Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 4, August 2021, pp. 2223~2230
ISSN: 2302-9285, DOI: 10.11591/eei.v10i4.2926

Convolutional neural networks framework for human hand gesture recognition

Aseel Ghazi Mahmoud[1], Ahmed Mudheher Hasan[2], Nadia Moqble Hassan[3]
[1]College of Nursing, University of Baghdad, Iraq
[2]Control and Systems Engineering Department, University of Technology, Iraq
[3]Computer Engineeirng Department, Mustansiraih University, Iraq

Article Info

Article history:

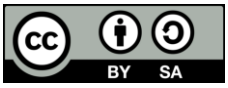
Received Feb 27, 2021
Revised May 20, 2021
Accepted Jun 16, 2021

Keywords:

Accurate recognition
Convolutional neural networks
Human hand gesture
Infrared images
Multiple layers

ABSTRACT

Recently, the recognition of human hand gestures is becoming a valuable technology for various applications like sign language recognition, virtual games and robotics control, video surveillance, and home automation. Owing to the recent development of deep learning and its excellent performance, deep learning-based hand gesture recognition systems can provide promising results. However, accurate recognition of hand gestures remains a substantial challenge that faces most of the recently existing recognition systems. In this paper, convolutional neural networks (CNN) framework with multiple layers for accurate, effective, and less complex human hand gesture recognition has been proposed. Since the images of the infrared hand gestures can provide accurate gesture information through the low illumination environment, the proposed system is tested and evaluated on a database of hand-based near-infrared which including ten gesture poses. Extensive experiments prove that the proposed system provides excellent results of accuracy, precision, sensitivity (recall), and F1-score. Furthermore, a comparison with recently existing systems is reported.



Corresponding Author:

Aseel Ghazi Mahmoud
College of Nursing
University of Baghdad
Baghdad, Iraq
Email: aseel@conursing.uobaghdad.edu.iq

1. INTRODUCTION

The recognition of human gestures represents the recognizing of category labels from a video or an image that includes gestures created via users. Human gestures are meaningful and expressive movements of the human body including physiological motions of the arms, hands, fingers, face, head, or body for the purpose of environmental interaction or conveying meaningful information. Amongst human gestures, hand gesture represents one of the most common, natural, and expressive kind of body language to convey emotions and attitudes of human interaction [1]-[3].

Generally, there are two categories of hand gesture recognition; static and dynamic. The static recognition concentrates on the inner information of an image (a hand of stable shape). While dynamic recognition works on exploring the characteristics of spatial-temporal (a series of hand movements) [4]. The studying of static recognition is very meaningful, since it is capable of conveying various shapes of hands into particular information without motion cues, also, reducing the problem of redundant frames that appears in the dynamic recognition [5]-[7].

Hand gesture recognition approaches provide an inspirational area of researches since they are capable of facilitating communications and offering a natural interaction means to be utilized in a diversity of

applications. These approaches can be categorized into sensor-based and vision-based approaches [8]-[10]. In sensor-based approaches, wearable sensors are attached directly to the hand on gloves for detecting the physical reaction of finger bending or hand movements. The data collected from these sensors are then processed by utilizing a microcontroller or a computer [11]. Despite the fact that the sensor-based approaches have granted good results, they have different drawbacks such as discomfort when wearing gloves for long periods, skin adverse reactions, infection, or damage in people who have sensitive skin, furthermore, some sensors are very expensive. While the vision-based approaches represent cost-effective approaches that didn't need uncomfortable gloves to be worn [8], [12], [13].

In recent years, the convolutional neural networks (CNN) overtake the complex pre-processing of images and assist in classifying and recognizing images, therefore, it is extensively utilized when handling images. Numerous researchers have started to implement CNN for recognizing human gestures and achieved good results [14]-[17]. In this paper, the proposed CNN framework works on recognizing static hand gestures for obtaining effective and accurate results. The remains of the paper are structured is being as; the recently existing related works are reviewed in section 2; the proposed CNN framework for recognizing hand gestures is described in section 3; the extensive experiments are explained and discussed in section 4; the conclusion and future work are stated in the last section.

