Computer Vision for Elderly Care Based on Deep Learning CNN and SVM

M Oudah¹, A Al-Naji¹,²,³ and J Chahl²

¹Electrical Engineering Technical College, Middle Technical University, Baghdad 10022, Iraq; Munir_aliraqi@yahoo.com
²School of Engineering, University of South Australia, Mawson Lakes SA 5095, Australia
³Correspondence: ali.al-naji@unisa.edu.au; Tel.: +9647710304768

Abstract: Computer vision has wide application in medical sciences such as health care and home automation. This study on computer vision for elderly care is based on a Microsoft Kinect sensor considers an inexpensive, three dimensional, non-contact technique, that is comfortable for patients while being highly reliable and suitable for long term monitoring. This paper proposes a hand gesture system for elderly health care based on deep learning convolutional neural network (CNN) that is used to extract features and to classify five gestures according to five categories using a support vector machine (SVM). The proposed system is beneficial for elderly patients who are voiceless or deaf-mute and unable to communicate with others. Each gesture indicates a specific request such as "Water", "Meal", "Toilet", "Help" and "Medicine" and translates as a command sending to a Microcontroller circuit that sends the request to the caregiver’s mobile phone via the global system for mobile communication (GSM). The system was tested in an indoor environment and provides reliable outcomes and a useful interface for older people with disabilities in their limbs to communicate with their families and caregivers.

Keywords: Hand gestures; deep learning; convolutional neural network (CNN); support vector machine (SVM); Microsoft Kinect sensor; global system for mobile communication (GSM).

1. Introduction

Home care is often preferable to patients and is usually less expensive for caregivers than hospitals. There are some ways for those people to communicate with others using specific gestures, such as body language, hand gestures, arm gestures and facial expressions. Hand gestures are a popular interaction technique that can be classified into three categories: static hand gestures by performing a special sign or pose using one hand[1], dynamic hand gestures by performing a special sign or pose using moving hands[2], and hybrids of both static and dynamic hand gestures which considers more complex gestures[3]. A wide range of research papers have been published about hand gesture based techniques. Previously, glove attached sensor were utilized to provide reliable interfacing to human-computer interaction (HCI) system[4]. However, this technique uses an expensive sensor, sometimes needs calibration and may be considered uncomfortable in cases requiring long-term monitoring.

Computer vision may offer a high performance and cost effective technique that overcomes the limitations related to glove sensors by providing natural interaction at a distance. Computer vision can use videos and images to identify a real-world scene using image processing techniques that can identify objects based on size, colour or location[5] [6]. In addition, tracking objects, estimating distance of objects from the camera, detecting and classifying objects in images and videos using techniques such as image subtraction or using algorithms, such as optical flow that is taking into account groups of pixels rather than individual pixels.

One advantage of computer vision systems based on deep learning is to detect and extract features in a cluttered scene for learning and matching processes. Many other proposed techniques successfully use hand gestures in applications such as leap motion[7], radar sensor[8], and time-of-flight camera...
However, these techniques have faced different challenges, including algorithms used for segmentation and recognition, number of performed gestures, gestures performed using a single hand or both hands, distance from the camera. An extensive review of these techniques with their advantages and disadvantages can be found in[3].

The main contribution of this paper is to investigate the feasibility of using hand gestures captured by the Microsoft Kinect sensor to help elderly people who have paraplegia and are voiceless, especially when they need to communicate with family members or caregivers for daily needs by sending their requests to the caregiver or family member mobile phone via SMS at any time.

The rest of this paper is arranged as follows: Section 2 presents the related works. Section 3 describes the materials and methods, including the participants and experimental setup, hardware design and software. Section 4 shows the experimental results and discusses the obtained outcomes. Finally, conclusion and future research directions are provided in Section 5.

2. Related Works

Artificial intelligence techniques based on deep learning are utilized in a wide range of modern applications. Deep learning uses multilayer neural networks for learning and requires large datasets for learning, validation and prediction. The dataset may be public or created by developers of the system according to the requirements of the system. The largest challenge facing this technique is the processing time for the learning process that depends on the number of layers and filters used. Once trained, however, the weights can be transferred to all similar systems. This section illustrates some of the preceding work with regard to deep learning techniques.

