

Using Eye Gaze to Forecast Human Pose in Everyday Pick and Place Actions

Haziq Razali and Yiannis Demiris¹

Abstract— Collaborative robots that operate alongside humans require the ability to understand their intent and forecast their pose. Among the various indicators of intent, the eye gaze is particularly important as it signals action towards the gazed object. By observing a person’s gaze, one can effectively predict the object of interest and subsequently, forecast the person’s pose. We leverage this and present a method that forecasts the human pose using gaze information for everyday pick and place actions in a home environment. Our method first attends to fixations to locate the coordinates of the object of interest before inputting said coordinates to a pose forecasting network. Experiments on the MoGaze dataset show that our gaze network lowers the errors of existing pose forecasting methods and that incorporating prior in the form of textual instructions further lowers the errors by a significant amount. Furthermore, the use of eye gaze now allows a simple multilayer perceptron network to directly forecast the keypose.^a

I. INTRODUCTION

Forecasting human intent is crucial for the deployment of collaborative robots. Intention can be defined by an end pose representing the goal, and a trajectory representing the preferred way of achieving the goal. As such, to achieve successful joint action, a robot must infer human behavioural cues to forecast their intentions in order for it to react appropriately and in a timely manner. Among the many cues, the eye gaze is known to be the most powerful predictor of human intent with empirical evidence showing that gaze information signals action intent and lead motor actions that follow [1]. For instance, a wearable eye tracker can effectively forecast the camera wearer’s actions at least 1 second in advance [2] and by attending to only 20% of the image contents centered around the wearer’s gaze point [3]. In the context of collaborative robots, the eye gaze allows the robot to take the perspective of a person and gain access to the mental states for intention forecasting [4]. Existing works incorporating eye gaze have done so for egocentric action recognition [3] and anticipation [2], perspective taking [4], and augmented reality control [5].

Studies show that there exist several types of gaze patterns such as fixations, saccades, and smooth pursuits [6]. Fixations stabilize the retina over a stationary point of interest. Saccades are rapid movements of the eyes that abruptly change the point of fixation made e.g. while gazing around a room. Finally, smooth pursuits are tracking movements when the eyes follow a moving target. Their usage thus depends

on the task at hand such as fixations for action recognition and smooth pursuits for visual object tracking.

Our focus in this paper is the use of eye gaze in everyday pick and place actions, specifically when a person performs a sequence of pick and place actions in a home environment. In this context, saccades and fixations become the most relevant types of gaze patterns. Saccades would be used mainly for exploring the environment in search of the object of interest during which intent cannot be inferred while fixations occur thereafter when the target object has been located and the eye gaze is kept on the object as the person moves towards it. The observed gaze patterns during a sequence of pick and place actions thus alternates between saccades and fixations, with both offering an insight into the person’s mental state, whether the person is searching for the object of interest, or has located it and is now moving towards it. We leverage this information and propose a neural network that attends to fixations and the fixated object, the latter of which is used to condition a pose forecasting network. In summary, our contributions are as follows:

- We present a novel method that operates on gaze information to identify the fixated object.
- The network is designed in such a way such that it can be augmented to any existing pose forecasting methods to lower their forecasting errors.
- The network can augment even a simple multilayer perceptron (MLP) to use pose information from a single timestep to directly forecast the pose the instant the person reaches the fixated object, differing largely from existing methods that use a recurrent architecture to forecast one step at a time.

II. RELATED WORK

Human Pose Forecasting: Existing works on pose forecasting output either the joint positions or joint rotations that are then converted to positions via forward kinematics. For instance, Martinez et al. [7] proposed a Gated Recurrent Unit (GRU) [8] that outputs the joint velocities. Li et al. [9] proposed to replace the GRU with convolutions to counteract the information loss that occurs in long term predictions. Corona et al. [10] modelled the relationship between joint positions and obstacles using a graph recurrent network. Fragkiadaki et al. [11] and Pavllo et al. [12] predict joint rotations that are parameterized using euler angles and quaternions respectively. There are also methods that incorporate stochasticity using variational autoencoders [13], [14]. A similarity shared by all the above-mentioned works however, is that they predict the pose one step at a time,

¹The authors are with the Personal Robotics Laboratory, Dept of EEE, Imperial College London, SW7 2AZ, UK {h.bin-razali20, y.demiris}@imperial.ac.uk YD is supported by a Royal Academy of Engineering Chair In Emerging Technologies.

^aCode available at www.imperial.ac.uk/personal-robotics/software

and do not use gaze information which is our contribution in this paper.

