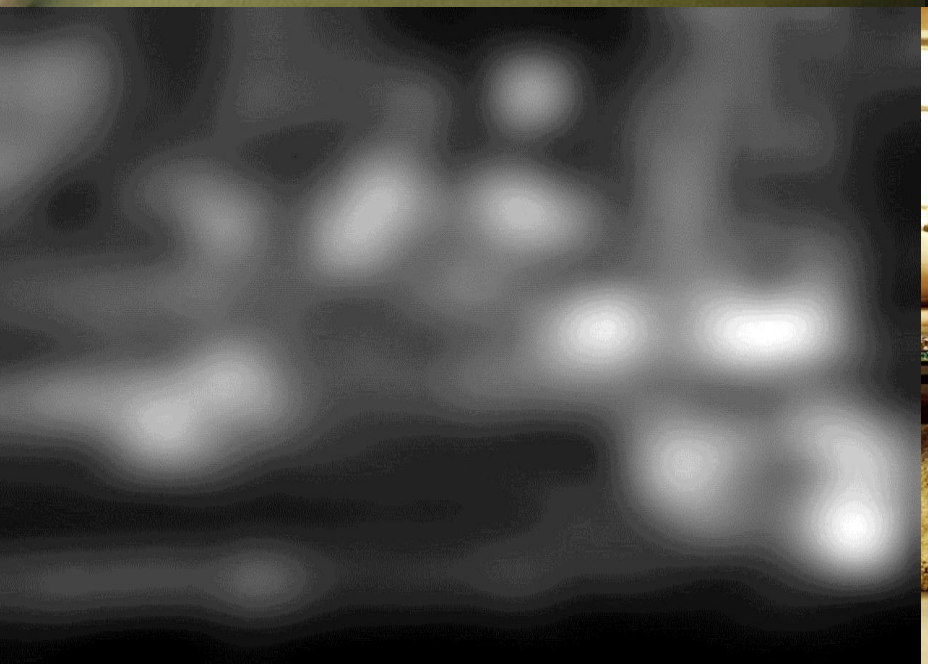
Learning to Predict Gaze in Egocentric Videos

Yin Li, Alireza Fathi, James M. Rehg

Outline:

- What is visual saliency (through Itti Koch & Torralba methods)
- Gaze distributions for GTEA Gaze + dataset (per user and per task)
- Local image excerpts around gaze points for different action labels
- Correlation between head motion and gaze
- Use of future frames to predict gaze in current frame
- Discussion and Conclusion



