Learning to Predict Where Humans Look

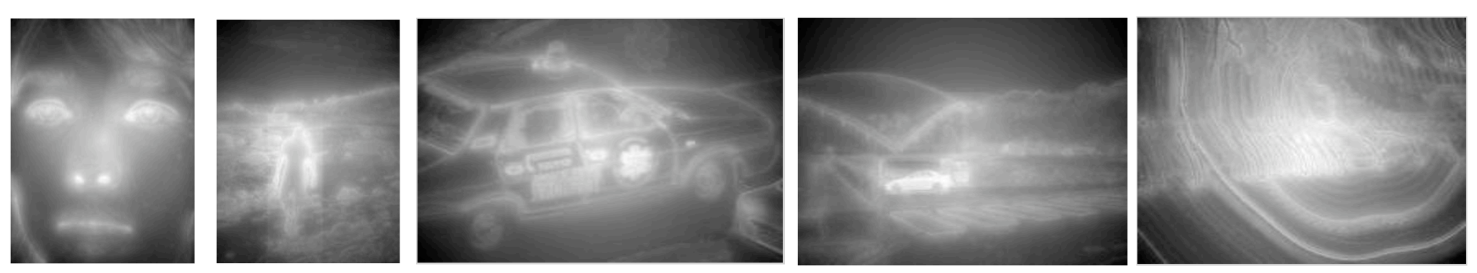
Tilke Judd, Krista Ehinger, Frédo Durand, Antonio Torralba
{tjudd,kehinger, fredo, torralba}@mit.edu
CSAIL MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY
Massachusetts Institute of Technology

Introduction

Teaser

Where do you look in these images?

This is where other people looked in eye tracking tests



This is where our model predicts you will look.

Applications

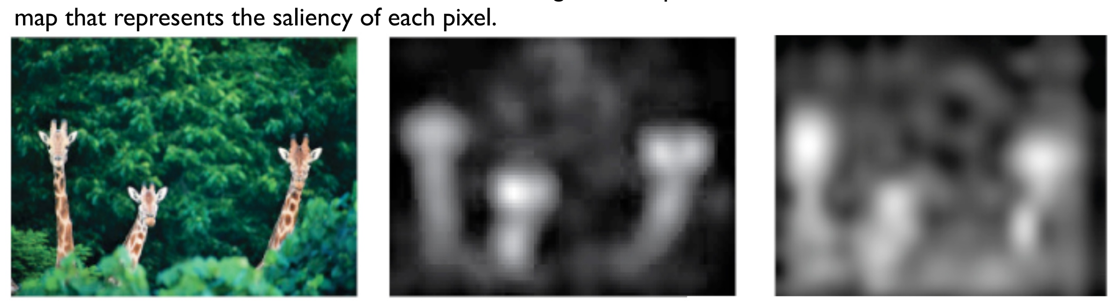
Understanding where people look has applications in graphics, advertising design and human computer interaction. In addition to enabling intelligent automatic cropping and thumbnailing and driving level of detail for non-photorealistic rendering, a good model of saliency can be used in seam carving and foveated image and video compression.

Level of detail for non-photorealistic rendering Doug DeCarlo and Anthony Santella [SIGGRAPH 2002]

(a) original (b) gaze-based (c) distracting crop
Automatic cropping Santella et al. [SIGCHI 2006]

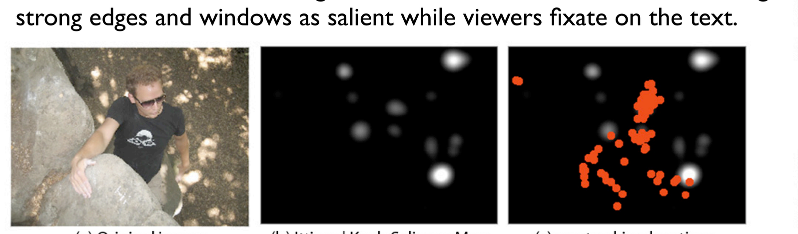
Previous models of saliency

Most models of saliency are based on a bottom-up computational model. Typically, multiple low-level visual features such as intensity, color, orientation, and texture are extracted from the image at multiple scales and combined in a linear or non-linear fashion into a master saliency map that represents the saliency of each pixel.



