Where Are They Really Looking?

Ambika Verma   Brady Zhou

University of Texas at Austin
{ambika,brady.zhou}@utexas.edu

December 2, 2016

Abstract

In this paper, we present a framework to predict where actors in an image are looking by utilizing joint attention and/or interaction characteristics. This approach is specifically useful in case of more than one person/actor being present in a scene. We use a combination of CNN [1] and MRF [2] models to accomplish this task. The approach is motivated from recent state of the art work in Gaze Following (CNN) and Social Attention (MRF) domains.

1 Introduction

In our day to day life we tend to interact with people and/or objects in a fairly predictable manner specifically in terms of social attention and/or interaction such as we tend to notice things that a group of people might be attending to. Additionally our gaze (direction we look in) is a strong indication of what we might do next. Thus, it is imperative for a computer or robot trying to predict the type of actions or interactions occurring in an image or video, to follow gaze of actors in a scene effectively. As a result, gaze following can be used for different high level tasks such as predicting visual saliency, activity recognition, active perception and behavioural analysis.

As noted earlier we solely concentrate on gaze following task for multiple actors in an image or video, with the idea in mind that joint gaze prediction can benefit the task of gaze following since humans interact in a predictive manner. This leads to some of the assumptions made in this paper (which are derived from social attention domain) [2], namely head orientation is a robust indicator of gaze direction and if another actor in the scene is looking at a particular location it is more likely for another person (with head orientation in the same direction) to looking at it as well.

1.1 Related Work

As noted by [2] there are only a few works which build upon the task of gaze following. There are three different scenarios where this task has been studied - Free-viewing Saliency, Social Attention/Interaction and true Gaze Following, we review each task briefly in this section.

Identifying type of Social Attention/Interaction occurring in a scene inherently limits itself to scenes with multiple people only [2][][][]. Though, it boasts of intuitive methods to perform joint gaze prediction. [2] develops MRF and CRF methods for their task of predicting type of interaction taking place in a scene, which can be applied to long first-person videos to filter out useful clips or subsets (similar to video summarization). This work utilizes an MRF model to predict where people are looking in a scene to later classify the type of interaction taking place using CRF. Thus, gaze following is utilized in such techniques in an implicit manner. The model developed utilizes joint predictions effectively, though is heavily biased to look at other faces in the scene. This bias is acceptable for identifying the type of interaction amongst people but does not generalize well for the gaze following task, wherein actors can be looking at another actor or an object.

On the other hand works in free-viewing saliency and gaze following tasks do include scenes with both multiple people and only one person. [2] is a CNN based model which is specifically tasked for gaze following utilizing head position and orientation along with saliency to accomplish the task. This method has several advantages such as it is robust in case of single actor being present in a scene, does not suffer from any evident bias of looking at people only (such as in [1]) and the network architecture does not require extensive hand-crafted functions (such as those required in probabilistic models such as MRF or CRF). Though, these merits are far reaching the model uses a naive approach in cases where multiple actors are present in the scene, it resorts to providing predictions for each actor individually.

Another relevant work which utilizes gaze following is [3] in which an actors gaze is used as an additional feature to generate more accurate free-viewing saliency maps. It has been noted in numerous works [],[],[],[] that the gaze of an actor in a scene influences where passive observers look during free-viewing. In this work the authors use head pose [] along with low level saliency [] to predict more accurately the semantically salient regions in a scene. This work does utilize joint predictions in case of multiple actors in the scene and it is worthy to note that considerable performance improvement is achieved over state of the art saliency methods []. Even though this method can be extended to gaze following task, it has not been evaluated on the same.

Works such as [1] and [3] utilize gaze following and joint predictions to achieve a secondary task. Thus, our approach is to combine the works of [2] and [1] in a constructive manner to add joint prediction capability to the model developed in [2] with the goal of achieving higher performance for cases with multiple actors in the scene.

2 Technical Approach

As in [1] we also assume that the head positions are provided and as demonstrated in [3] a head detector with pose estimation can be effectively used with a few modifications.

We plan to use the CNN network from [1] with post-processing added through an MRF model which is designed on the lines of model used in [2] to account for joint gaze predictions. The overall architecture of our approach is illustrated in Fig.

