2023 IEEE International Conference on Development and Learning (ICDL)
November 9-11, Macau, China

# What You See is What You Grasp: User-friendly Grasping Guided by Near-eye-tracking

Shaochen Wang*, Wei Zhang*, Zhangli Zhou*, Jiaxi Cao, Ziyang Chen, Kang Chen, Bin Li, Zhen Kan

***Abstract*— This study introduces an advanced human-robot interface designed to discern and execute manipulation tasks based solely on visual cues. The interface combines eye-tracking technology and robotic manipulation, facilitating actions like grasping or pick-and-place tasks. We have developed a head-mounted device for tracking eye movements, allowing the system to determine the user's focus and initiate sight-driven manipulation. Enhancing grasping efficiency, the system incorporates a transformer-based model, utilizing attention blocks for feature extraction and optimizing both channel capacity and spatial resolution of the feature maps. Our experiments confirm the system's capability in aiding users to perform tasks using only their gaze, suggesting significant implications for assistive robots in helping people with upper limb disabilities or the elderly with everyday activities.**

## I. Introduction

Modern advancements in artificial intelligence have catalyzed the extensive integration of robots [1]–[3] into numerous sectors, encompassing both industrial operations and routine daily activities. These systems are evolving beyond basic, repetitive functions, incorporating the ability to interpret human intentions for enhanced user assistance. This development highlights the necessity of creating interfaces that effectively convey human intentions to robots, thereby fostering a more integrated human-machine interaction [4].

Conventional manipulation techniques [5], [6], often reliant on joysticks, pose challenges for the elderly and those with upper limb disabilities. While recent strides in wearable technology, particularly brain-computer interfaces (BCI) [7], show promise in robotic assistance, they have limitations. Invasive BCI systems, requiring surgical implantation of microelectrodes in the cerebral cortex, carry inherent risks. Additionally, non-invasive BCIs are prone to noise interference and often come with high costs. Consequently, there is an urgent need for developing a new, both safe and user-friendly, human-robot interface.

Vision is the primary sensory input for humans, with over 80% of information obtained through it. This fact drives the advancement of eye-based robotic assistive systems for manipulation tasks. Incorporating eye-tracking technology, these systems show promise in diverse fields such as surgical diagnostics, rehabilitation, and research. They serve as an instinctive interface for those with physical disabilities, given that such conditions seldom impact vision. Despite these advantages, the widespread adoption of eye-tracking in practical applications remains limited. This is primarily due to challenges in precisely modeling eye motion to accurately determine gaze points. Most current eye-tracking robotic systems [8] use stationary cameras and are designed for desktop environments, limiting their broader applicability.

This study presents an innovative human-robot interface capable of interpreting and executing manipulation tasks based solely on visual input. Our integrated system combines near-eye tracking with robotic manipulation, facilitating user-directed actions such as grasping and pick-and-place activities. Our design is a head-mounted gadget that precisely monitors real-time eye movements to pinpoint the user's focus. Additionally, our transformer-based grasp detection framework enhances the robot's ability to perceive and execute user-directed manipulations. This framework employs self-attention to understand spatial relationships among pixels and a feature fusion pyramid to amalgamate multi-scale features for accurate grasping pose determination. Experimental results show low error rates in gaze estimation and high performance in various grasping tasks with our system.

The contribution of this work can be summarized as follows:

- We've developed a user-specific visual robotic aid, featuring a head-mounted intent detection system and a self-attention enhanced grasping tool for manipulative tasks.
- Introduction of an innovative human-robot interface that facilitates intuitive manipulation solely through eye-tracking technology.
- Comprehensive experimental validation confirming the efficiency of our robotic assistive system in various manipulation scenarios.

## II. Related Work

Robotic manipulation [9], [10] is a critical skill with applications spanning manufacturing, industry, and medicine. [11] addresses grasping's non-stationary dynamics to enhance success in grasp maneuvers. Concurrently, there is an increasing merger of reinforcement learning [12], [13] with robotics. Extensive research has focused on vision-based grasping methods. Pioneering this field, deep learning for grasp detection was first introduced by Lenz et al. [14]. Redmon et al. [15] later utilized a CNN for robotic grasp pose determination. Additionally, Morrison et al. [16]

* Contribute equally

This work was supported in part by National Key R&D Program of China under Grant 2022YFB4701400/4701403 and National Natural Science Foundation of China under Grant 62173314.

