# Learning Attributes from Human Gaze

Nils Murrugarra-Llerena and Adriana Kovashka
Department of Computer Science - University of Pittsburgh

IEEE 2017 Winter Conference on Applications of Computer Vision

## Introduction

We tackle the problem of improving attribute prediction with human knowledge. We represent human knowledge as a collection of gaze maps. Then, we create binary masks per attribute. Employing this localization information, we outperform six different baselines. Finally, we show two applications of our method.

## Motivation

Considering that attributes are defined by humans, **why not integrate humans more closely in the learning process?** Thus, we learn attributes using human gaze maps.


## What makes our work unique

- We are the first to use gaze for learning attribute models.
- Rationale approach **[1]** ask people to mark relevant regions associated with a category. Capturing these regions using gaze is **faster** than drawing, and also it uses **subconscious** information.
- While deep neural networks achieve great performance on attribute learning without exploiting human spatial support **[2]**, we **orthogonally to DNN** improve attribute accuracy of fc6 features using human gaze maps.
- **[3]** localizes relative attributes without involving humans. They could only find correlated parts without any human meaning. In contrast, we localize **binary attributes** using **humans' intuition**.

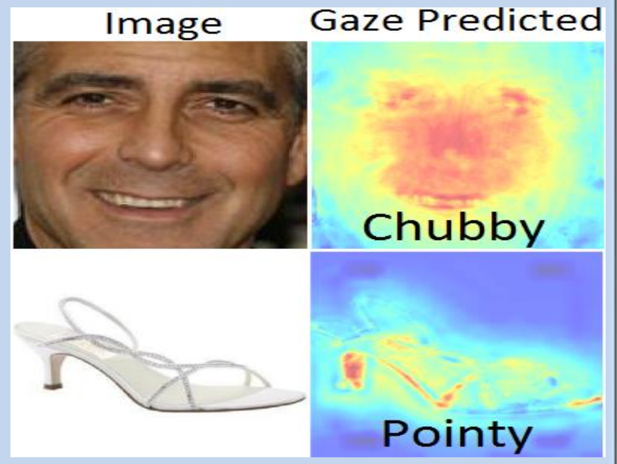
## Gaze dataset

- We employ a **GazePoint GP3 eye-tracker** to collect gaze data from 14 participants.
- To ensure quality, we have **validation images** and we split the whole experiment in **4 sub-sessions**. Between sub-sessions, participants are encouraged to take a break.
- Also, we considered **60 images per attribute** combining positive, negative and borderline annotations.
- Our data collection starts with a **screening phase**. We show ten images to participants and we record their gaze. Participants are required to look at specified regions.
- If their gaze are inside these previous regions, they start the **data collection task**. We show an image and we ask if an attribute is present or not. Then, we record participants' gaze and answers.
- Our dataset can be obtained from: **www.cs.pitt.edu/~nineil/gaze_proj/**

## Approach

- **Generate gaze templates**
  - We merge gaze maps from the same positive attribute labels with a *max* function and normalize them between [0, 1]. Then, we apply a 0.1 threshold obtaining a binary template. Finally, we mask selected cell from a 15x15 grid using our binary template creating ***gt***.
  - To capture **different attribute meanings and separate objects**, we perform **clustering** over positive annotated images before generating templates.
- **Attribute learning using fixed gaze templates**
  - **ST:** We mask train/test images with ***gt***, extract features and evaluate an SVM.
  - **MT:** Similar to ST, we train one classifier per cluster and predict a test image as positive if at least one positive classifier prediction exists.


- **Attribute learning using gaze prediction**
  Instead of using a fixed template, we predict gaze maps on our data using Judd's method **[4]**. We name these approaches as **STP/MTP**.




## Evaluation

We compared our methods: ST, MT, STP and MTP with five different baselines using **F-measure**:

- *Whole Image (WI)*, which extracts features without a mask.
- *Data-Driven (DD)*, which uses a binary mask created from an L1-regularizer over features extracted on a grid.
- *Unsupervised Saliency (US)*, which uses a binary mask from a state-of-the-art saliency predictor (Salicon).
- *Random grid (R)*, which employs a random binary mask from a 15x15 grid.
- *Random Ensemble grid (RE)*, which creates an ensemble of *R*.

