Gaze and Environmental Context-Guided Deep Neural Network and Sequential Decision Fusion for Grasp Intention Recognition

Bo Yang, *Student Member, IEEE*, Xinxing Chen, *Member, IEEE*, Xiling Xiao, Pei Yan, Yasuhisa Hasegawa, *Member, IEEE*, and Jian Huang, *Senior Member, IEEE*

***Abstract*— Grasp intention recognition plays a crucial role in controlling assistive robots to aid older people and individuals with limited mobility in restoring arm and hand function. Among the various modalities used for intention recognition, the eye-gaze movement has emerged as a promising approach due to its simplicity, intuitiveness, and effectiveness. Existing gaze-based approaches insufficiently integrate gaze data with environmental context and underuse temporal information, leading to inadequate intention recognition performance. The objective of this study is to eliminate the proposed deficiency and establish a gaze-based framework for object detection and its associated intention recognition. A novel gaze-based grasp intention recognition and sequential decision fusion framework (GIRSDF) is proposed. The GIRSDF comprises three main components: gaze attention map generation, the Gaze-YOLO grasp intention recognition model, and sequential decision fusion models (HMM, LSTM, and GRU). To evaluate the performance of GIRSDF, a dataset named Invisible containing data from healthy individuals and hemiplegic patients is established. GIRSDF is validated by trial-based and subject-based experiments on Invisible and outperforms the previous gaze-based grasp intention recognition methods. In terms of running efficiency, the proposed framework can run at a frequency of about 22 Hz, which ensures real-time grasp intention recognition. This study is expected to inspire additional gaze-related grasp intention recognition works.**

***Index Terms*— Grasp intention recognition, gaze, environmental context, object detection, hidden Markov model, deep neural networks.**

Manuscript received 22 March 2023; revised 18 July 2023 and 16 August 2023; accepted 5 September 2023. Date of publication 13 September 2023; date of current version 22 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grant U1913207, Grant 62103157, and Grant 62103180; in part by the Program for Huazhong University of Science and Technology (HUST) Academic Frontier Youth Team; and in part by the Fundamental Research Funds for the Central Universities under Grant HUST2022JYCXJJ002. *(Corresponding authors: Jian Huang; Xinxing Chen.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Union Hospital Tongji Medical College Huazhong University of Science and Technology under Application No. UHCT-IEC-SOP-016-03-01.

Bo Yang, Pei Yan, and Jian Huang are with the Key Laboratory of the Ministry of Education for Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: aleksibob@hust.edu.cn; yanpei@hust.edu.cn; huang_jan@mail.hust.edu.cn).

Xinxing Chen is with the Shenzhen Key Laboratory of Biomimetic Robotics and Intelligent Systems, Shenzhen 518055, China (e-mail: chenxx@sustech.edu.cn).

Xiling Xiao is with the Department of Rehabilitation, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China (e-mail: xiaoxiling23@126.com).

Yasuhisa Hasegawa is with the Department of Micro-Nano Mechanical Science and Engineering, Nagoya University, Chikusa-ku, Nagoya 464-8603, Japan (e-mail: hasegawa@mein.nagoya-u.ac.jp).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2023.3314503, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2023.3314503

NOMENCLATURE

Gaze Related Contents

n	Index of frame captured from scene camera.
k	Index of gaze point extracted from eye-tracker in each frame.
$\mathbf{f}_n$	The nth scene image frame.
$\mathbf{g}_n(k)$	kth unaligned gaze point on nth frame. $\mathbf{g}_n(k) = [g_{n,x}(k), g_{n,y}(k)]$.
$\mathcal{G}_n$	Set of unaligned gaze points on nth frame. $\mathcal{G}_n = \{\mathbf{g}_n(k) = [g_{n,x}(k), g_{n,y}(k)], k = 1, \ldots, K\}$.
$\bar{\mathbf{g}}_n(k)$	kth aligned gaze point on nth frame. $\bar{\mathbf{g}}_n(k) = [\bar{g}_{n,x}(k), \bar{g}_{n,y}(k)]$.
$\bar{\mathcal{G}}_n$	Set of aligned gaze points on nth frame. $\bar{\mathcal{G}}_n = \{\bar{\mathbf{g}}_n(k) = [\bar{g}_{n,x}(k), \bar{g}_{n,y}(k)], k = 1, \ldots, K\}$.
w_1	Gain of gaze map generation.
σ	Standard deviation of gaze map generation.
$\mathbf{img}_{n,k}$	The gaze map is generated by the gaze point $\mathbf{g}_n(k)$.
$\mathbf{gm}_n$	The gaze map is generated by the aligned gaze points set $\bar{\mathcal{G}}_n$.

Grasp Intention Related Contents

l_s	Sliding window size for gaze map generation.
l_z	Number of gaze points in the sliding window l_s.
l_w	Sliding window size for sequential fusion.
λ_j	HMM model. $\lambda_j = (\mathbf{A}_j, \mathbf{B}_j, \pi_j)$, $j = 1, 2$. λ_1 is the HMM model for IT and λ_2 is the HMM model for IA.
$\mathbf{A}_j$	The transition probability matrix.

$\mathbf{B}_j(n)$	The emission probability matrix given observation $\mathbf{o}_n$. $\mathbf{B}_j(n) = \left[p_j\left(\mathbf{o}_n \mid i_{n,j}=0\right),\ldots,p_j\left(\mathbf{o}_n \mid i_{n,j}=m_j\right)\right]$.
$p_j\left(\mathbf{o}_n \mid i_{n,j}\right)$	The emission probability of observing sample $\mathbf{o}_n$ given the latent state $i_{n,j}$.
$\mathcal{Q}_j$	The set of IT states. $\mathcal{Q}_1 = \{0,\ldots,m_1\}$ and $\mathcal{Q}_2 = \{0,\ldots,m_2\}$.
$i_{n,j}$	Subject's latent state of λ_j. $i_{n,1} \in \mathcal{Q}_1$ and $i_{n,2} \in \mathcal{Q}_2$.
$s_{n,j}$	The smoothed latent state. $s_{n,1} \in \mathcal{Q}_1$ and $s_{n,2} \in \mathcal{Q}_2$.
$s_{n,j}^{opt}$	The optimized latent state. $s_{n,1}^{opt} \in \mathcal{Q}_1$ and $s_{n,2}^{opt} \in \mathcal{Q}_2$.
$\mathbf{p}_j(s_{n,j})$	The probability distribution of the current smoothed state.
$\mathcal{O}$	The set of possible observations of two HMMs. $\mathcal{O} = \{\mathbf{o}_0,\ldots,\mathbf{o}_n\}$.
$\mathbf{o}_n$	The observation of HMM, i.e., the input sample of Gaze-YOLO. $\mathbf{o}_n$ is composed of the scene image and the gaze map concatenated in the channel dimension, which can be expressed as $\mathbf{o}_n = \mathbf{f}_n \oplus \mathbf{gm}_n$.

I. Introduction

THE upper limb assistive robots, such as prostheses [1], supernumerary robotic limbs [2], and exoskeletons [3], can help the elderly and infirm people with upper limb disabilities restore arm and hand functions. However, a barrier to using these assistive technologies is the lack of appropriate human-robot interaction (HRI) that allows people to express their grasp intentions intuitively and naturally. There have been many studies for upper limb intention recognition and prediction, utilizing electromyography (EMG) signals [4], Electroencephalogram (EEG) signals [5], etc. Although these biosignals can be effectively used for intention recognition, they can not be applied to all populations, such as stroke patients [6].

Eye-tracking is an emerging technology for users' gaze point estimation [7]. Even in severe hemiplegia and other motor disorders, the human oculomotor system typically remains intact [8], [9], making eye-tracking accessible to users with disabilities. It has been shown that gaze is related to intention and that a person's gaze can express intention and anticipate actions [10], [11]. The natural and intuitive link between intention and gaze makes it a promising approach to exploiting gaze for grasp intention recognition. In grasp intention recognition, there are two attributes that we need to focus on, one is the intentional target (IT), which is used to indicate the object that the user is interested in and viewing, and the other is the intentional action (IA), which is used to indicate whether the user has a grasp intention on this object [12], [13], [14], [15], [16], [17]. With IT and IA, the assistive robot can help the user with the grasping task.

Typically, studies on gaze-based upper limb assistive robots focused on two aspects: 1) gaze point estimation and trajectory planning and 2) intention recognition. The first category of studies explored gaze points to determine target position and plan robots' movement trajectories [9], [14], [15], [17], [18], [19], [20]. Faisal et al. proposed a 3D gaze calibration method utilizing continuous robotic arm trajectories [18]. Furthermore, Faisal et al. implemented a grasp assist task by leveraging the estimated 3D gaze points [14], [18]. Wang et al. proposed a method that combines a depth camera and an eye-tracker to estimate 3D gaze points [15]. However, the estimation of 3D gaze in real environments has certain limitations, such as being sensitive to the user's head motion or requiring additional optical devices to track head motion. Chen et al. developed a lightweight multi-model network for appearance-based eye gaze tracking. Their method fused eye and head features to improve gaze estimation accuracy and achieved a 27 × speedup [20]. Yang et al. introduced a set-membership filter based on eye-movement modality, which effectively improves the gaze signal quality [21]. Most of these studies primarily utilized machine learning techniques to estimate gaze points based on eye features or incorporated additional depth cameras to provide depth information. However, these methods do not resolve the problem of recognizing the user's grasp intention, including IT and IA.

