Learning to Predict Where Humans Look

Presented by: Chris Thomas
Paper By: Tilke Judd, et al.
Presentation Slides Loosely Based on Slides on site
What is common to these situations?

• In both situations:
  • Need to prioritize the visual field and decide what is important (i.e. salient)

• Why would you want to know where people look in an image?
  • Automatic Image Cropping (first example)
    ➢ Extracting thumbnails on computer
  • Image search
  • Compression
  • Advertising design
  • Non-photorealistic rendering
  • Object detection
  • Self-driving car / robot navigation
  • …. Many other applications

• How do you get the data?
  • Eye-tracking data = Expensive / time consuming
  • Computational approaches – Not as accurate
heat map of 200 people

6% of people look at the package

84% of people looked at the package
What is Saliency?

• Saliency is the quality that makes certain areas, objects, regions, etc. in the visual field stand out from the rest of the field and grab our attention

• This ability has developed evolutionarily
  • Critical to rapidly find predators, prey, etc.
  • Information overload in the visual field - too much to process at once
  • Solution: Do higher level processing serially
  • Problem: How do you know what to look at?

• Neuroscience research has shown that early visual neurons are specialized to detect certain types of basic “features” compared to their surroundings
  • Orientation contrast
  • Color contrasts
  • Lumination contrasts
  • Motion
  • Flicker
Bottom-Up Saliency Models

- Often based on biologically plausible linear filters
- Measure intensity, illumination, color, flicker, orientation contrast
- Lots of parameters need to be tuned
- Bottom-up
- Doesn’t always predict fixations well

Bottom-up saliency model. From Itti and Koch [2001]
Bottom-Up Influences
Bottom Up Influences
Find the pedestrian
There is no pedestrian
Where do people actually look?

- Top-down task and scene dependent cues
  - Semantic understanding, memories, task
- Bottom-up saliency cues
Many models for saliency exist

<table>
<thead>
<tr>
<th>Biologically Inspired</th>
<th>Mathematically Inspired</th>
<th>Add top-down features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cerf et al. (2007)</td>
<td>Graphical Model</td>
<td>(search task)</td>
</tr>
<tr>
<td>SUN model</td>
<td>Itti and Baldi (2006)</td>
<td></td>
</tr>
<tr>
<td>Goferman et al. (2009)</td>
<td>“Surprise” model</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Elazary and Itti (2010)</td>
<td></td>
</tr>
</tbody>
</table>
Problems and Contributions of Paper

- Many of the bottom up saliency models have many, many parameters that need artificially tuned
- Top-down attention is ignored is many models
- Lots of models exist, but no large standard benchmark existed
- Paper aims to Address these Problems:
  - Integrate many of the ideas from many existing bottom-up and top-down algorithms in one system trained using machine learning
  - Consider low-level, mid-level, and high-level visual features
  - Provide the largest scale eye-tracking dataset as a benchmark

- **Goal:** Learn where people look directly from the eye tracking data
1003 Natural Images from Flickr and LabelMe
Eye-Tracking Experiments

- 15 People free viewed the images (actually done in dark room)
- Calibrate check every 50 images
- Divide the experiment into two sessions of 500 randomly ordered images
- Initial fixation is discarded
Fixations for One Person

Fixations Vs Saccades
Fixations from 15 People
Fixation Map

• Just convolve a Gaussian over the fixation positions

• Can do this for a single person, or an aggregate of many people

• Can also threshold:
  • Choose top n %
  • Yields a binary map
  • See Figure 1D
Analysis: Fixation Consistency Depends on Image Content

Low entropy saliency maps

High entropy saliency maps

Avg of all saliency maps

Entropy
Observation: Fixations are Center Biased

- 40% of fixations are within the center 11% of the image
- 70% of fixations lie within the center 25% of the image

Why?
- **Photographic Bias** – Photographer puts objects of interest near center
- **Viewing Bias** – People expect to see the objects of interest near the center
- **Experiment Bias** – Viewers sitting right in front of the screen (not angled)
- **Viewing Strategy** – Center is a good place to start exploring
- **Orbital Reserve** – The eyes tend to look straight ahead (muscle laziness)
Evaluating Saliency Maps

- Saliency map from fixations of one person is thresholded to become a binary classifier
  - 1 if pixel at x,y over threshold, 0 otherwise
  - Compare against fixations from the other 14 people
- By varying the threshold we can get a ROC curve
  - Receiver Operating Characteristic Curve
  - Average over all users and all images
Human Fixations on Saliency Map

Human Fixations
ROC Curve - Humans

We calculate the percentage of fixations that lie within the salient portion of the map.
**ROC Curve - Humans**