The authors are with the University of Science and Technology of China, Hefei, 230026, China.

Zhen Kan is the corresponding author (zkan@ustc.edu.cn).

2023 IEEE International Conference on Development and Learning (ICDL) | 978-1-6654-7075-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDL55364.2023.10364491

developed a generative grasping CNN that uses depth data to generate grasp candidates.

Assistive robotic arms, aiding users with upper limb impairments in everyday tasks like object grasping and water pouring, are gaining popularity. Traditional control methods, such as joysticks, present challenges for elderly or upper limb disabled individuals. Conversely, visual interaction offers a more intuitive approach for those with physical, mobility, or speech impairments to operate robotic systems. Eye-tracking technology, evolving considerably over the past century, now enables manipulation of robotic devices through gaze tracking.

The concept of the attention mechanism draws inspiration from human visual perception, highlighting the significance of sight in directing attention. Hollenstein et al. [17] enhanced an annotation model by incorporating human sight, demonstrating its effectiveness in conveying semantic information for entity models. Similarly, in computer vision, Karessli et al. [18] utilized sight as an auxiliary task, notably improving zero-shot task classification accuracy. Eye-tracking technology's application extends to augmented reality [19], deep learning [20], [21], and mixed reality [22]. These advancements serve as a foundation for our research in employing sight for human-robot manipulation.

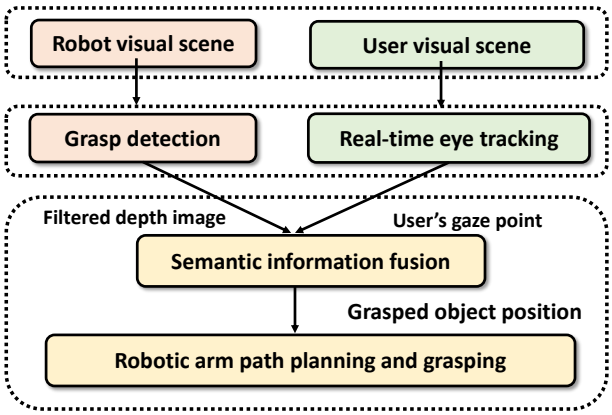
## III. Method

### A. System Overview

This section outlines the integration of near-eye tracking with robotic manipulation for enhanced human-robot interaction. The system utilizes a head-mounted eye-tracking device, enabling users to control a robotic arm for object manipulation using their sight. Fig. 1 depicts the process of interpreting human eye gaze into robotic actions. The methodology comprises three key steps: i) Utilizing a head-mounted near eye-tracker equipped with economical cameras to track the people's gaze. Combining a biological representation of the human eye using computer vision techniques, this procedure identifies the orientation of the pupil and precisely locates the gaze coordinates within the three-dimensional space. ii) Concurrently, an advanced hierarchical transformer-based visual model has been devised to enhance feature extraction during grasping tasks by harnessing the power of attention for global perception. A feature pyramid within the transformer network assimilates multi-scale sensory data, culminating in an optimal grasping configuration. iii) These subsystems merge their data, utilizing a grasping quality maps to process the gaze point and human attention, ultimately deriving the grasping pose configurations for the intended object.

### B. Eye-tracking System

As a sophisticated optical organ, human eyes generate images on the retina through complex reflection, utilizing its refractive elements (cornea, lens and vitreous humor). These elements, each with a distinct refractive index, create varied refractive surfaces, thus enhancing the complexity of the optical system. For modeling purposes, corneal refraction




:

Fig. 1. System Pipeline Overview: The subsystem dedicated to detecting grasps is denoted by the pink section, while the module for eye-tracker is represented in green. The fusion module, highlighted in yellow at the bottom, is where objects are chosen for grasping based on user gaze.

during pupil imaging is typically simplified, treating the cornea as a uniformly curved sphere. Our approach adopts the model from Nagamatsu et al. [23], which conceptualizes the eye using two concentric ellipsoids - a larger one for the ocular body and a smaller, rotatable ellipsoidal plane for the cornea. The pupil serves as the primary light entry channel, with its orientation reflecting the eyeball's rotation and the gaze direction. Central pupil coordinates are essential for accurate sight tracking. Our methodology for locating the pupil center comprises two phases: coarse and refined localization. The initial phase applies radial symmetry transformations for rapid pupil identification, excluding anomalies like blink-induced distortions. Following this, we apply Canny edge detection, along with edge filtering, to accurately pinpoint the center of the pupil with refined localization.