The second category of studies mainly focused on intention recognition. Li et al. constructed a naive Bayesian model for intention inference based on the correlation between objects and intentions [22]. Koochaki et al. used a density-based spatial clustering of applications with noise (DBSCAN) to extract gaze features and infer intention [23]. Such studies are only suitable for activities of daily living (ADL) intention inference, not grasp intention recognition. Another part of the studies focused on the foundation of intention recognition–grasp intention recognition. In [15] and [16], fixations with dwell times longer than two seconds were utilized to determine the grasp intention. In [24], a network based on the egocentric view termed VIDEO-Net was introduced to recognize the IA but has not determined IT. In [12], a weakly-supervised network was used to recognize IT, and then an extra long-short term memory (LSTM) was used to identify IA. The Earth Mover's Distance (GazeEMD) was exploited to evaluate the similarity between gaze points and target saliency to determine the IT by Shi et al. [25]. In [26], a gaze point motion model TAGMM was used to process the gaze data, and then multiple features were proposed for identifying IA and IT.

While these studies have produced positive outcomes, they still have some issues. The first issue is that gaze and environmental context (scene image) interact minimally and do not efficiently integrate. Typically, these algorithms require target detection techniques to detect objects in the scene, followed by the construction of a set of features for intention recognition based on the object coordinates and the gaze points. As a result, intention recognition is influenced by object detection performance and gaze data quality. For example, gaze points that fall outside the bounding box due to noise but are close to the boundaries may still indicate that the user's IT is the object. However, a method relying on the bounding box would misclassify this object as not being IT. The ideal approach is to fuse gaze data and scene images to extract effective features to identify intention. Convolutional neural networks

have proven to be a powerful technique for extracting features in image tasks. However, due to dimensional inconsistencies, discrete gaze point coordinates are challenging to be fed into a 2D convolutional network. Moreover, the gaze point can only provide information about a single pixel point, which does not reflect the actual human vision characteristics. Human eyesight is an area rather than a single point. Consequently, there is a dearth of an effective approach to extract gaze and scene features and perform intention recognition.

Another issue is the underuse of temporal information, which may result in a lack of accuracy in gaze-based grasp intention recognition. As previously reported, existing gaze-based grasp intention recognition methods had an accuracy of less than 76% [12], [25]. Human upper limb intentions are continuous in a grasping task–the user's gaze usually stays on the object until the grasping action is completed. Therefore, a sequence model could be established to fuse temporal information to increase grasp intention recognition performance, which has been proven feasible in human behavior prediction [27], gesture recognition [28], and intention recognition [29]. Neural networks, such as LSTM and gated recurrent units (GRU), are effective approaches for fusing temporal information. While these models improve the accuracy of human intention recognition, they are data-intensive and require training. Consider the grasp intention is classified into several classes, each of which can be described in probabilistic terms. Bayesian models provide an alternative method to fuse temporal information and optimize sequential decisions in scenarios involving probabilities. Additionally, the probabilistic model is interpretable.

From the previous analysis, the challenges in achieving gaze-based grasp intention recognition are as follows:

1) How to develop a framework for integrating gaze data and environmental context (scene images) to simultaneously detect IT and recognize the corresponding IA.
2) How to fuse sequential decisions to improve the accuracy of intention recognition.

Our objective is to establish a gaze-based framework for object detection and its associated grasp intention recognition based on multimodal information (gaze data and environmental context). To achieve this objective, we designed a gaze-based grasp intention recognition and sequential decision fusion framework (GIRSDF). This framework is composed of a gaze attention map generation method, a Gaze-YOLO network, and sequential decision models. The main contributions of the present paper include the following:

1) In terms of grasp intention recognition, a novel end-to-end deep neural network Gaze-YOLO is designed. This network employs gaze data and scene images as the inputs for scene objects detection and corresponding intention detection.
2) In terms of Gaze-YOLO inputs, a gaze attention map generation method based on human visual properties is proposed to align the representation of gaze data with the scene image.
3) In terms of sequential decision optimization, models that (HMM) do not require training and models (LSTM and GRU) that need training are constructed to fuse sequential decisions and improve intention recognition accuracy.
4) A dataset named Invisible is established. The dataset containing data from seven healthy individuals and two hemiplegic patients. The proposed framework's performance is evaluated on the dataset.

The rest of the paper is organized as follows: Section II describes GIRSDF. Section III introduces the experimental results of the proposed framework. Section IV presents the discussion. Section V concludes the paper.

II. The Methodology

The proposed GIRSDF is shown in Fig. 1, including a gaze attention map generation approach, a Gaze-YOLO intention recognition network, and a sequential decision fusion model. The following discussion will introduce each component of the grasp intention recognition framework.

A. Eye-Tracker Output

The eye-tracker (Pupil-Invisible: Pupil Labs, Berlin, Germany), which comprises eye cameras and a scene camera, delivers the user's two-dimensional (2D) gaze point coordinates on the scene image. The 2D gaze point coordinates are indicated as $\mathbf{g}_n(k) = [g_{n,x}(k), g_{n,y}(k)]$, where n is the scene image index, and k is the gaze point index in one scene image. The scene camera is sampled at 30 Hz and the eye camera is sampled at approximately 200 Hz. Thus, a scene image contains multiple gaze points. A snapshot of the outputs of the eye-tracker are displayed in Fig. S1 in the Supplementary Document.

B. Gaze Attention Map Generation

During the use of the eye tracker, the subject's head movement will cause the gaze points move across in different scene frames. The gaze points should be aligned to the same scene image to eliminate the effects of the subject's head movement. Consider a video clip $\mathcal{F} = \{\mathbf{f}_n, n = 1, \ldots, N\}$ and $\mathbf{f}_n$ associated gaze points $\mathcal{G}_n = \{\mathbf{g}_n(k), k = 1, \ldots, K_n\}$, where $\mathbf{f}_n$ is scene image frame. Then we adopt Diaz's approach [12] to align the gaze points.

$$\bar{\mathcal{G}}_n = align(\mathcal{G}_n, \mathbf{f}_n). \tag{1}$$

The utilization of gaze points encounters two difficulties. The first is that the gaze point coordinates cannot be fed into the image processing 2D convolutional neural network. Another difficulty is that the gaze point can only provide information about a single point, which is inconsistent with visual properties. As demonstrated in Fig. 2, when a person stares at a target, the eyesight is focused on the region of interest rather than a single location. The region's center is the most concerned and interested area determined by the brain, and the attention progressively attenuates to the surroundings [30]. To solve these two problems, we propose a method for generating gaze attention maps from gaze points.

We utilize Gaussian functions for generating gaze attention maps to model human visual attention's decay process from

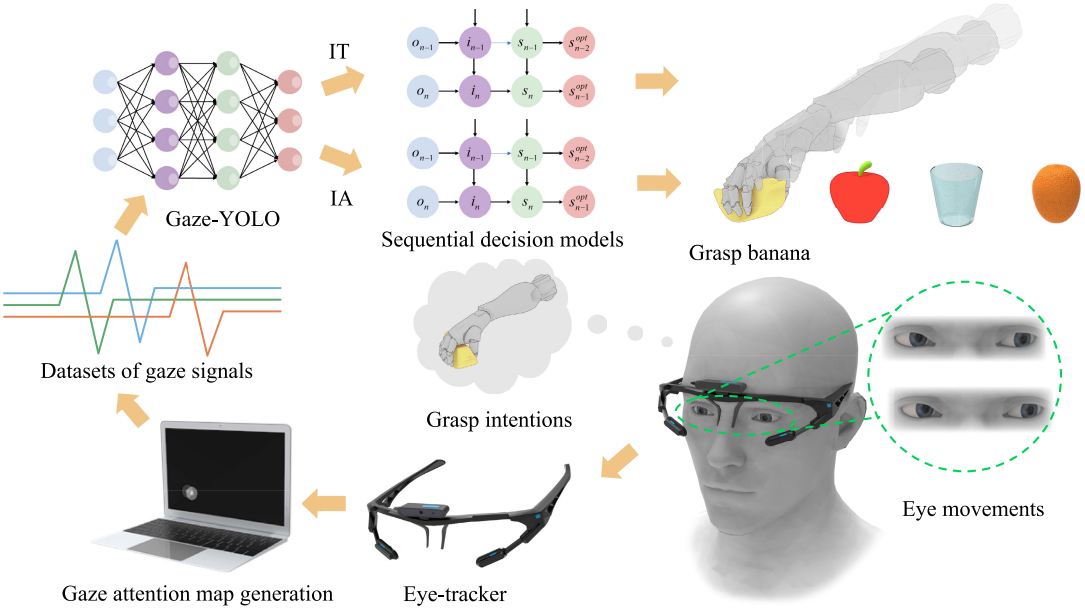


Fig. 1. GIRSDF framework. This framework includes a gaze attention map generation approach, a Gaze-YOLO intention recognition network, and sequential models. Gaze-YOLO's input is the gaze attention map and the corresponding scene image; Gaze-YOLO completes the object detection in the scene and recognizes the intention for each object. The grasp intention results are converted into probabilities of different types of IT and IA, respectively. Then the sequential models fuses the probabilities of the current sample and previous samples to estimate the ultimate intention decision.

