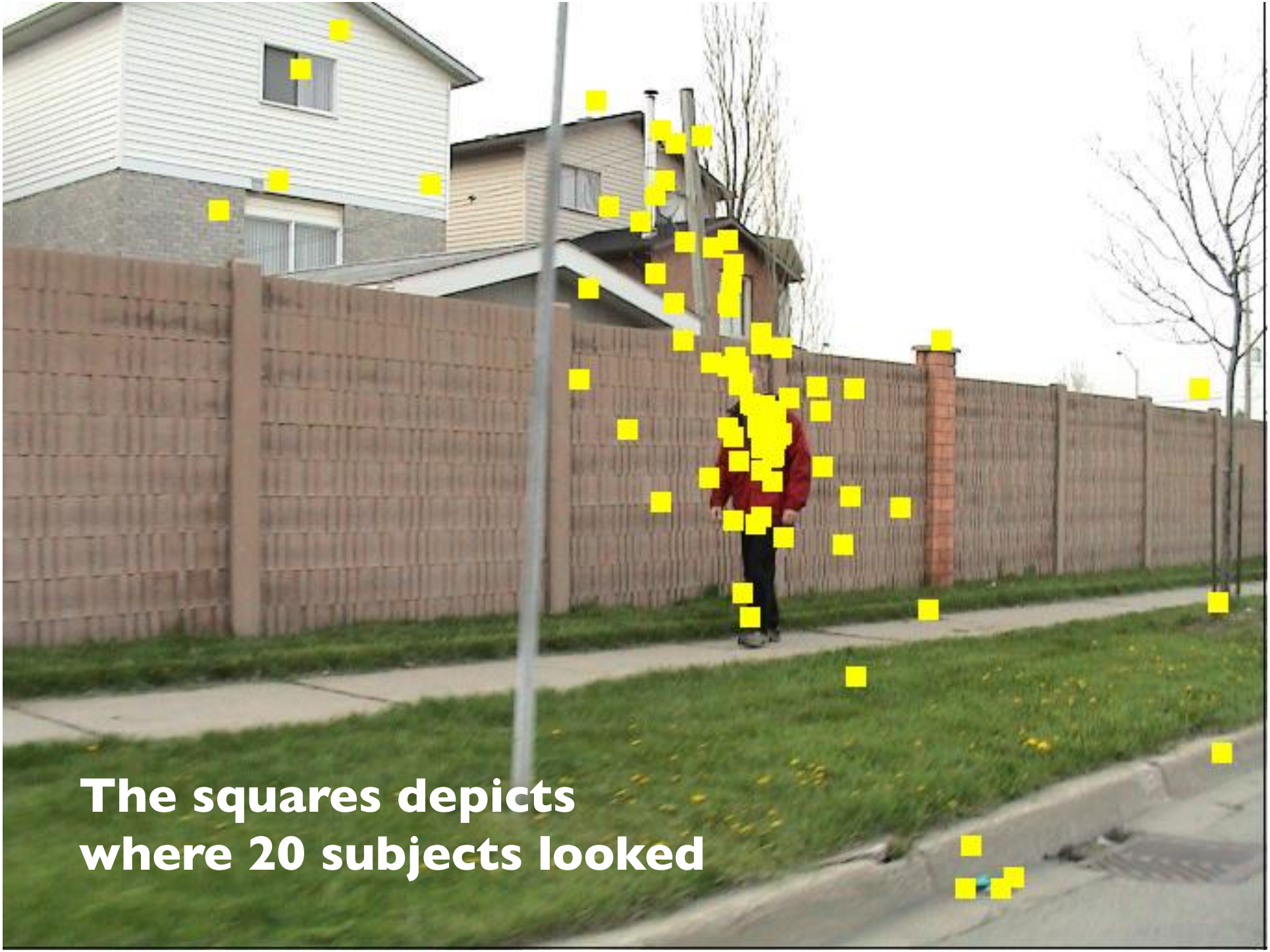
BIL 717
Image Processing

Erkut Erdem
Dept. of Computer Engineering
Hacettepe University

Visual Saliency


Where do you look on this image?


The squares depicts
where 20 subjects looked


Adapted from T. Judd

Adapted from T. Judd


Adapted from T. Judd

Adapted from T. Judd


Adapted from T. Judd

Adapted from T. Judd

Adapted from T .Judd

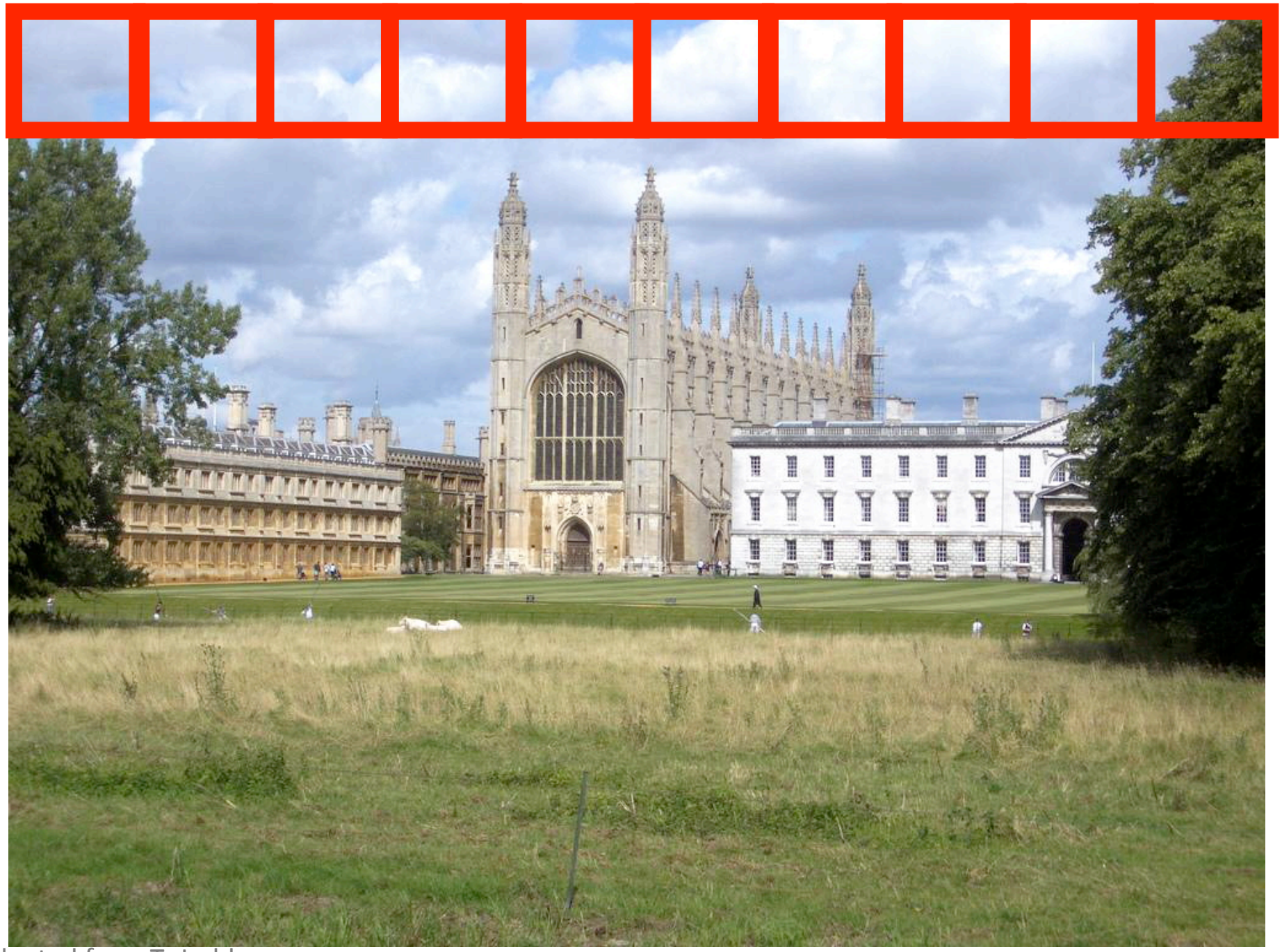
Adapted from T. Judd

Adapted from T. Judd

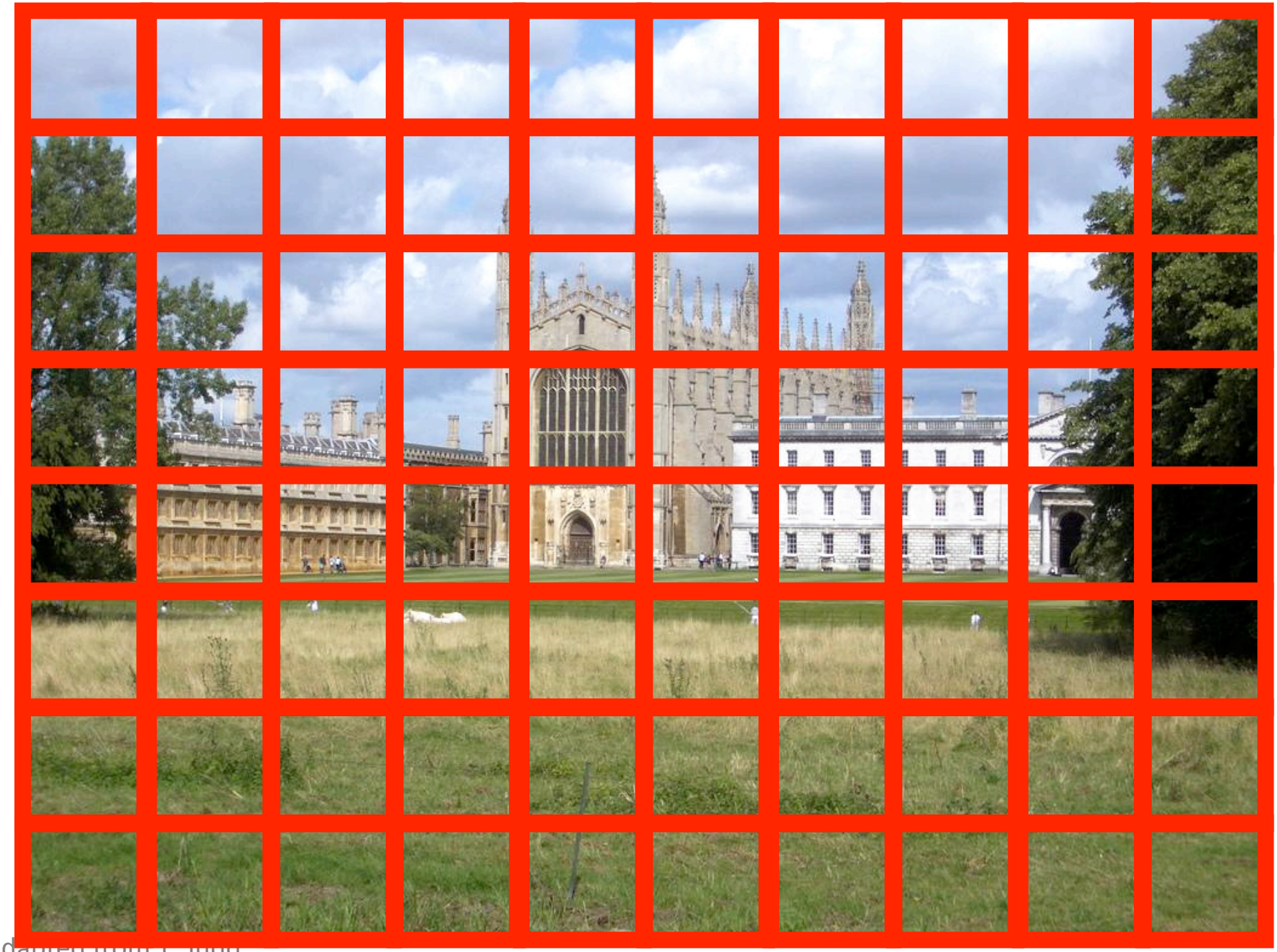
Adapted from T. Judd

Adapted from T. Judd

Adapted from T. Judd

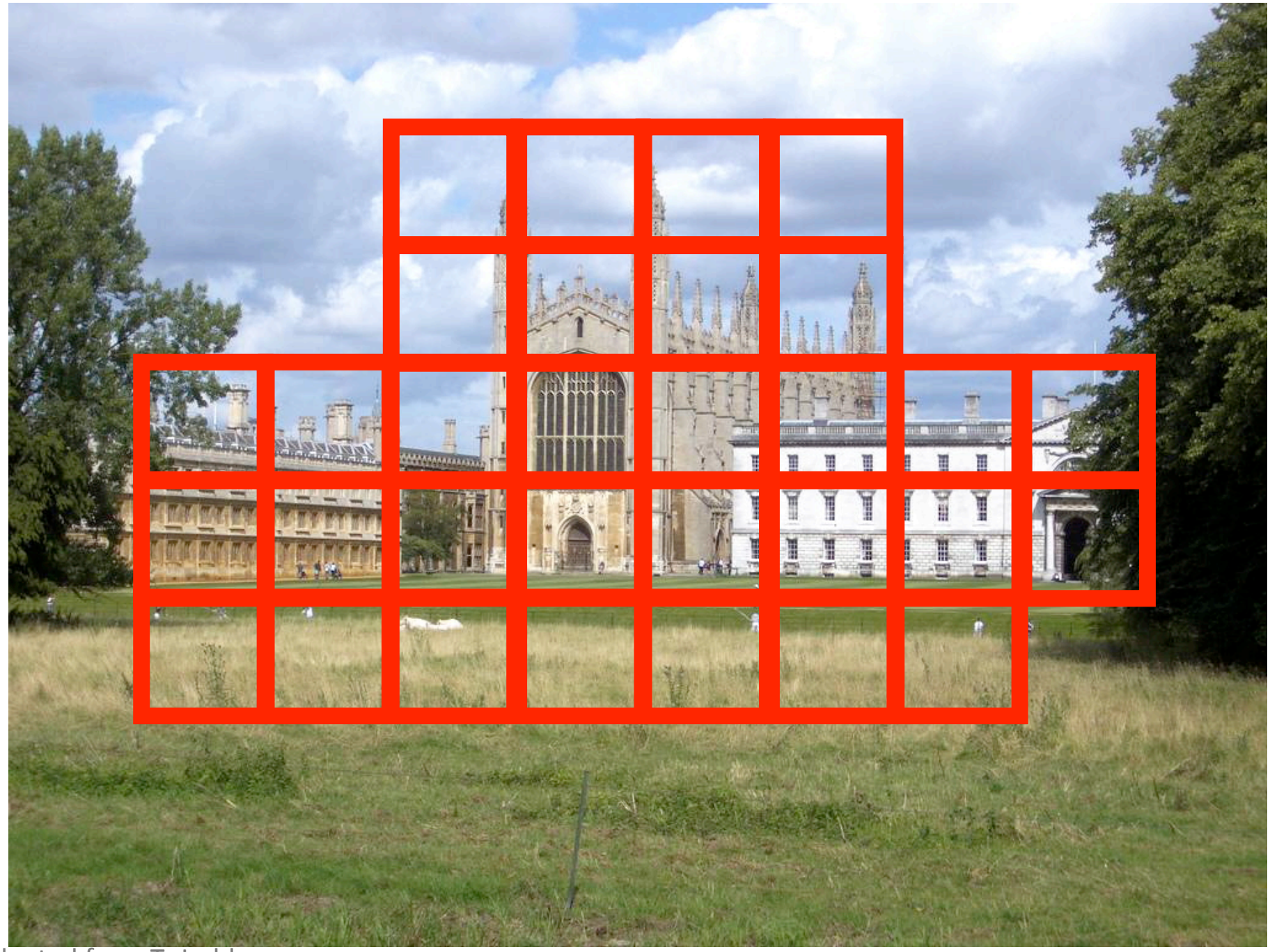


Adapted from T. Judd



Adapted from T. Judd

Adapted from T. Judd

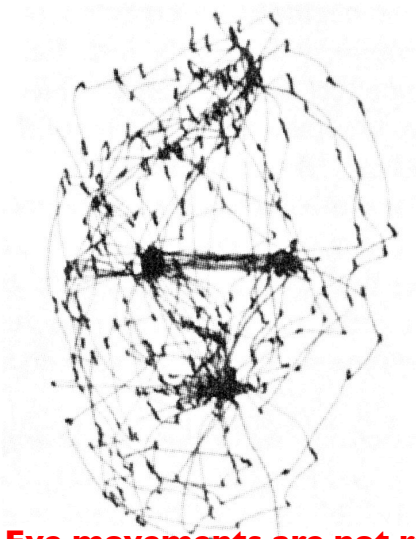


Adapted from T. Judd

Adapted from T. Judd

Scene Analysis and Eye Movements

- Our visual system processes an enormous amount of data coming from the retina. ~10^9 bits/sec (Itti, PhD Thesis, 2000)
- We have developed some **selection mechanisms** to find the most relevant parts of a scene.
  - Warning (noticing predators, sudden motion, etc.)
  - Exploration (finding preys, locating objects, etc.)



Eye movements are not random

Adapted from T. Judd, M. S. Lewicki

Attention and Scene Analysis

- **Attention** is a complex set of interrelated processes.
  - selection of information (bottom-up)
  - integration of that information with existing knowledge (top-down)
- Why do perceptual systems **need attention**?
  - limited resources
  - Even though we have 10^{12} neurons, the brain is still not sufficient to process all the information coming out of the retina
  - simplifies the problem computationally by selecting information
  - perceptual constancy by separating the "foreground" from the "background"

Adapted from M. S. Lewicki

Terminology

- **Gaze** refers to coordinated eye-head movements during shifts in visual attention.
- **Fixation** is the maintaining of the visual gaze on a single location.
- **Saliency** describes the distinctive nature of regions/ objects about how they stand out in relation to their surroundings (grab our attention).

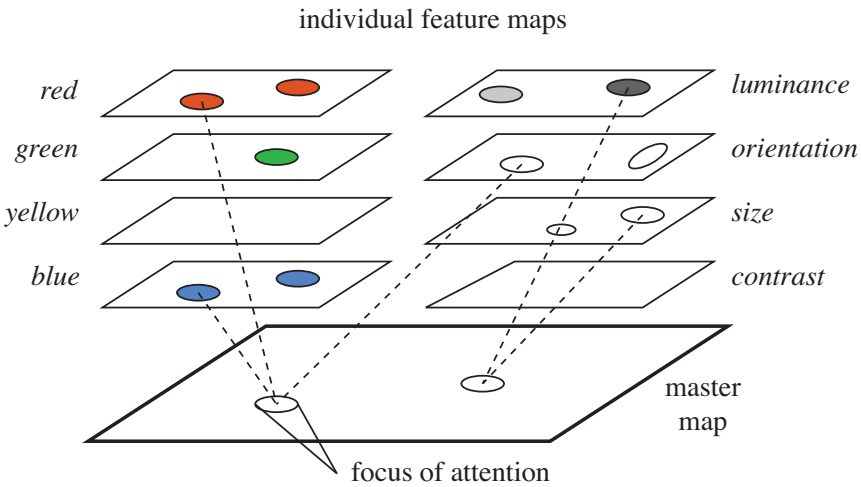
What is attention?

