Hand-Object Interaction Detection based on Visual Attention for Independent Rehabilitation Support

Adnan Rachmat Anom Besari[1,2]

[2] *Department of Information and Computer Engineering, Politeknik Elektronika Negeri Surabaya,*
Surabaya, Indonesia.
adnan-rachmat-anom-besari@ed.tmu.ac.jp, anom@pens.ac.id

Azhar Aulia Saputra[1], Wei Hong Chin[1], Naoyuki Kubota[1]

[1] *Department of Mechanical Systems Engineering, Faculty of Systems Design, Tokyo Metropolitan University,*
Tokyo, Japan.
aa.saputra@tmu.ac.jp, weihong@tmu.ac.jp, kubota@tmu.ac.jp

Kurnianingsih[3]

[3] *Department of Electrical Engineering, Politeknik Negeri Semarang,*
Semarang, Indonesia.
kurnianingsih@polines.ac.id

***Abstract*— Hand rehabilitation in post-stroke patients with visual impairment is currently not supported by the availability of a cyber-physical-social system (CPSS) that can monitor physical development during daily activities. This paper discusses how to extract hand activity information on objects based on visual attention in the task-specific reach-to-grasp cycle. We used perception-based egocentric vision to observe hand-object interactions (HOI) in grasping tasks. Our approach combines object detection with hand skeletal model estimation and visual attention to validate HOI detection. We choose a multilayer Gated Recurrent Unit (GRU) based on Recurrent Neural Networks (RNN) architecture to classify the four main activities when the hand interacts with an object (wonder-reach-grasp-release). We evaluated the algorithm quantitatively on the new dataset we introduced for cup grasping activity. This method can validate the HOI detection with 97.0% precision with less training time for small data. Further research will use these results to increase self-efficacy for independent hand-eye coordination rehabilitation support in community-centric systems. The code and dataset are available at https://github.com/anom-tmu/hoi-attention/.**

Keywords—post-stroke rehabilitation, hand-eye coordination, egocentric vision, hand gesture, reach-to-grasp cycle

I. INTRODUCTION

In recovering from neurological diseases, patients need rehabilitation to improve their condition. One type of rehabilitation that therapist usually uses to enhance physical function is occupational therapy. This therapy aims to improve the patient's self-care ability and make the patient able to carry out daily activities independently [1]. However, about two-thirds of stroke survivors have a visual impairment related to visual field loss, double vision, and perceptual problems. The impact of visual problems included a loss of confidence, being a burden to others, increased collisions, and fear of falling. They find it difficult to reach and grasp objects due to limited hand movement and visual function [2]. To facilitate the rehabilitation process for these patients, it would be pleasant and comfortable if the patient could carry out the therapy process at home instead of going to the hospital. However, it is not easy for therapists to visit patients' homes during the COVID-19 pandemic. This problem will make patients reluctant to do recovery therapy at home independently.

On the other hand, medical personnel must monitor the patient's rehabilitation results. Tracking patient rehabilitation progress can be done online via telemedicine technology [3]. However, patients cannot continuously carry this system due to limited therapy personnel or privacy concerns. Many studies have discussed the application of cyber-physical-social systems (CPSS) to monitor the hand development of post-stroke patients in performing daily activities [4]. Most researchers use the contact method, where patients must use electronic gloves or sensors attached to their hands and fingers to carry out daily activities. However, this is felt uncomfortable for most patients. And then, other studies turn to non-contact methods to address this issue. They usually use camera sensors facing the patient or attached to the body. Using cameras with egocentric vision, such as smart glasses [5], is the primary choice to reduce privacy issues.

Research in the egocentric vision for monitoring hand activity with particular objects is commonly known as hand-object interaction (HOI) detection. HOI detection is currently used in tracking patients' rehabilitation progress in post-stroke recovery [6]. However, the application of this research field is limited to the detection of hands and objects without considering visual attention. HOI application in post-stroke patients with visual impairment requires eye focus and hand gestures towards objects. The location of the hands, objects, and the focal point on the image pixels is needed for the health monitoring system. For example, a virtual HOI detection with a reach-to-grasp cycle [7] defines physical interactions in the digital world. It is important to include visual attention to enhance this approach in a real-world application.

The main contributions of this paper are as follows. First, we propose a framework for an independent hand rehabilitation monitoring system using feature extraction based on the egocentric vision in our previous work [8]. We developed this framework to get the feature from hands and objects. Then, we define the interaction by using the distance of each fingertip to the center of the object. Second, we developed the active perception ability of HOI detection based on features in hand gripping gestures. We used changing the distance of the tumb fingertip with the other four fingertips. Third, we evaluate cognitive ability using a multilayer GRU-based RNN [9] with input features obtained in the previous stage. This ability is needed to get symbolic information on four main activities: wonder, reach, grasp, and release. This hand activity cycle is necessary to validate the results of HOI detection for use in the knowledge domain in our previous works [10].

