Predicting Locomotion Intention using Eye Movements and EEG with LSTM and Transformers

Gianni Bremer* Markus Lappe†

Institute for Psychology
University of Muenster

Abstract

Predicting future locomotion based on intrinsic data serves many purposes, including optimizing the utilization of physical space in virtual reality environments and enhancing the control of electronic aids for patients with motor impairments. However, predicting human locomotion intentions proves challenging due to the inherent difficulty arising from the highly complex and nonlinear interactions among the relevant parameters. Deep neural networks offer a significant advantage over conventional approaches in addressing this challenge. We treat this task as a time series prediction problem and compare LSTM networks to transformer models. A distinctive aspect of our work is our approach's emphasis on eye movements as a central feature, contributing to its novel predictive capabilities. Besides gaze data, we evaluate the addition of EEG as a data source for this prediction task to be used in brain-computer interfaces. To achieve this, we conducted two data collection experiments in custom virtual environments that feature different tasks utilizing joystick control. We present these novel datasets in conjunction with this work. The results demonstrate that gaze data proves to be a valuable tool for locomotion prediction in different contexts, even when there is not a strong and direct connection between gaze and future waypoints. Transformer models were able to achieve better performance than LSTM networks, and we conclude that successful prediction across diverse situations requires datasets containing a wide range of movement scenarios.

Index Terms: Virtual Reality, Eye Tracking, Eye-Tracking, Locomotion, LSTM, Transformer, Path Prediction, Machine Learning, Deep Learning, Gaze.

1 Introduction

In the extended reality (XR) domain, the convergence of human-computer interaction and immersive environments has opened up new avenues for user engagement and exploration. One of the key challenges in this context is understanding and predicting a user's future movements. In Virtual Reality (VR), successfully anticipating and interpreting user movements in the virtual realm holds significant promise for improving VR experiences and developing more intuitive and responsive systems. In Augmented Reality (AR), the ability to predict the locomotion intentions of a user can be helpful for control interfaces of machinery or robotic aids. For example, an intriguing opportunity arises from incorporating a predictive model into the control system of an electronic wheelchair for patients with motor impairments. In such a device, prediction of the user's intention could be used to move the wheelchair in the respective direction and AR could play a crucial role as a visualization interface, presenting real-time predictions directly in the user's field of view, providing users with a transparent and immediate understanding of the wheelchair's anticipated movements. The interplay between predictions and AR visualization establishes a symbiotic loop: the user's cognitive intentions guide the wheelchair's movements, and the resulting actions are seamlessly displayed in the augmented environment. This integration not only holds promise for improving wheelchair navigation but also opens up new possibilities for inclusive and adaptive technologies. Prioritizing user comfort, autonomy, and safety in various mobility scenarios becomes a focal point.

A contemporary paradigm in trajectory prediction involves the utilization of artificial neural networks. In the area of time series prediction, a particular focus lies on recurrent neural networks and transformer neural networks. In the domain of human motion prediction, recurrent neural networks, such as Long Short-Term Memory networks (LSTMs), have gained prominence [43, 13, 57, 46, 8]. LSTMs, introduced by Hochreiter and Schmidhuber [22], exhibit the ability to retain and selectively forget information over extended sequences, making them particularly adept at modeling sequential dependencies. This ability to manage long-range dependencies makes LSTMs well-suited for tasks where understanding temporal patterns is essential, such as human motion prediction. Applied to predicting user positions in VR scenarios [11] and controlling redirected walking [37], LSTMs have showcased their versatility in capturing intricate spatial and orientational dynamics within dynamic environments.

Transformer networks [59], in contrast, focus on attention mechanisms rather than iterative memory processes. Transformers capture global dependencies within sequences by dynamically attending to relevant parts of the input sequence. This attention-based approach allows transformers to efficiently model relationships between distant elements in a sequence, thereby offering advantages in parallelization and scalability. Among many other areas of application, these powerful networks have also achieved success in the field of human trajectory prediction [16, 61]. While these networks have not been used for locomotion prediction in VR specifically, transformers have been successfully applied to forecast motion and behaviour in VR [10, 40]

For such time series prediction of locomotion the history of locomotor data, i.e., prior position or velocities, are the most important input feature, as human movements are typically smooth. In addition, gaze plays a pivotal role in motor action, as our eyes naturally gravitate towards targets of interest to gather essential visual information for effective action control [35]. Given that eye movements often precede other motor actions [36, 20], they serve as valuable indicators for predicting action intentions [2, 17, 65, 66]. Walkers, for instance, consistently align their gaze with a target just before approaching it [25, 14].

However, gaze behavior is multifaceted, extending beyond a singular focus on ultimate targets. Walkers also direct their gaze towards obstacles, and when navigating uneven terrain, they frequently glance at the ground a few steps ahead to ensure secure foot placements [24, 23, 9, 56, 44]. While gaze behavior may not exclusively pinpoint the ultimate target, it provides valuable insights into

*e-mail: gianni.bremer@uni-muenster.com
†e-mail: mlappe@uni-muenster.com

waypoints that walkers anticipate using on a short timescale, typically within the next few steps. One of the aims of the current work is to see whether gaze can be a useful feature even in situations in which users are actively scanning their environment and often look away from the target of their locomotion. To do this, we compare a navigational task with a visual search task in the scene.

Additionally, eye movements are intricately linked to changes in direction, with walkers adjusting their gaze inward during curved trajectories [19, 28]. Eye movements also contribute to decision-making processes, including choosing between alternative targets and searching for targets amid distractors [65, 62, 32], showcasing the adaptability of gaze behavior to varying task demands [58].

In essence, while gaze offers valuable information about future actions, leveraging this information to predict future locomotion behavior poses a complex challenge. Deep learning models emerge as promising tools for unraveling user intentions and forecasting locomotion directions, given the intricate interplay of gaze with dynamic environmental cues and task-specific demands.

Electroencephalography (EEG) has emerged as a valuable tool for investigating the neural correlates of locomotion and understanding the intricate interplay between the brain and the control of movement. Research has delved into the connection between EEG signals and locomotor behavior, shedding light on the neural processes underlying human mobility [18, 30].

EEG provides insights into cognitive functions associated with locomotion. Studies have demonstrated that distinct EEG patterns are associated with different phases of walking, such as the initiation, execution, and termination of gait cycles[60]. EEG signals correlate with alterations in walking speed, direction, and obstacle avoidance [33]. These patterns offer a window into the neural mechanisms governing motor planning, coordination, and execution during ambulation.

1.1 Related Work

Various algorithms utilized gaze data in the prediction of different locomotion tasks in VR [65, 17, 21]. Some contemporary approaches employ deep learning models for predicting future positions and have shown promise in diverse contexts, such as public pedestrian traffic [1, 64] and the prediction of future gaze directions [15, 63, 26, 27, 12]. Redirected Walking is a classic paradigm in VR, where the displayed path is different from the actual path [49]. When a virtual reality user encounters the boundaries of their physical environment, it can disrupt the immersive experience. To address this issue, redirected walking methods aim to guide the users along a different walking path by altering the virtual environment causing users to unknowingly adjust their direction and speed of movement. Robust path prediction can be very helpful here as it allows planning and applying the manipulation in advance. Deep Learning models for locomotion prediction in redirected walking have gained recent success, showing that locomotion prediction in VR is a worthwhile endeavor [11, 54, 7, 29, 45]. Gaze can be a valuable feature here [5, 54, 6, 7]. Even for the objective of locomotion intention prediction specifically, gaze has been proven to be a valuable feature for algorithms [41]. Another application for motion prediction in VR is gesture classification. Gesture classification uses head-mounted displays (HMDs) to recognize body actions and head gestures [69, 68, 67, 38].

