Prediction and Generation of Gaze in Conversation using Deep learning

Vidya Somashekarappa
PhD Fellow, CLASP
University of Gothenburg
Sweden
vidya.somashekarappa@gu.se

Christine Howes
Senior Lecturer, CLASP
University of Gothenburg
Sweden
christine.howes@gu.se

Asad Sayeed
Associate Professor, CLASP
University of Gothenburg
Sweden
asad.sayeed@gu.se

Abstract

A crucial social characteristic that facilitates human-robot cooperation is gaze following. We propose a new approach to estimate gaze using a neural network architecture, while considering the dynamic patterns of real world gaze behaviour in natural interaction. The main goal is to provide a basis for robot or avatar to communicate with humans using multimodal natural dialogue. We generated 2.4M gaze predictions of various types of gaze in a more natural setting (GHI-Gaze). The predicted and categorised gaze data can be used to automate contextualized robotic gaze-tracking behaviour in interaction. We evaluate the performance on a manually-annotated data set and a publicly available gaze-follow dataset. Compared to previously reported methods our model performs better with the closest angular error to that of a human annotator.

1 Introduction

Numerous research studies on the application of gaze following in HRI have been conducted, including investigations into the efficiency of gaze following for enhancing robot social presence and communication. Currently, robotic gaze systems are reactive in nature but our Gaze-Estimation framework can perform unified gaze detection, gaze-object prediction and object-landmark heatmap in a single scene, which paves the way for a more proactive approach. As future work, we propose an implementable gaze architecture for a social robot from Furhat robotics.

In order to associate visual targets or regions outside of the image scene, Storey et al. (2018), advocated a visual center of attention task. To determine the gaze direction and target in images, semantic information was integrated such as pose estimation, object detection, or facial landmark detection. Additionally, gaze-following has been extended to video analysis, which predicts the gaze target between frames (Jin et al., 2021). In this study, we focus on the gaze-following problem in anticipating the gaze target position exclusively within the image by utilizing pretrained models.

2 Gaze Annotation in Discourse

Manual gaze annotation involves manually marking the gaze direction in each frame of an image or video dataset. This is typically done by a human annotator who looks at the data and marks the gaze direction by placing a dot or other marker on the eye or gaze direction in each frame. Manual gaze annotation is often considered to be more accurate than automatic gaze annotation, but it is also time-consuming and costly (Somashekarappa et al., 2020).

Automatic gaze annotation, on the other hand, involves using algorithms to automatically label the gaze direction in a dataset. This can be done using a variety of methods such as template matching, feature-based methods, or deep learning-based approaches (Wood and Bulling, 2014). Automatic gaze annotation is typically faster and less expensive than manual gaze annotation, but the accuracy of the labels may be lower than with manual annotation.

3 Visual attention and Gaze-Target Prediction

The complete image, the subject's cropped face, and the location of the subject's face that requires attention estimation act as the input to the model. The two stills from the video are enlarged to 224x224 so that the network can see the face with improved resolution.