Itti & Koch Method

Original Image

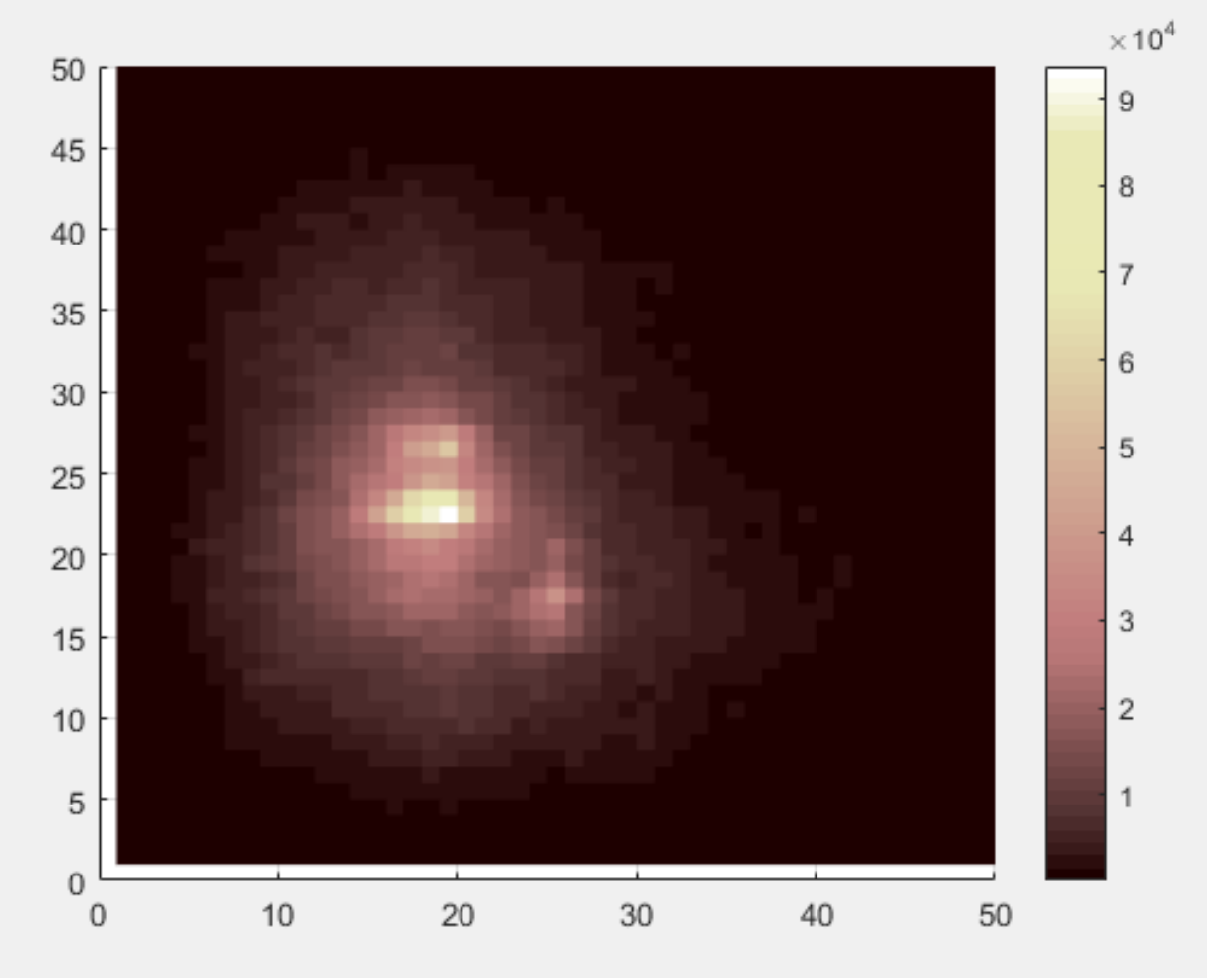
SaliencyMap

Original Image

SaliencyMap

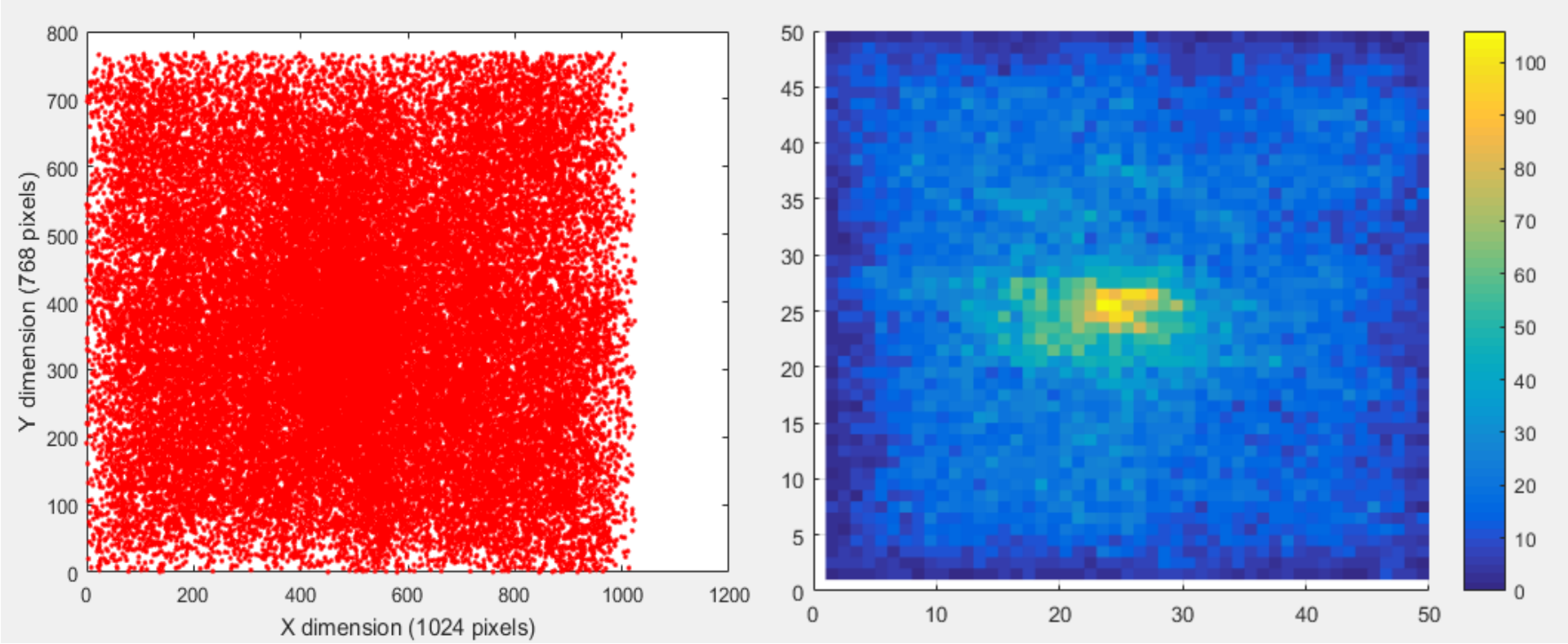
Torralba Saliency Model Prediction

Images vs Egocentric data



MIT Dataset

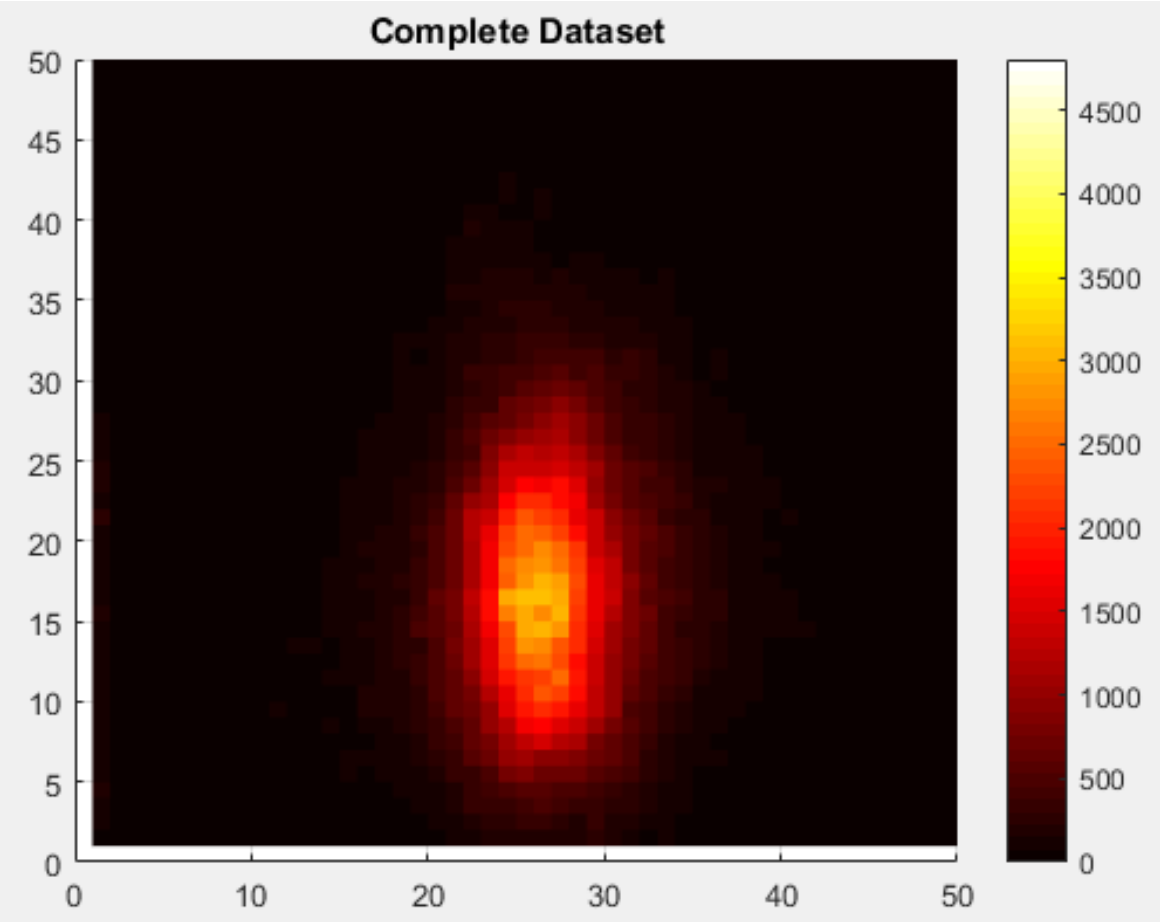
- 1003 images
- 15 subjects
- The figure is a histogram plot of all the eye gaze data, captured using eye tracker glasses
- From the figure the center bias can be observed
- This bias is attributed to the setup of experiments where the users are placed centrally in front of screen and to the fact that human photographers tend to place objects of interest in the center of the photograph.



