Hand Gesture Recognition Using CNN for Post-Stroke People

Norah Alnaim[1,2*], Abdullrahman Albar[1] Maysam Abbod[1]

[1] *Department of Computer Science/College of Science and Humanities in Jubail, Imam Abdulrahman Bin Faisal University, Saudi Arabia*
[2] *Department of Electronic and Computer Engineering, College of Engineering, Design and Physical Sciences Brunel University London, London, UK*
* *Norah.Alnaim@brunel.ac.uk*

Abstract – A human gesture is a non-verbal form of communication and is critical in human-robot interactions. Vision-based gesture recognition methods play a key role to detect hand motion and support such interactions. Hand gesture recognition allows a convenient and usable interface between devices and users. Hand gestures can be used for various fields which makes it be able to be implemented for communication and further. Hand gesture recognition is not only useful for people who are hearing-impaired or disabled but also for the people who experienced a stroke, as they need to communicate with other people using different common essential gestures such as the sign of eating, drink, family and, more. In this paper, an approach for recognizing hand gesture based on Convolutional Neural Network (CNN) is proposed. The developed method is evaluated and compared between training and testing modes based on several metrics such as execution time, accuracy, sensitivity, specificity, positive and negative predictive value, likelihood and root mean square. Results show that testing accuracy is 99% using CNN and is an effective technique in extracting distinct features and classifying data.

Keywords – Convolutional Neural Network, Deep Learning, Hand Gesture Recognition, Gesture Recognition, Stroke

I. Introduction

Recently, direct contact is the predominant form of communication between the user and the machine. The communication channel is based on devices such as a mouse, keyboard, remote control, touch screen, and other direct contact methods. Human to human communication is accomplished through more natural and intuitive non-contact methods, for example, sound and physical movements. The flexibility and efficiency of these non-contact communication methods have led many researchers to consider employing them to support human-computer interaction. The gesture is an important non-contact human communication method which forms a substantial part of the human language. Historically, wearable data gloves were regularly used to capture the angles and positions of each joint in the user's gesture. The difficulty and cost of a wearable sensor have restricted the widespread use of such a method. Gesture recognition can be defined as the ability of a computer to understand the gestures and perform certain commands based on those gestures. The main goal of Gesture recognition is to develop a system that can identify and understand specific gestures and communicates information from them [1].

Gesture recognition methods based on the non-contact visual inspection are currently popular. This is due to their low cost and convenience to the user. A hand gesture is an expressive communication method used in healthcare, entertainment and education industry, in addition to assisting users with special needs and the elderly. Hand tracking is vital to perform hand gesture recognition, involves undertaking various computer vision operations including hand segmentation, detection, and tracking.

Sign language uses hand gestures to convey feelings or information within the hearing impairment communication. The main problem is that an ordinary person would easily misunderstand the meaning conveyed. The advancement in AI and computer vision can be adapted to recognize and learn the sign language [2]. The modern systems can help an ordinary person to recognize and understand the sign language. This article presents a method which is related to the recognition of hand gestures using deep learning.

Stroke is a disease that affects arteries leading to and within the brain. Stroke is the fifth leading death cause as well as a cause of disability. A stroke occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot or bursts. Certain security measures keep the privacy implemented and protects an important part of the profile. This information has gathered people to demand skilled and capable information. Networks have a medical diagnosis system to allow the users in the expertise and experiences of groups and individual. This project shows that hand gesture is a very beneficial way to convey information and a very rich set of feelings and facts can be interpreted from gestures.

The rest of the paper is structured as follows: The material and method used in this paper found in Section II. A literature review of hand gesture detection techniques and methods used are shown in Section III. Theoretical concept of CNN is provided in Section IV. Section V concentrates on the details of the proposed system's implementation. Section VI describes the discussion and presentation of the results obtained. The conclusion and future work are discussed in section VII.