We calculate the percentage of fixations that lie within the salient portion of the map.
We calculate the percentage of fixations that lie within the salient portion of the map.
We calculate the percentage of fixations that lie within the salient portion of the map.
ROC Curve - Humans

Thresholded Saliency Map

Receiver Operating Characteristic curve

Percent Salient (False positives)

% of human fixations (True positives)
ROC Curve Humans

ROC curve always starts at 0
ends at 1
ROC Curve - Center Bias

Thresholded Center Map

Receiver Operating Characteristic curve

% of human fixations (True positives)

Percent Salient (False positives)
ROC Curve - Center Bias

Thresholded Center Map

Receiver Operating Characteristic curve

% of human fixations (True positives) vs. Percent Salient (False positives)
ROC Curve - Center Bias

Thresholded Center Map

Receiver Operating Characteristic curve

% of human fixations (True positives)

Percent Salient (False positives)
ROC Curve - Random Noise

Random Noise map

Receiver Operating Characteristic curve

% of human fixations (True positives)

Percent Salient (False positives)
ROC Curve – Random Noise

Random Noise map

Receiver Operating Characteristic curve

Chance performance

% of human fixations (True positives)

Percent Salient (False positives)
**ROC Curve - Optimal**

- **Perfect saliency map**
- **Receiver Operating Characteristic curve**
- **Best performance**
- **Chance performance**

**Legend**:
- % of human fixations (True positives)
- Percent Salient (False positives)
ROC Curve in the Paper - Figure 5
Where do people actually look?

- Labeled faces / text with bounding box and horizon line (if any)

Based on the annotations:
- 10% fixation on faces
- 11% on text

Animals, cars, human body parts
Size of Regions of Interest

Drew a bounding box around clustered fixations and revealed that the radius of ROIs tend to be similar.

Close-ups fixate on particular parts of the face, whereas in the other example focus on whole face.
Learning Classifier from Eye-Tracking Data

• Compute the features for a resized image of 200x200 to train

• Feature Types:
  • Low Level Features: Illumination, Orientation, Color…
    ➢ Steerable pyramid filters, Itti-Koch, Torralba, RGB values and probabilities of each
  • Mid-Level Features: Horizon Line
    ➢ Use gist features for this
  • High Level Features: Face detectors, person detectors
    ➢ Viola-Jones face detector
    ➢ DPM person detector
  • Center Prior – Distance to image center for each pixel
Low Level Image Features

Use subbands of the steerable pyramid filters (Simoncelli and Freeman 1995)
4 orientations

Simple saliency map from Torralba based on steerable filters

3 scales

Color Orientation Intensity

Itti-Koch
Low-Level Image Features (2)

<table>
<thead>
<tr>
<th>Red</th>
<th>Green</th>
<th>Blue</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
</tr>
<tr>
<td><img src="image4" alt="Image" /></td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
</tr>
<tr>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
<td><img src="image9" alt="Image" /></td>
</tr>
<tr>
<td><img src="image10" alt="Image" /></td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
</tr>
</tbody>
</table>

- **Color channels**
- **Probability of colors**
- **Compute probability of colors from 3D color histograms at different scales**
All Features (2)
Training Procedure

33 features

Support Vector Machine

10 salient samples from top 20%
10 non-salient samples from bottom 70%
from each of 903 training images

Learns weights for each feature that best predict labels
Use linear kernels

Model
Testing / Training Details

• Training:
  • Observations:
    - Adding more than 10 samples did not improve the performance
    - Using different ratios of positive / negative did not affect performance
  • Features were normalized to zero mean and unit variance before training
  • 9030 positive and 9030 negative training instances
  • Train models for each one of the types of features to compare against each other

• Testing:
  • For each test image, compute the 33 features.
    - For every pixel in the image
      - Give SVM the features at that pixel location from all the feature maps
      - Directly use the SVM score as the saliency per pixel (don’t just do a binary classification)
    - Threshold the image and compute ROC scores
  • Test set size 100
Second Experiment

• Motivation – Want to understand the impact of the center bias

• Experiment Design:
  • Design the image into a circular central region surrounded by a peripheral region
  • Central region is 0.42 (normalized) units from center or less
    ➢ This region was labelled positive and the rest was peripheral was negative
  • Divide the samples to consider those inside and outside of the center and test separately to see how good each model is at predicting samples in those regions
  • Also, use the face annotations from before to test how good the model is at finding samples inside face annotated regions and outside face annotated regions

• Report performance as the average of true positive and true negative rates
The center model performs only as well as chance for the subsets of samples.
All features without center does roughly the same as center on the total set, but on the subsets, it is more robust.
The object detector model performs best on the subsets with faces (as expected).
The All Features model (which uses the center prior) is outperformed on the subsets inside and outside.