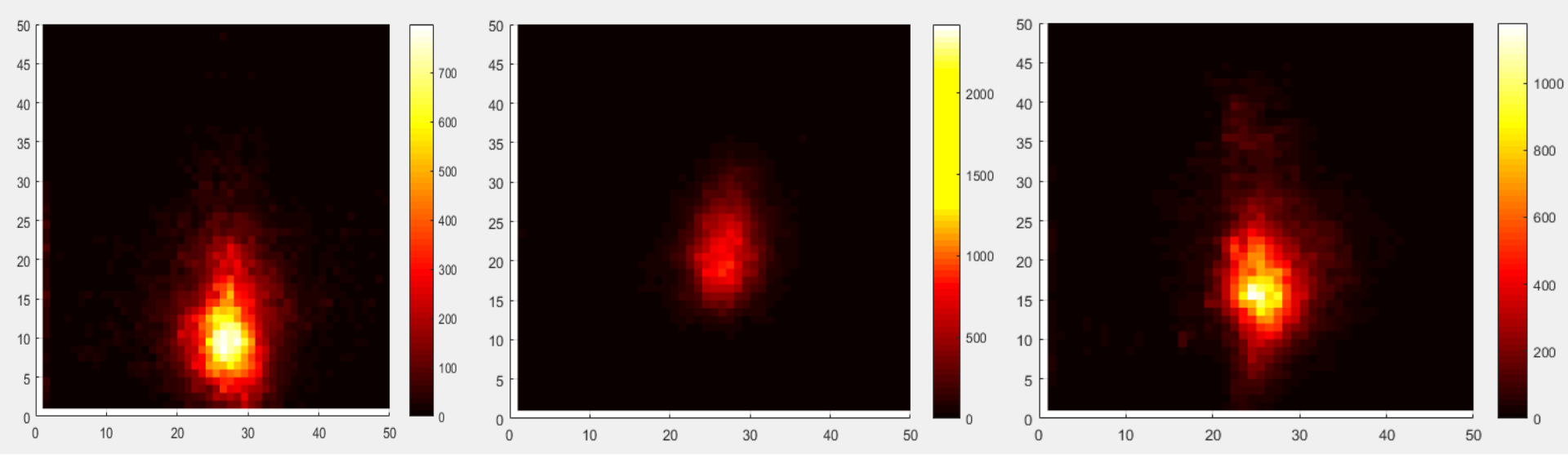
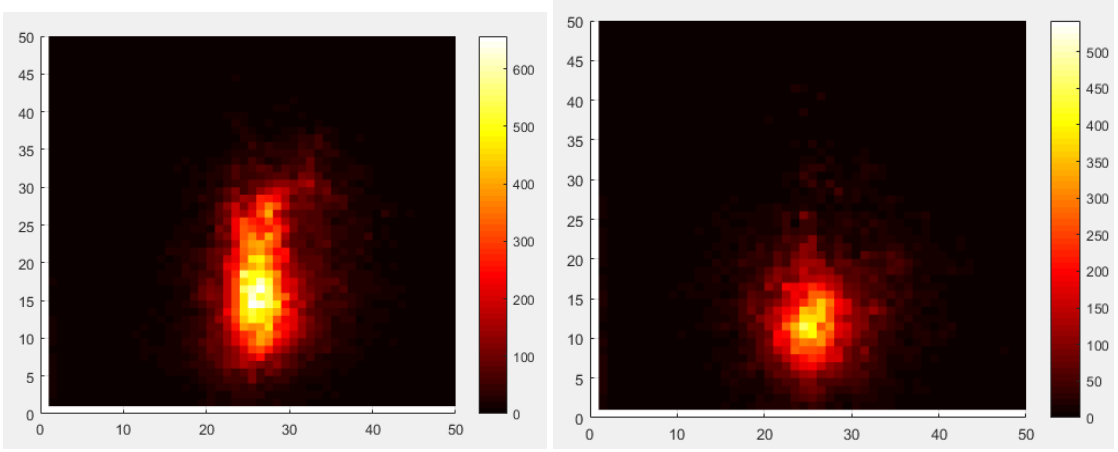
KTH Koostra Dataset :

- 99 images (1024x768 pixels)
- 31 subjects
- Original gaze points plotted on left, histogram of (x,y) positions on the right
- The central bias is again clearly visible

