Wearable Sensor-based Hand Gesture Recognition of Construction Workers

X. Wang[a] and Z. Zhu[a]

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

Abstract –

Maintaining good communication is important for keeping the construction site safe and the project running smoothly and on schedule. Hand gestures, as one of the common ways to communicate, are widely used on construction sites due to their simple but effective nature. However, the meaning of these hand gestures was not always captured precisely, which would lead to construction errors and even accidents. This paper presented a feasibility study on investigating whether the hand gestures could be captured and interpreted automatically with wearable sensors. A new dataset which is made of the accelerometer and gyroscope data is created. The created dataset contains 8 classes of hand gestures for instructing tower crane operations and is employed to compare two state-of-the-art deep learning networks, namely, Fully Convolutional Neural Network (FCN) and ResNet, and measure their hand gesture recognition performance. The comparison results indicate that a high classification accuracy (96.9%) could be achieved. Further, a pilot study was conducted in a laboratory environment to test whether the methods used in this study could serve as an interface to help workers control and/or interact with construction machines.

Keywords –

Hand Gesture Recognition; Wearable Sensor; Dataset Creation; Performance Comparison

1 Introduction

In the construction field, maintaining good communication is very important since it keeps the site safe and the project running smoothly and on schedule [1–3]. As one of the common ways to communicate, hand gestures are widely employed on construction sites due to their simple but effective nature [3–5]. They aid workers from different backgrounds and cultures to express their thoughts without difficulty [6]. Besides, consider that words may not be heard clearly on construction sites due to the noisy environment. Hand gestures provide a standard mode for workers to receive correct directions without the need for complicated devices [7].

However, hand gestures may not always be captured or interpreted correctly in the fields, which easily leads to worker injuries/fatalities, work interruption, and stoppage, etc. For example, it was reported that when a crew chief entered an active work area on an All-Terrain Vehicle (ATV), he was asked to leave by a foreman using a hand gesture. However, the gesture was not captured by the crew chief. The result was that the ATV was hit by a bulldozer and the chief suffered a fractured leg [8]. Another accident was noted when a driver was driving a concrete truck. He misread the hand gestures given by an officer and hit a 27-year-old electrician, who was working on the replacement of traffic lights from the buck of his truck [9].

These accidents indicate the necessities to automatically capture and interpret hand gestures in construction fields. So far, there are many research studies proposed for hand gesture recognition based on wearable sensors. The employed wearable sensors included Inertial Measurement Unit (IMU) [10], surface electromyography (sEMG) sensors [11], etc. The methods in these studies were employed to recognize sports referee gestures [10], promote human-computer interactions [11], understand sign languages [12], etc. They were based on either hand-crafted features [11] or deep neural networks [13]. The results illustrated the potential of using deep neural networks to capture and recognize hand gestures with excellent learning ability.

Although the performance of existing methods for hand gesture recognition is promising, it remains unclear whether they could be applied in the construction field to capture and interpret hand gestures made by construction workers. This paper presented a feasibility study on investigating whether hand gestures in the construction field could be recognized automatically with wearable sensors. A new dataset containing 8 classes of hand gestures for instructing tower crane operations was created. To measure the recognition performance with the created dataset, two state-of-the-art deep neural networks, namely, Fully Convolutional Neural Network (FCN) and ResNet were employed to achieve hand gesture recognition. The recognition results demonstrated that a high classification accuracy (96.9%) could be achieved, which illustrated the feasibility and potential of wearable sensors to automate the hand gesture recognition in the construction field.

2 Related Work

Currently, various research studies have been developed to achieve hand gesture recognition. They could be classified into two categories, vision-based methods [14,15] and wearable sensors-based methods [16,17], depending on the type of data source they relied on.

2.1 Vision-Based Hand Gesture Recognition

So far, many efforts have been dedicated to hand gesture recognition based on the video data. They either relied on hand-crafted features or through deep learning. Traditional methods generally relied on hand-crafted features, such as Improved Dense Trajectories (iDT) [18] and Mix Features Around Sparse Keypoints (MFSK) [19]. Besides these features, many research studies were focused on deriving novel features to represent the appearance, shape, and motion changes of a gesture [20–22]. Currently, the use of deep learning technologies has become a mainstream in gesture recognition. For example, Miao et al. [23] proposed a multimodal gesture recognition method using a Res3D network. The extracted spatiotemporal features from the Res3D were combined through canonical correlation analysis and then the final recognition was made by a linear SVM classifier. 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. [25] 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. [14] proposed a hierarchical CNN structure to realize the real-time hand signal 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.