II. Materials and Method

The objective of this study is to present the effectiveness of CNN technique to extract features and classify various images. In this study, CNN method is evaluated and compared between training and testing. A hand gesture recognition system was developed based on deep learning method. The performance of hand gesture recognition method was evaluated and compared using several factors like execution time, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood, negative likelihood and root mean square.

III. Literature Review

Convolution Neural Networks (CNNs) are used to evaluate hand gesture recognition, where depth-based hand data was employed with CNN to obtain successful training and testing results [2]. Another CNN method was proposed [3] that uses a skin model, hand position calibration and orientation to train and test the CNN. The authors proposed a gesture recognition system using Micro Electro-Mechanical System (MEMS) accelerometer and consists of a microcontroller, ADXL335 accelerometer and a display unit with speaker. The reason stated for choosing MEMS accelerometer is that it is easy to wear, and no special training is needed to be given to person wearing it [1]. A hand gesture recognition sensor is introduced in study uses ultra-wide band impulse signals. Each gesture has their own reflected waveform and CNN is used for gesture classification. Six gestures from ASL (American Sign Language) have been used for the experiment. The results show 90% accuracy using CNN and proves its effectiveness recognising the gesture [4].

A Pattern Recognition model is proposed in article [5] for dynamic hand gesture recognition which combines CNN with weighted fuzzy min-max neural network. The model also performs feature extraction, feature analysis and spatiotemporal template data representation based on the motion information of target is designed. The efficiency of classifier is increased by performing feature analysis technique using weighted fuzzy min-max neural network. The results show that influence caused by feature point's spatial and temporal variation can be reduced using the proposed implementation.

In this study, the authors present their study in process and methods related to sign language recognition using Deep Learning. 3D CNN was used for recognising images received through Kinect sensor. The method using 3D CNN was found the be very effective and the highest accuracy was found to be 91.23% [6]. CNN are used to recognise Indian sign language gestures. The capture method used was Selfie mode continuous sign language video, where a hearing-impaired person would operate the sign language recogniser mobile independently. The datasets were not available for mobile use, hence, the authors created datasets with five subjects which performed 200 signs in 5 different viewing angles under several background environments. Various CNN architectures were designed and tested and 92.88% was the best recognition rate obtained on the dataset [7].

Multi-class SVM and k-NN classifier are used to monitor seven gestures for residential rehabilitation of the post-stroke patients. The gestures were performed on seventeen young subjects. The results were assessed using k-fold cross validation method. Results shows that multi-class SVM and k-NN classifier achieved an accuracy of 97.29% and 97.71%, respectively [8].

Authors used webcam to track region of interest (ROI) which is hand region gestures. The kernelized correlation filters (KCF) algorithm is used to track the detected ROI. The image is resized and input to deep CNN to identify multiple hand gestures. Two deep CNN architectures modified from AlexNet and VGGNet, respectively are implemented. This process of tracking is continuously repeated, and gestures are recognised until the hand leaves camera range. The training data set reached recognition rate of 99.90%, and the test data set got recognition rate of 95.61% [8].

CNN and back propagation methodologies are used to recognise gestures to help disables. The machine is able to understand the images and identify what the images are that is really helpful in many ways [10]. In this study, Adapted Deep Convolutional Neural Network (ADCNN) is proposed to recognise the hand gestures. Data augmentation is applied to increase the size of dataset and increase robustness of deep learning. The images are input into ADCNN in presence of RELU and Softmax, L2 regularization is used to eliminate overfitting. This method has been proved to be efficient to recognise hand gestures. The model is first trained using 3750 images with several variations in features like rotation, translation, scale, illumination and noise. Compared to baseline CNN, ADCNN had a accuracy of 99.73%, and a 4% improvement over the baseline CNN model (95.73%) [11].

IV. Theory

Artificial Intelligence is bridging the gap between capabilities of humans and machines. Researchers are working on several fields to make great things happen. One such field is Computer Vision. The goal of computer vision is to make machines view the world as humans and perceive it the same way as humans and use this knowledge for amazing tasks such as Image recognition, Image analysis and classification, video recognition, Media recreation, natural language processing, etc. One such algorithm that has played a major role in advancements of computer vision and deep learning is Convolutional Neural Network (CNN).