This paper is structured as follows. Section 2 discusses the related works on using technology for monitoring independent hand rehabilitation. Section 3 proposes our method for developing HOI detection based on visual attention. Section 4 shows the results and discusses the effectiveness of the proposed method. Finally, section 5 presents the conclusions and future works of the research.

978-1-7281-8671-9/22/$31.00 ©2022 IEEE

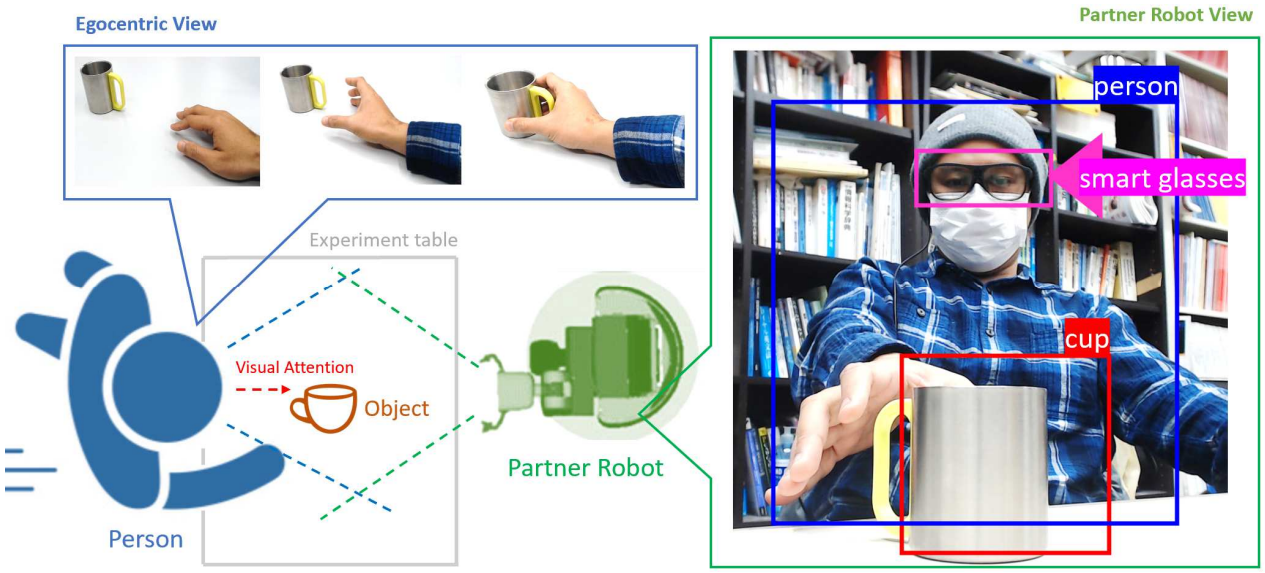


Fig. 1. Triadic interaction between Person-Partner Robot-Object

II. Related Works

Many researchers have developed independent hand rehabilitation monitoring studies. There are two main categories in a survey on hand monitoring for rehabilitation: contact and non-contact. The contact method is a wearable technology that researchers widely choose because it gives accurate data using several sensor technologies such as flex sensor, accelerometer, hall-effect, stretch-sensor, and magnetic sensor [11]. Developers have widely provided wearable technology to support this research area. However, this method has drawbacks, including the high cost of equipment and being less comfortable if used for too long. Therefore, researchers are looking for other alternatives by using the non-contact method even though the data provided has less accuracy than the contact method.

Our previous studies have discussed nonverbal communication based on instructed learning for socially embedded robot partners [12]. This study aims to determine a person's intentions and abilities when reaching and grasping objects. However, it is challenging to detect a person's intentions by using the robot partner's point of view as a third person [13]. The third-person perspective depends on many possibilities of point of view and is constrained by the occlusion. It is necessary to support the existing system using a first-person or egocentric point of view. This approach is inspired by Tomasello's concept of joint attentional interaction between two people with an object [14]. Figure 1 shows the idea of triadic interaction between the Person-Partner Robot-Object and the robot vision of the person and the object.

