2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

GazeMotion: Gaze-guided Human Motion Forecasting

Zhiming Hu, Syn Schmitt, Daniel Häufle, and Andreas Bulling

***Abstract*— We present *GazeMotion* – a novel method for human motion forecasting that combines information on past human poses with human eye gaze. Inspired by evidence from behavioural sciences showing that human eye and body movements are closely coordinated, GazeMotion first predicts future eye gaze from past gaze, then fuses predicted future gaze and past poses into a gaze-pose graph, and finally uses a residual graph convolutional network to forecast body motion. We extensively evaluate our method on the MoGaze, ADT, and GIMO benchmark datasets and show that it outperforms state-of-the-art methods by up to 7.4% improvement in mean per joint position error. Using head direction as a proxy to gaze, our method still achieves an average improvement of 5.5%. We finally report an online user study showing that our method also outperforms prior methods in terms of perceived realism. These results show the significant information content available in eye gaze for human motion forecasting as well as the effectiveness of our method in exploiting this information.**

I. Introduction

Understanding and forecasting human motion – coarse activities such as walking, or fine-grained movements such as reaching or grasping – is a long-standing research challenge in mobile robotics and human-robot interaction [1], [2]. Given the inherent sequential nature of human motion, much previous work on motion forecasting has focused on using recurrent neural networks, showing significant performance improvements [3], [4]. Other methods for human motion forecasting include Transformer-based architectures [5], graph convolutional networks (GCNs) [6], or multi-layer perceptrons (MLPs) [7]. Common to all of these methods is that they formulate motion forecasting as a sequence-to-sequence task in which future motion is predicted *solely* from past human movements or poses [4]–[7].

In a parallel line of work, studies in the cognitive and behavioural sciences have revealed the strong correlation between human eye gaze and body movements during daily activities [8], [9]. For example, when navigating, human visual attention is strongly correlated with the movement direction [10] and it tends to precede corresponding hand and arm movements when reaching for an object [11]. Despite the close coordination between human eye gaze and body movements, this information has only recently started to be explored for human motion forecasting. A method proposed in [12] requires rich information about the full 3D environment and objects therein.

We present *GazeMotion* – the first learning-based method for human motion forecasting that combines information on past human poses with eye gaze without requiring such information. Our method consists of three main components: a convolutional neural network to predict future eye gaze from historical gaze, a pose-gaze graph that fuses pose and gaze data, and a novel residual graph convolutional network consisting of three spatial-temporal modules that forecasts future body poses from the pose-gaze graph. We evaluate our method for motion forecasting at different future time horizons of up to 1 second (future 30 frames) on the MoGaze [13], ADT [14], and GIMO [12] benchmark datasets. We show that our method outperforms several state-of-the-art methods by a large margin: We achieve 7.4% improvement on MoGaze, 6.4% on ADT, and 6.1% on GIMO in terms of mean per joint position error (MPJPE). We further report an online user study that shows that our method outperforms other methods in terms of precision and perceived realism of the predicted human motion. Considering that eye gaze information is not always available in real applications, we further use head direction as a proxy to gaze and show that our method still achieves significantly better performances than prior methods. The full source code and trained models are available at zhiminghu.net/hu24_gazemotion.

The specific contributions of our work are three-fold:

- We propose a novel learning-based method that predicts future eye gaze from past gaze, fuses the predicted future gaze and past poses into a gaze-pose graph, and forecasts future poses through a novel residual graph convolutional network.
- We report extensive experiments on three public datasets for motion forecasting at different future time horizons and demonstrate significant performance improvements over several state-of-the-art methods.
- We conduct an online user study and validate that our method outperforms prior methods in both precision and realism.

II. Related Work

A. Coordination of Eye and Body Movements

The coordination of human eye and body movements has been extensively studied in multiple fields, including cognitive science and human-centred computing. Hu et al. [15]–[17] have demonstrated that head movements are strongly correlated with eye movements in many daily activities,

Zhiming Hu, Syn Schmitt, and Andreas Bulling are with the University of Stuttgart, Germany. E-mail: {zhiming.hu@vis.uni-stuttgart.de, schmitt@simtech.uni-stuttgart.de, andreas.bulling@vis.uni-stuttgart.de}. Daniel Haeufle is with Heidelberg University and University of Tuebingen, Germany. E-mail: daniel.haeufle@ziti.uni-heidelberg.de. Syn Schmitt, Daniel Haeufle, and Andreas Bulling are with the Center for Bionic Intelligence Tuebingen Stuttgart (BITS), Germany. Zhiming Hu is the corresponding author. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016.