CNN is defined as a multi-layer neural network with a unique architecture used for deep learning [12]. A CNN architecture is comprised of three essential layers: Convolutional Layer, Pooling Layer, and Fully-Connected Layer. CNN is generally used in recognizing objects, scenes, and carrying out image detection, extraction and segmentation. CNN has been significantly used in the last few years ago due to the following three aspects: (1) the necessity for feature extraction by using image processing tools is removed since CNN can directly learn the image data, (2) Exceptionally good for recognition of results and can be easily re-trained for new recognition purposes, and (3) CNN can be built on the pre-existing network [12].

The convolutional layer does the heavy computational tasks and is the core building block of CNN [13]. The convolutional layer's parameters contain a set of learnable filters. We pass each filter across the height and width of the input volume and calculate the dot product between filter's entries and input at all the positions during the forward pass. When the filter is slide across the input volume, a 2D activation map that gives responses of the filter at every spatial position is produced.

Now we have each set of filters in each convolutional layer. Each layer will produce a separate activation map and these maps are stacked with depth dimensions produce the output volume. Three hyperparameters that control the size of output volume are depth, stride, zero-padding.

Depth corresponds to the number of filters the user wants to use. Each filter will look to learn something different in the input. Stride refers to the stride with which we slide filter. If stride is given 1 then filters tend to move one pixel at a time, when it is 2 they jump two pixels at a time. Sometimes it is convenient to pad input volume with zeros around border. This is nothing but zero-padding hyper parameter. Zero padding allows controlling the spatial size of output volumes.

The formula for calculating the number of neurons "fit" is **(W-F+2P)/S+1**, where W = input volume size, F = receptive field size of the Convolution Layer neurons, S = stride with which they are applied, and P = amount of zero padding used on the border. Let's take an example, 7x7 input and a 3x3 filter with stride 1 and pad 0 we would get a 5x5 output.

Pooling Layer is a common practice to have a pooling layer between convolution layers in a convolutional neural network architecture. The main task of the pooling layer is to reduce the spatial size to reduce hyperparameters and in turn the computation in network. This also controls the overfitting problem. These layers act independently on different each depth slice of the input and resizes them spatially using MAX operation.

For Pooling layers, it is not common to pad the input using zero-padding. It is worth knowing that pooling units can also perform functions like such as average pooling or L2-norm pooling. But max-pooling has been proven to work best in practice.

In fully connected layers, the Neurons have full connections to all activations in the previous layer. The activations can be computed with a matrix multiplication followed by a bias offset. Converting a fully connected layer to convolutional layers stage; It is worth noting that the main difference between Fully connected and convolutional layers is that the neurons in Convolutional layer are connected only to a local region in the input, and many of the neurons in a convolutional volume share parameters. The neurons in both layers still compute dot products making their functional form is identical. Hence, it is possible to convert between fully connected and convolutional layers.

We have seen that a convolutional neural network consists of three layers: convolutional, pooling and fully connected layer. In practice, we also write the RELU activation function as a layer. Common practice stacks a few convolutional-RELU layers, followed by pooling layers, and repeat this pattern until the image has been merged spatially to a small size. The last fully-connected layer holds the output. The common CNN architecture follows the pattern [14]:

INPUT -> [[CONV -> RELU]*N -> POOL?]*M -> [FC -> RELU]*K -> FC

where the * indicates repetition, and POOL? indicates an optional pooling layer. Also, N >= 0 (and N <= 3), M >= 0, K >= 0 (and K < 3). Figure 1 shows a simple CNN architecture which takes m*n*1 size input. This input is passed through combination of several layers like Convolutional, Pooling and ReLu before it reaches fully connected layer and finally the gesture in the image is recognised in the output layer.