Gaze Distribution for GTEA Gaze+ dataset

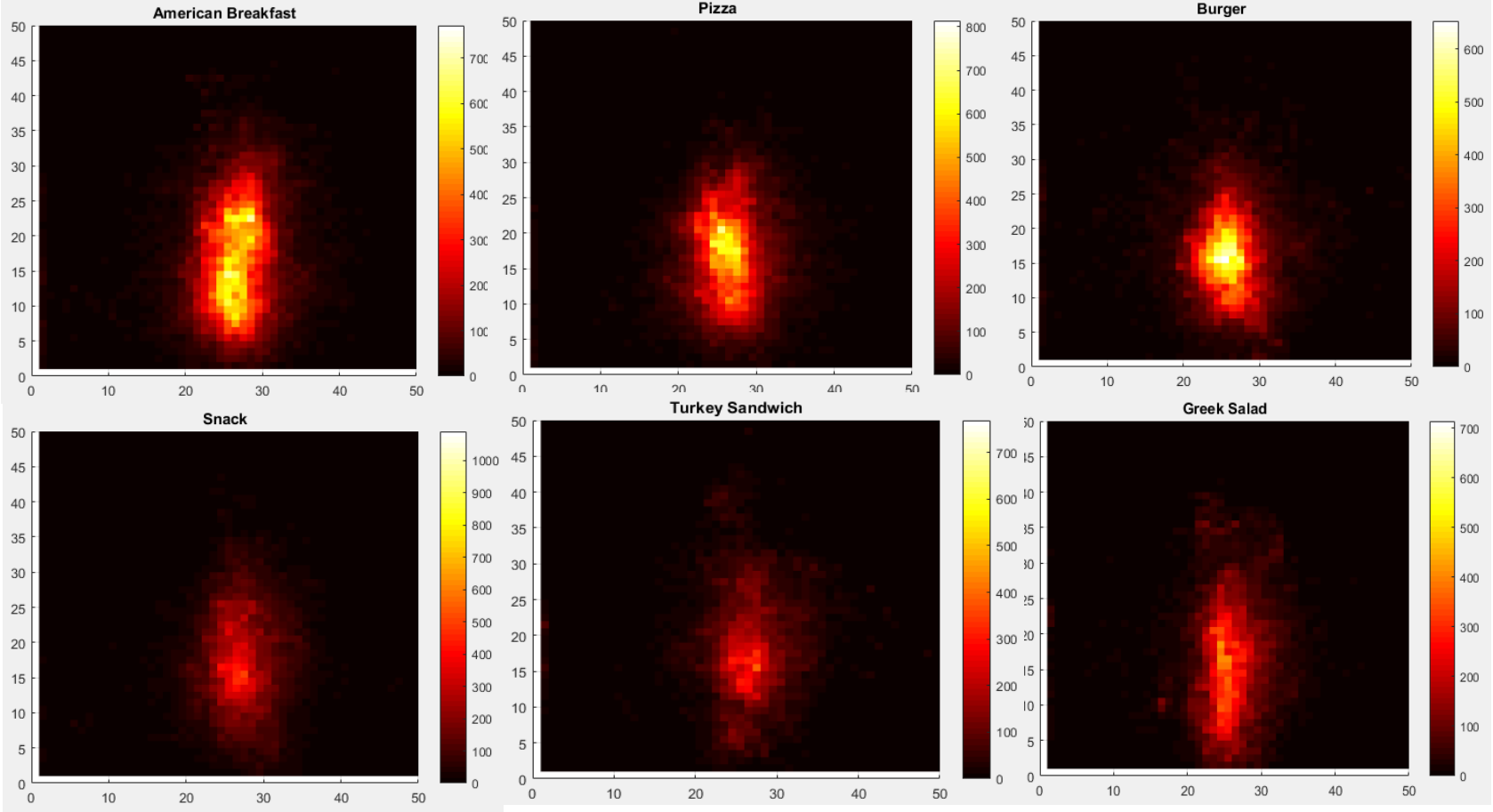


There is less variance and mostly concentrated in the bottom half due to the task at hand (meal preparation)

GTEA Gaze+ dataset :

- 7 videos
- 5 subjects
- Above are histograms of the gaze locations for each subjects



Above are histograms of the gaze locations for each meal preparation task.

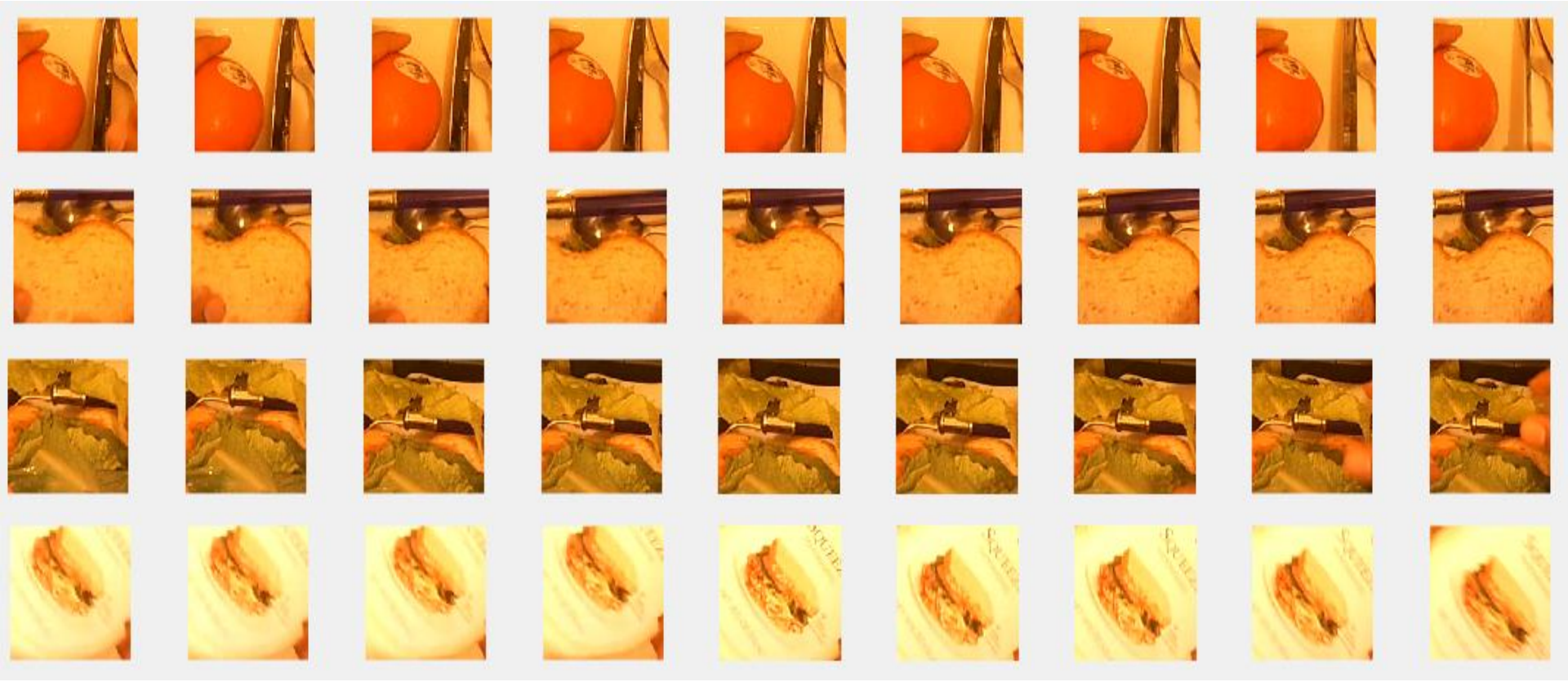
From the plots on this slide and previous slide we can say –

