Neural Network Implementation of Gaze-Target Prediction for Human-Robot Interaction

Vidya Somashekarappa[1], Asad Sayeed[2] and Christine Howes[3]

***Abstract*— Gaze cues, which initiate an action or behaviour, are necessary for a responsive and intuitive interaction. Using gaze to signal intentions or request an action during conversation is conventional. 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 foundation for robot/avatar to communicate with humans using natural multimodal-dialogue. 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. 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. As future work, we propose an implementable gaze architecture for a social robot from Furhat robotics[1].**

I. INTRODUCTION

A crucial social characteristic that facilitates human-robot cooperation is gaze-following (HRI). Robots that can track a person's gaze are better able to comprehend that person's attention, interest, and intentions. Gaze following enables us to make eye contact which can enhance our social presence and naturalness. To infer human visual attention, it is essential to carefully consider head posture, gaze direction, scene organization, and saliency [1].

The main objectives and contributions of the paper are:

1) Automating robot gaze behaviour using machine learning.
2) Classifying elements of the dialogue based only on gaze behaviours (such as dialogue acts, intimacy regulation and referencing objects)
3) Presenting a dataset containing annotations of attention targets with complex patterns of gaze behaviour and out-of-scene target predictions
4) Developing an implementable-model of gaze in dialogue for a conversational robot/avatar to interpret and produce human-like gaze behaviour

Eye gaze supports and augments other social behaviours such as speech/gesture, and the mental states or cognitive effort can substantially influence gaze. Since speech is a dominant mode of communication in human interactions, it is non-viable to separate gaze from speech in human-robot dialogue. Researchers have shown that gaze improves speech-based interactions such as disambiguating object references, maintaining engagement, conversation and narration, guiding attention, managing partners, and influencing turn-taking [2], [3]. We suggest a non-wearable eye-gaze detection technique that uses videos to study gaze in interaction by making use of the new developments in deep learning.

The majority of the research on gaze following in HRI points to its potential as a useful tool for enhancing robots capacity for social interaction and communication. However, there are still a lot of difficulties and unanswered problems, such as how to design and apply gaze in a reliable way, as well as how to gauge its efficiency in various HRI scenarios. Our method utilizes the manually annotated gaze predictions denoting various types of gaze to train the network that results in a more efficient and precise detection of gaze targets to its corresponding object in space at a given time. The rapid advancements in the field of robotic technologies presses the importance of social robots' prominence in the future, such as robots that are built for interacting with people and are designed for various applications such as therapy and education alongside industry.

II. CORRESPONDING WORK

Recent research has focused on developing algorithms for automatic gaze estimation using various approaches:

Deep Learning-based approaches: Deep learning algorithms, such as convolutional neural networks (CNNs) have been used to predict gaze locations in images or video data. These approaches have shown promising results in terms of accuracy and efficiency [4], [5], [6].

Appearance-based models: Focuses on developing models that take into account the appearance of a person's eyes and face to predict gaze locations. These models have been shown to outperform traditional gaze estimation algorithms that rely solely on image data [7].

Multi-modal fusion: Emphasizes on fusing multiple sources of information, such as eye movements, head movements, and facial expressions, to make more accurate gaze predictions. The multi-modal fusion approaches have been