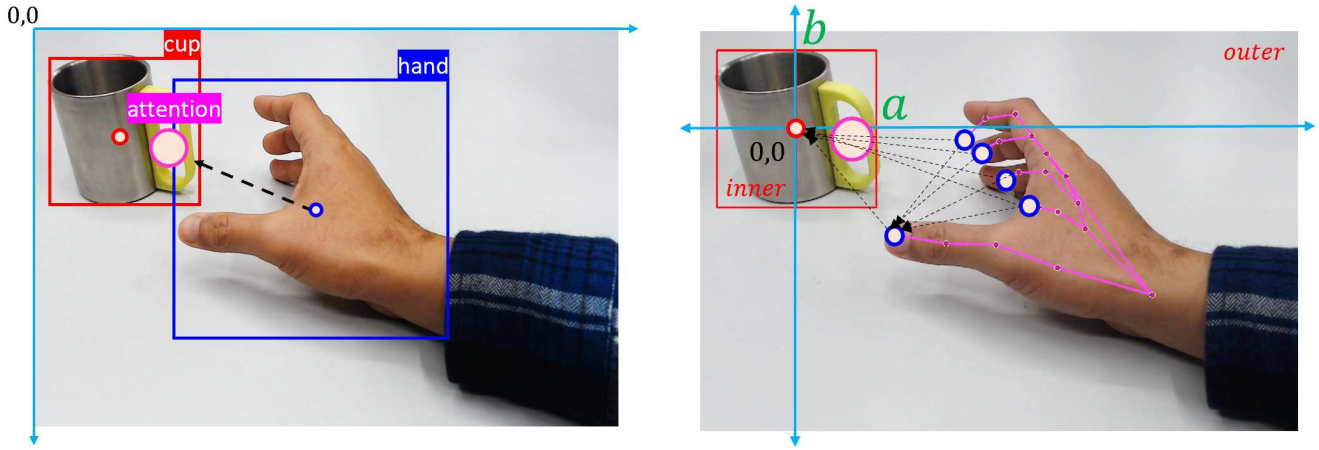
The use of egocentric vision-based sensors to recognize hand movements is a research trend that has developed recently [6]. Hand gesture recognition can be done by these vision sensors mainly by using 3D model estimation based on appearance-based techniques. The recognition is developed based on the kinematic model of the fingers. This volumetric model can obtain the required palm position and joint angle parameters. The main idea is to get the parameters of the hand by comparing the possible 2D view as projected by the 3D hand model and the input image from the camera.

Many vendors have developed various types of cameras to support the research on hand gesture recognition. Advanced camera devices such as stereo infra-red cameras on the Leap Motion Controller (LMC) and the other RGB-D camera have also begun to be developed by many AI companies to improve hand measurement accuracy [15]. However, these devices are pretty expensive and constrained in a real-world implementation. So many researchers are still struggling with RGB cameras, which are now commonly installed on smartphones, smart glasses, and spy camera necklaces. They consider wearable cameras to have future potential to record all human activities, including maintaining health.

Using an RGB camera is easy to use as input to estimate hand poses which consists of the position of the finger joints and the angle value. Some researchers use marker sets for hand motion capture, color markers, or skin-colored sections to get precise predictions of hand poses. However, the current deep learning algorithm supports real-time hand tracking without markers. Popular pre-trained models such as MediaPipe [16] already support low computing and are free. Thus, it becomes a friendly choice for health developers to use for artificial intelligence applications by using widely available devices.

The combination of hand poses and their interactions with objects is necessary [17], especially for assessing rehabilitation. Research on HOI detection has an excellent opportunity to detect objects compared with complete human interaction. Problems in detecting full-human interactions with objects often involve complex processing, multi-interpretation, and privacy concerns. Meanwhile, the issue of HOI detection is more straightforward because only hands and objects are detected, especially when using egocentric vision [18]. Egocentric or first-person vision is a computer vision approach that analyzes the image from a wearable camera. The advantages of egocentric vision include reduced privacy concerns, mobile supervisory, and finding attention during activities.

However, using egocentric vision, the development of HOI detection also faces some problems, especially in 2D images. The first problem that researchers often encounter is



(a) (b)

Fig. 2 Proposed method for HOI validation: (a) Conventional HOI point; (b) Object-centered coordinate with visual attention in HOI detection

the lack of validity of the detection results of hand and object interactions. This issue arises because the data in 2D images cannot describe depth as in 3D data. That way, the system cannot ascertain the position of the hand and object. Several methods can overcome the above problems, such as learning at the interaction point [19]. However, this problem is only limited to one type of object, and errors can occur in detecting many objects.

It is necessary to prove the existence of hand action when interacting with objects, especially during grasping. Grasping is an essential part of limb movement activities. In the International Classification of Functioning (ICF) on Disability and Health [20], WHO included grasping in the "Carrying, Moving and Handling Object" category, especially in the Fine Hand Use section (d440). In the "grasp" subsection, several studies have reduced it to four sequential activities: the reach-grasp-transport-release known as the reach-to-grasp cycle. However, it is infrequent for research to solve the problem of HOI detection using a medical approach like the one in the ICF. Therefore, this study aims to develop and use HOI detection techniques for hand rehabilitation in the case of the reach-to-grasp cycle.

