Predicting where humans look

Adrià, Andrea, Artem and Lawrence
03/06/2019

Contents

INTRODUCTION

Where do you look?

Where do you look?

Where do you look?

Where do you look?

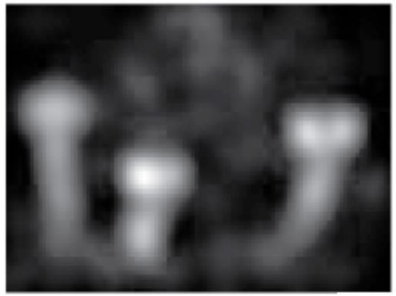
Why?

Why is it important to understand where people look?

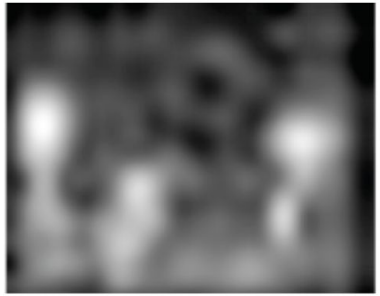
- Automatic image cropping
- Image compression and adaptation to screen
- Advertisement

Models of Saliency

(a) Original image

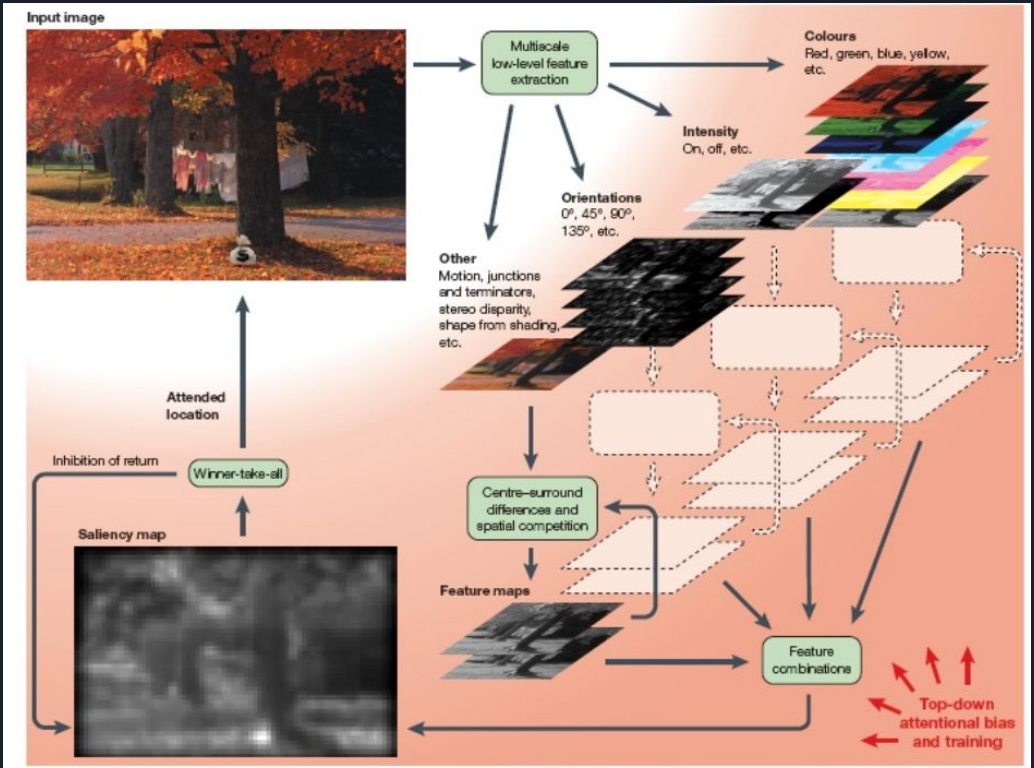


(b) Hou and Zhang



(c) Itti and Koch

A saliency model



Bottom-up saliency model. From Itti and Koch [2001]

Problem with saliency models

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

Solution?

Model of where people look directly from eye tracking data

How?

1) Collect eye tracking dataset
2) Supervised learning model of saliency

ORIGINAL METHOD

Description of Study

Study done by team from MIT (Tilke Judd, Krista Ehinger, Fredo Durand, Antonio Toralba)

Paper published in IEEE International conference on Computer vision(ICCV) 2009

Main objective: Train and evaluate a saliency model based on low, mid and high level features on eye tracking data on selected images

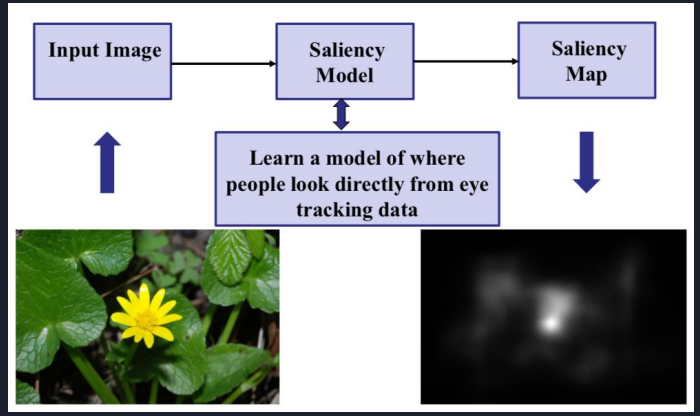
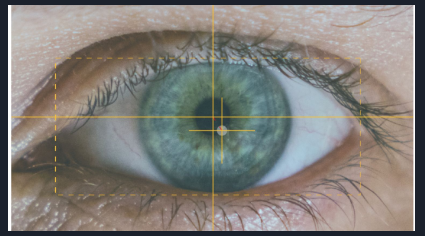


PHOTO CREDIT: CLASS PRESENTATION BY Khimya Khetarpal

Experimental Setup

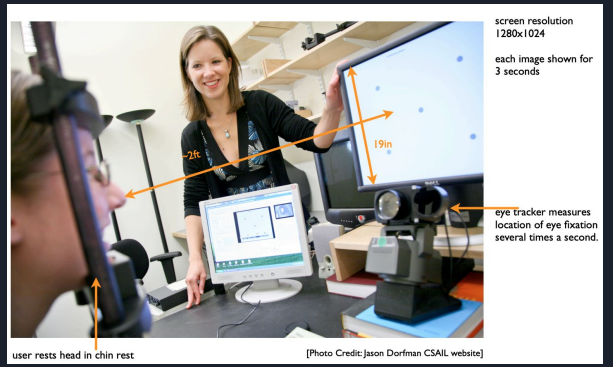
A large database of eye tracking data with labels and analysis generated, and available publicly.