2. RELATED WORKS

Recently, the existing researches have concentrated on the power of deep learning and its effectiveness in extracting and classifying high-level features of data. S. Hussain *et al*. [18], utilized the visual geometric group16-convolutional neural networks (VGG16-CNN) model which includes thirteen convolution layers succeeded by three fully-connected layers. However, this model requires to be modified for reaching the desired outcomes. Therefore, two layers have been changed with a set of layers for classifying eleven hand gesture classes. The Classifier utilized a dataset that includes more than 55000 self-acquired images (from 7 different volunteers), 70% were utilized for training, and 30% for testing. When the recognized hand gesture is dynamic then it will be traced for detecting motion. the obtained accuracy of this model was 93.09%.

Chaudhary and Raheja [19], proposed an ANN-based system for recognizing light invariant hand gestures in which unique features for the hand gestures were identified by using orientation histogram and classified using artificial neural network (ANN). The designed ANN includes eighteen neurons in the input layer, nine hidden neurons, and six neurons at the output. For each gesture, there are fourteen different images regarding six kinds of gestures collected from different sources that have been utilized for training ANN, and the achieved accuracy was 92.86%. However, ANN is criticized since it takes a long time to decide the optimum number of hidden layers and the number of nodes in each layer which makes it impractical for real-time implementations.

Sahoo *et al*. [20], proposed a deep CNN feature-based static hand gestures recognition system in which deep features are extracted using fully connected layers of pre-trained artificial neural network (AlexNet), then the redundant features are reduced by using the principal component analysis (PCA). After that, a support vector machine (SVM) as a classifier was utilized for classifying the poses of hand gestures. The system performance was evaluated on 36 gesture poses using american sign language (ASL) dataset, and the obtained average accuracy was 87.83%.

Wang *et al*. [21], presented a recognition model of hand gestures based on CNN for analyzing human behavior in the scenario of double teachers' classroom learning and instruction. The recognized hand gestures of instructors can be exploited for analyzing the nonverbal behaviors of teachers that attract the attention of learners and improve their learning results. In this model, the features of hand gesture images are extracted using a non-linear neural network that includes four convolution layers. The CNN with three convolution layers is designed for achieving robust recognition. This model is tested and evaluated using a dataset of 38425 infrared hand gesture images which represent the key frames extracted from the infrared videos. These images are labeled into two kinds, pointing and non-pointing gestures. The dataset of infrared hand gestures is separated into 80% and 20% for training and testing data, respectively. The obtained ratio of recognition accuracy for this model was more than 92%.

Song *et al*. [22], presented a recognition model of hand gestures in which multiple channel features are extracted for describing a large number of hand gestures, then an algorithm of local-global feature fusion is constructed for combining these multiple features, and the weights of features are tuned automatically. After that, an image kernel of a huge scale is constructed for integrating fused features and consequently fed to the support vector machine classifier to understand the hand gesture. The experiments of this model are accomplished using the halearn gesture dataset (CGD) which includes over 50000 images of hand gestures with 249 gesture labels, and the obtained average of recognition accuracies was 84.32%.

Most of the indicated related works are tried to classify multi-poses of human hand gestures. However, the accurate recognition of hand gesture poses is still a difficult task due to several aspects like the small size of the dataset, and low illumination of the acquired hand gesture images. For overcoming these challenges, this paper proposes an accurate and effective CNN framework which deals with a large dataset of infrared images for recognizing ten kinds of hand gesture poses.

3. THE PROPOSED CNN FRAMEWORK FOR RECOGNIZING HAND GESTURES

The proposed CNN framework is designed to obtain the best results for human hand gesture recognition. The CNN framework architecture is shown in Figure 1, and its details are summarized in Table 1. The first layer in the proposed CNN framework is the input layer which provides the input data to the subsequent layers. After this layer, there are two phases; The first phase is the feature extraction and the second one is the classification. These phases include multiple layers, each of which holds specific characteristics that require to be investigated.