Transformer networks have been successfully employed for human trajectory forecasting in different context, especially pedestrian behaviour [16, 64]. Alternatively transformer networks and LSTMs can be combined [61].

EEG data based locomotion prediction has been used to construct brain-computer interfaces (BCIs) for assistive devices, especially wheelchairs [48]. Deep Learning architectures like LSTMs have been successfully employed here [42]. EEG data has not been used for locomotion prediction in VR. However, Kritikos et al. have used LSTM models to forecast arm movements in VR [34].

1.2 Aim of this work

The objective of this study is to forecast individuals' intended locomotion intention. The research will specifically concentrate on sequence-to-sequence long-term prediction, recognizing that the navigational success of a control system is contingent on user feedback grounded in long-term prediction. We disregard lower limb activities, emphasizing an immersive perspective facilitated by data acquired through a Virtual Reality (VR) headset. This methodology transcends conventional screen-based applications, such as video games, by closely aligning with sensory-motor loops pertinent to vehicular operations, including cars or wheelchairs.

We will evaluate four key questions:

1. The first aspect of our investigation involves assessing the potential enhancement of deep neural networks through the incorporation of eye-tracking and EEG data as additional features. In the context of wheelchair control, predictive models leveraging eye-tracking and EEG data offer unique value, particularly for individuals with limited motor abilities. This evaluation focuses on improving locomotor behavior predictions and navigation target forecasts, aiming to contribute to advancements in assistive technologies.

2. For the second question, we undertake a comparative evaluation of two distinct datasets. The first dataset promotes a varied spectrum of locomotion and eye movements, allowing users to freely select locomotion targets. In contrast, the second dataset imposes stricter constraints on locomotion by providing predetermined paths and, simultaneously, adds a concurrent visual search task that prompts eye movements towards objects in scene that are not related to the locomotion. Our objective is to evaluate these datasets on both performance and generalizability metrics. By assessing how well the predictive models perform on each dataset and how effectively they generalize to the other scenarios, we aim to discern the impact of dataset characteristics on the overall robustness and applicability of the models.

3. The third aspect of this work builds upon the evaluation new features and of these two distinct datasets. We want to compare the value of additional eye-tracking features for our model in the two datasets separately. While the first tasks allows for relatively unconstrained eye movements that should be tightly coupled with intended locomotion goals, the second dataset forces the users to redirect their gaze and attention towards a non-locomotor task.

4. For the last aspect, we plan to evaluate the Long Short-Term Memory (LSTM) and transformer models, both in relation to each other and against simpler prediction approaches. This comparative analysis is essential for identifying potential advantages inherent in different architectural designs. The decision to explore LSTM and transformer models is motivated by their prominence in sequence-to-sequence time series prediction tasks, aligning with our primary objective of predicting individuals' intended movements

2 Data Acquisition

Two data collection experiments were conducted in order to acquire a suitable amount of training data for specific locomotor situations. Both datasets used for this work were obtained from Virtual Reality joystick experiments conducted together. All raw data files are freely available online. The forest experiment can be found at `https://osf.io/ney6v/`. The course and visual search double task experiment can be found at `https://osf.io/4g9pw/`.

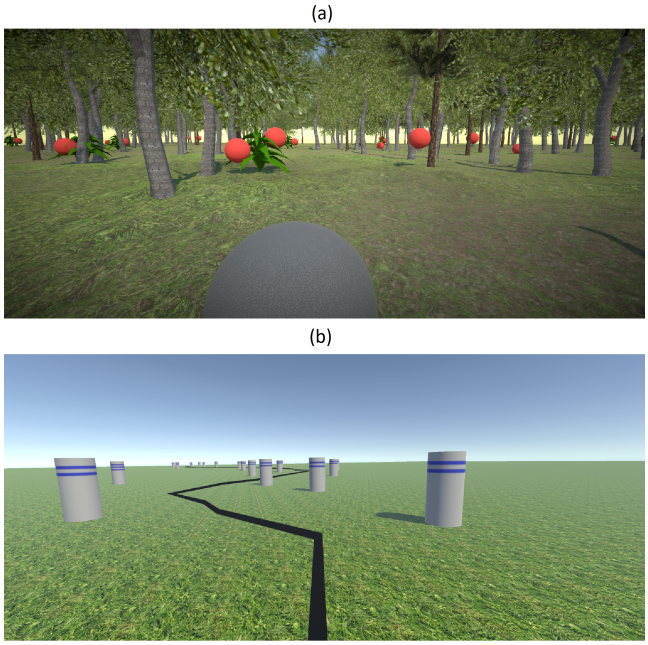

Figure 1: Screenshots from the virtual environments of the data collection experiments (a) The forest experiment (b) The course and visual search double task experiment

2.1 Participants

Twenty healthy participants (twelve female) completed the experiment. Their age ranged from 18 to 50 years (M = 24.4, SD = 7.09). The experimental procedures were approved by the Ethics Committee of the University of Muenster and all participants gave informed written consent. Four participants were left-handed and everybody had normal or corrected to normal vision. The participants were naïve to the intents of both experiments. Participants could be rewarded either by course credit or money (€10 per hour).

2.2 Materials

The virtual environment was presented in an HTC Vive Pro Eye HMD with a resolution of 1440×1600 pixels per eye, a frame rate of 90 Hz and a field of view of 110 degrees. Six Vive Lighthouses 2.0 were used to create a tracking area of 6×11 m. The virtual environment was built with Unity3D and was running with an Intel Core i9 processor and an NVIDIA GTX2080 graphics card.

The EEG was a 32-channel wet electrode mobile and wireless EEG produced by Brainproducts. An Easycap recording cap with a 1020 montage was used.

2.2.1 Eye Tracking Quality

For the Vive Pro Eye, the manufacturer HTC reports a spatial accuracy of 0.5°–1.1°. The accuracy seems to fall in that range for the central region of the field of view, but can be larger in the periphery [52, 50]. The peripheral regions are rarely used, thus the average accuracy seems to fall around or below 2° [54, 50] and there is evidence, that this accuracy remains stable over the course of an experiment [54]. The device exhibits eye tracking data delays between approximately 50 ms [55] and 60 ms [52]. All in all, these errors and delays are small enough to extract general directions as we want in this work.

2.2.2 Sensor Synchronization

To develop a model suitable for real-time application, we had to synchronize sensors with inherently different measurement latencies as they would be in a real-world scenario. Utilizing LabStreamingLayer, we live streamed EEG data to Python while concurrently streaming Virtual Reality data, including the Vive eye tracking data obtained in Unity, to Python. These timings were then used to construct the datasets.

2.3 Procedure

Prior to each experiment, participants received detailed task instructions, and demographic data was collected. The electrode cap was positioned, connected, and the impedance of all electrodes was verified. After applying electrode gel, participants wore a head-mounted display over the electrode cap. The participants were seated on a chair with the joystick in front of them. For every participation, the eye tracking was calibrated and checked with a custom validation tool. This process was repeated up until the validation showed a sufficient eye-tracking calibration. To create a controlled environment, ensure the cleanest possible eye-tracking data and minimize the potential for motion sickness, subjects were instructed to maintain a stable head position using a chin rest. The participants were instructed to control the joystick with their main hand. They could freely choose their preferred grip style.

2.3.1 Forest Task

In the first experiments, participants could freely move through a virtual forest environment. The forest was randomly auto-generated with new parts appearing if the world border was getting closer than 75m. The forest environment contained a slightly uneven ground, a set of similar trees and red balls that could symbolise fruits (see Fig. 1 (a)). These balls appeared in four variations: low on ground level, high on a tree, on a plant between these low and high points, and as a double ball on a plant. The occurrence rate of these four variations was identical.