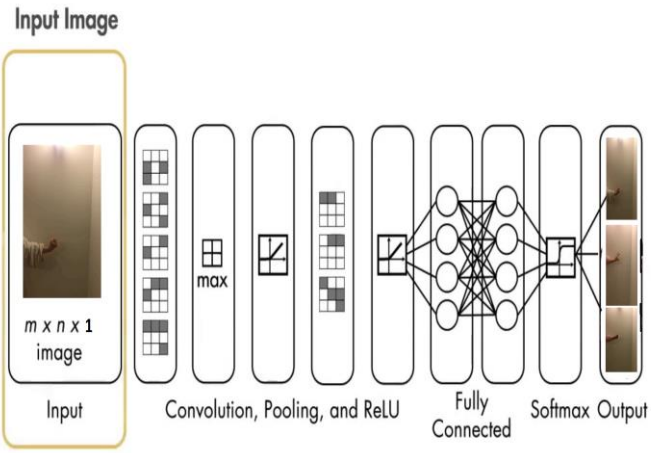
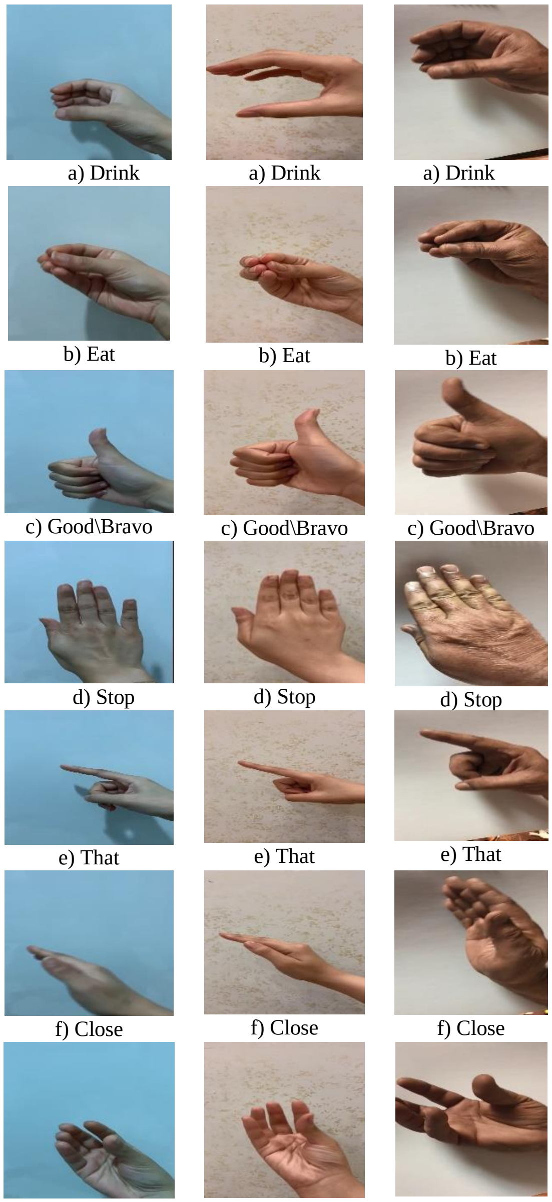


Fig. 1 Convolutional Neural Networks architecture

V. Implementation

A. Hand Gestures Input

Hand gestures represent a hundred and forty gestures is composed of seven gestures for twenty people as an input to the gesture recognition method evaluated and compared in this study. Figure 2 illustrates three examples of twenty people showing seven 2-D and 3-D universal common hand gestures with three different mobile cameras, backgrounds, illumination, the position of the hand and the shape of the hand. The mobile camera used to record the first gesture are iPhone8 and Samsung Galaxy S10 is used to record second gesture and the last gesture is recorded using iPhone8. The first background is light blue, the second background is lightly floral and the last one is plain. The illumination of the first example is less than the second and third example. The position of the hand is also slightly different as well as the shape of the hands. A first-hand gesture is for a young woman in the late of twenty, another hand gesture is for a young woman in the mid of thirty, the last one is for an old man in the mid of seventy. They are recorded within short distances and used in the study's experimental work. The hand gestures signs are referenced from Simple hand sign communication cards used by Single hand communications and are shown in Figure 3.