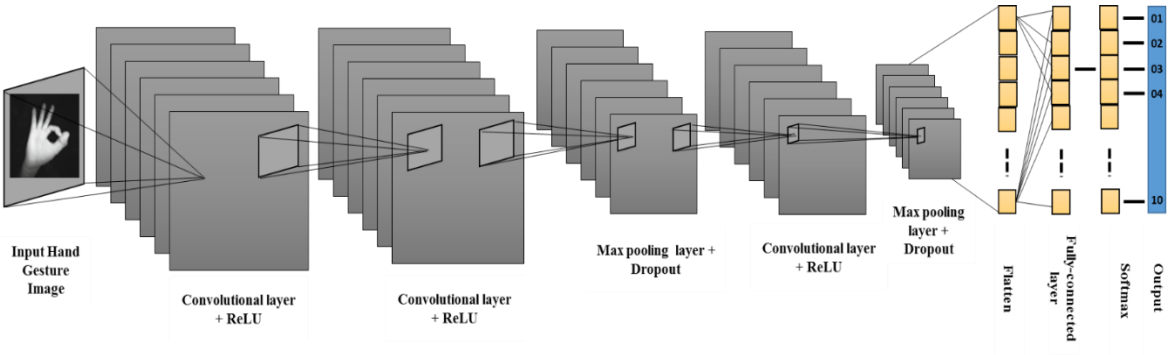


Figure 1. The CNN framework architecture

Table 1. The details of CNN framework architecture

Type of Layer	Output	Parameters
2D Convolutional layer	(None, 48, 48, 32)	320
Activation	(None, 48, 48, 32)	0
2D Convolutional layer	(None, 46, 46, 32)	9248
Activation	(None, 46, 46, 32)	0
Max pooling layer	(None, 23, 23, 32)	0
Dropout	(None, 23, 23, 32)	0
2D Convolutional layer	(None, 21, 21, 64)	18496
Activation	(None, 21, 21, 64)	0
Max pooling layer	(None, 10, 10, 64)	0
Dropout	(None, 10, 10, 64)	0
Flatten layer	(None, 6400)	0
Dense layer	(None, 256)	1638656
Dense layer	(None, 10)	2570
Total Parameters: 1,669,290		
Trainable Parameters: 1,669,290		
Non-trainable Parameters: 0		

3.1. Feature extraction phase

The first phase works on extracting features from the input hand gesture images, and it includes three convolutional layers. Each convolutional layer requires making a convolution on the input via utilizing a (3×3) kernel to produce a feature map. In the convolution process, each kernel is slid over the input and the stride size is considered as one (i.e; kernel moves pixel by pixel). The matrix multiplication is performed at each place and the output is added into a specific feature map. Each input grayscale image is transformed into a 2D matrix with specified height and width. Many convolutions are conducted on an input matrix with various kernels for generating diverse feature maps. These diverse feature maps are aggregated to obtain the convolutional layer output. Each convolutional layer is followed by the rectified linear unit (ReLU). ReLU represents an activation function that works on thresholding the inputs (changing the inputs to zero when their values less than zero) and generating non-linear output as in the following equation:

$$Af(v) = max\ (v, 0) \quad (1)$$

In this phase of CNN architecture, the second, and third convolutional layers are followed by the pooling layer. The main reason for using the pooling layer is to minimize the dimensionality and reduce computations with fewer parameters. Moreover, it is working on regulating the overfitting and reducing the time of training. In this layer, the max-pooling is used which selects the maximum value in each window (2×2), therefore, the size of the feature map is reduced while keeping the significant information. The dropout approach can be added to the max-pooling layers to decrease the overfitting and provide good improvement in the predictions in which a predefined ratio of the neurons in a hidden layer is randomly dropped per each iteration of the training phase.

3.2. Classification phase

The second phase of CNN architecture represents the classification phase which includes fully connected layers. In the fully connected layer, the neurons hold complete connections for every activation from the former layer. The fully connected layer performs its functions via implementing the same basics of a typical neural network. But, the 1D data can only be accepted via this layer. To transform 2D data to 1D data, the flatten function is utilized. The softmax layer works on taking the output of the final fully connected layer and transforming the real value into a distribution of probability. The SoftMax function can be given in (2) [23], [24]:

$$S_i = \frac{e^{b_i}}{\sum_{i=1}^{n} e^{b_i}} \quad (2)$$

Where S_i indicates the number of softmax output *i*, b_i indicates the output *i* before softmax, and *n* is number of output nodes. For the final layer, the size of the output is equal to the number of hand gesture classes (ten classes).