Two convolutional (conv) pathways make up

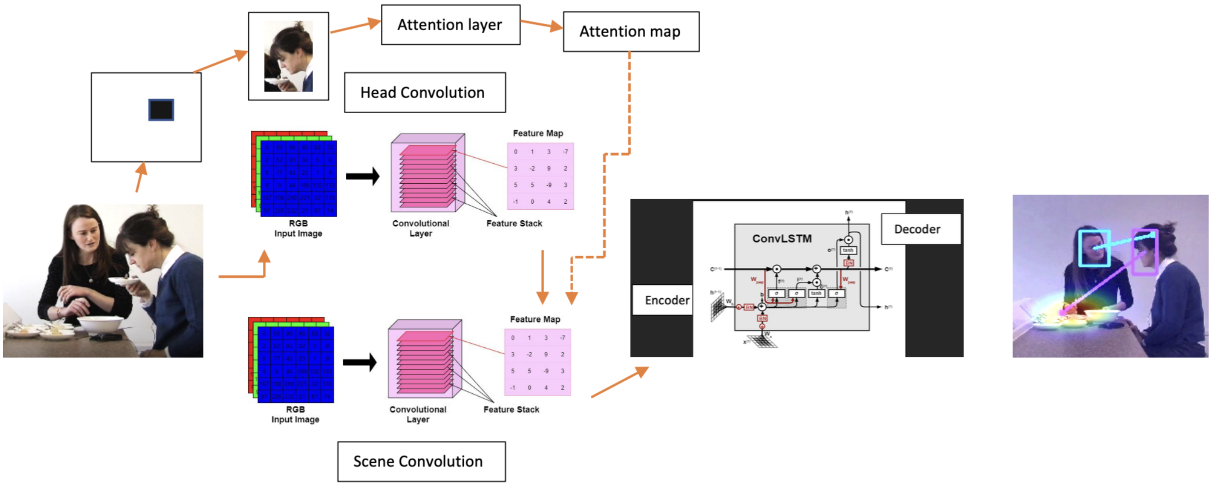

Figure 1: Network Architecture

the model: a scene pathway and a face pathway (see figure 1). The conv pathways employ ResNet 50 (He et al., 2016) as their backbone network and for each of the conv paths, all conv layers of ResNet50 are specifically used. Later three convolutional layers (1x1, 3x3, and 1x1) with ReLu and batch normalization with stride 1 and no padding after each ResNet50 block are added. The filter depths of the convolutional layers are 512, 128, and 1, respectively which results in extracted features' dimensionality being reduced.

In the face pathway, the feature vector computed with the face input image goes through a fully connected layer to predict the gaze angle represented using yaw and pitch intrinsic Euler angles. In the scene pathway, the feature vectors extracted from the whole image as well as from the face image are concatenated with the face position input vector to learn the person-centric heatmap. Similarly to face position, the ground truth used for learning the heatmap is available as a gaze target position in (x,y) coordinates which is quantized into 10 grids in each dimension.

4 Evalutation

In total for the 24 videos, 48 individualistic gaze predictions were generated accounting for close to 80k predictions per video. The heatmaps give us a clear understanding of the region of interest on which the gaze attention occurs. Two experiments were conducted to evaluate the performance of the model. We compare gaze annotation from GHI dataset (coded for various types of gaze) and GazeFallow dataset to our generated images. The analysis shows that the GHI-Gaze, predictions fine tuned with human annotations and attention maps perform better with AUC of 0.924 and the angular error of 13.7° compared to the previous results. The human metrics have the best performance measure with 0.924 AUC and 11° of angular error.

5 Discussion and Applications

The main goal of the paper is to improve the accuracy of gaze estimation and prediction. We propose a novel neural network architecture to simultaneously and accurately detect gaze target on the intended object for multiple people in a single scene. Our results show an improvement in the performance compared to previous methods and provide specific information of the type of gaze in a given scene.

In a customer service scenario, gaze follow behavior can be used to create a more personal and attentive interaction between the human and the robot. By tracking the human's gaze, the robot can direct its attention to the human and show that it is actively listening and engaging with them. In situations, such as in a robot assistant or companion, gaze follow behavior can be used to create a more intuitive and responsive interaction. By following the human's gaze, the robot can respond to their actions and cues in a way that is similar to how a human might respond. In general, gaze follow behavior can play an important role in creating a positive and engaging interaction between a human and a robot, but its specific importance will depend on the goals and context of the interaction.

References

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Tianlei Jin, Zheyuan Lin, Shiqiang Zhu, Wen Wang, and Shunda Hu. 2021. Multi-person gaze-following with numerical coordinate regression. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE.

Vidya Somashekarappa, Christine Howes, and Asad Sayeed. 2020. An annotation approach for social and referential gaze in dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 759–765.

Gary Storey, Ahmed Bouridane, and Richard Jiang. 2018. Integrated deep model for face detection and landmark localization from “in the wild” images. *IEEE Access*, 6:74442–74452.

Erroll Wood and Andreas Bulling. 2014. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the symposium on eye tracking research and applications*, pages 207–210.