Based on the pupil center coordinates obtained through eye image processing and rooted in the principles of camera imaging theory [24], the positioning of pupil centers is determined by the following equation:

$$\begin{bmatrix} u_p \\ v_p \\ 1 \end{bmatrix} = \frac{1}{P_z} A \begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} = \frac{1}{P_z} \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix}. \tag{1}$$

In near-eye image analysis, the estimated gaze location is represented as $[u_p, v_p]$, while in the world coordinate framework, it's denoted as $[P_x, P_y, 1]$. The transformation matrix $P_z$ and the camera's intrinsic matrix $A$, compatible with OpenCV [25], facilitate this coordinate translation. For simplicity, the image plane-camera distance is typically set to 1, which succinctly defines the pupil's 3D coordinates, leading to the following expression for the pupil's 3D coordinates:

$$\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} = \begin{bmatrix} P_x \\ P_y \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{u_p - c_x}{f_x} \\ \frac{u_p - c_y}{f_y} \\ 1 \end{bmatrix}. \tag{2}$$

In a similar fashion, the central point of the corneal reflection

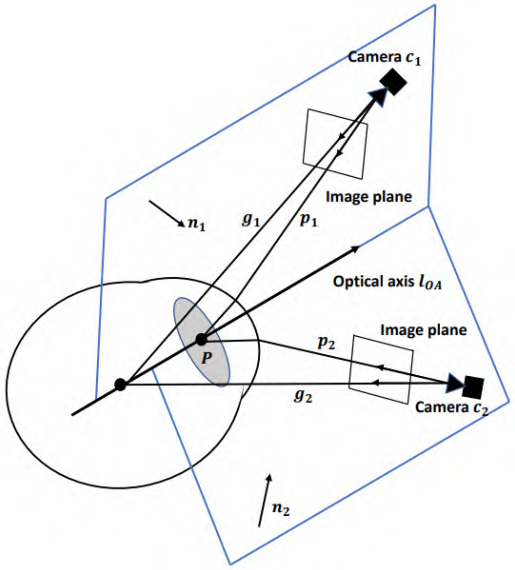



Fig. 2. Illustration of the near-eye-tracking.

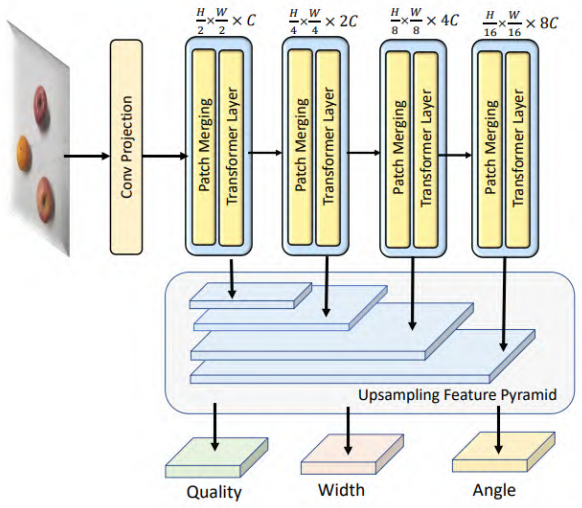



Fig. 3. Overview of Transformer-Based Grasp Detection Model.

is located using 3D coordinates. Vectors $p_1$ and $p_2$ are derived from the line of sight joining the center of the camera's lens to the centers of the pupils, while vectors $g_1$ and $g_2$ originate from the line connecting the camera's center to the corneal reflection's center. Given the ocular optic axis, the corneal reflection, and the camera center all lie in the same plane, the corresponding normal vectors of these planes, depicted in Fig. 2, are computed as

$$\begin{cases} \mathbf{n_1} = \frac{\mathbf{p_1}\times\mathbf{g_1}}{|\mathbf{p_1}\times\mathbf{g_1}|}, \ \mathbf{n_2} = \frac{\mathbf{p_2}\times\mathbf{g_2}}{|\mathbf{p_2}\times\mathbf{g_2}|}. \end{cases} \tag{3}$$

The optical axis of the eye, found at the intersection of these planes, is ascertained through the normalized cross product $\frac{n_1\times n_2}{|n_1\times n_2|}$, which allows for the precise derivation of the axis' coordinates.