(a) Original image (b) Hou and Zhang (c) Itti and Koch

Problems with previous models
Current saliency models do not accurately predict human fixations. Below, the low-level model selects bright spots of light as salient while viewers look at the human. On the right, the low level model selects the building's strong edges and windows as salient while viewers fixate on the text.



(a) Original image (b) Itti and Koch Saliency Map (c) eye tracking locations

(a) Original image (b) Itti and Koch Saliency Map (c) eye tracking locations

Our Contributions

1) Create a large public database of eye tracking experiments that show were people actually look in images.
2) Create a supervised learning model of saliency that combines both bottom up saliency-based cues and top-down image semantic cues

Our Experimental Setup

Image database
We collected a large database of 1000 natural images from Flickr and LabelMe.

Eye tracking experiment
We ran a large eye tracking experiment with 15 users and 1000 images. This is the largest eye tracking database of natural images that we know about and has been made available to the public.

screen resolution 1280x1024

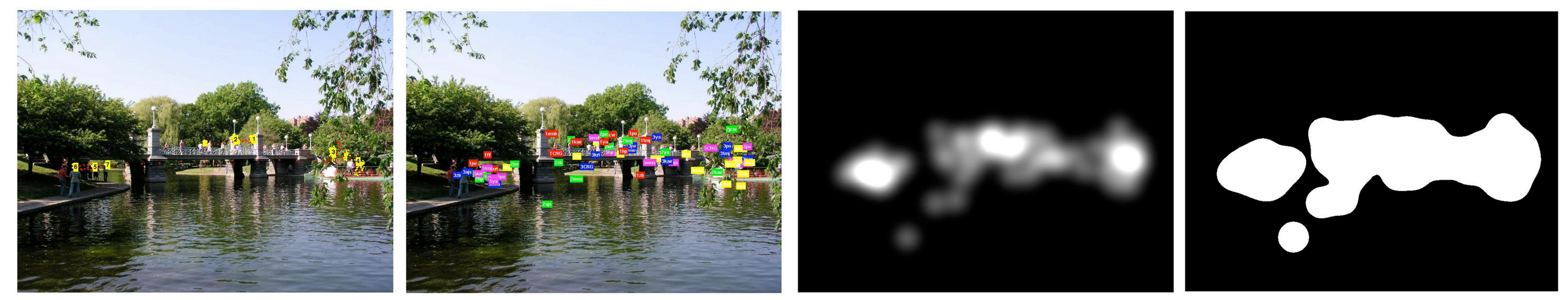
each image shown for 3 seconds

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

user rests head in chin rest

Our free and public eye tracking database

Eye tracking fixation information

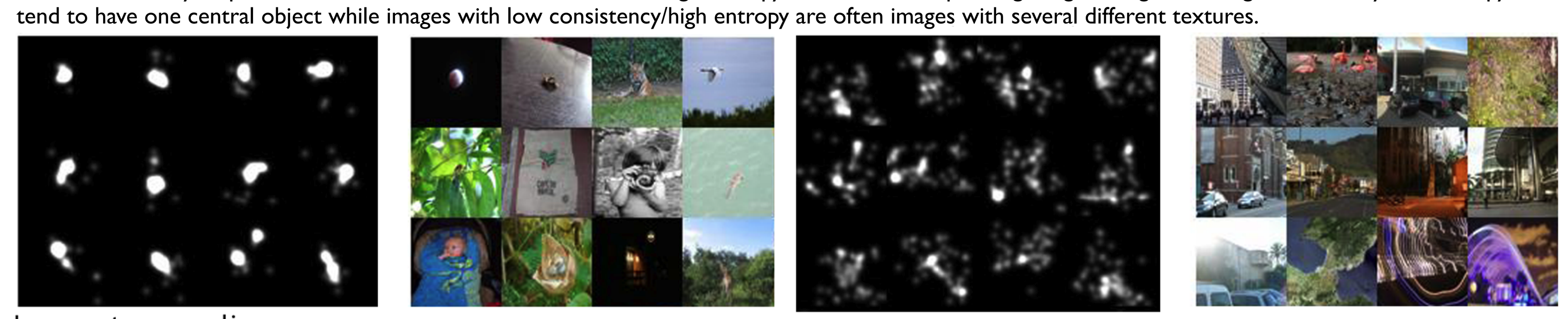


Fixation information
Colored squares indicate locations that 15 viewers fixated on when viewing this photograph, We stored data about the path and timing of user's fixations through the image.