Eye Gaze: There exists a large body of work on eye gaze including estimation [15], tracking [16], forecasting [17], and using eye gaze for applications such as action recognition [3] and human-robot interaction [18], [19]. We briefly cover applications that make use of eye gaze. Li et al. [3] used CNNs on first-person video to output a gaze distribution map that is used as a mask over the video for action recognition. Huang et al. [18] trained an SVM on gaze features such as gaze duration and number of glances to predict which ingredient the customer would choose for the smoothie. Richter et al. [19] inferred gaze using facial landmarks to help the robot predict the recipient of verbal communication in a multi-person setting. Shafti et al. [20] used an eye tracker to determine if the person is looking at the right-hand side of an object for assistive grasping. In our work, we assume the gaze to have already been estimated e.g. either via a wearable pupil tracking device or the method developed by Fischer et al. [21] and now want to identify the fixated object.

III. DATASET

The MoGaze dataset [22] enables the study on the correlation between eye gaze and human motion. It contains gaze data obtained via a wearable device, and motion capture data of a person performing pick and place actions around an indoor environment 3×4 metres large. The environment contains colored kitchenware items such as cups and plates, and furnitures such as tables and shelves that are all tracked via MoCap. The person is tasked to pick the items one at a time from one furniture and to place it onto another given instructions such as "put all cups on the table". The person thus alternates between a pick and place action and is not allowed hold on to more than 1 item at any time. In the following sections, we refer to objects that can be picked as pick-objects and locations for the person to place the items on as place-locations.

IV. METHOD

The goal of our method is to utilize the person's input pose and gaze to predict the pick-object or place-location and subsequently, the future pose. In short, we want to learn the expression $p(P_{\text{out}}, O^{(k)} | P_{\text{in}}, E_{\text{in}}, \mathbf{O}, s, a)$, where $a = \text{pick}$ or place , s the given instructions, $\mathbf{O}$ the label and coordinates $\in \mathbb{R}^{3 \times N}$ of all N pick-objects when $a = \text{pick}$, or N place-locations when $a = \text{place}$, $P_{\text{in}} \in \mathbb{R}^{3 \times M \times T1}$ and $E_{\text{in}} \in \mathbb{R}^{3 \times T1}$ the input pose with M joints and eye gaze respectively over $T1$ timesteps, $O^{(k)} \in \mathbb{R}^3$ the coordinates of the predicted pick-object or place-location depending on a , and $P_{\text{out}} \in \mathbb{R}^{3 \times M \times T2}$ the future pose over $T2$ timesteps.

However, because studies have shown that the eye gaze typically fixates on an object before the execution of any movement or manipulative action towards said object [23], we can thus further factorize our initial objective into the following expression:

$$\log p(P_{\text{out}} | P_{\text{in}}, O^{(k)}) + \log p(O^{(k)} | \mathbf{O}, E_{\text{in}}, s, a) \quad (1)$$

The result is more reflective of human behaviour in that it first uses gaze information to locate the pick-object or place-location before forecasting the pose. Note that due to our formulation, the values of $\mathbf{O}$ given a will be constant throughout time. In the following, we describe our method for these two components.

A. Eye Gaze Network

Attending to Fixations: Predicting $O^{(k)}$ given a sequence of gaze data is not as straightforward as it seems due to the complexity of visually guided behaviour that occurs in pick and place actions. The human gaze in this context is composed of independent subtasks such as object search followed by obstacle avoidance whilst guiding the person towards $O^{(k)}$ [24]. After locating $O^{(k)}$, the human gaze may occasionally deviate from the target due to distractors or noise, or even fixate on the ground ahead for path planning. As such, there is a need to identify and emphasise sections of the gaze events that correspond to intent and to likewise, deemphasise the less informative ones. The attention mechanism [25] in neural networks lend us the ability to do just so by learning to enhance important parts of the data while suppressing the rest. Specifically, we first compute the gaze relative to each of the N pick-objects or place-locations at time t as $\bar{E}_t - \bar{O}^{(n)}$ where the bar indicates that the vector has been normalized. The attention mechanism then aggregates the vectors according to the expression:

$$h^{(n)} = \phi_1(\bar{E}_{\text{in}} - \bar{O}^{(n)}) \quad (2)$$

$$\alpha_t^{(n)} = \frac{e^{\phi_2(h_t^{(n)})}}{\sum_t e^{\phi_2(h_t^{(n)})}} \quad (3)$$