Eye tracking experiment performed with 15 users and 1000 images

- Images size - 405 - 1024 px
- Users(M/F) - 18 - 35 yrs
- Camera calibration checked every 50 images for high quality tracking results
- Memory test to motivate users to pay attention

*Images collected from Flickr and LabelMe



Experimental Setup

Data stored about the spatial and temporal fixations on the image.

Saliency map from a user considered as binary classifier on every pixel, by thresholding portions of the image are classified as fixated or not.

Verified using the saliency maps from the remaining users as the label.

Master saliency map determined "*by convolving a gaussian over the top n(6 used in this case) fixation locations for all 15 users*".

These saliency maps are used as the human ground truth or labels, indicating where humans actually look

Analysis of Dataset

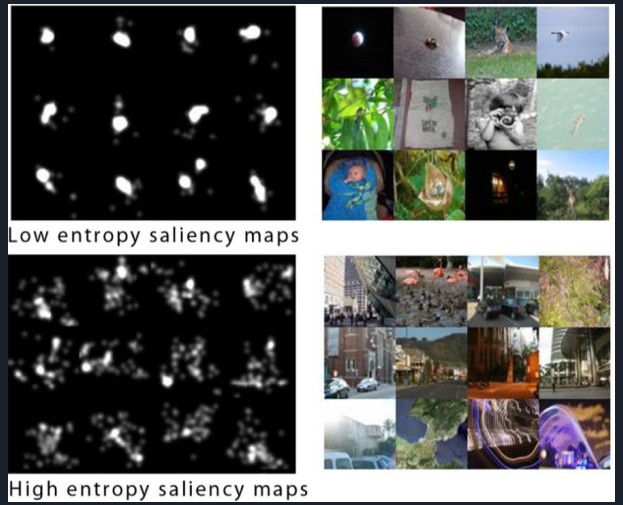
Locations of fixations different for images.

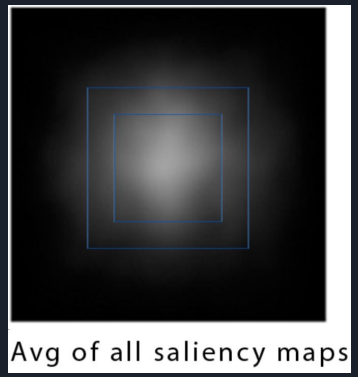
Consistency of fixations over images determined by the entropy of the average continuous saliency map for all the viewers.

Entropy level of image usually correlated with presence of central object or different textures.

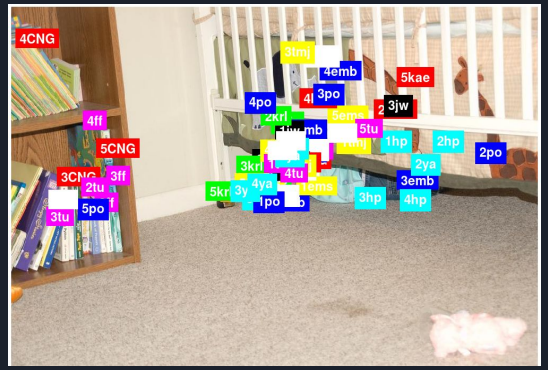
Strong bias towards center of image noticed in saliency maps (*“about 70% of the fixations lie within the centre 25% of the image”*)

Possibly influenced by setup of experiment and propensity of objects of interest to be in the center of photographs

