Integrated Sensor-Based Interface for Human-Robot Collaboration in Construction

X. Wang[a], D. Veeramani[b] and Z. Zhu[a]

[a]Department of Civil and Environmental Engineering, University of Wisconsin-Madison, 1415 Engineering Drive, Madison, WI 53706, USA
[b]Department of Industrial and Systems Engineering, University of Wisconsin-Madison, 1513 University Avenue, Madison, WI 53706, USA
E-mail: xwang2463@wisc.edu, raj.veeramani@wisc.edu, and zzhu286@wisc.edu

Abstract –

Construction robots have the potential to increase construction productivity at job sites and can help overcome industry challenges such as labor shortage and safety risks. User-friendly interfaces are critical for advancing human-robot work collaboration and increasing use of construction robots. However, human-robot interfaces in the context of construction industry applications have been investigated to a limited extent only. This paper proposes a novel sensor-based framework which integrates eye tracking and hand gesture recognition for human-robot interaction in construction. Specifically, it begins with visual detection of construction machines in the first-person views. Then, the machine-of-interest is determined based on the detection results and human gaze points. Finally, a real-time hand gesture recognition system is employed for conveying messages to the machine to guide its operations. So far, the proposed framework was tested in a laboratory setting using a robotic dump truck. The results showed that the proposed framework could serve an effective interface to support the interactions between workers and construction machines.

Keywords –

Wearable Sensors; Eye Tracking; Hand Gesture Recognition; Human-Robot Interface

1 Introduction

The construction industry is facing a unique set of challenges, such as low productivity, poor safety records, labor shortage [1,2], etc. Through years of development, construction robots and autonomous machines have demonstrated their potential to improve the construction industry [3]. They have the functional ability to perform construction tasks that are impossible, undesirable, or unsafe for human workers [4]. Also, construction robots have the potential to enhance quality and efficiency of job site operations [5].

Recent advances in robotics make it possible for human-robot collaboration on construction sites [6,7]. This collaboration helps workers transfer some of their current duties to robots and instead devote their effort on high-level planning and cognitive work as robot supervisors [8]. Human workers can also benefit from the assistance of robots in performing repetitive physically-demanding tasks [6]. To maximize the benefits from human-robot work collaboration, a user-friendly interface is critical to support their interactions. However, human-robot interfaces in the context of construction is a less explored field [9].

A variety of interfaces, including visual displays, hand gestures, speech language, and eye tracking, have been developed for human-robot interactions in various industries [10–13]. Among them, non-verbal communication, such as hand gestures and eye tracking, is deemed to be an effective channel in noisy construction environments [12]. As natural and intuitive interfaces, they can provide a standard mode for workers from different backgrounds and cultures to convey correct instructions to a robot [14].

There are many research studies proposed for developing human-robot interfaces based on different types of sensors. The employed sensors include electromyography (sEMG) sensors [15], Inertial Measurement Unit (IMU) [16], radar sensors [17], infrared technology [18], etc. These studies aimed to interpret subjects' intentions [16], understand sign language [19], express human emotions [18], etc. They either relied on hand-crafted features [15] or deep neural networks [20]. The results illustrated the potential of deep neural networks for performing recognition with excellent learning ability.

Although the performance of existing interfaces is promising, one significant challenge they face is in dealing with uncertainty and ambiguity that commonly arise in unstructured and dynamic environments such as

construction sites. It has been well noted that one type of sensor data may not be enough to address the uncertainty and resolve unambiguity. This paper proposes a sensor-based framework which integrates eye tracking and hand gesture recognition for human-robot interaction in construction. In this approach, visual detection of construction machines is first conducted using first person view frames. Based on the detection results and gaze points, the machine-of-interest is then defined. Finally, a real-time system is employed for hand gesture recognition. The recognized gesture would be sent to the machine-of-interest. The effectiveness of the framework was tested in a laboratory study to interact with a robotic dump truck. The results showed that the proposed framework can be used to serve as an effective interface for workers to interact with construction machines.

2 Related Work

Various research studies have been conducted to develop human-robot interfaces. They relied on hand gesture recognition, eye tracking, smart glasses, etc. An overview of these studies is provided below.