the gaze point coordinates to the surrounding environment. The gaze attention map is generated as follows:

$$\mathbf{img}_{n,k}(x, y) = w_1 \exp\left(-\frac{L\left([x, y], \mathbf{g}_n(k)\right)^2}{2\sigma^2}\right), \quad (2)$$

where $\mathbf{img}_{n,k}$ represents the gaze attention map of $\mathbf{g}_n(k)$ and $\mathbf{img}_{n,k}(x, y)$ is the pixel value in the xth row and yth column of this grayscale image. σ^2 is the Gaussian function's variance, which represents the attention decay rate and w_1 is the gain factor. $L(\cdot)$ denotes the Euclidean distance. Assume an aligned gaze point set $\bar{\mathcal{G}}_{buf} = \{\bar{\mathcal{G}}_{n+1-l_s}, \ldots, \bar{\mathcal{G}}_n\}$ contains l_z gaze points, where l_s is the size of the sliding window. As a result, the gaze attention map of $\bar{\mathcal{G}}_{buf}$ is synthesized by multiple gaze points:

$$\begin{aligned} \mathbf{gm}_n &= \sum_{k=1}^{K_{n+1-l_s}} \mathbf{img}_{n+1-l_s,k} + \ldots + \sum_{k=1}^{K_n} \mathbf{img}_{n,k}, \\ l_z &= K_{n+1-l_s} + \ldots + K_n, \end{aligned} \quad (3)$$

where $\mathbf{gm}_n$ is the gaze attention map of $\bar{\mathcal{G}}_{buf}$. It corresponds to the reference scene image $\mathbf{f}_n$. Fig. 2 depicts the procedure for generating the gaze attention map. The gaze attention maps mimic human vision and furnish a more detailed visual attention distribution and information than a single gaze point.

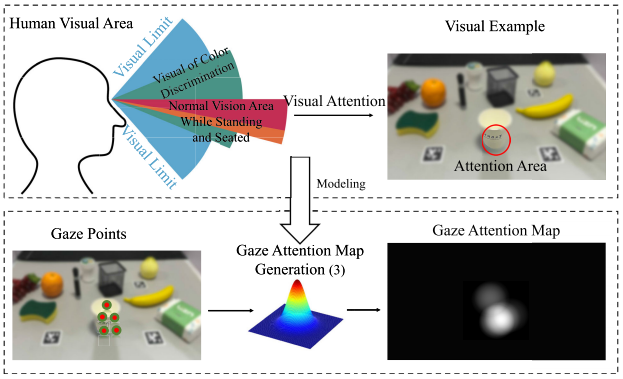


Fig. 2. Gaze attention map generating process. The upper part represents the human visual attributes, and a Gaussian function is used to model this process; the lower part represents the gaze attention map generation process based on the Gaussian function. The gaze attention maps are represented as grayscale images (the Apriltags pasted on the table is not relevant to this paper).

C. Gaze-YOLO

1) Network Architecture: In this subsection, we designed an intention recognition network Gaze-YOLO inspired by YOLO [31], which was a high-efficiency network, but was only applicable to object detection. A spatial pyramid pooling (SPP) module is integrated in Gaze-YOLO to achieve feature fusion of different scales. The detail of the proposed Gaze-YOLO is shown in Fig. S2 in the Supplementary Document.

The gaze attention map $\mathbf{gm}_n$ and the corresponding scene image $\mathbf{f}_n$ are concatenated in the channel dimension as the input of Gaze-YOLO. Gaze-YOLO predicts object boxes on three different scales. The input image is divided into grids on each scale. Then three prediction boxes are generated for each grid. For object detection, each box is responsible for detecting an object, which is composed of bbox coordinates (i.e., t_x, t_y, t_w, t_h), objectness scores p_{obj} (i.e., whether the box contains an object), and object class scores $p_{o,0}, \ldots, p_{o,m_1-1}$ (i.e., possibilities that the object belongs to different classes).

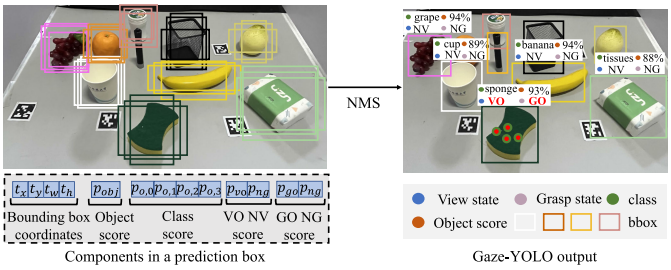


Fig. 3. The left part represents the components in a Gaze-YOLO prediction box. Multiple prediction boxes in the left part are processed by NMS, and then the Gaze-YOLO output in the right part is obtained.

Besides the object detection properties, in the prediction box of each grid, four dimensions are added and denoted as p_{vo}, p_{nv}, p_{go}, and p_{ng}, respectively, as shown in Fig. 3. p_{vo} determines the probability that the object is the IT, while p_{nv} determines the probability that the object is not the IT. The probability of grasp intention on the object is determined by p_{go}, while the probability of no grasp intention on the object is determined by p_{ng}. vo and nv mean viewing object and not viewing object, respectively. go and ng mean grasping object and not grasping object, respectively. It is worth noting that if the subject has grasp intention for the object, the object must be IT.

2) Loss Function: The proposed Gaze-YOLO model generalizes the loss function of YOLO by introducing the losses of grasp intention recognition. The entire loss function is shown in Eq.(4)

$$\begin{aligned} Loss = {} & \alpha_1 L_{\text{coord}} + \alpha_2 L_{\text{obj}} + \alpha_3 L_{\text{cls}} \\ & + \alpha_4 L_{\text{IT}} + \alpha_5 L_{\text{IA}}, \end{aligned} \tag{4}$$

where L_{coord} is the localization loss, L_{obj} is the object loss, and L_{cls} is the class loss. α represents the gain factor of the losses. The object detection losses are inherited from YOLO. Except the object detection losses, we design intention losses, including the IT loss L_{IT} and the IA loss L_{IA}. Both of them use the binary cross-entropy loss, which can be expressed as

$$\begin{aligned} L_{IT} = {} & -\sum_{i=0}^{S^2-1}\sum_{j=0}^{B-1} I_{ij}^{obj} \sum_{v\in\{vo,n\ v\}} \Big[\hat{p}_i(v)\log\left(p_i(v)\right) \\ & + \left(1-\hat{p}_i(v)\right)\log\left(1-p_i(v)\right)\Big], \\ L_{LA} = {} & -\sum_{i=0}^{S^2-1}\sum_{j=0}^{B-1} I_{ij}^{obj} \sum_{g\in\{go,n\ g\}} \Big[\hat{p}_i(g)\log\left(p_i(g)\right) \\ & + \left(1-\hat{p}_i(g)\right)\log\left(1-p_i(g)\right)\Big]. \end{aligned} \tag{5}$$

S^2 denotes the number of grid and B denotes the number of prediction boxes generated by each grid. I_{ij}^{obj} indicates whether the jth prediction box of the ith grid contains an object, and its value is 1 if it does and 0 if it does not. $\hat{p}_i(\cdot)$ represents the intention label. $p_i(\cdot)$ represents the intention prediction. Multiple prediction boxes are processed by Non-Maximum Suppression (NMS) to obtain the Gaze-YOLO output. In our work, there are a total of m_1 classes of objects numbered $0 \sim m_1 - 1$ ($m_1 = 10$). The prediction box with the highest score in each class is selected and output in the NMS process. The VO score vector $\mathbf{P_{vo}} = \left[p_{vo,0}, \ldots, p_{vo,m_1-1}\right]$ and the NV score vector $\mathbf{P_{nv}} = \left[p_{nv,0}, \ldots, p_{nv,m_1-1}\right]$ of all $m_1 - 1$ objects are utilized to determine the IT (i,e., whether or not the subject is looking at this object). When an object is missing from a scene image, the p_{vo} corresponding to the missing object is set to 0 and the p_{nv} is set to 1. The GO score vector $\mathbf{P_{go}} = \left[p_{go,0}, \ldots, p_{go,m_1-1}\right]$ and the NG score vector $\mathbf{P_{ng}} = \left[p_{ng,0}, \ldots, p_{ng,m_1-1}\right]$ are utilized to determine IA (i.e., whether or not the subject wants to grasp this object). As shown in the lower-left part of Fig. 3, "VO" denotes IT and "NV" denotes non-IT. "GO" means the subject wants to grasp this object and "NG" means the subject has no grasp intention for this object.