The head-mounted near-eye assistance device, illustrated in Fig. 5, integrates compact eye cameras for acquiring high-definition images close to the eye. As depicted in Fig. 2, for each eye, dual sight lines are constructed to detect the corneal reflection and to ascertain the spherical center of the cornea and its 3D optical axis.

Using a spherical cornea model and near-eye cameras, this system captures user-viewpoint images with a scene camera. It determines corneal center, aligns the sight line by detecting the pupil's center, and deduces sight direction from the pupil's 3D coordinates connected to the corneal center, based on geometric eye principles [23].

### C. Grasp Detection

Grasp detection, in contrast to standard object detection, predominantly utilizes smaller rectangles and exhibits heightened sensitivity to spatial positioning and rotational orientations. To facilitate a comprehensive interpretation of this model, this study employs a hierarchy of transformer layers, instead of conventional convolutional kernels with static receptive fields. These layers serve as the primary structure for progressively distilling features from coarse to fine granularity. Fig. 3 illustrates that the initial image $\mathcal{I}$, existing in the space $\mathbb{R}^{W\times H}$, where W and H denote its width and height, undergoes an initial division into unique, non-overlapping segments through a convolutional projection layer. In this context, each image patch is analogous to a word token. Echoing the approach in [26], the model incorporates four sequential stages for the extraction of semantically enriched features. Every phase includes both a layer for merging patches and a swin transformer layer. The process of patch merging mirrors the pooling function in CNNs, aiming to diminish image resolution while concurrently amplifying feature channel depth. Fig. 3 also details the feature dimensions at each stage. At the heart of the swin transformer layer lies the attention mechanisms, which linearly transforms input features to produce query, key, and value elements. This leads to the computation of self-attention as detailed subsequently:

$$\text{Attention}(Q,K,V) = \text{SoftMax}(\frac{QK^T}{\sqrt{d}})V, \tag{4}$$

where $\sqrt{d}$ acts as the scaling factor. Self-attention in the swin transformer layer is confined to a localized window, significantly reducing computational requirements. Additionally, it utilizes a shifted window technique to encapsulate global interconnections. The computational sequence within the Swin Transformer framework is mathematically articulated as follows:

$$\begin{aligned} \hat{\mathbf{u}}^l &= \text{W-MSA}\left(\text{LN}\left(\mathbf{u}^{l-1}\right)\right) + \mathbf{u}^{l-1}, \\ \mathbf{u}^l &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{u}}^l\right)\right) + \hat{\mathbf{u}}^l, \\ \hat{\mathbf{u}}^{l+1} &= \text{SW-MSA}\left(\text{LN}\left(\mathbf{u}^l\right)\right) + \mathbf{u}^l, \\ \mathbf{u}^{l+1} &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{u}}^{l+1}\right)\right) + \hat{\mathbf{u}}^{l+1}. \end{aligned} \tag{5}$$

In this sequence, the feature $\mathbf{u}^{l-1}$, derived from the preceding layer, is initially processed through a layer normalization (LN) procedure and subsequently engaged in the W-MSA

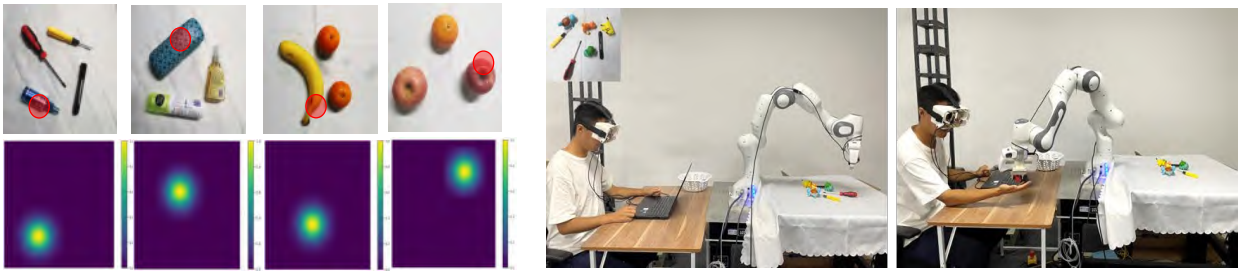



Fig. 4. (a) illustrates the visual representation of scene images and corresponding human attention as captured by our method. The images, in sequence from left to right, highlight human gaze fixation on a blue bottle, eyeglass box, banana, and apple. (b) describes an experimental setup for real-time grasping based on human gaze tracking.