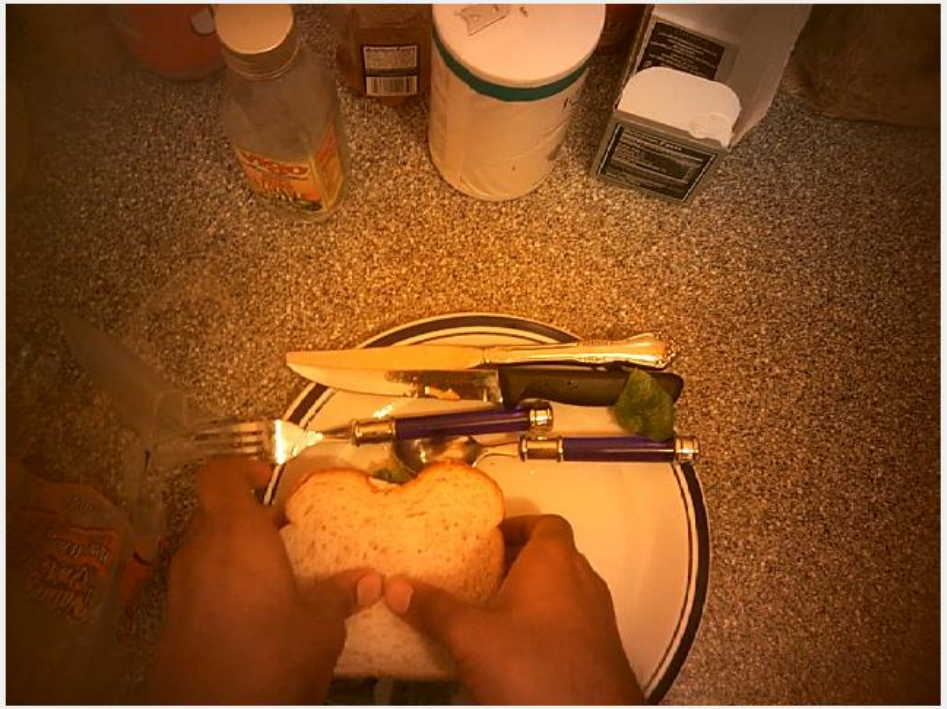
- Each subject has a low variance gaze distribution (though the bias varies).
- This difference in bias for each subject shows up as higher variance for each task in the above distributions.

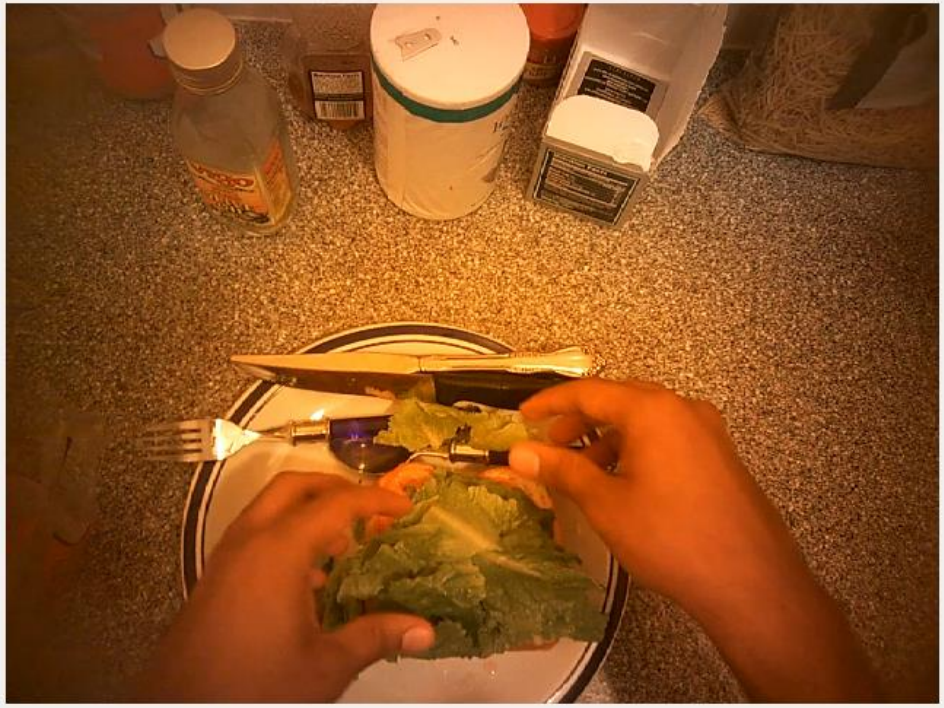
Region around the Gaze point

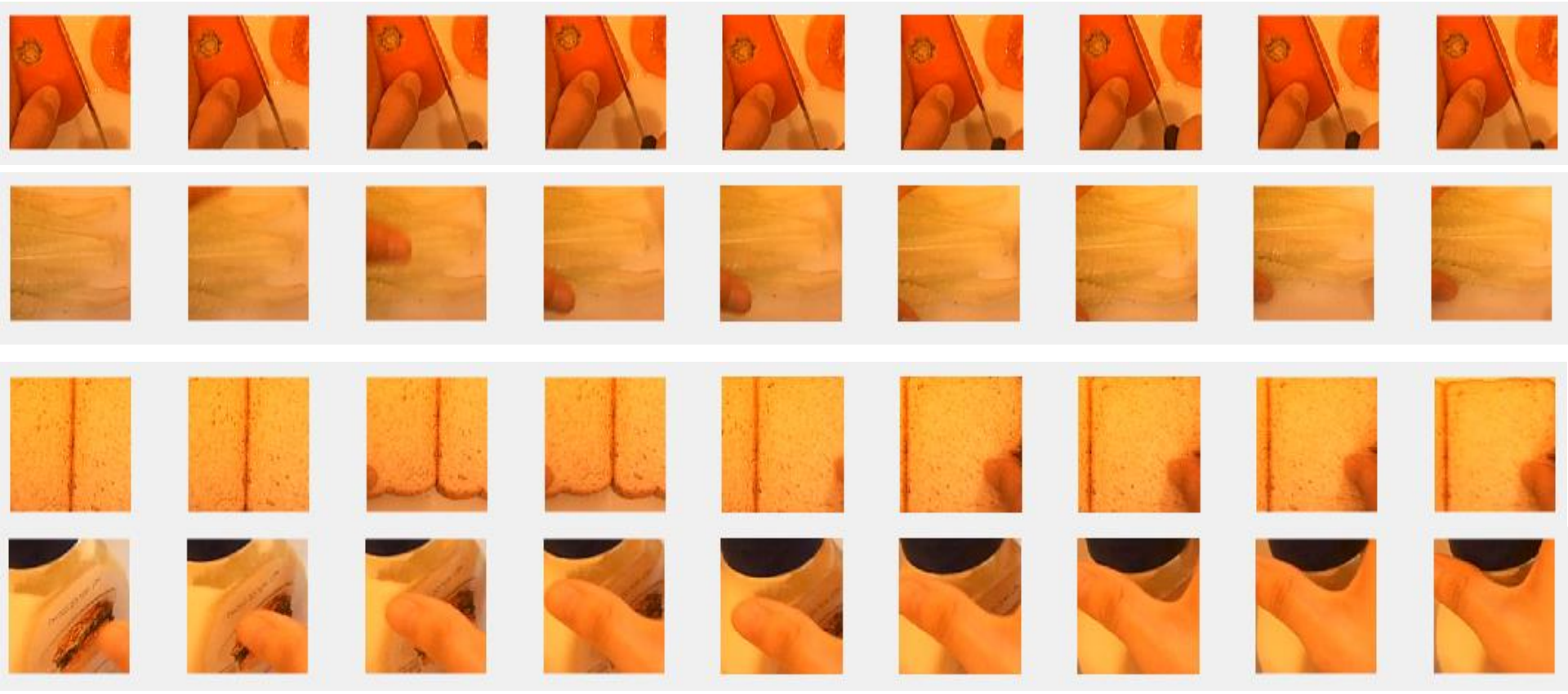
Gaze points are useful if they are reflective of the object being manipulated.