The user's IT which is recognized by Gaze-YOLO, denoted as $i_{n,1} \in \mathcal{Q}_1 = \{0, \ldots, m_1\}$, where $0 \sim m_1 - 1$ indicate different IT types, and m_1 indicates "no target". $i_{n,1}$ is computed according to

$$\begin{aligned} i_{n,1} = {} & \begin{cases} \arg\max\limits_{j}\left(p_{go,j}\right), & \text{if } p_{go,j} \geq 0.5 \\ m_1, & \text{if } p_{go,j} < 0.5, \end{cases} \\ & j = 0, \ldots, m_1 - 1. \end{aligned} \tag{6}$$

The user's IA which is recognized by Gaze-YOLO, denoted as $i_{n,2} \in \mathcal{Q}_2 = \{0, m_2\}$, where number 0 indicates the intention of grasping, and number $m_2 = 1$ indicates the intention of no grasping. $i_{n,2}$ is computed according to

$$i_{n,2} = \begin{cases} 0, \text{if } p_{go,i_{n,1}} > p_{ng,i_{n,1}} \\ 1, \text{if } p_{go,i_{n,1}} < p_{ng,i_{n,1}} \text{ or } i_{n,1} = m_1. \end{cases} \tag{7}$$

D. Sequential Decision Fusion of Intention Recognition

Considering that the use of short-term information to recognize human intentions is not robust. For instance, people's blinks cause a sudden change in the gaze points; or head movement may cause the camera to capture error images, which will cause errors in intention recognition. An intuitive approach is to fuse temporal information to improve the accuracy of intention recognition, Two HMMs $\lambda_1 = (\mathbf{A}_1, \mathbf{B}_1, \pi_1)$ and $\lambda_2 = (\mathbf{A}_2, \mathbf{B}_2, \pi_2)$ are constructed to describe the transition relationship of IT and IA, respectively. The HMM $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$ consists of a state transition probability matrix $\mathbf{A}$, an emission probability matrix $\mathbf{B}$, and an initial probability π. The initial state is not taken into account for sequential decision fusion in this work. The subject's IT is regraded as the latent state $i_{n,1}$. The subject's IA is regraded as the latent state $i_{n,2}$. The set of possible observations for both HMMs are denoted as $\mathcal{O} = \{\mathbf{o}_0, \ldots, \mathbf{o}_n\}$, where $\mathbf{o}_n$ is the observation; i.e., it is the input sample of Gaze-YOLO. $\mathbf{o}_n = \mathbf{f}_n \oplus \mathbf{gm}_n$, where $\oplus$ represents the concatenation operation of the channel dimension. Since the two HMMs are in a similar form, only the detail of λ_1 is presented in the following paragraphs, but the readers can take it as a reference for both HMMs.

The emission probability $p_1(\mathbf{o}_n \,|i_{n,1})$ is calculated from the output of Gaze-YOLO with Eq. (8).

$$\begin{aligned} & p_{vo,m_1} = 1 - \max\left(p_{vo,j}\right), j = 0, \ldots, m_1 - 1, \\ & \mathbf{p_{vo}} = \text{softmax}\left(\left[p_{vo,0}, \ldots, p_{vo,m_1}\right]\right), \\ & p_1(\mathbf{o}_n \,|i_{n,1} = j) = \mathbf{p_{vo}}\left[j\right], j = 0, 1, .., m_1, \end{aligned} \tag{8}$$

where p_{vo,m_1} denoted the VO score of "no target". Thus the emission probability matrix of observing sample $\mathbf{o}_n$ can be defined as

$$\begin{aligned}\mathbf{B}_1(n) &= \mathbf{p}_1(\mathbf{o}_n \,|i_{n,1}) \\ &= \left[p_1(\mathbf{o}_n \,|i_{n,1} = 0), \ldots, p_1(\mathbf{o}_n \,|i_{n,1} = m_1)\right]. \end{aligned} \quad (9)$$

Since the input samples are determined at each instant, the emission probability matrix is determined and calculated based on the VO score. The estimated category of $i_{n,1}$ may not be robust due to the blinks or the blurred scene image. To make the system tolerant of errors, we introduce a smoothed state $s_{n,1} \in \mathcal{Q}_1$ to substitute $i_{n,1}$, by calculating the average probability distribution in the sliding window:

$$p(s_{n,1} = j) = \begin{cases} p_1(\mathbf{o}_n \,|i_{n,1} = j), & \text{if } n \le l_w \\ \sum_{h=n-l_w+1}^{n} p_1(s_{h,1} = j)/l_w, & \text{if } n > l_w, \end{cases}$$

$$\mathbf{p}_1(s_{n,1}) = \left[p(s_{n,1} = j)\right], j = 0, \ldots, m_1. \quad (10)$$

In HMM, transferring between two adjacent latent states is characterized by transition probability. The transition probabilities between different states constitute the transition probability matrix, which can be constructed from our life experience. Empirically, we have the following assumptions on the transition probability matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, whose elements a_{ij} represent the transition probabilities from the previous state i to the next state j.

For the transition probability matrix $\mathbf{A}_1$ of IT, we have the following empirical rules:

- The probabilities of IT remaining at the same objects are higher than the probabilities of switching to other objects ($a_{ii} > a_{ij}, i \neq j$).
- The probabilities of IT remaining at the same objects are almost equal, and the probabilities of IT staying at no target are lower than the probabilities of staying at a object ($a_{ii} > a_{m_1 m_1}, i = 0, \ldots, m_1 - 1$).
- The probabilities of IT switching from no target to objects are almost the same, which are higher than the probability of IT switching between different objects ($a_{m_1 j} > a_{ij}, i \neq j, i = 0, \ldots m_1 - 1, j = 0, \ldots m_1 - 1$).
- The probabilities of IT switching from other objects to no target are the same, which are higher than the probabilities of IT switching between different objects ($a_{i m_1} > a_{ij}, i \neq j, i = 0, \ldots m_1 - 1, j = 0, \ldots m_1 - 1$).
- The probabilities of IT switching between different objects are almost the same and are the lowest ($a_{ij}, i \neq j, i = 0, \ldots, m_1 - 1, j = 0, \ldots, m_1 - 1$).

For the transition probability matrix $\mathbf{A}_2$ of IA, we have the following empirical rule:

- The probabilities of IA remaining at the same actions are higher than the probabilities of switching to the other action ($a_{ii} > a_{ij}, i \neq j$).

The two transition probability matrices are then initialized as shown in Supplementary Document Fig. S3.

We used the modified Viterbi algorithm [27] to implement the sequential decision and estimate the smoothed state s_n. Due to the similarity of estimating IT and IA, we will only discuss the sequential decisions of IT. The posterior probability distribution of the last smoothed state $s_{n-1,1}$ can be calculated with Eq. (11):

$$\hat{p}_1(s_{n-1,1} = i) = \sum_{j=0}^{m_1} p_1(s_{n-1,1} = i) \times a_{ij,1} \times p_1(\mathbf{o}_n \,|i_{n,1} = j)$$

$$\hat{\mathbf{p}}_1(s_{n-1,1}) = \left[\hat{p}_1(s_{n-1,1} = i)\right], i = 0, 1, \ldots, m_1. \quad (11)$$

Then, the smoothed state with the max probability is chosen as the latest smoothed state

$$s_{n-1,1}^{opt} = \arg\max_i \left(\hat{p}_1\left(s_{n-1,1} = i\right)\right) \quad (12)$$

The posterior probability distribution of the current smoothed state is updated by:

$$\begin{aligned}\hat{p}_1(s_{n,1} = j) &= p_1(s_{n-1,1} = s_{n-1,1}^{opt}) \\ &\quad \times a_{s_{n-1}^{opt} j,1} \times p_1(\mathbf{o}_n \,|i_{n,1} = j), \\ \hat{\mathbf{p}}_1(s_{n,1}) &= \left[\hat{p}(s_{n,1} = j)\right], j = 0, 1, \ldots, , m_1. \end{aligned} \quad (13)$$

Finally, the current smoothed state probability is normalized as:

$$\begin{aligned}\mathbf{p}_1(s_{n,1}) &= \left[p_1(s_{n,1} = j)\right], j = 0, 1, \ldots, m_1, \\ p_1(s_{n,1} = j) &= \hat{p}_1(s_{n,1} = j) / \sum_{i=0}^{m_1} \hat{p}_1(s_{n,1} = i). \end{aligned} \quad (14)$$

The sequential decision fusion process is shown in Fig. 4 and the whole framework is summarized in Algorithm 1.

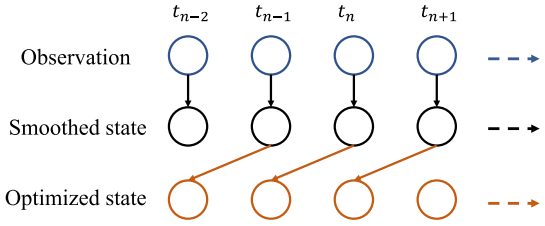


Fig. 4. The sequential decision fusion process. The observed samples are first utilized to compute the smoothed state and then output the optimized state at the latest moment.

To verify the effectiveness of sequential decision fusion, two neural networks, LSTM and GRU, were designed as comparisons. The outputs of Gaze-YOLO were converted into the probability of 10 (objects) × 2 (intentions) + 1 (intention-free) = 21 classes of intentions. These probabilities were used to train LSTM and GRU. Both the LSTM and GRU consisted of an input layer, a hidden layer, and a fully-connected output layer with a softmax activation function. The input feature size and sequence size were set to 21 and l_w, respectively. The hidden layer size was set to 128, and the output layer size was set to 21 to output the probability of each type of intention.

III. Experiments and Results

This study aims to discover the underlying mechanism underpinning grasp intention recognition from gaze. Therefore, we collected data from healthy and hemiplegic subjects and conducted trial-based experiments and subject-based experiments.

