Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 171 (2020) 2353–2361

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

A Deep Convolutional Neural Network Approach for Static Hand Gesture Recognition

Adithya V.*[a], Rajesh R.[a]

[a]*Department of Computer Science, Central University of Kerala, Kasaragod, Kerala, India*

Abstract

The communication barrier and the hearing majority are the key social concerns of the deaf-dumb community that prevent them from accessing the basic and essential services of the life. Eventhough the problem has been addressed with the innovations in automatic sign language recognition, an adequate solution has not yet been attained due to a number of challenging factors. Most of the existing works try to develop vision based recognizers through classical pattern analysis approach by deriving complex hand crafted feature descriptors from the captured images of the gestures. But the efficiency of those methods are very limited to work with large sign vocabulary captured in complex and uncontrolled background conditions. This paper proposes a methodology for the recognition of hand gestures, which is the prime component in sign language vocabulary, based on an efficient deep convolutional neural network (CNN) architecture. The method has been tested on two publicly available datasets (NUS hand posture dataset and American fingerspelling A dataset) and achieved better recognition accuracies.

Keywords: Sign Language, Hand Gestures, Complex Backgrounds, Deep Learning, Convolutional Neural Networks

1. Introduction

Sign language enables the smooth communication in the community of people with speaking and hearing difficulty (deaf and dumb). They use hand gestures along with facial expressions and body actions to interact with each other. But, as it is not an international language, only very few people learn the sign language gestures [1]. The communication barrier that arises when the deaf and dumb people want to interact with the hearing people who do not know sign language is a major concern in the society. This undeniable gap in communication is usually filled up by the help of interpreters who translates the sign language to spoken language and vice versa. This system is very expensive

* Corresponding author. Tel.: +91-8129897716.
E-mail address: adithyaushas88@gmail.com

10.1016/j.procs.2020.04.255

and may not be available throughout the life of a deaf person. So the developments in automatic recognition of sign language gestures will be very beneficial to the deaf and dumb community as it will lead towards breaking the existing communication barrier [2].

Hand gestures constitute the major component of the sign language vocabulary, whereas facial expressions and body actions play the roles of giving emphasis to the words and phrases expressed by hand gestures. Hand gestures can be static or dynamic [3] [4]. Static hand gestures are otherwise known as hand postures and are formed of various shapes and orientations of hands without representing any motion information. Dynamic hand gestures are constituted by a sequence of hand postures with associated motion information. Hand postures mainly constitute the fingerspelling of the sign language vocabulary, which are used for the letter by letter signing of names, place names, age, numbers, date, year and words that doesn't have predefined signs in the vocabulary. Visual interfacing using hand postures have also received wide acceptance in varied application fields (human computer interaction (HCI) [5], human robot interaction (HRI) [6], virtual reality systems [7] and medical procedures [8]) as it avoids the physical contact with the traditional interfacing devices. Thus automatic hand posture recognition has been a hot research area and many works exist on the same using vision based approaches and electronic signal based approaches [6] [9]. Among those, the vision based approaches seem to be more user friendly and convenient than others when considering the complexity of data acquisition process.

Most of the existing works on vision based hand gesture recognition mainly follow the steps of classical pattern analysis that goes through image/video pre-processing, feature extraction and classification [4] [10] [11]. Ali et al. [12] presented a multilayer perceptron for recognizing the letters and numbers in Persian Sign Language (PSL) using the features derived with discrete wavelet transform (DTW) and achieved a classification rate of 94.06%. Al-Rousan et al. [13] used discrete cosine transform (DCT) features and HMM for user independent Arabic Sign Language Recognition (ArSLR). They claim a recognition accuracy of 87%. Cao et al. [14] proposed a hand posture recognition approach with heterogeneous feature fusion and obtained a recognition accuracy of 99.16% on publicly available Triesh data set. They trained a multiple kernel support vector machine classifier with feature vectors obtained by combining the shape context feature, pyramidal HOG feature and SIFT based bag of feature for recognizing the various static hand gestures. Nasser et al. [15] presented a novel hand gesture recognition method using multiclass support vector machine with bag of features derived through scale invariant feature transform (SIFT) descriptors. They achieved a classification rate of 96.23%. Pugeault et al. [16] presented a multiclass random forest classifier to recognize 24 static signs in ASL (American Sign Language) alphabet with the features extracted through Gabor filters in four levels. They claimed a recognition rate of 49%. Pramod et al. proposed a method for hand posture recognition in presence of complex background objects. They utilized the feature based visual attention with the shape, texture and color descriptors extracted from images and the classification using a multiclass support vector machine gives an accuracy of 94.36% [17]. In addition to the hand gestures, face expressions and body postures are also getting great importance in gesture communication. So the development of a recognition system that integrates all the three categories of gestures is also an ongoing research topic. Yang and Lee [18] proposed a work for the combined recognition of hand gestures and facial expressions for automatic British Sign Language (BSL) recognition.