The participants aim was to gather red balls. This could be done by steering close to a ball up until the ball would turn grey in 2 meters distance. Once a ball was grey, the participants could gather it by releasing the joystick or putting it back to the rest position. A ball that was gathered would turn white and could not be gathered again.

The environment and task were designed to allow a diverse set of navigation paths, including turning, backwards navigation and obstacle avoidance, while still allowing for the occurrence of natural behaviour. While it was the objective to gather the balls, no target number or speed requirements were given. Instead, the participants should move smoothly. No instructions regarding gaze behaviour were given. Participants were free to choose different gaze and movement targets in this environment.

2.3.2 Course Task

In the second experiment, participants were following a black path while counting blue stripes on grey cylinders. Participants used the joystick to navigate a minimalist virtual environment along a predefined path (see Fig. 1 (b)). These black navigation paths were generated by concatenating walking paths from the Microsoft Geolife dataset [70, 72, 71]. The first path was oriented to lay right in front of the participant. The last known orientation was used when concatenating two paths. The subjects were instructed to follow the paths but were not instructed to follow them as close as possible.

Gray cylindrical objects were placed randomly on both sides of the virtual path in 2 meters distance to the path. These objects had either two or three dark blue lines, with the majority featuring two lines. While following the path, the participants also had to count the amount of objects with three lines that appeared on the sides. There could be up to four objects with three lines. This was the visual search task that had to be solved simultaneously with the navigation task, resulting in a division of attention and visual tar-

Figure 2: A visualization of the full model architecture including all features. Blue shows Eye-Tracking. Light brown shows positional information. Green shows EEG information. The pink model represents the transformer architecture. For the LSTM, there would be a replacement here.

gets. The distance between the objects varied relative to the random paths. Most spacings were between 3 and 4 meters.

Each participant followed five distinct paths within the VR environment, with a gray cylinder featuring six lines marking the endpoint for virtual locomotion. The double task experiment was designed to force the participants to make a diverse set of eye movements, aiming both at future waypoints and at the individual objects. The choice of locomotion targets was more restricted here and visual attention was required for non-movement targets as well.

3 Prediction Model

Using these two datasets, we constructed prediction models with the aim of forecasting future locomotion intentions, specifically future locomotion trajectories. Both datasets accounted for approximately half of the data (52.7% for the forest experiment and 47.3% for the course and visual search experiment). Different types of models were built, but the same preprocessing was applied each time.

3.1 Preprocessing

We binned the motion and eye data into 60ms steps. For each prediction, we used 30 input points (1.8s) and 70 output points (4.2s). All positional data was transformed to a unifying reference frame. The origin of the coordinate system was reset to the point of prediction. The forward axis (sagittal) was set to the yaw orientation at the point of prediction. The lateral axis was built orthogonally to the yaw orientation at this point. Finally, we calculated velocities between each time step for both axes separately. These two-dimensional velocities enable a homogeneous input and output for the model.

The 70 output points consisted of the two-dimensional velocity in future time steps. The input consisted of 3 types of features: two-dimensional velocity of 30 past time steps, the history of yaw and pitch eye-tracking measurements, the preprocessed EEG features (see Fig. 2). Finally, all data was z-score normalized (standardized). After excluding data with missing values, 294,840 input-output-pairs were obtained.

3.1.1 EEG Preprocessing

First, the 500 Hz EEG was downsampled to obtain 450 4ms EEG values for our 1.8 second input. These were corrected with the channel mean of one input sequence and filtered with a High-pass filter of 1 Hz and a low-pass filter of 80 Hz. 80 Hz was chosen as the monitor refresh rate of the HMD was 90Hz creating an artifact. The EEG was z-standardized and extreme outliers beyond 4 standard deviations were clipped.

On the filtered signal, time frequency analysis was conducted using multitaper for the spectral density estimation. The minimum frequency was 1 Hz, while the maximal frequency was 30Hz. We obtained 5 frequency estimates for the 1.8s history of data, each one 360ms after one another.

Our aim was to create a model that would be able to perform online in a real-time scenario. Thus, complex preprocessing steps like independent component analysis (ICA) filtering were not possible. Nevertheless, we created a dataset where ICA was used to filter artefacts to obtain a better understanding of the EEG components. This dataset could not be used online. We employed the help of the ICLabel classifier [39] to automate the process and exclude artifacts such as eye blink, muscle, and movement activity.

3.2 Architecture

We compare different types of approaches to the temporal nature of this data and to determine the most accurate model. These include basic autoregression, RNNs and Transformer models.

To process the EEG data for our deep learning models, we transform the 450 4ms time steps to 30 60ms timesteps. This matches the sampling length of the velocities, that we use for input, output and eye data, we employed two one-dimensional convolutional layers with a ReLU activation function. The first convolutional layer has a kernel size of 5, 32 input channels and 22 output channels. The second has a kernel size of 3, 22 input channels and 12 output channels. We use two linear layers with sigmoid activation functions to change the temporal size to 30 and the feature size to 8.

The 48 features of 30 time steps are then either processed by an LSTM or a transformer network. To prevent overfitting on the first and easy to predict second, ensemble modeling was employed. The first 15 of the 70 Outputs of the deep neural network Y^{NN} were mixed with autoregression results Y^{AR} according to a weighted average with linearly changing weights:

$$\mathbf{Y}_i = \frac{15-i}{15} \cdot \mathbf{Y}_i^{NN} + \frac{i}{15} \cdot \mathbf{Y}_i^{AR} \quad \text{for } i = 1,2,\ldots,15. \qquad (1)$$

The Transformer architecture consist of a linear embedding layer with 48 input and output dimensions a positional encoding layer and the transformer itself with 4 attentional heads, 2 encoder layers and 2 decoder layers and a feedforward dimension of 24. We predict 70 outputs corresponding to 4.2 seconds. Figure 2 depicts this architecture.

The LSTM has a feature size of 48 and a 0.1 dropout rate. Our LSTM model had 48 hidden units. The output of the LSTM layer went through a dropout layer (p = 0.1) [53] resulting in the final linear dense layer with two outputs, one for each label coordinate and 70 time steps.

We used adam as the optimizer [31]. The learning rate was set to $4e-5$ and to prevent overfitting, a weight decay of $5e-6$ was applied. The model was trained for 30 epochs using a batch size of 128. Then the epoch with the lowest validation error was selected.

3.2.1 Cost function

Forecasting velocities for each time step provides a consistent and reliable outcome, as each step within the time sequence maintains a comparable magnitude. Typically, we also input velocities for a navigation system like a joystick (as opposed to positions).

However, for the cost function we want to accumulate those velocities and compare positions. If a critical turn is predicted a little too early or a little too late, this is much better than if it is entirely overlooked. If we fail to predict a critical turn and compare velocities, we might only notice an error in the immediate timestep in which the turn occurs. If we compare positions, the error becomes evident in every subsequent timestep. Using positions in the cost function enables more stable long-term predictions. Thus, we calculate the cumulative sum of all velocities and compare outputs and labels. Due to the higher variance of larger distances, steps further away from the initial positions would cause larger errors. Thus, we weight each step by the inverse of the step number.

$$P_i = \sum_{k=1}^{i} V_k * \Delta t \tag{2}$$

$$E = \sum_{i=1}^{N} \frac{1}{i}((P_{x,i} - P_{x_{\text{true}},i})^2 + (P_{z,i} - P_{z_{\text{true}},i})^2) \tag{3}$$

The predicted arrays of velocities V_x and V_z will be compared with the true arrays of velocities V_x and V_z. First, we use the cumulative sum to get to an array of positions. Then we get the squared differences between true and predicted. Lastly we sum it all up for the total error.