Algorithm 1 GIRSDF (HMM)

Input: Frames: $\mathbf{f}_n$, unaligned gaze points on the nth frame: $\mathcal{G}_n = \{\mathbf{g}_n(k) = \{[g_{n,x}(k), g_{n,y}(k)], k = 1, \ldots, K\}$, IT transition probability: $a_{ij,1}$, IA transition probability: $a_{ij,2}$;
Initialize: Sequential fusion sliding window size l_w and gaze map generation sliding window size l_s;
Output: Optimized smoothed state of the last time $s^{opt}_{n-1,1}$ and $s^{opt}_{n-1,2}$, the probability distribution of IT current smooth state $\mathbf{p}(s_{n,1})$, the probability distribution of IA current smooth state $\mathbf{p}(s_{n,2})$;
Obtain aligned gaze points by (1);
if $n \geq l_s$ **then**
    Establish aligned gaze points buff $\bar{\mathcal{G}}_{buf} = \{\bar{\mathcal{G}}_{n+1-l_s}, \ldots, \bar{\mathcal{G}}_n\}$;
    Generate gaze maps by (2)-(3);
    Calculate VO, NV, GO, NG scores:
    $\mathbf{P}_{vo}, \mathbf{P}_{nv}, \mathbf{P}_{go}, \mathbf{P}_{ng} =$ Gaze-YOLO$(\mathbf{f}_n, \mathbf{gm}_n)$;
    Calculate the emission probability $\mathbf{p}_1(\mathbf{o}_n \,|\, i_{n,1})$ distribution using (8)-(9);
    Calculate the smooth state probability distribution $\mathbf{p}_1(s_{n,1})$ using (10)
    Calculate the posterior probability distribution $\hat{\mathbf{p}}_1(s_{n-1,1})$ of the last smooth state by (11);
    Choose the smooth state $s^{opt}_{n-1,1}$ with the max probability as the latest smooth state (12);
    Update the posterior probability distribution $\hat{\mathbf{p}}_1(s_{n,1})$ of the current smooth state using (13);
    Normalizing the current smooth state probability $\mathbf{p}_1(s_{n,1})$ using (14);
    // The sequential decision fusion process for IA is similar to IT.
end if

A. Dataset and Experiment Setup

We conducted visually guided natural grasping and viewing experiments to establish datasets for grasp intention recognition. Seven healthy subjects and two hemiplegic subjects were recruited and instructed to wear a eye-tracker and perform tasks. Their gaze and actions were recorded to build the dataset. Each subject was asked to perform two categories of tasks: grasping and viewing. Gaze-based human-robot interaction often encounters the Midas touch problem [22], [25]. This issue pertains to a situation where, when a user attempts to interact with a target using gaze, two possibilities arise: either the user is "just looking at the target," or the user is "intending to interact with the target." To address the Midas touch problem in the gaze interface, subjects were also instructed to perform an intention-free viewing task. This task merely requires the subjects to look at the object without engaging in any intentional interaction. The data gathered during this task enabled GIRSDF to overcome the Midas touch problem. There were ten objects in our experiments. All participants signed an informed consent that was approved by the ethical committee of UHCT (UHCT-IEC-SOP-016-03-01). Details of the experiments and information of datasets are shown in Supplementary Document Section III. Datasets can be found at Dataset.

TABLE I
GAZE-YOLO PERFORMANCE ON OUR DATASET

	YOLO	Gaze-YOLO
AP@[IoU=0.5 | area=all](%)	98.49±0.12	98.36±0.36
AP@[IoU=0.75 | area=all](%)	88.79±2.21	88.66±2.17
Time (ms/frame)	23.19	23.25
Intention detection (IT and IA)	×	✓

Two different experiments were conducted to verify the effectiveness of GIRSDF.

1) Trial-based experiments: Each subject completed repeated trials of each task. One repetition was selected as the test set, and the rest were used as the training set. To obtain statistically significant results, we utilized a five-fold cross-validation procedure.

2) Subject-based experiments: one subject's data were used as the test set, and the left subjects' data were utilized as the training set. We conducted experiments with each subject left out by turn to verify the intention recognition framework.

We further investigated the impact of data size and diversity on sequential decision fusion. For data size, training and test sets were divided according to trial-based experiments. Specifically, we utilized the $p\%$ ($p = 5,\ 15, \ldots,\ 100$) of data from the training set to train the sequential model. For data diversity, training and test sets were divided according to the subject-based experiments. The data collected from different numbers (1)-(8) of subjects were used to train the sequential model. In addition, to verify the validity of the gaze-attention map, we conducted both comparison and ablation experiments, the details of which are provided in Supplementary Document Section V.

The sliding window size l_s was set to 18 for generating gaze attention maps. Furthermore, w_1 was set to a suitable value so that the maximum value of the gaze attention map pixels was about 255. σ was set to 25 to reduce the pixel value to 0 at a diameter of 90 pixels centered on the gaze point, similar to the clear region of human vision. The sequential fusion sliding window l_w was set to 5, taking about 0.5 seconds to initialize.

B. Object Detection Results of Gaze-YOLO

First we evaluated the performance of Gaze-YOLO on object detection and the results were shown in Table I. From the results, we know that the object detection performance of Gaze-YOLO is close to that of YOLO with only a slight degradation (no statistical difference), but the network can perform intention recognition for each object.

C. Statistical Analysis

Statistical tests were performed in groups. Metrics (accuracy, F1 score, and success rate) corresponding to different parameters (factors) are divided into groups, e.g., accuracy values for different numbers of neurons (e.g., 32 and 64). The trial-based experiment group size was 5 (5 folds), while the subject-based experiment group size was 9 (9 subjects). The Shapiro-Wilk Test was initially conducted on each group

TABLE II

GRASP INTENTION RECOGNITION RESULTS IN THE TRIAL-BASED AND SUBJECT-BASED EXPERIMENTS. THE BOLDED DATA DENOTES THE OPTIMAL RESULTS, AND THE UNDERLINED DATA DENOTES THE SUBOPTIMAL RESULTS. ASTERISKS INDICATE SIGNIFICANT DIFFERENCES COMPARED WITH GIRSDF (HMM)

		Trial-based			Subject-based		
		Accuracy	F1 score	Success rate	Accurcy	F1 score	Success rate
	Fixaton [14]	55.45±2.37*	54.64±2.41*	50.54±4.65*	55.34±5.23*	54.75±5.14*	50.31±4.68*
	VIDEO-Net [24]	64.28±2.65*	63.37±2.84*	42.51±2.95*	60.40±7.28*	57.84±7.54*	43.72±7.98*
	MIDAS-Net [32], [33]	77.37±3.32*	76.48±3.27*	56.53±3.43*	65.40±6.49*	63.67±6.47*	47.94±5.88*
	TAGMM [26]	87.31±3.27*	87.37±3.19*	69.15±3.11*	78.43±4.19*	78.69±3.79*	60.74±4.72*
	GDOD-COMB [12][1]	75.30±3.30*	-	-	-	-	-
Our method	Gaze-YOLO	91.68±1.50*	91.70±1.44*	79.76±2.28*	85.11±3.37*	85.28±2.95*	69.44±5.70*
	GIRSDF (HMM)	94.04±2.05	94.06±2.00	**89.34±2.61**	**88.12±3.13**	**88.13±2.90**	79.48±5.49
	GIRSDF (LSTM)	**94.15±1.51**	**94.18±1.48**	86.83±2.54	87.68±3.59	87.69±3.32	79.70±5.31
	GIRSDF (GRU)	94.08±1.49	94.10±1.45	87.54±2.66	87.81±3.41	87.82±3.18	**79.87±5.18**

[1] This algorithm uses the GITW dataset corresponding to the text. GITW: four healthy subjects, 404 video recordings.
[2] Other algorithms use the Invisible dataset. Invisible: seven healthy subjects and two hemiplegic subjects, 797 video recordings.

of data to test whether it followed a normal distribution. For data following normal distributions, the analysis of variance (ANOVA) was applied to detect whether there was an overall significant difference. Suppose an overall significant difference was found; the T-Test was then conducted to perform the pairwise comparison (other parameters versus reference parameters). For data that did not follow a normal distribution, a non-parametric test (Kruskal-Wallis H Test) was performed to check for overall significant differences among groups. Suppose an overall significant difference was found; the Wilcoxon signed-rank Test was subsequently performed for pairwise comparisons. The differences were considered significant if $p < 0.05$ was achieved.

D. Trial-Based Grasp Intention Recognition

Three metrics including success rate, accuracy, and F1 score are introduced as evaluation metrics. The success rate quantified the proportion of successful trials to the total number of trials. A successful trial means no errors occur from the moment when the correct intention is identified to the end of the trial. The performance of our GIRSDF framework was compared with other approaches. Additionally, three sequential decision fusion strategies were compared.