a) Drink a) Drink a) Drink

b) Eat b) Eat b) Eat

c) Good\Bravo c) Good\Bravo c) Good\Bravo

d) Stop d) Stop d) Stop

e) That e) That e) That

f) Close f) Close f) Close

g) Family g) Family g) Family

Fig. 2 Three examples of seven universal hand gestures for three different hands

Fig. 3 Simple hand signs cards

The framework model shown in Figure 4 illustrated the system implementation steps. Using different high standard mobile cameras with two resolutions which are HD and 4k, the hand motions shown in Figure 3 are recorded. Each recording lasts from 2 to 10 seconds and the resolution of the recorded video is vary.

B. Computing Platform Specification

The experiment was performed using a Dell desktop C2544404 with processor Generation Intel ® Core™ i7-6700 CPU @ 3.40 GHz, memory type DDR4 16 GB, speed 2133 MHz, 512 GB storage hard drive, 24 inch (60.9 cm) Full HD (1920 × 1200) Widescreen Ultrasharp IPS Panel display w/Adobe RGB colour space and touch. Windows 10 (64 bits) operating system was used and the system is implemented using MATLAB R20187bV language.

C. Convolutional Neural Network Implementation

CNN forms an integral part of deep learning as is used to train data without using any image processing tool. In this experiment, a new directory is created for each video. Hundred and forty videos are read to generate 24,698 image frames. The method to transfer the image frame from RGB to grey and resize it to 227×227 from the original image size is shown in Figure 5. Each video has a different number of frames between 3394 to 3670 frames. The image's data is split into training and testing datasets. The number of training frames is 2485 which is 70%. The CNN topology is created in seven layers with each layer having the following functionality and size: ImageInputLayer Input size [227,227,1], Convolution2-DLayer Filter size [5,20], ReLULayer (Rectified Linear Unit), MaxPooling2-DLayer Pool size [2,2], FullyConnectedLayer Input size [auto] and Output size [7], SoftmaxLayer and ClassificationOutputLayer Output size [auto]. The hyperparameters of CNN are generated inside the training options function. The value of epochs parameter is set to 50 epochs.

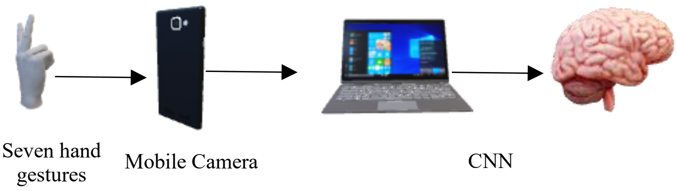
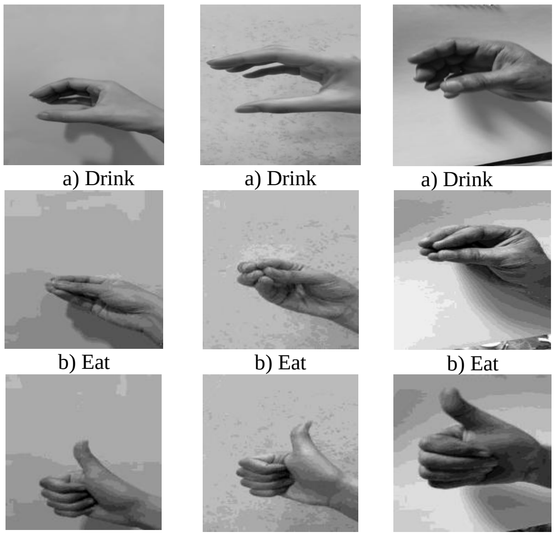