4. RESULTS AND DISCUSSION

In this proposed system, the dataset of the hand-based near-infrared [25] was used in which various poses of right-hand gestures have been gained for ten different subjects (five men and five women) by utilizing a sensor of leap motion located on a table. This dataset involves 20000 hand gesture images of size 50×50. Figure 2 shows the samples of hand gesture images' poses from the hand-based near-infrared dataset. The dataset of the hand-based near-infrared is separated into 70% (training) and 30% (testing) with 32 batch size, 20 epochs, and Adam optimizer. Table 2 illustrates the classes description and labels of the near-infrared hand gestures poses.

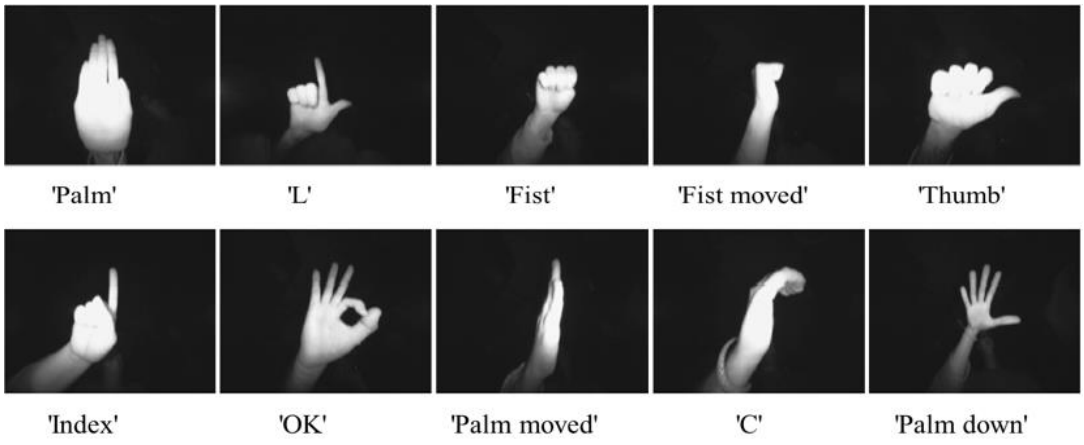


Figure 2. Samples of hand gesture images classes from the hand-based near-infrared dataset

The main aim of this paper is to construct a network with less complexity and high accuracy. Therefore, the proposed CNN framework can be evaluated by utilizing several criteria; precision, sensitivity (Recall), F1-score, and accuracy. These criteria are given in the following equations:

$$Precision = \frac{True_P}{True_P + False_P} \quad (3)$$

$$Sensitivety\ (Recall) = \frac{True_P}{True_P + False_N} \quad (4)$$

$$F1 - Score = \frac{2 \times (Sensitivety \times Precision)}{(Sensitivety + Precision)} \quad (5)$$

$$Accuracy = \frac{(True_P + True_N)}{(True_P + True_N + False_N + False_P)} \quad (6)$$

Where, $True_P$ indicates the true positive, $True_N$ indicates the true negative, $False_N$ indicates the false negative, and $False_P$ indicates the false positive.

Table 2. The classes description and labels of the hand-based near-infrared dataset

The Classes of Hand Gestures	Associated Label
Open palm- aligned to sensor	'Palm': 01
Closed palm, extended index finger, and thumb finger	'L': 02
Closed palm	'Fist': 03
Fist vertical to sensor	'Fist moved': 04
Closed palm with extended thumb	'Thumb': 05
Closed palm with extended index	'Index': 06
Open palm with thumb and index drawing a circle	'OK': 07
Open palm vertical to the sensor	'Palm moved': 08
The semi-closed palm of a shape C	'C': 09
Open palm with separate fingers	'Palm down': 10

Table 3 demonstrates the outcomes of selected criteria that validate the effectiveness and accuracy of the proposed system. In order to optimal understand the system behavior regarding per hand gesture class recognition, the confusion matrix is shown in Figure 3. Even though the utilized hand gesture images are approximately similar and not easy to be distinguished, the main observation is that all the gestures have excellent scores.