*This work is supported by the Swedish Research Council

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

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

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

[1]https://furhatrobotics.com/

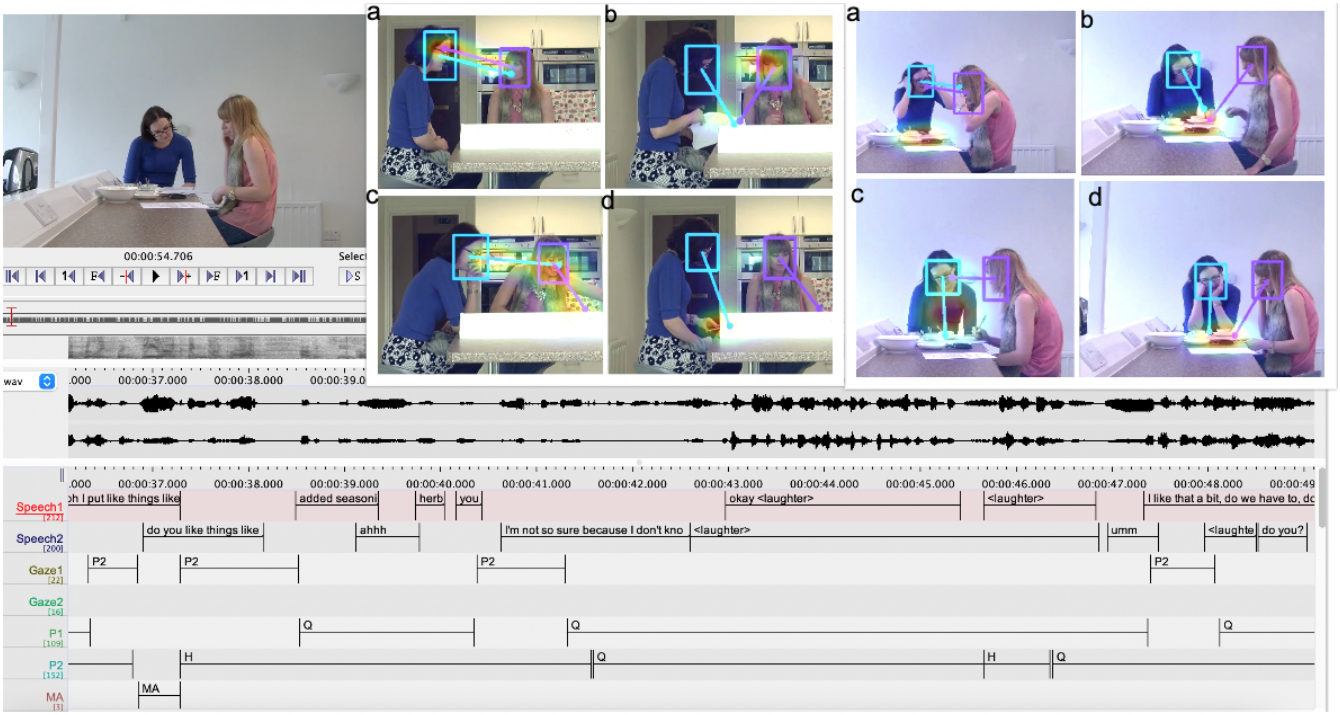

Fig. 1. Various gaze estimates captured in two different angles in a particular scene. a. Mutual attention (looking at each other), b. Joint attention (gaze on hummus and questionnaire), c. Gaze aversion (gaze on partner while the partner looks away), d. Individual gaze (attention on different objects in space)

shown to improve the accuracy of gaze prediction compared to single-modality methods [8]

Gaze correction in virtual reality: Virtual reality (VR) environments introduce new challenges for gaze estimation, as the participant's gaze is not directly visible. The main focus is on developing gaze correction methods that can correct for misalignment between the gaze location and the virtual scene in VR environments [9].

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.

III. Modeling Gaze Behaviour in a Robot

Modeling gaze behavior in a robot often requires a combination of computer vision, and robotic control system, as well as a thorough understanding of human gaze behavior. It can be challenging, particularly when the gaze behavior is complex or subtle.

Some of the most difficult types of gaze to model include: **Mutual gaze/ Direct gaze**, where the robot directly looks at the human, simulating eye contact, because it requires the robot to respond in real-time to the human gaze, while also adjusting its own gaze in a natural and engaging way [11]. Longer mutual attention can be considered eerie and induce uncanny valley effects. **Averted gaze**, where the robot looks away from the human, requires the robot to simulate a lack of attention or interest, which can be difficult to do in a way that is convincing to humans [10]. **Gaze cues**, where the robot uses gaze to initiate an action or behavior, often requires the robot to respond to human gaze in a sophisticated and context-sensitive way [12], [13]. **Scanning gaze**, where the robot moves its gaze around its environment, simulating exploration or attention to multiple objects or people. **Follow gaze**, where the robot tracks the movement of a human's gaze, implying attention or interest in what the human is looking at.

In a human-robot interaction, the prediction of a gaze target can be used as input for the determination of the best robot action in response to a human action, given the circumstances. The challenges arise from the need to simulate human-like gaze behavior in a way that is realistic, engaging, and responsive to human actions. Follow gaze, where the robot tracks the movement of a human gaze, can be easier to model compared to other types of gaze behavior in a robot. This is because gaze follow often requires less sophisticated modeling of human gaze and more straightforward control of the robot's gaze mechanism. These different types of gaze can be used in combination to create more sophisticated and nuanced interactions between the robot and the human.