This is because it takes a hit because of the center bias, however all features without center remains roughly at the top for inside, outside, and faces inside, and almost for faces outside.
NPR Application
Conclusions

• The paper presents an approach combining top-down and bottom-up cues to compute a saliency map of an image to determine where people are likely to look in the image.

• The model is trained using human eye-tracking data and features fed into a SVM classifier to predict the saliency of each pixel.
  • Could be considered a weakness ⇒ We need to have eye-tracking data.

• Future Work:
  • Integrating context/scene – What objects are salient may depend on the context.
  • How to integrate surprise? (objects which don’t fit in that context).
  • Task?
  • Why not have a text detector? They know that text really matters so why not just detect text as another feature. This shouldn’t be too difficult.
  • Compare against truly top-down models of saliency
    ⇒ High level features here are detections from DPM but more robust Top down models exist.
  • Objectness.
Understanding and Predicting Importance In Images

Secondary Paper

Berg, et al.

Presented by Chris Thomas
Problem Overview

- Problem: What do people consider important in an image?
  - Different problem than where people look in an image
Problem Overview

• Problem: What do people consider important in an image?
  • Different problem than where people look in an image

“A raft with 3 adults and two children in a river.”
“Four people in a canoe paddling in a river lined with cliffs.”
“Several people in a canoe in the river.”
Problem Overview

- Problem: What do people consider important in an image?
  - Different problem than where people look in an image

Why don’t people describe this? Why isn’t it important?
The authors propose a number of factors for predicting importance:

- Compositional Factors – Size / location
- Content Semantics – Object category (human vs rock) / scene category + strength
- Contextual Factors – Object-Scene / Attribute-Object context

They evaluate the models by looking at image content labels and descriptions of images written by people of images.

Learn models to predict what people would describe:

- Given what is in the image (algorithm told what is there)
- Automatically detect image content and then predict what is important

**Motivations**: Image search / generating natural descriptions for images
Approach Overview

• Collect Dataset

• Content Labels: Object labels, scene labels, attribute labels
  • Use Mturk to get scene labels and attribute labels

• Mapping from content to description
  • Need to know which object “child” or “bearded man” refers to
    ➢ Also scene category labels + attribute labels
  • Perform this by hand
  • Also devise an automatic method based on matching description to label

• Evaluation of the factors

• Building / Evaluating predictive models of importance
Mapping Labels to Descriptions

Human description:
A man is riding a brown horse

Amazon Mechanical Turk
Result: Compositional Factors

![Bar Graphs showing Location and Size with probability of being described on the y-axis and central/peripheral or smaller/larger on the x-axis.](image)
Result: Compositional Factors

Objects further from center are less likely to be mentioned
Result: Compositional Factors

Larger objects are more likely to be mentioned, unless it is too big
## Semantic Factors

<table>
<thead>
<tr>
<th>Top10</th>
<th>Prob</th>
<th>Last10</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>firework</td>
<td>1.00</td>
<td>hand</td>
<td>0.15</td>
</tr>
<tr>
<td>turtle</td>
<td>0.97</td>
<td>cloth</td>
<td>0.15</td>
</tr>
<tr>
<td>horse</td>
<td>0.97</td>
<td>paper</td>
<td>0.13</td>
</tr>
<tr>
<td>pool</td>
<td>0.94</td>
<td>umbrella</td>
<td>0.13</td>
</tr>
<tr>
<td>airplane</td>
<td>0.94</td>
<td>grass</td>
<td>0.13</td>
</tr>
<tr>
<td>bed</td>
<td>0.92</td>
<td>sidewalk</td>
<td>0.11</td>
</tr>
<tr>
<td>person</td>
<td>0.92</td>
<td>tire</td>
<td>0.11</td>
</tr>
<tr>
<td>whale</td>
<td>0.91</td>
<td>smoke</td>
<td>0.09</td>
</tr>
<tr>
<td>fountain</td>
<td>0.89</td>
<td>instrument</td>
<td>0.07</td>
</tr>
<tr>
<td>flag</td>
<td>0.88</td>
<td>fabric</td>
<td>0.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Prob-ImageCLEF</th>
<th>Prob-Pascal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animate</td>
<td>0.91</td>
<td>0.84</td>
</tr>
<tr>
<td>Inanimate</td>
<td>0.53</td>
<td>0.55</td>
</tr>
</tbody>
</table>