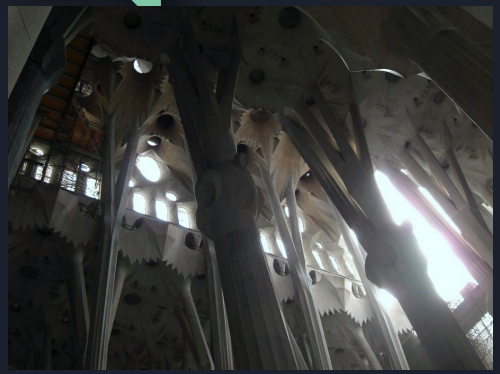




Analysis of Dataset

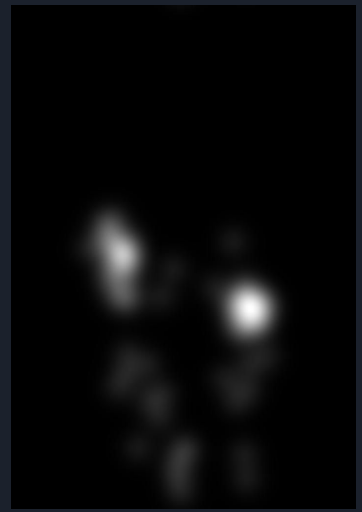


Analysis of Dataset

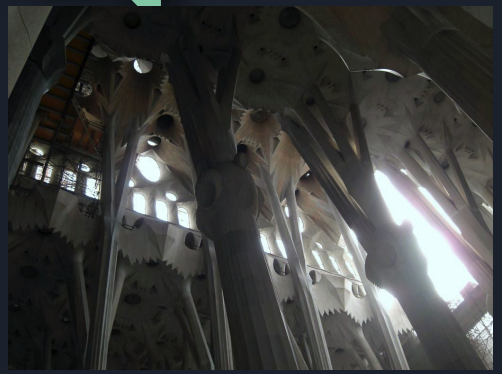


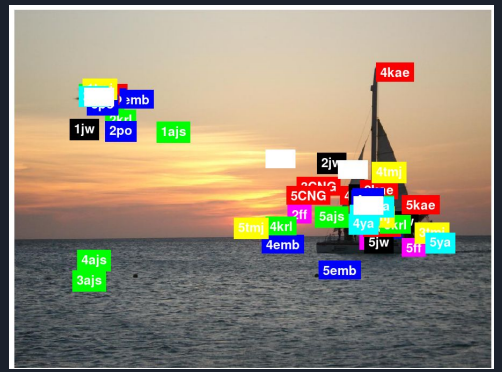


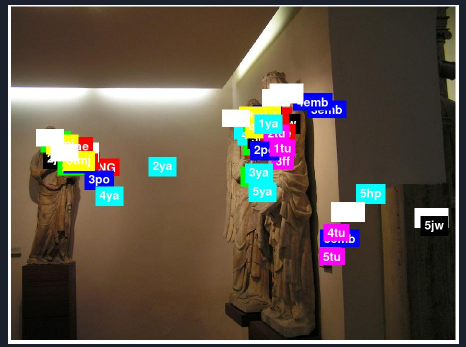


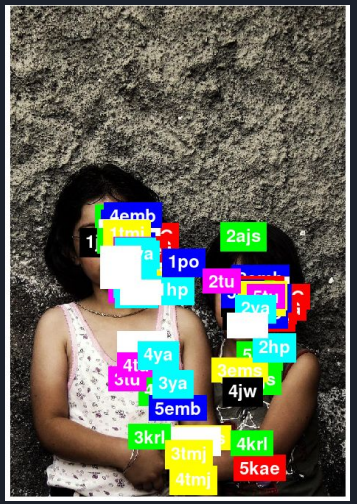



Analysis of Dataset

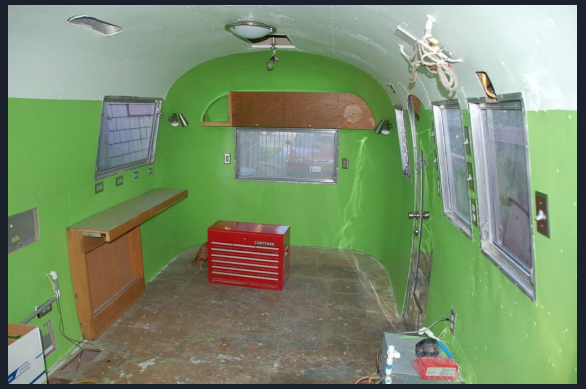


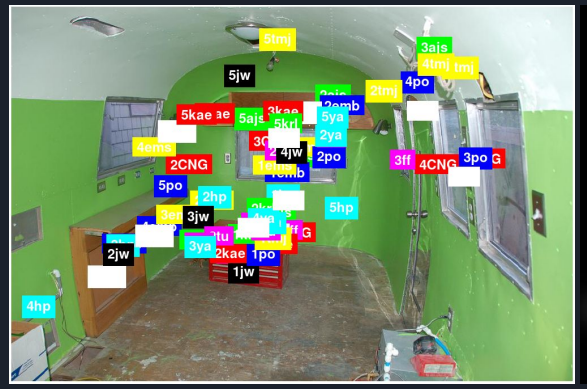




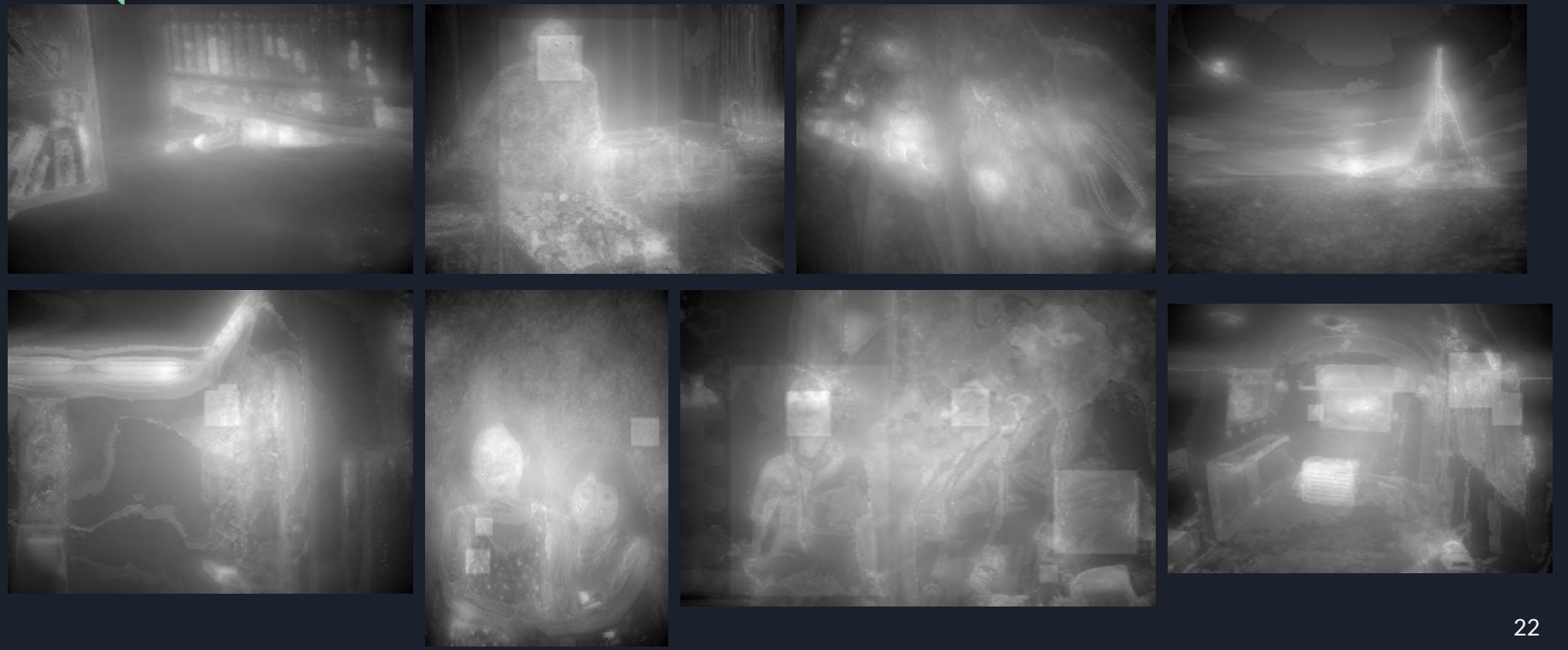


Analysis of Dataset





Analysis of Dataset - *Saliency Maps*

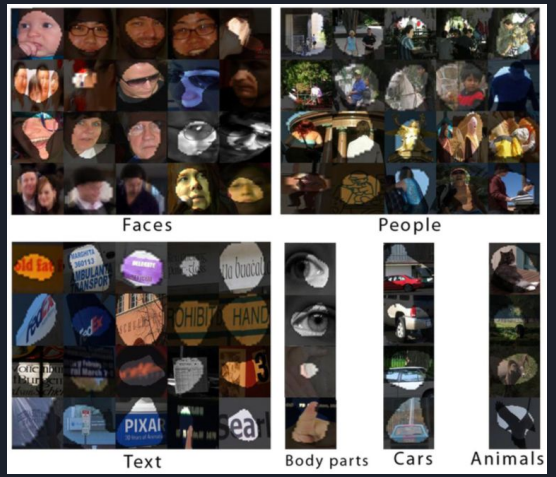


Analysis of Dataset

90% of human fixations(from experiment) are within the top 20% salient locations of a novel viewer's saliency map.

Further analysis of fixations through hand-labelling reveal:

- 10% on faces
- 11% on text
- Most fixation on people and representations of people, animals, cars , human body parts



Model Building - *Feature Selection*

Collection of features assumed to carry the most predictability of where people look in the images

Low level	Mid- Level	High-Level
- Pyramid subbands in 4 orientations and 3 scales - Intensity - Orientation - Color Contrast - Toralba Saliency	- Horizon line **Center Prior** - Distance to center for each pixel	- Face detection feature - Person Detection Feature - Car detection feature

Features (33 in total) precomputed for every pixel of every image

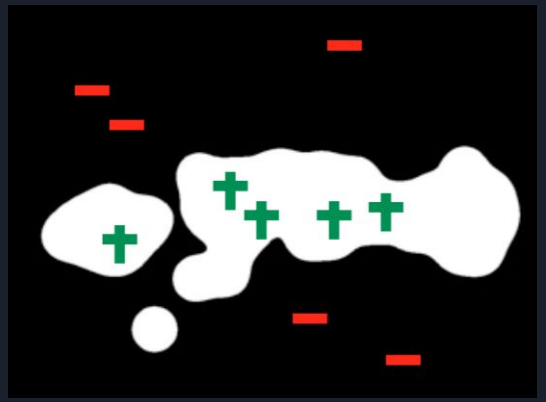