Some examples of a 200x200 region around gaze point (for a given action label, across frames)clearly depict their usefulness.

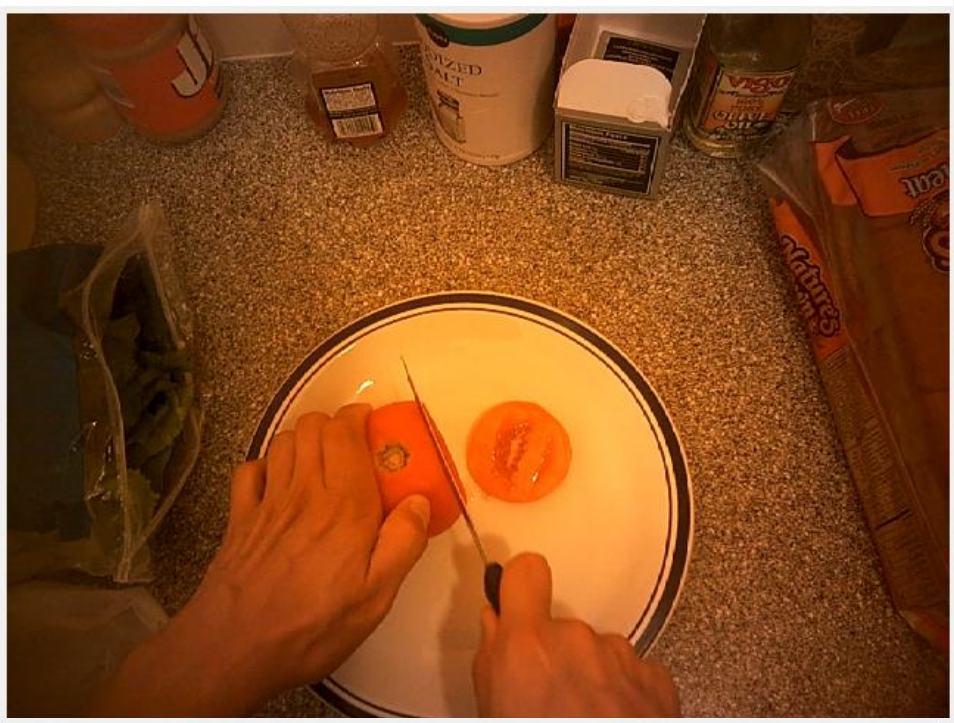






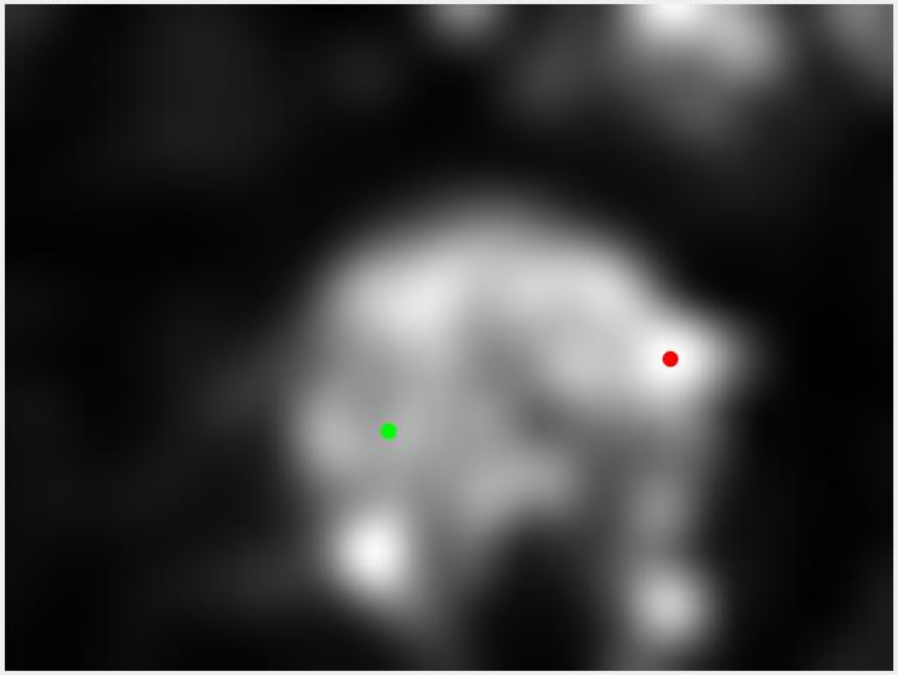


*another subject, same action labels

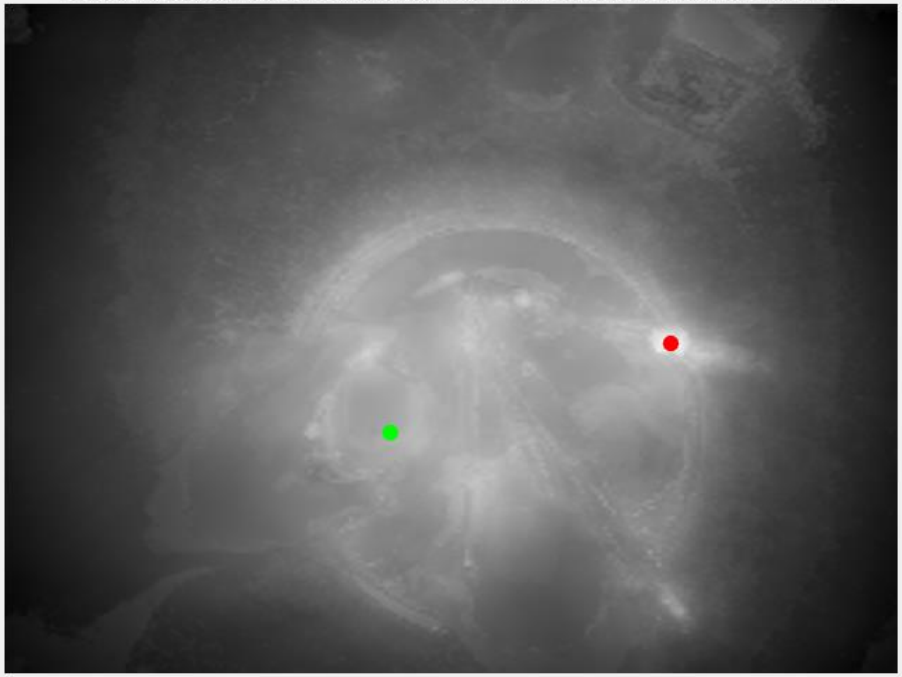


Frame

Itti & Koch prediction (green-original gaze point, red-predicted most salient point)

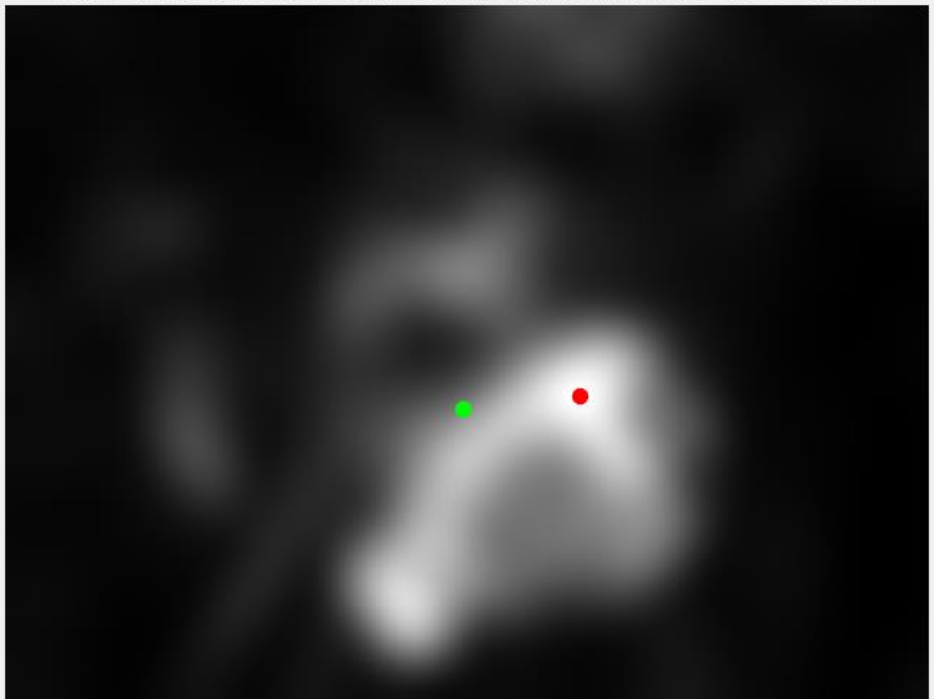


Torralba prediction (green-original gaze point, red-predicted most salient point)