2.1 Hand Gesture Recognition

Hand gestures, as a common way to express intent, have various applications in human machine interaction due to their simple, yet effective, nature [21–24]. Various research studies have been conducted to achieve hand gesture recognition. They can be classified into two categories, vision-based methods [23,25] and wearable sensors-based methods [26,27], depending on the type of data source they relied on. Vision-based methods generally relied on hand-crafted features, such as Improved Dense Trajectories (iDT) [28] and Mix Features Around Sparse Keypoints (MFSK) [29]. With technical development, the use of deep learning technologies has become mainstream in gesture recognition. For example, Molchanov et al. [24] combined 3D Convolutional Neural Network (CNN) with recurrent layers to perform simultaneous detection and classification of dynamic hand gestures. The recurrent 3D-CNN enabled the gesture classification without requiring explicit pre-segmentation. Cao et al. [30] presented a framework of C3D+LSTM+RSTTM which augmented C3D with a recurrent spatiotemporal transform module. The presented framework could not only capture short-term spatiotemporal features but also model long-term dependencies. Köpüklü et al. [23] proposed a hierarchical CNN structure to realize the real-time hand gesture recognition. The proposed architecture firstly employed a detector which was a lightweight 3D-CNN to detect the existence of hand gestures and then utilized deep 3D-CNNs to classify the detected gestures.

Motion sensory data provide an alternative data source for hand gesture recognition. For instance, Su et al. [31] presented a robust hand gesture recognition framework based on random forests. The random forests were established using improved decision trees which included the pre-classifiers to avoid the misclassification of gestures with similar features. Côté-Allard et al. [32] applied CNNs on aggregated data from multiple users to identify hand gestures. In their work, CNNs were combined with transfer learning to decrease the data requirement of the training model. Fang et al. [27] designed a new CNN architecture named SLRNet to achieve dynamic gesture recognition. The CNN architecture extracted the features of two hands and fused the features into the fully connected layer. Yuan et al. [20] proposed an improved deep feature fusion network to detect long distance dependency in complex hand gestures. In their work, a LSTM model with fused feature vectors was introduced to classify complex hand motions into corresponding categories.

2.2 Eye Tracking

Conventionally, eye tracking has been regarded as one of the most visible cues for user behavior/intention recognition [33]. There are many efforts dedicated to developing reliable eye tracking-based methods. Zhang et al. [34] presented a novel eye tracking-learning-detection algorithm with tracking feedback. The detection area was adjusted adaptively and narrowed by the tracking feedback to adapt to situations where the human eye was partially blocked or had morphological changes. Santini et al. [35] introduced a novel method named Pupil Reconstructor with Subsequent Tracking (PuReST) for fast and robust pupil tracking. The PuReST consists of three distinct parts: initial pupil detection, shared tracking preamble, outline and greedy tracker. Laddi and Prakash [36] proposed an unobtrusive and calibration-free framework for an eye gaze tracking based interface for a desktop environment. The proposed eye gaze tracking involved a hybrid approach wherein the unsupervised image gradients method computed the iris centers over the eye regions extracted by the supervised regression-based algorithm. Cubero and Rehm [37] relied on eye tracking to obtain eye gazes and then developed an LSTM-based machine learning model to classify human intent. As the technology matures, commercial eye tracking products such as Tobbi glasses 3 [38] and Pupil Core [39] are becoming available in the market and have various potential fields of application.

2.3 Smart Glasses

There are other commonly used human-robot

interfaces including augmented reality (AR) / mixed reality (MR) / virtual reality (VR) glasses, etc. For example, Wang et al. [40] presented a method of manufacture assembly fault detection based on AR. The augmented information interactions made the manufacturing and assembly inspection process more visual and intuitive. Du et al. [41] proposed a novel teleoperation method that allowed users to guide robots through a combined form of AR glasses and Leap Motion Controllers. Users could observe the virtual robot from an arbitrary angle, which enhanced the users' interactive immersion and provided more natural human-machine interaction. Wallmyr et al. [42] employed MR interfaces to display information within the excavator operators' field of view, which enhanced information detectability through quick glances. This practice could help lower operator's mental workload together with an improved rate in detection of presented information. However, their main adoption limitation lies in the expensive hardware and training; and also the AR and MR technologies behind those smart glasses are still not mature and/or suitable enough for engineering and construction [43].