2.2 Wearable Sensors-Based Hand Gesture Recognition

Motion sensory data provide an alternative data source to achieve hand gesture recognition. They usually can be collected by various wearable sensors attached on human bodies or placed near hands, such as surface electromyography (sEMG) sensors, Inertial Measurement Units (IMU), etc. Currently, many methods have been developed to recognize hand gestures based on wearable sensors. For sEMG sensors, Su et al. [26] 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. Allard et al. [27] 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. For IMU sensors, Fang et al. [16] 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. Jirak et al. [28] introduced an echo state network (ESN) framework for continuous gesture recognition. The framework included LSTM layers to achieve the automatic detection of the start and end phase of a gesture.

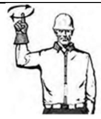
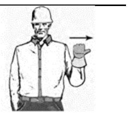
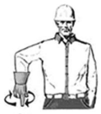
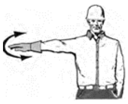
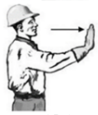
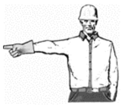
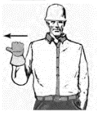
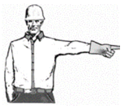
3 Methodology

3.1 Dataset Design

In this study, the accelerometer and gyroscope signals captured from wearable sensors are employed as raw data. The accelerometer signals are used to measure the vibration or acceleration of hand/finger movements while the gyroscope signals refer to the rotational motions of hand/finger. During data collection, the subject who makes hand gestures was requested to wear the sensor on his hand at first. To capture the characteristics of construction site environments, the way to make hand gestures is considered when creating the dataset. The subject was moving and making hand gestures synchronously.

The hand gestures made by the subject are those commonly seen on construction sites. For example, tower cranes are the most frequently shared resources [29,30], which are mainly used for lifting heavy things and transporting them to other places. 8 classes of hand gestures for directing tower crane operations were selected here, as indicted in Table 1.

Table 1. Hand gestures for instructing tower crane operations adapted from [4,5]

Hand gesture	Examples	Hand gesture	Examples
Hoist		Trolley travel left	
Lower		Stop	
Tower travel		Swing right	
Trolley travel right		Swing left	

In addition, several techniques are employed to preprocess the raw data. First, all the sensors are unified to sample at 5HZ for synchronization since the original sampling rates may be different for various sensors. Besides, the raw signals coming from sensors may be affected by external noise. To eliminate noise, the mean and standard deviation of raw signals are firstly subtracted from the data to obtain the offset of the signals. Then, Z score standardization (Eq. 1) is utilized to rescale the raw data, which allows all signals to be considered with equal importance.

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \tag{1}$$

where z_i is the standard score for *i*-th signal channel, x_i refers to the raw data for *i*-th signal channel, μ_i and σ_i are the mean and standard deviation of *i*-th signal channel, separately.

3.2 Hand Gesture Recognition

Based on the findings from existing studies of recognition methods, deep learning networks illustrated an excellent learning ability among all the methods [13,31]. Thus, in this study, the recognition methods are selected from state-of-the-art deep learning networks which achieved high recognition accuracy. In previous studies, FCN and ResNet networks achieved better recognition performance compared to other deep learning networks [32,33]. Therefore, they are selected here to test the performance on the created dataset. The characteristics of the selected methods are summarized as below.

The architecture of FCN is first composed of 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 whole 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. Figure 1 shows an overview of FCN architecture. More details of the network architecture could be found in the work of Wang et al. [33].