A study by Nunez et al. [9] proposed a hybrid approach based on combining deep learning with CNN and long-short term memory (LSTM) to build two stages for extensive learning and to detect 3D pose obtained from the entire body for hand skeletons and human body activity. Their approach was evaluated using six datasets and the best results were obtained for small datasets. The training time of the approach was between 2 and 3 hours using a GPU. This approach offers some limitation such as suitable to use with small datasets and need powerful computer to execute the task using GPU. Another study by Devineau et al. [10] introduced a 3D hand gesture recognition model based on deep learning using a parallel CNN. This model was used to process hand-skeletal joints positions by parallel convolutions and could achieve a classification accuracy of 91.28% for 14 gestures and 84.35% for the 28 gestures. Another study by Hussain et al. [11] presented a new approach based on 6-static and 8-dynamic hand gestures which linked with specific commands executed by a computer. The CNN based on classifier was trained through a transfer-learning process using pre-trained CNN model that trained initially on a large dataset to recognize hand shape and gestures type. The system was recorded accuracy of 93.09%.

John et al. [12] proposed a hand gesture recognition algorithm for intelligent vehicle applications based on long-recurrent convolutional neural network for gesture classification. The algorithm was achieved to improve classification accuracy and computational efficiency by extracting three representative frames from the video sequences using semantic segmentation-based deconvolutional neural network. Another study by Li et al. [13] proposed a hand gesture recognition system based on unsupervised learning CNN where the characteristics of CNN were used to avoid the feature extraction process and reduce the number of parameters that needed to be trained. The error back-propagation algorithm was used in the network to modify the threshold and weights of neural networks and to reduce the error of the whole model. The support vector machine classifier was added to the CNN to improve the classification function of the CNN. Nagi et al. [14] introduced a hand gesture for human-robot interaction (HRI) using deep learning-based combining CNN and max-pooling MPCNN for supervised feature learning and classification of hand gestures to mobile robots using colourd gloves.

Molchanov et al. [15] proposed a dynamic hand gesture recognition system with 3D CNN and fused motion volume of normalized depth and image gradient values. The system utilised spatiotemporal data augmentation to avoid overfitting and was tested with the VIVA dataset, where it achieved a classification rate of 77.5%. Another study by Oyedotun et al. [16] proposed a complex deep learning hand gesture recognition for the whole 24 american sign language (ASL) obtained from the Thomas Moeslund’s gesture recognition database based on CNN with lower error rates. Sun et al. [17] extracted hand gestures from a complicated background by establishing the skin colour model using AdaBoost.
classifier based on haar according to the particularity of skin colour and CamShift algorithm. The CNN was used in their study to recognise 10 common digits and showed 98.3% of accuracy. Finally, a new model was proposed by Xing et al. [18] based on deep learning and improved the accuracy of EMG-based hand gesture recognition by using parallel architecture with five convolution layers. The current study was proposed a hand gesture system for elderly health care based on deep learning CNN for feature extraction and used these features to classify five gestures according to five categories using a SVM.

3. Materials and Methods
This section explains the hardware used and participants to evaluate the performance of the proposed system where is described by the following subsections.

3.1. Participants and Experimental Setup
The experiment was evaluated with three elderly participants, including two males and one female between ages of 60 and 73 and one adult (34 years), and it was tested in the home environment and repeated at various times for approximately 0.5 hour for each of them. This study adhered to the Declaration of Helsinki ethical principles (Finland 1964) where written informed consent forms were obtained from all participants after a full explanation of the experimental procedures. The Kinect V2 sensor was installed at distance ranging from 0.5 to 4.5 m with an angle of 0° in front of the elderly patient. Only the depth sensor was used to capture video frames with a resolution of 512×424 and a frame rate of 30 fps. The Kinect sensor was connected to a Laptop with a conversion power adapter and had installed a standard development kit (Microsoft Kinect for Windows SDK 2.0).