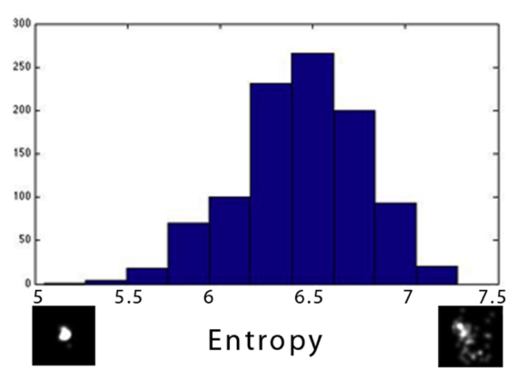
Human saliency map
We convolve a gaussian over the fixation locations from all 15 viewers to create a ground truth saliency map which shows the likelihood of a human to look at a certain location. This saliency map can be thresholded to show the most salient percent (here 20%) of the image.

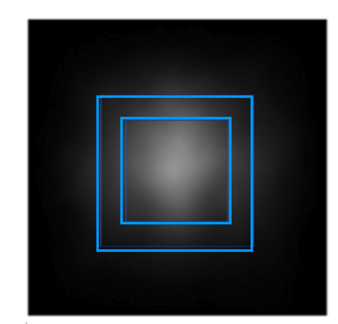
Analysis of database

Measuring the spread of fixations
These are saliency maps made from human fixations with low and high entropy and their corresponding images. Images with high consistency/low entropy tend to have one central object while images with low consistency/high entropy are often images with several different textures.



Low entropy saliency maps High entropy saliency maps





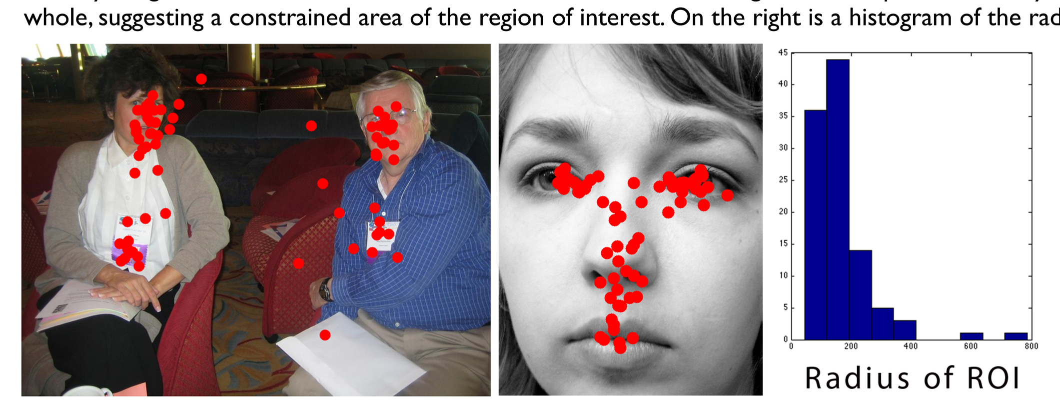
Avg of all saliency maps

Histogram of saliency map entropies (left)

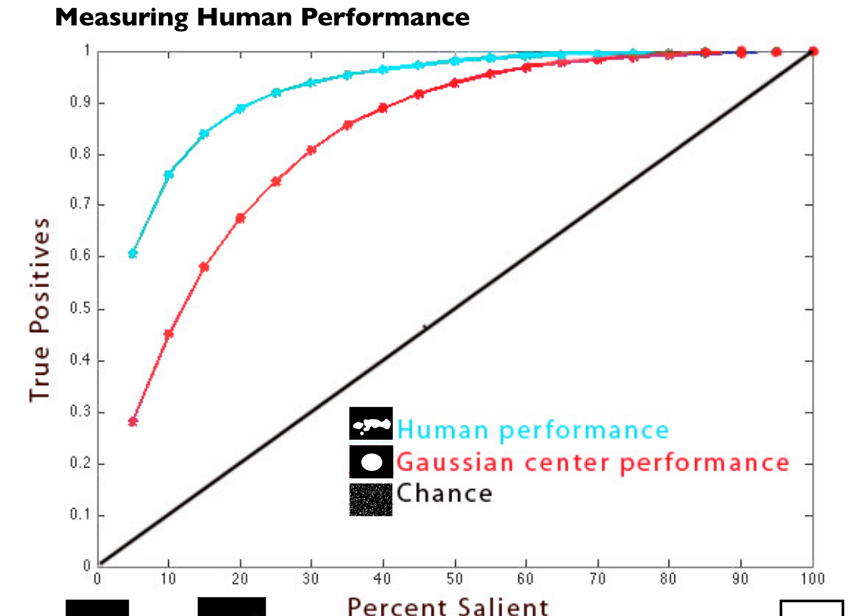
Strong center prior (right)
This is a plot of all the saliency maps from human eye fixations indicating a strong bias to the center of the image. 40% and 70% of fixations lie within the indicated rectangles.