3 Proposed Framework

The overview of the proposed framework is illustrated in Figure 1. The framework consists of three components: visual detection, machine-of-interest generation, and hand gesture recognition. Specifically, the visual detection of construction machines from first person view frames is first conducted. Based on the detection results and gaze points, the machine-of-interest is then generated. Finally, a real-time system is employed to achieve hand gesture recognition.

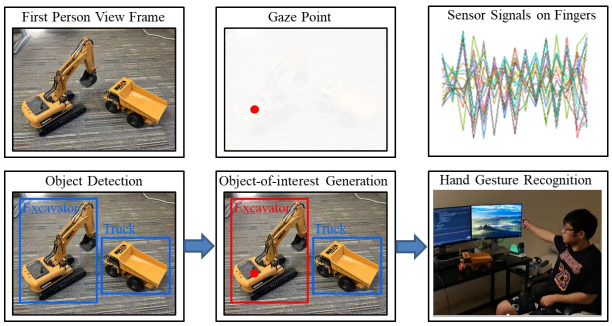

Figure 1. Framework for human-robot collaboration

3.1 Visual Detection

In this component of the framework, an object detection algorithm is employed to extract the regions of construction machines in video sequences. YOLOv3 [44] is selected in this study to detect the construction machines because many research results have verified the high performance of YOLOv3 in various construction object detection scenarios [45,46]. The YOLOv3 system can be generally divided into two steps: feature extraction and detection. First, Darknet-53 is applied to extract features of the whole image and obtain feature embeddings at different scales. Then, these features are fed into different branches of the detector to get bounding boxes and class information. The coordinates of bounding boxes from the detection results are then used as the input for the object-of-interest generation process.

3.2 Machine-of-Interest Generation

In this component of the framework, the machine-of-interest is generated based on the bounding boxes of construction machines and gaze points. This component can be divided into three steps: synchronization for the bounding boxes and gaze points, determination of the machine-of-interest, and interaction mode triggering. First, the bounding boxes and gaze points are synchronized based on a unified timestamp since they are produced or derived from different sensors. Then, the machine is determined as the machine-of-interest if the gaze point resides in its bounding box. As for the triggering of the interaction mode, if the gaze points stay in the bounding box of the machine-of-interest for a duration longer than a threshold τ, the machine will enter the interaction mode and the hand gesture recognition component can then be applied to convey messages to the machine; otherwise, it means that the framework is not confident regarding which machine the user desires to interact with. It should be noted that the selection of τ depends on how likely the user intends to trigger the interaction mode. Here, based on preliminary trials, τ has been set as 0.3 second.

3.3 Hand Gesture Recognition

The purpose of this component of the framework is to apply a hand gesture recognition system to convey messages to the object-of-interest. Specifically, the accelerometer and gyroscope signals are directly captured from sensors attached on fingers as raw data. Several techniques including sampling rate synchronization and Z score normalization are used to preprocess the raw data. Then, a sliding window approach is designed to achieve real-time classification of hand gestures. With the signals coming in continuously, the window moves through the whole set of signals and the preprocessed data in the latest window are fed into a Fully Convolutional Network (FCN)-based classifier to achieve hand gesture recognition. If the highest probability of the classifier is more than a threshold θ, the identification of the hand

gesture is confirmed. Here, based on preliminary trials, θ has been set as 0.95.

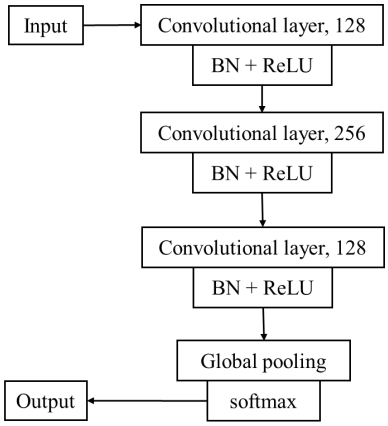

Figure 2. The architecture of FCN classifier

The FCN developed by Wang et al. [47] is selected here since it is superior for multivariate time series classification tasks compared to other deep learning networks [48]. Figure 2 shows an overview of the FCN architecture. It comprises three convolutional blocks where each block contains three operations: a convolution followed by a batch normalization whose result is fed to a ReLU activation function. The result of the third convolutional block is averaged over the entire time dimension which corresponds to the Global Average Pooling (GAP) layer. Finally, a traditional softmax classifier is fully connected to the GAP layer's output.