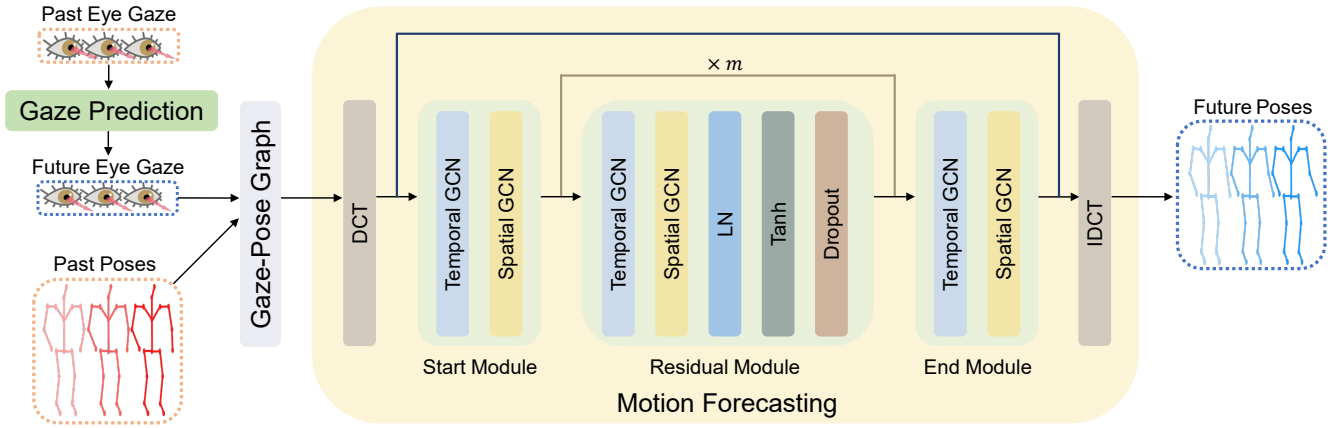

Fig. 1. Our method first forecasts future eye gaze from past gaze using a 1D convolutional neural network, then fuses the predicted gaze and past body poses into a gaze-pose graph, and finally applies a graph convolutional network consisting of a start, a residual, and an end module to forecast body motion.

such as free viewing or searching for an object. Emery et al. [11] investigated the coordination of eye, hand, and head movements in a virtual environment. In this work, we demonstrate that findings on the strong link between human motion and eye gaze can be successfully transferred to the task of motion forecasting and lead to significant performance improvements.

B. Eye Gaze Prediction

Human eye gaze is significant for many important applications including human-robot interaction [1], [2] and human action anticipation [18], [19]. In this background, eye gaze prediction has become a popular research topic in the areas of robotics and human-centred computing. Kim et al. focused on robot manipulation scenarios and proposed a Transformer-based method to predict eye gaze based on sequential visual input [20]. Hu et al. concentrated on a visual search task and used the scene content and task-related information to forecast future eye gaze [16]. Existing methods usually rely on additional information, e.g. image content and task-related variables, making it difficult to apply their methods to other situations. In contrast, our method only employs historical eye gaze to forecast future gaze. In addition, we are the first to combine gaze prediction with motion forecasting, i.e., we first forecast future eye gaze and then use the predicted future gaze to forecast future motion.

III. Method

We define gaze-guided human motion forecasting as the task of predicting future body motion, for a certain time horizon, jointly from past body poses and eye gaze information. Human pose p is represented by the 3D positions of all human joints $p \in R^{3\times n}$, where n is the number of joints. Eye gaze direction is indicated using a unit vector $g \in R^3$. Given a sequence of past human poses $P_{1:t} = \{p_1, p_2, ..., p_t\}$ and human eye gaze $G_{1:t} = \{g_1, g_2, ..., g_t\}$, the task is to forecast human poses in the future $P_{t+1:T} = \{p_{t+1}, p_{t+2}, ..., p_T\}$. Our method consists of three main components: a gaze prediction network that forecasts future eye gaze from past gaze, a gaze-pose fusion procedure that fuses predicted future gaze and past poses into a gaze-pose graph, and a motion forecasting network that forecasts future motions from the fused gaze and motion data (see Figure 1 for an overview of our method).