3.2. Hardware Design
The practical circuit of the proposed system is shown in Figure 1. The hardware components, including the Microsoft Kinect sensor V2, Arduino Nano microcontroller, GSM module Sim800l and DC-DC step down chopper (buck).

![Figure 1. The practical circuit of the proposed system: (a) Top view; (b) Side view.](image)

3.2.1 Microsoft Kinect Sensor.
The Kinect sensor comes with a factory calibration between the depth camera and the RGB camera. This factory calibration provides the necessary parameters to undistort the depth and colour images and map depth pixels to colour pixels. Moreover, the Kinect sensor provides information to detect and track the location of the body joints of up to six people. The high-resolution RGB camera is 1920 x 1080 pixel and depth sensor resolution is 512 x 424 pixel, making it suitable for broad research fields[19] [20]. Figure 2 shows the outer structure of the Kinect V2.
3.2.2 Arduino Nano Microcontroller.
The Arduino microcontroller board type Nano is suitable to use in this work because it has specific characteristics, such as being small, globally available, well supported for development, inexpensive and has 14 digital I/O pins, 8 analogue pins with clock frequency of 16MHz[21]. The Nano microcontroller controls sending messages through GSM according to data acquired from the PC in real-time via a Mini USB serial cable as can be seen in Figure 1.

3.2.3 GSM Module Sim800l.
The GSM sim800l module is a cost-effective, small size (approximately 0.025 m²), and low power consumption device that can be used for executing tasks, such as making voice calls, sending GPS information, and sending messages. It operates at a voltage range between 3.4 V and 4.4 V and it is suitable for the study purpose to send 5 different messages controlled by the microcontroller using AT commands. The module is connected to the microcontroller via two digital ports with module pins (TX, RX), while the VCC and GND terminal of the module are connected with DC chopper LM2596 that supplies the module with the proper voltage of 3.7 volts as shown in Figure 1.

3.2.4 DC-DC Chopper (buck).
LM2596 dc-dc buck converter step-down power module with high-precision potentiometer for adjusting output voltage was used. This converter is capable of driving a load of up to 3A with high efficiency and an output voltage of less than or equal to the input voltage. It is used to reduce the 5V voltage supply to 3.7 V to be suitable for the GSM sim800l module[6]. Figure 1 shows the outer view of the buck converter.

3.3. Hand Gesture Classification Using CNN and SVM
This method utilised the CNN that was previously trained to extract image features and classify input images according to five categories. CNN is a widely applied machine-learning-tool from deep-learning. It is trained by using five images folders and every folder contains a large number of images related to specific gestures which help the CNN to efficiently train any gesture. The benefits of using pre-trained CNN is to reduce the processing time and effort. In this study, the images were classified according to its categories using a multi-class linear support-vector-machine (SVM) which is trained by using image features extracted via CNN. This method gives an accuracy rate close to 100% compared to other systems, such as HOG or SURF. This method includes machine learning, statistics, and a model for resnet-50 networks.

3.3.1. Image Data
The category-classifier was trained on images from the hand gesture data store folder created by the author using the Kinect sensor V2 to capture hand images that include five different hand gestures corresponding to 1 to 5. Figure 3 represents the five hand gestures used in this experiment.
3.3.2. Image Loading.
The hand gestures dataset folder includes five different categories: 1, 2, 3, 4, and 5. The SVM was trained to recognise these categories as follow:

1- Create an image data store which helps to store image categories and train the network. The image data store uses file location for reading images, where the image reading is only loaded into memory. This method is considered suitable for large image collections.

2- The image data store includes category labels related to all images, where the folder-name is used to automatically assign labels.

3- It must adjust the number of images per category to maintain the equilibrium of the trained network, where every category contains 130 images for training network and validation.