Model Building - *Training and Test Samples*

Dataset divided into 903 training and 100 testing images

For each image, using human ground truth saliency maps, randomly choose:

- 10 positively labelled pixels from top 20% salient locations
- 10 negatively labelled pixels from bottom 70% salient locations

In order to have strong positive and strong negative samples



Total of 18060 and 2000 training and testing samples, respectively

Features normalized (mean = 0, variance =1) for both train

and test samples

Model Building - *Training Model*

Support Vector Machine algorithm used with linear kernels

Model trained to find weights for linear combination of the features leading to the most accurate prediction of saliency

Misclassification cost *c*, set at 1.

(No impact on performance in [1,1000])

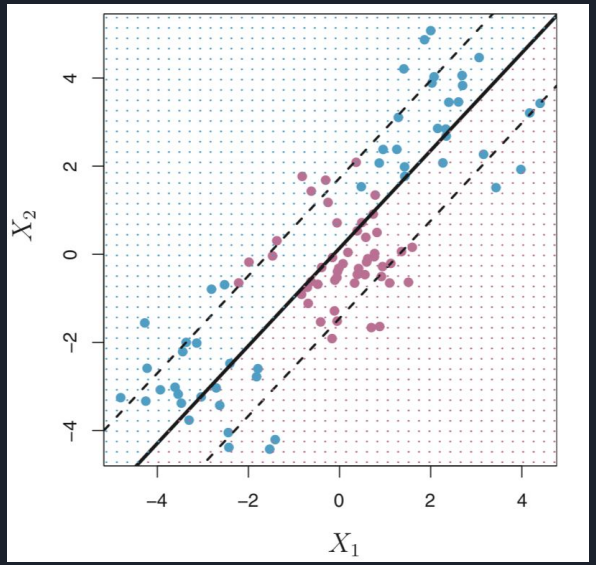


Photo credit: ISLR

RESULTS

Results

Researches measure performance of saliency models in two ways:

- Performance of each model by its ROC curve.

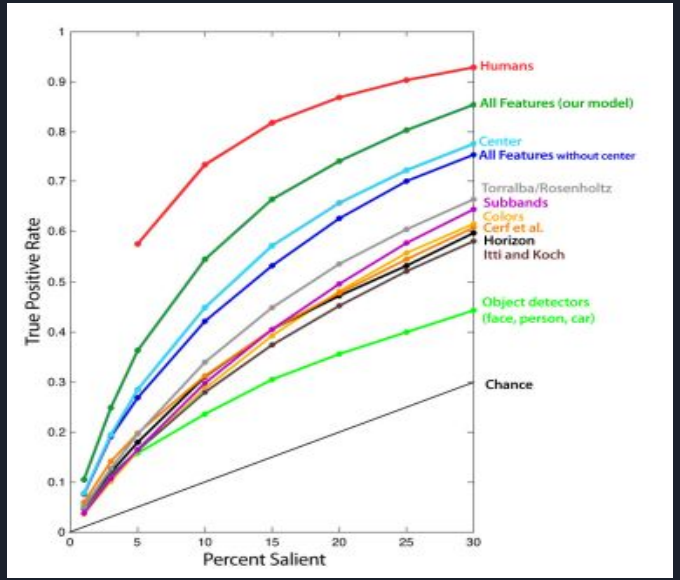
- Examine the performance of different models on specific subsets of samples: samples inside and outside a central area of the image and on faces.

Results

Prediction of the saliency per pixel using a specific trained model.

For each map, we find the percentage of human fixations within the salient areas of the map as the measure of performance.