III. Proposed Method

This section will discuss the stages in the development of task-specific reach-to-grasp cycle-based HOI detection to support independent rehabilitation. These stages consist of the experimental setup, object-centered coordinate transformation, and validation of HOI detection based on the task-specific reach-to-grasp cycle.

A. Experimental Setup

We use a camera in smart glasses facing the experimental table with egocentric vision. The smart glasses used is a Tobii Pro Glasses 3 used by a participant facing down straight at the object. The smart glasses camera has a 1920 x 1080 pixels resolution with 25 fps. We use this camera to get pictures of hands and objects above the table. We installed this camera on a computer with Intel Core i7-10875H CPU @ 2.30GHz specifications (16 CPUs), 16GB RAM, and NVIDIA GEFORCE RTX 2080 (8GB GDDR6 VRAM). We chose a small cup with a handle as the hand will interact with this object. We selected the cup because people in daily activities often use it, and there are many ways to hold it. We will get many data variations on the reach-to-grasp cycle activity. Figure 2(a) shows an egocentric view with HOI detection using conventional hand-object interaction points. This image uses standard pixel coordinates, where the initial value *(0,0)* is at the top left corner of the plane.

After doing hardware preparation and setup, we do software setup. To get information on the bounding box of hands and objects, we processed the results of capturing image frames in egocentric vision using YOLO (You Only Look Once) [21]. And then, we have improved this detection with the Simple Online Realtime Tracking (SORT) algorithm [22]. This framework has outstanding capabilities for learning representation and applying it in object detection and tracking applications. There are two things we can get from this object detection application. First, we can search to identify the object in a particular image, and second, we can determine the exact location of the object in the two-dimensional image. We use MediaPipe [16] hand tracking from Google Research to get the estimated hand pose data. A framework designed explicitly for complex perceptual channels utilizing accelerated inference in real-time. We only use hand pose prediction as supporting data to validate HOI detection from this framework.

B. Object-Centered Coordinate Transformation

After we get information from object detection and estimation of the position of the finger joint feature, the next step is to perform a coordinate transformation. We have achieved a coordinate transformation centered on the object to simplify the validation process and get less input [23]. The purpose of centering the object is to move the initials of the image coordinates *(0,0)* to the center of the object's bounding box (x_{center}, y_{center}). Figure 2(b) shows the object-centered coordinates with visual attention in HOI detection. To get a new center (0,0) for each new frame, we need to find the x_{center} and y_{center} values with the following equation:

$$x_{center} = x_2 - \frac{(x_2 - x_1)}{2} \quad (1)$$

$$y_{center} = y_2 - \frac{(y_2 - y_1)}{2} \quad (2)$$

Then we can determine the position of the new coordinates (a_n, b_n) with the following equation:

$$a_n = x_n - x_{center} \quad (3)$$

$$b_n = y_n - y_{center} \quad (4)$$

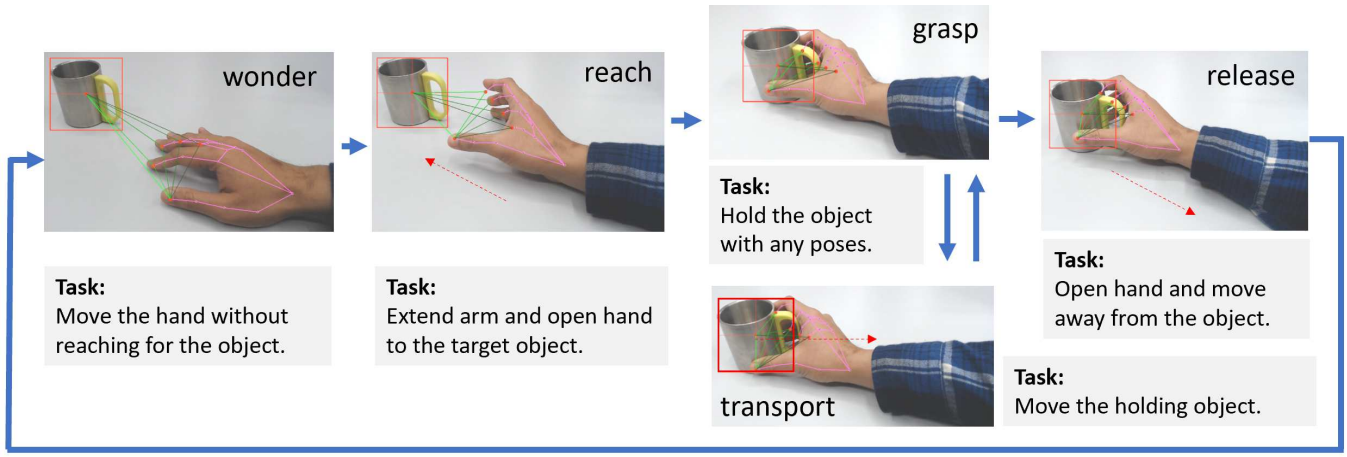


Fig. 3. Design of the system and the architecture for HOI validation.