3.3.3. Load Pretrained Network.
There are many preceding training networks trained on images data, such as the "ResNet-50" model[22], AlexNet[23], GoogLeNet[24], VGG-16[25] and VGG-19[26]. Figure 4 shows the architecture of ResNet-50 that is composed of 50 layers.

The pre-trained networks can be summarized as follows:
1. While every CNN has its own dimensions, the first layer specifies the image dimension which utilised in this method is 224-by-224-by-3.
2. The middle layer of the CNN is larger than others because it is composed of series-convolutional-layers separated by rectified linear units (ReLU) and max-pooling layers[27].
3. The last layer gives an output of the classification where the CNN model is trained to solve and classify 1000 tasks. In this study, it was used for classifying image data from the hand gesture dataset that includes five different categories of hand gestures.

3.3.4. Preparing Training and Test Image Sets.
The data set is divided into training data and validation data. Where 30% of the image data set from every category is used for training and 70% for validation. The randomisation process was used to avoid the biasing results.

3.3.5. Pre-processing Images for CNN.
The augmented image data store was served to resize and change the format for every input binary image to non-format on the fly to avoid the resaving of all images in the category folders.

3.3.6. Extract Training Features Using CNN.
Each CNN layer gives a special response to the input image. A few-layers on CNN are suitable for image-feature-extraction. The first network layer extracts simple features, such as edge, blob using the network filter weights from the first CNN. The learned filter through the first CNN layer which captures blob and edge features, where these features are feed to the deeper layer to merge these features and form high-level image features which are suitable for the recognition task[28]. The activation method can easily extract features from deeper layers where the layer locates before classification in the network. It can work with GPU if available or depend only on the CPU. The output is arranged as a column to increase multiclass-linear SVM training for the classification task.

4. Experimental Results & Discussion
The experiment results of the proposed system were carried out by the Matlab R2019b (Image processing toolbox, Computer vision toolbox, Deep learning toolbox), Microsoft standard development kit (SDK) for Kinect V2 and Arduino (IDE) software. The outcomes of CNN used to train SVM-classifier where the stochastic-gradient-descent-solver was used for training SVM linearly. This served to increase the training process when working with a high-dimension CNN feature vector. Figure 5 shows the outcomes of the proposed system at five hand gestures based on CNN and SVM.

Based on the results of Figure 5, the analysis of five results of the total number of tested gestures per participant in terms of the number of recognized gestures, number of un-recognized gestures, the percentage of correct recognition and the percentage of fault recognition of the total number of sample gestures is shown in Table 1.
Figure 5. The experimental results of the proposed system at five hand gestures (a), (b), (c), (d) and (e).
Table 1. The analysis of five results of the total number of tested gestures per participant.

| Hand gesture type | Total number of sample per tested gesture | Number of recognized gestures | Number of unrecognized gestures | Percentage of correct recognition for the total number of sample gestures % | Percentage of fault recognition for the total number of sample gestures % |
|-------------------|------------------------------------------|------------------------------|-------------------------------|-------------------------------------------------|-------------------------------------------------|
| 1                 | 65                                       | 63                           | 2                             | 96.92                                           | 3.08                                             |
| 2                 | 65                                       | 61                           | 4                             | 93.85                                           | 6.15                                             |
| 3                 | 65                                       | 64                           | 1                             | 98.46                                           | 1.54                                             |
| 4                 | 65                                       | 62                           | 3                             | 95.38                                           | 4.62                                             |
| 5                 | 65                                       | 64                           | 1                             | 98.46                                           | 1.54                                             |
| Total             | 325                                      | 314                          | 11                            | -----                                            | -----                                            |

Figure 6 analyses the results of Table 1 based on the confusion matrix, where every type of gesture was tested 65 times. The observation of the overall recognition rate was 96.62%. As an example, the first hand gesture represents one hand gesture that was tested 65 times, which means the result of the tested gesture must indicate the desired result. In addition, Figure 6 gives 63 correct recognitions and two incorrect recognitions which falls at number two gesture. Also, Figure 6 shows the confusion matrix of the five results.