A. Gaze Prediction

Guided by the intuition that information on future eye gaze is more useful for motion forecasting than past eye gaze, we first forecast future gaze $G_{t+1:2t}$ from past gaze $G_{1:t}$ and then use the predicted future gaze to forecast future body motion. In light of the good performance of 1D CNN for processing eye gaze data [16], [21], we employed four 1D CNN layers to forecast eye gaze. More specifically, we used three 1D CNN layers, each with 32 channels and a kernel size of three, to extract features from the historical eye gaze data. Each CNN layer was followed by a layer normalisation (LN) and a Tanh activation function. After the three CNN layers, we used a 1D CNN layer with three channels and a kernel size of three, and a Tanh activation function to predict future eye gaze from the gaze features. The predicted gaze directions were finally normalised to unit vectors: $\hat{G}_{t+1:2t} = \{\hat{g}_{t+1}, \hat{g}_{t+2}, ..., \hat{g}_{2t}\} \in R^{3\times t}$.

B. Gaze-Pose Fusion

We first padded the past poses and predicted gaze up to T by repeating the last pose and gaze for $T - t$ times following prior works [6], [22]. Each pose was represented using the 3D coordinates of all the human joints: $P_{1:T} = \{J^1_{1:T}, J^2_{1:T}, ..., J^n_{1:T}\} \in R^{3\times n\times T}$. We further cloned the predicted eye gaze data $\hat{G}$ for $n - 1$ times to make the gaze data have the same size as the motion data. We then concatenated the motion data and gaze data along the spatial dimension and obtained $X \in R^{3\times 2n\times T}$. We modelled the fused motion and gaze data X as fully-connected spatial and temporal graphs, jointly called the *gaze-pose graph*, to learn the relationship between human eye gaze and human motion. The spatial graph represents each joint J^k and each gaze $\hat{G}$ as a separate node and thus contains $2n$ nodes: n joint

nodes and n gaze nodes. The temporal graph consists of T nodes, corresponding to the fused data at different times: $X_1, X_2, ..., X_T$. The spatial and temporal graphs are both fully-connected with their adjacency matrices measuring the weights between each pair of nodes.

C. Motion Forecasting

We first applied discrete cosine transform (DCT) [6], [7] to encode the fused data X in the temporal domain using DCT matrix $M_{dct} \in R^{T\times T}$:

$$X_d = XM_{dct}. \tag{1}$$

We then proposed three novel GCN modules, i.e., a start module that mapped $X_d \in R^{3\times 2n\times T}$ into feature space, a residual module that extracted the features, and an end module that mapped the features to the original data space.

Start Module: The start module first used a temporal GCN to extract the temporal features from the transformed data. To this end, the temporal GCN learned the weighted adjacency matrix $A^T \in R^{T\times T}$ of the fully-connected temporal graph and calculated temporal convolution using

$$X_d^1 = X_d A^T. \tag{2}$$

$X_d^1 \in R^{3\times 2n\times T}$ was then permuted to $X_d^2 \in R^{T\times 2n\times 3}$. A weight matrix $W^{start} \in R^{3\times 16}$ was used to convert the input node features (3 dimensions) to latent features (16 dimensions):

$$X_d^3 = X_d^2 W^{start}. \tag{3}$$

After the weight matrix, a spatial GCN was employed to extract the spatial features. It learned the weighted adjacency matrix $A^S \in R^{2n\times 2n}$ of the fully-connected spatial graph and performed spatial convolution using

$$X_d^4 = A^S X_d^3. \tag{4}$$

$X_d^4 \in R^{T\times 2n\times 16}$ was further permuted to $X_d^5 \in R^{16\times 2n\times T}$. Considering that repeating important features is beneficial for motion forecasting [6], [22], we copied the output of the start module along the temporal dimension ($R^{16\times 2n\times T} \rightarrow R^{16\times 2n\times 2T}$) and used it as input to the residual module.

Residual Module: The residual module consisted of m GCN blocks with each block containing a temporal GCN that learned the temporal adjacency matrix $A_i^T \in R^{2T\times 2T}$, a weight matrix $W_i^{res} \in R^{16\times 16}$ that extracted the latent features, a spatial GCN that learned the spatial adjacency matrix $A_i^S \in R^{2n\times 2n}$, a layer normalisation, a Tanh activation function, and a dropout layer with dropout rate 0.3 to prevent overfitting. m was set to 16 and a residual connection was added for each GCN block. We cut the output of the residual module in half in the temporal dimension ($R^{16\times 2n\times 2T} \rightarrow R^{16\times 2n\times T}$) and input it to the end module.