ImageCLEF

UIUC

84
Some object types are more likely than others to be mentioned.

<table>
<thead>
<tr>
<th>Top10</th>
<th>Prob</th>
<th>Last10</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>fireworks</td>
<td>1.00</td>
<td>hand</td>
<td>0.15</td>
</tr>
<tr>
<td>turtle</td>
<td>0.97</td>
<td>cloth</td>
<td>0.15</td>
</tr>
<tr>
<td>horse</td>
<td>0.97</td>
<td>paper</td>
<td>0.13</td>
</tr>
<tr>
<td>pool</td>
<td>0.94</td>
<td>umbrella</td>
<td>0.13</td>
</tr>
<tr>
<td>airplane</td>
<td>0.94</td>
<td>grass</td>
<td>0.13</td>
</tr>
<tr>
<td>bed</td>
<td>0.92</td>
<td>sidewalk</td>
<td>0.11</td>
</tr>
<tr>
<td>person</td>
<td>0.92</td>
<td>tire</td>
<td>0.11</td>
</tr>
<tr>
<td>whale</td>
<td>0.91</td>
<td>smoke</td>
<td>0.09</td>
</tr>
<tr>
<td>fountain</td>
<td>0.89</td>
<td>instrument</td>
<td>0.07</td>
</tr>
<tr>
<td>flag</td>
<td>0.88</td>
<td>fabric</td>
<td>0.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Prob-ImageCLEF</th>
<th>Prob-Pascal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animate</td>
<td>0.91</td>
<td>0.84</td>
</tr>
<tr>
<td>Inanimate</td>
<td>0.53</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Animate Objects Are more Likely to be mentioned than Inanimate.
Semantic Factors - Scene Level

<table>
<thead>
<tr>
<th>scene type</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>office</td>
<td>0.29</td>
</tr>
<tr>
<td>airport</td>
<td>0.13</td>
</tr>
<tr>
<td>kitchen</td>
<td>0.36</td>
</tr>
<tr>
<td>dining room</td>
<td>0.21</td>
</tr>
<tr>
<td>field</td>
<td>0.16</td>
</tr>
<tr>
<td>living room</td>
<td>0.13</td>
</tr>
<tr>
<td>street</td>
<td>0.18</td>
</tr>
<tr>
<td>river</td>
<td>0.1</td>
</tr>
<tr>
<td>restaurant</td>
<td>0.28</td>
</tr>
<tr>
<td>sky</td>
<td>0.18</td>
</tr>
<tr>
<td>forest</td>
<td>0.0</td>
</tr>
<tr>
<td>mountain</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Probability of description for each scene type

<table>
<thead>
<tr>
<th>rating</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>0.21</td>
</tr>
<tr>
<td>3</td>
<td>0.21</td>
</tr>
<tr>
<td>4</td>
<td>0.22</td>
</tr>
<tr>
<td>5</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Scene Depiction Strength and Rating Probability
Context Factors

The diagram represents the context factors with entries such as "office", "airport", "kitchen", etc. Each entry is color-coded with purple indicating 'Low' and brown indicating 'High'. The diagram visualizes the occurrence and importance of various contexts for different items.
Context Factors

Unusual Object-Scene Pairs

Low
High
Attribute-Object Context

![Graphs showing the context of Person and Cow attributes](image_url)
Person-Riding is a rather unusual attribute-object pair and influences importance
Predicting Importance

Known Image Content

Estimated Content

Random Split

Object

Scene

Attribute

Predict Accuracy

Discriminative Models (Logistic Regression)

Repeat 10 times, measure mean and standard deviation
### Results

We are trying to predict whether an object will be mentioned in a description.

Baseline: Predict everything

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Accuracy% (std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (ImageCLEF)</td>
<td></td>
<td>57.5 (0.2)</td>
</tr>
<tr>
<td>Log Reg (ImageCLEF)</td>
<td>$K^s_o + K^l_o$</td>
<td>60.0 (0.1)</td>
</tr>
<tr>
<td>Log Reg (ImageCLEF)</td>
<td>$K^c_o$</td>
<td>68.0 (0.1)</td>
</tr>
<tr>
<td>Log Reg (ImageCLEF)</td>
<td>$K^c_o + K^s_o + K^l_o$</td>
<td>69.2 (1.4)</td>
</tr>
<tr>
<td>Baseline (UIUC-Kn)</td>
<td></td>
<td>69.7 (1.3)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Kn)</td>
<td>$K^s_o + K^l_o$</td>
<td>69.9 (0.6)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Kn)</td>
<td>$K^c_o$</td>
<td>79.8 (1.4)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Kn)</td>
<td>$K^c_o + K^s_o + K^l_o$</td>
<td>82.0 (0.9)</td>
</tr>
<tr>
<td>Baseline (UIUC-Est)</td>
<td></td>
<td>76.5 (1.0)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Est)</td>
<td>$E^s_o + E^l_o$</td>
<td>76.9 (1.1)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Est)</td>
<td>$E^c_o$</td>
<td>78.9 (1.4)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Est)</td>
<td>$E^c_o + E^s_o + E^l_o$</td>
<td>79.52 (1.2)</td>
</tr>
</tbody>
</table>
We are trying to predict whether an object will be mentioned in a description.