$$f^{(n)} = \sum_t \alpha_t^{(n)} (\bar{E}_t - \bar{O}^{(n)}) \quad (4)$$

where $\bar{O}^{(n)}$ is repeated $T1$ times to match the shape of $\bar{E}_{\text{in}}$. ϕ_1 can be any neural network that operates on a 2D tensor such as GRUs and ϕ_2 a fully connected layer that produces a scalar. We have ϕ_1 in equation 2 process $\bar{E}_{\text{in}} - \bar{O}^{(n)}$ instead of only E_{in} since we find that the gaze vectors alone do not convey fixations when measured on a moving person. To explain, although $O^{(n)}$ remains stationary, the location of the eyes and consequently the gaze vectors E_{in} , continue to change as the person moves around the environment even if the gaze is fixated on $O^{(n)}$. E_{in} alone thus offers no patterns for the neural network to learn. Their normalized difference $\bar{E}_{\text{in}} - \bar{O}^{(n)}$ however, remains rather constant and small during fixations making it easier for the network to learn. $f^{(n)}$ thus tells us the resulting gaze vector with respect to object n after aggregating them via the attention scores. We aid the model in learning to identify fixations by injecting prior knowledge into the loss function as follows:

$$L_{\text{gaze}} = \sum_n ||f^{(n)} - \text{median}(\bar{E}_{\text{in}} - \bar{O}^{(n)})||_2^2 \quad (5)$$

where median is the geometric median over the vectors. In short, equation 5 acts as an auxiliary loss [26] that informs the model that the fixation of a gaze sequence tend

to be the vector that points towards the median. Note that the weight for L_{gaze} is kept small since it serves as an auxiliary loss.

Attending to Fixated Objects: Given the aggregated gaze vectors relative to each object $f^{(1)}$ to $f^{(n)}$, our task is to now determine the pick-object or place-location $O^{(k)}$. Although this appears to be a classification problem that can be optimized with respect to a one-hot target, the reality is that a person’s gaze towards a specific object may not reflect his true intent, especially when the scene is cluttered with many objects or distractors nearby [27]. For instance, a person who intends to look for and pick a red cup will fixate on several other colored cups and objects that are in close proximity while performing this visual search. The placement of these fixations may also not lie within each object and are actually influenced by factors such as color and shape with respect to the target object [28]. Together with noisy gaze measurements, it is a possibility that the gaze vector is in closer proximity to the object the person does not intend to pick. We tackle this issue by converting the problem from classification to regression. Specifically, we optimize the following loss function:

$$L_{\text{pick-place}} = \left\| \sum_n \beta_n O^{(n)} - O^{(k)} \right\|_2^2 \quad (6)$$

where β_n is computed as follows:

$$\beta_n = \frac{e^{\phi_3(f^{(n)})}}{\sum_n e^{\phi_3(f^{(n)})}} \quad (7)$$

where ϕ_3 is a fully connected layer that produces a scalar. The term $\sum_n \beta_n O^{(n)}$ is in fact, another attention weighting over $O^{(n)}$ where the weights β_n are produced by feeding the attended fixation vectors $f^{(n)}$ obtained from equation 4. Prediction can then be made via the attention scores:

$$\hat{O}^{(k)} = \text{argmax}_{\beta_n} O^{(n)} \quad (8)$$

Equation 6 encourages the model to assign higher probability scores to $O^{(n)}$ located near $O^{(k)}$ where the ideal scoring is $\beta_k = 1$. The advantage of optimizing it this way is that the loss is scaled according to how far $\sum_n \beta_n O^{(n)}$ is from $O^{(k)}$. In contrast, using a one-hot target with a classification-based objective such as the cross entropy loss excludes the notion of distance and thus incurs equal penalties for wrong predictions no matter how close they are to $O^{(k)}$. Solving via regression thus alleviates the issue of the uncertainty in gaze measurements.

To summarize, equations 2 to 5 encourage the model to attend to fixations relative to each pick-object or place-location while equations 6 and 7 encourage the model to attend to the pick-objects or place-locations that are most likely being fixated on.