(Window Multi-Head Self Attention). This is followed by the inclusion of a residual connection between each of the modules. The process is then replicated in the SW-MSA (Shifted Window Multi-Head Self Attention) layer. The employment of the Swin Transformer as the backbone network is primarily driven by its capacity to simultaneously accommodate global and local perceptual abilities. This architecture also presents a reduction in computational complexity, especially in comparison to the conventional self-attention mechanism.

In the proposed model, illustrated at the base of Fig. 3, a feature fusion pyramid is strategically implemented to amalgamate features derived from each layer within the backbone network. This pyramid structure facilitates a multi-scale integration of features, thereby enriching the contextual information encompassing both semantic and spatial aspects. The fusion module within this architecture employs a concatenation strategy to merge these diverse features. Through $1\times1$ convolutional kernels, the network yields three unique outputs: grasping quality, width, and angle heads, each maintaining the original input image's dimensional size.

The grasping quality head assigns a success probability from 0 to 1 for each point, reflecting the grasp's likelihood at that image location. For grasping angles, two components, $cos2\theta$ and $sin2\theta$, are used, with the angle calculated as $\frac{1}{2}arctan\frac{cos2\theta}{sin2\theta}$. The network then locates the highest quality grasp point within the quality head's outputs, determining it as the grasp center and identifying the corresponding gripper rotation angle and width.

The model's loss function, $\mathcal{L}$, is defined as $\mathcal{L} = w_1\mathcal{L}pose + w_2\mathcal{L}angle + w_3\mathcal{L}width$, where $\mathcal{L}i$ denotes the mean square error between predicted and ground truth values for each loss component. Weight factors $w_1, w_2$, and $w_3$ adjust each component's impact. For instance, the pose loss $\mathcal{L}pose$ is calculated as $\sum_{i=1}^{N}|\tilde{G}_i - G_i^*|^2$, with $\tilde{G}_i$ being the grasp quality head's output and $G_i^*$ its ground truth.

## IV. Experiment

### A. Dataset and Model Implementation Specifications

For the purpose of assessing our grasp detection model's performance, we utilize the Cornell Grasping Dataset [14]. This dataset comprises images, each uniformly cropped to a dimension of $224\times224$. The complete grasp detection model has been developed using the PyTorch framework. The model training is conducted with a batch size of 32, employing the AdamW optimizer. The learning rate for this process is set at $1e$-4.

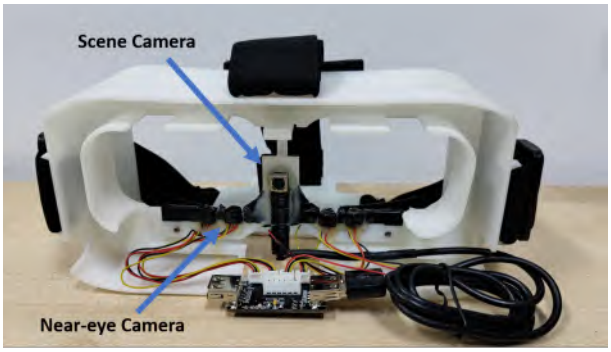



Fig. 5. Development of the Head-Mounted Eye-Tracking Device.

**Evaluation Criteria.** Consistent with the evaluation benchmarks established in [14], [27], [28], we adopt the grasping rectangle metric to quantify the effectiveness of the grasping predictions. Accuracy in a predicted grasp is determined by meeting two key criteria:

i) The rotational angle deviation between the predicted grasp and the ground truth is limited to a maximum of $30°$.

ii) The predicted grasp's Jaccard index, in comparison to the ground truth, is over $0.25$. The Jaccard index is delineated as:

$$J\left(\mathcal{R},\mathcal{R}\right) = \frac{|\mathcal{R}\cap\mathcal{R}|}{|\mathcal{R}\cup\mathcal{R}|}, \tag{6}$$

In this equation, $\mathcal{R}$ represents the region of the predicted grasping rectangle, while $\mathcal{R}$ denotes the ground truth region. The terms $\mathcal{R}\cap\mathcal{R}$ and $\mathcal{R}\cup\mathcal{R}$ refer to the intersection and union of these regions, respectively.