In gaze follow, the robot uses computer vision algorithms to detect the position and movement of the human gaze, and then adjusts its own gaze accordingly. It does not need to respond in real-time or make sophisticated judgments about the context of the interaction. Instead, the robot simply follows the human gaze as it moves. However, it is important to note that gaze follow is still a challenging task, particularly when the gaze is fast-moving or unpredictable and goal/context dependent. The results of the study have a direct application on improving the contextualized gaze on a social robot and in this paper, we discuss the gaze

architecture for implementation on a robot.

IV. Method

A. Data

The data consists of videos of pairs of participants, working at the Good Housekeeping Institute[2] (GHI, a consumer product testing organisation in the UK), who taste eight different types of hummus and rate them. Participants are seated at 90° angle to each other, with separate cameras and radio microphones capturing each participant. The setup is designed to record clear eye movements, facial expressions, gestures and speech.

In total, 24 high definition videos lasting between 24-30 minutes were recorded at 30 frames per second. Participants were allowed to spend considerable amount of time for tasting and rating each hummus, while choosing their strategies in performing the task and organising the interaction. They are task directed dialogues and not completely spontaneous. This type of dialogue is ideal for our purposes as it allows the internal dynamics of the conversation to be entirely free while the task creates an external trigger about which participants are communicating, meaning that both referential and interactive aspects of gaze ought to be present.

B. Annotation and Transcription

For the video transcription principles of Gesprächsanalytisches Transkriptionssystem (GAT) were considered. Annotation was done using ELAN software[3]. The orthographical transcription was done in 2 different tiers *speech1* and *speech2* for each participant(figure 1). These tiers contained metadata indicating the beginning and end of the excerpt with respective spoken utterance per unit encoded. In some situations a short description of the interactional context in unicode such as cough, umm, laughter etc.

We were able to study interactional dynamics by using the recordings specific to face-to-face dialogue while also understanding the collaborative processing and generation of language. Henceforth, during the recording sessions participants had to perform a collaborative task while having free range of conversations yielding natural interaction.

C. Neural Network Architecture

The videos that were manually annotated were later used to predict robust gaze target location.

1) Convolutional layers: The architecture consist of head convolutional layers for head feature extraction (ResNet-50). It is followed by an additional residual layer and an average pooling layer to reduce the spatial dimension of the feature maps. The blue and the purple pixels shown in figure 2, denote the head bounding box of each person and the white pixels denote rest of the image. They are reduced using three max pooling operations. The scene feature extraction network computes scene feature map similar to the head convolution module. Head feature map and head position are then concatenated while the scene convolution is the concatenation of the head position and the scene image.

2) Dense layers: The final layers of the architecture consist of a fully connected layer where the attention layer computes attention maps by passing the two concatenated layers. Lastly, two convolution layers encode the features in the encoder module.

$$y = W_d * h + b_d \tag{1}$$

where y is the gaze prediction, W_d is the weight matrix of the final dense layer, h is the output of the previous layer, and b_d is the bias term of the final dense layer.

$$h_i = f(W_i * x + b_i) \tag{2}$$

where h_i is the output of the i^{th} convolutional layer, f is the activation function, W_i is the weight matrix of the i^{th} layer, x is the input image, and b_i is the bias term of the i^{th} layer.

3) Recurrent layers: The scene information providing head position as spacial referencing enables the model to learn faster. Subsequently, the architecture includes convolutional Long Short-Term Memory (Conv-LSTM) to capture temporal dependencies in the eye movement data from the sequence of frames. Four deconvolution layers makeup the deconv module to up-sample the features computed by the convLSTM into a full sized feature map.

$$h_t = f_r(W_r * h_{t-1} + U_r * x_t + b_r) \tag{3}$$

where h_t is the hidden state at time step t, f_r is the activation function of the recurrent layer, W_r and U_r are the weight matrices of the recurrent layer, x_t is the input image at time step t, and b_r is the bias term of the recurrent layer.