Forecasting $O^{(k)}$ for Pick and Place Actions: Till now, our description of the method only allows for pick-actions since we assumed $O^{(n)}$ to be the center of object n . A similar representation for the place-action i.e. using the center of

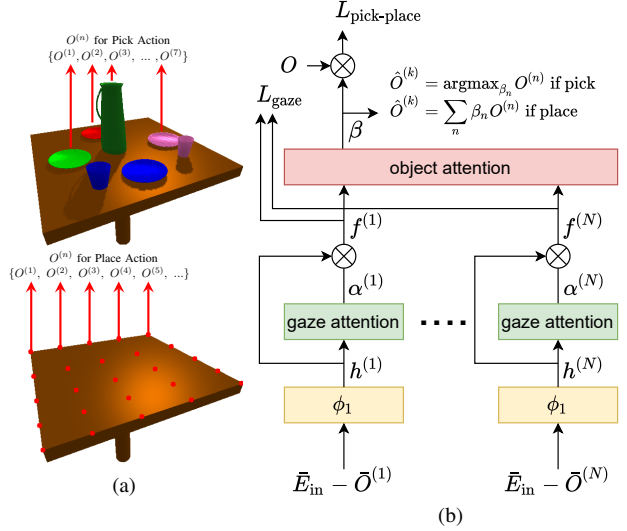


Fig. 1: (a) $O^{(n)}$ for pick and place actions (b) Our gaze network. The gaze attention block attends to fixations relative to each object while the object attention block attends to objects that are most likely being fixated on.

the table or the shelf for $O^{(n)}$ will not do since the person is at liberty to choose the exact place-location which may not be the center. We circumvent this issue by discretizing the surfaces into a grid of coordinates. Then, because the person may also not place the object at the center of any grid, we replace the argmax in equation 8 with a weighted sum over the grid. This lets the model predict a continuous value for the place-location instead of one of the grid’s centroid. Specifically, the output of our neural network is defined according to the current action a :

$$\hat{O}^{(k)} = \text{argmax}_{\beta_n} O^{(n)} \text{ if } a = \text{pick} \quad (9)$$

$$\hat{O}^{(k)} = \sum_n \beta_n O^{(n)} \text{ if } a = \text{place} \quad (10)$$

and where $O^{(n)}$ denotes the coordinates of kitchenware item n such as cups and plates during a pick-action, and the coordinates of point n on the grid during a place-action. We provide an illustration of $O^{(n)}$ for pick and place actions in Figure 1a and our gaze network that outputs $O^{(k)}$ in Figure 1b.

Utilizing Textual Instructions as Prior: Although a person’s gaze provides sufficient information about the intended pick-object or place-location $O^{(k)}$, having prior knowledge of the person’s goal can further benefit the model by reducing the number of candidates in a cluttered environment, which consequently improves accuracy. This often occurs in collaborative tasks when a person voices his intentions to the other person e.g., ”let me get a cup for you”. The cups then form the list of potential pick-objects. In the MoGaze dataset, this prior exists in the form of instructions read by the participant such as ”place all cups on the table”. Similarly, the cups and table would then form the list of pick-objects and place-location respectively.

TABLE I: MoGaze instructions and their parsed tuples.

Instruction	Parsed Tuple <Pick-Object, Quantity, Place-Location>
Set the table for 1/2/3/4 persons	< [cup, plate, jug, bowl], 1/2/3/4, table >
Put all cups on big shelf	< cup, all, big shelf >
Put all green objects on big shelf	< green, all, big shelf >
Put jug and bowl on small shelf	< [jug, bowl], 1, small shelf >

We augment our gaze network with a very simple instruction parsing tool. Specifically, we scan the text for keywords belonging to 3 categories: (a) the pick-objects such as cups and plates, (b) the place-location such as a table, and (c) the quantity to be moved. The list of pick-objects or place-locations to be fed to the gaze network are then filtered based on their labels and locations. For example, the instruction ”place all cups on the table” would be first converted into the tuple <cup,all,table>. Items other than cups that are not on the table, or cups that are on the table will be then removed from the list of pick-objects. Likewise, place-locations other than the table will not be fed to the neural network if the person is currently holding a cup. We chose to hardcode the instructions instead of learning them in order to avoid overfitting since there are only a total of 439 instructions in the 3 hour dataset. We present some instructions and their parsed tuples in Table I.

B. Pose Forecasting Network

We make a simple extension to existing recurrent architectures built for pose forecasting by concatenating $\hat{O}^{(k)}$ to P_{in} at every timestep. This information lets the model learn to correlate P_{out} to both P_{in} and $\hat{O}^{(k)}$, conditioning it to direct its output towards $\hat{O}^{(k)}$. Note that our gaze network is designed to work directly with pose networks that forecast the joint positions in world coordinates [7], [9], [10], [13] but not the joint rotations [11], [12] since the latters’ absence of world coordinates in their inputs makes it far more difficult to correlate the rotation values to coordinates $\hat{O}^{(k)}$.