### B. Grasping Performance Analysis

Utilizing inverse kinematics, the robot accurately formulates the grasping trajectory based on defined coordinates. Our methodology is rigorously evaluated against contemporary approaches, as delineated in Table II using the Cornell Grasping Dataset. Despite the relatively narrow performance differentials among leading-edge models,

978-1-6654-7075-9/23/$31.00 ©2023 IEEE

TABLE I

SUCCESS RATES IN EXPERIMENTAL ROBOTIC GRASPING FOR VARIOUS OBJECTS.

| Category | Object | Detected / Total | Grasp % | Object | Grasp % |
|---|---|---|---|---|---|
| Seen Objects | Mouse | 15 / 15 | 13 / 15 (87%) | Remote Control | 13 / 15 (87%) |
| | Apple | 14 / 15 | 12 / 15 (80%) | Pencil | 13 / 15 (87%) |
| Familiar Objects | Orange | 14 / 15 | 12 / 15 (80%) | Knife | 11 / 15 (73%) |
| | Staples Box | 15 / 15 | 12 / 15 (80%) | Screwdriver | 14 / 15 (93%) |
| Unseen Objects | Scissor | 12 / 15 | 11 / 15 (73%) | Toothpaste Box | 12 / 15 (80%) |
| | Razor | 13 / 15 | 11 / 15 (73%) | Toy | 9 / 15 (60%) |

TABLE II

THE ACCURACY ON CORNELL GRASPING DATASET.

| Authors | Approach | Accuracy (%) |
|---|---|---|
| Jiang [29] | Fast Search | 60.5 |
| Asif [30] | GraspNet | 90.2 |
| Redmon [15] | AlexNet, MultiGrasp | 88.0 |
| Guo [31] | ZF-net | 93.2 |
| Asif [32] | STEM-CaRFs | 88.2 |
| Wang [28] | Two-stage closed-loop | 85.3 |
| Kumra [27] | ResNet-50x2 | 89.2 |
| Morrison [16] | GG-CNN | 73.0 |
| Lenz [14] | SAE, struct. reg. | 73.9 |
| Zhou [33] | FCGN, ResNet-101 | 97.7 |
| Karaoguz [34] | GRPN | 88.7 |
| Our | GraspFormer-D | 96.28 |
| | GraspFormer-RGB | 97.72 |
| | GraspFormer-RGB-D | **98.86** |

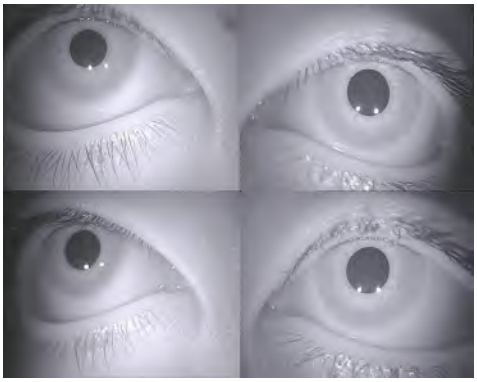


Fig. 6. Eye images captured by cameras positioned near the eye..

our approach demonstrates superior results. Specifically, our transformer-based model for grasp detection exhibits an accuracy of 96.28% when solely employing depth images, and an elevated accuracy of 98.86% with RGB-D input. A notable aspect of our model is its capability to directly ascertain the quality, angle, and width of the grasping rectangles. This feature significantly reduces the need for designing specific anchors for varying targets, streamlining the grasp detection process.

To assess our model's generalizability to new environments, objects were rearranged in varied positions and orientations, and categorized into three groups: those in the dataset, similar to those in the dataset, and entirely new. Each group contained a minimum of four objects. These objects underwent multiple grasping trials, with successful attempts being recorded. The results, detailed in Table I, indicate strong performance for known objects, effective generalization for similar objects, and notable accuracy improvements in grasping unseen objects in complex scenes.