4) Object detection: The feature map is then modulated by a scalar that defines whether the gaze attention of the person is within the bounds of the scene or out-of frame, higher the value of the scalar the focus is within the frame. It consists of two convolution layers and a fully connected layer while the element-wise subtraction from the feature map normalization is performed. Following, heatmap with minimum values greater than or equal to zero are cropped resulting in the final heatmap that can be visualized with intensity maps of the object prediction.

V. Implementation of Neural Networks

A. Multiface processing pipeline

The technique uses a two-stage strategy. First, by considering the scene and head data as a separate network input that share the same scene properties. The scene image serves as a discrete input to the scene-channel for a one-shot feature extraction regardless of the presence of the individuals.

The feature extraction backbone network is Resnet50 [14] for network verification, where the head channel considers the number of persons present in an input image into account (dyadic in this particular situation). To forecast the gaze target, the two participants' head images and head locations

[2]https://www.goodhousekeeping.com/uk/ the-institute/

[3]https://archive.mpi.nl/tla/elan

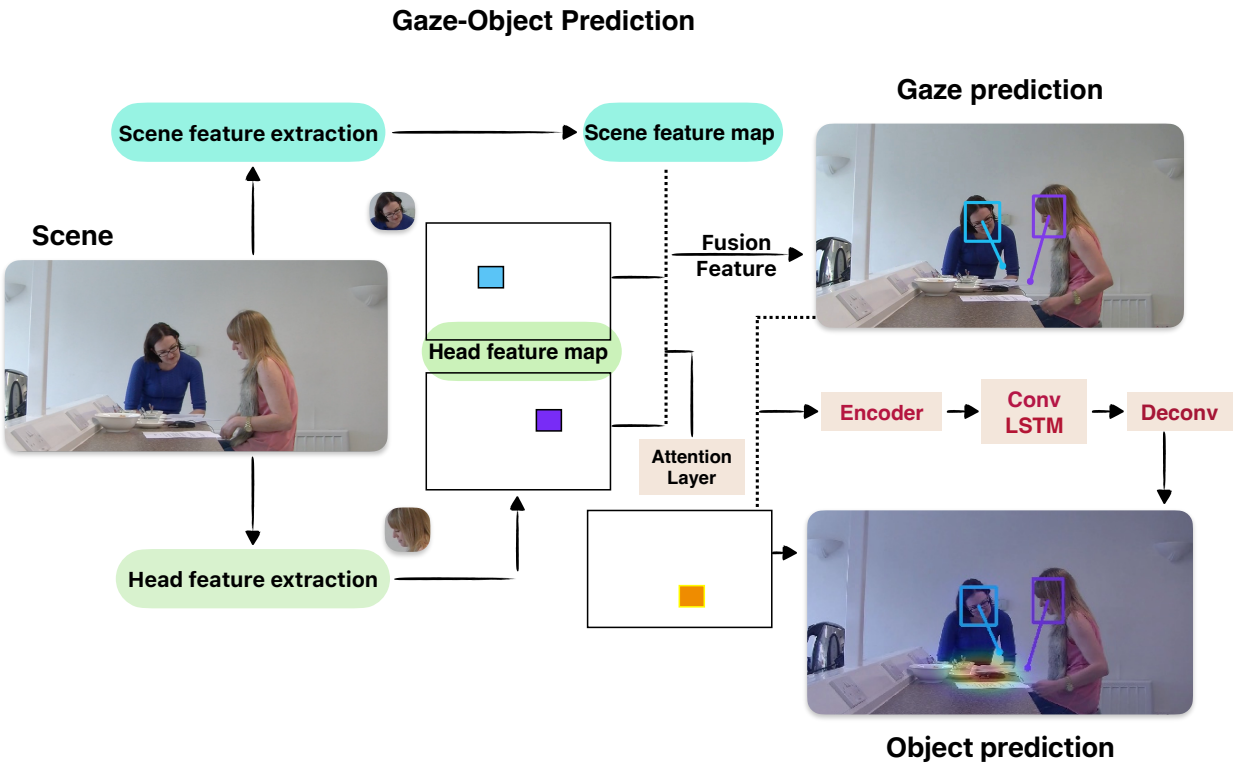

Fig. 2. In head feature extraction, location mask of the head image performs multiple feature extraction (in this case two people) and the scene image acts an an independent output for a one time feature extraction. Following, head features are concatenated with the shared scene features. The fusion feature predicts the the final gaze output of the respective person. The object detection maps the corresponding gaze target point generating heatmaps

are used as binary masks. Each head image location mask serves as the head channel input.