The grasp intention recognition results of trial-based experiment are shown in Table II and Fig. 5. The proposed GIRSDF outperforms other gaze-based grasp intention recognition methods in trial-based experiments. The best accuracy achieved by GIRSDF is 94.15% (LSTM), and the best success rate of GIRSDF is 89.34% (HMM). The confusion matrix depicting the intention recognition results is displayed in Fig. 6. The utilization of various decision fusion methods eliminated some of the errors and increased the accuracy of most classes. Consequently, the overall success rate of the framework is improved. Here, only a comparison of the two methods is presented. For additional methods and their respective confusion matrices, please refer to Fig. S5 and Fig. S6 in the Supplementary Document. The results indicate that all three sequential decision fusion methods significantly improve the performance of GIRSDF compared with Gaze-YOLO (all $p < 0.01$), especially the success rate. This improvement can be attributed to sequential decision fusion can effectively correct unexpected intention recognition

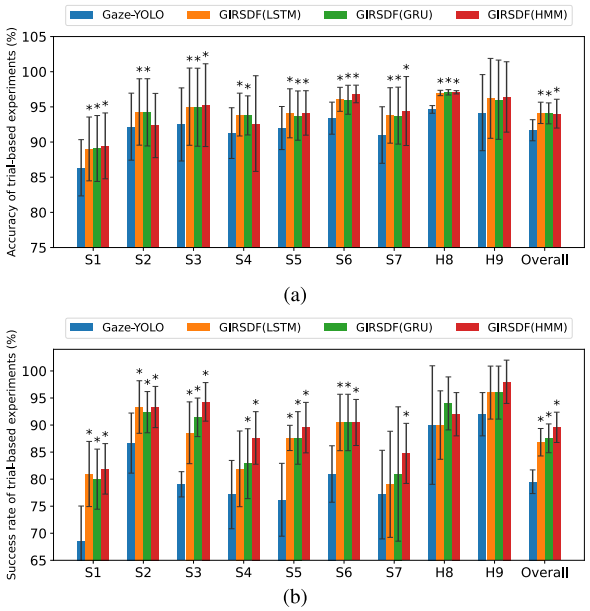
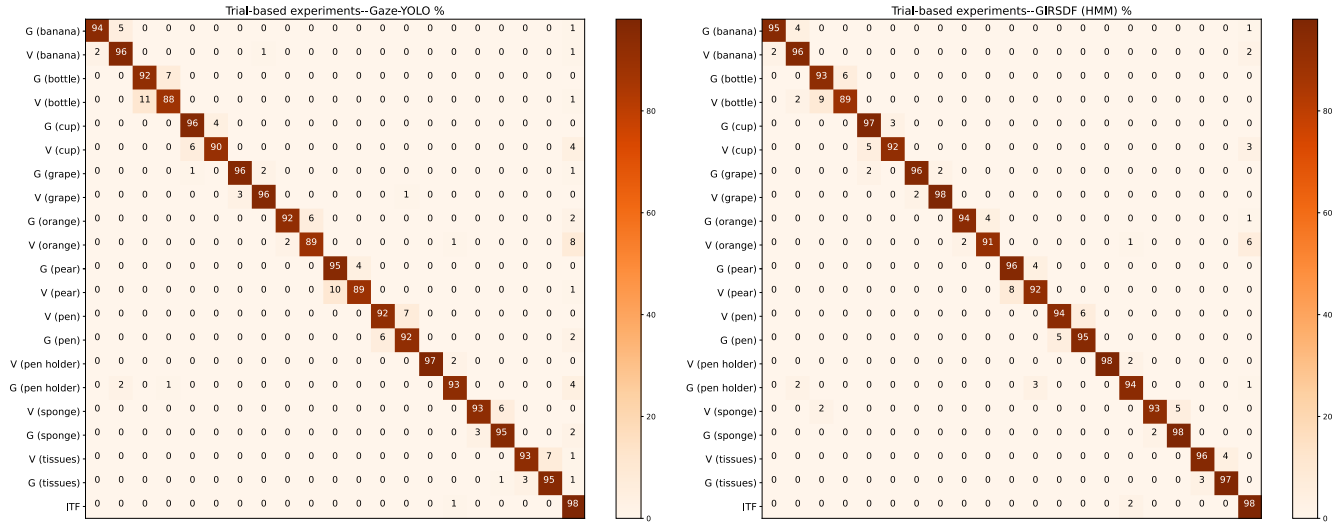


Fig. 5. The accuracy and success rate in the trial-based experiments with the Invisible dataset. Error bars represent mean ± one standard deviation in five repetitions. Asterisks indicate significant differences compared with Gaze-YOLO. Seven healthy subjects (S1-S7) and two hemiplegic subjects (H8-H9) participated in the experiments.

errors, which may arise from blurred scene images or outliers in the gaze points. Moreover, there is no significant difference between the results of the three sequential decision methods (accuracy: $p = 0.93$; F1 score: $p = 0.93$; success rate: $p = 0.39$). This result suggests that both the training-free HMM and trained LSTM and GRU can optimize intention recognition and achieve comparable performance. Notably, the HMM has a lower computational burden than LSTM and GRU, owing to its simpler design. Although the advantage is not obvious, the proposed GIRSDF has advantages over LSTM and GRU in that HMM is simple in design and of a low computational burden.

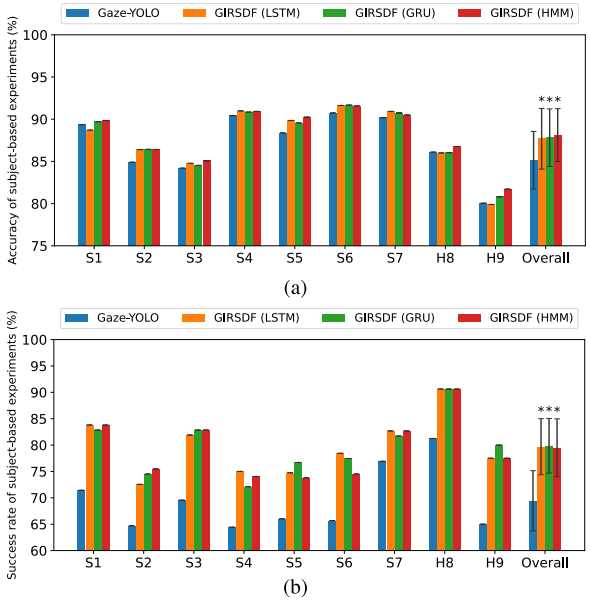
E. Subject-Based Grasp Intention Recognition

The subject-based experiment results are provided in Table II. The accuracy and success rate of each subject are presented in Fig. 7. The proposed GIRSDF achieves the best



(a) (b)

Fig. 6. The confusion matrices for Gaze-YOLO and GIRSDF (HMM) in the trial-based experiments. G and V are the abbreviations of grasp and view. ITF is the abbreviations of intention-free.



(a)

(b)

Fig. 7. The accuracy and success rate in the subject-based experiments. Error bars for the overall results represent mean ± one standard deviations in different subjects. Asterisks indicate significant differences compared with Gaze-YOLO.

accuracy of 88.12% (HMM) and the best success rate of 79.87% (GRU), demonstrating similar characteristics between grasp intentions and gaze among different subjects, including hemiplegic patients (H8 and H9). The results for hemiplegic patients H8 (acc: 86.77%; success rate: 90.63%) and H9 (acc: 81.72%; success rate: 80.00%) demonstrate the feasibility of using gaze to recognize grasp intentions for people who retain eye-movement control ability.

Compared with Gaze-YOLO, the accuracy of GIRSDF improved by 3.01% (HMM: $p < 0.01$), 2.75% (LSTM: $p < 0.01$), and 2.70% (GRU: $p < 0.01$), respectively, after incorporating different sequential decision fusion strategies. Similarly, the success rate improved by 10.04% (HMM: $p < 0.01$), 10.26% (LSTM: $p < 0.01$), and 10.43% (GRU: $p < 0.01$), respectively. This result proves the effectiveness of sequential decision fusion in enhancing grasp intention recognition performance. Moreover, there is no overall significant difference in the results of the three sequential decision fusions methods (accuracy: $p = 0.96$; F1 score: $p = 0.95$; success rate: $p = 0.98$).

Although the performance of GIRSDF is degraded compared to trial-based experiments, the proposed GIRSDF still outperforms other gaze-based grasp intention recognition methods. The degradation of grasp intention recognition performance in the cross-subject case is caused by a lack of data from the test subject when tuning the model.

F. Intention Recognition Results in Healthy and Hemiplegic Subjects

In this subsection, the variability in intention recognition on healthy and hemiplegic subjects are analyzed. In the trial-based experiments, the results on healthy subjects are combined with those on the hemiplegic subjects H8 and H9, respectively, for statistical analysis. The results are presented in Table III.