We also show that knowledge of $\hat{O}^{(k)}$ now makes it possible to use a simple MLP that uses the pose from a single timestep to directly forecast the pose the instant the pick or place action is completed, and the person either comes into contact with the pick-object or places the object onto the place-location respectively. In short, we are now able to solve a variant of equation 1:

$$\log p(P_k | P_i, O^{(k)}) + \log p(O^{(k)} | \mathbf{O}, E_{in}, s, a) \quad (11)$$

where P_k denotes the above-described key-pose and P_i the input pose at time i which corresponds to the final timestep of E_{in} . We solve this using an MLP that similarly, uses the concatenation of P_i and $O^{(k)}$ as its input to output the joint positions. Note that we do not take obstacles into account.

V. EXPERIMENTS

A. Setup

We evaluate our gaze network on the MoGaze dataset [22] described in section III. We set $L_{gaze} = 0.1$ and $L_{pick-place} = 1$. We then freeze the weights and attach it

to several pose-forecasting architectures [13], [9], [10] as described in section IV-B, and train them to forecast the next 2 seconds (20 frames) using the first 0.5 seconds (5 frames) as input. We also train an MLP with 3 hidden layers to directly forecast the keypose using a single input pose from the final gaze timestep. Finally, we scale up the number of parameters of all models to match [10] at 3 million.

We develop all models on PyTorch and train them to convergence using ADAM [29] with a learning rate of 1e-3, batch size of 32, and default hyperparameters. We preprocess each segment by subtracting the coordinates of every joint and object by the pelvis of the first frame. We do not augment the data in any way.

B. Quantitative Results

We first present the quantitative results for pose forecasting with and without eye gaze and instructions in Table II. We report the L2 errors at various timesteps including the key-pose. Note that the key-pose does not occur at a fixed timestep.

It can firstly be seen that the inclusion of eye gaze lowers the errors the further predictions are made into the future. The errors are further reduced when the model makes use of instructions. These numbers provide strong evidence that using instructions help reduce clutter by narrowing down the list of pick-objects or place-locations, especially when the eye gaze is unreliable. Finally, the table also indicates the improvements we get by directly optimizing over the key-pose as opposed to minimizing the errors between the predicted and ground truth poses at every timestep. Having knowledge of $O^{(k)}$ thus offers an alternative in applications where the intermediate poses are not required and what is needed is instead the key-pose when the pick or place action is completed.

C. Qualitative Results

We present some qualitative results of the MLP model that forecasts the key-pose in Figure 2. It can be observed throughout the outputs, especially in Figures 2.i.a and 2.ii.a that our model assigns higher scores to gaze vectors that lie in closer proximity to each other and lower scores to vectors that are further away from the group. Similarly, our model scores pick-objects or place-locations based on their proximity to the attended fixation vector. These results indicate that our model has learnt what fixations are and that the pick-object or place-location tend to be the one that is closest to said fixations.

Column b of Figure 2 illustrates the importance of prior knowledge. The person who has been instructed to ”put all cups on the big shelf” has just placed an item on said shelf and is about to look for the next cup to pick. The model is then given a challenging task to predict the next pick-object at a very early stage. Naturally, the model without access to instructions fail and assumes the pick-object to be the red cup on the big shelf that is being gazed at. When incorporating instructions however, the model is able to first filter out all non-relevant pick-objects including the cups that are already

TABLE II: L2 errors averaged over the joints in meters of the various methods for the pick and place actions on the MoGaze dataset. The use of eye gaze and prior such as instructions both decrease the errors by a substantial amount. Knowledge of $O^{(k)}$ through the eye gaze also allows a simple MLP or any model to directly predict the key pose.

	Eye Gaze	Instructions	$a = \text{Pick}$				$a = \text{Place}$			
			50ms	100ms	200ms	Key-Pose	50ms	100ms	200ms	Key-Pose
Chung et al. [13]			0.269	0.491	0.768	1.027	0.245	0.444	0.726	0.926
Li et al. [9]			0.221	0.426	0.847	1.281	0.233	0.433	0.820	1.075
Corona et al. [10]			0.235	0.458	0.706	0.948	0.213	0.435	0.758	0.865
Chung et al. [13]	✓		0.226	0.409	0.673	0.920	0.217	0.388	0.609	0.803
Li et al. [9]	✓		0.216	0.399	0.686	0.985	0.209	0.364	0.611	0.784
Corona et al. [10]	✓		0.223	0.398	0.654	0.923	0.218	0.376	0.593	0.734
MLP	✓		-	-	-	0.771	-	-	-	0.650
Chung et al. [13]	✓	✓	0.230	0.406	0.537	0.602	0.204	0.343	0.443	0.505
Li et al. [9]	✓	✓	0.226	0.394	0.542	0.664	0.199	0.326	0.454	0.556
Corona et al. [10]	✓	✓	0.231	0.392	0.505	0.596	0.213	0.347	0.456	0.486
MLP	✓	✓	-	-	-	0.490	-	-	-	0.318

on the big shelf. The result is that the model is now able to correctly predict the pick-object i.e. the green cup on the table even though the gaze is not directed towards it, because it is the only relevant item left. The same observation can be made for column c where the model is tasked to predict the place-location at a very early stage. These two figures thus stress the importance of incorporating prior knowledge when the gaze is not reliable, apart from helping reduce clutter.