Table 3. The outcomes of precision, sensitivity (recall), F1-score, and accuracy

Hand Gesture Classes	Precision	Sensitivity (Recall)	F1-Score	No. of Tested Images
'Palm'	1.00	1.00	1.00	595
'L'	1.00	1.00	1.00	604
'Fist'	1.00	1.00	1.00	627
'Fist moved'	1.00	1.00	1.00	596
'Thumb'	1.00	1.00	1.00	621
'Index'	1.00	1.00	1.00	574
'OK'	1.00	1.00	1.00	621
'Palm moved'	1.00	1.00	1.00	578
'C'	1.00	1.00	1.00	617
'Palm down'	1.00	1.00	1.00	567
Accuracy			1.00	6000
Macro Average	1.00	1.00	1.00	6000
Weighted Average	1.00	1.00	1.00	6000

The experiments are accomplished via specifying different numbers of training epochs (from epochs 0 to 20) to obtain excellent results of accuracies. Figure 4 illustrates the increase of validation accuracy with the final result of 100%, while Figure 5 illustrates the decrease of validation loss with the final result of 0.0021. The obtained results of the proposed system and the previously indicated hand gesture recognition models using different datasets are summarized in Table 4. It is noticeable that the proposed system outperformed and the achieved accuracy is 100%. While working on an infrared hand gesture dataset of 38425 images, the accuracy of the recognition model in [21] is 92%. Based on these obtained accuracies, we notice that our proposed CNN framework is effective and achieved excellent results while using 20000 images.

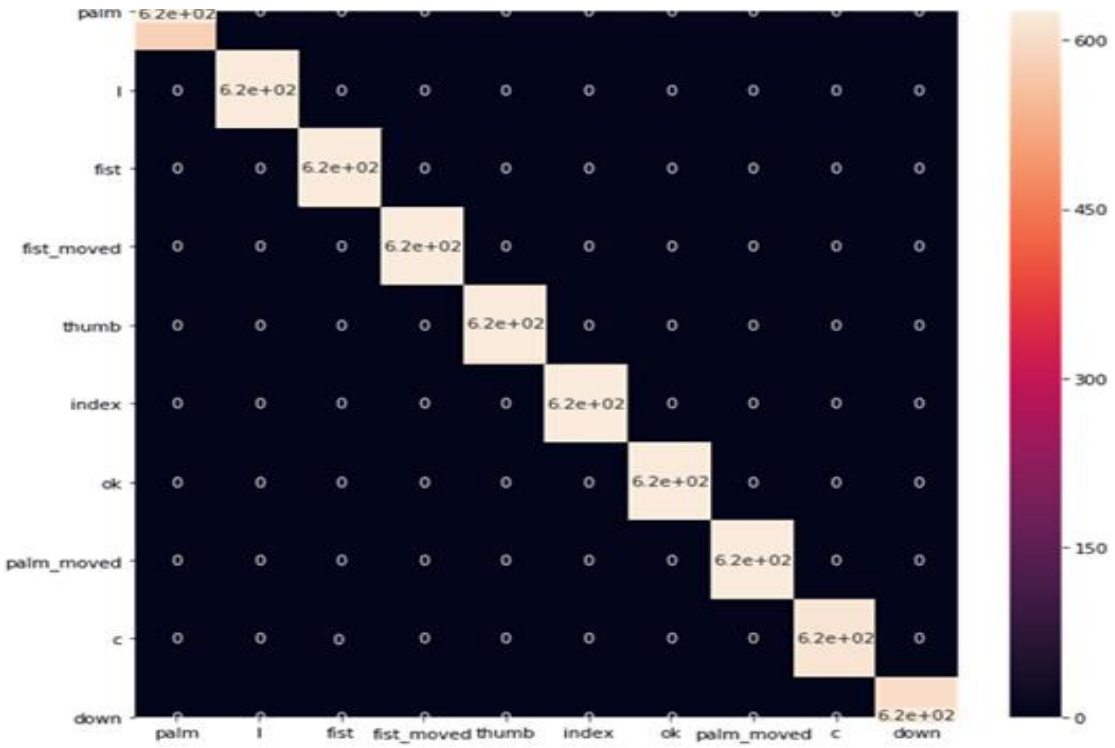


Figure 3. The confusion matrix of hand gesture images classification results

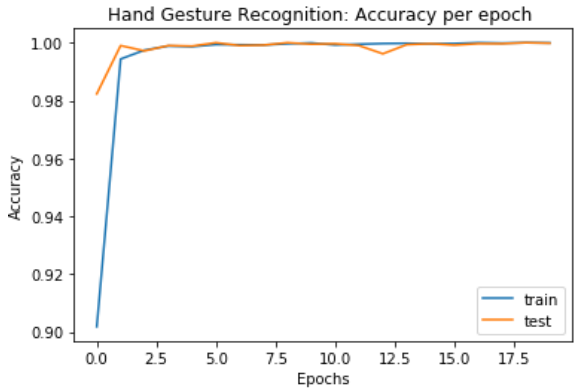


Figure 4. Accuracy validation change against training epochs (20-Epoch)