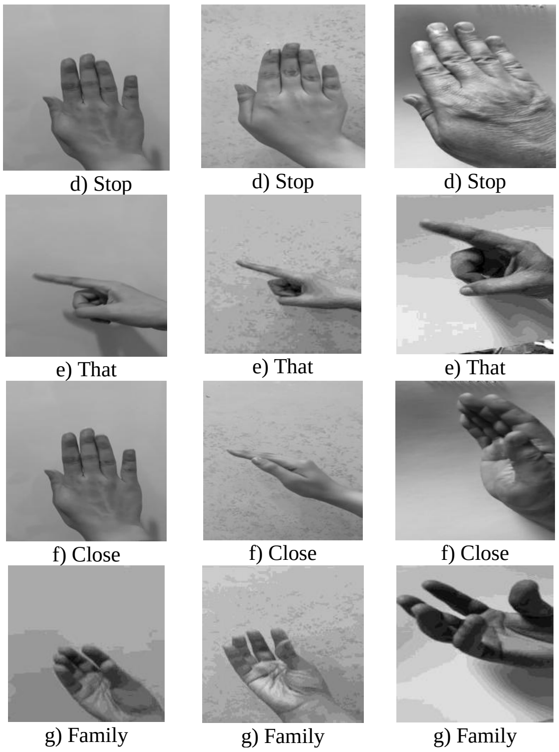
Fig. 4 framework model of system implementation



a) Drink a) Drink a) Drink

b) Eat b) Eat b) Eat

c) Good\Bravo c) Good\Bravo c) Good\Bravo



d) Stop d) Stop d) Stop

e) That e) That e) That

f) Close f) Close f) Close

g) Family g) Family g) Family

Fig. 5 Three examples for seven universal common hand gestures for three different hands

D. Parameters Comparison

The performance of CNN algorithm is compared between training and testing using several parameters including execution time, that is the duration taken by the software to process the given task. Sensitivity measures the percentage of positives which are properly identified. Specificity is a measure of the false positive rate. The PPV and NPV are the percentages of positive and negative results in diagnostic and statistics tests which also describe the true positive and true negative results. The LR+ and LR− are known measures in diagnostic accuracy.

VI. Results

The experiments were executed ten times to obtain the mean of seven-hand gestures. Two different training and testing modes were presented and compared to find the best result. Training accuracy is achieved by implementing a model on the training data and determining the accuracy of the algorithm. A summary of the values obtained for various parameters in training and testing approach is listed in Table 1. It can be noticed that the execution time of training and testing is equalled. The accuracy result of training is 100% compared to that of testing. The value of sensitivity in training is a bit higher than testing. Specificity in training is 100% whereas in testing is 0.9989. The PPV and NPV of testing is lower than training. The best value for LR+ and LR− are recorded for training. For RMS, the value of training and testing are matched. The parameter values of training in CNN are constant for all categories. Its execution time is approximate 15,598 seconds, which is duration to train and test the system using seven hand gestures which are used in the experiment. Overall, training has the best values in most parameters.

Table I. CNN Training and testing results

Factors	**CNN**	
	Training	**Testing**
Exe Time ± SD (sec)	15,598± 244.9784	15,598± 244.9784
Accuracy ± SD	1± 0	0.9912± 0.0086
Sensitivity ± SD	1± 0	0.9934±0.0042
Specificity ± SD	1± 0	0.9989±0.0023
Positive Predictive Value (PPV) ± SD	1± 0	0.9934±0.0040
Negative Predictive Value (NPV) ± SD	1± 0	0.9989±0.0021
Positive Likelihood (LR+)± SD	1± 0	884.4175±37.5328
Negative Likelihood (LR−)± SD	1± 0	0.0066±1.7920
RMS ± SD	1± 0	1± 0

VII. Conclusions and Future Work

Hand gesture detection is fundamental to provide a natural HCI skill. It is now known that in gesture recognition, the most essential aspects are detection, segmentation and tracking. In this experiment, a system has been created for hand gestures recognition using features extraction and classification in CNN technique. Seven 2-D and 3-D motions with different mobile cameras, backgrounds, illumination, position of hand and the shape of hand are recorded within short distances. Experiments were performed to compare the performance of training and testing in CNN method. Results showed that training provides better accuracy compared to testing. In future work, the number of gestures will be extended to ten common gestures using 3-D Holoscopic imaging technique camera.