D. Ablation Studies

We conduct several ablation studies where we trained multiple variants of our MLP and gaze network with different gaze lengths, with and without instructions. We then select the top N object that is closest to the ground truth for the pick action, and vary the grid spacing for the place action, and evaluate their pick-object, place-location, and key-pose errors. The results are summarized in Figures 3 and 4.

It can firstly be seen throughout the figures that models utilizing the given instructions achieve significantly lower errors than their counterparts without access to this information. It shows the importance of utilizing prior knowledge as it allows the model to filter out irrelevant objects from the scene, reducing clutter and consequently, lowering errors. However, although the errors coming from models without access to instructions appear significantly higher, we point out that the models were evaluated over the entire sequence and so the large errors come mainly from the start of the action segments when the person is looking in a different direction as shown in Figure 2 column b. This is further evidenced by the plots in Figure 4 where we evaluate the models at $N = 1$ and a grid spacing of 16/15/11 cm, and bin the errors based on the fixation angles, in steps of 5 degrees, from 5 to 150 degrees. Specifically, a fixation angle of 10 degrees indicates that the sequence contains a frame where the person was looking 10 degrees away from $O^{(k)}$. The bar chart then indicates the percentage that fall within each bin. Figure 4 shows that models without instructions degrade rapidly the further away the person is looking from the object. It makes sense since it is impossible to determine the object being fixated on if the person is looking in the opposite direction. Being able to filter the non-relevant objects via

instructions can help get a good estimate even when the gaze is unreliable.

Next, the extreme left of Figure 3 shows the average distance between the true pick-object and nearest object to be at 0.4m, shown by the horizontal red dashed line. This value helps us intuitively understand how cluttered the environment is and lets us gauge the performance of our model in this environment. In other words, we can summarize that the model trained with instructions tend to make predictions that is "1 object away" from the true pick-object when $N = 1$.

Figure 3 shows that the pick-object, place-location and key-pose errors all decrease as we increase N or decrease the grid size. Naturally, a predicted pick-object that is closer to the true pick-object will lead to a predicted key-pose that is closer to the true key-pose. Similarly, a finer grid makes it easier for the model to localize the coordinates of the place-location. The errors in Table II would thus be lower if the metrics were reported with $N > 1$ or an even finer grid resolution.

Finally, all figures indicate the benefits of using longer gaze lengths, particularly when instructions are not available. We see the curves over all figures improving with longer gaze lengths. Naturally, models with longer gaze lengths perform better since they are less susceptible to noisy measurements or distractors. Interestingly, the use of instructions seem to suggest the lack of effect the gaze lengths have on the errors. It again, highlights the importance of utilizing prior knowledge if available as it helps combat unreliable gaze information.

VI. CONCLUSION

We have presented a method that uses eye gaze to improve pose forecasting. Our contribution is a gaze network that runs a cascade of attention networks on the gaze vectors followed by the pick-objects or place locations. The network can be easily augmented to existing pose forecasting architectures. The main takeaways are as follows. (i) Knowledge of the pick-object or place-location identified via the gaze network can improve existing recurrent based methods, and more importantly, allow the key-pose to be directly predicted. (ii) Longer gaze lengths help lower the overall errors since the