The algorithm uses the above equation for any pixel point (x_n, y_n) in the image plane. For example, the location of the bounding box, which consists of the length (a_0) and width (b_0) of the object, or the position of the fingertip or finger joint in the new coordinate plane. With the further bounding box location information, we will get the inner and outer boundaries of the object. Then we can determine the distance between each point (a_n, b_n) and the center of the object coordinates *(0,0)* with the following Pythagorean equation:

$$d_n = \sqrt{a_n^2 + b_n^2} \tag{5}$$

The distance d_n will be required to determine the distance between the fingertip or finger joint to the object-centered coordinate. Next, we will use a_n, b_n, and d_n to validate the hand-object interaction.

C. *Validate the HOI detection*

As discussed in the previous section, validation of HOI detection is limited to the reference grasp of the acceptable hand use section in ICF for hand rehabilitation, particularly in the case of the reach-to-grasp cycle [7]. There are four specific tasks in the process, and we will define them one by one. The first is the "wonder" task as the initial status, which indicates that the person moves the hand without reaching for the object. The second task is the "reach" task which is person extends the arm and opens his hand to the object. The third is the "grasp" task which is person holds the object with any pose. This task has transport as an additional state when the person moves the holding object. The fourth is the "released" task, when the person's open hand moves away from the object. Figure 3 shows the phases of the task-specific reach-to-grasp cycle.

As mentioned earlier, we use a single object as a reference. The object we chose is a medium-sized cup with several possible hand poses. We operate ten features consisting of five elements of the distance of each fingertip to the center of the object $(d_o, d_1, d_2, d_3, d_4)$ four elements of the distance of each fingertip to the thumb fingertip (e_1, e_2, e_3, e_4) and one visual attention (f_0). With our computer specifications, we acquire 50 fps which becomes our standard for determining the amount of a data sequence. To get the real-time result, we process 10 data in every single series of the image capture. We use this data as input for the learning system in our neural networks.

We used recurrent neural networks (RNN) architecture based on a multilayer GRU for multivariate time-series classification [9]. The multilayer GRU architecture that we chose has several parameters such as *input_size,* which is the number of features in the input *x*; *hidden_size,* which is the number of features in the hidden state *h*; and *number of_layers* which is the number of recurrent layers. For each element in the input sequence of multilayer GRU, each layer computes the following function:

$$r_t = \sigma(W_{ii}x_t + b_{ir} + W_{hr}h_{(t-1)} + b_{hr}) \tag{6}$$

$$z_t = \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{(t-1)} + b_{hz}) \tag{7}$$

$$n_t = tanh(W_{in}x_t + b_{in} + r_t * (W_{hn}h_{(t-1)} + b_{hn})) \tag{8}$$

$$h_t = (1 - z_t) * n_t + z_t * h_{(t-1)}) \tag{9}$$

Where h_t is the hidden state at time t, x_t is the input at time t, $h_{(t-1)}$ is the hidden state of the layer at time $t-1$ or the initial hidden state at time o, and r_t, z_t, n_t are the reset, update, and new gates, respectively. σ is the sigmoid function and $*$ is the Hadamard product. In our multilayer GRU, the input $x_t^{(l)}$ of the l -th layer $(l >= 2)$ is the hidden state $h_t^{(l-1)}$ of the previous layer multiplied by dropout $\delta_t^{(l-1)}$ where each $\delta_t^{(l-1)}$ is a Bernoulli random variable which is 0 with a probability of dropout. Figure 4 shows the design of the system and the architecture for HOI validation.

After developing the multilayer GRU-based RNN architecture, the next step is to prepare a dataset to validate each action on HOI detection. We capture each data by recording 1-2 seconds of video with a minimum of 50 fps per sequence. We collected 100 videos of hands interacting with objects with various possibilities. We made the video capture with the following division: 25 data for wonder-task, 25 data for reach-task, 25 data for grasp-task including transport, and 25 data for release-task. As training and testing data, we divide it in a ratio of 80:20 randomly. We considered this division sufficient because the data we have collected is subjective. In this experiment, we involved a single respondent. The system uses 80 videos for training and 20 videos for testing. We will discuss the training and testing results in more detail in the results and discussion section.