End Module: The end module consisted of a temporal GCN that learned the temporal adjacency matrix, a weight matrix $W^{end} \in R^{16\times 3}$ that mapped the latent features to 3 dimensions, and a spatial GCN that learned the spatial adjacency matrix. We added a global residual connection to improve the network flow. The output of the end module $Y_d \in R^{3\times 2n\times T}$ was converted back to the original representation space using an inverse discrete cosine transform (IDCT) matrix $M_{idct} \in R^{T\times T}$:

$$Y = Y_d M_{idct}. \tag{5}$$

The predicted future poses $\hat{P}_{t+1:T} \in R^{3\times n\times T-t}$ were obtained from the joint nodes in $Y \in R^{3\times 2n\times T}$.

D. Loss Function

We trained the gaze prediction and motion forecasting networks separately. For the gaze prediction network, we used the angular error between the predicted eye gaze direction $\hat{G}$ and the ground truth gaze direction G as the loss function:

$$\ell = \arccos(\hat{G} \cdot G). \tag{6}$$

For the motion forecasting network, we used a combination of motion loss ℓ_m and velocity loss ℓ_v as our loss function ℓ:

$$\ell = \ell_m + \ell_v. \tag{7}$$

ℓ_m measures the mean per joint position error between the predicted future poses and the ground truth [6]:

$$\ell_m = \frac{1}{n(T-t)} \sum_{j=t+1}^{T} \sum_{k=1}^{n} \|\hat{p}_{j,k} - p_{j,k}\|^2, \tag{8}$$

where $\hat{p}_{j,k} \in R^3$ represents the 3D coordinates of the k^{th} joint at the future time of j while $p_{j,k} \in R^3$ is the corresponding ground truth. ℓ_v measures the mean per joint velocity error between the predicted future poses and the ground truth:

$$\ell_v = \frac{1}{n(T-t-1)} \sum_{j=t+1}^{T-1} \sum_{k=1}^{n} \|\hat{v}_{j,k} - v_{j,k}\|^2, \tag{9}$$

where $\hat{v}_{j,k} \in R^3$ represents the velocity of the k^{th} joint at the future time of j while $v_{j,k} \in R^3$ is the corresponding ground truth. The velocity is computed using the time difference: $\hat{v}_{j,k} = \hat{p}_{j+1,k} - \hat{p}_{j,k}$ and $v_{j,k} = p_{j+1,k} - p_{j,k}$.

IV. Experiments and Results

A. Datasets

a) MoGaze dataset [13]: The MoGaze dataset contains human motion and eye gaze data recorded at 120 Hz from six people performing *pick* and *place* actions. We down-sampled the human pose and gaze data to 30 Hz for simplicity [6], [7] and represented human poses using the 3D coordinates of 21 human joints. We used a leave-one-person-out cross-validation: We trained on five participants and tested on the remaining one, repeated the experiment six times with a different participant for testing, and calculated the average performance across all six iterations.

b) ADT dataset [14]: The ADT dataset contains 35 sequences of human pose and gaze data performing various indoor activities including *room decoration*, *meal preparation*, and *work*. Each human pose consists of the 3D coordinates of 21 human joints recorded at 30 Hz. For experiments on ADT, we randomly selected 25 sequences for training and 10 sequences for testing.

c) GIMO dataset [12]: The GIMO dataset is collected in various indoor scenes and contains motion and gaze data of 11 participants performing daily activities. Each human pose consists of the 3D coordinates of 23 human joints recorded at 30 Hz. We used the default train/test split [12]: motion and gaze data from 12 scenes were used for training and data from 14 scenes (12 known scenes from the training set and two new environments) for testing.