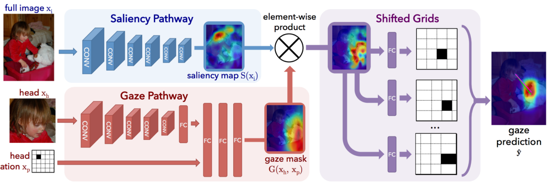

Figure 1: The network architecture from [1].

A given image along with head position is provided as input to the CNN. We only use images where there are multiple actors in the scene. The CNN then iterates through the image for each actor and gives as an output a two-dimensional position say (gx,gy) normalized by the image dimensions to lie between 0 and 1. This prediction is then used as an initial seed for MRF model. In addition the MRF model is also supplied with head location and orientation vectors, where orientations is obtained using the head position along with predicted gaze location and is jittered to replicate errors one would observe using a head detector. The MRF model then optimizes the unary and pairwise potentials to jointly refine the CNN predictions based on the predictions obtained for other actors in the scene.

Following sections discuss the CNN and MRF models in detail.

2.1 GazeFollow CNN Model

The CNN model consists of two pathways, namely Gaze pathway and Saliency pathway. As noted in [1], the gaze pathway tends to learn predicting the direction of gaze based on the head orientation (this is the reason why we use the CNN prediction and head position to generate orientation vectors for the MRF model) and the saliency pathway tends to learn to find salient regions in the scene. The model is implemented in Caffe [] and the two pathways are equivalent to the first five layers of AlexNet []. The saliency pathway is initialized with weights from Places-CNN [] and gaze pathway is initialized with ImageNet-CNN []. The saliency and gaze masks obtained are then combined through an element-wise layer and final prediction is made through shifted grids. More details can be found in [1].

2.2 Social Interactions MRF Model

Markov random fields are a form of probabilistic graph models that have been heavily used throughout the computer vision field to solve classification problems. Markov random fields have proved to be effective in certain applications like image segmentation and parsing text (NLP). The goal of using an MRF is find a labeling $\{p_1 = l_1,\ p_2 = l_2,\ \dots,\ p_n = l_n\}$ such that the joint probability $P(p_1 = l_1,\ p_2 = l_2,\ \dots,\ p_n = l_n)$ is maximized.

To perform energy minimization for the MRF, direct inference is intractable. The number of possible labelings for n people using and a grid size of 25×25 is $(25 \times 25)^n$. The complexity grows exponentially with the number of people with a high constant and a direct solver is not possible to use. We use a graph cut algorithm [9] for approximate energy minimization that requires an initial labeling for each person. The method, alpha expansion, then iterate upon this initial labeling until convergence.

We use the graphical model from [2] to refine the gaze predictions from [1]. The MRF model is used to utilize the relationship between the multiple actors in a scene. The motivation for this approach stems for example from scenes where multiple actors are looking at the same location. In case of individual predictions there is no way for benefitting from the information we have about the gaze of other actors in the scene, whereas joint prediction allows us to give higher weights to locations which lie in the direction of gaze of an actor and are also looked at by other actors in the scene.

The MRF model implementation uses the following unary and pairwise potentials. The unary potentials allows the model to predict the direction of the gaze based on head position and orientation vectors and on the other hand the pairwise potentials capture the interaction between actors in the scene.

$$\phi_u = \phi_1 \cdot \phi_2 \cdot \phi_3 \tag{1}$$

The unary potential is a product over three different potential functions. Each of these potentials seeks to score the probability of a gaze (higher is better).

$$\phi_1(l_i = l, d_i, f_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\big\{ - \frac{\|d_i - (l - f_i)\|^2}{2\sigma^2} \big\} \tag{2}$$

The first unary term is a guassian distribution of how well the predicted gaze aligns with the head orientation. This scores the labeling of a person i's gaze to a location l. l is a label, the cell of the predicted gaze. d_i is a unit vector, the head orientation. f_i is the position of the subject's face. σ is a constant.

$$\phi_2(l_i = l, f_i) = \frac{1}{1 + \exp - c_t \cdot \|l - f_i\|} \tag{3}$$

The second unary term is thresholding function that assigns the score to 0 if the predicted gaze is too close to a person's face. This is used to prevent extremely shot gazes (labeling a person's gaze to their own face). c_t is a constant.

$$\phi_3(l_i = l) = \begin{cases} c_p, & \text{if } l \in \{f_1,\ f_2, \dots,\ f_n\}. \\ 1, & \text{otherwise.} \end{cases} \tag{4}$$

The third unary term is to bias subjects' gazes towards other faces in the scene. c_p is a constant.