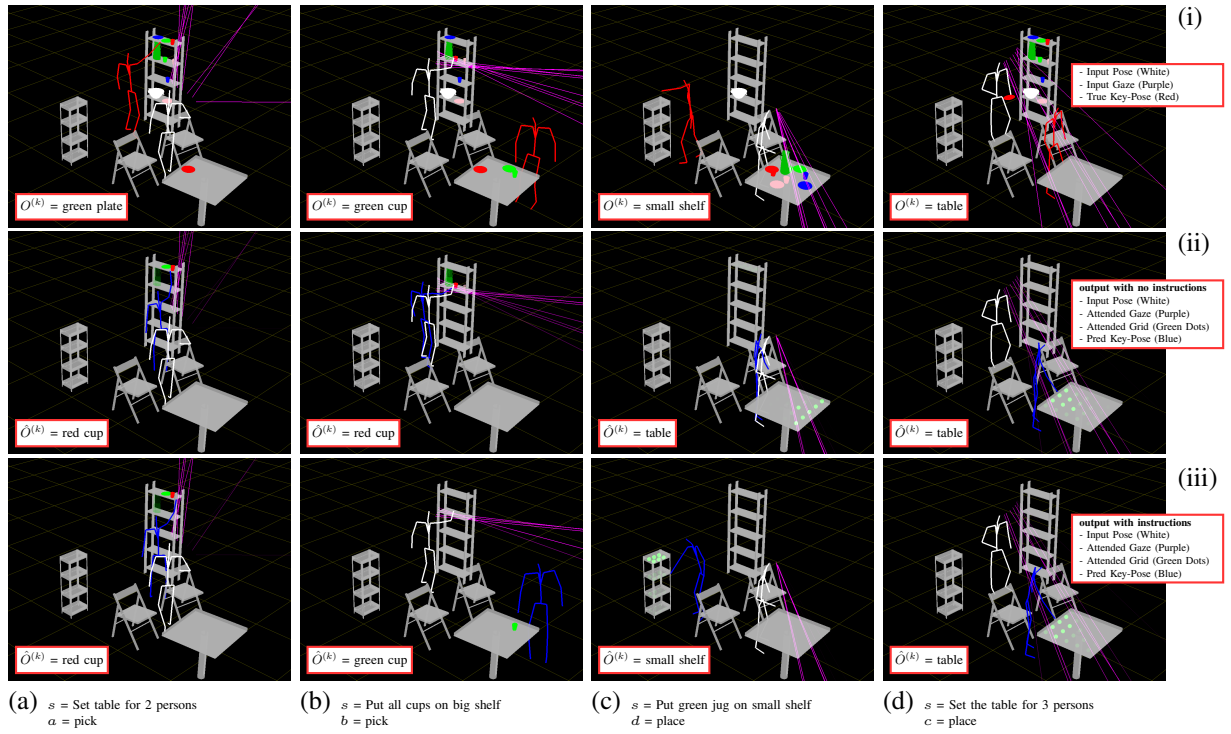


Fig. 2: Each column illustrates the input and outputs of a single segment. The caption under each column describes the segment’s input instruction s and action a . The caption to the right of each row describes the information presented and the assigned color, and the output with or without instructions. The text box in each subfigure indicates either $O^{(k)}$ or $\hat{O}^{(k)}$. Finally, the brightness of the gaze and pick-object or place-locations in rows 2 and 3 have been scaled according to their attention scores. Our model has learnt what fixations are and can use them to predict $\hat{O}^{(k)}$. The model is aided by the given instructions when the gaze is unreliable.

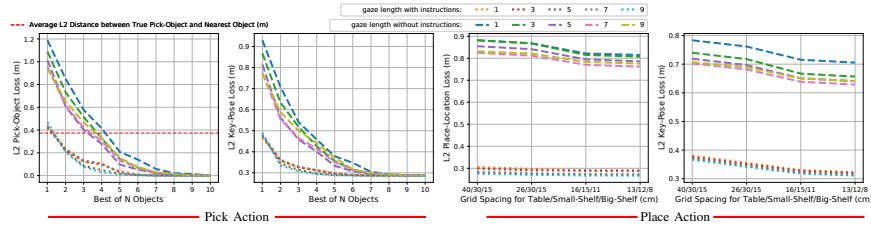


Fig. 3: Pick-Object, Place-Location, and Key-Pose errors at various settings. The errors improve as N increases or as the grid is made finer. The use of instructions also results in errors that are lower and plateau faster.

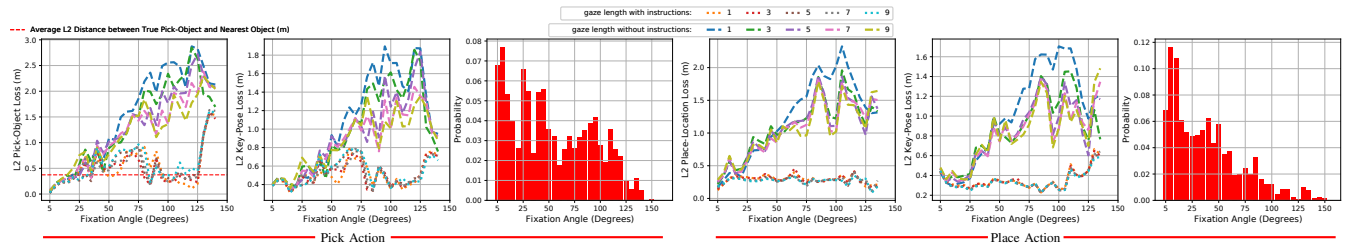


Fig. 4: Pick-Object, Place-Location, and Key-Pose errors binned into fixation angle intervals. The plots show the benefit of using longer gaze lengths to reduce the errors at higher angles when instructions are not available. They again show the importance of incorporating the instruction prior.

model is less susceptible to distractors (iii) The use of prior information such as instructions can significantly reduce errors either by reducing clutter or by making mistakes less

severe when the gaze is not reliable.