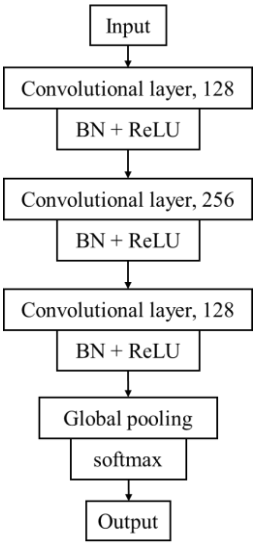


Figure 1. The architecture of FCN

In addition, ResNet is composed of three residual blocks followed by a GAP layer and a final softmax classifier whose number of neurons is equal to the number of classes in a dataset [34]. The main characteristic of ResNet is the shortcut residual connection between consecutive convolutional layers. Each residual block is first composed of three convolutions whose output is added to the residual block's input and then fed to the next layer. The number of filters for all convolutions is fixed to 64, with the ReLU activation function that is preceded by a batch normalization operation. In each residual block, the filter's length is set to 8, 5 and 3 respectively for the first, second and third convolution. The architecture of ResNet is indicated in Table 2.

Table 2. The architecture of ResNet

Layer name	Number of blocks
Conv1	192
Conv2	384
Conv3	384
Global pooling, fc layer with softmax	---

The recognition performance is evaluated in terms of gesture classification accuracy. The classification accuracy is defined as the percentage of correctly labeled gesture samples by the recognition method.

4 Results

4.1 Dataset Preparation

To create the dataset, Tap Strap 2 [35] is selected as the wearable sensor. As shown in Figure 2, it 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. There are totally 21 signal channels captured by the Tap sensor.

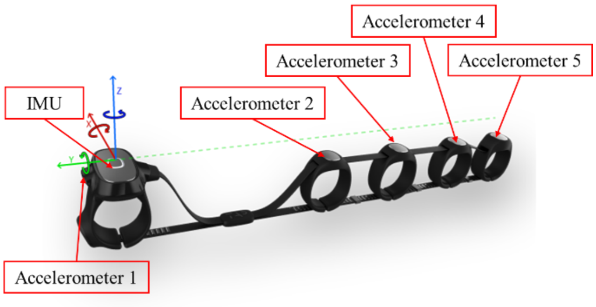


Figure 2. The structure of Tap Strap 2

During data collection, each class of gesture was performed 20 times by the subject, resulting in 160 (20 × 8) gestures in the dataset. The duration for each gesture is 10 s. Examples of the collected data for one gesture could be found in Figure 3. The details of the dataset after preprocessing are listed in Table 3. The average signal values for these 8 classes of gestures are -0.043, -0.873, 0.200, -0.069, 0.148, 0.042, 0.233 and 0.361, separately.

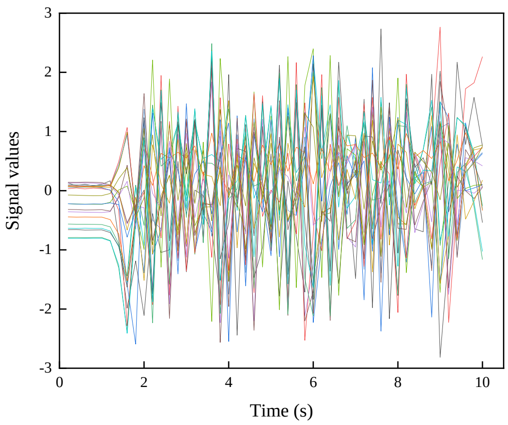


Figure 3. Examples of the data for one gesture

Table 3. Dataset description

Gesture class	Min	Mean	Max	Variance
Hoist	-4.287	-0.043	3.513	1.027
Lower	-4.876	-0.873	3.670	1.098
Tower travel	-3.760	0.200	4.126	1.008
Trolley travel right	-3.664	-0.069	4.056	1.061
Trolley travel left	-3.240	0.148	3.699	0.897
Stop	-4.531	0.042	4.294	1.143
Swing right	-3.381	0.233	2.759	0.498
Swing left	-3.306	0.361	2.970	0.470
Total	-4.876	0.000	4.294	1.000

4.2 Training for Recognition Methods

The recognition methods have been implemented on an Ubuntu Linux 64-bit operating system. The hardware configuration includes 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). For training, the dataset is randomly split into training (80%) and testing (20%) sets, resulting in 128 training and 32 testing gestures.

As for training details, the learning rate and the batch size are set as large as possible. When the loss is steady, the learning rate is reduced with a fixed decay factor which is set to 10. Stochastic gradient descent (SGD) is employed as the optimizer. Specifically, the learning rate of FCN and ResNet is set as 0.0001 while the batch sizes of these two methods are 16 and 64, separately. Figure 4 shows the loss reduction along with the training progress.