- *"Everyone knows what attention is."*

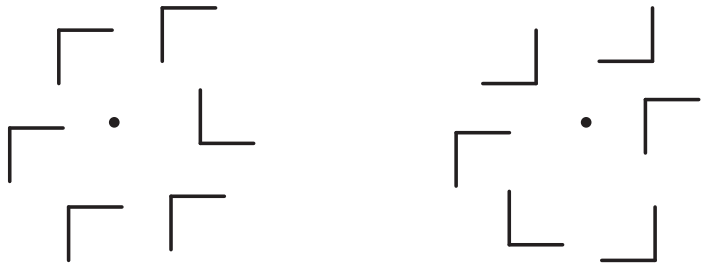
  William James

- Theories of attention
  - Feature-Integration Theory [Treisman and Gelade 1980]
  - Textons [Julezs 1981]
  - Guided Search Theory [Wolfe 1989]
  - Similarity [Duncan and Humphreys 1987]

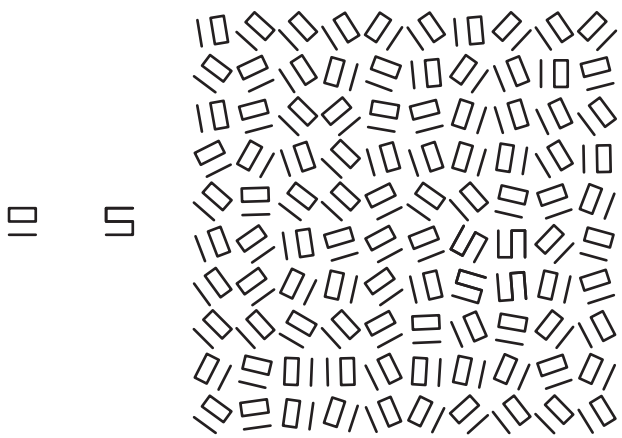
Theories of Visual Attention



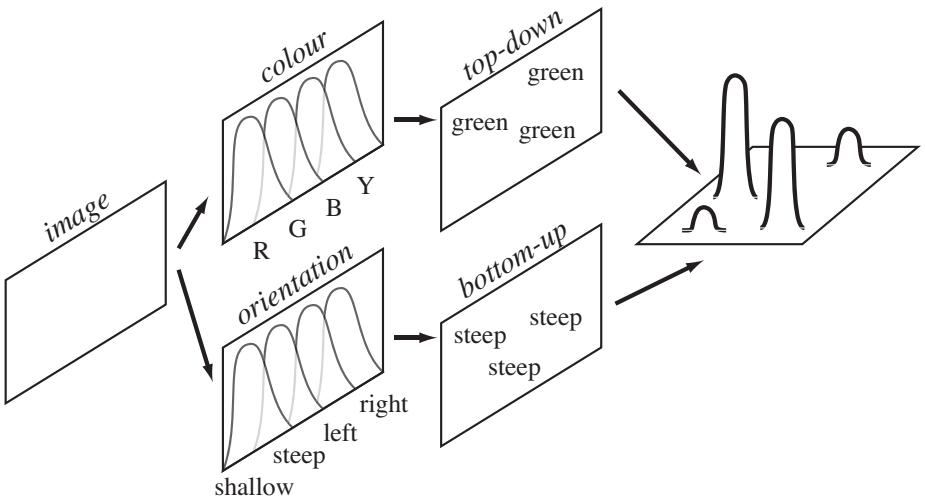
- Feature-Integration Theory [Treisman and Gelade 1980]



- Similarity [Duncan and Humphreys 1987]



- Textons [Julezs 1981]



- Guided Search Theory [Wolfe 1989]

[Healey and Enns 2011]

Attention Mechanisms

- **Attention** is a complex set of interrelated processes
  - selection of information (bottom-up)
  - integration of that information with existing knowledge (top-down)
- Bottom-up attention
  - very rapid, task-independent
- Top-down attention
  - slower, task dependent

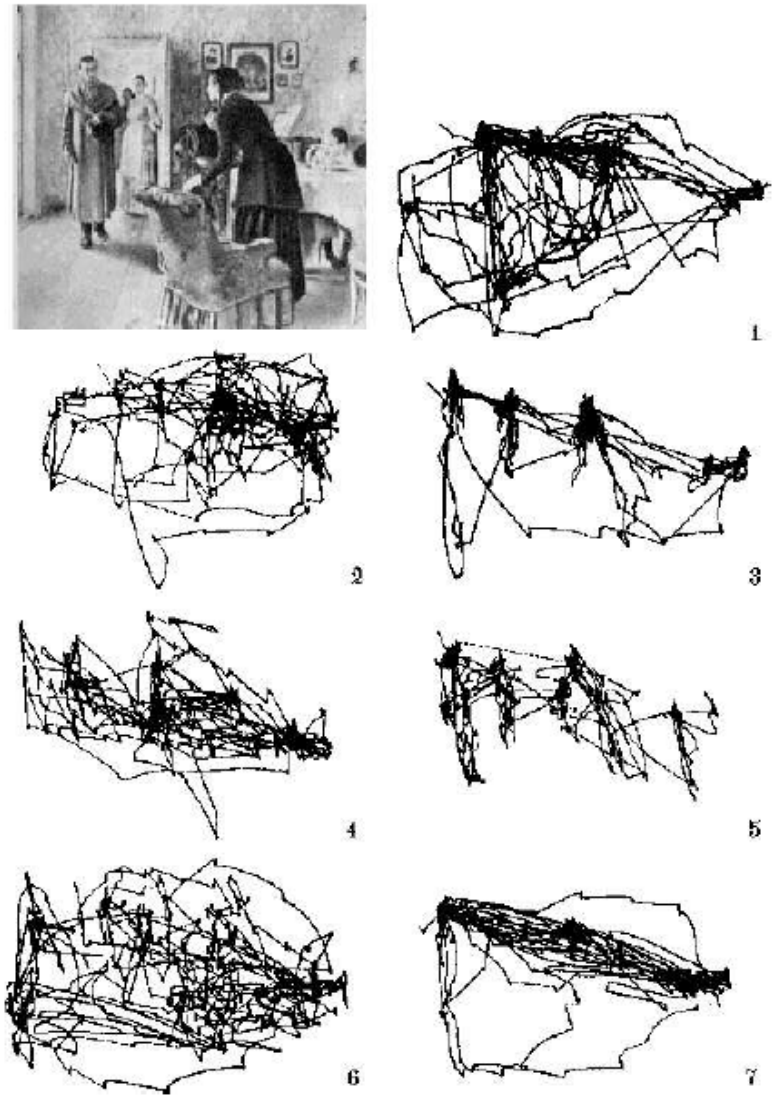
Top-down Attention

"They did not expect him" by Repin

Yarbus (1967) was the first to show that *task* influences fixation locations

Adapted from T. Judd

Saccades and Fixations



Free Examine

What are the material circumstances of the family?

What are their ages?

What were they doing before arrival?

Remember the clothes

Remember object and person positions

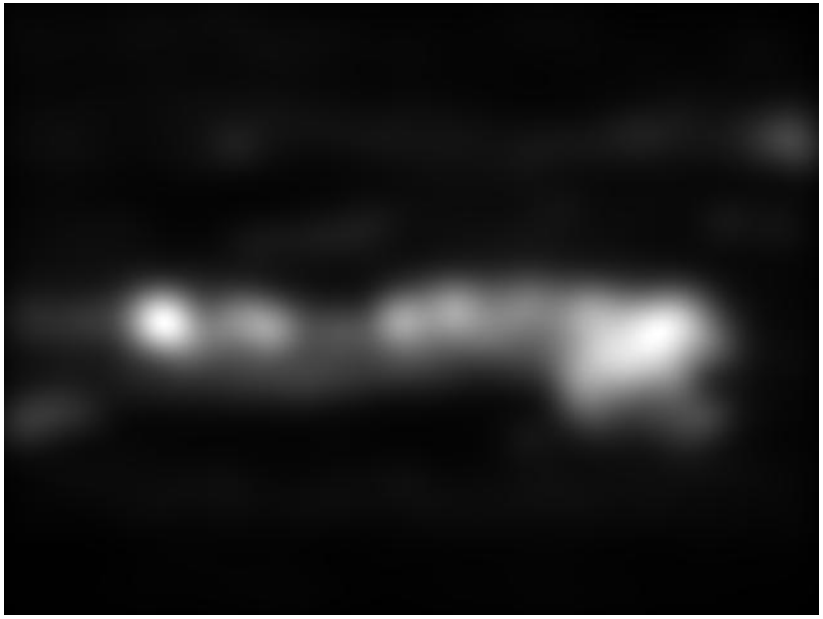
How long has the unexpected visitor been away?

Slide credit: D. Hoeim

[Yarbus 1967]

Computational Models of Visual Attention

- Can machines predict where the humans look at a given image?



Many models of saliency have been introduced

Biologically Inspired

Itti and Koch (1998)

Cerf et al. (2007)

Hou and Zhang (2007)

Rosenholtz (1999)

Itti and Baldi (2006)

Le Meur et al. (2006)

Seo & Milanfar (2009)

Zhang & Cottrell (2008)
SUN model

Goferman et al. (2009)

Achanta (2010)

Mathematically Inspired

Harel et al. (2007)
Graphical Model

Avraham and Lindenbaum (2009)
Esaliency

Bruce and Tsotsos (2009)
Information theoretic approach

Kienzle et al., (2007)

Gao and Vasconulos (2005)

Itti and Baldi (2006)
"Surprise" model

Navalpakkam and Itti (2005)

Elazary and Itti (2010)

Add top-down features

Ehinger et al., (2009)
(search task)

Oliva et al. (2003)

Torralba et al. (2006)

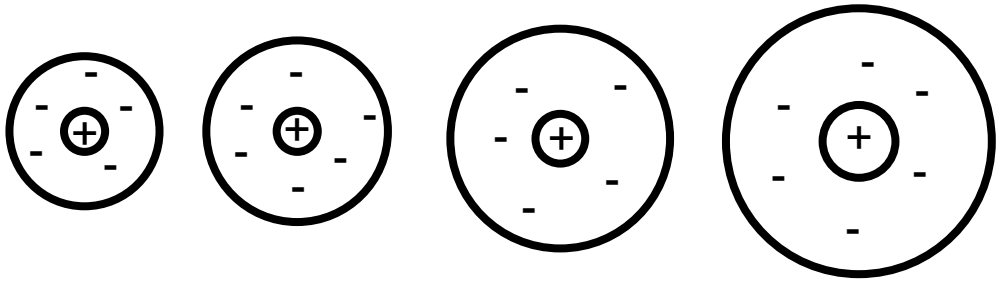
Zhang et al. (2008)

Kanan et al. (2009)

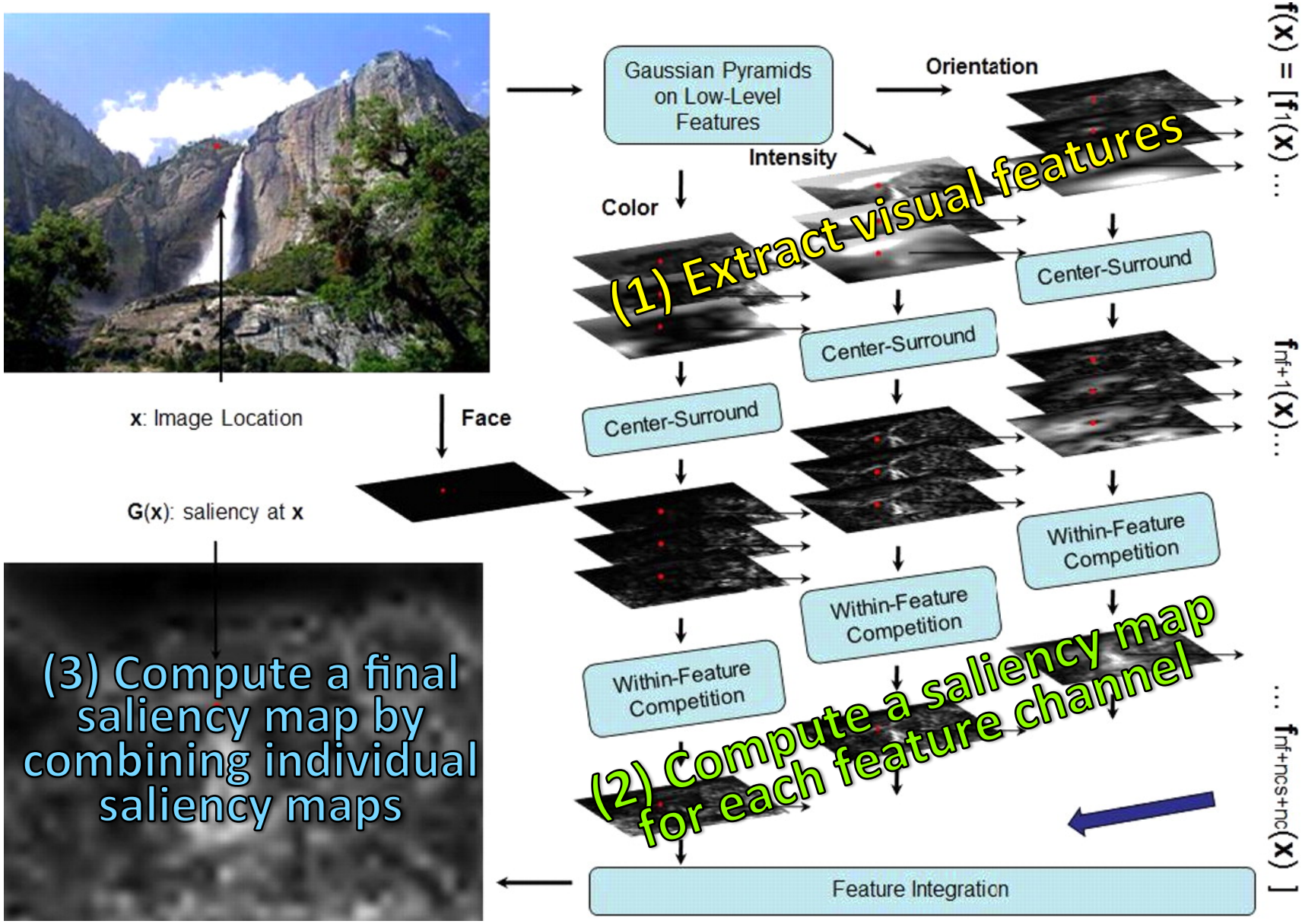
Adapted from T. Judd

Itti-Koch Model (1998)

- *First computational model of visual attention to* predict where people look
- An implementation of Koch & Ullman, 1985
- employs a multi-scale center-surround mechanism which imitates the workings of the retinal receptive field.



Bottom-up Saliency Models



Zhao Q , Koch C J Vis 2012;12:22

Feature Extraction

- Three basic visual features:

1. Intensity Contrast

$$I = (r+g+b)/3$$

2. Color

$$R = r - (g+b)/2 \qquad B = b - (r+g)/2$$
$$G = g - (r+b)/2 \qquad Y = (r+g)/2 - |r-g|/2 - b$$

3. Orientation

Feature Extraction

- Feature maps are extracted by considering center-surround differences between a "center" fine scale *c* and a "surround" coarser scale *s*:

$$c \in \{2, 3, 4\} \qquad s = c + \delta,\ \delta \in \{3, 4\}$$

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)|$$

$$\mathcal{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|$$

$$\mathcal{BY}(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|$$

Color double-opponent model

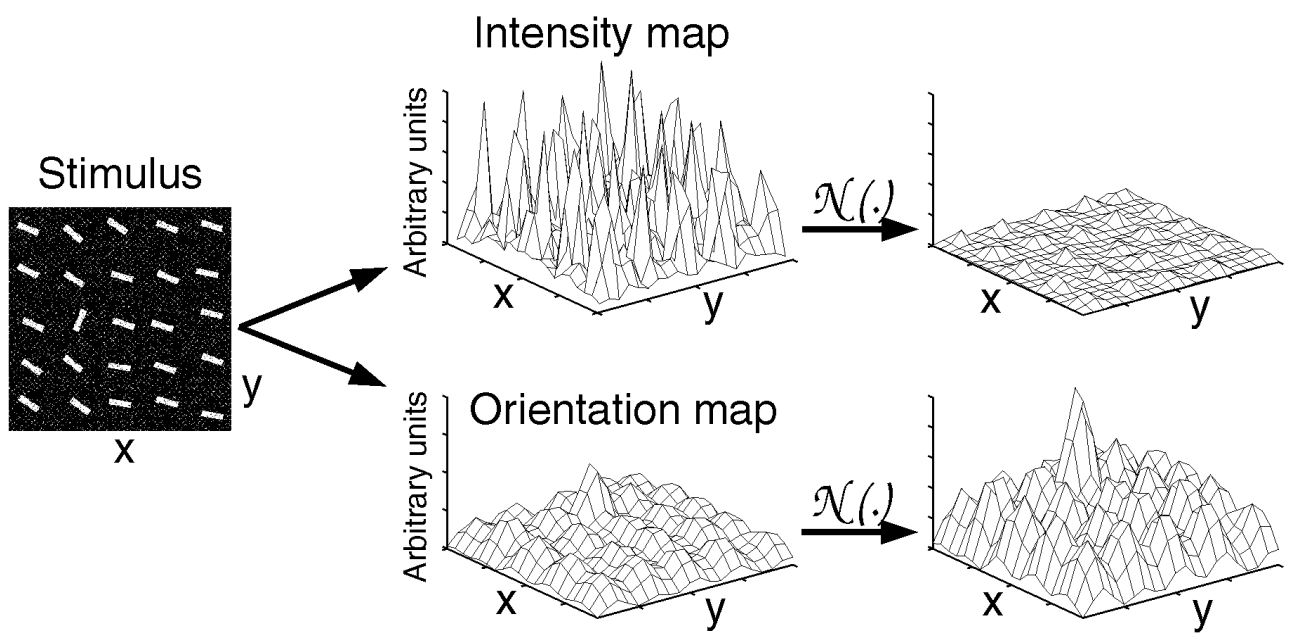
$$\mathcal{O}(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|$$

$$\theta \in \{0°, 45°, 90°, 135°\}$$

In total, 42 different feature maps are estimated: 6 for intensity, 12 for color, and 24 for orientation.

Normalization operator

- Normalize values to fixed range [0..M]
- Compute the average of local maxima: m
- Multiply the feature map by $(M-m)^2$



Saliency Estimation

- Compute individual saliency maps for each visual feature:

$$\overline{\mathcal{I}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} \mathcal{N}\left(\mathcal{I}(c,s)\right)$$

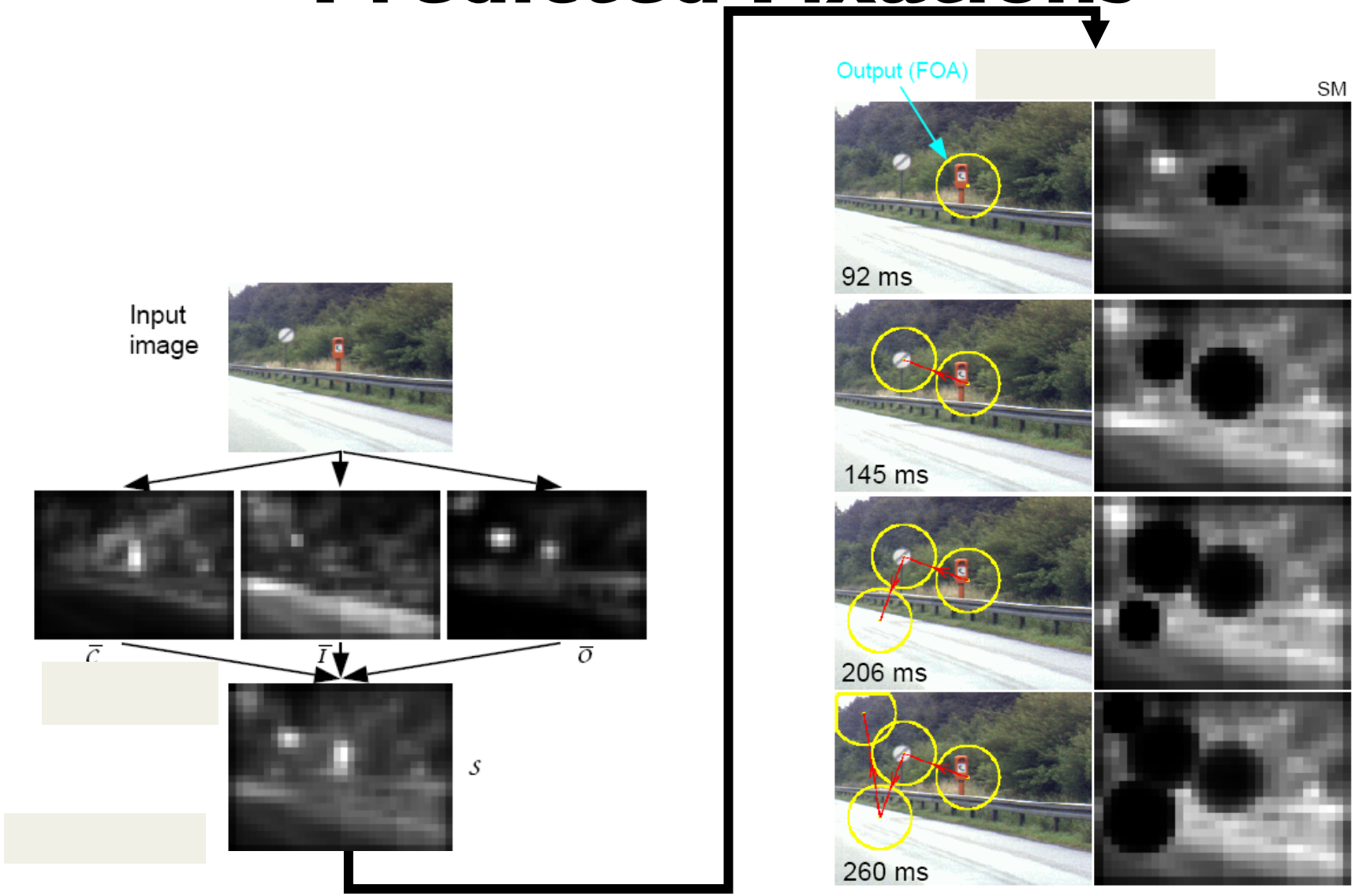
$$\overline{\mathcal{C}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \left[\mathcal{N}\left(\mathcal{R}\mathcal{G}(c,s)\right) + \mathcal{N}\left(\mathcal{B}\mathcal{Y}(c,s)\right)\right]$$

$$\overline{\mathcal{O}} = \sum_{\theta \in \{0°, 45°, 90°, 135°\}} \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}\left(\mathcal{O}(c,s,\theta)\right)\right)$$

- Combine these individual saliency maps by averaging:

$$\mathcal{S} = \frac{1}{3}\left(\mathcal{N}\left(\overline{\mathcal{I}}\right) + \mathcal{N}\left(\overline{\mathcal{C}}\right) + \mathcal{N}\left(\overline{\mathcal{O}}\right)\right)$$

Predicted Fixations



Slide credit: D. Hoeim

[Itti Koch Niebur 1998]