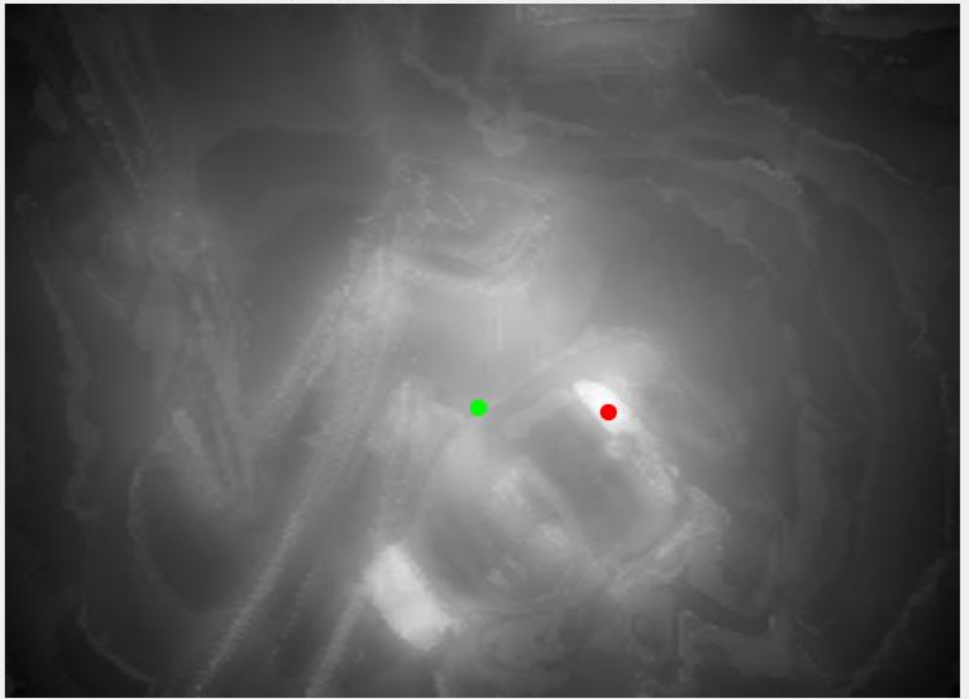


Frame

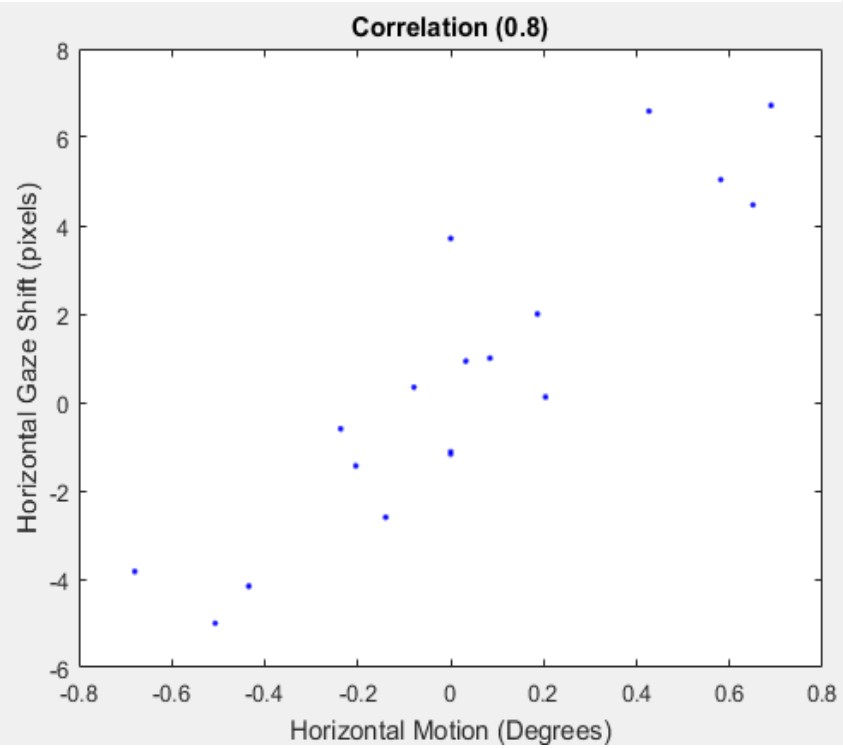
Itti & Koch prediction (green-original gaze point, red-predicted most salient point)

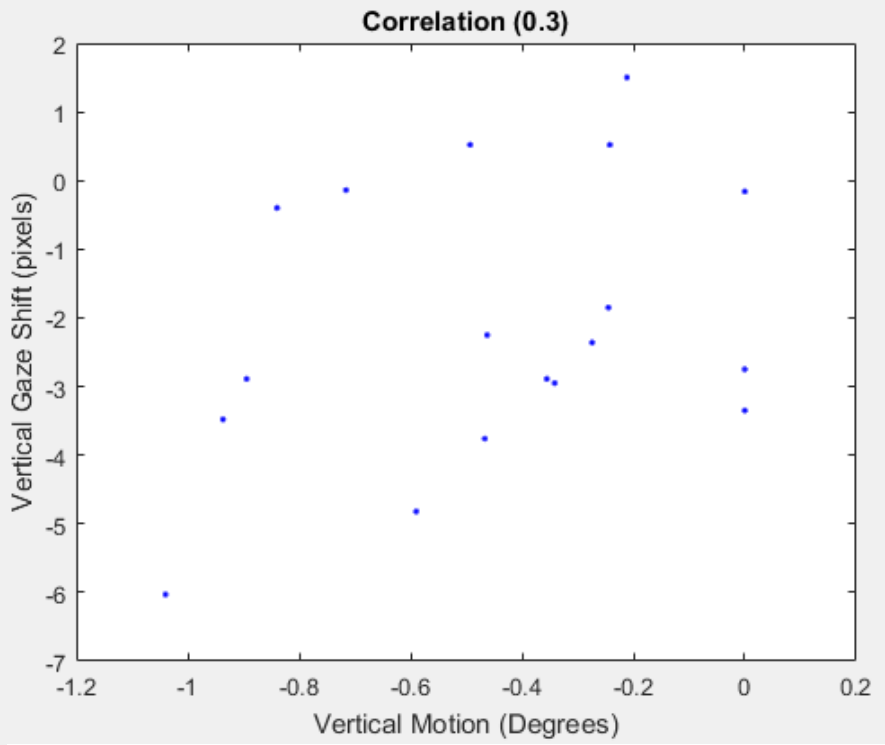


Torralba prediction (green-original gaze point, red-predicted most salient point)



Correlation of Head Motion with Eye Gaze





Similar characteristic as reported in the paper are obtained for small sections of the videos (20-30 frames)

But for higher number of frames no correlation is obtained.

Conclusion: head motion tracks gaze shifts for small durations of time and thus does not perform well for gaze prediction task (as noted in the paper).

Gaze ahead of hands

(a) Gaze in f

(b) FG of f

(c) FG of $f+t$ to f

(d) FG of $f+t$

(e) Gaze in f

(f) FG of f

(g) FG of $f+t$ to f

(h) FG of $f+t$

A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In ECCV, pages 314–327, 2012.


FLANDERS
16 Quarter Pound
Beef
Patties
Family Pack
Keep Frozen

Discussion & Conclusion

- Head motion and Hand position are effective descriptors for Gaze Prediction
- Gaze prediction helps in localizing the area in a frame where activity (object manipulation) is occurring
- Limited data is available in the realms of Egocentric videos with Gaze (A new dataset is EgoSum + with more diverse activities)
- The method is considerably dependent on the presence of hands in the frame (which is a limiting factor for a widespread application.
- Accuracy could possibly be further increased by incorporating the method of predicting gaze in current frame via using future frame descriptors (head motion, hand position, etc.). Better gaze predictions would lead to higher activity recognition accuracy

References

J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency", Proceedings of Neural Information Processing Systems (NIPS), 2006

J. Harel, A Saliency Implementation in MATLAB: http://www.klab.caltech.edu/~harel/share/gbvs.php

Learning to Predict Where Humans Look
Tilke Judd and Krista Ehinger and Fredo Durand and Antonio Torralba
IEEE International Conference on Computer Vision (ICCV)
2009
http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html

Kootstra, G., de Boer, B., and Schomaker L.R.B. (2011) Predicting Eye Fixations on Complex Visual Stimuli using Local Symmetry. *Cognitive Computation*, 3(1):223-240.

M. F. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? Vision Research, 41:3559 – 3565,2001

T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. IEEE TPAMI, 33(3):500–513, 2011

Some recent work

- Gaze-enabled Egocentric Video Summarization via Constrained Submodular Maximization, Jia Xuy, Lopamudra Mukherjeex, Yin Liz, Jamieson Warnery, James M. Rehgz, Vikas Singh, CVPR 2015