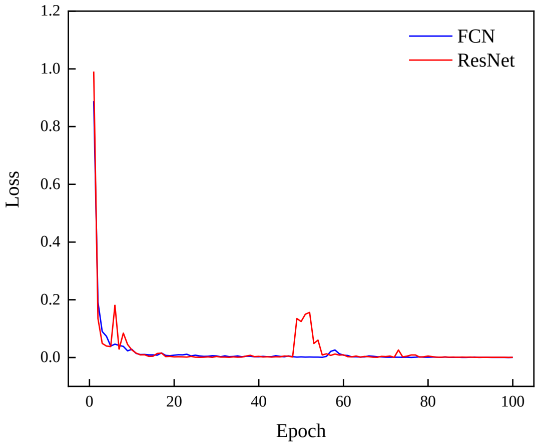


Figure 4. The loss reduction along with the training progress

4.3 The Recognition Performance

Table 4 presented the recognition performance of FCN and ResNet. For both FCN and ResNet, the networks achieve the training accuracy of 100% and the validation accuracy of 96.9%. The training losses of FCN and ResNet are 0.000 and 0.001, separately, while the validation losses of these two networks are 0.030 and 0.015, respectively. Both of them achieve satisfying recognition performance.

Table 4. The Recognition performance of best models

Indexes	FCN	ResNet
Training loss	0.000	0.001
Validation loss	0.030	0.015
Training accuracy	100%	100%
Validation accuracy	96.9%	96.9%

4.4 Pilot Study

Further, a pilot study was conducted in a laboratory environment to test whether the methods used in this study could serve as an interface to help workers control and/or interact with construction machines. Specifically, the subject was asked to perform hand gestures, which were captured by a Tap sensor connecting to a computer. The motion data captured by the Tap sensor were input into the deep learning networks 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 showed an example of using the deep learning networks to remotely control a toy truck to move and lift its dump box. The subject firstly made the hand gesture of "swing right" to request the truck model to turn right. 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 drove towards the right gradually. After a short pause, the subject then performed the gesture of "hoist" to request the truck model to lift its dump box. The truck model received the corresponding instruction and then lifted its dump box.