![Confusion matrix](image)

Figure 6 Confusion matrixes of the obtained results.

Although the proposed system succeeded to recognise the five different hand gestures and provided a useful interface for older people to communicate with their families and caregiver via SMS messages. It has some challenges that need to be considered in future work. These challenges can be summarised as follows:

- The system works at real-time and gives a good recognition rate. Moreover, it takes time initially during network training because the ResNet-50 is composed of 50 layers for deep learning.
- The experiment provides a good recognition rate but suffered due to distance limitations related to the range sensor used when the dataset was created.
- The hand segment based on joint tracking was provided by the Kinect V2 is cropped and tracked and use as an input image. However, if the cropped hand while tracking process perform a gesture in front of the body, and overlap the body which may produce wrong outcome.
• The distance limitation range for detection between 0.5 to 4.5 m due to the Kinect sensors ranges limitation. Table 2 provides a summary table for the results obtained from the proposed system.

| No. of gestures | Type of gestures | Principle                                           | Recognition rate | Distance from camera |
|-----------------|------------------|-----------------------------------------------------|------------------|----------------------|
| Five gestures   | Hand gesture     | Feature extraction using CNN and gesture classify using (SVM) | 96.62 %          | 2 ~ 2.1 meter        |

5. Conclusion
This study explored the feasibility of extracting five hand gestures in real-time using the Microsoft Kinect V2 sensor based on joint tracking using deep learning CNN and SVM to extract and classify different gesture categories. The depth metadata provided by the Kinect was used to track hand joints and achieve dataset matching. The hardware design of the proposed system was carried out by using the Arduino Nano Microcontroller, GSM Module Sim800l and DC-Dc chopper. The experimental evaluation of the proposed system has been conducted in real-time for all participants in their home environment and showed reliable results. The experimental results were recorded and analysed by using the confusion matrix and gave acceptable outcomes with an overall recognition rate of 96.62%, making this system as a promising method for future home care applications.

Author Contributions
Conceptualization, Ali Al-Naji and Munir Oudah; Methodology, Munir Oudah and Al-Naji.; Data analysis and software, Munir Oudah and Ali Al-Naji; Supervision, Ali Al-Naji and Javaan Chahl; Writing—original draft preparation, Munir Oudah; Writing—review and editing, Ali Al-Naji and Javaan Chahl. All authors have read and agreed to the published version of the manuscript.

Conflict of interest
The authors of this manuscript have no conflict of interest relevant to this work.

Reference
[1] Bamwenda J and Özerdem M. S June 2019 Recognition of Static Hand Gesture with using ANN and SVM. *DUJE*. vol. 10. no.2. pp. 561-568.
[2] Pal D. H. and Kakade S 2016. Dynamic hand gesture recognition using kinect sensor. in *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPIC)*, pp. 448-453.
[3] Oudah M, Al-Naji A and Chahl J 2020 Hand Gesture Recognition Based on Computer Vision: A Review of Techniques. *Journal of imaging*. vol. 6. p. 73.
[4] Bukhari J, Rehman M, Malik S. I, Kamboh A. M and Salman A 2015 American sign language translation through sensory glove; signspeak *International journal of u-and e-service, science and technology*. vol. 8. pp. 131-142.
[5] Oudah M, Al-Naji A and Chahl J 2020 Hand Gestures for Elderly Care Using a Microsoft Kinect *Nano biomed. eng.* vol. 12. pp. 197-204.
[6] Oudah M, Al-Naji A and Chahl J 2021 Elderly Care Based on Hand Gestures Using Kinect Sensor. *Computers*. vol. 10, p. 15.
[7] Marin G, Dominio F and Zanuttigh P 2014 Hand gesture recognition with leap motion and kinect devices *IEEE International conference on image processing (ICIP)*. pp. 1565-1569.
[8] Wan Q, Li Y, Li C and Pal R 2014 Gesture recognition for smart home applications using portable radar sensors *36th annual international conference of the IEEE engineering in medicine and biology society*. pp. 6414-6417.