GBVS (Harel et al, 2007)

- A graph-based bottom-up model
- Several feature maps are extracted at multiple spatial scales and then represented as fully connected graphs.
  - Vertices: grid positions,
  - Edges: relationships between pairs of vertices

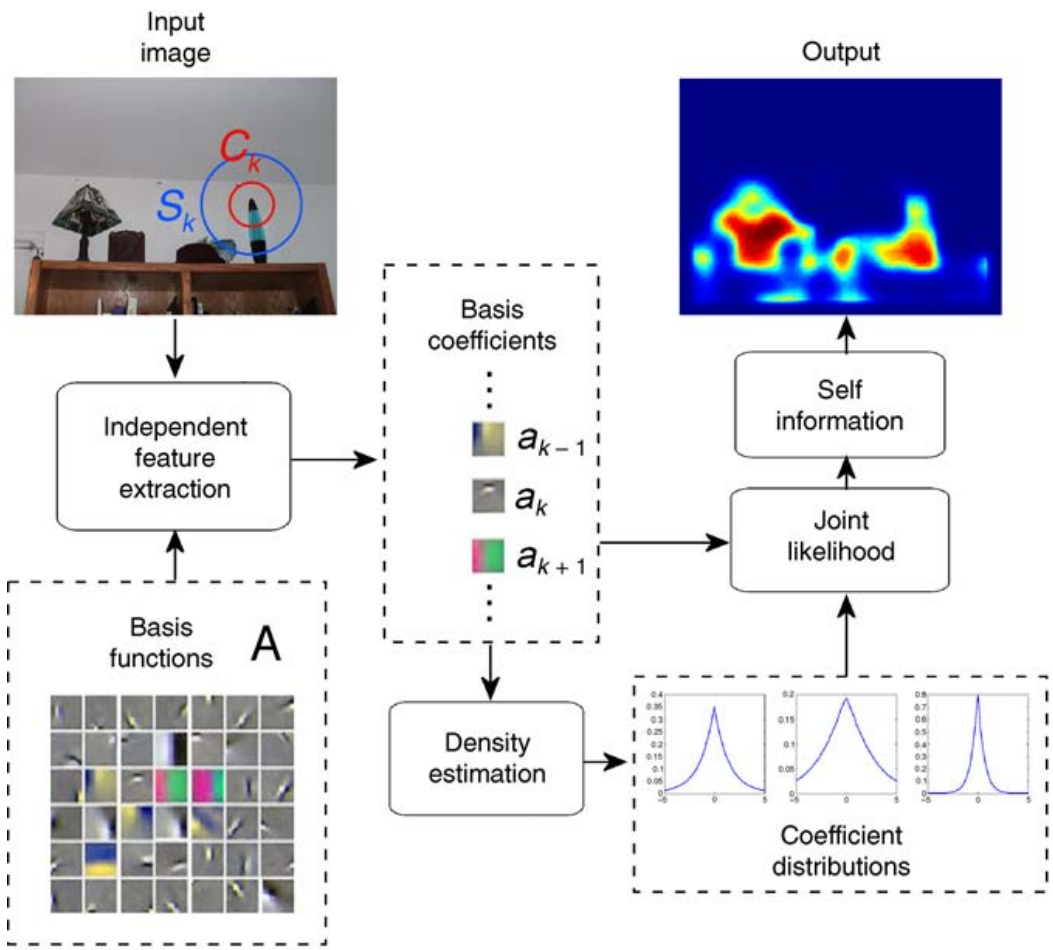
$$d((i,j)||(p,q)) \triangleq \left|\log \frac{M(i,j)}{M(p,q)}\right|$$

$$\begin{aligned} w_1((i,j),(p,q)) &\triangleq d((i,j)||(p,q)) \cdot F(i-p, j-q) \\ F(a,b) &\triangleq \exp\left(-\frac{a^2+b^2}{2\sigma^2}\right). \end{aligned}$$

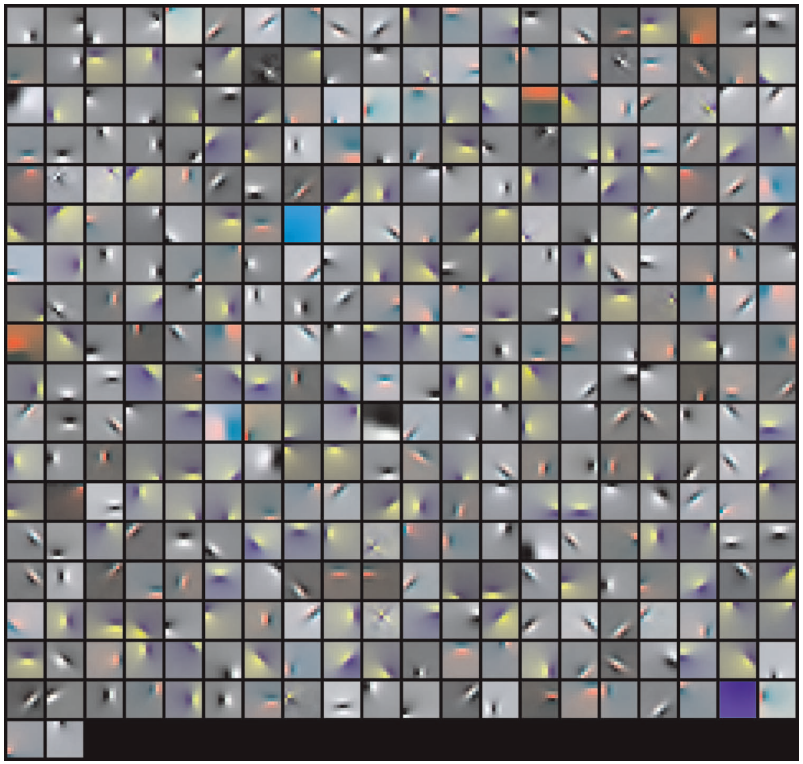
- The resulting graphs are used to define Markov chains in which their equilibrium distribution is treated as in the activation and saliency maps.

AIM (Bruce and Tsotsos, 2006, 2009)

- A bottom-up saliency model based on an information-theoretic perspective. (Shannon's measure of Self-Information)
- A sparse set of basis functions is learned from the input image itself via ICA.
- Features representing the center region and its surrounding regions are described by means of these basis functions.



SUN (Zhang et al., 2009)



- Similar to AIM model, it employs high-level features derived via a visual dictionary, but these features are learned from a training set of natural images.
- Two sets of features based on: **1.** Difference of Gaussians (DoG) filters and **2.** ICA
- A global definition of saliency in which the salient parts are estimated by considering the global rarity of the local visual features in the entire image.

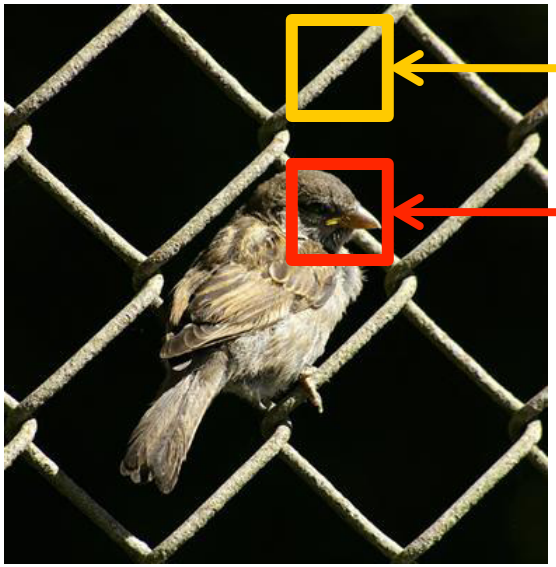
CSD (Goferman et al., 2010)

- A context-aware saliency model based on four principles:

1. Local color and contrast information
2. Frequently occurring global features (maintain unique features)
3. Visual organization rules (Gestalt)
4. High-level features (Faces, Objects, People, etc.)

CSD (Goferman et al., 2010)

- Unique appearance ➔ salient
- Position is important!



Not salient

salient

$$d(p_i, p_j)=d_{color}(p_i, p_j)/(1+c.d_{pos}(p_i, p_j))$$

High for K most similar

salient

Adapted from S. Goferman

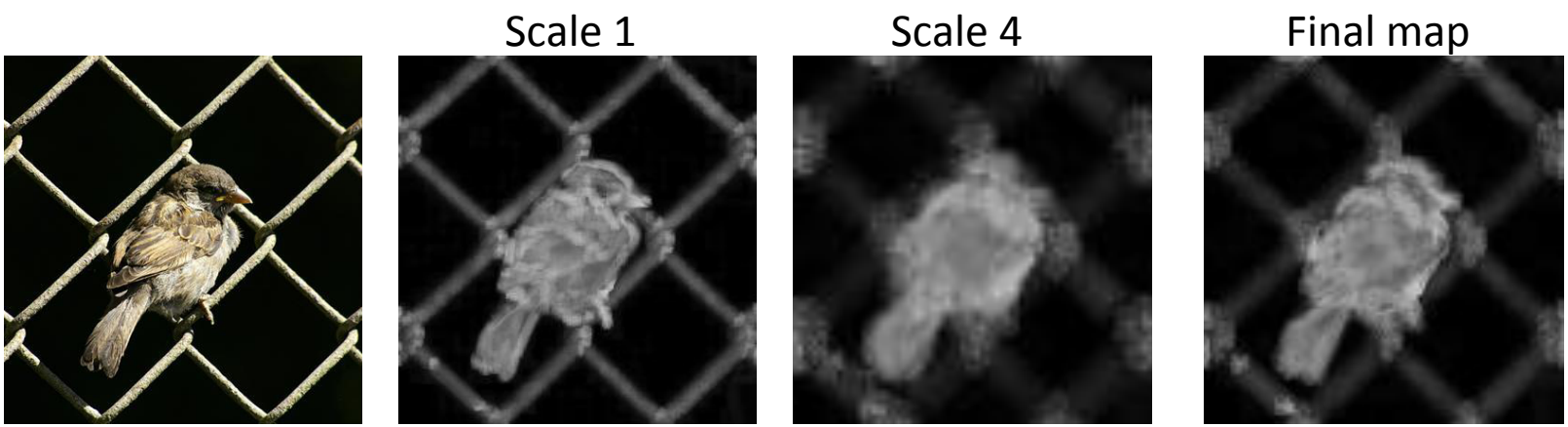
Single and multi-scale saliency

- A multi-scale analysis for saliency estimation:

$$S_i^r = 1 - exp\{-\frac{1}{K}\sum_{k=1}^{K} d\left(p_i^r, q_k^{r_k}\right)\} \qquad \bar{S}_i = \frac{1}{M}\sum_{r \in R} S_i^r$$

Single scale saliency formula

multi-scale saliency formula

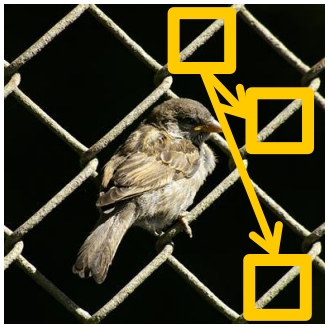


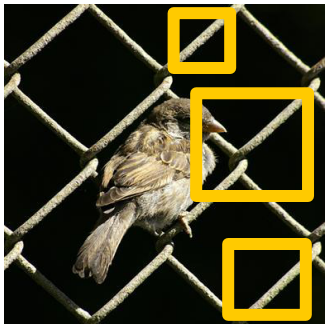
Adapted from S. Goferman

CSD (Goferman et al., 2010) - Summary

- Single-scale saliency

$$S_i^r = 1 - exp\{-\frac{1}{K}\sum_{k=1}^{K} d\left(p_i^r, q_k^{r_k}\right)\}$$



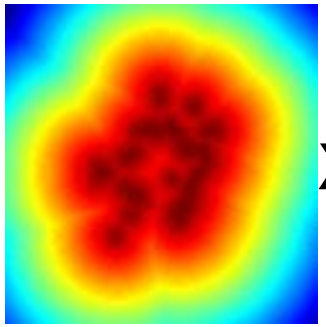


- Multiple scales

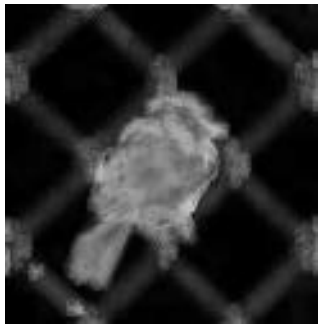
$$\bar{S}_i = \frac{1}{M}\sum_{r \in R} S_i^r$$

- Final saliency

$$\hat{S}_i = \bar{S}_i(1 - d_{foci}(i))$$



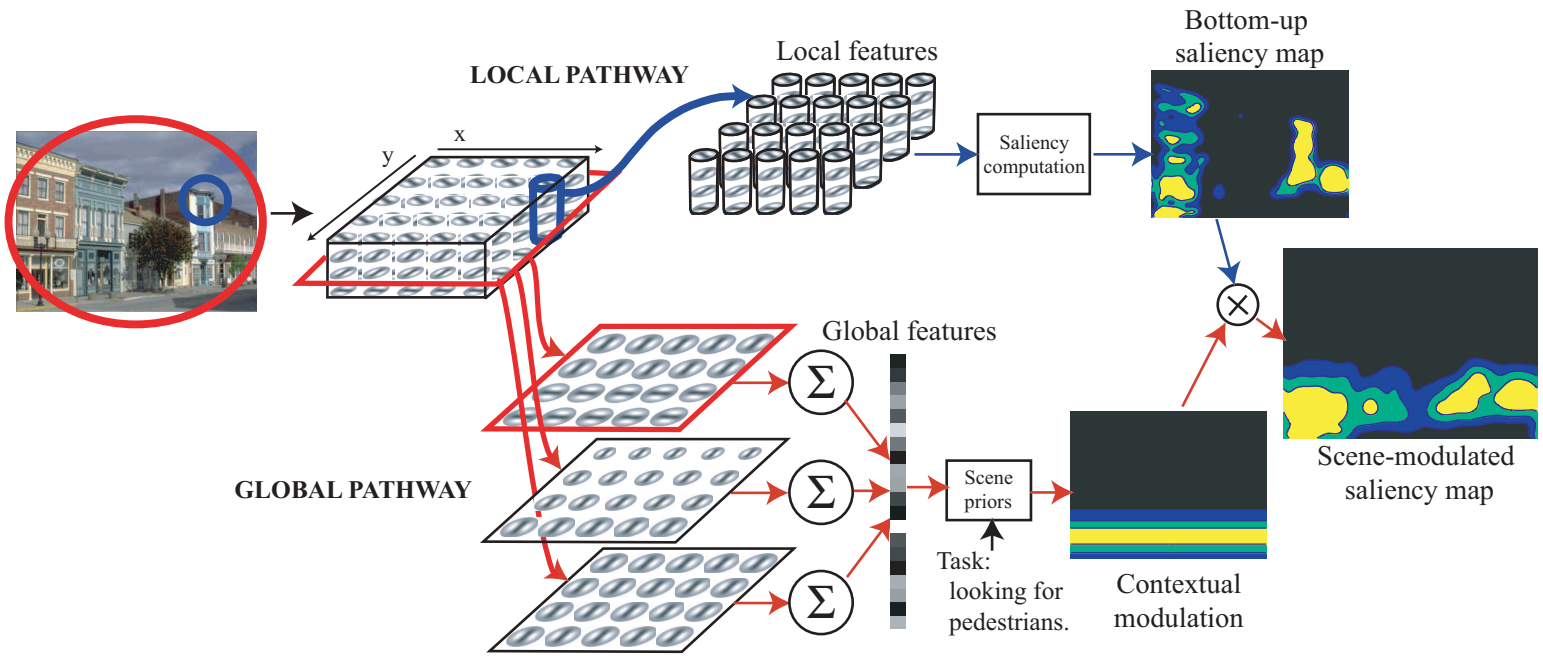
x



Adapted from S. Goferman

Torralba et al., 2006

- A Bayesian contextual guidance model for visual search tasks which combines low-level salience and scene context:



$$p(O=1,X|L,G) = \frac{1}{p(L|G)} p(L|O=1,X,G)\,p(X|O=1,G)\,p(O=1|G)$$

Bottom-up factors — Top-down target knowledge — Context-based priors — Target presence

Learning Saliency (Judd et al, 2009)

- Formulate saliency estimation as a supervised learning problem.
- Learn a classification function $f : R^D \rightarrow \{+1, -1\}$ with D denoting the dimension of the features, which returns the label (+1) for salient (fixated) points, and (−1) for the non-salient points.
- Based on training SVMs to determine the saliency of a pixel.
- Simple low-level features like intensity, orientation and color along with some mid- and high-level features such as responses of horizon line detectors, face and pedestrian detectors.
- <u>Learning of specific feature weights and normalization schemes in the integration step.</u>

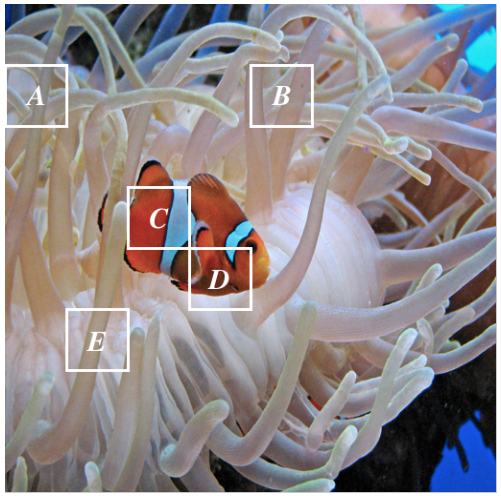
Feature Integration

- It still remains an open question how different feature dimensions contribute to the overall saliency.
- Common strategies are:
  - summation, max, etc.
- Some studies try to overcome this issue by finding optimal values for feature weights in a supervised manner as in Judd et al., 2009.

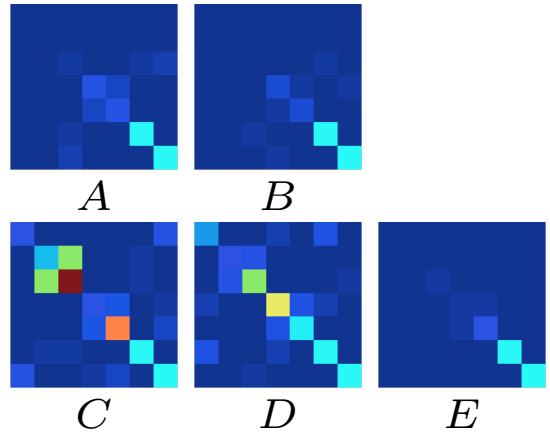
CovSal (Erdem and Erdem, 2012)

- Covariance matrices of simple image features (aka region covariance; Tuzel, Porikli, & Meer, 2006) are used as meta-features for saliency estimation.
- Region covariances capture local image structures better than standard linear filters.
- More importantly, they naturally provide nonlinear integration of different features by modeling their correlations.

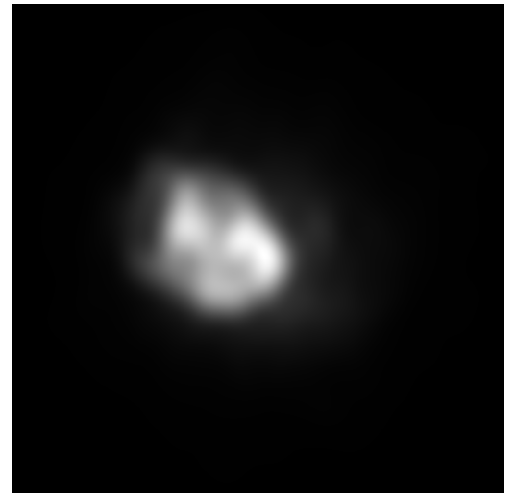
CovSal (Erdem and Erdem, 2012)



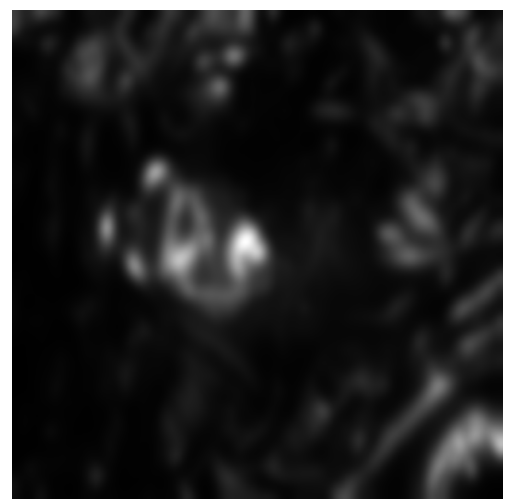
(a) Input image



(b) Covariance matrices for the regions A-E



(c) Our saliency



(d) Itti's saliency

- Very simple visual features:

$$F(x,y) = \left[x \quad y \quad L(x,y) \quad a(x,y) \quad b(x,y) \quad \left| \frac{\partial I(x,y)}{\partial x} \right| \quad \left| \frac{\partial I(x,y)}{\partial y} \right| \right]^T$$

Region Covariances

- Let I denote an image, and F be the feature image extracted from I:

$$F(x,y) = \Phi(I,x,y)$$

  with Φ denoting the d-dimensional function of features.