4 Implementation and Results

4.1 Sensor Selection

Pupil Core [39] is employed as the eye tracking device to get the first person view frames and track the eye gaze points. It is selected since Pupil Core is an open-sourced software, which is conducive for user developments. The structure of Pupil Core is shown in Figure 3. A scene camera is mounted on the front of the eye tracking device to get the first-person view frames. Two eye cameras are facing towards two eyes, separately, to obtain their gaze points.

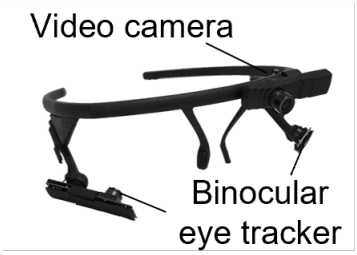

Figure 3. The structure of Pupil Core (adapted from [39])

To capture the hand motions, Tap Strap 2 [49] is selected as the wearable sensor. Compared to other wearable sensors like data gloves which are not easy or comfortable to wear, the Tap sensor is portable, lightweight and easy to wear on the fingers. This is beneficial for the construction workers to complete the tasks using their hands. As shown in Figure 4, the Tap sensor includes five 3-axis accelerometers and one IMU (3-axis accelerometer + 3-axis gyroscope). The five accelerometers are located at five fingers, separately, while IMU is placed on the thumb. In total, there are 21 signal channels captured by the Tap sensor.

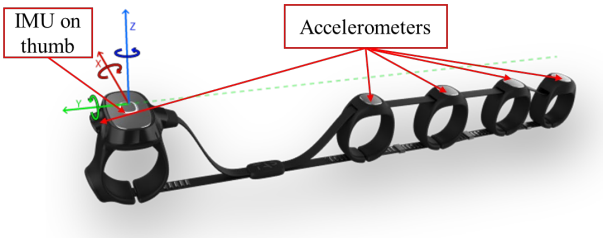

Figure 4. The structure of Tap Strap 2

4.2 Offline Training for Hand Gesture Recognition

The offline training has been conducted on an Ubuntu Linux 64-bit operating system. The hardware configuration is as follows: an Intel® Core™ i7-4820K CPU (Central Processing Unit) @ 3.70 GHz, a 32 GB memory, and an NVIDIA Titan Xp DDR5X @ 12.0 GB GPU (Graphics Processing Unit).

The dataset created in [50] was employed to conduct offline training for hand gesture recognition. The dataset is randomly split into training (66.7%), validation (16.7%) and test (16.6%) sets, resulting in 128 training, 32 validation and 32 test gestures.

For training, the learning rate and the batch size are set as large as possible, i.e., 0.0001 and 16, respectively. When the loss is steady, the learning rate is reduced with a fixed decay factor which is set to 10. Stochastic gradient descent is employed as the optimizer. Table 1

provides a summary of the recognition performance of FCN. The accuracies on validation and test sets are 96.9% and 87.5%, respectively. The inference time achieved on validation and test sets are 0.13 second and 0.14 second, respectively.

Table 1. Recognition performance of FCN

Indexes	Validation set	Test set
Accuracy (%)	96.9	87.5
Inference time (s)	0.13	0.14

4.3 Laboratory Study

A laboratory study was conducted to test whether the proposed framework could serve as an interface to help workers control and/or interact with construction machines. Specifically, the user was asked to stare at the construction machine he/she intended to interact with and then perform hand gestures. The first-person view frames and gaze points were captured by a Pupil core while the hand motions were obtained by a Tap sensor. All these data were transferred to a computer and input into the framework in real time. Based on the recognition results, the corresponding instructions would be sent to a remote controller, where the control signals would be transmitted to operate the truck model remotely.

Figure 5 shows an example of using the proposed framework to remotely control a toy truck to lift its dump box. The user first stared at the truck and made the hand gesture of "hoist" to request the truck model to lift its dump box. The gesture was captured by the framework and the corresponding instruction was sent to the truck model through the remote controller. Following the instruction, the truck model lifted its dump box gradually (Frames 185 and 221). After a short pause, the user stared at an irrelevant place and performed the gesture of "hoist" again. Since the truck did not enter the interaction mode, no recognition results were incurred (Frames 374 and 401).