$$f_h = C_h(I_h), f_s = C_s(I_s) \tag{4}$$

Second, in order to predict the gaze, same head (f_h) and scene (f_s) features are concatenated to a fusion layer which goes through several up sampling and convolution layers to predict the position heatmap of the gaze targets. To avoid exorbitant consumption of computational resources for scene and face feature extraction we opted for lightweight ghostnet module which uses inexpensive operations to generate feature maps similar to convolutional layers, although it does not transcend the convolution operation.

B. Face detection and heatmap generation

Single-stage methods for multi-stage face recognition are preferred for real-time applications due to their light weight and high accuracy. For example, face recognition methods apply a single-layer architecture to design more efficient modules for facial features.

We created an extensive face dataset using a large number of manual annotations and use a one-step gaze estimation method. The input to the model is a complete image with two faces and the output is the gaze directions of the people in the scene. Instead of processing each face individually, we propose a model that estimates the gaze of multiple people simultaneously with assistance from attention heatmaps within the image.

Heatmap generation for gaze predictions using machine learning involves creating visual representations of gaze data. Heatmaps are generated by plotting the gaze data onto a 2D image and color coding the points based on the density of gaze data in that area [15]. Gaze heatmaps can be used to evaluate the performance of gaze-based algorithms by comparing the generated heatmaps with ground-truth data.

C. Gaze estimation

The model was trained on NVIDIA RTX TITAN GPU and implemented in Pytorch[4]. The network is optimized by the Adam algorithm for 100 epochs and the batch size was set as 32, and the initial learning rate was 10^{-4}. The scene and the masked header image, acts as input to the model, scaled to 224*224. The output was a heat map of size 64*64 and the ground truth heat map was generated with 2D Gaussian weights around the ground truth of the tracked target. During training, to ensure a repeatable balance of the two channels,