[9] Nunez J. C, Cabido R, Pantrigo J. J, Montemayor A. S and Velez J. F 2018 Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition *Pattern Recognition*. vol. 76. pp. 80-94.

[10] Devineau G, Moutarde F, Xi W and Yang J 2018 Deep learning for hand gesture recognition on skeletal data *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. pp. 106-113.

[11] Hussain S, Saxena R, Han X, Khan J. A and Shin H 2017 Hand gesture recognition using deep learning *International SoC Design Conference (ISOCC)*. pp. 48-49.

[12] John V, Boyali A, Mita S, Imanishi M and Samna N 2016 Deep learning-based fast hand gesture recognition using representative frames *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. pp. 1-8.

[13] Li G, Tang H, Sun Y, Kong J, Jiang G, Jiang D, Tao B, Xu S, Liu H 2019 Hand gesture recognition based on convolution neural network *Cluster Computing*. vol. 22. pp. 2719-2729.

[14] Nagi J, Ducatelle F, Di Caro G. A, Cireşan D, Meier U, Giusti A, Nagi F, Schmidhuber J, Gambardella L. M 2011 Max-pooling convolutional neural networks for vision-based hand gesture recognition *IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. pp. 342-347.

[15] Molchanov P, Gupta S, Kim K and Kautz J 2015 Hand gesture recognition with 3D convolutional neural networks in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1-7.

[16] Oyedotun O. K. and Khashman A 2017 Deep learning in vision-based static hand gesture recognition *Neural Computing and Applications*. vol. 28. pp. 3941-3951.

[17] Sun J.-H, Ji T.-T, Zhang S.-B, Yang J.-K and Ji G.-R 2018 Research on the hand gesture recognition based on deep learning in *2018 12th International Symposium on Antennas, Propagation and EM Theory (ISAPE)*. pp. 1-4.

[18] Xing K, Ding Z, Jiang S, Ma X, Yang K, Yang C, Li X, Jiang F 2018 Hand gesture recognition based on deep learning method in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*. pp. 542-546.

[19] Samir M, Golkar E and Rahni A. A. A 2015 Comparison between the Kinect™ V1 and Kinect™ V2 for respiratory motion tracking in *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. pp. 150-155.

[20] Yang L, Zhang L, Dong H, Aalelaawi A and Saddik A. El 2015 Evaluating and improving the depth accuracy of Kinect for Windows v2. *IEEE Sensors Journal*. vol. 15. pp. 4275-4285.

[21] Sudhan R. H, Kumar M. G, Prakash A. U, Devi S and Sathiya P 2015 Arduino ATMEGA-328 microcontroller *International journal of innovative research in electrical, electronics, instrumentation and control engineering*. vol. 3. pp. 27-29.

[22] Deng J, Dong W, Socher R, Li L.-J, Li K and Fei-Fei L 2009 Imagenet: A large-scale hierarchical image database *IEEE conference on computer vision and pattern recognition*, pp. 248-255.

[23] Alom M. Z, Taha T. M, Yakopcic C, Westberg S, Sidike P, Nasrin M. S, Van B. C, Awwal A. A, Asari V. K 2018 The history began from alexnet: A comprehensive survey on deep learning approaches *arXiv preprint arXiv:1803.01164*.

[24] Ballester P and Araujo R. M 2016 On the performance of GoogLeNet and AlexNet applied to sketches in *Thirtieth AAAI Conference on Artificial Intelligence*.

[25] Guan Q, Wang Y, Ping B, Li D, Du J, Qin Y, Lu H, Wan X, Xiang J 2019 Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study *Journal of Cancer*. vol. 10. p. 4876.

[26] Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition *arXiv preprint arXiv:1409.1556*.

[27] Krizhevsky A, Sutskever I and Hinton G. E 2012 Imagenet classification with deep convolutional neural networks in *Advances in neural information processing systems*. pp. 1097-1105.
[28] Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T 2014 Decaf: A deep convolutional activation feature for generic visual recognition in *International conference on machine learning*. pp. 647-655.