- Then, a region R inside F can be represented with a d×d covariance matrix C_R of the feature points:

$$\mathbf{C}_R = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{f}_i - \boldsymbol{\mu})(\mathbf{f}_i - \boldsymbol{\mu})^T$$

- Distance between two covariances C_1 and C_2 can be estimated using special metrics:

$$\rho(\mathbf{C}_1,\mathbf{C}_2) = \sqrt{\sum_{i=1}^{n} ln^2 \lambda_i(\mathbf{C}_1,\mathbf{C}_2)}$$

Covariance matrices do not lie on Euclidean space.

Model 1: Saliency using covariance features

- The saliency of a region R_i is defined as the weighted average of the dissimilarities between R_i to the m-most similar regions around it:

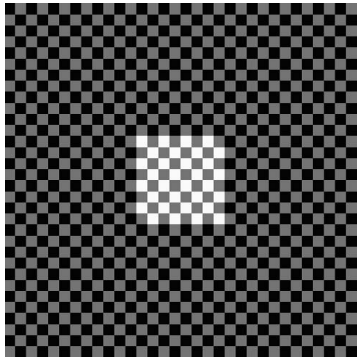
$$S(R_i) = \frac{1}{m}\sum_{j=1}^{m} d(R_i, R_j)$$

- Dissimilarity measure $d(R_i, R_j)$ is defined as:

Weighting the covariance distance by inverse spatial distance makes the nearby regions have more influence on the saliency computation

$$d(R_i, R_j) = \frac{\rho(\mathbf{C}_i, \mathbf{C}_j)}{1 + ||\mathbf{x}_i - \mathbf{x}_j||}$$

Model 2: Saliency using covariance and mean features



- First-order statistics can be also valuable for saliency.
- <u>Sigma Points Representation:</u>
  Every SPD matrix has a unique factorization.
- Let C be a d×d covariance matrix,
  the corresponding set of Sigma Points can be computed as:

$$\mathbf{s}_i = \begin{cases} \alpha\sqrt{d}\mathbf{L}_i & \text{if } 1 \leq i \leq d \\ -\alpha\sqrt{d}\mathbf{L}_i & \text{if } d+1 \leq i \leq 2d \end{cases}$$

  where L_i is the ith column of the lower triangular matrix L obtained with the Cholesky decomposition $C=LL^T$.

- First-order statistics can be easily incorporated to obtain an enriched feature vector: $\Psi(\mathbf{C}) = (\mu, \mathbf{s}_1, \ldots \mathbf{s}_d, \ , \mathbf{s}_{d+1}, \ldots, \mathbf{s}_{2d})^T$

Model 2: Saliency using covariance and mean features

- The saliency of a region R_i is given b:

$$S(R_i) = \frac{1}{m}\sum_{j=1}^{m} d'(R_i, R_j)$$

- Dissimilarity measure $d(R_i, R_j)$ is defined as:

$$d'(R_i, R_j) = \frac{||\Psi(\mathbf{C}_i) - \Psi(\mathbf{C}_j)||}{1 + ||\mathbf{x}_i - \mathbf{x}_j||}$$

Incorporating center bias

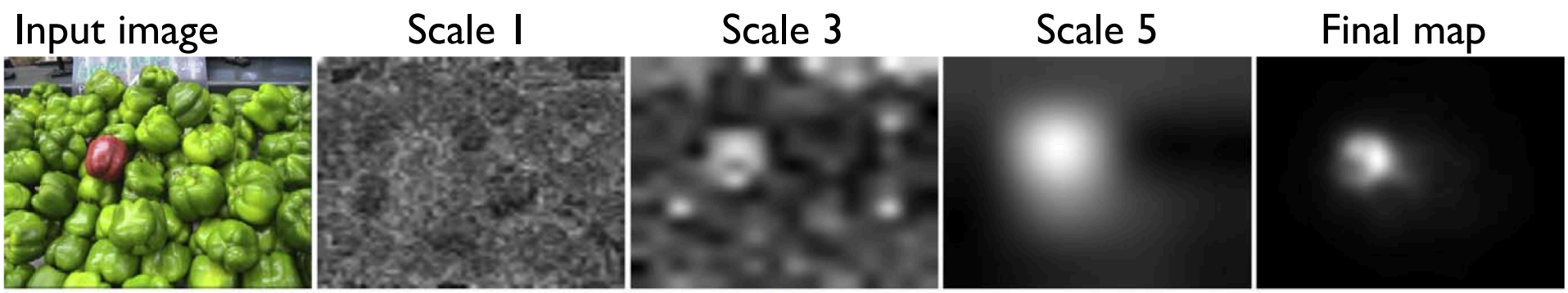
- Experiments on human eye fixations demonstrate that there is a tendency in humans to look towards the image center, which is called the center bias.
- This bias is mainly explained by several factors:
  - Photographer bias,
  - Viewing strategy,
  - Motor bias
- We included a center bias into our second model by defining the saliency of region R_i as follows:

$$S'(R_i) = \left(1 - \frac{||\mathbf{x}_i - \mathbf{x}_c||}{Z}\right) \cdot S(R_i)$$

Scale-space extension

- The objects that can be treated as salient in an image can and do appear over a wide range of scales.
- Saliency detection should be carried out simultaneously at all possible scales.
- Employ a fusion strategy to combine single-scale maps to come up with one final saliency map:

$$S(x) = G_\sigma(x) * \prod_{k \in K} \hat{S}^k(x)$$



Many models of saliency have been introduced

Biologically Inspired

Itti and Koch (1998)

Cerf et al. (2007)

Hou and Zhang (2007)

Rosenholtz (1999)

Itti and Baldi (2006)

Le Meur et al. (2006)

Seo & Milanfar (2009)

Zhang & Cottrell (2008)
SUN model

Goferman et al. (2009)

Achanta (2010)

Mathematically Inspired

Harel et al. (2007)
Graphical Model

Avraham and Lindenbaum (2009)
Esaliency

Bruce and Tsotsos (2009)
Information theoretic approach

Kienzle et al., (2007)

Gao and Vasconulos (2005)

Itti and Baldi (2006)
"Surprise" model

Navalpakkam and Itti (2005)

Elazary and Itti (2010)

Add top-down features

Ehinger et al., (2009)
(search task)

Oliva et al. (2003)

Torralba et al. (2006)

Zhang et al. (2008)

Kanan et al. (2009)

Adapted from T. Judd

Which one is the best?

Biologically Inspired

Itti and Koch (1998)

Cerf et al. (2007)

Hou and Zhang (2007)

Rosenholtz (1999)

Itti and Baldi (2006)

Le Meur et al. (2006)

Seo & Milanfar (2009)

Zhang & Cottrell (2008)
SUN model

Goferman et al. (2009)

Achanta (2010)

Mathematically Inspired

Heral et al. (2007)
Graphical Model

Avraham and Lindenbaum (2009)
Esaliency

Bruce and Tsotsos (2009)
Information theoretic approach

Kienzle et al., (2007)

Gao and Vasconulos (2005)

Itti and Baldi (2006)
"Surprise" model

Navalpakkam and Itti (2005)

Elazary and Itti (2010)

Add top-down features

Ehinger et al., (2009)
(search task)

Oliva et al. (2003)

Torralba et al. (2006)

Zhang et al. (2008)

Kanan et al. (2009)

Adapted from T. Judd

Which one is the best?

Biologically Inspired

Itti and Koch (1998)

Cerf et al. (2007)

Hou and Zhang (2007)

Rosenholtz (1999)

Itti and Baldi (2006)

Le Meur et al. (2006)

Seo & Milanfar (2009)

Zhang & Cottrell (2008)
SUN model

Goferman et al. (2009)

Achanta (2010)

Mathematically Inspired

Heral et al. (2007)
Graphical Model

Avraham and Lindenbaum (2009)
Esaliency

Bruce and Tsotsos (2009)
Information theoretic approach

Kienzle et al., (2007)

Gao and Vasconulos (2005)

Itti and Baldi (2006)
"Surprise" model

Navalpakkam and Itti (2005)

Elazary and Itti (2010)

Add top-down features

Ehinger et al., (2009)
(search task)

Oliva et al. (2003)

Torralba et al. (2006)

Zhang et al. (2009)

Kanan et al. (2009)

Benchmark needed!

Adapted from T. Judd

How benchmark data sets are formed?

screen resolution 1280x1024

each image shown for 3 seconds

eye tracker measures location of eye fixation several times a second.

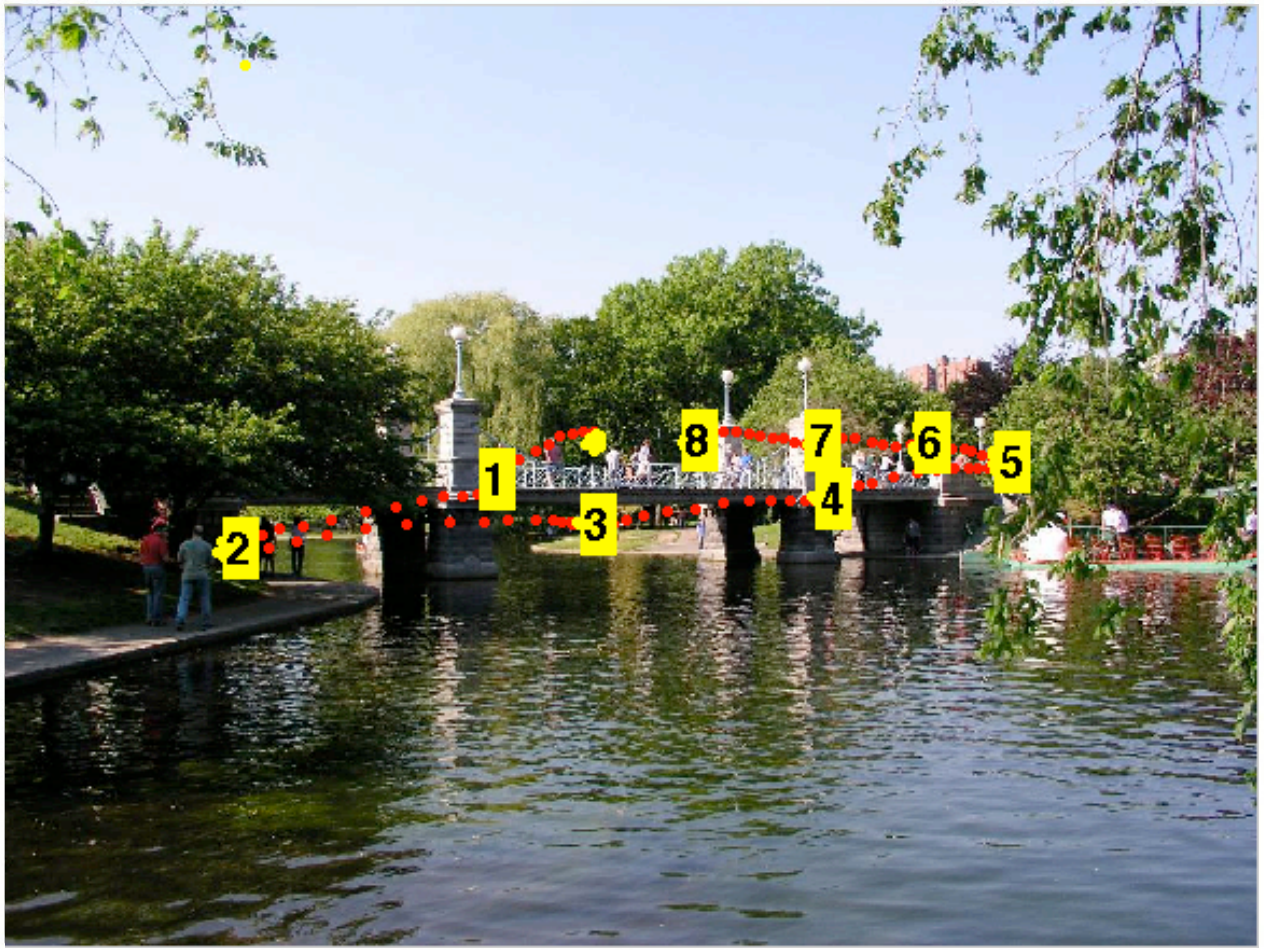
user rests head in chin rest

[Photo Credit: Jason Dorfman CSAIL website]

Through eye tracking experiments

Adapted from T. Judd

Fixations for one observer

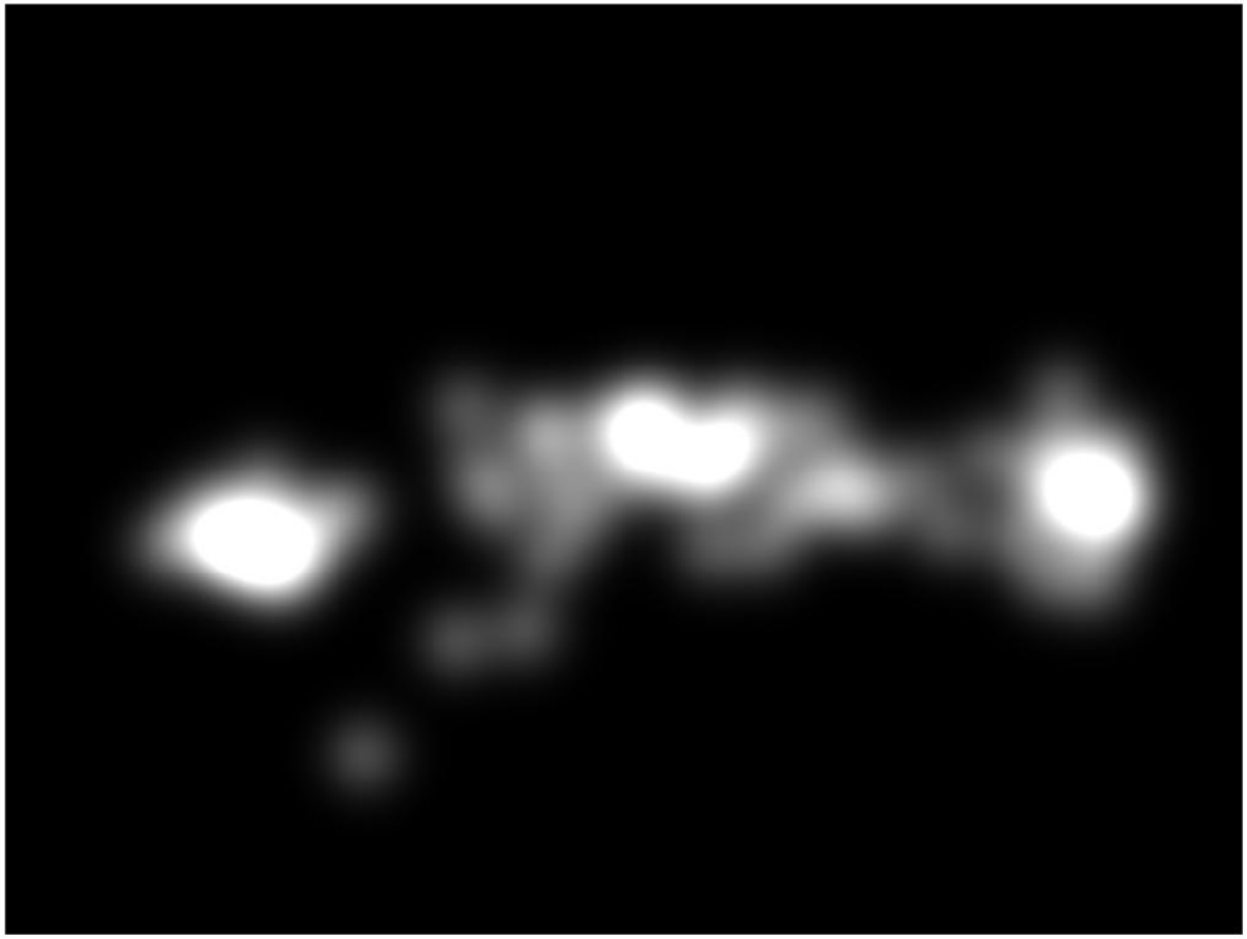


Adapted from T. Judd

Fixations from 15 observers

Adapted from T. Judd

Fixation map created from gaussian convolution over fixations



Fixation map

Adapted from T. Judd

Human Eye Fixations Data Sets

- Bruce data set [Bruce and Tsotsos 2006]
  - 120 color images of indoor and outdoor scenes
  - Eye movement data from 20 subjects
- MIT 1003 data set [Judd et al. 2009]
  - 1003 natural color images (779 landscape + 228 portrait)
  - Eye movement data from 15 subjects
- MIT 300 data set [Judd et al. 2012]
  - 300 natural images (223 landscape + 77 portrait)
  - Eye movement data from 39 subjects

Sample Images

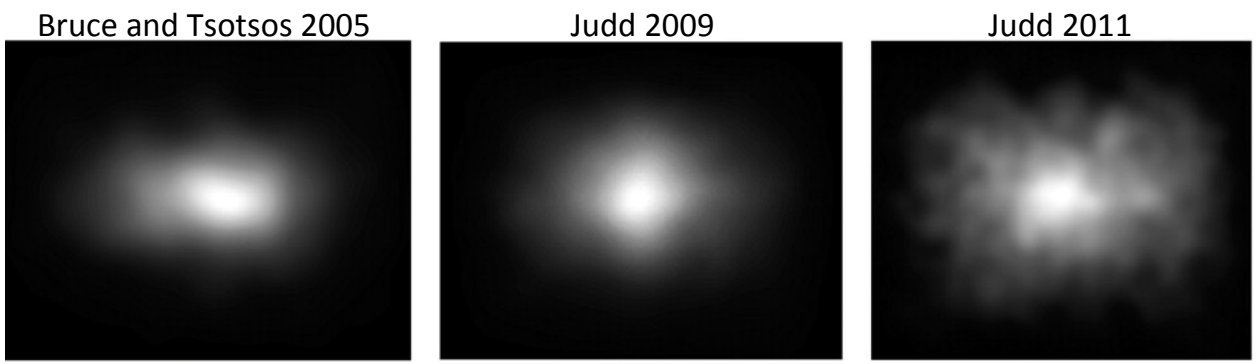
Bruce data set

MIT 1003 data set

MIT 300 data set

Why are fixations center biased?