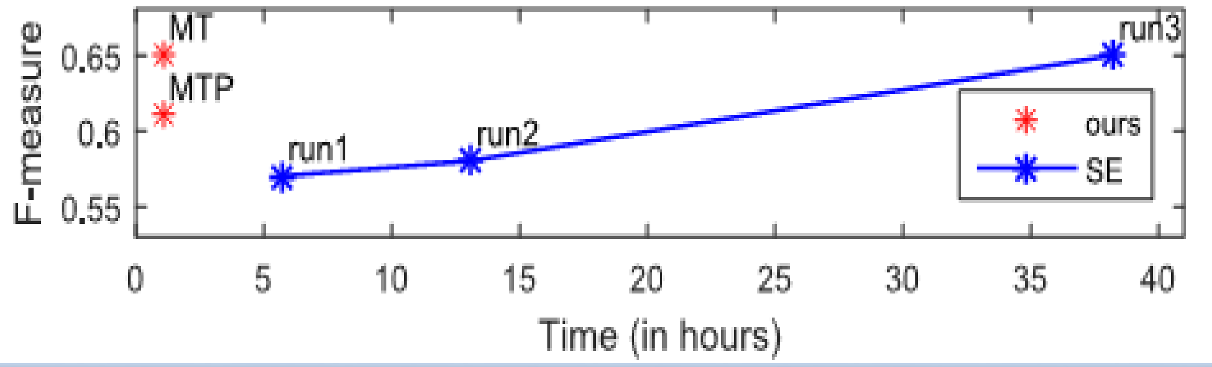
| HOG-GIST features | | | | | | | fc6 features | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| WI | ST | **MT** | DD | US | R | RE | WI | ST | **MT** | US | R | RE |
| 0.52 | 0.52 | **0.57** | 0.48 | 0.51 | 0.51 | 0.50 | 0.51 | 0.49 | **0.54** | 0.48 | 0.46 | 0.47 |

| Dense-SIFT | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| WI | ST | **MT** | STP | **MTP** | DD | US | R | RE |
| 0.52 | 0.53 | **0.57** | 0.49 | 0.53 | 0.45 | 0.45 | 0.49 | 0.50 |

- Our **MT** approach outperforms all baselines. It boosts performance because It captures different meanings of attributes and their possible locations.

## Comparison with Spatial Extent approach [3]

Our approach runs **faster** that Spatial Extent (SE) approach achieving **similar performance**. We tried three different parameter configurations.

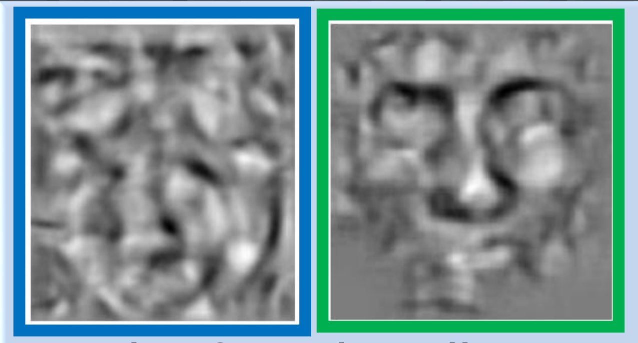



## Adaptation for scene attributes

We associate relevant objects with attributes using human gaze and an R-CNN deep neural network [5].

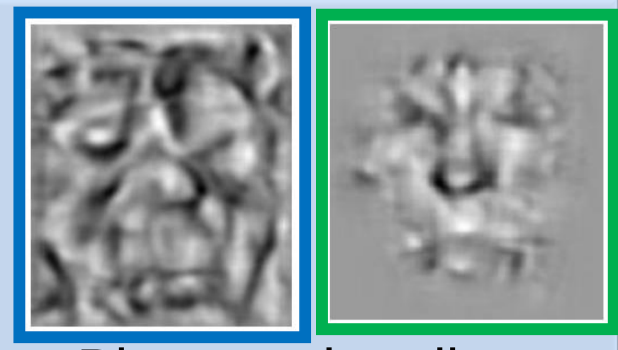
| Attribute | Objects | Attribute | Objects |
|---|---|---|---|
| climbing | mountain, sky, tree, trees, building | sunny | sky, tree, building, grass, trees |
| open area | sky, trees, grass, road, tree | driving | sky, road, tree, trees, building |

## Visualizing attribute models



Baby-faced attribute

Gaze-based templates (**ST**) produces more meaningful attribute visualizations compared to the whole image (**WI**) approach.



Big-nosed attribute

## Finding schools of thought

Researchers [6] find that users perceive attributes differently. A usual approach factorizes an (annotator, image) binary table. We enhance this table clustering gaze maps on positive and negative annotations separately.

| Original | Gaze-based |
|---|---|
| 0.37 | **0.40** |

**F-measure**

**References**
[1] J. Donahue and K. Grauman. Annotator rationales for visual recognition. In *ICCV*, 2011.
[2] S. Shankar et al. Deep-carving: Discovering visual attributes by carving deep neural nets. In *CVPR*, 2015
[3] F. Xiao and Y. J. Lee. Discovering the spatial extent of relative attributes. In *ICCV*, 2015.
[4] T. Judd et al. Learning to predict where humans look. In *ICCV*, 2009.
[5] R. Girshick. Fast R-CNN In *ICCV*. 2015.
[6] A. Kovashka and K. Grauman. Discovering attribute shades of meaning with the crowd. *IJCV*, 114(1):56–73, 2015.