[4]The code will be made available on GitHub

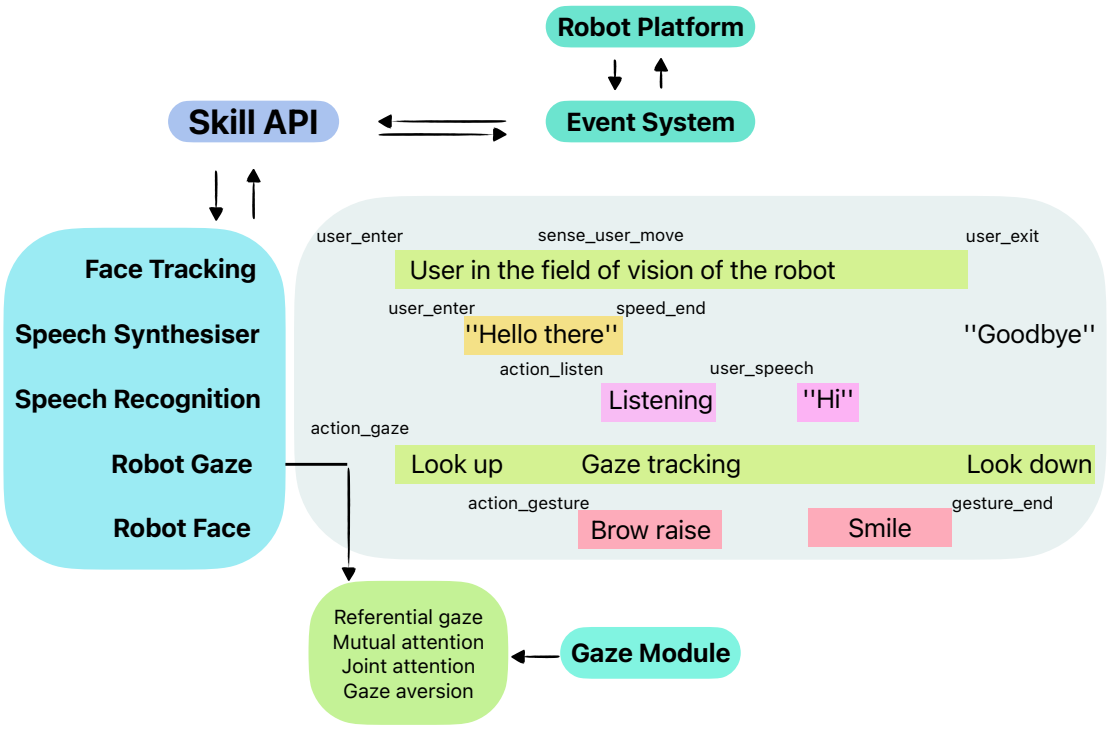

Fig. 3. Real-time Human-Robot Interaction architecture for a Social Robot

a single randomized head image was selected from the scene during each iteration.

In addition, weight loss varied depending on the duration of the training. In the early stages, we expect the output heatmaps to reflect the tracked target points, like previous gaze tracking methods. In the final stage, we focused on refining heat maps and minimizing errors through regression. Thus, in our implementation, the value of (alpha) increases as the number of training epochs increases from 0 to 0.5. During inference, to track multiple people's gazes, the scene-channel extracts features from the scene only once, while the main channel runs multiple times for different people.

VI. Experiment and Evaluation

The manual annotation has been done on 4 videos for the GHI corpus and we automate gaze for all 24 videos ranging between 24-30 minutes, and the generated images for each video is between 40k to 60k. Therefore, the resulting gaze prediction dataset (GHI-Gaze) approximately consists of 2.4 million images with facial landmark, gaze information and heatmaps were generated with two different angles for the same session (figure 1). The various types of gaze behaviour can be extracted by collecting the specific coded temporal information.

Two experiments were conducted to evaluate the performance of the model. We compare gaze annotation from GHI dataset (coded for various types of gaze) [16] to our automatically generated gaze estimate images from videos (GHI-Gaze). Evaluating the accuracy of automatic and manual gaze predictions involves comparing the predicted gaze locations to the ground truth, which is typically obtained through manual annotation by a human annotator. Following, we evaluated our models performance on the GazeFollow dataset that is publicly available. ImageNet, PASCAL and MSCOCO are used to build the GazeFollow dataset which contains 122k images of different scenes and 130k annotations.

We use four performance measures that are key indicators based on previous gaze-following methods to evaluate the model. **AUC** (Area Under the Curve) is the metric used to evaluate the performance of a binary classifier i.e the similarity between the predicted versus the ground truth heatmap. AUC ranges from 0 to 1, with a value of 1 indicating perfect classifier performance, and 0.5 indicating random performance. The averaged difference between the coordinates of the predicted gaze target point and the coordinates of the ground truth point that some annotators have assigned a label is the average distance, **Avg Dist**. The shortest distance between the anticipated point and the closest labeled point is the Minimum Distance, **Min Dist**. The angular error between the predicted and ground truth gaze direction from the head position to the attention target in the image is referred to as **Ang**.

Similar to the recent gaze prediction approaches [22], [23], [24], [25], [26], [27] for feature extraction we adopt resnet50 network. The resulting comparison with previous methods is

TABLE I
EVALUATING ON GAZEFOLLOW DATASET

Method	AUC ↑	Avg. Dist ↓	Min. Dist ↓	Ang ↓
Recasens et al.(2015)	0.878	0.190	0.113	24.0°
Lian et al.(2018)	0.906	0.145	0.082	17.6°
Chong et al.(2020)	0.921	0.137	0.077	-
Dai et al.(2021)	0.922	0.133	-	16.1°
Jin et al.(2021)	0.919	0.126	0.076	-
Tu et al.(2022)	0.917	0.133	0.069	-
GHI-Gaze (ours)	0.920	0.112	0.059	13.7°
Human	0.924	0.096	0.040	11.0°

shown in Table 1. 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.

In Figure 1, "*a*" represents mutual gaze where the individuals are looking at one another. Due to a clear view of the face it is easier to visualize the gaze in the first image. While in the second generated image of "*a*" due to face occlusion it is impossible for human annotators to recognize the gaze, but the model accurately detects the gaze taking into consideration other factors such as head pose estimates and attention heatmaps.