Figure 5. Loss validation change against training epochs (20-Epoch)

Table 4. A comparison between the proposed CNN-based recognition system and the recently existing related works

Author/(s) name, Ref. No., Year	The utilized dataset	Accuracy
S. Hussain *et al*. [18], 2017	Self-acquired dataset (more than 55000 images)	93.09%
Ankit Chaudhary and J. L. Raheja [19], 2018	Dataset acquired from different sources (14 different images for six kinds of gestures)	92.86%
J. P. Sahoo *et al*. [20], 2019	ASL dataset (36 gestures from 5 subjects)	87.83%
Jixin Wang *et al*. [21], 2020	Infrared hand gesture dataset (38425 images for two kinds of gestures)	92%
Tao Song *et al*. [22], 2021	CGD dataset (50000 images for 249 kinds of gestures)	84.32%
The proposed system	Hand-based near-infrared dataset (20000 images for ten kinds of gestures)	100%

5. CONCLUSION

In this paper, an accurate and effective deep learning framework is proposed for recognizing static hand gestures based on CNN. This framework includes two main phases; feature extraction and classification. These phases include multiple layers, each of which was designed to obtain the best results for human hand gesture recognition. The results of extensive experiments demonstrate that the proposed CNN framework of multiple layers achieves excellent performance results using a large-size database of hand-based near-infrared images. Also, it is significant to highlight dealing with infrared images to avoid the problem of low illumination. The comparison between the proposed system and other related works proved that the proposed system is more effective and accurate than others. In future work, this proposed CNN framework will be prepared to be utilized for recognizing dynamic gestures.

REFERENCES

[1] H. Tang, H. Liu, W. Xiao, and N. Sebe, "Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion," *Neurocomputing*, vol. 331, pp. 424-433, February 2019, doi: 10.1016/j.neucom.2018.11.038.

[2] G. Devineau, F. Moutarde, W. Xi, and J. Yang, "Deep Learning for Hand Gesture Recognition on Skeletal Data," *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 106-113, doi: 10.1109/FG.2018.00025.

[3] M. W. Cohen, N. B. Zikri, and A. Velkovich, "Recognition of Continuous Sign Language Alphabet Using Leap Motion Controller," *2018 11th International Conference on Human System Interaction (HSI)*, 2018, pp. 193-199, doi: 10.1109/HSI.2018.8430860.

[4] P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision-based hand gesture recognition: A review," *Computer Vision and Image Understanding*, vol. 141, pp. 152-165, December 2015, doi: 10.1016/j.cviu.2015.08.004.

[5] Y. Li, X. Wang, W. Liu, and B. Feng, "Deep attention network for joint hand gesture localization and recognition using static RGB-D images," *Information Sciences*, vol. 441, pp. 66-78, May 2018, doi: 10.1016/j.ins.2018.02.024.

[6] S. Riofrío, D. Pozo, J. Rosero, and J. Vásquez, "Gesture Recognition Using Dynamic Time Warping and Kinect: A Practical Approach," *2017 International Conference on Information Systems and Computer Science (INCISCOS)*, 2017, pp. 302-308, doi: 10.1109/INCISCOS.2017.36.

[7] F. A. Raheem and A. W. A. A. Hussain, ''Deep Learning Convolution Neural Networks Analysis and Comparative Study for Static Alphabet ASL Hand Gesture Recognition,'' *Journal of Xidian University*, vol. 14, no. 4, pp. 1871-1881, 2020, doi: 10.37896/jxu14.4/212.

[8] M. Oudah, A. Al-Naji, and J. Chahl, "Hand Gesture Recognition Based on Computer Vision: A Review of Techniques," *Journal of Imaging*, vol. 6, no. 8, p. 73, July 2020, doi: 10.3390/jimaging6080073.