3.3 Evaluation

To prevent the overlap of input sequences between the training and test sets and ensure the model's generalizability to new data, group-level cross-validation, specifically leave-one-out cross-validation, was implemented. Prior to training, both features and labels underwent z-standardization. In employing this cross-validation methodology, individual prediction errors were computed for each participant. Furthermore, to assess whether a model surpasses a reference model, a significance test was used to obtain more informative results than a simple comparison of average errors.

To distill a singular value for our predicted sequence that facilitates comparison, a quadratic function $ax+bx+c$ was fitted to the error function. It was ensured that each fit achieved an R^2 above 97%. Subsequently, the fitted a and b values were compared for our significance test.

Addressing the issue of non-independence between individual results due to overlapping training and test sets in the cross-validation process, the method proposed by Nadeau and Bengio [47] was employed to correct for this. Consequently, the paired t-test with the Nadeau and Bengio correction was utilized.

However, it is important to approach the results of these significance tests with caution, as highlighted by concerns raised by Bouckaert and Frank [4] regarding the replicability of test methods dependent on the data partitioning in the cross-validation process.

The chosen alpha level for statistical significance was set to 0.05. However, to avoid barely significant spurious findings, we will only consider a difference significant if both fit parameters reach that alpha level. All tests were two-sided. The assumption of normally distributed data was verified through a Shapiro-Wilk test [51] conducted beforehand. The Benjamini-Hochberg correction [3] was applied to the p-values of the multiple tests comparing different features to avoid underestimation of the p-value due to multiple testing for both tested parameters.

To provide an estimate of prediction performance we calculated normalized errors between the true values (labels) and the predictions, i.e. we divide the mean squared errors by the mean squared distance traveled for each of the predicted time steps. This gives an estimate that is relative to the distance the subjects had moved.

For comparison, we also present an autoregression model for all 70 steps, to check whether our deep learning architectures beat this simple technique. For reference, we also show a null model which just predicts the average velocity for every step of the way.

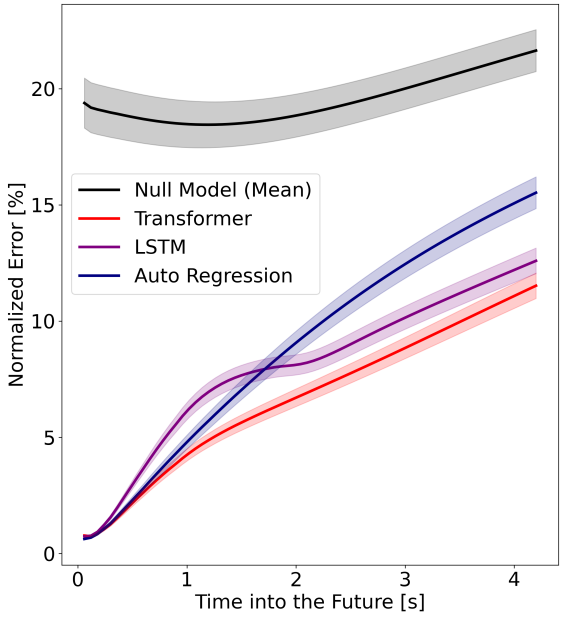

Figure 3: The normalized prediction error of the different models over the time course of 4.2 predicted seconds. Error bars denote between-subject standard errors.

Table 1: Table of the prediction errors at the very last time step after 4.2 seconds. The lower models are all transformer models. Between-Subject Standard Deviations in Brackets.

Model	Normalized Error [%]	Mean Squared Error [m²]	Euclidean / Absolute Error [m]
Transformer	11.49 (2.40)	5.55 (1.03)	1.91 (0.19)
LSTM	12.59 (2.46)	6.08 (1.01)	2.46 (0.20)
Autoregression	15.52 (3.05)	7.48 (1.19)	5.35 (0.87)
Mean	21.64 (4.03)	10.39 (1.33)	7.72 (1.08)
just Position	13.35 (3.02)	6.43 (1.21)	2.52 (0.24)
just Eye	15.38 (2.86)	7.40 (1.01)	2.71 (0.19)
just EEG	17.26 (2.90)	8.31 (0.99)	2.88 (0.17)
Eye + EEG	15.59 (2.96)	7.51 (1.09)	2.73 (0.20)
Position + EEG	12.65 (2.62)	6.09 (1.06)	2.04 (0.20)
Position + Eye	11.35 (2.43)	5.47 (1.00)	1.91 (0.18)

4 Results

The participants traveled a mean distance of 6.41 meters in the output sequence length of 4.2 seconds. The average movement speed was 1.62 m/s. Tab. 1 shows the final prediction errors for different models and different metrics at the end of the output sequence after 4.2 seconds.

4.1 Architecture

After 4.2 seconds the transformer proved superior with a normalized error of 11.49% on average (the absolute error was 1.91 m; the squared error was 5.55 m²) compared to the LSTM with 12.59% (the absolute error was 2.46 m; the squared error was 6.08 m²). The transformer network gave a more accurate prediction for each participant and each time step. The autoregression (normalized error 15.52%) gave a worse prediction for each participant and each time step. The intercept model only reached a normalized error of 21.64% after 4.2 seconds. A significant difference was found when comparing the transformer model with the LSTM network for the fit parameters a ($t(19) = -2.13, p = 0.023$) and b

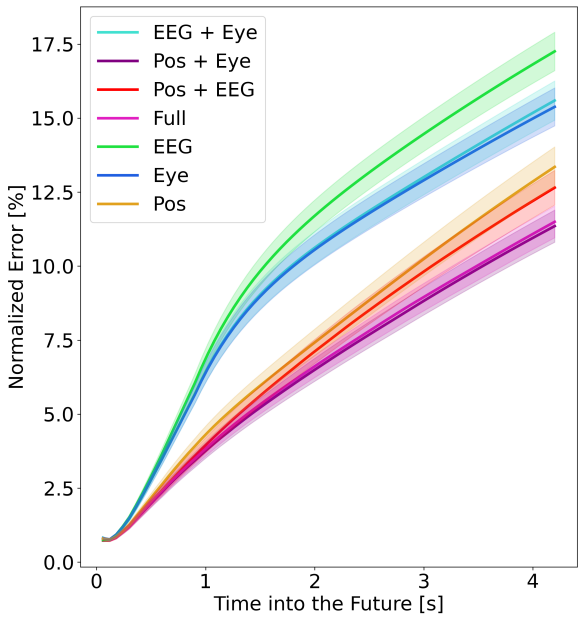

Figure 4: The normalized prediction error of models using different feature combinations over the time course of 4.2 predicted seconds. Error bars denote between-subject standard errors.

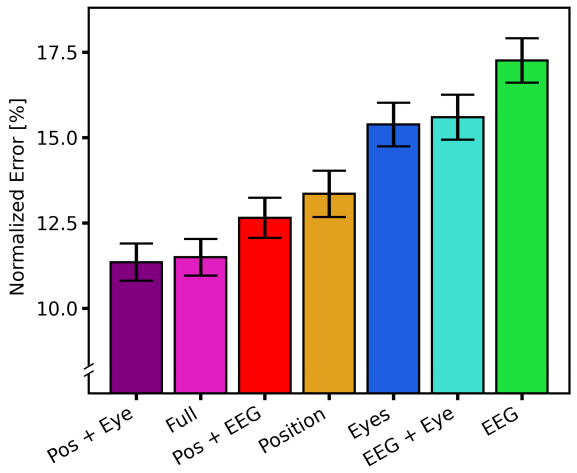

Figure 5: The normalized prediction error of models using different feature combinations at the last time step after 4.2 predicted seconds. Error bars denote between-subject standard errors.