Acknowledgment

The first author would like to show her gratitude to the financial support received from King Faisal University.

References

[1] V. Vardhan1 and P. Prasad, “Hand Gesture Recognition Application for Physically Disabled People,” International Journal of Science and Research, vol. 3, no. 8, pp. 765–796, Aug. 2014.

[2] N. Soodtoetong and E. Gedkhaw, "The Efficiency of Sign Language Recognition using 3D Convolutional Neural Networks," 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Rai, Thailand, 2018, pp. 70-73.

[3] J. Pyo, S. Ji, S. You and T. Kuc, "Depth-based Hand Recognition using Convolutional Neural Networks," 13th International Conference on Ubiquitous Robots and Ambient Intelligence, pp. 225-227, Xi'an, China, 2016.

[4] H. Lin, M. Hsu and W. Chen, "Human Hand Gesture Recognition using a Convolution Neural Network," IEEE International Conference on Automation Science and Engineering (CASE), pp. 1038-1043, Taipei, Taiwan 2014.

[5] S. Y. Kim, H. G. Han, J. W. Kim, S. Lee and T. W. Kim, "A Hand Gesture Recognition Sensor Using Reflected Impulses," in IEEE Sensors Journal, vol. 17, no. 10, pp. 2975-2976, 15 May15, 2017. doi: 10.1109/JSEN.2017.2679220

[6] Ho-Joon Kim, J. S. Lee and J. Park, "Dynamic hand gesture recognition using a CNN model with 3D receptive fields," 2008 International Conference on Neural Networks and Signal Processing, Nanjing, 2008, pp. 14-19. doi: 10.1109/ICNNSP.2008.4590300

[7] G. A. Rao, K. Syamala, P. V. V. Kishore and A. S. C. S. Sastry, "Deep convolutional neural networks for sign language recognition," 2018 Conference on Signal Processing And Communication Engineering Systems (SPACES), Vijayawada, 2018, pp. 194-197. doi: 10.1109/SPACES.2018.8316344

[8] W. Li, C. Hsieh, L. Lin and W. Chu, "Hand gesture recognition for post-stroke rehabilitation using leap motion," 2017 International Conference on Applied System Innovation (ICASI), Sapporo, 2017, pp. 386-388. doi: 10.1109/ICASI.2017.7988433

[9] H. Chung, Y. Chung and W. Tsai, "An Efficient Hand Gesture Recognition System Based on Deep CNN," 2019 IEEE International Conference on Industrial Technology (ICIT), Melbourne, Australia, 2019, pp. 853-858. doi: 10.1109/ICIT.2019.8755038

[10] [A8] K. S. Varun, I. Puneeth and T. P. Jacob, "Hand Gesture Recognition and Implementation for Disables using CNN'S," 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, pp. 0592-0595. doi: 10.1109/ICCSP.2019.8697980

[11] A. A. Alani, G. Cosma, A. Taherkhani and T. M. McGinnity, "Hand gesture recognition using an adapted convolutional neural network with data augmentation," 2018 4th International Conference on Information Management (ICIM), Oxford, 2018, pp. 5-12. doi: 10.1109/INFOMAN.2018.8392660

[12] S. Pang, J. del Coz, Z. Yu, O. Luaces and J. Díez, "Deep Learning to Frame Objects for Visual Target Tracking," *Engineering Applications of Artificial Intelligence*, vol 65, pp. 406-420, 2017.

[13] Simonyan, Karen and Zisserman, Andrew, "Two-Stream Convolutional Networks for Action Recognition in Videos," Advances in Neural Information Processing Systems - editors Z. Ghahramani and M. Welling and C. Cortes and N. D. Lawrence and K. Q. Weinberger, pp 568--576,

[14] "CS231n Convolutional Neural Networks for Visual Recognition," github. [Online]. Available: http://cs231n.github.io/convolutional-networks/. [Accessed: 20-Sep-2019].