### C. Design of the Integrated System

Our system integrates two core modules: an eye-tracking unit and a grasping mechanism. Experiments are conducted using a Franka Emika Panda robot, equipped with a RealSense D435i RGB-D camera attached to its gripper. The camera's depth images undergo preprocessing as outlined in [35]. The Panda robot features a parallel-finger gripper with a 10 cm operational range and a maximum load capacity of 3 kg. This eye-tracking setup features four infrared-illuminated eye cameras and a scene camera, all mounted on a head frame. The under-eye cameras capture detailed images, using near-infrared light for precise 3D eye modeling based on corneal reflections.

### D. System Limitations

The system faces challenges due to complexities in eye-tracking and grasping. For eye-tracking, accurately determining pupil center coordinates is difficult due to factors like corneal refraction and asphericity, which introduce hard-to-calibrate variables. To address this, our module employs a simplified eye model, disregarding corneal refraction and assuming a spherical corneal surface, although in reality, the curvature varies across the eye's surface. These approximations may result in gaze estimation inaccuracies. In the grasping subsystem, limitations arise while handling transparent objects, as the RealSense camera's depth perception for such materials is inadequate. Experiments indicate that objects with complex or smooth surfaces tend to slip from the grippers, highlighting a need for improvement in handling diverse object textures and shapes.

## V. CONCLUSIONS AND FUTURE DIRECTIONS

This study introduces a novel contactless human-robot interface, facilitating robotic manipulation through visual cues. Our system incorporates a head-mounted eye-tracker to pinpoint objects under human observation. Utilizing the user's gaze data, the transformer-based grasp model effectively discerns the user's focus area, capitalizing on its global perception capabilities. Empirical results indicate that the developed gaze-directed robotic arm adeptly performs tasks such as object relocation and precise grasping, guided by near-eye tracking.

978-1-6654-7075-9/23/$31.00 ©2023 IEEE

Further examination through ablation studies confirms the satisfactory tracking accuracy of our eye-tracking module. This research lays the groundwork for more intuitive and efficient human-robot collaboration, presenting potential for further enhancements in accuracy and versatility in varying operational contexts.

## REFERENCES

[1] J. B. Sol, "Effective grasping enables successful robot-assisted dressing," *Sci. Robotics*, vol. 7, no. 65, 2022.

[2] Z. Li, G. Li, X. Wu, Z. Kan, H. Su, and Y. Liu, "Asymmetric cooperation control of dual-arm exoskeletons using human collaborative manipulation models," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 12 126–12 139, 2022.

[3] Y. Xia, S. Wang, and Z. Kan, "A nested u-structure for instrument segmentation in robotic surgery," in *International Conference on Advanced Robotics and Mechatronics*, 2023, pp. 994–999.

[4] S. Wang, Z. Zhou, B. Li, Z. Li, and Z. Kan, "Multi-modal interaction with transformers: bridging robots and human with natural language," *Robotica*, pp. 1–20, 2023.

[5] R. Rahman, M. S. Rahman, and J. R. Bhuiyan, "Joystick controlled industrial robotic system with robotic arm," in *IEEE International Conference on Robotics, Automation, Artificial-intelligence and Internet-of-Things (RAAICON)*, 2019, pp. 31–34.

[6] S. Wang, Z. Zhou, H. Wang, Z. Li, and Z. Kan, "Unsupervised representation learning for visual robotics grasping," in *International Conference on Advanced Robotics and Mechatronics*, 2022, pp. 57–62.

[7] F. Sun, W. Zhang, J. Chen, H. Wu, C. Tan, and W. Su, "Fused fuzzy petri nets: A shared control method for brain-computer interface systems," *IEEE Trans. Cogn. Dev. Syst.*, vol. 11, no. 2, pp. 188–199, 2019.

[8] Y.-S. L.-K. Cio, M. Raison, C. L. Ménard, and S. Achiche, "Proof of concept of an assistive robotic arm control using artificial stereovision and eye-tracking," *IEEE Trans. Neural Syst. and Rehabilitation Engineering*, vol. 27, no. 12, pp. 2344–2352, 2019.

[9] Z. Zhou, S. Wang, Z. Chen, M. Cai, H. Wang, Z. Li, and Z. Kan, "Local observation based reactive temporal logic planning of human-robot systems," *IEEE Transactions on Automation Science and Engineering*, pp. 1–13, 2023.