For visualization was used ROC metric



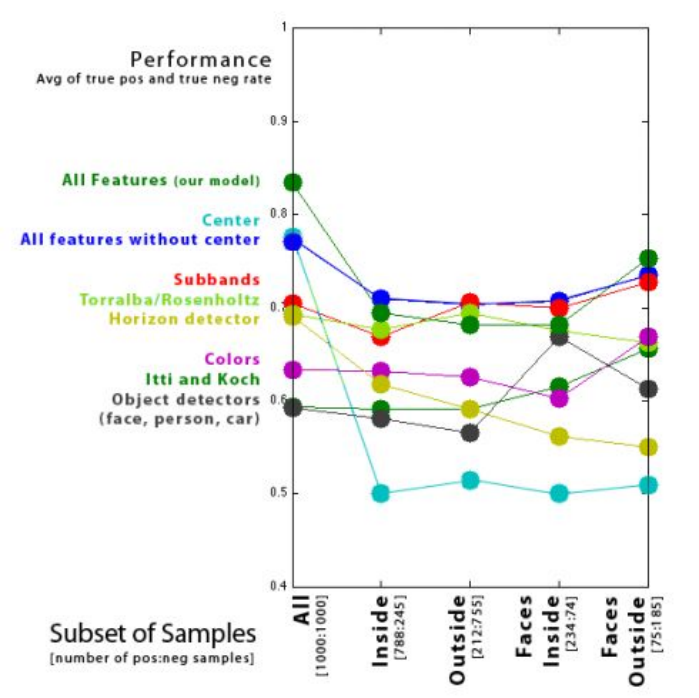
ROC metric is a metric to evaluate the performance of human saliency maps to predict eye fixations. Using this method, the saliency map from the fixation locations of one user is treated as a binary classifier on every pixel in the image. Saliency maps are thresholded such that a given percent of the image pixels are classified as fixated and the rest are classified as not fixated.

Results

Each image divided into a circular central and a peripheral region

In figure was shown the average rate of true positives and true negatives for SVMs trained with different feature sets on different subsets of samples.

This value is equivalent to the performance of the model if there were an equal number of positive and negative samples in each subset.



Discussion

In general, the fixation locations of several humans is strongly indicative of where a new viewer will look.

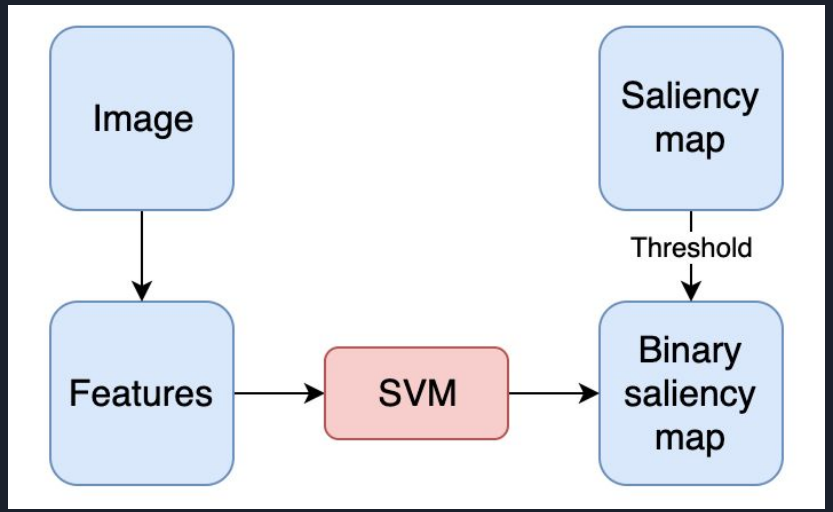
So far, computer generated models have not matched humans' ability to predict fixation locations

As text, face, person and other object detectors get better, models of saliency which include object detectors will also get better.

OUR APPROACH

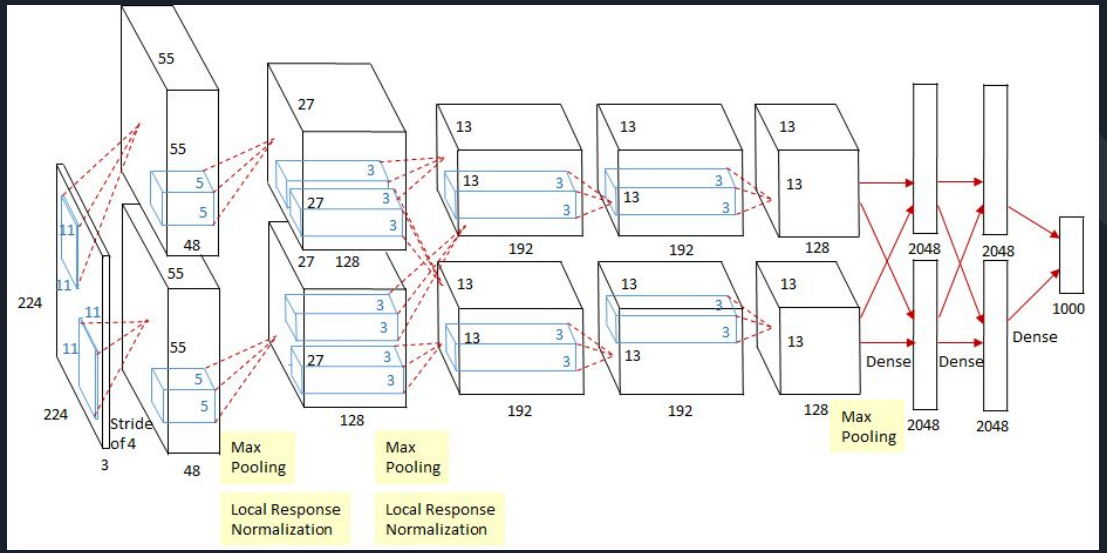
Related Work

- Judd, Tilke, et al. "Learning to predict where humans look." 2009 IEEE 12th international conference on computer vision. IEEE, **2009**.



