Where are they really looking?

Ambika Verma Brady Zhou
University of Texas at Austin

Introduction

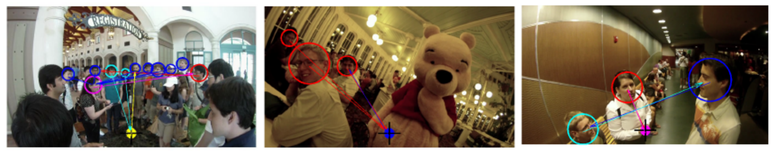
The GazeFollow (1) method achieves the goal of predicting where people are looking in images by making singular predictions using their CNN based model. We propose to extend this work by accounting for the following scenarios (2) –

- Head orientation plays a major role in predicting gaze direction
- More likely for a person to look at another person in the image
- If others are looking in the same direction, then it is more probable for someone to be looking at it as well.

Thus, we build an MRF model to utilize social attention and interaction factors.

Related Work

(2) Uses a MRF model to preform gaze following to predict the type of interaction between people. It is heavily dependent on First Person Vision (FPV) videos, 3D head locations and requires multiple people in the scene. Though, the method can predict common attention even if people are not looking at a person.



(3) Builds upon a similar idea as (1) by combining low-level saliency and head pose information without using deep learning methods and depends heavily on extensive annotations of images.

Figure 1: Social Attention examples from (2)

Experimental Setup

Dataset:

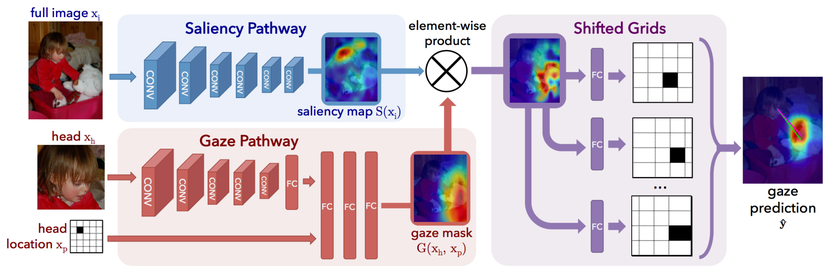
GazeFollow (1) consists of 120,000 train and 3000 test images. These images are labeled with ground truth gazes for every head in the image.

We filter out this dataset to a subset with the properties that each image contains more than one person. This comes out to approximately 3500 train and 1000 test images.

Results are evaluated on the test images based on L2 distance (normalized) and angular error (degrees) between ground truth and predicted gaze points with respect to the head position.

Our Approach

Our MRF model closely follows the work of (2). We assume that head positions and orientations are given, similar to (1), which can be obtained by using a head detector (4). The image is discretized into cells, to reduce model complexity. Unary potentials are used to learn the likelihood of looking in a particular direction based on head orientation and position and bias to look at other faces. Pairwise potential capture the interaction between people in a scene.



Where Are They Looking? CNN Architecture

$$\phi_U = \phi_1 . \phi_2 . \phi_3$$

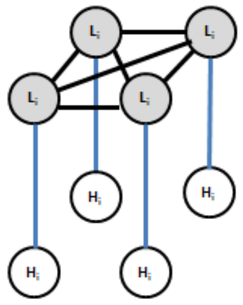
$$\phi_1 = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left|H_{oi} - (l - H_{pi})\right|^2}{2\sigma^2}\right\}$$

$$\phi_2 = \frac{1}{1 + \exp\{-c.||l - H_i||\}}$$

$$\phi_3 = \begin{cases} p & l = H_i \\ 1 & otherwise \end{cases}$$

$$\phi_P = \begin{cases} c_P & l_1 = l_2 \\ 1 - c_P & otherwise \end{cases}$$

Potential Functions



Markov Random Field

H_i are head position and orientation vectors. L is the gaze point on the discretized grid. ϕ_1 is a Gaussian function that penalizes gazes that do not agree with the head orientation. ϕ_2 penalizes gazes that are are too close to the head. ϕ_3 is a function that is grants a bias towards gazes that are looking at other faces in the scene, and the pairwise potential ϕ_p represents that multiple people in a scene tend to look towards the same object. Constants c, p, c_p and σ are learned through cross validation on the subset of GazeFollow with multiple subjects in the image. Graph inference is performed through alpha expansion, a method for energy minimization across graph cuts. We initialize the prediction labels with the gaze predictions obtained from the GazeFollow CNN (1).

Results

	L2 Distance	Angular Error
Random	0.452	67.12°
MRF-Chance	0.310	32.62°
MRF-Unary	0.278	28.43°
CNN	**0.199**	**26.50°**
Ours	0.266	26.82°

Table 1: Quantitative Results. L2 distance is measured on normalized image coordinates [0, 1].

Figure 2: Qualitative Results.
CNN predictions plotted in cyan, ours plotted in green, ground truth in red.
The top row shows success cases, bottom row shows failure cases.

References

1. Where are they looking? Khosla, Recasens, Vondrick, Torralba. NIPS 2015.
2. A.Fathi, J. Hodgins and J. Rehg, “Social Interactions: A First-Person Perspective”, CVPR 2012
3. D. Parks, A. Borji, L. Itti, “Augmented saliency model using automatic 3D head pose detection and learnedgaze following in natural scenes”, Vision Research, 2014.
4. "Here's looking at you, kid." Detecting people looking at each other in videos. Proceedings of the BritishMachine Vision Conference, 2011