how they relate to point process theory, and argue that thinking in terms of point processes substantially extends the range of analyses that can be performed and clarify their interpretation.

many methods that process the raw data and turn it into a more manageable format, checking for calibration, distinguishing saccades from other eye movements (e.g., Engbert & Mergenthaler, 2006; Mergenthaler & Engbert, 2010). Our focus is rather on fixation locations.

In the kind of experiment that will serve as an example throughout the paper, subjects were shown a number of pictures on a computer screen, under no particular instructions. The resulting data are a number of points in space, representing what people looked at in the picture—the fixation locations. The fact that fixations tend to cluster shows that people favor certain locations and do not simply explore at random. Thus one natural question to ask is why are certain locations preferred?

We argue that a very fruitful approach to the problem is to be found in the methods of spatial statistics (Diggle, 2002; Illian, Penttinen, Stoyan, & Stoyan, 2008). A sizeable part of spatial statistics is concerned with how things are distributed in space, and fixations are "things" distributed in space. We will introduce the concepts of point processes and latent

32

Summary

Visual saliency and modelling it has attracted a lot of interest, not least because it has technological and commercial applications, from navigation (landmark detection) to autonomous driving (traffic sign detection), from the military (break camouflage) to image and video compression (only encode salient objects with high quality) to the prediction where people will look at first on WWW-pages (advertising) ... .

Difficult problem because human fixation patterns are clearly stochastic in nature—almost all current models are deterministic, however.

Currently highly successful models on the **mit saliency benchmark** are based on deep convolutional neural network, incorporating high-level features such as faces—do they really measure bottom-up saliency?

SVM-based ML approach suggests the bottom-up component in (boring) images may be based on a rather simple contrast-normalised centre-surround computation (Kienzle et al., 2009).