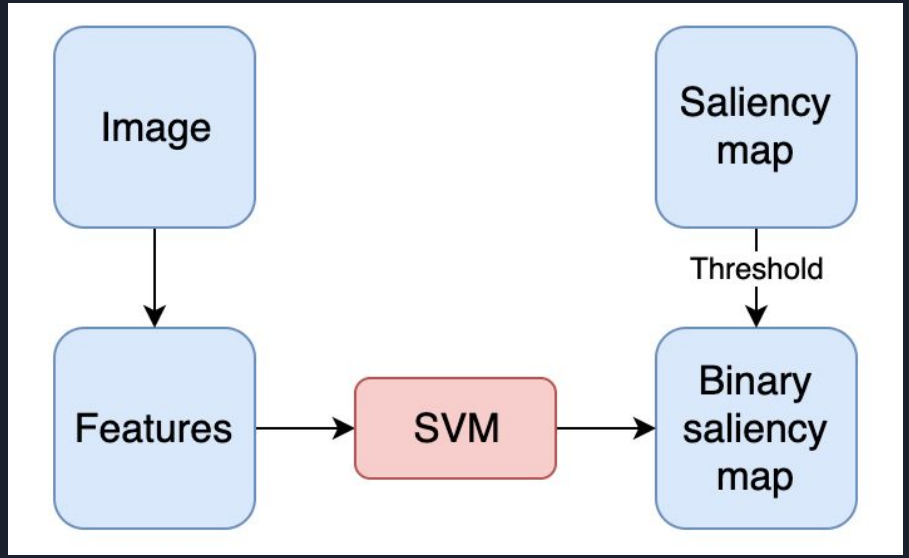
Related Work

- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. **2012**.