B. Evaluation Settings

a) Evaluation Metric: As is common in human motion forecasting [6], [7], [12], we used the mean per joint position error (see Equation 8) in millimeters as our metric to evaluate the different motion forecasting methods.

b) Baselines: We compared our method with the following state-of-the-art baseline methods for motion forecasting:

- *Res-RNN* [4]: *Res-RNN* is a RNN-based method that applies a residual connection between the input pose and output pose to improve performance.
- *siMLPe* [7]: *siMLPe* is a light-weight MLP-based method that applies discrete cosine transform and residual connections to improve performance.
- *HisRep* [5]: *HisRep* is a Transformer-based method that extracts motion attention to capture the similarity between the current motion context and the historical motion subsequences.
- *PGBIG* [6]: *PGBIG* is a GCN-based method that employs a multi-stage framework to forecast human motions where each stage predicts an initial guess for the next stage.

c) Time Horizons of Input and Output Sequences: For experiments on the MoGaze, ADT, and GIMO datasets (30 Hz), we used 10 frames of data as input to forecast human poses in the future 30 frames (i.e., up to one second into the future), following the common evaluation settings [5], [6].

d) Implementation Details: We trained the baseline methods from scratch using their default parameters. To train our gaze prediction network, we used the Adam optimiser with an initial learning rate of 0.01 that we then decayed by 0.9 every epoch. We used a batch size of 32 to train the gaze prediction network for a total of 50 epochs. For our motion forecasting network, the Adam optimiser with an initial learning rate of 0.01 was used and the learning rate was decayed by 0.95 every epoch. A batch size of 32 was employed to train the motion forecasting network for 100 epochs. Our method was implemented using the PyTorch framework.

C. Motion Forecasting Results

Results on MoGaze: Table I summarises the performances of different methods on the MoGaze dataset. The table shows the average MPJPE error (in millimeters) over all 30 frames as well as the prediction errors for different future time horizons: 200 ms, 400 ms, …, 1000 ms. As can be seen from the table, our method consistently outperforms the state-of-the-art methods at different future time intervals, achieving an average improvement of 7.4% (75.9 *vs.* 82.0) over the state of the art. A paired Wilcoxon signed-rank test was used to compare the performances of our method with the state of the art and the results validated that the differences between our method and the state of the art are statistically significant ($p < 0.01$). Figure 2 shows an example of the predicted poses from different methods. We can see that our method achieves significantly better performances than other methods. See supplementary video for more prediction results.

TABLE I

MPJPE ERRORS (UNIT: MILLIMETERS) OF DIFFERENT METHODS FOR MOTION FORECASTING ON THE MOGAZE, ADT AND GIMO DATASETS. BEST RESULTS ARE IN BOLD.

Dataset	Method	200 ms	400 ms	600 ms	800 ms	1000 ms	Average
MoGaze	*Res-RNN* [4]	53.1	91.3	136.8	187.5	240.8	124.3
	siMLPe [7]	40.6	72.0	108.8	152.6	201.0	99.5
	HisRep [5]	31.4	60.5	95.4	135.3	177.9	85.3
	PGBIG [6]	29.4	57.7	92.0	130.7	171.5	82.0
	Ours *w/o gaze*	27.2	55.3	88.9	126.9	167.1	79.0
	Ours	**25.8**	**53.3**	**85.8**	**122.0**	**160.0**	**75.9**
ADT	*Res-RNN* [4]	35.6	55.7	77.8	100.0	122.5	70.1
	siMLPe [7]	29.9	48.3	69.1	93.8	120.7	63.8
	HisRep [5]	15.5	30.5	47.6	66.8	88.2	42.3
	PGBIG [6]	14.5	28.7	45.4	64.4	85.8	40.6
	Ours w/o *gaze*	12.0	26.6	44.0	63.8	85.3	39.1
	Ours	**11.7**	**25.8**	**42.8**	**62.1**	**82.8**	**38.0**
GIMO	*Res-RNN* [4]	82.6	126.4	170.2	212.9	255.4	152.8
	siMLPe [7]	42.8	78.3	114.6	150.7	188.5	100.3
	HisRep [5]	41.8	78.1	115.0	152.7	192.4	100.2
	PGBIG [6]	38.0	68.6	101.9	136.1	172.2	89.2
	Ours w/o *gaze*	33.7	66.1	99.7	134.4	170.4	86.8
	Ours	**32.6**	**64.1**	**97.0**	**130.0**	**162.4**	**83.8**

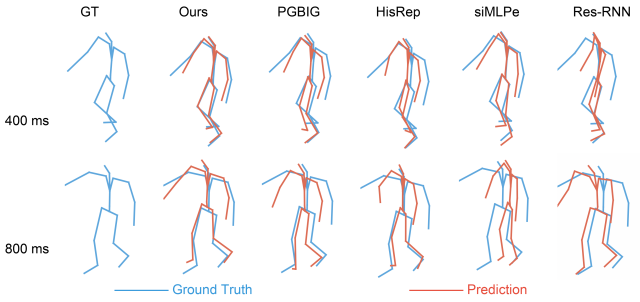

Fig. 2. Visualisation of the predictions of different methods on MoGaze [13]. Our method consistently outperforms other methods when predicting 400ms and 800ms into the future.