Images represented as "*b*" denote gaze on the same object in the scene, and "*c*" averted gaze where an individual looks at the partner and the partner looks away. Finally, "*d*" represents gaze on different objects with in the scene.

VII. GAZE INTERACTION ARCHITECTURE FOR A ROBOT

The Furhat Robot platform consists of input and output interfaces (projector, neck servo motors, touchscreen, etc.) and software modules for automatic speech recognition (ASR), text-to-speech synthesis (TTS), face tracking, etc. The Event System mediates all of the sensory inputs, modules, and actuators in the Robot Platform. To create a gaze plan, the Gaze Planner advocates high-level events such as the user's position, speech input, and the positioning of objects on the touchscreen. The Gaze Controller then takes advantage of this strategy to generate actions that causes the robot's head to turn and eyes to move. Using the Skill API, which defines all of the interaction-specific details, is where the interactions can be implemented.

The interaction is modelled using state charts where the dialog contexts are defined as hierarchical states and the generic behaviors are defined on higher levels in the hierarchy (figure 3). While the more specific behaviors are defined further down in the hierarchy, and may override generic behaviors. The intent classification (NLU) dynamically takes the current hierarchical contexts enabling multiple intent in each utterance. Complex behaviors may be defined in their own state charts, and reused across applications. Computer vision platform tracks real time multi-user face and estimates head pose from the video stream. The architecture provided in figure 3, describes the dialogue with gaze module implementation from the current work. A specific type of gaze, acts as an input for the robotic gaze based on the speech intent and face tracking by assessing the temporal predictions.

VIII. DISCUSSION AND FUTURE WORK

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. We compare the results firstly with manually annotated data from GHI corpus and then with the popular GazeFollow dataset. 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.

We faced challenges such as head pose variations, occlusions, and cluttered backgrounds, but with the help of extensive manual annotation data made available it has been possible to reduce error while also adopting open-sourced pre-trained models. Most current gaze prediction methods use visual information only [17]. However, incorporating linguistic modalities such as dialogue could lead to improved performance and more natural gaze predictions [18].

It is possible to identify when someone is inattentive by observing how they look at an object, following their gaze, and even identifying if they are maintaining mutual gaze. Yet, there remains a complex, open challenge in automatically detecting and quantifying these types of visual attention from images and videos.

Gaze information can be used as an input for human-Robot interaction (figure 3), and the work could focus on developing more sophisticated gaze-based interaction methods that are more natural and intuitive. Gaze can provide insights into human behavior, such as attention, memory, and emotions [19], [20], [21]. Hence, by developing new methods for analyzing gaze data, it is possible to gain a better understanding of human behavior and its underlying cognitive processes. We plan to implement the gaze results obtained from the study on a Furhat Robot.

Many current gaze estimation and prediction methods are computationally intensive and not suitable for real-time applications. Developing fast and efficient algorithms for gaze estimation and prediction is an important area for future work.

Overall, the field of gaze estimation and prediction has the potential to revolutionize the way we interact with technology and gain insights into human behavior. There is much work to be done to reach these goals, but the potential impact is significant.

ACKNOWLEDGMENT

The research reported in this paper was supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg, Sweden.

REFERENCES

[1] M. Staudte and M. W. Crocker, 'Visual attention in spoken human-robot interaction', in Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, 2009, pp. 77–84.

[2] G. Storey, A. Bouridane, and R. Jiang, 'Integrated deep model for face detection and landmark localization from "in the wild" images', IEEE Access, vol. 6, pp. 74442–74452, 2018.

[3] T. Jin, Z. Lin, S. Zhu, W. Wang, and S. Hu, 'Multi-person gaze-following with numerical coordinate regression', in 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), 2021, pp. 01–08.

[4] D. Apgar and M. R. Abid, 'Multi-Model Face Liveness Detection Via Gaze Detection and Convolutional Neural Networks', in 2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2022, pp. 0255–0261.

[5] P. Li, J. Su, A. N. Belkacem, L. Cheng, and C. Chen, 'Steady-state visually evoked potential collaborative BCI system deep learning classification algorithm based on multi-person feature fusion transfer learning-based convolutional neural network', Frontiers in Neuroscience, vol. 16, 2022.