REFERENCES

- [1] M. Land, N. Mennie, and J. Rusted, "The roles of vision and eye movements in the control of activities of daily living," *Perception*, vol. 28, no. 11, pp. 1311–1328, 1999.
- [2] M. Liu, S. Tang, Y. Li, and J. M. Rehg, "Forecasting human-object interaction: joint prediction of motor attention and actions in first person video," in *European Conference on Computer Vision*. Springer, 2020, pp. 704–721.
- [3] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 619–635.
- [4] T. Fischer and Y. Demiris, "Computational modeling of embodied visual perspective taking," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 4, pp. 723–732, 2019.
- [5] R. Chacón-Quesada and Y. Demiris, "Augmented reality controlled smart wheelchair using dynamic signifiers for affordance representation," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4812–4818.
- [6] D. Purves, "Neuroscience. 2nd edition," 2001.
- [7] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2891–2900.
- [8] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.
- [9] C. Li, Z. Zhang, W. Sun Lee, and G. Hee Lee, "Convolutional sequence to sequence model for human dynamics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5226–5234.
- [10] E. Corona, A. Pumarola, G. Alenya, and F. Moreno-Noguer, "Context-aware human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6992–7001.
- [11] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4346–4354.
- [12] D. Pavllo, C. Feichtenhofer, M. Auli, and D. Grangier, "Modeling human motion with quaternion-based neural networks," *International Journal of Computer Vision*, pp. 1–18, 2019.
- [13] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," *Neural Information Processing Systems*, 2015.
- [14] Y. Yuan and K. Kitani, "Dlow: Diversifying latent flows for diverse human motion prediction," in *European Conference on Computer Vision*. Springer, 2020, pp. 346–364.
- [15] S. Ghosh, A. Dhall, M. Hayat, J. Knibbe, and Q. Ji, "Automatic gaze analysis: A survey of deep learning based approaches," *arXiv preprint arXiv:2108.05479*, 2021.
- [16] K. Wang, R. Zhao, H. Su, and Q. Ji, "Generalizing eye tracking with bayesian adversarial learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 907–11 916.
- [17] M. Zhang, K. Teck Ma, J. Hwee Lim, Q. Zhao, and J. Feng, "Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4372–4381.
- [18] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," in *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 2016, pp. 83–90.
- [19] V. Richter, B. Carlmeyer, F. Lier, S. Meyer zu Borgsen, D. Schlangen, F. Kummert, S. Wachsmuth, and B. Wrede, "Are you talking to me? improving the robustness of dialogue systems in a multi party hri scenario by incorporating gaze direction and lip movement of attendees," in *Proceedings of the Fourth International Conference on Human Agent Interaction*, 2016, pp. 43–50.
- [20] A. Shafti, P. Orlov, and A. A. Faisal, "Gaze-based, context-aware robotic system for assisted reaching and grasping," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 863–869.
- [21] T. Fischer, H. J. Chang, and Y. Demiris, "Rt-gene: Real-time eye gaze estimation in natural environments," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 334–352.
- [22] P. Kratzer, S. Bihlmaier, N. B. Midlagajni, R. Prakash, M. Tous-saint, and J. Mainprice, "Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze," *arXiv preprint arXiv:2011.11552*, 2020.
- [23] T. Foulsham, "Eye movements and their functions in everyday tasks," *Eye*, vol. 29, no. 2, pp. 196–199, 2015.
- [24] M. F. Land and M. Hayhoe, "In what ways do eye movements contribute to everyday activities?" *Vision research*, vol. 41, no. 25–26, pp. 3559–3565, 2001.
- [25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [27] K. Rayner, "The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search," *Quarterly journal of experimental psychology*, vol. 62, no. 8, pp. 1457–1506, 2009.
- [28] C. C. Williams, J. M. Henderson *et al.*, "Incidental visual memory for targets and distractors in visual search," *Perception & Psychophysics*, vol. 67, no. 5, pp. 816–827, 2005.
- [29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.