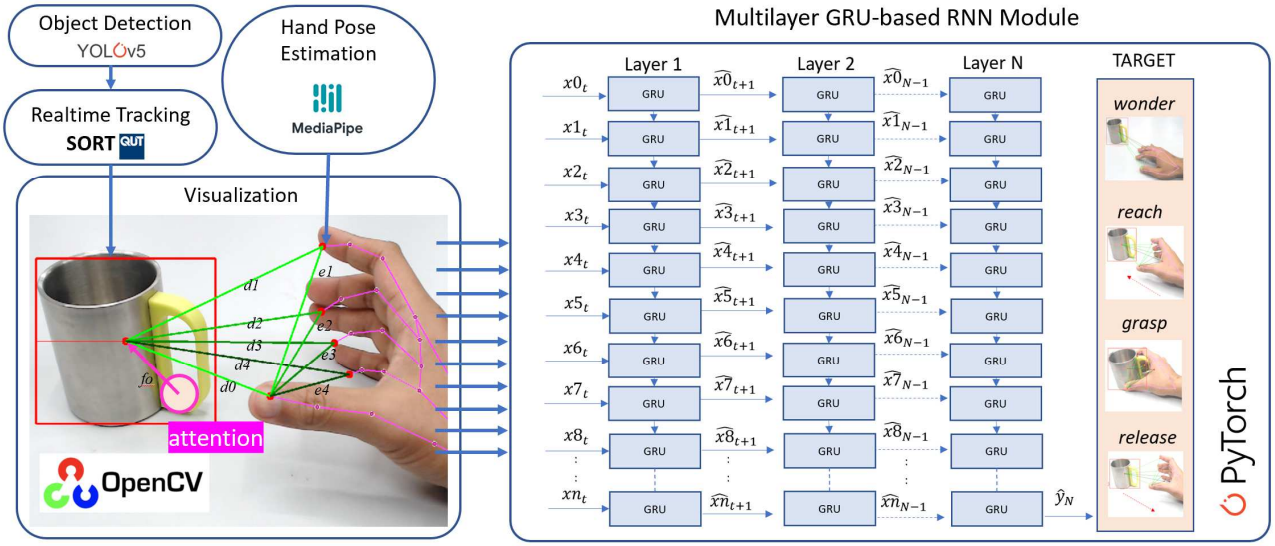


Fig. 4. Design of the system and the architecture multilayer GRU-based on RNN module for HOI validation.

IV. Result and Discussion

We have evaluated the proposed frameworks through a series of experiments. We conducted the experiments on a single person to do the task-specific reach-to-grasp cycle. First, we discuss the feature extraction abilities of egocentric vision that developed from our previous work [8]. We have used object-centered transformation to simplify the extraction of these features. However, there were some technical issues in feature extraction that arose in the development of our system. The first problem we have encountered is that the estimation of hand poses with MediaPipe predicts only one frame.

In contrast to object detection with YOLO integrated with SORT tracking algorithms. The occlusion in some gripping poses becomes less accurate because it does not consider the previous data. This problem can be solved using hand tracking methods and finger pose estimation filters like in the Leap Motion Controller [24]. Another problem is that the data obtained is in 2D pixel units while the egocentric point of view is in the perspective of 3D space. With all the limitations, the 2D camera is sufficient to produce uniform features as long as the range taken is as long as the hand's reach, so there is no need for precise data, for example, in millimeters. Thus, the future development of this low-cost technology can be carried out by research massively.

Second, we evaluate the test results in active perception ability for the task-specific reach-to-grasp cycle. We have conducted some experiments with five-time training using three variants of RNN (vanilla-RNN, GRU, and LSTM). The best training results are predicted in the 50th epoch, where the GRU wins with 13.03 seconds in average training time compared to RNN (20.44 seconds) and LSTM (20.48 seconds). All of the RNN-based learning systems can be classified quite well. The best recognition results can be seen from the system's accuracy, 97.0% for GRU, 96% for RNN, and 94% for LSTM. Table I shows the comparison of RNN-based models in the 50th epoch. In this experiment, the GRU slightly outperformed the traditional RNN. If we compare the result with LSTM, GRU uses fewer tensor operations. It takes less time to train. However, the results of these three RNN-based models are almost the same. Figure 4 shows the training result at the 50th epoch of three RNN types for HOI validation.

TABLE I
Comparison of RNN-based model in 50th epoch.

No.	The architecture of RNN	Learning Time (sec)	Accuracy (%)
1.	RNN	20.44	96.0
2.	LSTM	20.48	94.0
3.	GRU	13.03	97.0

Third, we examine the cognitive ability using a multilayer GRU-based RNN to solve a multivariate time-series classification problem. Based on benchmarking multivariate time series classification study [25], this GRU addresses the vanishing and exploding gradient problem of conventional RNN. GRU is rated better than vanilla-RNN and LSTM. Several experiments show that the model on the multilayer GRU integrates quickly and provides advanced time-series recognition performance for a relatively small model. This learning algorithm improves accuracy compared to conventional methods such as standard RNN and the LSTM. With this result, we get sufficient accuracy even though we use few features for training.