Although the laboratory study illustrated the feasibility of using the proposed framework as human-robot interface, there are still several technical challenges to be addressed before it can be applied at construction sites. First, the performance of the framework highly depends on visual detection of construction machines. Considering that construction sites are complex and cluttered with tools, materials, workers, etc. The trained detection model needs to be robust to accommodate such complicated characteristics of the environment. Second, the gaze point accuracy is critical for determining which machine the user intends to interact with. However, several complicating factors at construction sites, such as diverse weather conditions and sunlight intensities, pose challenges for accurate estimation of eye gaze points.

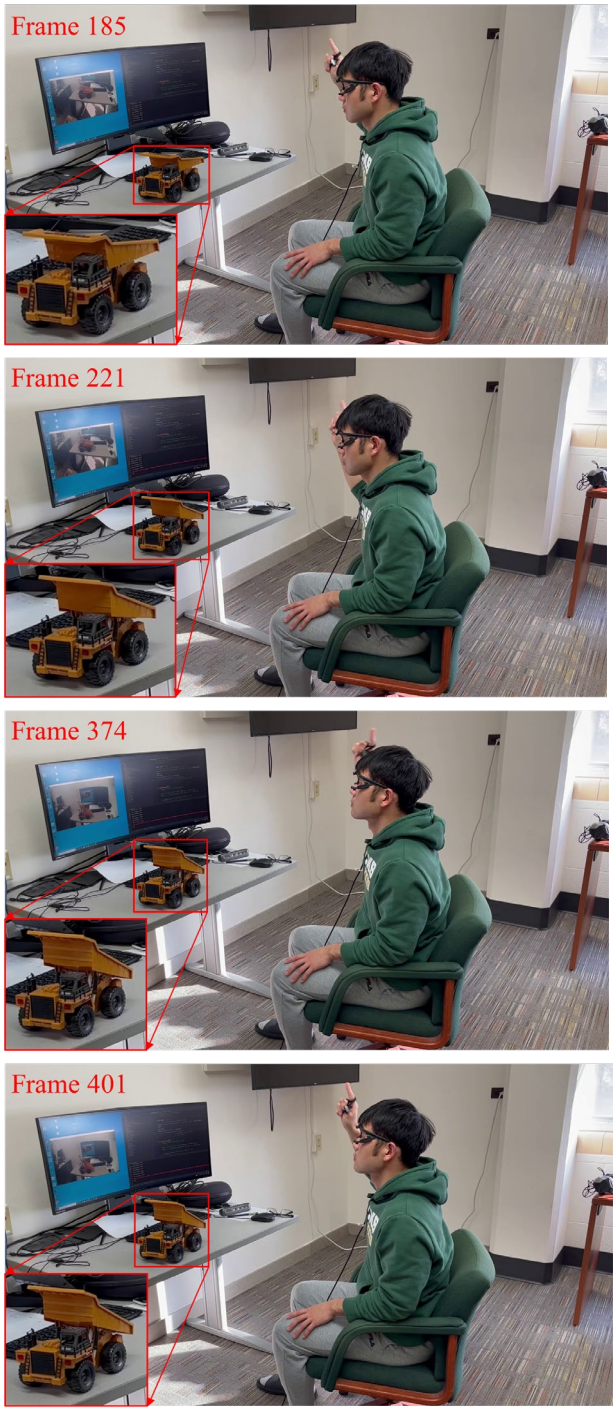

Figure 5. Demonstration of integrated eye-tracking and gesture-based control

5 Conclusions and Future Work

While construction robots have the potential to offer significant benefits to the construction industry, their increased adoption will require user-friendly interfaces for human-robot collaboration. So far, work on human-robot interfaces in the context of the construction domain is limited. This paper has proposed a sensor-based framework which integrates eye tracking and hand gesture recognition for human-robot interaction in construction. The framework comprises three components: visual detection, machine-of-interest generation, and hand gesture recognition. The effectiveness of the framework was tested with a laboratory study to interact with a robotic dump truck. The results show that the proposed framework is suitable for developing an interface to help workers interact with construction machines.