($t(19) = -3.05, p = 0.003$).

Fig. 3 depicts the normalized errors of these architectures. Around the 2-second mark, the error of the LSTM model appears to rise more rapidly. This trend is less evident in other error metrics.

The transformer model used 0.881 million floating point operations (FLOPs) per call and had 15.242 parameters. The presented LSTM was larger with 1.732 million FLOPs and 24.356 parameters.

As the transformer network proved to outperform other architectures for every following result, the transformer architecture was being used for all following models. For this full model, the within-subject standard deviations went from 3.03% in the first time step to 16.30% in the last time step which corresponds to 0.59 cm in the first time step to 1.35 m in the last time step.

4.2 Features

After 4.2 seconds, the models using all features (normalized error = 11.49%) and only position and eye-tracking (normalized error = 11.35%) were very close in performance. No significant difference was achieved when comparing these two models for the fit parameters a ($t(19) = 0.85, p = 0.20$) and b ($t(19) = -0.19, p = 0.43$).

The models using only position (normalized error = 13.35%) performed worse, while the model using position and EEG data came in between (normalized error = 12.65%). We can compare the model with all features with the model only using positional and EEG data to check for the impact of EEG data. A significant difference was found for the fit parameters a ($t(19) = -5.61, p < 0.001$) and b ($t(19) = 5.05, p < 0.001$). The model with eye data was consistently better beginning 360 milliseconds after the time of prediction. The same is true, when comparing the best performing model with positional and eye-tracking information to the model only using positional and EEG data for the fit parameters a ($t(19) = -7.05, < 0.001$) and b ($t(19) = 5.40, p < 0.001$). This model with just positional and eye data was consistently better than the model with positional and EEG data beginning 240 milliseconds after the time of prediction.

Models without positional information were much worse after 4.2 seconds. The model only using eye-tracking data (normalized error = 15.38%) beat the model only using EEG (normalized error = 17.26%). The model using both eye and EEG data came very close (normalized error = 15.59%). No significant difference was achieved when comparing the model only using eye-tracking information and the model using EEG data as well for the fit parameters a ($t(19) = -0.87, p = 0.20$) and b ($t(19) = 1.07, p = 0.18$).

The difference in performance between different feature combinations measured with the normalized Errors is shown over time in Fig. 4. The accumulated errors at the last time-step are additionally depicted in Fig. 5.

4.2.1 EEG as an Input

While the addition of EEG data did not improve the models based on eye and on eye and positional data, with a normalized error of 12.65% after 4.2 seconds, the model using position and EEG data predicted future paths significantly more accurately than the model using only positional data (normalized error = 13.35%) for both the fit parameters a ($t(19) = 3.07, p = 0.005$) and b ($t(19) = -3.02, p = 0.005$). The model with EEG data was consistently better from start to end of the 4.2 second time window.

To investigate which EEG channels were being used, we calculated the absolute weights of the EEG channel in the input of the first convolutional network layer, where the filtered EEG channel data stream went into. Figure 6 shows a projection of the EEG channels. Notably large weights could be observed at the electrodes, that were placed right above the eyes.

Lastly, we tested how the model based on ICA-cleaned data compared to the model using only our standard EEG data. With a normalized error of 22.49% it proved to be worse than the already poor performance of the model using only filtered and transformed EEG data (normalized error = 17.26%). This relationship proved to be statistically significant for both the fit parameters a ($t(19) = 5.70, p < 0.001$) and b ($t(19) = -4.72, p < 0.001$).

4.3 Generalization

To compare our datasets we calculated the errors for the forest and the course experiment without using the other dataset. After 4.2 seconds the normalized error in the forest task was 11.43% while the error in the course experiment was 11.58% The results over time are almost identical Fig. 7 a). No significant difference was achieved when comparing both models for the fit parameter b

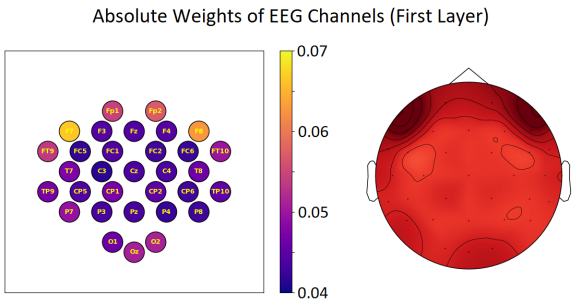

Figure 6: Left: A visualization of the 32 EEG electrodes. Yellow denotes high average absolute weights in the first convolutional layer of the full network. Blue denotes low average absolute weights. Right: a projection onto the head of a person. Darker parts indicate high weights in the first layer.

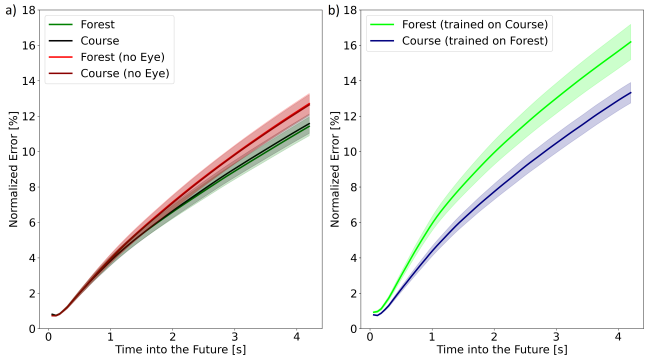

Figure 7: The normalized prediction error of the model over the time course of 4.2 predicted seconds. On the left, we compare an evaluation on the forest with an evaluation on the course. In addition to that we show the performance of models without access to eye data. On the right, we show the generalization of a model that was only trained on the forest experiment to the course experiment and vice versa. Error bars denote between-subject standard errors.

($t(19) = 1.17, p = 0.13$, for a: $t(19) = -2.09, p = 0.02$). In order to evaluate generalization, we also trained two models only using one of the two datasets and calculated the error in the other unseen experiment. For the model that was trained on the course the error after 4.2 seconds in the forest experiment was 16.19% For the model that was trained on the forest the error after 4.2 seconds in the course experiment was 13.32% Fig. 7 b). The difference between these two model variants was statistically significant for the fit parameters a ($t(19) = 6.39, p < 0.001$) and b ($t(19) = -4.66, p < 0.001$).

4.4 Eye Movements in different Experiments

Evaluating the models without eye-tracking data, we get a normalized error of 12.71% in the forest task and a normalized error of 12.64% in the course experiment. Figure Fig. 7 a) provides an illustration. As stated, adding eye data to the models the errors are similar as well (11.43% and 11.58%). No significant difference was achieved when comparing the ratios of errors between both models with and without eye-tracking data for the forest and course task for the fit parameters a ($t(19) = 0.41, p = 0.34$) and b ($t(19) = -0.72, p = 0.24$).

5 Discussion

In this study, we conducted an evaluation of various features and sequence-to-sequence models aimed at predicting locomotion intentions within the context of steering in Virtual Reality. We demonstrated that accurate predictions of human locomotion paths can be achieved without external information, relying solely on intrinsic data.

Firstly, we wanted to evaluate the impact of different features. The integration of eye-tracking data significantly enhanced prediction accuracy. The model combining the history of positional data with eye-tracking information yielded significantly smaller errors than the model only relying on positional information. This result further adds to the existing literature that employs eye-tracking as a pivotal feature for locomotion prediction [65, 17, 21, 54, 6]. This prominent result is not unexpected, as human gaze precedes locomotion targets and contains directional information before actions are performed [36, 20].