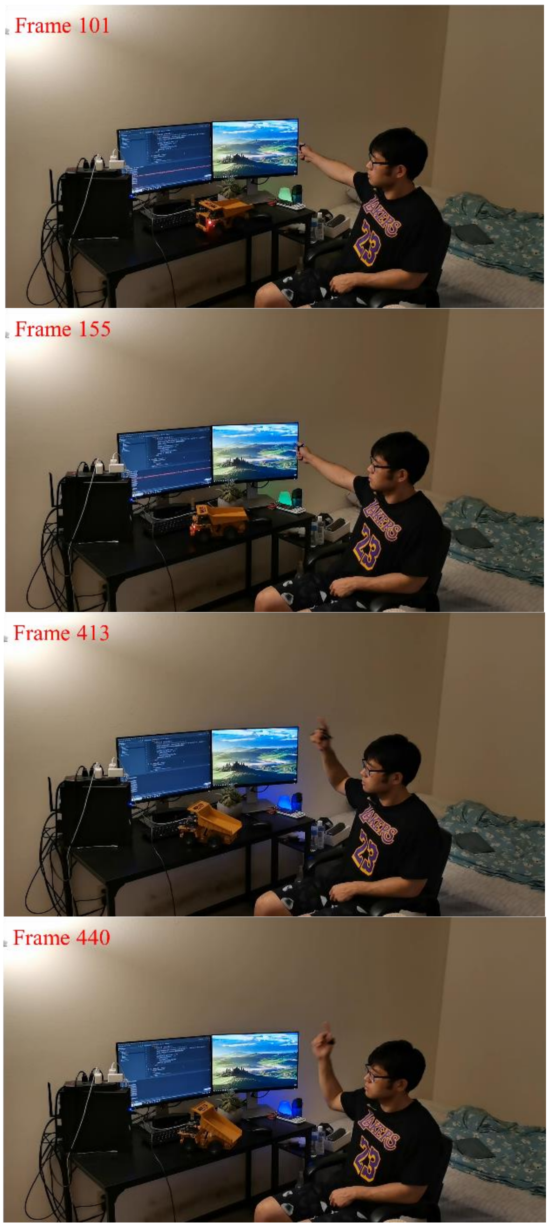


Figure 5. Examples of the test results in the pilot study

5 Conclusions and Future Work

In construction fields, it is common for workers to rely on hand gestures to communicate and express thoughts because of their simple but effective nature. However, the meaning of these hand gestures was not always captured precisely. As a result, it would lead to construction errors and even accidents. This paper investigated whether the recognition of hand gestures could be automated based on wearable sensors in construction. A new dataset with 8 classes of hand gestures in construction is introduced and employed to evaluate two state-of-the-art recognition networks, FCN and ResNet. The results indicated a high classification accuracy (e.g. 96.9%) could be achieved. Further, a pilot study was conducted in a laboratory environment to test whether the methods used in this study could serve as an interface to help workers control and/or 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. Besides, it will investigate how to use the gesture detection and classification to control construction machines.

Acknowledgment

This paper is based in part upon the work supported by the Wisconsin Alumni Research Foundation (WARF) under Project No. AAD5524 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] The Off-highway Plant and Equipment Research Centre, Hand signals for when excavations are used as cranes: A Voluntary code of practice, Birmingham City University, 2019.

[2] Kines P., Andersen L.P.S., Spangenberg S., Mikkelsen K.L., Dyreborg J. and Zohar D. Improving construction site safety through leader-based verbal safety communication, Journal of safety research, 41(5):399-406, 2010.

[3] Neitzel R.L., Seixas N.S. and Ren K.K. A Review of Crane Safety in the Construction Industry, Applied occupational and environmental hygiene, 16(12):1106-1117, 2001.

[4] The American Society of Mechanical Engineers, Safety Standard for Cableways, Cranes, Derricks, Hoists, Hooks, Jacks, and Slings, 2012.

[5] National Commission for the Certification of Crane Operators, Signalperson Reference Manual, 2014.

[6] Bust P.D., Gibb A.G.F. and Pink S. Managing construction health and safety: Migrant workers and communicating safety messages, Safety science, 46(4):585-602, 2008.

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

[8] ENFORM, D8 Bulldozer Contact with Surveyor on ATV. Online: http://www.energysafetycanada.com/files/safety-alerts/SA05-13-ATV-Bulldozer.pdf, Accessed: 13/07/2021.

[9] Reakes, Traffic Signal Worker Thrown From Bucket In Stamford. Online: https://dailyvoice.com/connecticut/stamford/news/traffic-signal-worker-thrown-from-bucket-in-stamford/732557/, Accessed: 10/06/2021.

[10] Pan T.Y., Chang C.Y., Tsai W.L. and Hu M.C. OrsNet: A hybrid neural network for official sports referee signal recognition. In Proceedings of the 1st International Workshop on Multimedia Content Analysis in Sports, pages 51-58, New York, NY, USA, 2018.

[11] Su H., Ovur S.E., Zhou X., Qi W., Ferrigno G. and De Momi E. Depth vision guided hand gesture recognition using electromyographic signals, Advanced Robotics, 34(15):985-997, 2020.

[12] Khomami S.A. and Shamekhi S. Persian sign language recognition using IMU and surface EMG sensors, Measurement, 168:108471, 2021.

[13] Yuan G., Liu X., Yan Q., Qiao S., Wang Z. and Yuan L. Hand Gesture Recognition Using Deep Feature Fusion Network Based on Wearable Sensors, IEEE Sensors Journal, 21(1):539-547, 2021.

[14] Köpüklü O., Gunduz A., Kose N. and Rigoll G. Real-time hand gesture detection and classification using convolutional neural networks, In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pages 1-8, Lille, France, 2019.

[15] Koller O., Camgoz N.C., Ney H. and Bowden R. Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos, IEEE transactions on pattern analysis and machine intelligence, 42(9):2306-2320, 2020.

[16] Fang B., Lv Q., Shan J., Sun F., Liu H., Guo D. and Zhao Y. Dynamic gesture recognition using inertial sensors-based data gloves, In 2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM), pages 390-395, Toyonaka, Japan, 2019.