The parameters $\sigma,\ c_t,\ c_p$ are learned through training images with more than one actor in the scene through cross validation. Additionally, the image space which is the set of all possible gaze points is discretized into cells, we chose 50×50 empirically for performance reasons.

3 Experiments

3.1 Dataset

GazeFollow dataset [1] is used as part of this work. The dataset contains 122,143 training images and 4,782 test images drawn from varied other datasets such as SUN [4], MS COCO {lin2014microsoft, Actions 40 [5], PASCAL [6], ImageNet challenge [7], and Places dataset [8].

Each image in training set is annotated with ground truth head position and ground truth gaze point.

Each image in test set is annotated with ground truth head position and 10 ground truth gaze points to account for human consistency. We use average of these 10 gaze points in all further experiments.

To assess performance on joint gaze prediction, we test our model on a subset of the GazeFollow dataset that contains multiple subjects.

The CNN model [1] is trained on the GazeFollow dataset. 3500 images with multiple actors in the scene are used to train the MRF model and 1000 images are used for testing purpose.

3.2 Evaluation Metrics

We use two error metrics to quantify the performance of our method and to conduct comparative studies.

L_2 **distance** - The images are normalized to 1x1 and the distance between ground truth and predicted gaze point is taken. For a single image with multiple people, we report the average of the errors, that is, $\frac{1}{n}\sum_{i=1}^{n}\sqrt{(p_x^{(i)}-g_x^{(i)})^2+(p_y^{(i)}-g_y^{(i)})^2}$.

Angular error - The predicted gaze point is connected with the eye position to create a unit vector v, and the ground truth gaze point is connected with the eye position to create a unit vector u. The angular error is defined as $u^T v$, and converted to degrees.

3.3 Baselines

To see how well the MRF approach refines the predictions from [1], we will use the original predictions from [1] as a baseline, as well as the MRF model from [2]. The intuition behind our method is that the initial seeds for the energy minimization algorithm affect the predicted gazes, and a "better" initial prediction will increase performance. We include a few extra baselines. The first is random chance, which is assigning a random label to each subject. Next, another baseline is setting the initial labeling is set to labels that optimizes the unary terms. This is closest to [2]. Our model sets the initial labeling to the prediction acquired from the GazeFollow CNN. We use these different methods to compare and assess just how large of a factor the initial seeds are for energy minimization.

3.4 Experimental Results

We evaluate the performance of the described models on the GazeFollow [1] dataset.

3.5 Qualitative Results

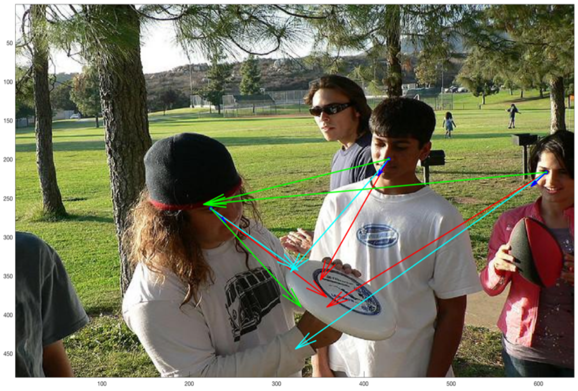

Figure 2: This is a caption

References

[1] A. Recasens*, A. Khosla*, C. Vondrick, and A. Torralba, "Where are they looking?" in *Advances in Neural Information Processing Systems (NIPS)*, 2015, * indicates equal contribution. 1, 2, 3, 4

[2] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1226–1233. 1, 2, 3, 4

[3] D. Parks, A. Borji, and L. Itti, "Augmented saliency model using automatic 3d head pose detection and learned gaze following in natural scenes," *Vision research*, vol. 116, pp. 113–126, 2015. 2

[4] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010, pp. 3485–3492. 3

[5] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1331–1338. 3

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010. 3

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. 3

[8] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495. 3

[9] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001. 3