In the trial-based experiments, no significant differences are found between the hemiplegic and healthy subjects for most metrics. Interestingly, the hemiplegic subjects exhibited higher accuracy, F1 score, and success compared with the healthy subjects, indicating more remarkable behavioral similarity between them. During our experiments, we discovered that the hemiplegic patients demonstrated similar behaviors when performing the same task. In the subject-based experiments, the accuracy and F1 scores on H8 and H9 are lower than those on the healthy subjects, which can be explained in two aspects. First, the hemiplegic patients may exhibit different visual behaviors during grasping and viewing tasks compared to the healthy individuals. Second, the calibration accuracy of the eye-tracker varies between subjects, which results in differences in the recorded data even though the subjects' visual behaviors are similar. The closer the calibration accuracy is

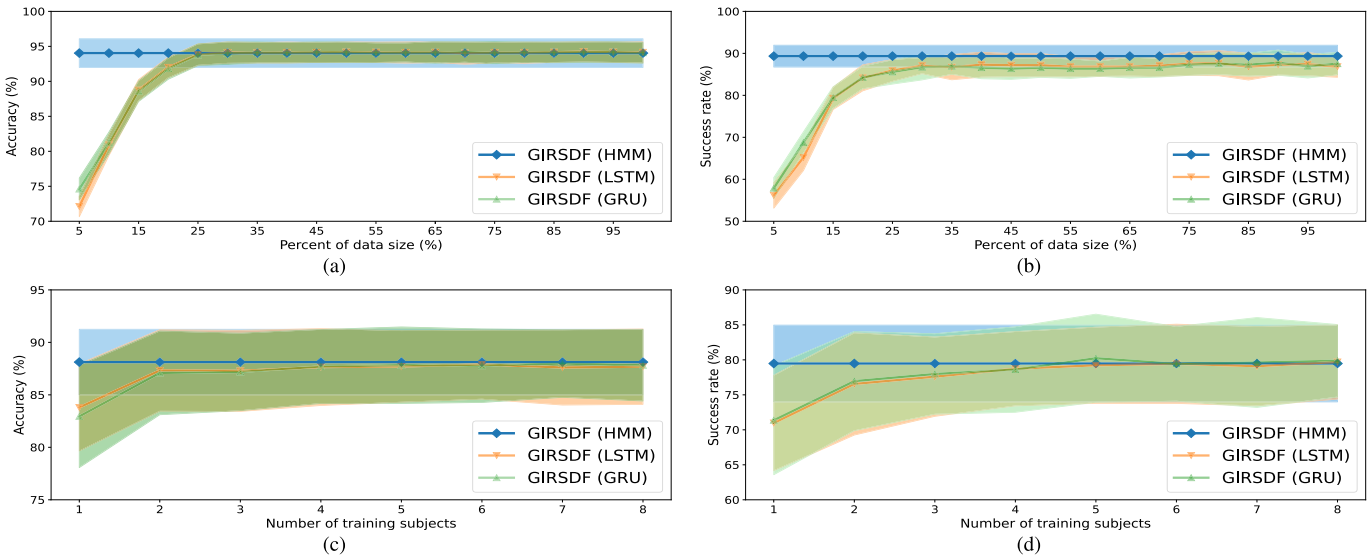


Fig. 8. Accuracy and success rates for different data size and diversity. The shaded area indicates the one standard deviation.

to that of the healthy subjects, the greater possibility that the accuracy will be high. However, only two hemiplegic patients participated in our experiment. To gain a deeper understanding of the dissimilarity between hemiplegic and healthy subjects, further research with a number of hemiplegic patients is needed.

G. Comparison of Data Size and Data Diversity

The average accuracy and F1 scores with different data sizes and diversities are shown in Fig. 8. When the percentage of training data is increased from 5% to 20%, the accuracy and success rate improve significantly (LSTM accuracy:$p < 0.01$, success rate: $p < 0.01$; GRU accuracy:$p < 0.01$, success rate: $p < 0.01$) as shown in Fig. 8(a) and (b). This result demonstrates that LSTM and GRU are sensitive to training data size. When the data size exceeds 20%, the performance improvement is not significant (LSTM accuracy:$p = 0.08$, success rate: $p = 0.22$; GRU accuracy:$p = 0.07$, success rate: $p = 0.12$). There is no overall significant difference (accuracy: $p = 0.93$; F1 score: $p = 0.93$; success rate: $p = 0.39$) between the performance of the three sequential decision fusion methods when the 100% of the training data is utilized.

As shown in Fig 8(c) and (d), the performance of LSTM and GRU is significantly improved when the training data of more than one subjects is utilized (LSTM accuracy:$p = 0.03$, success rate: $p < 0.01$; GRU accuracy:$p = 0.03$, success rate: $p = 0.02$). The reason for the poor performance when the models are trained on only one subject's data may be the insufficient size of the training data. After the training data with more than two subjects, the performance of LSTM and GRU improves but is not statistically significant (LSTM accuracy:$p = 0.75$, success rate: $p = 0.34$; GRU accuracy:$p = 0.70$, success rate: $p = 0.36$). This indicates that increasing data diversity (data from different subjects) has no appreciable effect on the results of intention recognition. There is no significant difference (accuracy: $p = 0.96$; F1 score: $p = 0.95$; success rate: $p = 0.98$) between the performance of the three sequential decision fusion methods when the training subjects are all the remaining subjects.

The experimental results demonstrate that LSTM and GRU require training and are data-intensive. When the training data is sufficient, LSTM and GRU perform well. Additionally, LSTM and GRU are not user-specific, and the models trained on various trainging subjects can be applied to the test subjects. HMMs utilize pre-built models that do not require explicit training and are suitable for situations with insufficient training data. A suitable sequential decision fusion method can be selected according to the training data size for good grasp intention recognition performance.

H. Time Consumption

The intention recognition framework's time consumption is quantified and summarized in Table S1 in the Supplementary Document. Gaze-YOLO's training times are approximately 14 mins per epoch, while LSTM and GRU's are approximately 11.7 seconds. HMMs do not require training. After training, the inference times of LSTM, GRU, and HMM are 3.4 ms, 3.39 ms, and 0.02 ms per frame, respectively. The results reveal that HMM has the lowest computational complexity compared with LSTM and GRU. Considering the gaze attention map generation, the proposed GIRSDF runs with a frequency of about 22 Hz, which can satisfy the real-time requirements.

IV. Discussion

A. Datasets

In the Invisible dataset, ten representative kinds of objects were chosen. Different subjects, including healthy individuals and hemiplegics, participated in the experiment. Notably, our dataset only contains one object for each class. If multiple similar objects exist, it is difficult for the annotator to correctly identify which one among them is to be grasped during the annotation process. Therefore, different kinds of objects are selected to speed up the labeling process.

TABLE III
RESULTS OF GRASP INTENTION RECOGNITION ON THE HEMIPLEGIC PATIENTS AND HEALTHY SUBJECTS. ASTERISKS INDICATE SIGNIFICANT DIFFERENCES ($p < \mathbf{0.05}$) FROM THE HEALTHY SUBJECTS. NO STATISTICAL ANALYSIS RESULTS ARE AVAILABLE FOR THE SUBJECT-BASED EXPERIMENTS

		Trial-based			Subject-based		
		Accuracy	F1 score	Success rate	Accuracy	F1 score	Success rate
GIRSDF (HMM)	Healthy subjects	93.58±5.25	92.67±6.26	88.84±6.12	89.23±2.29	89.15±2.30	78.17±4.32
	H8	97.07±0.23	97.09±0.23	92.00±4.00	86.77	86.35	90.63
	H9	96.41±5.00	97.11±3.68*	98.00±4.00*	81.72	82.82	77.50
GIRSDF (LSTM)	Healthy subjects	93.74±4.52	93.11±5.20	85.99±7.96	89.04±2.39	88.93±2.44	78.45±4.12
	H8	96.97±0.38	97.03±0.32	90.00±6.32	86.01	85.51	90.62
	H9	96.21±5.68	96.98±4.22	96.00±4.90*	79.90	81.19	77.50
GIRSDF (GRU)	Healthy subjects	93.65±4.55	93.02±5.25	86.53±8.17	89.07±2.42	88.98±2.46	78.32±3.95
	H8	97.07±0.38	97.13±0.32	94.00±4.90	86.04	85.52	90.63
	H9	96.02±5.63	96.70±4.36	96.00±4.90*	80.81	82.01	80.00

For all samples of each trial, we assigned the identical intention label. In practice, subjects may not gaze at the target object for a short period of time before the start or after the end of the trial (e.g., the user's gaze is not on the target object at first but moves quickly to the target object from elsewhere after locating it and then performing the task). It is possible to reduce the weights of these samples to eliminate the potential effects on the training process.

B. GIRSDF

Considering the possible connection between gaze and grasp intentions, we designed GIRSDF for grasp intention recognition. Trial-based and subject-based experiments on Invisible dataset were organized to demonstrate the generalization performance and effectiveness of the framework. The proposed framework only relies on gaze to recognize grasp intentions and does not require the user to learn specific behaviors, which is promising to be applied to hemiplegics and the elderly.

The gaze maps generated by all the gaze points are combined to create the final gaze map, which effectively reduces the influence of abnormal gaze points. When outliers occur, the pixel values in the generated gaze map are extremely low, among which the maximum is roughly 2. This low value has a smaller effect than the gaze maps generated from the normal gaze points, which makes Gaze-YOLO insensitive and robust to outliers.

GIDSDF can be easily extended to handle multiple objects. By augmenting other object categories in the dataset (the non-intention samples are labeled as NV and NG), Gaze-YOLO can adapt to scenes containing more kinds of objects. The transition probability matrix of HMM presented in this study is constructed using a fixed number of objects. It is possible to develop an adaptive HMM construction approach by combining the number of detected objects in the scene with the rules. The transition probability matrix is built on empirical rules that are interpretable and valid. Compared to LSTM and GRU models, HMMs impose a low computational burden and do not require training. With sufficient training data, LSTM and GRU may achieve better performance. A suitable sequential decision fusion method can be selected for optimal grasp intention recognition performance according to the training data size.