While the EEG device data was a valuable addition to the history of positional data, it was not able to provide further benefits when combined with eye-tracking features. Given this result and the fact that the usage of EEG channels that seems to value lateral frontal electrodes (see Fig. 6), we hypothesize that eye artifacts contained in the EEG data were a major information source. An EEG inadvertently measures the moving polarized retina. Given that we wanted to only rely on preprocessing steps that can be performed live, precise removal of eye muscle artifacts is very difficult. The model results that make use of an ICA-filtered EEG data stream perform substantially worse. This further adds to this notion. Deep Learning models are a black box, which makes it hard to answer this with certainty, but we provide some indications that eye muscle activity was responsible for the increase in prediction accuracy of the model with positional and EEG data compared to the model with only positional information. Potentially this also means, that EOG measurements could replace infrared camera based eye tracking in this context.

Secondly, we measured the prediction accuracy and generalization performance of two datasets that differed in the constraints imposed on locomotion and eye movements. Despite these different constraints, the course and forest experiment provided similar predictability, as the errors of the main model are almost identical in both experiments. When we looked at generalization, however, the model trained on the forest was much better in predicting the course than the other way around (see Fig. 7). Thus, the generalization performance of the model that was trained on the forest was much better. This improvement likely results from a more balanced and comprehensive set of motions achievable in the forest task, facilitating superior model generalization.

For our third question, we looked at the improvement due to the inclusion of eye-tracking data for both datasets separately. The addition of eye-tracking features to the model is advantageous for both datasets. The results are very similar here. This means that gaze contains valuable information even in more restricted settings. The second experiment features a double task in which subjects not only had to follow a path but also solve a visual search task. Apparently, there is still enough statistical structure in the link between eye movements and locomotion to be picked up by the model.

Lastly, our findings indicate the superiority of the transformer architecture over a more computationally expensive LSTM prediction. The primary transformer model, incorporating all features, exhibited consistently lower errors across all 70 time steps within our 4.2-second time course. Notably, both deep learning architectures surpassed autoregression in performance, highlighting their predictive capabilities. The efficiency of the transformer architecture may stem from its attentional mechanisms, particularly advantageous for handling features like eye-tracking, which exhibit abrupt changes during events such as saccades rather than linear progression.

These findings collectively contribute valuable insights into en-

hancing the accuracy and understanding the underlying mechanisms of locomotion intention prediction in Virtual Reality.

In our predictive modeling, we utilized features related to users' bodily movements and the direction of their gaze. These features pertain specifically to the perspective of the user, known as egocentric features, and did not encompass information regarding the surrounding environment. Although it might seem intuitive to include environmental factors to enhance predictive accuracy, we intentionally confined our analysis to egocentric features. This decision was driven by our objective to develop a system capable of predicting locomotion across diverse environments in a generalized manner. By not incorporating environmental layout information, our model can be applied across different virtual reality (VR) settings and even extended to non-VR environments, provided accurate measurements of input features are available. This includes to augmented or extended reality scenarios, where obtaining environmental data may be challenging.

5.1 Limitations

Since our motion data was collected using a joystick, the generalizability of our findings to real walking scenarios could be restricted. Our primary emphasis was on virtual motion and locomotor impairments, aligning the joystick data domain with our research objectives. Predicting walking paths with a model that used joystick data as the ground truth for training could lead to compromised performance and vice versa.

Given that our motion data was acquired using a joystick, the applicability of our work to real walking has some limitations. We wanted to focus on virtual motion and locomotor impairments and thus this domain joystick data gives as a higher applicability. If a model that was trained on with joystick data as ground truth is being used for walking path predictions, the performance can suffer and vice versa. While observing healthy behavior is necessary to define the target behavior of a model, only patient data can provide definitive insights into real-world usability. Future work should therefore incorporate a study of patients with mobility impairments.

In this study, we employed standard implementations of transformer and LSTM models. It is plausible that a more intricate approach could yield further enhancements in prediction accuracy. Moreover, expanding the dataset to encompass a more diverse range of environments might contribute to refining model training.

For EEG as a feature, the quantity of data used in our study might have hindered its efficacy in revealing complex and subtle relationships with deep learning. Notably, when compared to positional and eye-tracking features, EEG data provides a substantially greater number of features at a higher sampling rate. This discrepancy in data dimensions warrants consideration in interpreting the results.

Nevertheless, we applied these models to two distinct tasks. While our results highlight successful generalization with the forest task, the application of these findings to a broader context may require careful consideration of task-specific nuances and complexities.

5.2 Applications

The applications of models developed in this study extend to both virtual reality (VR) environments and real-world / augmented reality scenarios. For real-walking VR interfaces prediction of locomotor intention holds promise for collision avoidance and redirected walking [11, 54, 45], contributing to a safer and more immersive virtual experience. For joystick-based navigation predicting the future position of a user within the subsequent frames could serve to augment methodologies aimed at optimizing rendering for applications such as video games. In interactive video games predictions like these could also be used to modify the scene, for example to populate it with objects relevant to the game. For interactive cooperative scenarios, predictions of where a partner is headed could also provide real-time feedback to an online collaborator. Particularly, in cooperative scenarios involving two individuals, knowing what the other person intends to do would be advantageous. For this, a visual interface could be developed that presents the partners predicted future directory graphically as, for example in Fig. 8 (Videos can be found in the Supplemental Materials).

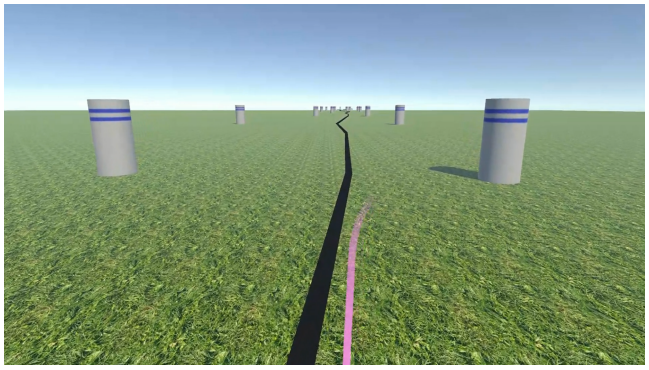

Figure 8: A screenshot of a pseudo-online run of the model showing a possible visual interface.

Such an interface, along with the prediction of locomotor intention, could also be useful in augmented reality scenarios. One application on which we are currently working is in assistive locomotor devices for patients with motor disabilities. The implementation of prediction models could significantly enhance the control of electronic wheelchairs for paraplegic individuals, for example by steering the wheelchair automatically in the predicted direction, or by gently subserving joystick control in the predicted direction. AR interfaces (similarly visualized as in Fig. 8) would then provide feedback to the user and establish a communication loop between the system and the user. This interaction would foster trust and enables precise control, creating a symbiotic relationship between the predictive system and the user.

In conjunction with our research, we introduce two novel datasets derived from distinct VR experiments. The forest task dataset captures a diverse array of locomotor actions within a pseudo-natural setting, while the course task dataset forces a complex set of eye behavior. We envision that these datasets will contribute to the growing repository of public VR datasets, fostering collaborative research endeavors and serving as valuable resources for future investigations in the field.

6 Conclusion

Eye-tracking data proved to be a valuable feature for locomotion prediction, even in the absence of gait and steps. This is also true when gaze and locomotion targets are not tightly linked. Successful generalization requires datasets encompassing diverse movement scenarios. Transformer networks emerged as superior to LSTM networks for locomotion prediction.

Supplemental Materials

All supplemental materials are available on OSF. Videos of a pseudo-online run of the model are available at `https://osf.io/6xnzf`. A second video shows different possibilities of visualization. The forest experiment can be found at `https://osf.io/ney6v/`. The course can be found at `https://osf.io/4g9pw/`.