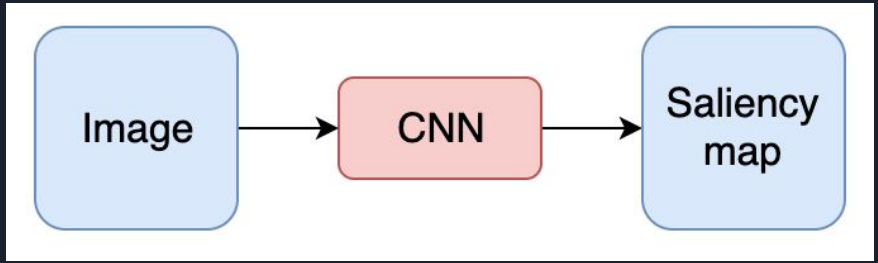
Related Work

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



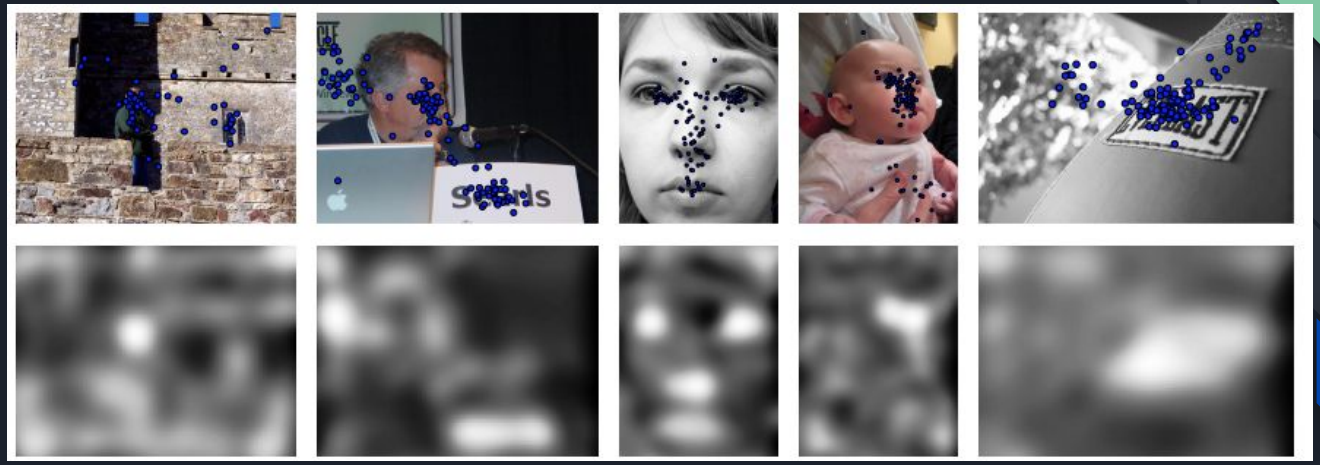
Related Work

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



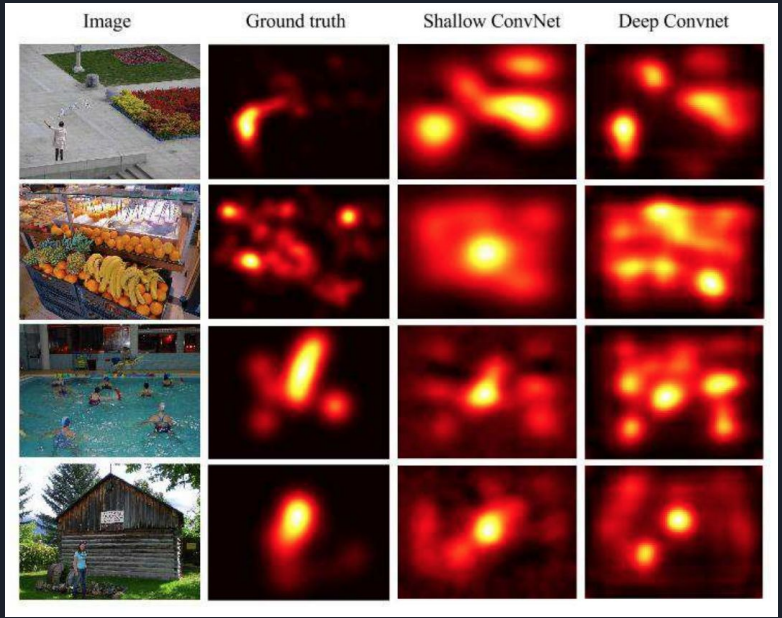
Related Work

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



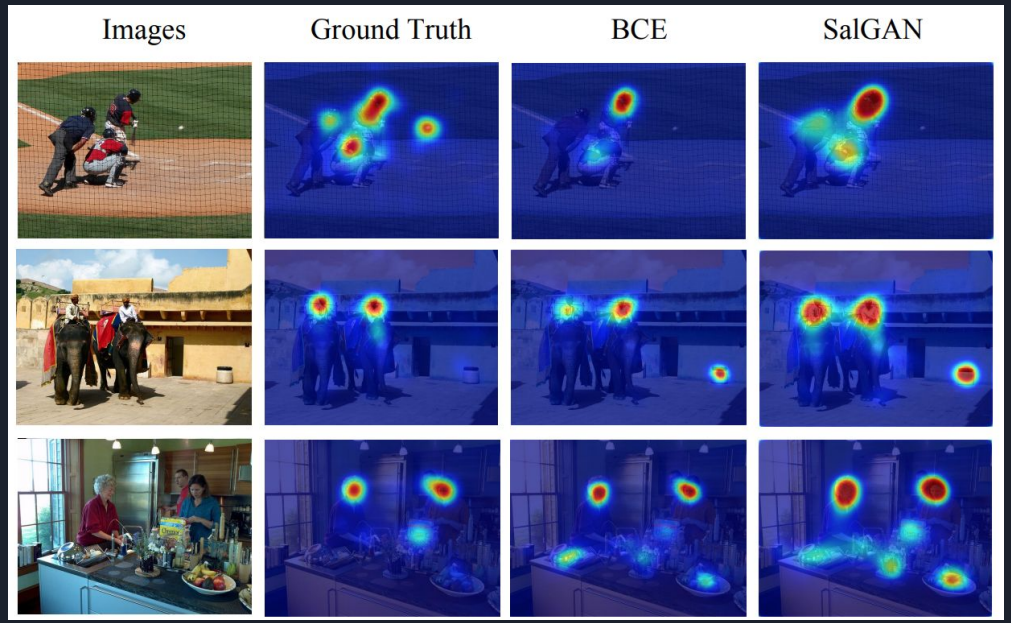
Related Work

- Pan, Junting, et al. "Shallow and deep convolutional networks for saliency prediction." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. **2016**.



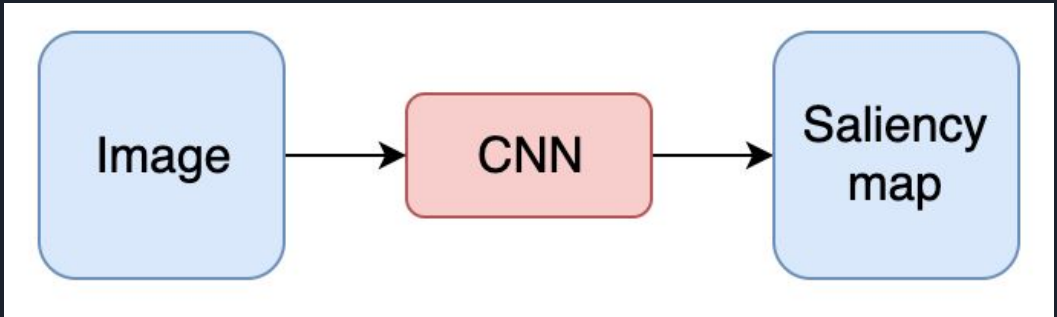
Related Work

- Pan, Junting, et al. "Salgan: Visual saliency prediction with generative adversarial networks." arXiv preprint arXiv:1701.01081. **2017**.



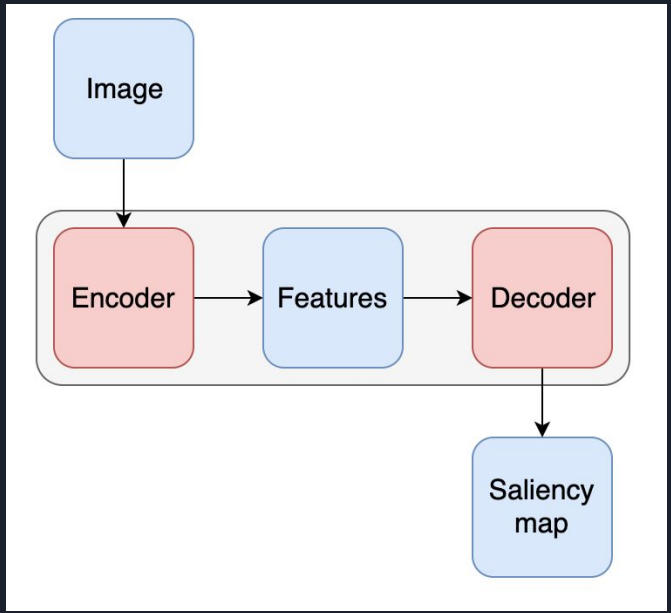
Our Methodology

- **Convolutional neural network** based implementation.



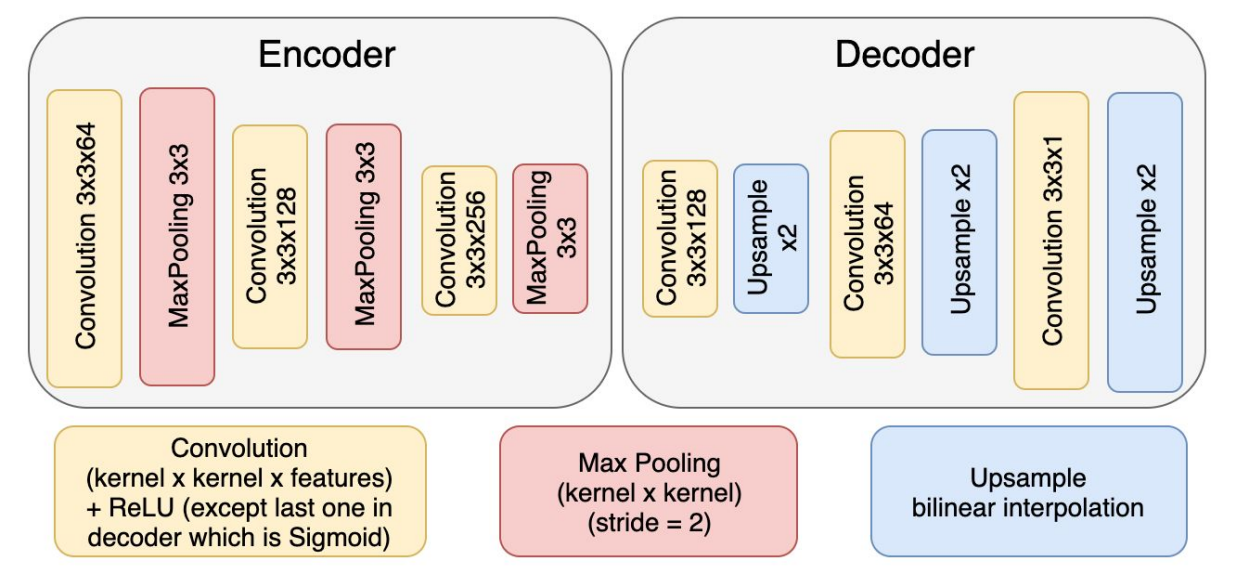
Our Methodology

- We use a **encoder** network to extract features from the image. Then we process this features by means of a **decoder** in order to obtain a saliency map.