- photographer bias
- viewing strategy



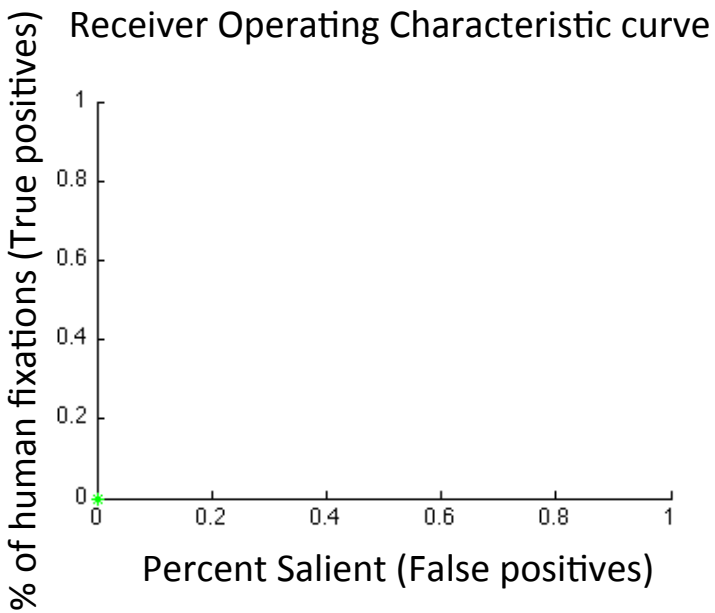
Adapted from T. Judd

Evaluation Metrics

- Area under ROC curve (AUC)
- Earth Mover's Distance (EMD)
- Normalized Scanpath Saliency (NSS)
- Similarity Score
- …

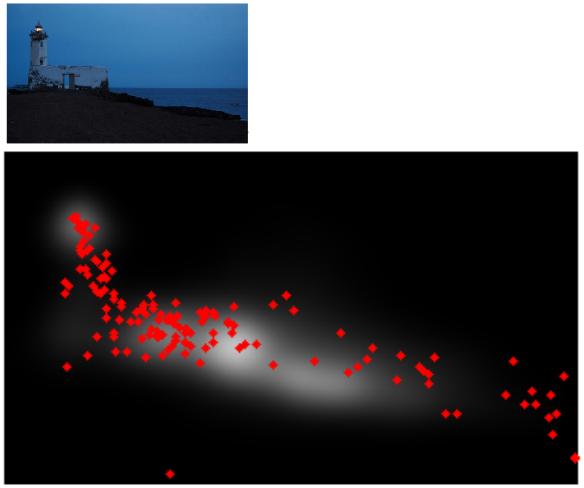
Area under ROC curve (AUC) as a performance measure

Saliency Map

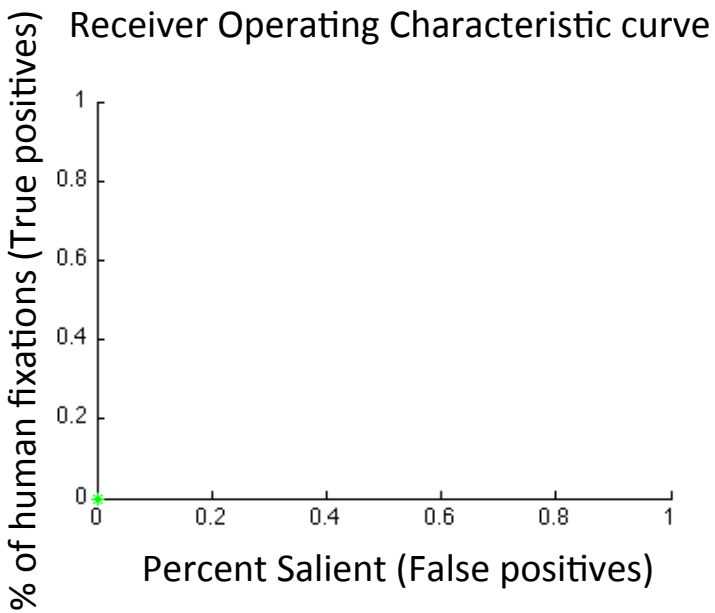


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

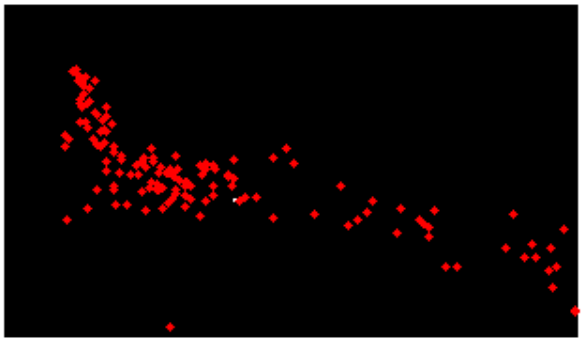


Human Fixations

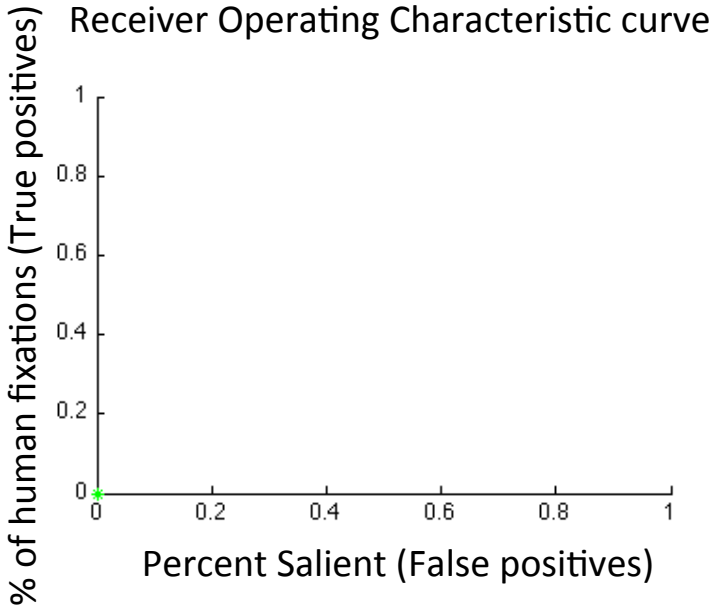


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

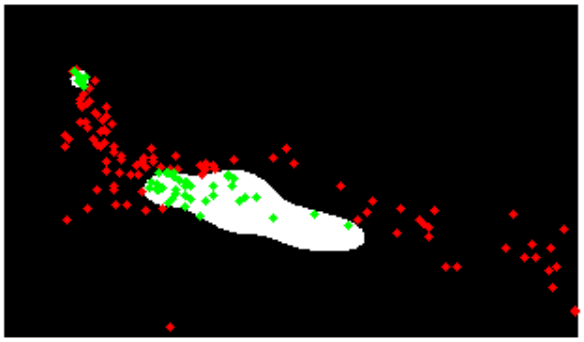


Thresholded Saliency Map

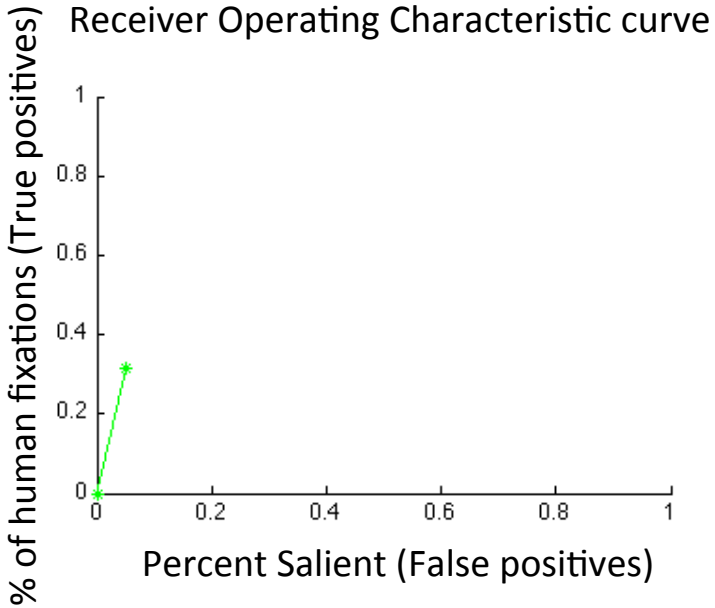


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

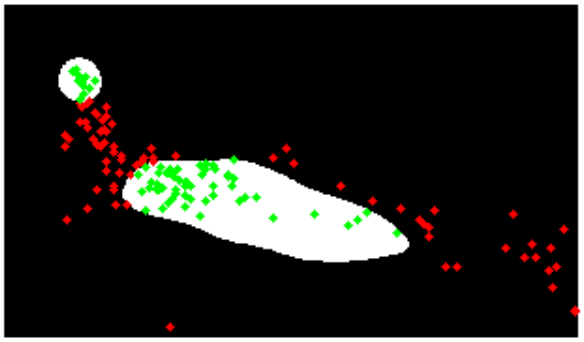


Thresholded Saliency Map

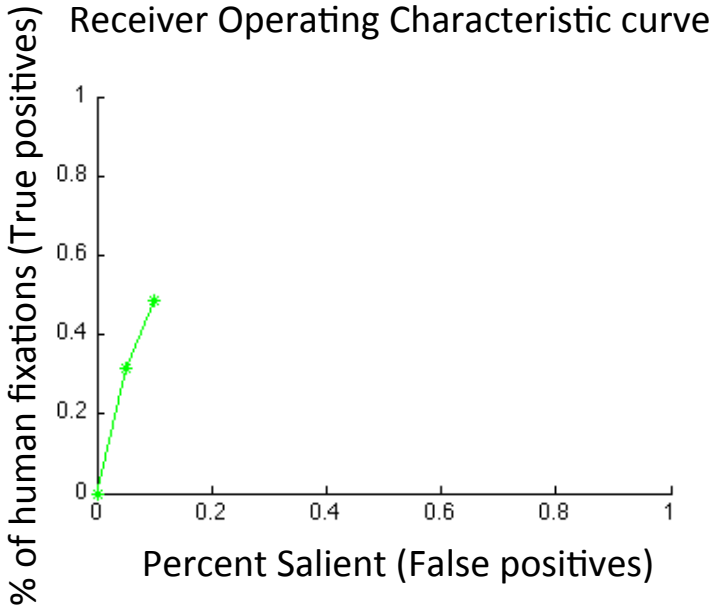


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

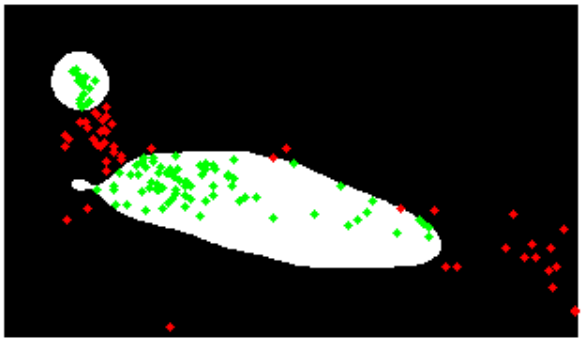


Thresholded Saliency Map

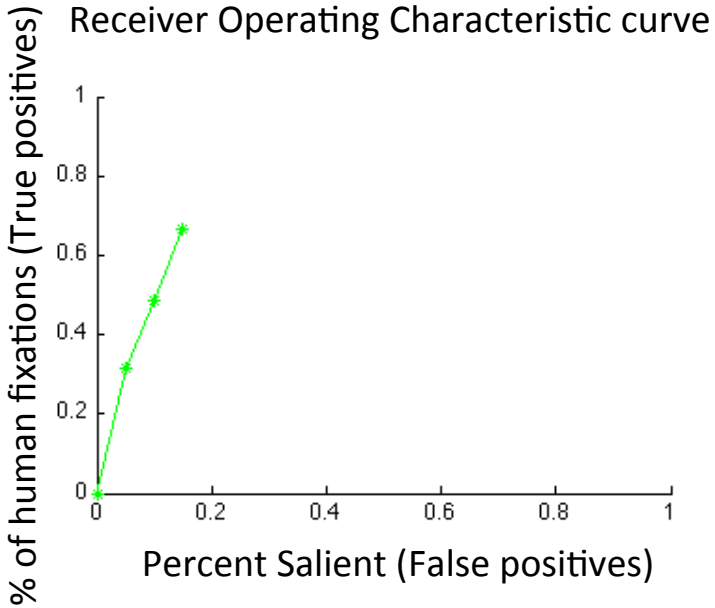


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure



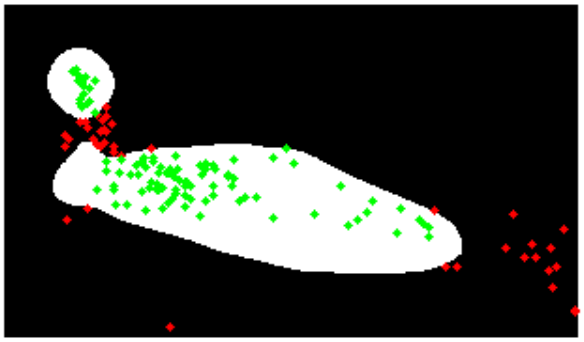
Thresholded Saliency Map



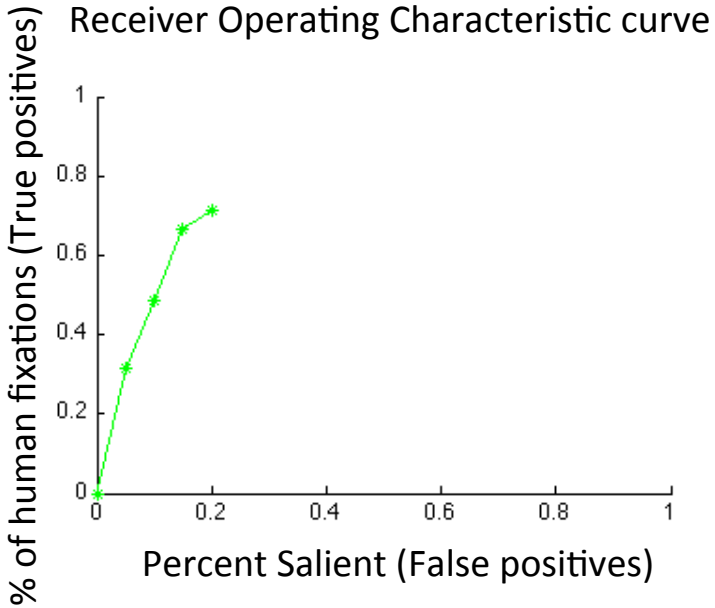
Percentage of fixations are calculated
that lie within the salient portion of the map

Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

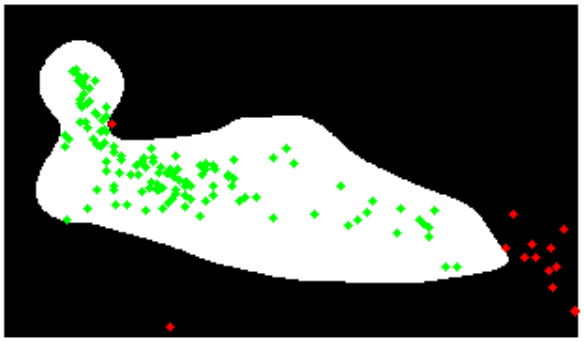


Thresholded Saliency Map

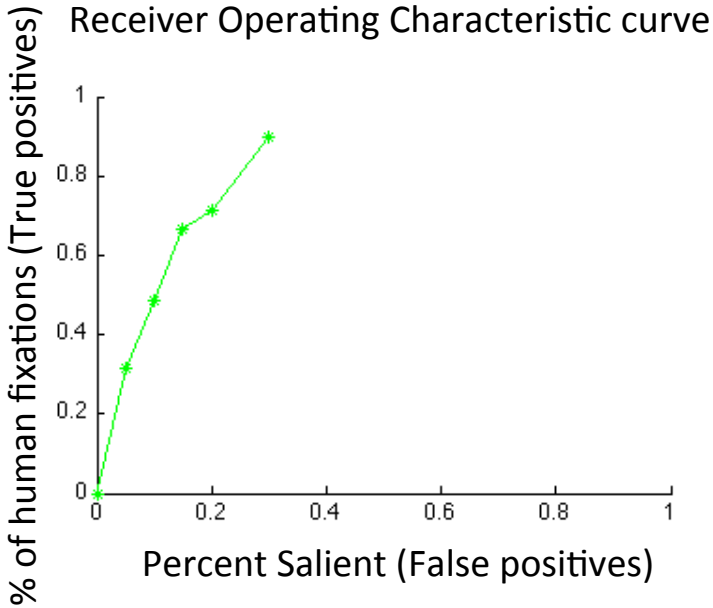


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

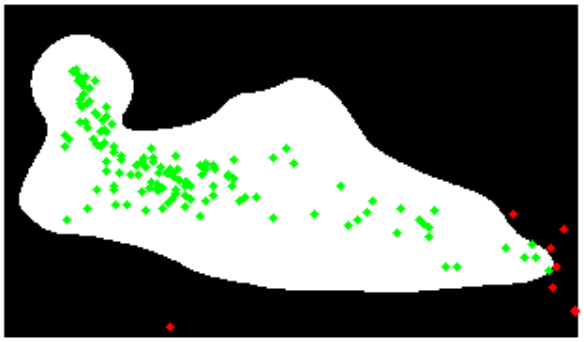


Thresholded Saliency Map

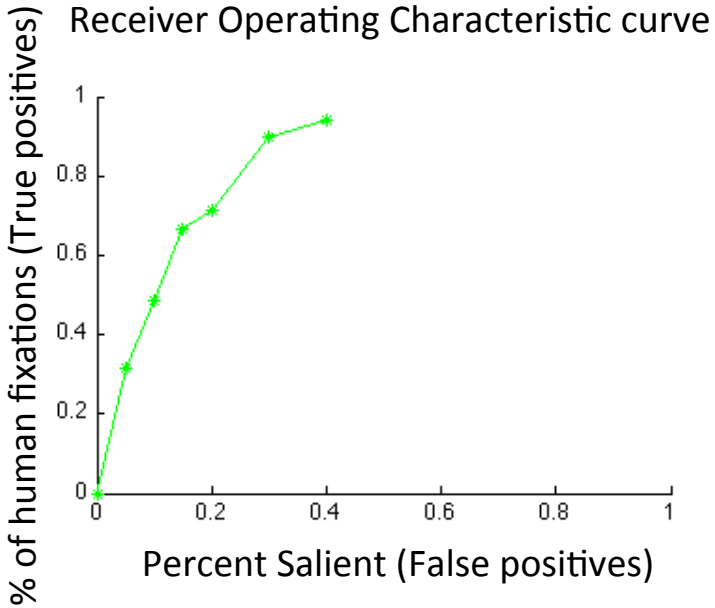


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

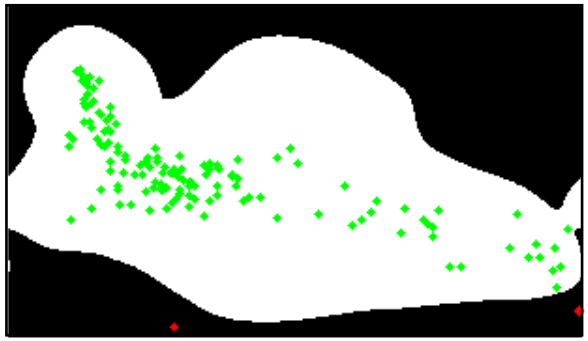


Thresholded Saliency Map

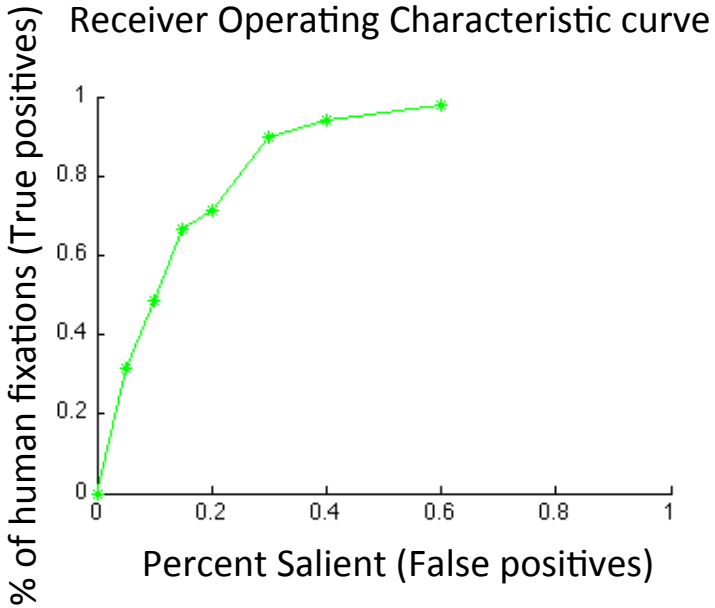


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

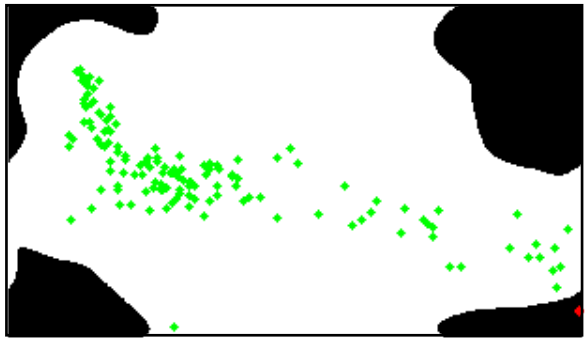


Thresholded Saliency Map

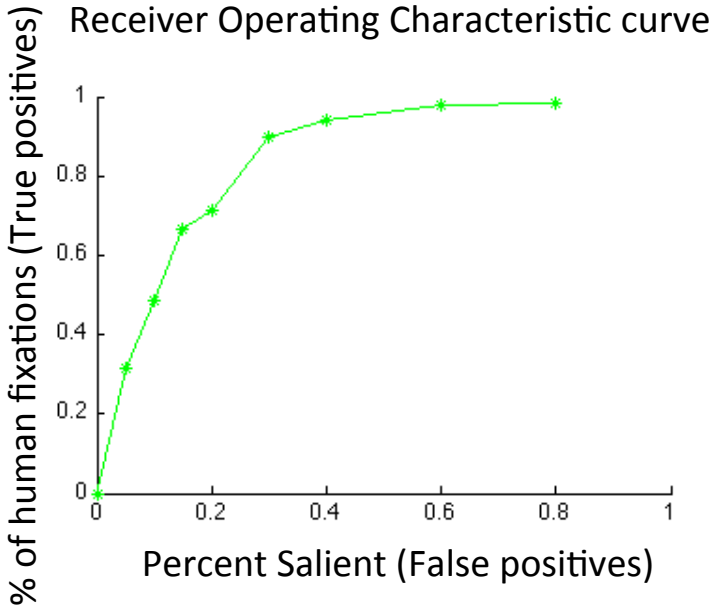


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

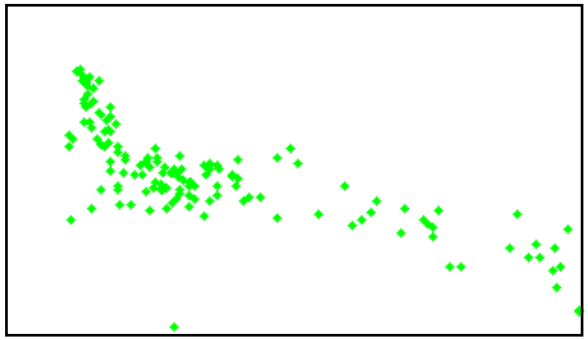


Thresholded Saliency Map

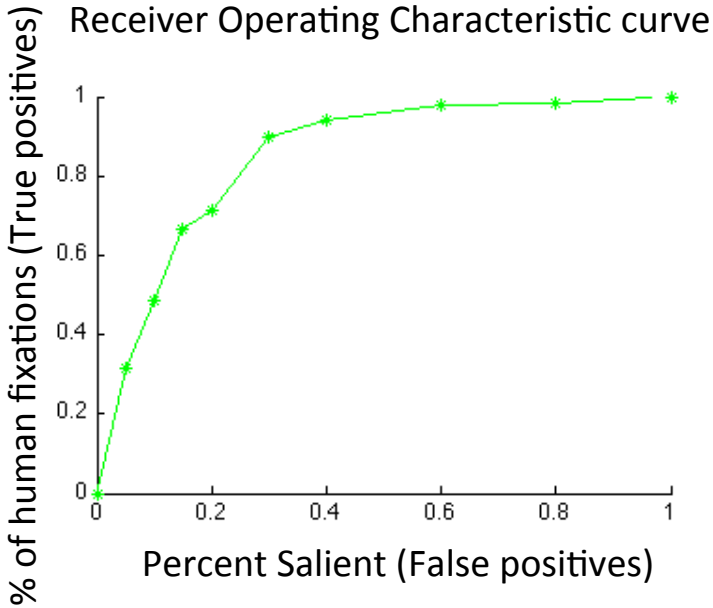


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure



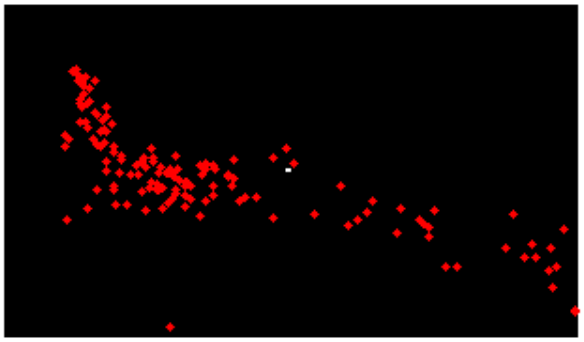
Thresholded Saliency Map



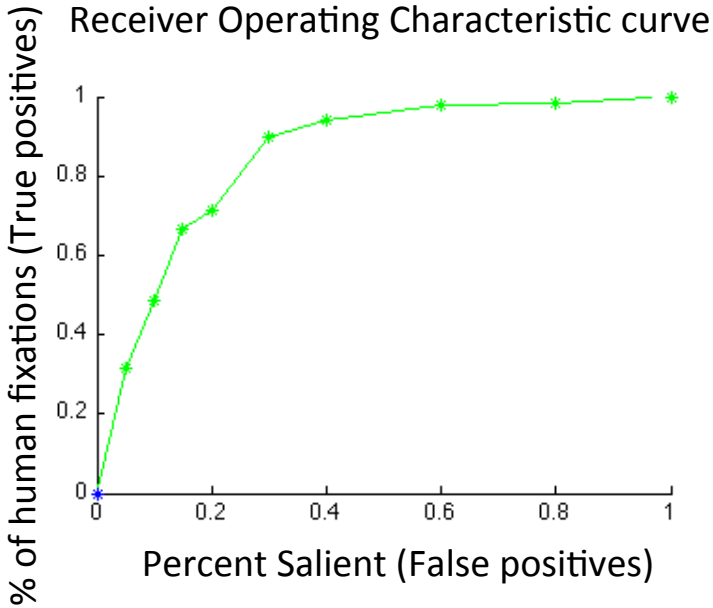
ROC curve always starts at 0 ends at 1

Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

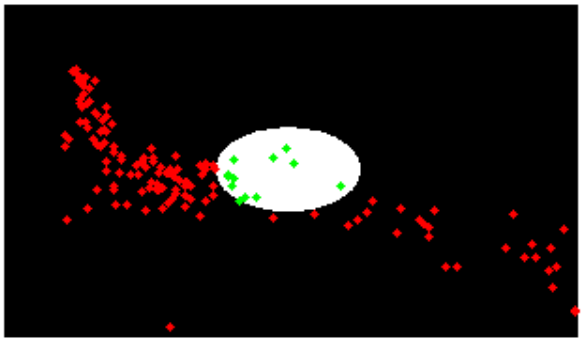


Thresholded Center Map

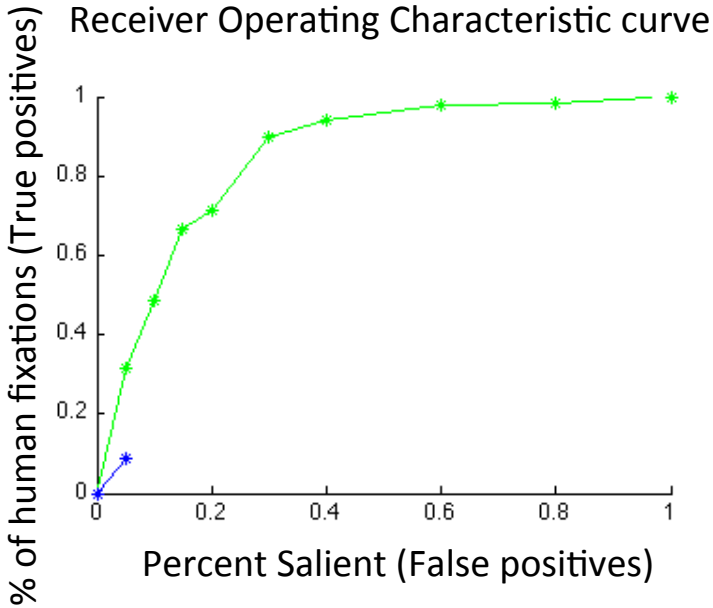


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

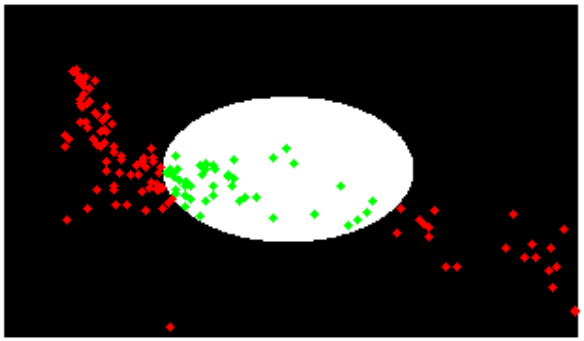


Thresholded Center Map

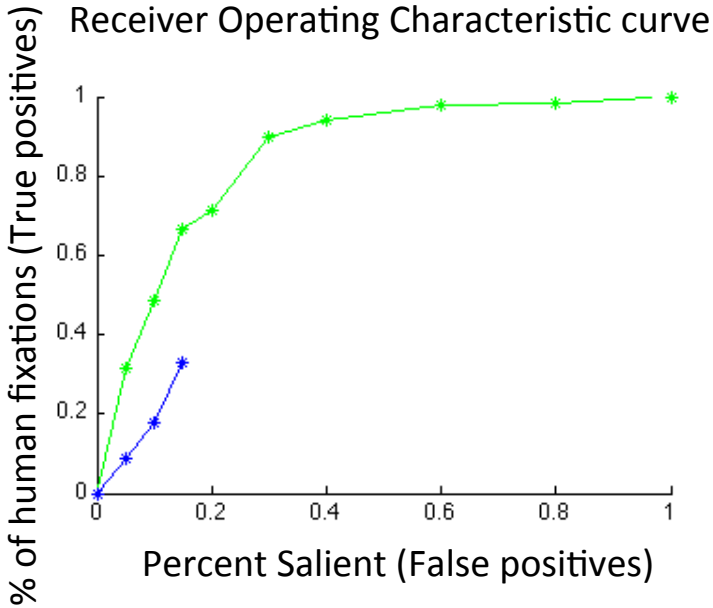


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

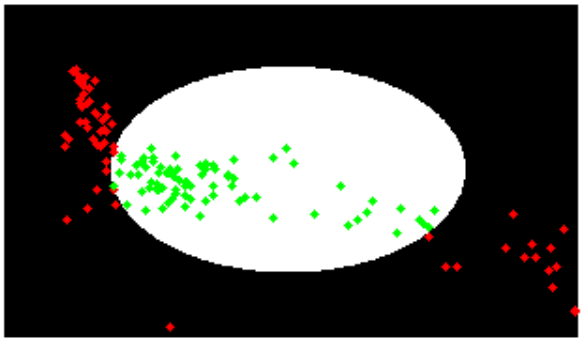


Thresholded Center Map

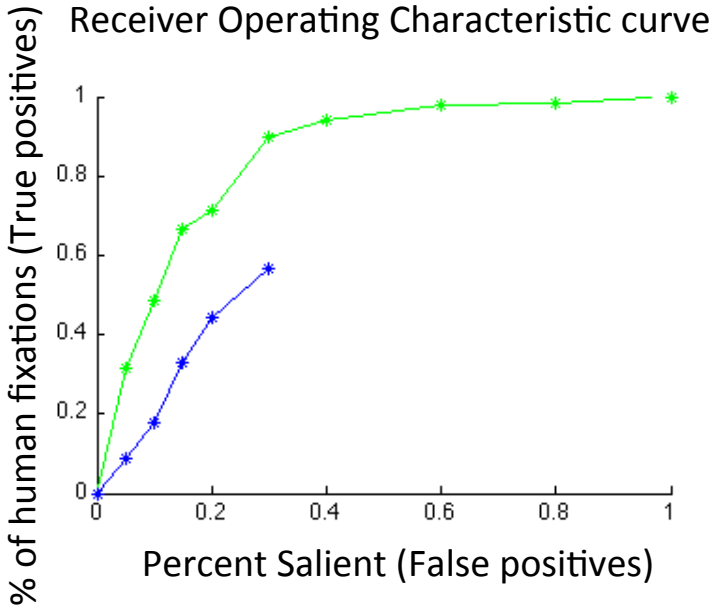


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

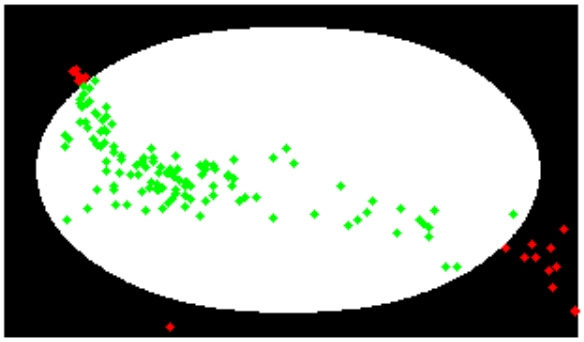


Thresholded Center Map

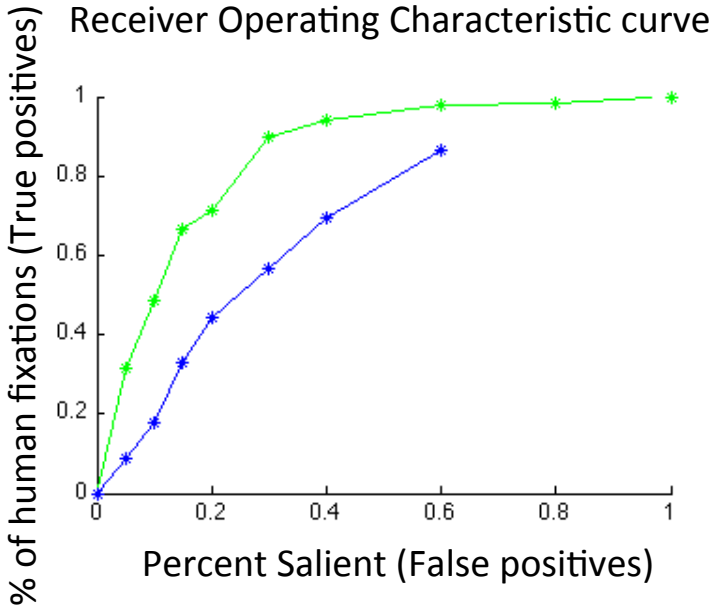


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

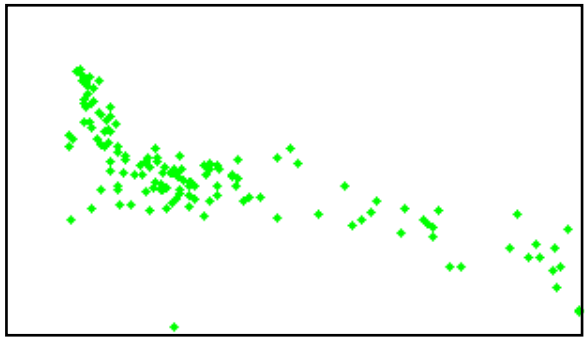


Thresholded Center Map

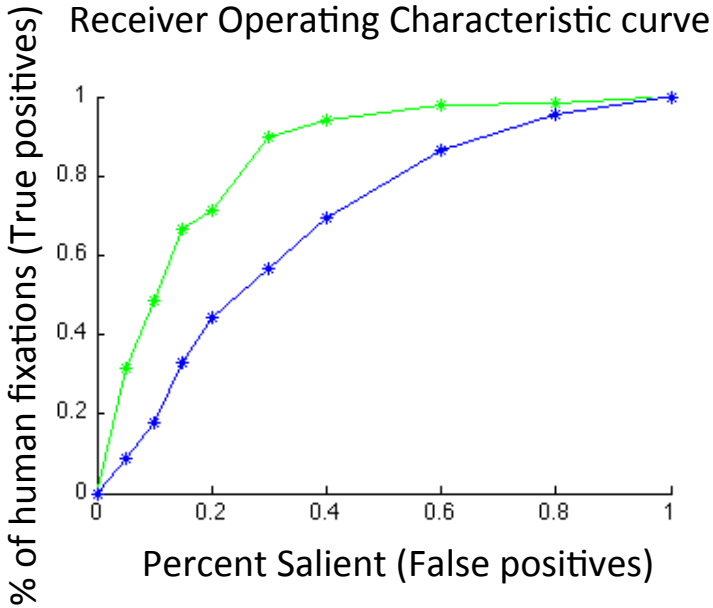


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

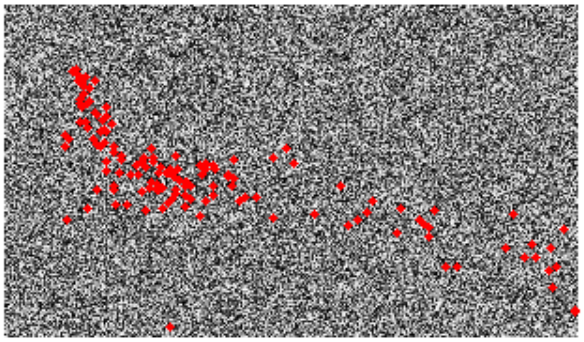


Thresholded Center Map

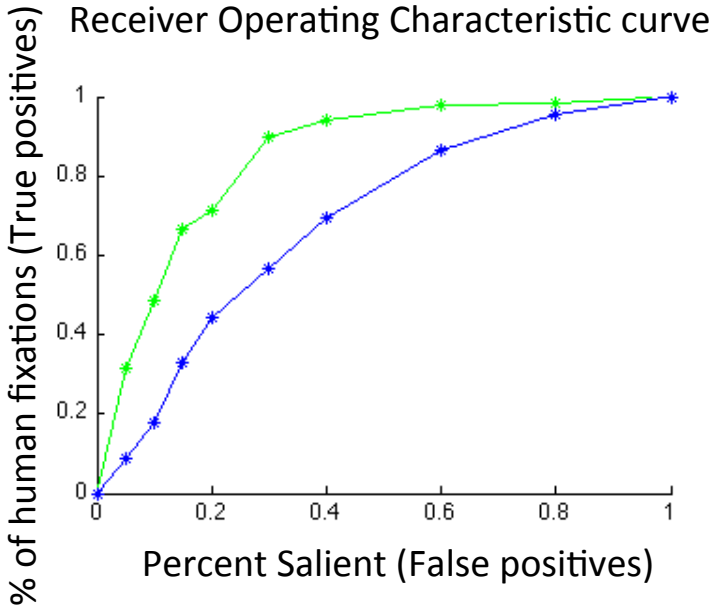


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

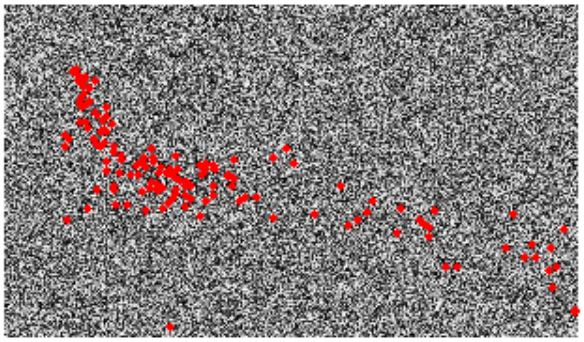


Random Noise map

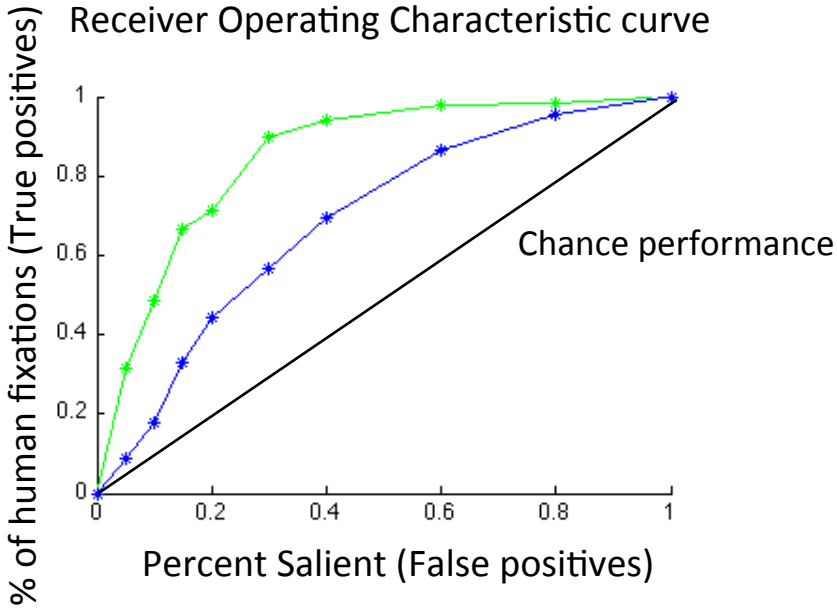


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

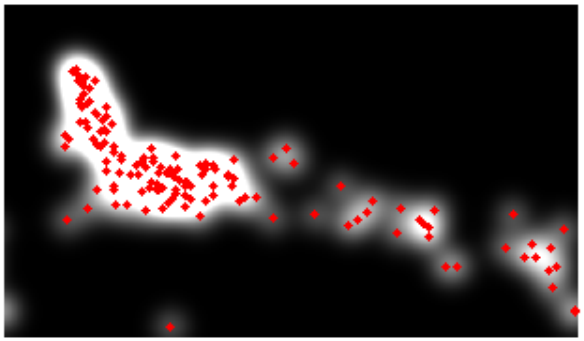


Random Noise map

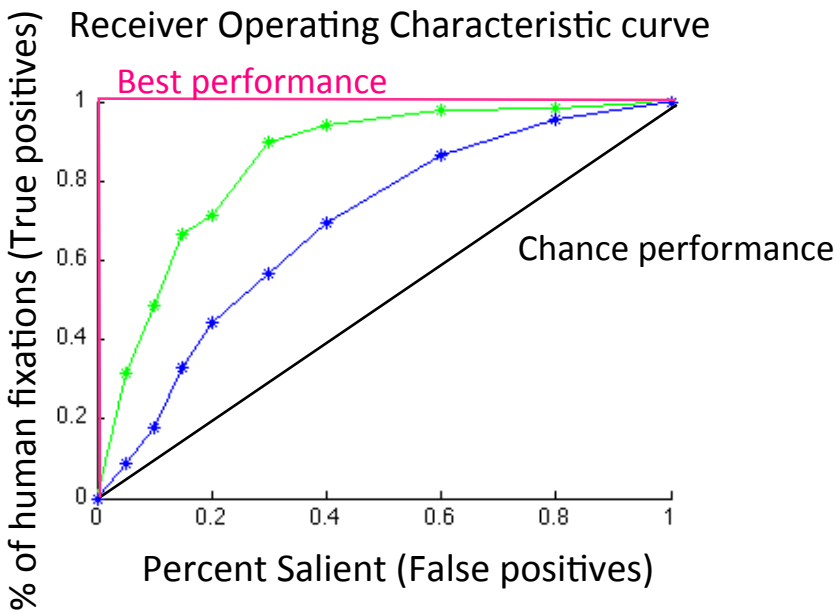


Adapted from T. Judd

Area under ROC curve (AUC) as a performance measure

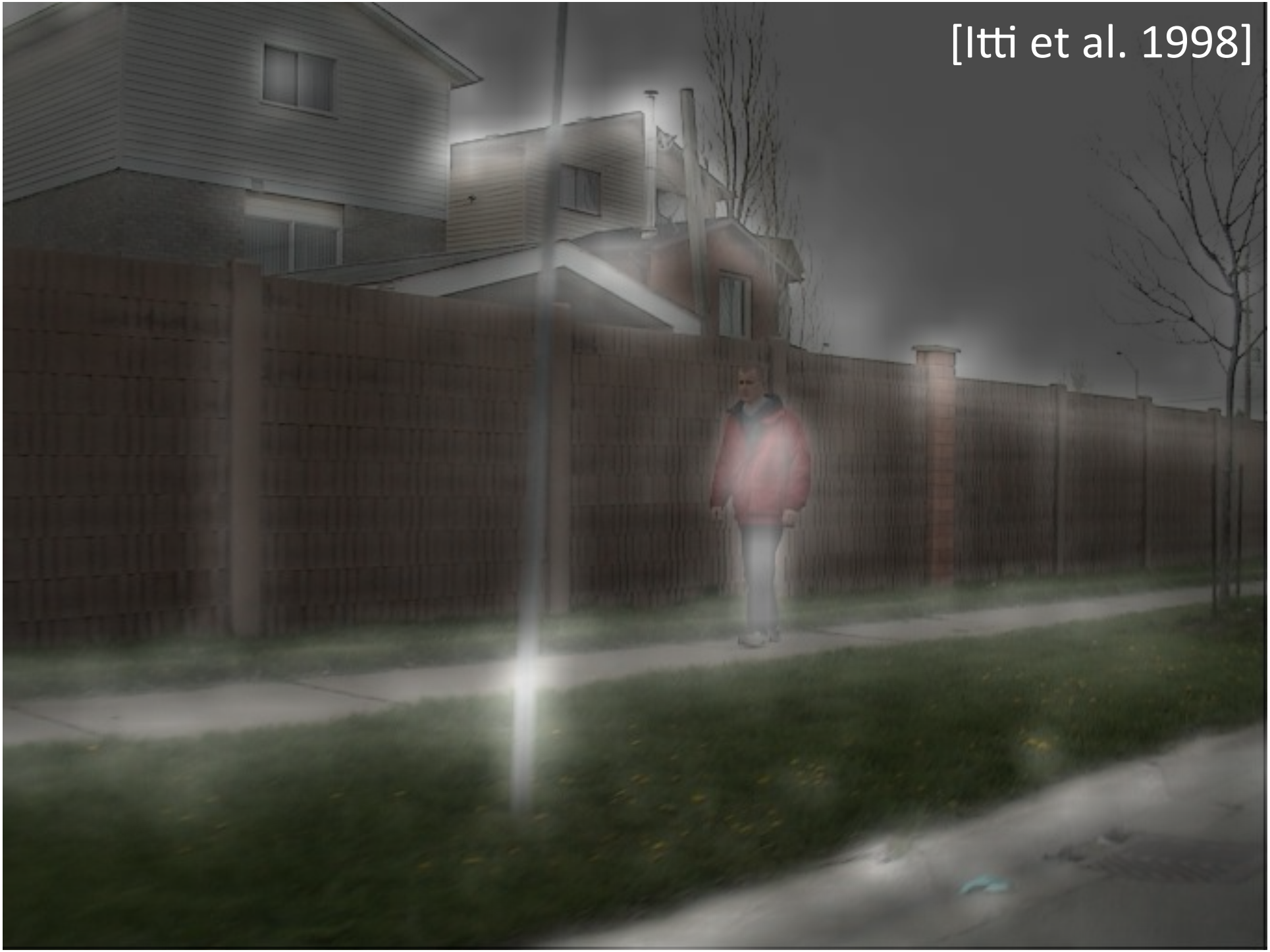


Perfect saliency map



Adapted from T. Judd


Density map from human fixation data


[Itti et al. 1998]