Results on ADT: Table I shows the MPJPE errors of different methods for different future time horizons as well as the average error. We can see that our method significantly outperforms prior methods (paired Wilcoxon signed-rank test, $p < 0.01$), achieving an overall average improvement of 6.4% (38.0 *vs.* 40.6).

Results on GIMO: We can see from Table I that our method outperforms state-of-the-art methods with an overall average improvement of 6.1% (83.8 *vs.* 89.2) and our improvements are statistically significant (paired Wilcoxon signed-rank test, $p < 0.01$).

TABLE II
ABLATION STUDY ON THE MOGAZE DATASET.

Method	200 ms	400 ms	600 ms	800 ms	1000 ms	Average
w/o *spatial GCN*	30.9	62.1	96.3	133.8	173.1	84.7
w/o *temporal GCN*	46.6	74.0	107.9	147.0	188.0	99.3
w/o *copy*	26.8	54.5	87.1	123.2	161.0	77.0
w/o *global residual*	32.9	58.7	90.7	127.0	165.0	81.9
w/o *velocity loss*	26.3	53.9	86.6	122.9	161.1	76.6
w/o gaze	27.2	55.3	88.9	126.9	167.1	79.0
past gaze	26.3	54.3	87.2	123.8	162.0	77.1
past head	26.3	54.4	87.9	125.4	164.1	77.8
future head	26.1	54.1	87.5	124.8	163.7	77.5
Ours	**25.8**	**53.3**	**85.8**	**122.0**	**160.0**	**75.9**

D. Head Direction as a Proxy to Eye Gaze

Our method requires eye gaze as input but gaze may not always be available, thus limiting the range of our method's possible applications. To increase the usability of our method, we propose to use head direction (forward direction of head) as a proxy to eye gaze, inspired by the strong correlations between eye and head movements [15], [23]. Specifically, we first predicted head direction from past head direction, then fused the predicted future head direction and past body poses, and finally applied our motion forecasting network to predict future motions. Our method using head direction achieves an average improvement of 5.5% (77.5 *vs.* 82.0), 5.2% (38.5 *vs.* 40.6), and 3.8% (85.8 *vs.* 89.2) over the state of the art on MoGaze, ADT, and GIMO, respectively. These results demonstrate that our method can use head direction as a proxy to eye gaze and still obtain significantly better performances than the state of the art.

E. Ablation Study

a) Effectiveness of Eye Gaze for Motion Forecasting: We tested different ways of using eye gaze information, i.e., 1) don't use eye gaze (*w/o gaze*), 2) use past head direction in the gaze-pose fusion procedure (*past head*), 3) use predicted future head direction (*future head*), 4) use past eye gaze (*past gaze*), and 5) use predicted future eye gaze (Ours). Table II shows the motion forecasting results of different ways of using eye gaze on the MoGaze dataset. As can be seen from the table, using eye gaze or head direction as a proxy to eye gaze achieves significantly better performances than not using eye gaze (paired Wilcoxon signed-rank test, $p < 0.01$), validating that eye gaze information can help improve the performance of motion forecasting. We also find that using predicted future eye gaze or head direction achieves better performances than directly using past eye gaze (75.9 *vs.* 77.1) or head direction (77.5 *vs.* 77.8), revealing the significant potential of applying gaze prediction methods to improve the performances of motion forecasting.

b) Effectiveness of Our GCN Architecture: From Table I we can see that even without using eye gaze, our method still significantly outperforms the state of the art (79.0 *vs.* 82.0 on MoGaze, 39.1 *vs.* 40.6 on ADT, 86.8 *vs.* 89.2 on GIMO, paired Wilcoxon signed-rank test, $p < 0.01$), validating the effectiveness of our GCN architecture. Furthermore, we respectively removed *spatial GCN*, *temporal GCN*, *copy* in start module, *global residual*, and *velocity loss*, and retrained the ablated methods. We can see from Table II that each component helps improve our method's motion forecasting performance. We also find that *temporal GCN* is most important to the success of our method, revealing the significance of temporal features for motion forecasting.

c) Training Parameters: We used different dropout rate to train our model and the performances of dropout rate 0.1, 0.2, 0.3 (Ours), and 0.4 on MoGaze are 77.7, 76.9, 75.9, and 76.5, respectively. We also added a weighting constant for ℓ_v in Equation 7 and the performances of weighting constant 0.25, 0.5, 1.0 (Ours), and 1.5 on MoGaze are 76.0, 76.5, 75.9, and 76.3, respectively. These results validate that our training parameters are optimal in practice.