[9] O. K. Oyedotun and A. Khashman, "Deep learning in vision based static hand gesture recognition," *Neural Computing and Applications*, vol. 28, no. 12, pp. 3941-3951, December 2017, doi: 10.1007/s00521-016-2294-8.

[10] D. Satybaldina and G. Kalymova, "Deep learning based static hand gesture recognition," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 1, pp. 398-405, January 2021, doi: 10.11591/ijeecs.v21.i1.pp398-405.

[11] S. Jiang, Q. Gao, H. Liu, and P. B. Shull, "A novel, co-located EMG-FMG-sensing wearable armband for hand gesture recognition," *Sensors and Actuators: A. Physical*, vol. 301, p. 111738, January 2020, doi: 10.1016/j.sna.2019.111738.

[12] S. F. Chevtchenko, R. F. Vale, V. Macario, and F. R. Cordeiro, "A convolutional neural network with feature fusion for realtime hand posture recognition," *Applied Soft Computing*, vol. 73, pp. 748-766, December 2018, doi: 10.1016/j.asoc.2018.09.010.

[13] V. Ranga, N. Yadav, and P. Garg, "American sign language fingerspelling using hybrid discrete wavelet transform-gabor filter and convolutional neural network," *Journal of Engineering Science and Technology*, vol. 13, no. 9, pp. 2655-2669, 2018.

[14] Z. Hu, Y. Hu, J. Liu, B. Wu, D. Han, and T. Kurfess, "3D separable convolutional neural network for dynamic hand gesture recognition," *Neurocomputing*, vol. 318, pp. 151-161, November 2018, doi: 10.1016/j.neucom.2018.08.042.

[15] A. Hassan and A. Mahmood, "Convolutional Recurrent Deep Learning Model for Sentence Classification," in *IEEE Access*, vol. 6, pp. 13949-13957, 2018, doi: 10.1109/ACCESS.2018.2814818.
[16] A. A. Abdulhussein and F. Raheem, ‘‘Hand Gesture Recognition of Static Letters American Sign Language (ASL) using Deep Learning,’’ *Engineering and Technology Journal*, vol. 38, no. 6A, pp. 926-937, 2020, doi: 10.30684/etj.v38i6A.533.
[17] O. Mazhar, S. Ramdani, and A. Cherubini, ‘‘A Deep Learning Framework for Recognizing Both Static and Dynamic Gestures,’’ *Sensors*, vol. 21, no. 6, p. 227, March 2021, doi: 10.3390/s21062227.
[18] S. Hussain, R. Saxena, X. Han, J. A. Khan, and H. Shin, "Hand gesture recognition using deep learning," *2017 International SoC Design Conference (ISOCC)*, 2017, pp. 48-49, doi: 10.1109/ISOCC.2017.8368821.
[19] A. Chaudhary and J. L. Raheja, "Light invariant real-time robust hand gesture recognition," *Optik*, vol. 159, pp. 283-294, April 2018, doi: 10.1016/j.ijleo.2017.11.158.
[20] J. P. Sahoo, S. Ari, and S. K. Patra, "Hand Gesture Recognition Using PCA Based Deep CNN Reduced Features and SVM Classifier," *2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS)*, 2019, pp. 221-224, doi: 10.1109/iSES47678.2019.00056.
[21] J. Wang, T. Liu, and X. Wang, "Human hand gesture recognition with convolutional neural networks for K-12 double-teachers instruction mode classroom," *Infrared Physics & Technology*, vol. 111, p. 103464, December 2020, doi: 10.1016/j.infrared.2020.103464.
[22] T. Song, H. Zhao, Z. Liu, H. Liu, Y. Hu, and D. Sun, "Intelligent human hand gesture recognition by local-global fusing quality-aware features," *Future Generation Computer Systems*, vol. 115, pp. 298-303, February 2021, doi: 10.1016/j.future.2020.09.013.
[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016, pp. 2818-2826.
[24] K. He, X. Z. S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016, pp. 770-778.
[25] ‘‘Hand Gesture Recognition Database’,’ *Acquired by Leap Motion*. Data accessed 1/2/2021 [Online]. Avaliable at: https://www.kaggle.com/gti-upm/leapgestrecog.