V. Conclusion and Future Work

This paper has discussed extracting hand activity information on objects based on visual attention for the task-specific reach-to-grasp cycle. We conducted this study as a proposed framework for independent hand rehabilitation in post-stroke patients with visual impairment who have experienced grasping difficulties. We have developed an egocentric vision to observe hand-object interactions in real-world tasks. We have successfully created object detection based on active perception and hand skeletal model estimation. Then, we successfully applied RNN with multilayer GRU-based architecture to classify four main activities: wonder-reach-grasp-release. We evaluated this algorithm quantitatively on a new data set for cup grasping activity. The results of our experiments have shown that our proposed method can validate HOI detection with a good level of precision. We will develop the proposed system to be faster at non-standard grasping poses in future work. We hope this research can adapt to various objects for eye-hand coordination rehabilitation needs significantly to increase the patients' self-efficacy.

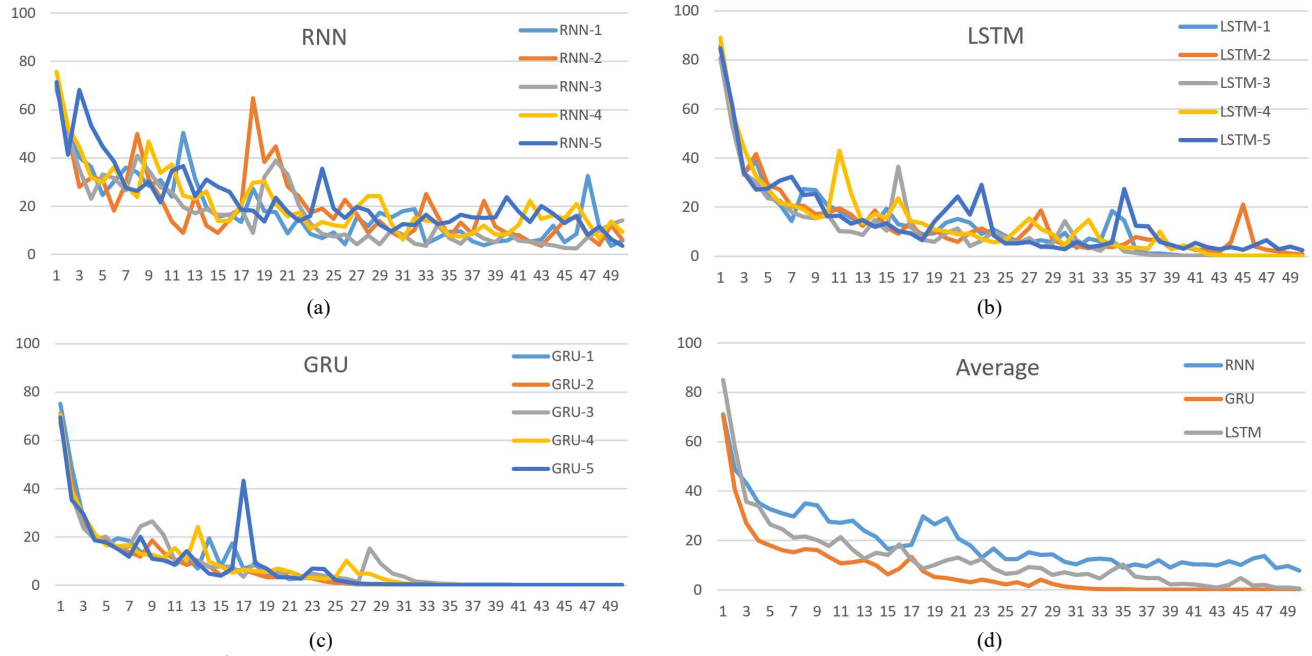


Fig.5. Training result at 50th epoch of three RNN type: (a) Vanilla-RNN; (b) LSTM; (c) GRU and (d) Comparison of the three types.

ACKNOWLEDGMENT

The authors would like to thank the Japan Ministry of Education, Culture, Sports, Science, and Technology (MEXT) for providing financial support. This work was partially supported by Japan Science and Technology Agency (JST), Moonshot R&D, with grant number JPMJMS2034.

REFERENCES

[1] K. A. Almhdawi, V. G. Mathiowetz, M. White, and R. C. delMas, "Efficacy of Occupational Therapy Task-oriented Approach in Upper Extremity Post-stroke Rehabilitation: Task-oriented Therapy Post-stroke," *Occup. Ther. Int.*, vol. 23, no. 4, pp. 444–456, Dec. 2016.