Acknowledgments

This work was supported by the German Research Foundation (LA 952/11-1) and has received funding from the European Union's Horizon 2020 and Horizon Europe research and innovation programmes under grant agreements No 951910 and No 101086206.

REFERENCES

[1] S. Becker, R. Hug, W. Hübner, and M. Arens. An evaluation of trajectory prediction approaches and notes on the trajnet benchmark. *arXiv preprint arXiv:1805.07663*, 2018. doi: 10.48550/arXiv.1805.07663 2

[2] A. Belardinelli, M. Y. Stepper, and M. V. Butz. It's in the eyes: Planning precise manual actions before execution. *Journal of Vision*, 16(1):18–18, 01 2016. doi: 10.1167/16.1.18 1

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x 5

[4] R. R. Bouckaert and E. Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 3–12. Springer, Sydney, NSW, Australia, 2004. doi: 10.1007/978-3-540-24775-3_3 5

[5] G. Bremer, N. Stein, and M. Lappe. Predicting future position from natural walking and eye movements with machine learning. In *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 19–28. IEEE, Taichung, Tawian, 2021. doi: 10.1109/AIVR52153.2021.00013 2

[6] G. Bremer, N. Stein, and M. Lappe. Do they look where they go? gaze classification during walking. In *NeuRIPS 2022 Workshop on Gaze Meets ML*, 2022. 2, 7

[7] G. Bremer, N. Stein, and M. Lappe. Machine learning prediction of locomotion intention from walking and gaze data. *International Journal of Semantic Computing*, 17(01):119–142, 2023. doi: 10.1142/S1793351X22490010 2

[8] A. Breuer, S. Elflein, T. Joseph, J.-A. Bolte, S. Homoceanu, and T. Fingscheidt. Analysis of the effect of various input representations for lstm-based trajectory prediction. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 2728–2735. IEEE, Auckland, NZ, 2019. doi: 10.1109/ITSC.2019.8917373 1

[9] D. Calow and M. Lappe. Efficient encoding of natural optic flow. *Network Comput. Neural Syst.*, 19(3):183–212, 2008. doi: 10.1080/09548980802368764 1

[10] F.-Y. Chao, C. Ozcinar, and A. Smolic. Transformer-based long-term viewport prediction in 360° video: Scanpath is all you need. In *MMSP*, pp. 1–6, 2021. doi: 10.1109/MMSP53017.2021.9733647 1

[11] Y.-H. Cho, D.-Y. Lee, and I.-K. Lee. Path prediction using lstm network for redirected walking. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 527–528. IEEE, Reutlingen, Germany, 2018. doi: 10.1109/VR.2018.8446442 1, 2, 8

[12] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018. doi: 10.1109/TIP.2018.2851672 2

[13] 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*, pp. 6992–7001. Seattle, WA, USA, 2020. doi: 10.48550/arXiv.1904.03419 1

[14] S. Durant and J. M. Zanker. The combined effect of eye movements improve head centred local motion information during walking. *PLOS ONE*, 15(1):e0228345, 2020. doi: 10.1371/journal.pone.0228345 1

[15] X. Feng, Y. Liu, and S. Wei. Livedeep: Online viewport prediction for live virtual reality streaming using lifelong deep learning. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 800–808. Atlanta, GA, USA, 2020. doi: 10.1109/VR46266.2020.00104 2

[16] L. Franco, L. Placidi, F. Giuliari, I. Hasan, M. Cristani, and F. Galasso. Under the hood of transformer networks for trajectory forecasting. *Pattern Recognition*, 138:109372, 2023. doi: 10.1016/j.patcog.2023.109372 1, 2

[17] J. Gandrud and V. Interrante. Predicting destination using head orientation and gaze direction during locomotion in vr. In *ACM Symposium on Applied Perception, SAP 2016*, pp. 31–38. Association for Computing Machinery, Inc, Anaheim, CA, USA, 2016. doi: 10.1145/2931002.2931010 1, 2, 7

[18] K. Gramann, J. T. Gwin, D. P. Ferris, K. Oie, T.-P. Jung, C.-T. Lin, L.-D. Liao, and S. Makeig. Cognition in action: imaging brain/body dynamics in mobile humans. *Reviews in the neurosciences*, 2011. doi: 10.1515/RNS.2011.047 2

[19] R. Grasso, P. Prévost, Y. P. Ivanenko, and A. Berthoz. Eye-head coordination for the steering of locomotion in humans: an anticipatory synergy. *Neuroscience Letters*, 253(2):115–118, 1998. doi: 10.1016/s0304-3940(98)00625-9 2

[20] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194, 2005. doi: 10.1016/j.tics.2005.02.009 1, 7

[21] C. Hirt, M. Ketzel, P. Graf, C. Holz, and A. Kunz. Short-term path prediction for spontaneous human locomotion in arbitrary virtual spaces. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 554–559. IEEE, 2022. doi: 10.1109/ISMAR-Adjunct57072.2022.00116 2, 7

[22] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735 1

[23] M. A. Hollands and D. E. Marple-Horvat. Visually guided stepping under conditions of step cycle-related denial of visual information. *Experimental Brain Research*, 109(2):343–356, 1996. doi: 10.1007/BF00231792 1

[24] M. A. Hollands, D. E. Marple-Horvat, S. Henkes, and A. K. Rowan. Human eye movements during visually guided stepping. *Journal of Motor Behavior*, 27(2):155–163, 1995. doi: 10.1080/00222895.1995.9941707 1

[25] M. A. Hollands, A. E. Patla, and J. N. Vickers. "look where you're going!": gaze behaviour associated with maintaining and changing the direction of locomotion. *Experimental Brain Research*, 143(2):221–230, 2002. doi: 10.1007/s00221-001-0983-7 1

[26] 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*, 27(5):2681–2690, 2021. doi: 10.1109/TVCG.2021.3067779 2

[27] 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*, 26(5):1902–1911, 2020. doi: 10.1109/TVCG.2020.2973473 2

[28] T. Imai, S. T. Moore, T. Raphan, and B. Cohen. Interaction of the body, head, and eyes during walking and turning. *Experimental Brain Research*, 136(1):1–18, 2001. doi: 10.1007/s002210000533 2

[29] S.-B. Jeon, J. Jung, J. Park, and I.-K. Lee. F-rdw: Redirected walking with forecasting future position. *arXiv preprint arXiv:2304.03497*, 2023. doi: 10.48550/arXiv.2304.03497 2

[30] A. Khajuria, R. Sharma, and D. Joshi. Eeg dynamics of locomotion and balancing: Solution to neuro-rehabilitation. *Clinical EEG and Neuroscience*, 55(1):143–163, 2024. doi: 10.1177/15500594221123690 2

[31] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv Preprint arXiv:1412.6980*, 2014. doi: 10.48550/arXiv.1412.6980 4

[32] D. Kit, L. Katz, B. Sullivan, K. Snyder, D. Ballard, and M. Hayhoe. Eye movements, visual search and scene memory, in an immersive virtual environment. *PLOS ONE*, 9(4):1–11, 04 2014. doi: 10.1371/journal.pone.0094362 2

[33] J. E. Kline, H. J. Huang, K. L. Snyder, and D. P. Ferris. Isolating gait-related movement artifacts in electroencephalography during human walking. *Journal of neural engineering*, 12(4):046022, 2015. doi: 10.1088/1741-2560/12/4/046022 2

[34] J. Kritikos, A. Makrypidis, A. Alevizopoulos, G. Alevizopoulos, and D. Koutsouris. Can brain–computer interfaces replace virtual reality controllers? a machine learning movement prediction model during virtual reality simulation using eeg recordings. In *Virtual Worlds*, vol. 2, pp. 182–202. MDPI, 2023. doi: 10.3390/virtualworlds2020011 2