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Accuracy% (std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (ImageCLEF)</td>
<td></td>
<td>57.5 (0.2)</td>
</tr>
<tr>
<td>Log Reg (ImageCLEF)</td>
<td>$K^s_o + K^l_o$</td>
<td>60.0 (0.1)</td>
</tr>
<tr>
<td>Log Reg (ImageCLEF)</td>
<td>$K^c_o$</td>
<td>68.0 (0.1)</td>
</tr>
<tr>
<td>Log Reg (ImageCLEF)</td>
<td>$K^c_o + K^s_o + K^l_o$</td>
<td>69.2 (1.4)</td>
</tr>
<tr>
<td>Baseline (UIUC-Kn)</td>
<td></td>
<td>69.7 (1.3)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Kn)</td>
<td>$K^s_o + K^l_o$</td>
<td>69.9 (0.6)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Kn)</td>
<td>$K^c_o$</td>
<td>79.8 (1.4)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Kn)</td>
<td>$K^c_o + K^s_o + K^l_o$</td>
<td>82.0 (0.9)</td>
</tr>
<tr>
<td>Baseline (UIUC-Est)</td>
<td></td>
<td>76.5 (1.0)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Est)</td>
<td>$E^s_o + E^l_o$</td>
<td>76.9 (1.1)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Est)</td>
<td>$E^c_o$</td>
<td>78.9 (1.4)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Est)</td>
<td>$E^c_o + E^s_o + E^l_o$</td>
<td>79.52 (1.2)</td>
</tr>
</tbody>
</table>

$K = \text{Known}$

$E = \text{Estimated}$
We are trying to predict whether an object will be mentioned in a description.

\[ K^s_o = \text{Known size} \]
\[ K^l_o = \text{Known location} \]
\[ K^c_o = \text{Known category} \]
Predicting Scenes in Sentences

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Accuracy% (std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (UTUC-Kn)</td>
<td></td>
<td>86.0 (0.2)</td>
</tr>
<tr>
<td>Log Reg (UTUC-Kn)</td>
<td>$K_s^c + K_s^r$</td>
<td>96.6 (0.2)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Est)</td>
<td>$E_s^d$</td>
<td>87.4 (1.3)</td>
</tr>
</tbody>
</table>

$K_s^c$ = Known scene category
$K_s^r$ = Known user rating of scene
$E_s^d$ = Estimated scene descriptor (classification scores for 26 scene categories)

Accuracy is predicting whether or not a scene will be mentioned
Predicting Attributes in Sentences

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Accuracy% (std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (UIUC-Kn)</td>
<td></td>
<td>96.3 (.01)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Kn)</td>
<td>$K_a^c + K_o^c$</td>
<td>97.0 (.01)</td>
</tr>
<tr>
<td>Log Reg (UIUC-Est)</td>
<td>$E_a^d + E_o^c$</td>
<td>96.7 (.01)</td>
</tr>
</tbody>
</table>

$K_a^c$ = Known attribute category  
$K_o^c$ = Known object category  
$E_a^d$ = Estimated attribute descriptor  
$E_o^c$ = Estimated object detection category

Accuracy is predicting whether or not a specific attribute type will be mentioned.
Conclusions

• The authors’ study examines how compositional, semantic, and contextual factors predict importance in images
  • Importance is defined here as the probability of image content appearing in descriptions of the image

• Strengths:
  • The paper attempts to build predictive models of importance (not just descriptive)
  • Their evaluation breaks down the problem into known / estimated
    ➢ This helps us eliminate poor CV performance as noise from the evaluation
  • The factors are examined and evaluated on a large scale (1k-20k) vs 97 images prior

• Weaknesses:
  • Why not consider object saliency as a metric?
  • Does task affect importance? Vs. Free viewing?
  • How about relative attributes? Could this help over binary attributes?
Discussion
References

- [http://www.scholarpedia.org/article/Visual_salience#Neural_and_computational_mechanisms](http://www.scholarpedia.org/article/Visual_salience#Neural_and_computational_mechanisms)
- [https://www.youtube.com/watch?v=jpIkMBsZdRY](https://www.youtube.com/watch?v=jpIkMBsZdRY)