[6] H. Xu, J. Zhang, H. Sun, M. Qi, and J. Kong, 'Analyzing students' attention by gaze tracking and object detection in classroom teaching', Data Technologies and Applications, 2023.

[7] V. Nagpure and K. Okuma, 'Searching Efficient Neural Architecture With Multi-Resolution Fusion Transformer for Appearance-Based Gaze Estimation', in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 890–899.

[8] J. R. Wilson, P. T. Aung, and I. Boucher, 'When to Help? A Multimodal Architecture for Recognizing When a User Needs Help from a Social Robot', in Social Robotics: 14th International Conference, ICSR 2022, Florence, Italy, December 13–16, 2022, Proceedings, Part I, 2023, pp. 253–266.

[9] N. Sendhilnathan, T. Zhang, B. Lafreniere, T. Grossman, and T. R. Jonker, 'Detecting Input Recognition Errors and User Errors using Gaze Dynamics in Virtual Reality', in Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, 2022, pp. 1–19.

[10] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu, 'Conversational gaze aversion for humanlike robots', in Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction, 2014, pp. 25–32.

[11] Y. Zhang, J. Beskow, and H. Kjellström, 'Look but don't stare: Mutual gaze interaction in social robots', in Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9, 2017, pp. 556–566.

[12] J.-D. Boucher et al., 'I reach faster when I see you look: gaze effects in human–human and human–robot face-to-face cooperation', Frontiers in neurorobotics, vol. 6, p. 3, 2012.

[13] V. Somashekarappa, C. Howes, and A. Sayeed, 'An annotation approach for social and referential gaze in dialogue', in Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 759–765.

[14] X. Cai et al., 'Gaze estimation with an ensemble of four architectures', arXiv preprint arXiv:2107. 01980, 2021.

[15] B. Wang, T. Hu, B. Li, X. Chen, and Z. Zhang, 'Gatector: A unified framework for gaze object prediction', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19588–19597.

[16] V. Somashekarappa, C. Howes, and A. Sayeed, 'A deep gaze into social and referential interaction', in Proceedings of the Annual Meeting of the Cognitive Science Society, 2021, vol. 43.

[17] Y. Yu, G. Liu, and J.-M. Odobez, 'Deep multitask gaze estimation with a constrained landmark-gaze model', in Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0–0.

[18] H. Kim, Y. Ohmura, and Y. Kuniyoshi, 'Memory-based gaze prediction in deep imitation learning for robot manipulation', in 2022 International Conference on Robotics and Automation (ICRA), 2022, pp. 2427–2433.

[19] M. H. Black et al., 'Mechanisms of facial emotion recognition in autism spectrum disorders: Insights from eye tracking and electroencephalography', Neuroscience and Biobehavioral Reviews, vol. 80, pp. 488–515, 2017.

[20] C. Scopa, L. Contalbrigo, A. Greco, A. Lanatà, E. P. Scilingo, and P. Baragli, 'Emotional transfer in human–horse interaction: New perspectives on equine assisted interventions', Animals, vol. 9, no. 12, p. 1030, 2019.

[21] F. Cassioli, L. Angioletti, and M. Balconi, 'Tracking eye-gaze in smart home systems (SHS): first insights from eye-tracking and self-report measures', Journal of Ambient Intelligence and Humanized Computing, vol. 13, no. 5, pp. 2753–2762, 2022.

[22] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, 'Where are they looking?', Advances in neural information processing systems, vol. 28, 2015.

[23] D. Lian, Z. Yu, and S. Gao, 'Believe it or not, we know what you are looking at!', in Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14, 2019, pp. 35–50.

[24] E. Chong, Y. Wang, N. Ruiz, and J. M. Rehg, 'Detecting attended visual targets in video', in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5396–5406.

[25] L. Dai, J. Liu, Z. Ju, and Y. Gao, 'Attention-Mechanism-Based Real-Time Gaze Tracking in Natural Scenes With Residual Blocks', IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 2, pp. 696–707, 2021.

[26] T. Jin, Z. Lin, S. Zhu, W. Wang, and S. Hu, 'Multi-person gaze-following with numerical coordinate regression', in 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), 2021, pp. 01–08.

[27] D. Tu, X. Min, H. Duan, G. Guo, G. Zhai, and W. Shen, 'End-to-end human-gaze-target detection with transformers', in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2192–2200.