[Harel et al. 2007]


[Hou and Zhang 2009]

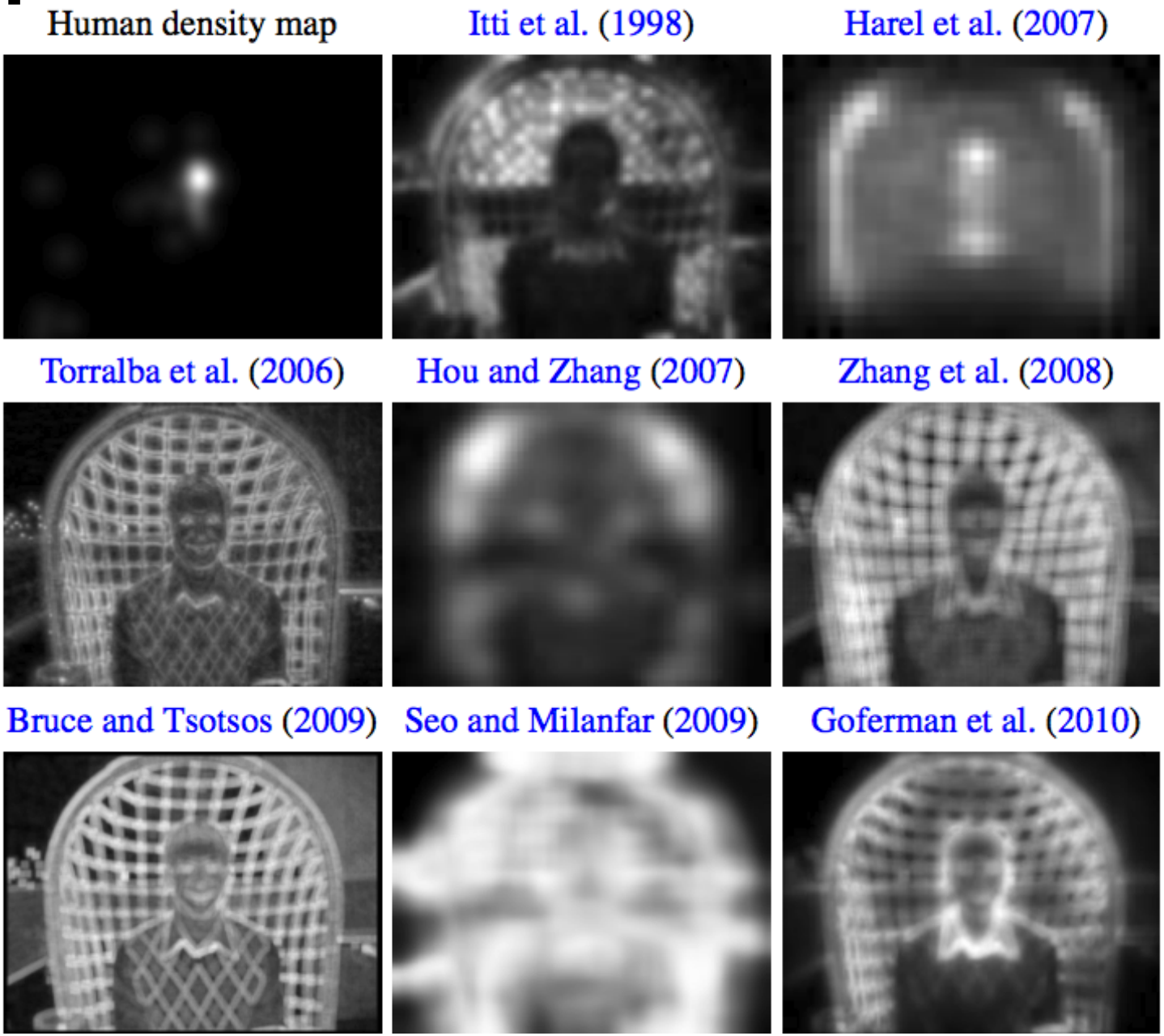

[Goferman et al. 2010]


Our Model 1


Our Model 2

Comparison to the state-of-the-art models

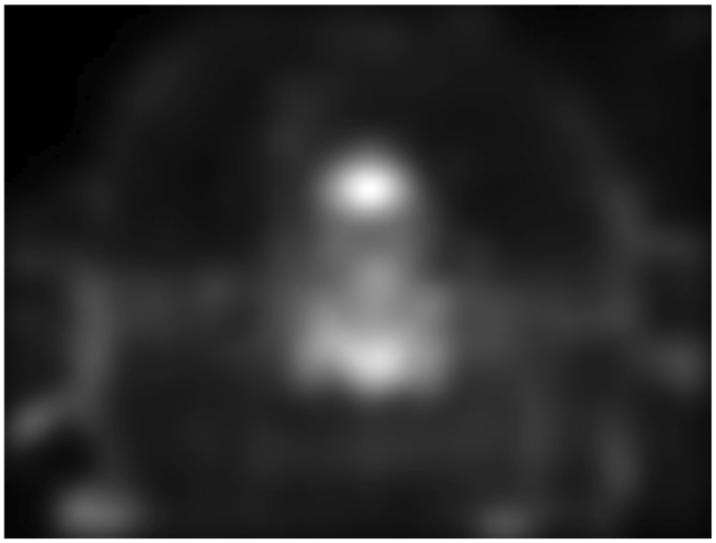
Comparison to the state-of-the-art models



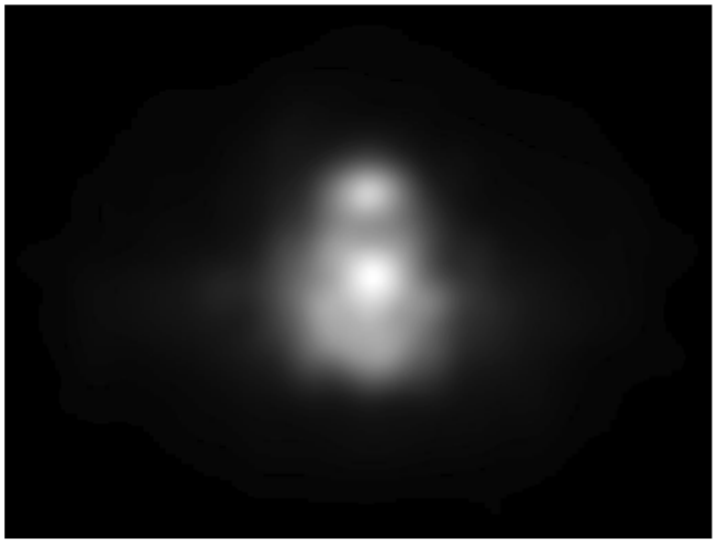
Comparison to the state-of-the-art models

Our approach with

covariances only

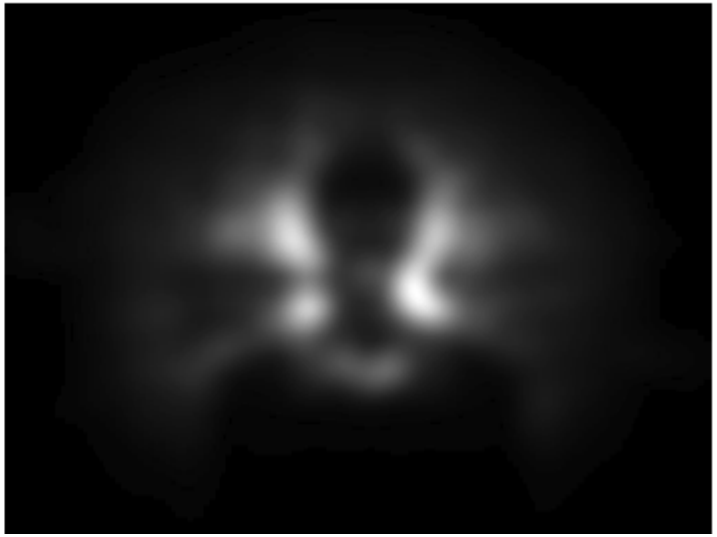


covariances + center



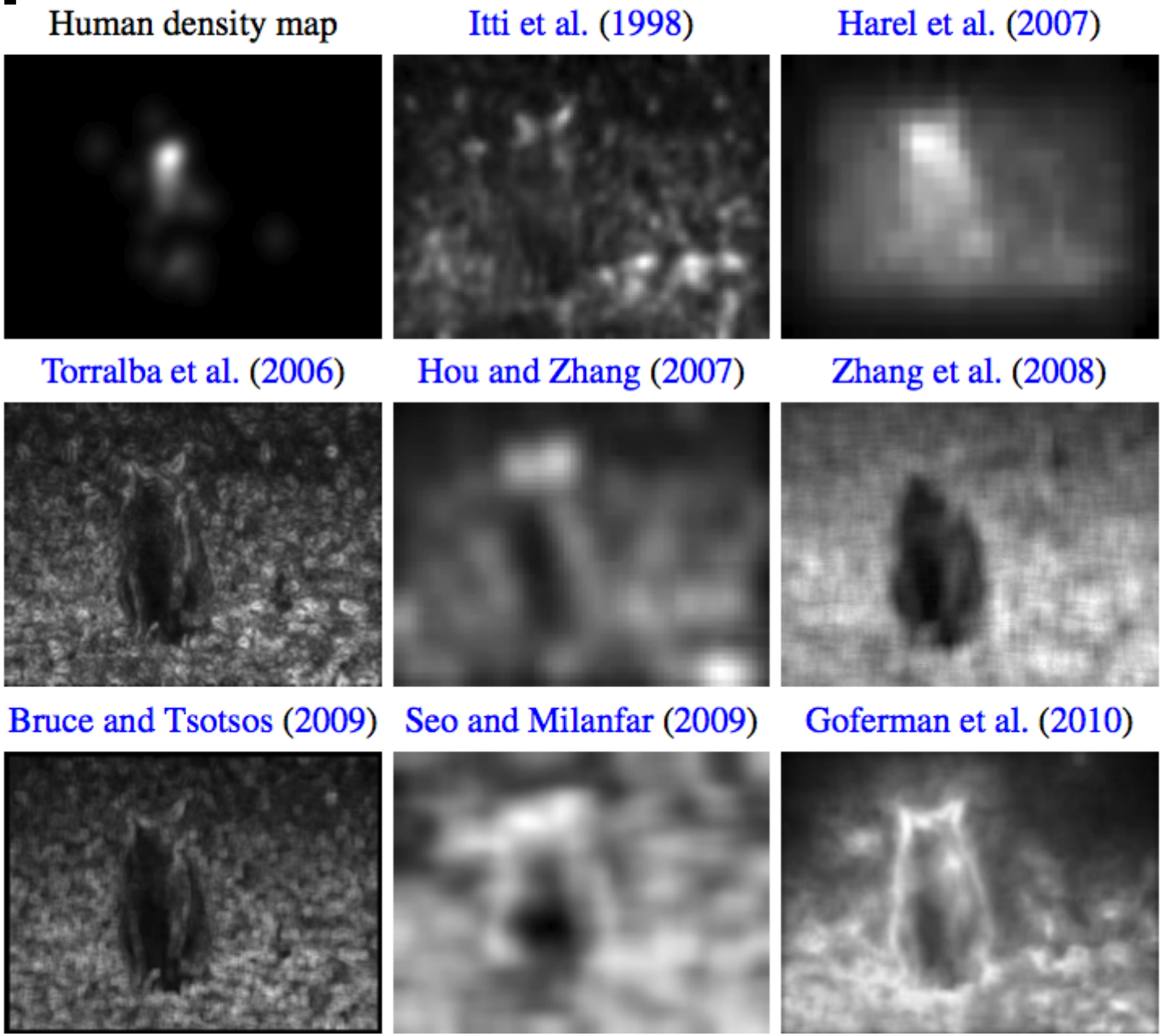
covariances + means

covariances + means + center



Comparison to the state-of-the-art models

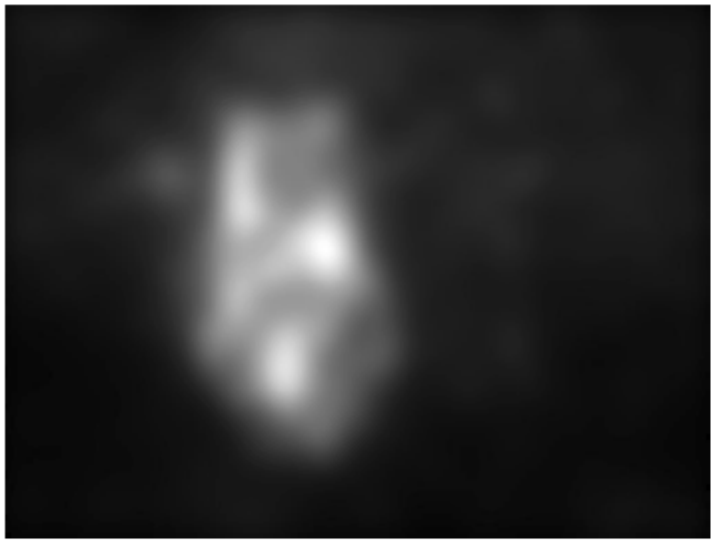
Comparison to the state-of-the-art models



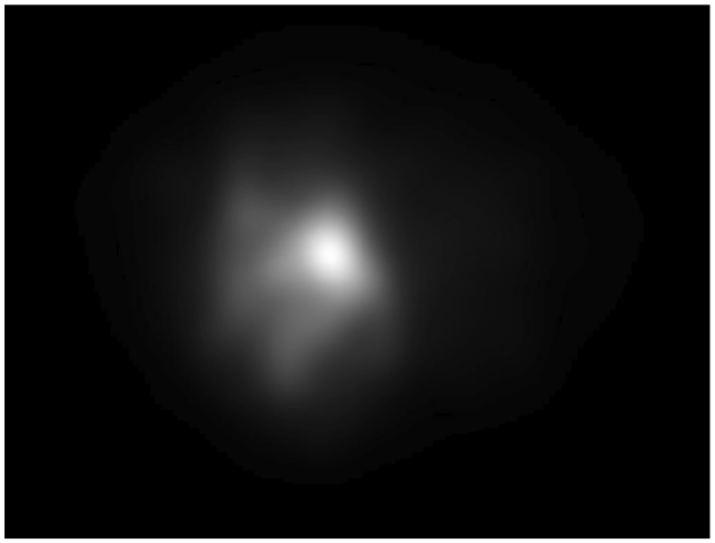
Comparison to the state-of-the-art models