Future work will focus on including more classes of construction gestures into the dataset to make the training and testing of hand gesture classifiers more robust. Additionally, it will investigate the development of a human-robot interaction system using the proposed framework.

Acknowledgment

This paper is based in part upon the work supported by the Wisconsin Alumni Research Foundation (WARF) under Project No. AAJ4872 and the M.A. Mortenson Company. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of WARF or Mortenson.

References

[1] X. Li, W. Yi, H.L. Chi, X. Wang, A.P.C. Chan, A critical review of virtual and augmented reality (VR/AR) applications in construction safety, Autom. Constr. (2018). https://doi.org/10.1016/j.autcon.2017.11.003.

[2] S. Kim, S. Chang, D. Castro-Lacouture, Dynamic Modeling for Analyzing Impacts of Skilled Labor Shortage on Construction Project Management, J. Manag. Eng. (2020). https://doi.org/10.1061/(asce)me.1943-5479.0000720.

[3] Q. Chen, B. García de Soto, B.T. Adey, Construction automation: Research areas, industry concerns and suggestions for advancement, Autom. Constr. (2018). https://doi.org/10.1016/j.autcon.2018.05.028.

[4] H. Ardiny, S. Witwicki, F. Mondada, Construction automation with autonomous mobile robots: A review, in: Int. Conf. Robot. Mechatronics, ICROM 2015, 2015. https://doi.org/10.1109/ICRoM.2015.7367821.

[5] T. Bock, The future of construction automation: Technological disruption and the upcoming ubiquity of robotics, Autom. Constr. (2015). https://doi.org/10.1016/j.autcon.2015.07.022.

[6] S. You, J.H. Kim, S.H. Lee, V. Kamat, L.P. Robert, Enhancing perceived safety in human–robot collaborative construction using immersive virtual environments, Autom. Constr. (2018). https://doi.org/10.1016/j.autcon.2018.09.008.

[7] A. Bauer, D. Wollherr, M. Buss, Human-robot collaboration: A survey, Int. J. Humanoid Robot. (2008). https://doi.org/10.1142/S0219843608001303.

[8] S. You, T. Ye, L.P. Robert, Team Potency and Ethnic Diversity in Embodied Physical Action (EPA) Robot-Supported Dyadic Teams, in: ICIS 2017 Transform. Soc. with Digit. Innov., 2018. http://aisel.aisnet.org/icis2017/HumanBehavior/Presentations/3/.

[9] J. Czarnowski, A. Dąbrowski, M. Maciaś, J. Główka, J. Wrona, Technology gaps in Human-Machine Interfaces for autonomous construction robots, Autom. Constr. (2018). https://doi.org/10.1016/j.autcon.2018.06.014.

[10] P. Majaranta, A. Bulling, Eye Tracking and Eye-Based Human–Computer Interaction, in: 2014. https://doi.org/10.1007/978-1-4471-6392-3_3.

[11] M.A. Goodrich, A.C. Schultz, Human-robot interaction: A survey, Found. Trends Human-Computer Interact. (2007). https://doi.org/10.1561/1100000005.

[12] J. Berg, S. Lu, Review of Interfaces for Industrial Human-Robot Interaction, Curr. Robot. Reports. (2020). https://doi.org/10.1007/s43154-020-00005-6.

[13] R.C. Luo, L. Mai, Human Intention Inference and On-Line Human Hand Motion Prediction for Human-Robot Collaboration, in: IEEE Int. Conf. Intell. Robot. Syst., 2019. https://doi.org/10.1109/IROS40897.2019.8968192.

[14] P.E. Hagan, J.F. Montgomery, J.T. O'Reilly, Accident prevention manual for business & industry: engineering & technology, National Safety Council, 2015.

[15] H. Su, S.E. Ovur, X. Zhou, W. Qi, G. Ferrigno, E. De Momi, Depth vision guided hand gesture recognition using electromyographic signals, Adv. Robot. (2020). https://doi.org/10.1080/01691864.2020.1713886.

[16] T.Y. Pan, C.Y. Chang, W.L. Tsai, M.C. Hu, OrsNet: A hybrid neural network for official sports referee signal recognition, in: MMSports 2018 - Proc. 1st Int. Work. Multimed. Content Anal. Sport. Co-Located with MM 2018, 2018. https://doi.org/10.1145/3265845.3265849.