[35] M. Land and B. Tatler. *Locomotion on foot*, pp. 100–115. Oxford University Press, 07 2009. doi: 10.1093/acprof:oso/9780198570943.003.0006 1

[36] M. F. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25-26):3559–3565, 2001. doi: 10.1016/S0042-6989(01)00102-X 1, 7

[37] D.-Y. Lee, Y.-H. Cho, and I.-K. Lee. Real-time optimal planning for redirected walking using deep q-learning. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 63–71. IEEE, Osaka, Japan, 2019. doi: 10.1109/VR.2019.8798121 1

[38] S.-H. Lee, H.-T. Joo, I. Chung, D. Park, Y. Choi, and K.-J. Kim. A novel approach for virtual locomotion gesture classification: Self-teaching vision transformer for a carpet-type tactile sensor. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 461–471. IEEE, 2024. 2

[39] A. Li, J. Feitelberg, A. P. Saini, R. Höchenberger, and M. Scheltienne. Mne-icalabel: Automatically annotating ica components with iclabel in python. *Journal of Open Source Software*, 7(76):4484, 2022. doi: 10.21105/joss.04484 4

[40] M. Li, N. K. Banerjee, and S. Banerjee. Using motion forecasting for behavior-based virtual reality (vr) authentication. *arXiv preprint arXiv:2401.16649*, 2024. doi: 10.48550/arXiv.2401.16649 1

[41] M. Li, B. Zhong, E. Lobaton, and H. Huang. Fusion of human gaze and machine vision for predicting intended locomotion mode. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1103–1112, 2022. doi: 10.1109/TNSRE.2022.3168796 2

[42] J.-S. Lin and B.-H. She. A bci system with motor imagery based on bidirectional long-short term memory. In *IOP conference series: Materials science and engineering*, vol. 719, p. 012026. IOP Publishing, 2020. doi: 10.1088/1757-899X/719/1/012026 2

[43] 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*, pp. 2891–2900. Honolulu, HI, USA, 2017. doi: 10.1109/CVPR.2017.497 1

[44] J. S. Matthis, J. L. Yates, and M. M. Hayhoe. Gaze and the control of foot placement when walking in natural terrain. *Current Biology*, 28(8):1224–1233, 2018. doi: 10.1016/j.cub.2018.03.008 1

[45] J. Mayor, P. Calleja, and F. Fuentes-Hurtado. Long short-term memory prediction of user's locomotion in virtual reality. *Virtual Reality*, 28(1):1–12, 2024. doi: 10.1007/s10055-024-00962-9 2, 8

[46] H.-S. Moon and J. Seo. Prediction of human trajectory following a haptic robotic guide using recurrent neural networks. In *2019 IEEE World Haptics Conference (WHC)*, pp. 157–162. IEEE, Tokyo, Japan, 2019. doi: 10.1109/WHC.2019.8816157 1

[47] C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003. doi: 10.1023/A:1024068626366 5

[48] A. Palumbo, V. Gramigna, B. Calabrese, and N. Ielpo. Motor-imagery eeg-based bcis in wheelchair movement and control: A systematic literature review. *Sensors*, 21(18):6285, 2021. doi: 10.3390/s21186285 2

[49] S. Razzaque, Z. Kohn, and M. C. Whitton. Redirected Walking. In *Eurographics 2001 - Short Presentations*. Eurographics Association, Manchester, UK, 2001. doi: 10.2312/egs.20011036 2

[50] I. Schuetz and K. Fiehler. Eye tracking in virtual reality: Vive pro eye spatial accuracy, precision, and calibration reliability. *Journal of Eye Movement Research*, 15(3), 2022. 3

[51] Shapiro and Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 12 1965. doi: 10.1093/biomet/52.3-4.591 5

[52] A. Sipatchin, S. Wahl, and K. Rifai. Eye-tracking for clinical ophthalmology with virtual reality (vr): A case study of the htc vive pro eye's usability. In *Healthcare*, vol. 9, p. 180. Mdpi, 2021. 3

[53] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 4

[54] N. Stein, G. Bremer, and M. Lappe. Eye tracking-based lstm for locomotion prediction in vr. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 493–503. IEEE, Christchurch, New Zealand, 2022. doi: 10.1109/VR51125.2022.00069 2, 3, 7, 8

[55] N. Stein, D. C. Niehorster, T. Watson, F. Steinicke, K. Rifai, S. Wahl, and M. Lappe. A comparison of eye tracking latencies among several commercial head-mounted displays. *i-Perception*, 12(1):2041669520983338, 2021. 3

[56] B. M. 't Hart and W. Einhauser. Mind the step: complementary effects of an implicit task on eye and head movements in real-life gaze allocation. *Experimental Brain Research*, 223(2):233–249, 2012. doi: 10.1007/s00221-012-3254-x 1

[57] Y. Tang, L. Ma, W. Liu, and W. Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamic. *arXiv preprint arXiv:1805.02513*, 2018. doi: 10.24963/ijcai.2018/130 1

[58] B. W. Tatler and S. L. Tatler. The influence of instructions on object memory in a real-world setting. *Journal of Vision*, 13(2):5–5, 02 2013. doi: 10.1167/13.2.5 2

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. doi: 10.48550/arXiv.1706.03762 1

[60] J. Wagner, T. Solis-Escalante, P. Grieshofer, C. Neuper, G. Müller-Putz, and R. Scherer. Level of participation in robotic-assisted treadmill walking modulates midline sensorimotor eeg rhythms in able-bodied subjects. *Neuroimage*, 63(3):1203–1211, 2012. doi: 10.1016/j.neuroimage.2012.08.019 2

[61] Q. Wang, H. Luo, L. Ye, A. Men, F. Zhao, Y. Huang, and C. Ou. Pedestrian heading estimation based on spatial transformer networks and hierarchical lstm. *IEEE Access*, 7:162309–162322, 2019. doi: 10.1109/ACCESS.2019.2950728 1, 2

[62] J. Wiener, O. De Condappa, and C. Holscher. Do you have to look where you go? gaze behaviour during spatial decision making. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 33. Boston, MA, USA, 2011. 2

[63] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao. Gaze prediction in dynamic 360° immersive videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT, USA, June 2018. doi: 10.1109/CVPR.2018.00559 2

[64] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *European Conference on Computer Vision*, pp. 507–523. Springer, Glasgow, United Kingdom, 2020. doi: 10.1007/978-3-030-58610-2_30 2

[65] M. Zank and A. Kunz. Eye tracking for locomotion prediction in redirected walking. In *2016 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 49–58. IEEE, Greenville, SC, USA, 2016. 1, 2, 7

[66] M. Zank and A. Kunz. Where are you going? using human locomotion models for target estimation. *The Visual Computer*, 32(10):1323–1335, 2016. doi: 10.1007/s00371-016-1229-9 1

[67] J. Zhao, M. Shao, Y. Wang, and R. Xu. Real-time recognition of in-place body actions and head gestures using only a head-mounted display. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 105–114. IEEE, 2023. 2

[68] L. Zhao, X. Lu, Q. Bao, and M. Wang. In-place gestures classification via long-term memory augmented network. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 224–233. IEEE, 2022. 2

[69] L. Zhao, X. Lu, M. Zhao, and M. Wang. Classifying in-place gestures with end-to-end point cloud learning. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 229–238. IEEE, 2021. 2

[70] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pp. 312–321, 2008. doi: 10.1145/1409635.1409677 3

[71] Y. Zheng, X. Xie, W.-Y. Ma, et al. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010. 3

[72] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pp. 791–800, 2009. doi: 10.1145/1526709.1526816 3