[10] S. Wang, Z. Zhou, and Z. Kan, "When transformer meets robotic grasping: Exploits context for efficient grasp detection," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8170–8177, 2022.

[11] Y. Pu, S. Wang, X. Yao, and B. Li, "Context-based soft actor critic for environments with non-stationary dynamics," *arXiv preprint arXiv:2105.03310*, 2021.

[12] S. Wang, R. Yang, B. Li, and Z. Kan, "Structural parameter space exploration for reinforcement learning via a matrix variate distribution," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 4, pp. 1025–1035, 2023.

[13] S. Wang and B. Li, "Implicit posterior sampling reinforcement learning for continuous control," in *Neural Information Processing*, Cham, 2020, pp. 452–460.

[14] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robotics Res.*, vol. 34, no. 4-5, pp. 705–724, 2015.

[15] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 1316–1322.

[16] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *Int. J. Robotics Res.*, vol. 39, no. 2-3, pp. 183–201, 2020.

[17] N. Hollenstein and C. Zhang, "Entity recognition at first sight: Improving ner with eye movement information," in *NAACL-HLT (1)*, 2019, pp. 1–10.

[18] N. Karessli, Z. Akata, B. Schiele, and A. Bulling, "Gaze embeddings for zero-shot image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4525–4534.

[19] J.-Y. Lee, H.-M. Park, S.-H. Lee, T.-E. Kim, and J.-S. Choi, "Design and implementation of an augmented reality system using gaze interaction," in *Int. Conf. on Information Science and Applications*, 2011, pp. 1–8.

[20] Y. Sugano and A. Bulling, "Seeing with humans: Gaze-assisted neural image captioning," *arXiv preprint arXiv:1608.05203*, 2016.

[21] Y. Pu, S. Wang, X. Yao, and B. Li, "Latent context based soft actor-critic," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[22] F. Bruno, L. Barbieri, and M. Muzzupappa, "A mixed reality system for the ergonomic assessment of industrial workstations," *Int. Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 14, no. 3, pp. 805–812, 2020.

[23] T. Nagamatsu, Y. Iwamoto, J. Kamahara, N. Tanaka, and M. Yamamoto, "Gaze estimation method based on an aspherical model of the cornea: surface of revolution about the optical axis of the eye," in *Proc. of Symposium on Eye-Tracking Research Applications*, 2010, pp. 255–258.

[24] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, 2000.

[25] G. Bradski and A. Kaehler, "Opencv," *Dr. Dobbs journal of software tools*, vol. 3, p. 120, 2000.

[26] L. Z. et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vision*, 2021, pp. 10 012–10 022.

[27] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst*, 2017, pp. 769–776.

[28] Z. Wang, Z. Li, B. Wang, and H. Liu, "Robot grasp detection using multimodal deep convolutional neural networks," *Advances in Mechanical Engineering*, vol. 8, no. 9, 2016.

[29] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2011, pp. 3304–3311.

[30] U. Asif, J. Tang, and S. Harrer, "Graspnet: An efficient convolutional neural network for real-time grasp detection for low-powered devices." in *IJCAI*, vol. 7, 2018, pp. 4875–4882.

[31] D. Guo, F. Sun, H. Liu, T. Kong, B. Fang, and N. Xi, "A hybrid deep architecture for robotic grasp detection," in *Proc. IEEE Int. Conf. Robot.Automat.*, 2017, pp. 1609–1614.

[32] U. Asif, M. Bennamoun, and F. A. Sohel, "Rgb-d object recognition and grasp detection using hierarchical cascaded forests," *IEEE Trans. on Robotics*, vol. 33, no. 3, pp. 547–564, 2017.

[33] X. Zhou, X. Lan, H. Zhang, Z. Tian, Y. Zhang, and N. Zheng, "Fully convolutional grasp detection network with oriented anchor box," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2018, pp. 7223–7230.

[34] H. Karaoguz and P. Jensfelt, "Object detection approach for robot grasp detection," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 4953–4959.

[35] K. Chen, S. Wang, B. Xia, D. Li, Z. Kan, and B. Li, "Todetrans: Transparent object depth estimation with transformer," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 4880–4886.