[17] Z. Zhang, Z. Tian, M. Zhou, Latern: Dynamic Continuous Hand Gesture Recognition Using FMCW Radar Sensor, IEEE Sens. J. (2018). https://doi.org/10.1109/JSEN.2018.2808688.

[18] J.Z. Lim, J. Mountstephens, J. Teo, Emotion recognition using eye-tracking: Taxonomy, review and current challenges, Sensors (Switzerland). (2020). https://doi.org/10.3390/s20082384.

[19] S.A. Khomami, S. Shamekhi, Persian sign language recognition using IMU and surface EMG sensors, Meas. J. Int. Meas. Confed. (2021). https://doi.org/10.1016/j.measurement.2020.108471.

[20] G. Yuan, X. Liu, Q. Yan, S. Qiao, Z. Wang, L. Yuan, Hand Gesture Recognition Using Deep Feature Fusion Network Based on Wearable Sensors, IEEE Sens. J. (2021). https://doi.org/10.1109/JSEN.2020.3014276.

[21] G. Yasmeen, S. Arun, J.N. Swaminathan, S.A.K. Jilani, S. Asif, Efficient Hand Gesture Recognition for Traffic Control System Using ti Sensor Tag, in: 2018 Int. Conf. Comput. Commun. Informatics, ICCCI 2018, 2018. https://doi.org/10.1109/ICCCI.2018.8441483.

[22] Z. Lu, X. Chen, Q. Li, X. Zhang, P. Zhou, A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices, IEEE Trans. Human-Machine Syst. (2014). https://doi.org/10.1109/THMS.2014.2302794.

[23] O. Köpüklü, A. Gunduz, N. Kose, G. Rigoll, Real-time hand gesture detection and classification using convolutional neural networks, in: Proc. - 14th IEEE Int. Conf. Autom. Face Gesture Recognition, FG 2019, 2019. https://doi.org/10.1109/FG.2019.8756576.

[24] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, J. Kautz, Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2016. https://doi.org/10.1109/CVPR.2016.456.

[25] O. Koller, C. Camgoz, H. Ney, R. Bowden, Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos, IEEE Trans. Pattern Anal. Mach. Intell. (2019). https://doi.org/10.1109/tpami.2019.2911077.

[26] A.A. Neacsu, G. Cioroiu, A. Radoi, C. Burileanu, Automatic EMG-based hand gesture recognition system using time-domain descriptors and fully-connected neural networks, in: 2019 42nd Int. Conf. Telecommun. Signal Process. TSP 2019, 2019. https://doi.org/10.1109/TSP.2019.8768831.

[27] B. Fang, Q. Lv, J. Shan, F. Sun, H. Liu, D. Guo, Y. Zhao, Dynamic gesture recognition using inertial sensors-based data gloves, in: 2019 4th IEEE Int. Conf. Adv. Robot. Mechatronics, ICARM 2019, 2019. https://doi.org/10.1109/ICARM.2019.8834314.

[28] H. Wang, D. Oneata, J. Verbeek, C. Schmid, A Robust and Efficient Video Representation for Action Recognition, Int. J. Comput. Vis. (2016). https://doi.org/10.1007/s11263-015-0846-5.

[29] J. Wan, G. Guo, S.Z. Li, Explore Efficient Local Features from RGB-D Data for One-Shot Learning Gesture Recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2016). https://doi.org/10.1109/TPAMI.2015.2513479.

[30] C. Cao, Y. Zhang, Y. Wu, H. Lu, J. Cheng, Egocentric Gesture Recognition Using Recurrent 3D Convolutional Neural Networks with Spatiotemporal Transformer Modules, in: Proc. IEEE Int. Conf. Comput. Vis., 2017. https://doi.org/10.1109/ICCV.2017.406.

[31] R. Su, X. Chen, S. Cao, X. Zhang, Random forest-based recognition of isolated sign language subwords using data from accelerometers and surface electromyographic sensors, Sensors (Switzerland). (2016). https://doi.org/10.3390/s16010100.

[32] U. Côté-Allard, C.L. Fall, A. Drouin, A. Campeau-Lecours, C. Gosselin, K. Glette, F. Laviolette, B. Gosselin, Deep Learning for Electromyographic Hand Gesture Signal Classification Using Transfer Learning, IEEE Trans. Neural Syst. Rehabil. Eng. (2019). https://doi.org/10.1109/TNSRE.2019.2896269.