The most salient objects
In our database, viewers frequently fixated on faces, people, and text. Other fixations were on body parts such as eyes and hands, cars and animals. We found theses salient image patches by sampling connected areas of the top 3% most salient pixels.

Size of regions of interest
In many images, viewers fixate on human faces. However, when viewing the close up of a face, they look at specific parts of a face rather than the face as a whole, suggesting a constrained area of the region of interest. On the right is a histogram of the radii of the regions of interest in pixels.



Measuring Human Performance



In this ROC curve, the y axis is the percent of fixations from 15 viewers that lie within the salient region of the image (or when measuring human performance, the percentage of fixations from the 14 other viewers).

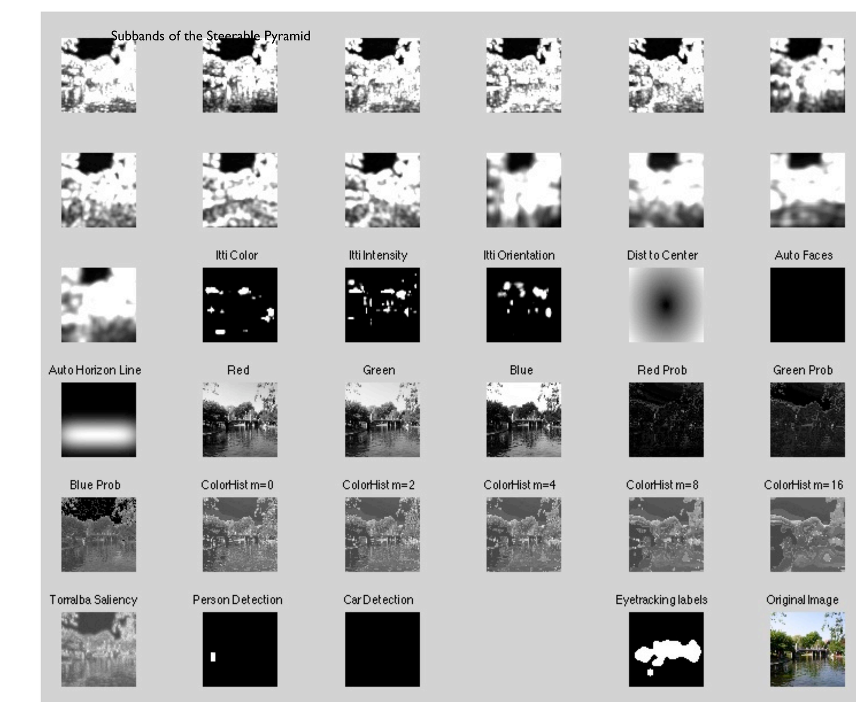
Human performance is very high demonstrating that the locations where a human looks are very indicative of where other humans have looked.

The gaussian center model performs much better than chance because of the strong bias of the fixations in the database towards the center.

Learning a model of saliency

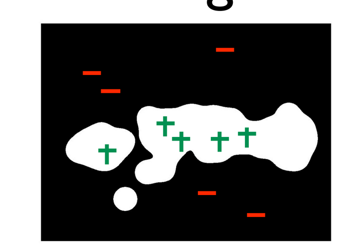
Features

We collect a set of features we believe might be predictive of where people look.

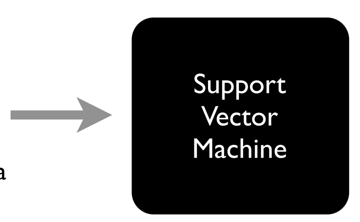


These features include:
low level image features
- illuminance, color, and orientation
high level image context features
- location of the horizon line,
- distance to the center of image,
- presence of a face, person, or car.