F. User Study

To further evaluate whether our method's improvements are significant in terms of qualitative evaluation, we conducted an online user study to compare our method with prior methods.

1) Stimuli: We randomly selected 24 motion forecasting samples from the MoGaze, ADT, and GIMO datasets (8 samples from each dataset) and used them as our stimuli. Each sample consisted of 30 frames of predictions (corresponding to future 1 second) and was visualised as a short video.

2) Participants: We recruited 20 participants (12 males and 8 females, aged between 21 and 36 years) to take part in our user study through university mailing lists and social networks. All of the participants reported normal or corrected-to-normal vision. The user study was approved by our university's ethical review board.

3) Procedure: We conducted our user study using a Google form. During the study, the ground truth future motions and the predictions of different methods were displayed to the participants in parallel using a layout that is similar to Figure 2. The names of different methods were hidden and the order of these methods were randomised. The visualisation videos of the ground truth and different methods were set to loop automatically, allowing participants to observe them with no time limit. During their observation, participants were required to rank different methods according to two criteria: *precision* and *realism*.

- *Precision*: check different methods to see whether they *align with the ground truth* and rank them based on your observation.
- *Realism*: check different methods to see whether they are *physically plausible* and rank them based on your observation.

We collected the participants' responses for further analysis.

4) Statistical Analysis: The means and standard deviations (SDs) of different methods' rankings are shown in Table III. We can see that our method outperforms the state of the art in terms of both precision (1.6 *vs.* 3.2) and realism (1.9 *vs.* 3.1) and the results are statistically significant (paired Wilcoxon signed-rank test, $p < 0.01$). The above results demonstrate that our method achieves significantly better performances over prior methods in qualitative evaluation.

TABLE III

STATISTICAL RESULTS OF DIFFERENT METHODS' RANKINGS IN OUR USER STUDY.

		Ours	*PGBIG* [6]	*HisRep* [5]	*siMLPe* [7]	*Res-RNN* [4]
Precision	Mean	**1.6**	3.2	3.2	3.3	3.7
	SD	0.9	1.2	1.2	1.3	1.3
Realism	Mean	**1.9**	3.3	3.1	3.3	3.5
	SD	1.3	1.2	1.3	1.3	1.4

G. Discussion

From the results in Table I, we find that the performances of all the methods deteriorate significantly with the increase of prediction time. This is a well-known problem in motion forecasting [6], [7] that all current methods suffer from since these methods only use historical motion information. Integrating more context information such as user's goal or task into motion forecasting has the potential to alleviate this problem. In addition, we only explored the effectiveness of eye gaze and head direction on motion forecasting but ignored other important body signals such as hand gestures or gait. Integrating such body signals into our pipeline to further improve the performance is an interesting avenue of future work. Furthermore, we are also looking forward to incorporating stochasticity into the human motion forecasting model to further improve the performance.

V. CONCLUSION

In this work we proposed a novel method for human motion forecasting that first predicts future eye gaze from past gaze, fuses future eye gaze and past body poses into a gaze-pose graph, and finally uses a spatio-temporal residual GCN. Through extensive experiments on three public benchmark datasets we showed that our method outperforms several state-of-the-art methods by a large margin. We also validated that our predictions are more precise and more realistic than prior methods through an online user study. We further showed that head direction can be a suitable proxy to eye gaze for use cases where eye gaze is not available, thereby further improving the applicability of our method. As such, our work reveals the significant information content available in eye gaze for human motion forecasting and paves the way for future research on this promising research direction.

REFERENCES

[1] A. Belardinelli, A. R. Kondapally, D. Ruiken, D. Tanneberg, and T. Watabe, "Intention estimation from gaze and motion features for human-robot shared-control object manipulation," in *Proceedings of the 2022 IEEE International Conference on Intelligent Robots and Systems*. IEEE, 2022, pp. 9806–9813.

[2] L. Shi, C. Copot, and S. Vanlanduit, "Gazeemd: Detecting visual intention in gaze-based human-robot interaction," *Robotics*, vol. 10, no. 2, p. 68, 2021.