[33] D.Y. Cho, M.K. Kang, Human gaze-aware attentive object detection for ambient intelligence, Eng. Appl. Artif. Intell. (2021). https://doi.org/10.1016/j.engappai.2021.104471.

[34] J. Zhang, Y. Wu, H. Huang, G. Hou, A New Human Eye Tracking Method Based on Tracking Module Feedback TLD Algorithm, in: Proc. - 20th Int. Conf. High Perform. Comput. Commun. 16th Int. Conf. Smart City 4th Int. Conf. Data Sci. Syst. HPCC/SmartCity/DSS 2018, 2019. https://doi.org/10.1109/HPCC/SmartCity/DSS.2

018.00071.
[35] T. Santini, W. Fuhl, E. Kasneci, PuReST: Robust pupil tracking for real-time pervasive eye tracking, in: Eye Track. Res. Appl. Symp., 2018. https://doi.org/10.1145/3204493.3204578.
[36] A. Laddi, N.R. Prakash, Eye gaze tracking based directional control interface for interactive applications, Multimed. Tools Appl. (2019). https://doi.org/10.1007/s11042-019-07940-3.
[37] C. Gomez Cubero, M. Rehm, Intention Recognition in Human Robot Interaction Based on Eye Tracking, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2021. https://doi.org/10.1007/978-3-030-85613-7_29.
[38] Tobii Inc., Tobii Pro Glasses 3, (2021). https://www.tobiipro.com/product-listing/tobii-pro-glasses-3/ (accessed February 3, 2022).
[39] Pupul Labs, Pupil Core, (2021). https://pupil-labs.com/products/core/ (accessed February 3, 2022).
[40] S. Wang, R. Guo, H. Wang, Y. Ma, Z. Zong, Manufacture assembly fault detection method based on deep learning and mixed reality, in: 2018 IEEE Int. Conf. Inf. Autom. ICIA 2018, 2018. https://doi.org/10.1109/ICInfA.2018.8812577.
[41] G. Du, B. Zhang, C. Li, H. Yuan, A novel natural mobile human-machine interaction method with augmented reality, IEEE Access. (2019). https://doi.org/10.1109/ACCESS.2019.2948880.
[42] M. Wallmyr, T.A. Sitompul, T. Holstein, R. Lindell, Evaluating Mixed Reality Notifications to Support Excavator Operator Awareness, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2019. https://doi.org/10.1007/978-3-030-29381-9_44.
[43] J.M. Davila Delgado, L. Oyedele, T. Beach, P. Demian, Augmented and Virtual Reality in Construction: Drivers and Limitations for Industry Adoption, J. Constr. Eng. Manag. (2020). https://doi.org/10.1061/(asce)co.1943-7862.0001844.
[44] J. Redmon, A. Farhadi, YOLOv3: An incremental improvement, ArXiv. (2018). https://arxiv.org/abs/1804.02767.
[45] F. Wu, G. Jin, M. Gao, Z. He, Y. Yang, Helmet detection based on improved YOLO V3 deep model, in: Proc. 2019 IEEE 16th Int. Conf. Networking, Sens. Control. ICNSC 2019, 2019. https://doi.org/10.1109/ICNSC.2019.8743246.
[46] X. Luo, H. Li, H. Wang, Z. Wu, F. Dai, D. Cao, Vision-based detection and visualization of dynamic workspaces, Autom. Constr. (2019). https://doi.org/10.1016/j.autcon.2019.04.001.
[47] Z. Wang, W. Yan, T. Oates, Time series classification from scratch with deep neural networks: A strong baseline, in: Proc. Int. Jt. Conf. Neural Networks, 2017. https://doi.org/10.1109/IJCNN.2017.7966039.
[48] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.A. Muller, Deep learning for time series classification: a review, Data Min. Knowl. Discov. (2019). https://doi.org/10.1007/s10618-019-00619-1.
[49] Tap Systems Inc., Meet Tap, (2021). https://www.tapwithus.com/.
[50] X. Wang, Z. Zhu, Automatic Recognition of Construction Workers' Hand Gestures Based on Wearable Sensors, Autom. Constr. In Review (2022).