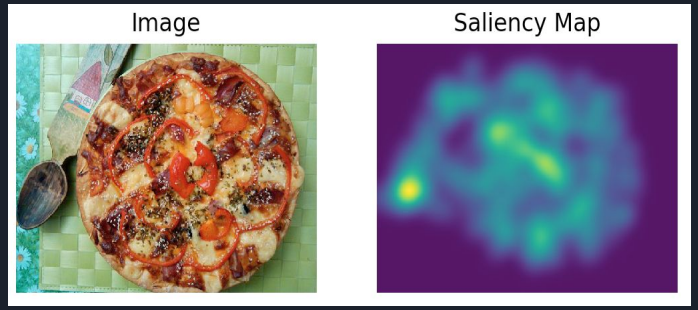
Our Methodology

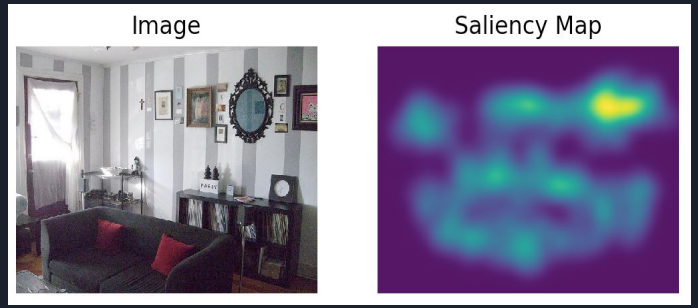
- We use a **encoder** network to extract features from the image. Then we process this features by means of a **decoder** in order to obtain a saliency map.

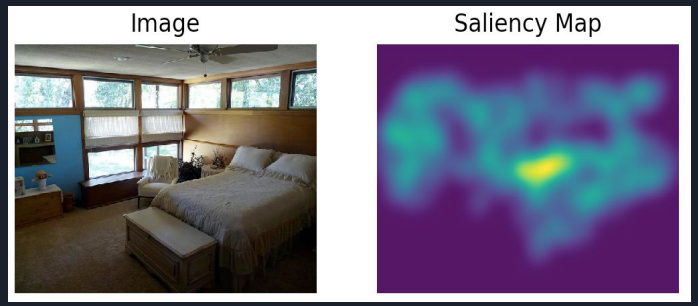


Dataset

- We trained the network with the saliency maps which we obtained from SALICON dataset: http://salicon.net/challenge-2017/







Our Results


Image
Saliency Map
Prediction


Image
Saliency Map
Prediction

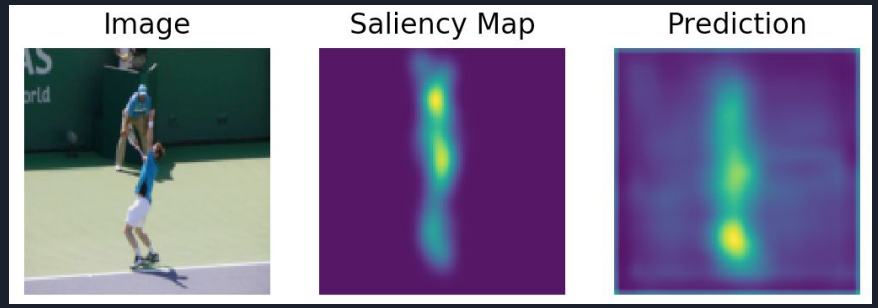


Image
Saliency Map
Prediction


Image
Saliency Map
Prediction


Image
Saliency Map
Prediction

Application

Modern eye track models are becoming increasingly common in studies concerning many fields of life, starting at navigating in a real-life shops and looking at shelves, perceiving the urban space, through watching outdoor advertisements and television commercial, driving a car, up to sports games

- Usability testing
- Ergonomics
- Psychology
- Advertising
- Studying the shelves in shops
- Public information systems
- Urban improvement

Application

Model can be used even in data visualization. Analyzing and predicting where people will look is important part of work of every data scientist. It is necesary to focus attention of auditory on details which are important for presentator. Using low-levels (colors, texture) features model can understand where people will look and using this information we can improve possibilities of visualization. For example this text is almost useless. I'm already talking for one minute but you still didn't finish read this text and didn't get the message. I can write some strange words here like cucumber, beach, bocadillo, tapas. Today is already third of june and i guess almost everyone thinking about the beach. Almost nobody mentioned it because your first attention was attracted to the center of the slide and then to the heading. Changing some elements on this slide model can predict will be eye tracking useful in presentation or not. Let's try?

Application

Model can be used even in data visualization. Analyzing and predicting where people will look is important part of work of every data scientist. It is necessary to focus attention of auditory on details which are important for presentator. Using low-levels (colors, texture) features model can understand where people will look and using this information we can improve possibilities of visualization. For example this text is almost useless. I'm already talking for one minute but you still didn't finish read this text and didn't get the message. I can write some strange words here like cucumber, beach, bocadillo, tapas. Today is already third of june and I guess almost everyone is thinking about the beach. Almost nobody mentioned it because your first attention was attracted to the center of the slide and then to the heading. Model can predict if changing some elements on this slide will be useful in presentation or not. Let's try?

Conclusions - TBD

- There is a database of eye tracking with possibility of improving in a future
- Can learn of models of saliency and BEYOND
- Future work will focus on enhance model, explore cropping
- Big field for implementation and application

References

- https://imotions.com/blog/eye-tracking-work/
- https://www.researchgate.net/publication/312578921_The_application_of_eye_tracking_in_business

END