It is notable that there is an apparent variation in the intention recognition accuracy in the trial-based experiments. This phenomenon is because the visual behavior of subjects may vary across replicate trials, which makes the trials with similar behavior highly accurate and the rest less accurate. The results of grasp intention recognition validated the effectiveness of GIRSDFD and demonstrated its applicability for assistive robot control. As reported in [14] and [15], the gaze-based grasp assistive robot will execute the grasping action after detecting successive identical intentions. With this premise, the success rate can reach 100%. The subject-based experiments further verify the GIRSDF's generalization ability and the existence of subject-to-subject similarity in gaze behavior. Even on hemiplegic subjects, satisfactory results are obtained. The generalization ability minimizes the need for new users' data and offers the possibility of recognizing the grasp intentions of new users, whose data are often difficult to obtain. In addition, there is no significant difference in the eye movements between healthy people and hemiplegic patients [34]. Therefore it is possible to apply the trained model to patients.

C. Limitations and Future Works

Although the proposed GIRSDF achieves the optimal grasp intention recognition results and good generalization, there are some limitations. As shown in the confusion matrices of Fig. S5 and S6 in Supplementary Document, most grasp intention recognition errors are the different IAs of the same IT (e.g., grasp cup and view cup are misidentified). This is due to the fact that vision is typically capable of reliably recognizing IT, but variations in the gaze signal can lead to IA recognition errors. Inspired by EEG signals in intention detection [5], we plan to incorporate EEG signals to improve IA identification ability. Second, GIRSDF has not been applied to control the assistive robot. In practical applications, a depth camera or a pose detection network [35] will be utilized to determine the position of the intentional target to accomplish the assistive grasping tasks.

V. CONCLUSION

In this work, a gaze-based generic framework GIRSDF is proposed for grasp intention recognition and performing sequential decision fusion. This framework consists of a gaze attention map generation approach, a Gaze-YOLO grasp intention recognition model, and sequential decision fusion models. A dataset Invisible containing healthy and hemiplegic subjects'

data is established to validate the performance of GIRSDF. Trial-based and subject-based experiments demonstrate the framework's effectiveness and generalization ability for grasp intention recognition. The experimental results further revel the similarity of different subjects' gaze behavior and grasp intention. Experiments on data size and data diversity illustrate the sensitivity of LSTM and GRU to data size. HMM employs pre-designed models that do not require training. The proposed framework can run at a frequency of about 22 Hz, which can satisfy the need for real-time intention recognition. Future work includes fusing EEG signals to improve intention recognition performance and applying GIRSDF to control the assistive robot for validation and evaluation.

REFERENCES

[1] Y. Huang, K. B. Englehart, B. Hudgins, and A. D. C. Chan, "A Gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 11, pp. 1801–1811, Nov. 2005.

[2] B. Yang, J. Huang, X. Chen, C. Xiong, and Y. Hasegawa, "Supernumerary robotic limbs: A review and future outlook," *IEEE Trans. Med. Robot. Bionics*, vol. 3, no. 3, pp. 623–639, Aug. 2021.

[3] T. Lenzi, S. M. M. De Rossi, N. Vitiello, and M. C. Carrozza, "Intention-based EMG control for powered exoskeletons," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 8, pp. 2180–2190, Aug. 2012.

[4] D. Xiong, D. Zhang, X. Zhao, and Y. Zhao, "Deep learning for EMG-based human-machine interaction: A review," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 3, pp. 512–533, Mar. 2021.

[5] J. Wang, L. Bi, W. Fei, and C. Guan, "Decoding single-hand and both-hand movement directions from noninvasive neural signals," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 6, pp. 1932–1940, Jun. 2021.

[6] M. Jochumsen, I. Khan Niazi, D. Taylor, D. Farina, and K. Dremstrup, "Detecting and classifying movement-related cortical potentials associated with hand movements in healthy subjects and stroke patients from single-electrode, single-trial EEG," *J. Neural Eng.*, vol. 12, no. 5, Aug. 2015, Art. no. 056013.

[7] H. Zhang, S. Wu, W. Chen, Z. Gao, and Z. Wan, "Self-calibrating gaze estimation with optical axes projection for head-mounted eye tracking," *IEEE Trans. Ind. Informat.*, early access, May 23, 2023, doi: 10.1109/TII.2023.3276322.

[8] W. W. Abbott and A. A. Faisal, "Ultra-low-cost 3D gaze estimation: An intuitive high information throughput compliment to direct brain–machine interfaces," *J. Neural Eng.*, vol. 9, no. 4, Jul. 2012, Art. no. 046016.

[9] E. A. Corbett, K. P. Körding, and E. J. Perreault, "Real-time evaluation of a noninvasive neuroprosthetic interface for control of reach," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 21, no. 4, pp. 674–683, Jul. 2013.

[10] B. C. Goldwater, "Psychological significance of pupillary movements," *Psychol. Bull.*, vol. 77, no. 5, pp. 340–355, 1972.

[11] G. Binsted, R. Chua, W. Helsen, and D. Elliott, "Eye–hand coordination in goal-directed aiming," *Human Movement Sci.*, vol. 20, nos. 4–5, pp. 563–585, Nov. 2001.

[12] I. González-Díaz, J. Benois-Pineau, J.-P. Domenger, D. Cattaert, and A. de Rugy, "Perceptually-guided deep neural networks for ego-action prediction: Object grasping," *Pattern Recognit.*, vol. 88, pp. 223–235, Apr. 2019.

[13] J. Jiang, Z. Nan, H. Chen, S. Chen, and N. Zheng, "Predicting short-term next-active-object through visual attention and hand position," *Neurocomputing*, vol. 433, pp. 212–222, Apr. 2021.

[14] A. Shafti, P. Orlov, and A. A. Faisal, "Gaze-based, context-aware robotic system for assisted reaching and grasping," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 863–869.

[15] M.-Y. Wang, A. A. Kogkas, A. Darzi, and G. P. Mylonas, "Free-view, 3D gaze-guided, assistive robotic system for activities of daily living," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 2355–2361.

[16] B. Yang, J. Huang, M. Sun, J. Huo, X. Li, and C. Xiong, "Head-free, human gaze-driven assistive robotic system for reaching and grasping," in *Proc. 40th Chin. Control Conf. (CCC)*, Jul. 2021, pp. 4138–4143.

[17] S. Li, X. Zhang, and J. D. Webb, "3-D-gaze-based robotic grasping through mimicking human visuomotor function for people with motion impairments," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 12, pp. 2824–2835, Dec. 2017.

[18] P. M. Tostado, W. W. Abbott, and A. A. Faisal, "3D gaze cursor: Continuous calibration and end-point grasp control of robotic actuators," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 3295–3300.

[19] N. E. Krausz, D. Lamotte, I. Batzianoulis, L. J. Hargrove, S. Micera, and A. Billard, "Intent prediction based on biomechanical coordination of EMG and vision-filtered gaze for end-point control of an arm prosthesis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 6, pp. 1471–1480, Jun. 2020.

[20] L. Chen et al., "Real-time gaze tracking with head-eye coordination for head-mounted displays," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2022, pp. 82–91.

[21] B. Yang and J. Huang, "Outlier-robust gaze signal filtering framework based on eye-movement modality recognition and set-membership approach," *IEEE Trans. Biomed. Eng.*, early access, Feb. 27, 2023, doi: 10.1109/TBME.2023.3249233.

[22] S. Li and X. Zhang, "Implicit intention communication in human–robot interaction through visual behavior studies," *IEEE Trans. Hum.-Mach. Syst.*, vol. 47, no. 4, pp. 437–448, Aug. 2017.

[23] F. Koochaki and L. Najafizadeh, "A data-driven framework for intention prediction via eye movement with applications to assistive systems," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 974–984, 2021.

[24] D. Kim et al., "Eyes are faster than hands: A soft wearable robot learns user intention from the egocentric view," *Sci. Robot.*, vol. 4, no. 26, Jan. 2019, Art. no. eaav2949.

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

[26] B. Yang, J. Huang, X. Chen, X. Li, and Y. Hasegawa, "Natural grasp intention recognition based on gaze in human–robot interaction," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 4, pp. 2059–2070, Apr. 2023.

[27] K. Zhang, W. Zhang, W. Xiao, H. Liu, C. W. De Silva, and C. Fu, "Sequential decision fusion for environmental classification in assistive walking," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 9, pp. 1780–1790, Aug. 2019.

[28] P. Gulati, Q. Hu, and S. F. Atashzar, "Toward deep generalization of peripheral EMG-based human–robot interfacing: A hybrid explainable solution for NeuroRobotic systems," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 2650–2657, Apr. 2021.

[29] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making sense of spatio-temporal preserving representations for EEG-based human intention recognition," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3033–3044, Jul. 2020.

[30] H. Strasburger, I. Rentschler, and M. Juttner, "Peripheral vision and pattern recognition: A review," *J. Vis.*, vol. 11, no. 5, p. 13, Dec. 2011.

[31] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[32] A. Shafti, P. Festor, P. Orlov, M. Li, and A. A. Faisal, "Deep learning human action intention classification from natural eye movement patterns," *J. Vis.*, vol. 21, no. 9, p. 2715, Sep. 2021.

[33] P. Festor, A. Shafti, A. Harston, M. Li, P. Orlov, and A. A. Faisal, "MIDAS: Deep learning human action intention prediction from natural eye movement patterns," 2022, *arXiv:2201.09135*.

[34] K. Nagai et al., "Multimodal visual exploration disturbances in Parkinson's disease detected with an infrared eye-movement assessment system," *Neurosci. Res.*, vol. 160, pp. 50–56, Nov. 2020.

[35] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial PoseNet: A structure-aware convolutional network for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1221–1230.