[17] Neacsu A.A., Cioroiu G., Radoi A. and Burileanu C. Automatic EMG-based hand gesture recognition system using time-domain descriptors and fully-connected neural networks, In 2019 42nd International Conference on Telecommunications and Signal Processing (TSP), pages 232-235, Budapest, Hungary, 2019.

[18] Wang H., Oneata D., Verbeek J. and Schmid C.

A Robust and Efficient Video Representation for Action Recognition, International journal of computer vision, 119(3):219-238, 2016.

[19] Wan J., Guo G. and Li S.Z. Explore Efficient Local Features from RGB-D Data for One-Shot Learning Gesture Recognition, IEEE transactions on pattern analysis and machine intelligence, 38(8):1626-1639, 2016.

[20] Singha J. and Das K. Recognition of Indian Sign Language in Live Video, arXiv preprint arXiv:1306.1301, 2013.

[21] Wang X., Xia M., Cai H., Gao Y. and Cattani C. Hidden-Markov-Models-based dynamic hand gesture recognition, Mathematical Problems in Engineering, 2012.

[22] Lin L., Cong Y. and Tang Y. Hand gesture recognition using RGB-D cues, In 2012 IEEE International Conference on Information and Automation, pages 311-316, Shenyang, China, 2012.

[23] Miao Q., Li Y., Ouyang W., Ma Z., Xu X., Shi W. and Cao X. Multimodal Gesture Recognition Based on the ResC3D Network, In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 3047-3055, Venice, Italy, 2017.

[24] Molchanov P., Yang X., Gupta S., Kim K., Tyree S. and Kautz J. Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks, In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4207-4215, Las Vegas, NV, USA, 2016.

[25] Cao C., Zhang Y., Wu Y., Lu H. and Cheng J. Egocentric Gesture Recognition Using Recurrent 3D Convolutional Neural Networks with Spatiotemporal Transformer Modules, In Proceedings of the IEEE International Conference on Computer Vision, pages 3763-3771, Venice, Italy, 2017.

[26] Su R., Chen X., Cao S. and Zhang X. Random forest-based recognition of isolated sign language subwords using data from accelerometers and surface electromyographic sensors, Sensors, 16(1):100, 2016.

[27] Côté-Allard U., Fall C.L., Drouin A., Campeau-Lecours A., Gosselin C., Glette K., Laviolette F. and Gosselin B. Deep Learning for Electromyographic Hand Gesture Signal Classification Using Transfer Learning, IEEE Transactions on Neural Systems and Rehabilitation Engineering, 27(4):760-771, 2019.

[28] Jirak D., Tietz S., Ali H. and Wermter S. Echo State Networks and Long Short-Term Memory for Continuous Gesture Recognition: a Comparative Study, Cognitive Computation, 1-13, 2020.

[29] Al-Hussein M., Athar Niaz M., Yu H. and Kim H. Integrating 3D visualization and simulation for tower crane operations on construction sites, Automation in Construction, 15(5):554-562, 2006.

[30] Yang J., Vela P.A., Teizer J. and Shi Z.K. Vision-based crane tracking for understanding construction activity, Journal of Computing in Civil Engineering, 28(1):103-112, 2011.

[31] Pan T.-Y., Tsai W.-L., Chang C.-Y., Yeh C.-W. and Hu M.-C. A Hierarchical Hand Gesture Recognition Framework for Sports Referee Training-Based EMG and Accelerometer Sensors, IEEE Transactions on Cybernetics, 2020.

[32] Ismail Fawaz H., Forestier G., Weber J., Idoumghar L. and Muller P.A. Deep learning for time series classification: a review, Data mining and knowledge discovery, 33(4):917-963, 2019.

[33] Wang Z., Yan W. and Oates T. Time series classification from scratch with deep neural networks: A strong baseline, In 2017 International joint conference on neural networks (IJCNN), pages 1578-1585, Anchorage, AK, USA, 2017.

[34] He K., Zhang X., Ren S. and Sun J. Deep residual learning for image recognition, In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770-778, Las Vegas, NV, USA, 2016.

[35] Tap Systems Inc. Meet Tap. Online: https://www.tapwithus.com/, Accessed: 05/02/2021.