Training the model



Training Samples
On a subset of images we chose several salient and non-salient locations as training samples. For each sample we have a label and a vector of feature values.



Learning a Model
We use our training samples to train linear models using a support vector machine. The models aim to find weights for combining features that leads to the most accurate prediction of the saliency label. We test the models on the remaining images in our database to asses performance.

Comparisons

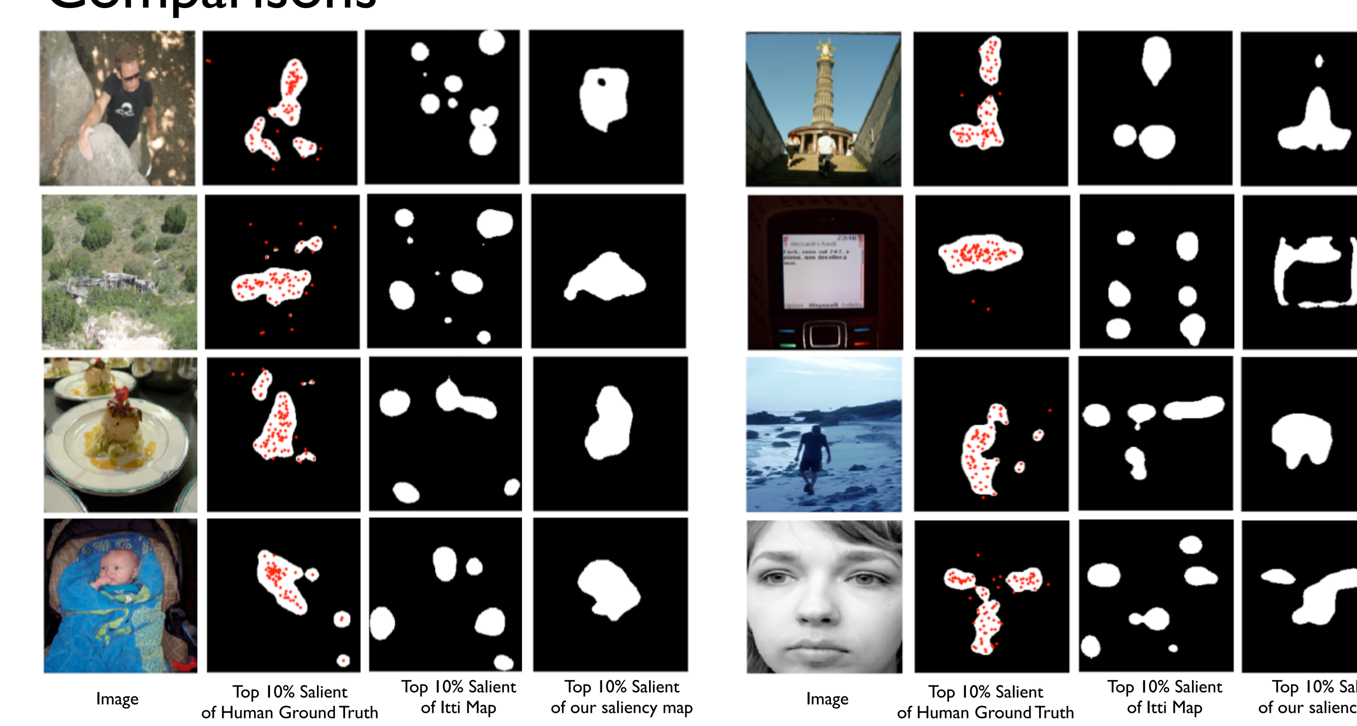
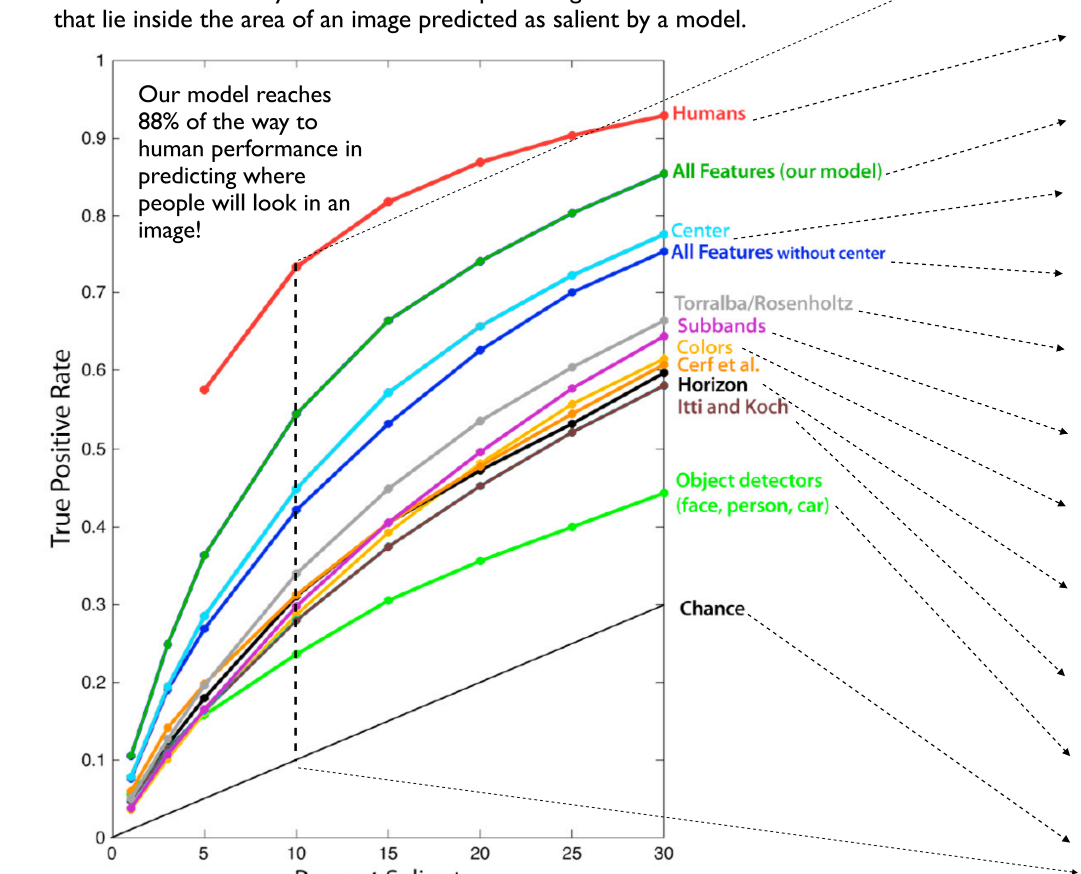
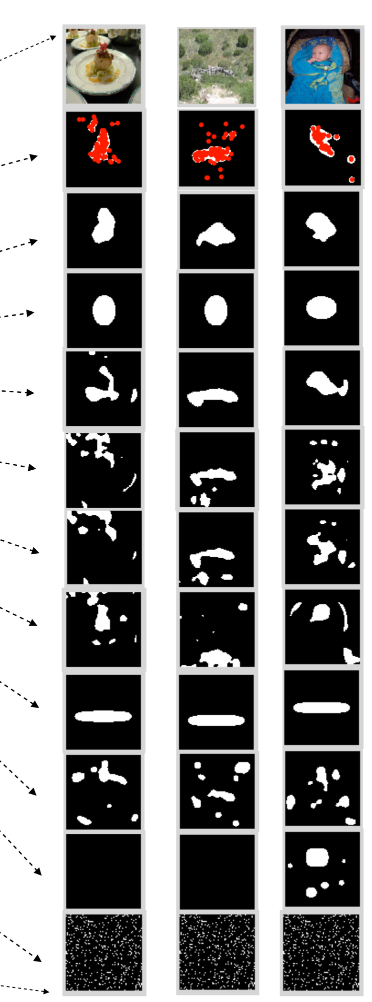


Image | Top 10% Salient of Human Ground Truth with fixation locations | Top 10% Salient of Itti Map | Top 10% Salient of our saliency map

Performance Results

This ROC curve compares the performance of models trained on different sets of features. The y axis indicates the percentage of human fixations that lie inside the area of an image predicted as salient by a model.





Top 10% salient maps from models trained on different features

Acknowledgements

Funding comes from NSF CAREER awards 0447561 and IIS 0747120, Microsoft Research New Faculty Fellowship, Sloan Fellowship, Royal Dutch Shell, Quanta T-Party, MIT-Singapore GAMBIT lab, Xerox graduate fellowship. Thanks to Aude Oliva, Barbara Hidalgo-Sotelo, Nicolas Pinto, Yann LeTallec