[3] A. T. Le, P. Kratzer, S. Hagenmayer, M. Toussaint, and J. Mainprice, "Hierarchical human-motion prediction and logic-geometric programming for minimal interference human-robot tasks," in *Proceedings of the 2021 IEEE International Conference on Robot and Human Interactive Communication*. IEEE, 2021, pp. 7–14.

[4] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2891–2900.

[5] W. Mao, M. Liu, M. Salzmann, and H. Li, "Multi-level motion attention for human motion prediction," *International Journal of Computer Vision*, vol. 129, no. 9, pp. 2513–2535, 2021.

[6] T. Ma, Y. Nie, C. Long, Q. Zhang, and G. Li, "Progressively generating better initial guesses towards next stages for high-quality human motion prediction," in *Proceedings of the 2022 IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6437–6446.

[7] W. Guo, Y. Du, X. Shen, V. Lepetit, X. Alameda-Pineda, and F. Moreno-Noguer, "Back to mlp: A simple baseline for human motion prediction," in *Proceedings of the 2023 IEEE Winter Conference on Applications of Computer Vision*, 2023, pp. 4809–4819.

[8] E. G. Freedman, "Coordination of the eyes and head during visual orienting," *Experimental brain research*, vol. 190, pp. 369–387, 2008.

[9] H. H. Goossens and A. V. Opstal, "Human eye-head coordination in two dimensions under different sensorimotor conditions," *Experimental Brain Research*, vol. 114, pp. 542–560, 1997.

[10] Q. Sun, A. Patney, L.-Y. Wei, O. Shapira, J. Lu, P. Asente, S. Zhu, M. McGuire, D. Luebke, and A. Kaufman, "Towards virtual reality infinite walking: dynamic saccadic redirection," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–13, 2018.

[11] K. J. Emery, M. Zannoli, J. Warren, L. Xiao, and S. S. Talathi, "Openneeds: A dataset of gaze, head, hand, and scene signals during exploration in open-ended vr environments," in *Proceedings of the 2021 ACM Symposium on Eye Tracking Research and Applications*, 2021, pp. 1–7.

[12] Y. Zheng, Y. Yang, K. Mo, J. Li, T. Yu, Y. Liu, K. Liu, and L. J. Guibas, "Gimo: Gaze-informed human motion prediction in context," in *Proceedings of the 2022 European Conference on Computer Vision*, 2022.

[13] P. Kratzer, S. Bihlmaier, N. B. Midlagajni, R. Prakash, M. Toussaint, and J. Mainprice, "Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 367–373, 2020.

[14] X. Pan, N. Charron, Y. Yang, S. Peters, T. Whelan, C. Kong, O. Parkhi, R. Newcombe, and Y. C. Ren, "Aria digital twin: A new benchmark dataset for egocentric 3d machine perception," in *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023, pp. 20 133–20 143.

[15] Z. Hu, C. Zhang, S. Li, G. Wang, and D. Manocha, "Sgaze: a data-driven eye-head coordination model for realtime gaze prediction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 5, pp. 2002–2010, 2019.

[16] Z. Hu, A. Bulling, S. Li, and G. Wang, "Fixationnet: forecasting eye fixations in task-oriented virtual environments," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 5, pp. 2681–2690, 2021.

[17] Z. Hu, J. Xu, S. Schmitt, and A. Bulling, "Pose2gaze: Eye-body coordination during daily activities for gaze prediction from full-body poses," *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[18] N. F. Duarte, M. Raković, J. Tasevski, M. I. Coco, A. Billard, and J. Santos-Victor, "Action anticipation: Reading the intentions of humans and robots," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4132–4139, 2018.

[19] Z. Hu, A. Bulling, S. Li, and G. Wang, "Ehtask: Recognizing user tasks from eye and head movements in immersive virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[20] 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)*. IEEE, 2022, pp. 2427–2433.

[21] Z. Hu, S. Li, C. Zhang, K. Yi, G. Wang, and D. Manocha, "Dgaze: Cnn-based gaze prediction in dynamic scenes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 5, pp. 1902–1911, 2020.

[22] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proceedings of the 2019 IEEE International Conference on Computer Vision*, 2019, pp. 9489–9497.

[23] L. Sidenmark and H. Gellersen, "Eye, head and torso coordination during gaze shifts in virtual reality," *ACM Transactions on Computer-Human Interaction*, vol. 27, no. 1, pp. 1–40, 2019.