Where are they looking?

Adria Recasens*, Aditya Khosla*, Carl Vondrick, Antonio Torralba

Presented by: Surbhi Goel
Where are they looking?

Follow the gaze of the person and identify the object being looked at
Demo: http://gazefollow.csail.mit.edu/demo.html
Experiments

● Dataset Visualizations
  ○ Images in the Dataset
  ○ Head Locations
  ○ Gaze Locations/Length

● Model Experiments
  ○ Qualitative Evaluation
  ○ Visualizing Gaze Mask and Saliency Map
  ○ Animal Gaze Following
  ○ Extending to Short Video
Dataset Visualizations
Training Set Images
Training Set Images
Heatmaps for Head Location

Train

Test
Heatmaps for Gaze Location

Train

Test
Heatmaps for Relative Gaze Location

Train

Test
Histogram for Length of Gaze

Train

Test
Observations

- Head/Gaze are concentrated for train and scattered for test
- Relative gaze is concentrated for both
- Gaze length relatively short (0.2 peak)
Model Evaluation
Good Cases
Good Cases
Bad Cases

Head fully tilted but missed
Bad Cases

Face forward but eyes tilted
No object of attention
Bad Cases

Back facing
Observations

- Handle groups well
- Gaze location is very accurate, head location often not
- Unable to capture eye movement independent of face orientation
- Fails at a lot of back facing cases
Gaze Mask and Saliency Map
Gaze Mask and Saliency Map

- Gaze Mask incorporates the general direction of gaze
- Saliency Map incorporates the salient objects in image
- Element-wise product captures locations that satisfy both
Gaze Mask and Saliency Map

Image with Gaze

Gaze Mask

Saliency Map
Animal Gaze Follow
Animal Gaze Follow
Animal Gaze Follow

Works (almost) for even birds
Animal Gaze Follow

Works even when more than one salient object
Animal Gaze Follow

- Model generalizes to animals
  - Initialized with ImageNet which has animal data

- Able to learn properties based on orientation of head

- Point of gaze is not always correct
Extension to a Short Video

Apply model per frame of video
Extension to a Short Video

Head detector often fails, could use temporal context to improve
Conclusions

- Can be confused with mixed orientations and back-facing
- Model generalizes well to animals
- Could be potentially extended to videos
- Could be applied to other domains?
Thank You!