[2] F. J. Rowe, "Stroke survivors' views and experiences on impact of visual impairment," *Brain Behav*, vol. 7, no. 9, p. e00778, Sep. 2017.

[3] M. Szekeres and K. Valdes, "Virtual health care & telehealth: Current therapy practice patterns," *Journal of Hand Therapy*, p. S0894113020302131, Nov. 2020.

[4] A. Laghari, Z. A. Memon, S. Ullah, and I. Hussain, "Cyber Physical System for Stroke Detection," *IEEE Access*, vol. 6, pp. 37444–37453, 2018.

[5] M. Dousty and J. Zariffa, "Tenodesis Grasp Detection in Egocentric Video," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1463–1470, May 2021.

[6] M.-F. Tsai, R. H. Wang, and J. Zariffa, "Identifying hand use and hand roles after stroke using egocentric video," *IEEE J. Transl. Eng. Health Med.*, pp. 1–1, 2021.

[7] Q. Cai, J. Li, and J. Long, "Effect of Physical and Virtual Feedback on Reach-to-Grasp Movements in Virtual Environments," *IEEE Trans. Cogn. Dev. Syst.*, pp. 1–1, 2021.

[8] A. R. Anom Besari, A. A. Saputra, W. H. Chin, N. Kubota, and Kurnianingsih, "Feature-based Egocentric Grasp Pose Classification for Expanding Human-Object Interactions," in *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, Kyoto, Japan, Jun. 2021, pp. 1–6.

[9] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling." arXiv, Dec. 11, 2014. Accessed: May 22, 2022.

[10] A. R. Anom Besari, W. H. Chin, N. Kubota, and Kurnianingsih, "Ecological Approach for Object Relationship Extraction in Elderly Care Robot," in *2020 21st International Conference on Research and Education in Mechatronics (REM)*, Cracow, Poland, Dec.2020, pp.1–6.

[11] A. Rashid and O. Hasan, "Wearable technologies for hand joints monitoring for rehabilitation: A survey," *Microelectronics Journal*, vol. 88, pp. 173–183, Jun. 2019.

[12] R. Tanaka, J. Woo, N. Kubota, and Intelligent Mechanical Systems, Graduate School of System Design, Tokyo Metropolitan University 6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan, "Nonverbal Communication Based on Instructed Learning for Socially Embedded Robot Partners," *JACIII*, vol. 23, no. 3, pp. 584–591, May 2019.

[13] M. Yani, A. R. A. Besari, N. Yamada, and N. Kubota, "Ecological-Inspired System Design for Safety Manipulation Strategy in Home-care Robot," in *2020 International Symposium on Community-centric Systems (CcS)*, Hachioji, Tokyo, Japan, Sep. 2020, pp. 1–6.

[14] M. Tomasello, *The cultural origins of human cognition*, 4. print. Cambridge, Mass.: Harvard Univ. Press, 2003.

[15] A. Ganguly, G. Rashidi, and K. Mombaur, "Comparison of the Performance of the Leap Motion ControllerTM with a Standard Marker-Based Motion Capture System," *Sensors*, vol. 21, no. 5, p. 1750, Mar. 20210.

[16] F. Zhang *et al.*, "MediaPipe Hands: On-device Real-time Hand Tracking," *arXiv:2006.10214 [cs]*, Jun. 2020, Accessed: Jan. 27, 2022.

[17] M. Cai, K. Kitani, and Y. Sato, "Understanding hand-object manipulation by modeling the contextual relationship between actions, grasp types and object attributes," *arXiv:1807.08254 [cs]*, Jul. 2018, Accessed: Jan. 31, 2022.

[18] R. Kaur and D. V. Sharma, "A Review of Vision-Based Techniques Applied to Detecting Human-Object Interactions in Still Images," *Journal of Computing Science and Engineering*, vol.15, no.1, p.16, 2021.

[19] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning Human-Object Interaction Detection using Interaction Points," *arXiv:2003.14023 [cs]*, Mar. 2020, Accessed: Jan. 27, 2022.

[20] Weltgesundheitsorganisation, Ed., *International classification of functioning, disability and health: ICF*. Geneva: World Health Organization, 2001.

[21] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv:2004.10934 [cs, eess]*, Apr. 2020, Accessed: Oct. 10, 2021.

[22] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple Online and Realtime Tracking," *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468, Sep. 2016.

[23] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2637–2646.

[24] A. Vysocký *et al.*, "Analysis of Precision and Stability of Hand Tracking with Leap Motion Sensor," *Sensors*, vol. 20, no. 15, p. 4088, Jul. 2020.

[25] A. P. Ruiz, M. Flynn, and A. Bagnall, "Benchmarking Multivariate Time Series Classification Algorithms," *Data Min Knowl Disc*, vol. 35, no. 2, pp. 401–449, Mar. 2021.