Our approach with

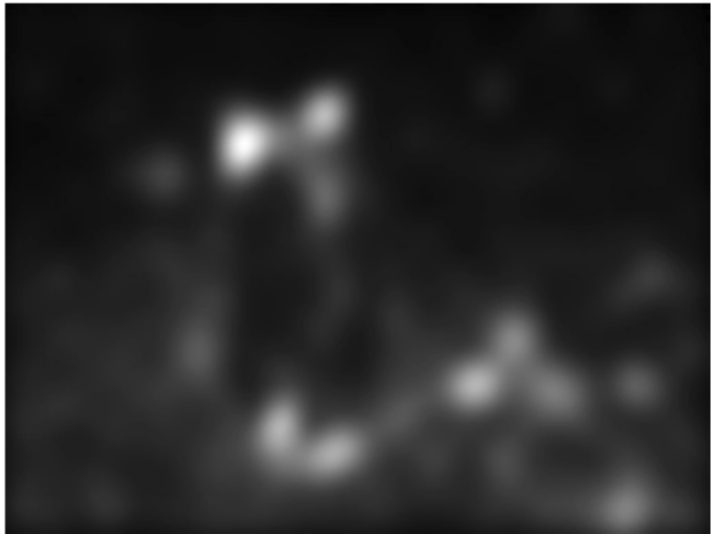
covariances only



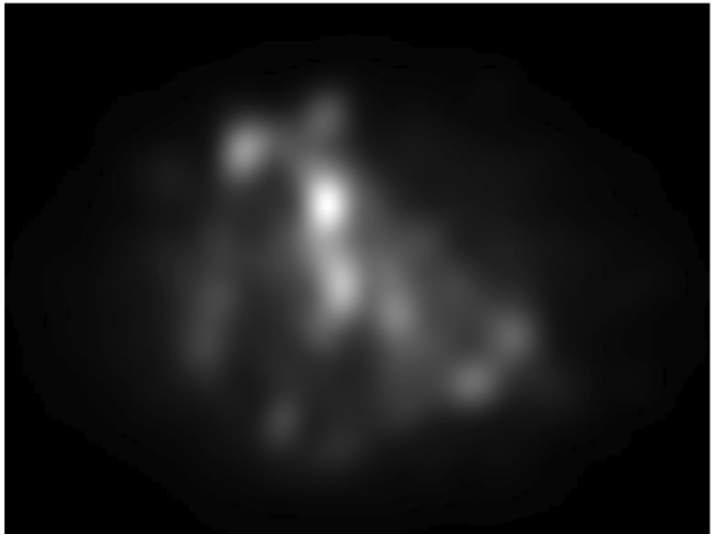
covariances + center



covariances + means

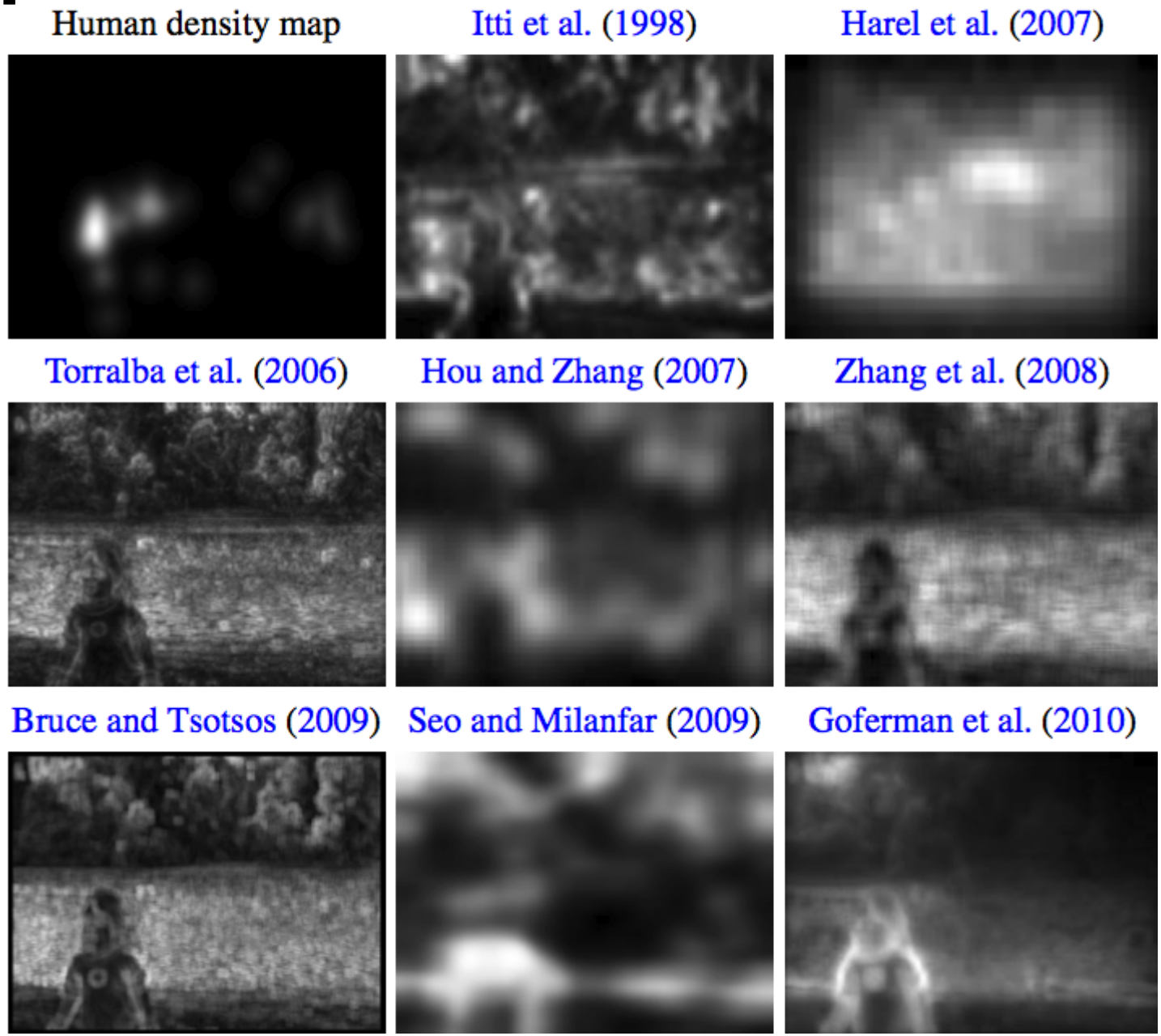


covariances + means + center



Comparison to the state-of-the-art models

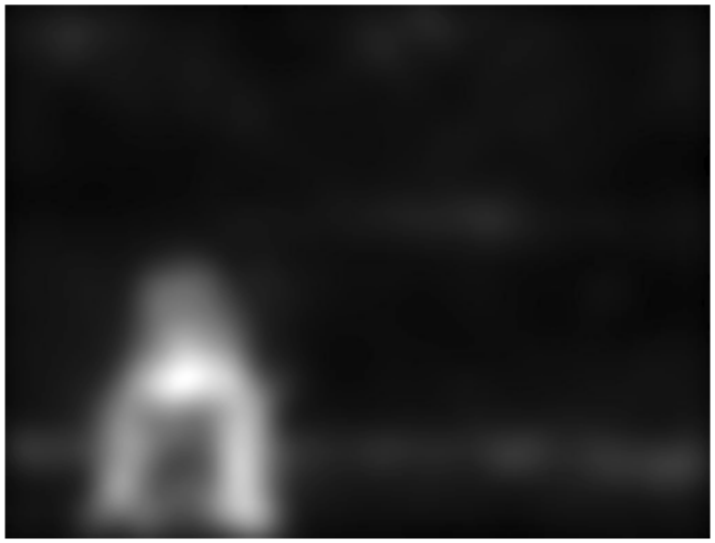
Comparison to the state-of-the-art models



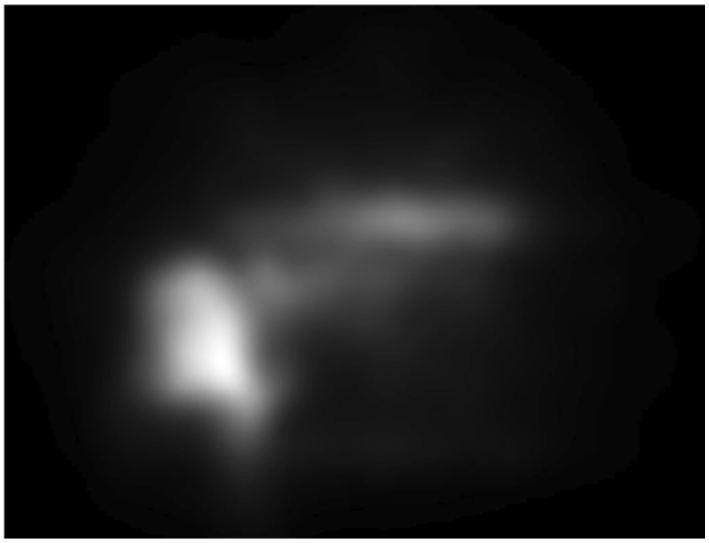
Comparison to the state-of-the-art models

Our approach with

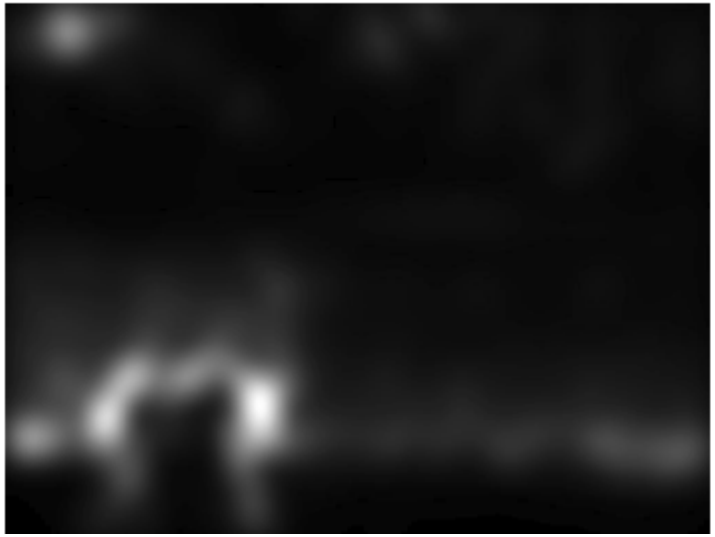
covariances only



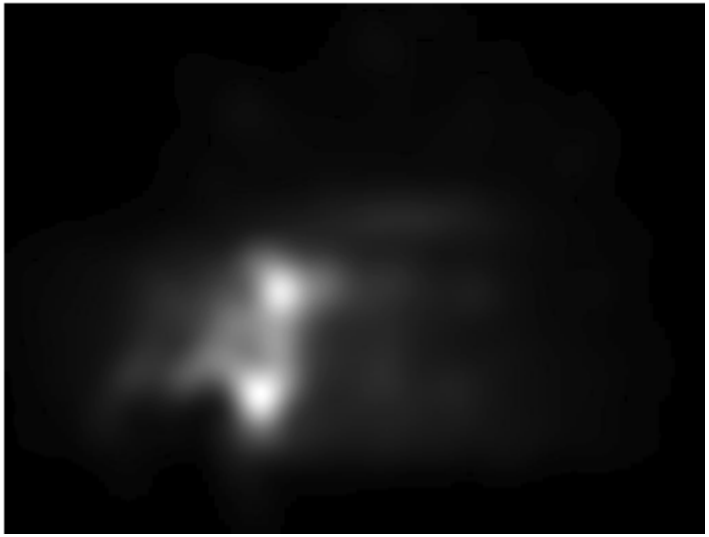
covariances + center



covariances + means



covariances + means + center



Quantitative Comparisons

	AUC		NSS		EMD		Similarity	
	Without CB	With CB	Without CB	With CB	Without CB	With CB	Without CB	With CB
Itti et al. (1998)	0.771	0.825	1.137	1.264	2.906	2.002	0.397	0.521
Harel et al. (2007)	0.829	0.835	1.533	1.533	2.014	1.886	0.519	0.556
Torralba et al. (2006)	0.710	0.832	0.805	1.185	3.467	1.868	0.330	0.528
Hou & Zhang (2007)	0.736	0.835	0.964	1.271	3.791	1.959	0.360	0.550
Zhang et al. (2008)	0.718	0.832	0.884	1.194	3.954	1.968	0.347	0.541
Bruce & Tsotsos (2009)	0.728	0.835	0.896	1.165	3.127	1.809	0.351	0.535
Seo & Milanfar (2009)	0.766	0.845	1.100	1.320	3.222	1.759	0.415	0.579
Goferman et al. (2010)	0.784	0.841	1.272	1.370	3.520	1.819	0.431	0.574
Our approach with								
Covariances only	0.767	0.834	1.184	1.342	3.142	1.931	0.408	0.546
Covariances + means	0.765	0.834	1.198	1.396	3.398	1.896	0.402	0.548
Covariances + center	0.840	0.840	1.753	1.753	1.901	1.901	0.561	0.561
Covariances + means + center	**0.851**	**0.851**	**1.891**	**1.898**	**1.728**	**1.728**	**0.581**	**0.581**
Center	–	0.803	–	0.969	–	2.401	–	0.478
Chance	0.505	0.803	−0.001	0.969	5.159	2.339	0.187	0.479

Table 1. Performance comparisons of the saliency models on the Toronto data set. Chance and Center are the baselines, which respectively stand for the random and the centered Gaussian models. CB denotes center bias. The best performing model is shown in bold type.

Quantitative Comparisons

	AUC		NSS		Similarity	
	Without CB	With CB	Without CB	With CB	Without CB	With CB
Itti et al. (1998)	0.741	0.827	0.921	1.170	0.273	0.402
Harel et al. (2007)	0.791	0.829	1.150	1.182	0.319	0.415
Torralba et al. (2006)	0.700	0.832	0.771	1.156	0.244	0.412
Hou & Zhang (2007)	0.713	0.833	0.855	1.200	0.264	0.421
Zhang et al. (2008)	0.703	0.834	0.829	1.177	0.261	0.418
Bruce & Tsotsos (2009)	0.709	0.835	0.813	1.148	0.254	0.415
Seo & Milanfar (2009)	0.712	0.836	0.826	1.171	0.263	0.424
Goferman et al. (2010)	0.758	0.840	1.053	1.241	0.297	0.431
Our approach with						
Covariances only	0.715	0.826	0.862	1.169	0.261	0.410
Covariances + means	0.740	0.832	0.940	1.240	0.287	0.417
Covariances + center	0.833	0.833	1.468	1.486	0.417	0.418
Covariances + means + center	**0.843**	**0.843**	**1.488**	**1.543**	**0.428**	**0.432**
Center	–	0.810	–	1.004	–	0.379
Chance	0.500	0.810	−0.000	1.004	0.131	0.383

Table 2. Performance comparisons of the saliency models on the MIT1003 data set. The best performing model is shown in bold type.

Quantitative Comparisons

	AUC		EMD		Similarity	
	Without CB	With CB	Without CB	With CB	Without CB	With CB
Itti et al. (1998)	0.750	0.806	4.560	3.394	0.405	0.493
Harel et al. (2007)	0.801	0.813	3.574	3.315	0.472	0.501
Torralba et al. (2006)	0.684	0.806	4.715	**3.036**	0.343	0.488
Hou & Zhang (2007)	0.682	0.804	5.368	3.200	0.319	0.487
Zhang et al. (2008)	0.672	0.799	5.088	3.296	0.340	0.473
Bruce & Tsotsos (2009)	0.751	**0.820**	4.236	3.085	0.390	0.507
Goferman et al. (2010)	0.742	0.815	4.900	3.219	0.390	**0.509**
Our approach with						
Covariances + center	0.800	0.800	3.422	3.422	0.487	0.487
Covariances + means + center	**0.806**	0.811	**3.109**	3.109	**0.502**	0.503
Center	–	0.783	–	3.719	–	0.451
Chance	0.503	0.783	6.352	3.506	0.327	0.482
Judd et al. (2009)	0.811	0.813	3.130	3.130	0.506	0.511

Table 3. Performance comparisons of the saliency models on the MIT300 data set. The best performing model is shown in bold type.

Understanding attention enables applications in computer graphics & vision, design

- image cropping / thumbnailing
- image and video compression
- non photorealistic rendering
- scene understanding
- advertising and package design
- web usability
- localization / recognition
- object detection
- navigational assistance
- robot active vision
- surveillance systems
- assistive technology for blind or low-vision people

Detecting Salient Objects

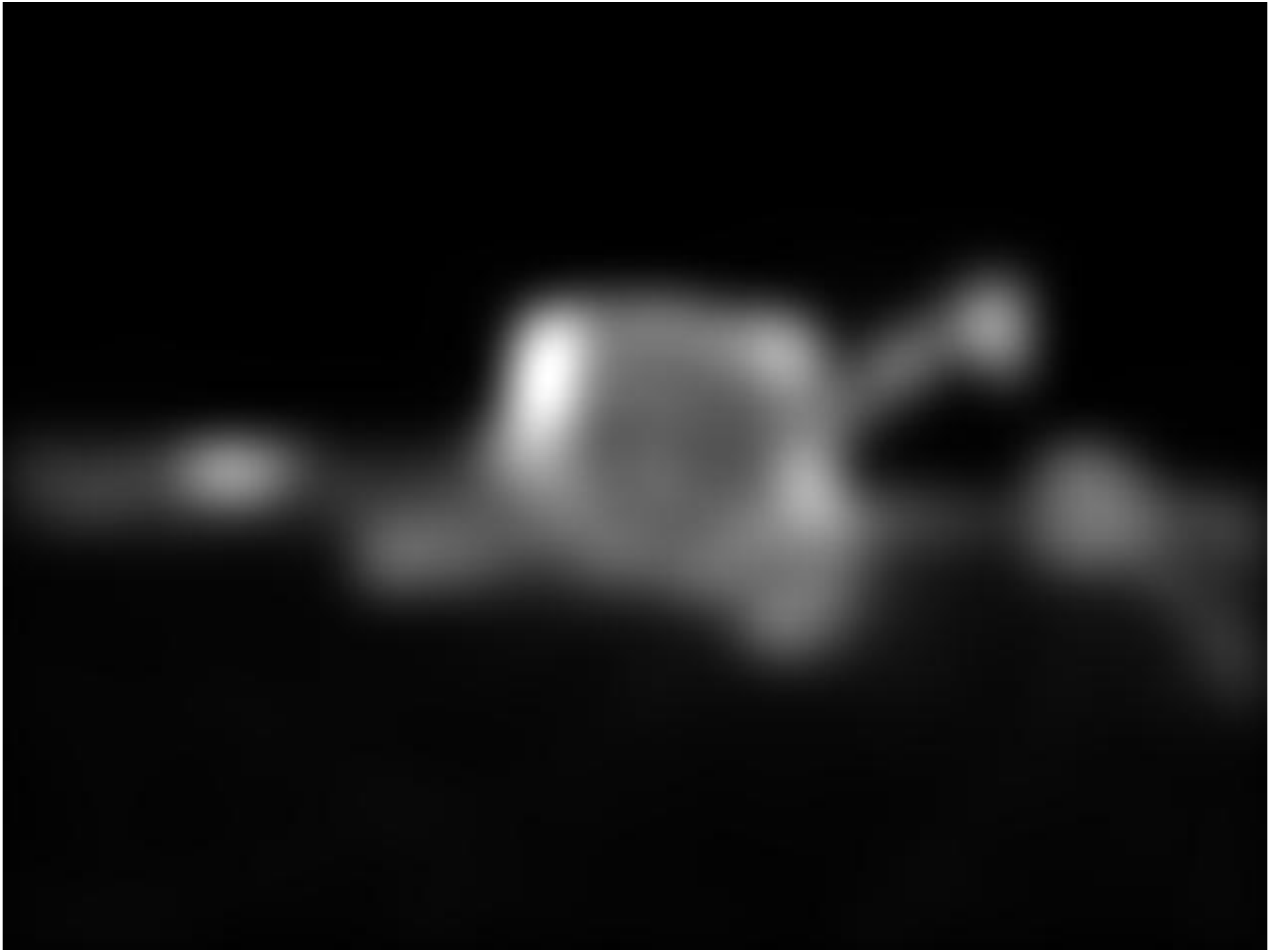
- The task is to detect foreground objects that attracts more attention in a given image

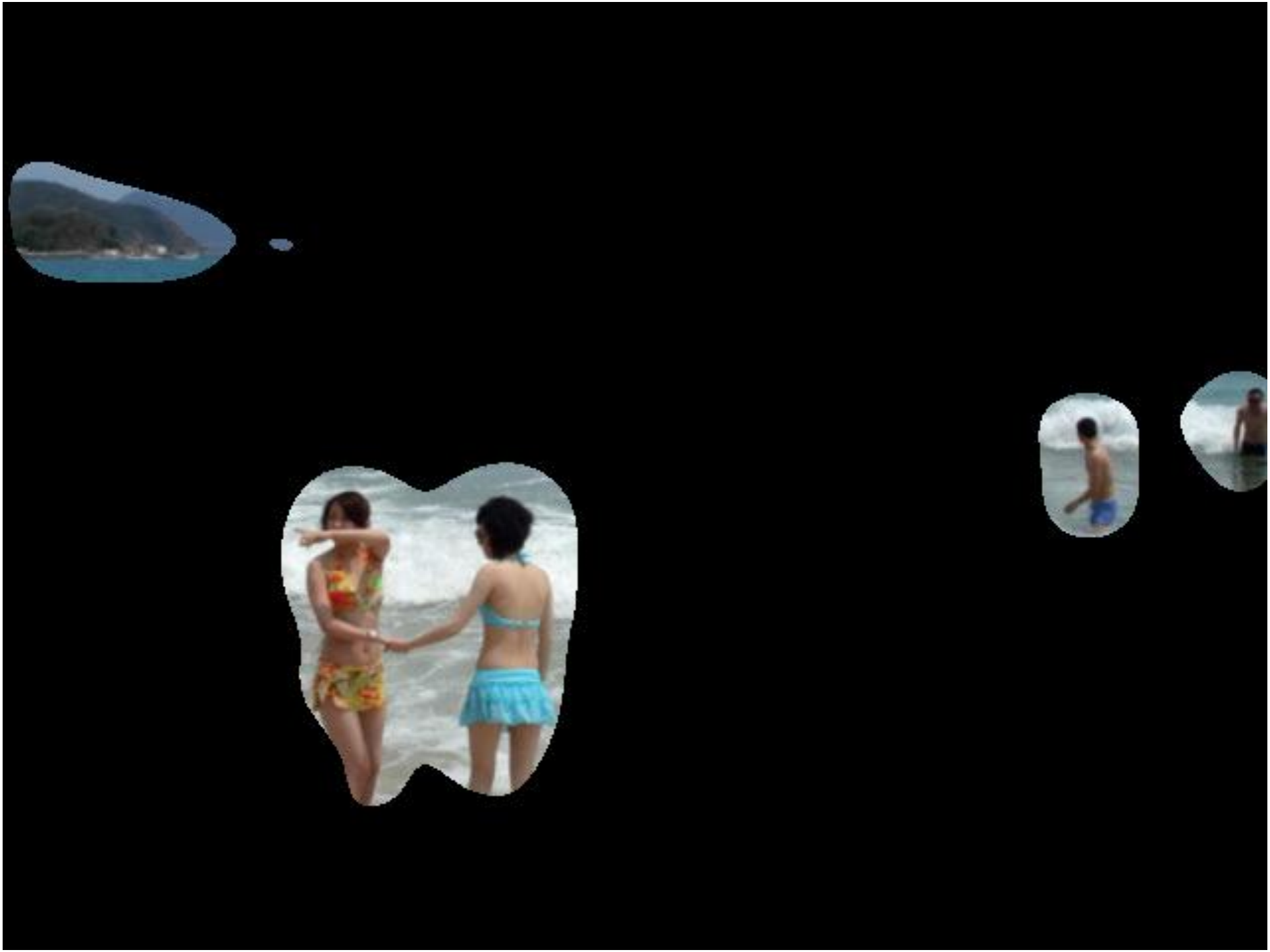
Detecting Salient Objects

- The task is to detect foreground objects that attracts more attention in a given image

Detecting Salient Objects

- ImgSal data set [Li, Levine, An, Xu & He, 2012]
- 235 natural images of size 480 x 640 pixels
- 19 subjects are asked to label the most salient objects in the images presented to them
- 6 different categories
  - large salient regions,
  - intermediate salient regions,
  - small salient regions,
  - cluttered backgrounds,
  - repeating distractors,
  - large and small salient regions)





Summary

- Can machines predict where the humans look at a given image?
- State-of-the-art computational models of visual attention reach a certain point of success.
- Selection mechanisms provided by the computational saliency models could improve several machine perception applications.

Thanks for your attention!