Despite the promising results, the classical methods fail to derive consistent feature descriptors for hand posture recognition in real time scenarios due to a number of challenging factors [10][4]. The reasons for these challenges are mainly due to the failure of conventional machine learning techniques to accurately learn the distinguishable information of patterns from the natural raw input data. One such challenge faced by the hand posture recognition approach is the detection and segmentation of hands from images captured with complex background conditions [19]. Another difficulty is in deriving the powerful features that characterizes the geometrical variations in appearance of the same hand posture shown by different individuals [4]. The existence of large number of gesture classes with very small interclass variation is another challenging issue, especially in automatic sign language recognition [4]. It requires computationally complex image/video analysis steps with considerable domain knowledge to transform the raw images into the most discriminative representation by which the classifier can detect and differentiate the patterns accurately. The lack of publicly available datasets with sufficient number of sample images is another blocking factor in the study of sign language recognition. The major reason for this issue is the variations in sign language vocabulary used in different countries as well as regions of the world. For example, ASL, BSL, CSL (Chinese Sign Language) , ISL (Indian Sign Language), PSL etc.

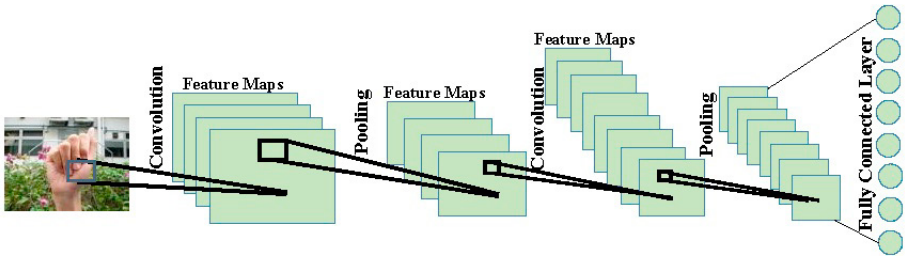

Fig. 1: The architecture of a typical CNN.

The recently emerged deep learning techniques, and advancements in convolutional neural networks (CNN) outweighs the classical approach to hand gesture recognition as it avoids the need of deriving complex hand crafted feature descriptors from images, following the conventional pre-processing and segmentation steps [20] [21] [22] [23]. CNNs automate the process of feature extraction by learning the high level abstractions in images and capture the most discriminative feature values using hierarchical architecture [24] [25]. Thus it solves the drawback of getting inconsistent feature descriptors, when working with large number of gesture classes with very slight interclass variations. Ameen et al. [26] proposed a recognition model for letters of ASL alphabet using CNN. They utilized the features extracted from both color and depth images of gestures using two parallel CNNs and achieved a recognition accuracy of 80.34% on the ASL fingerspelling bench mark dataset. A deep learning approach presented by Rastgoo et al. [27] utilized RBMs(Restricted Boltzmann Machine) to recognize the ASL fingerspelling with RGB and depth images. This model utilized CNNs for hand detection and the detected hand images are fed as input to the RBMs to recognize the sign labels. Their model has been tested on four publicly available datasets (Massey University Gesture Dataset, American Sign Language (ASL) and Fingerspelling Dataset from the University of Surrey's Center for Vision, Speech and Signal Processing, NYU, and ASL Fingerspelling A datasets) and achieved better recognition accuracies. Another deep learning framework with CNN has been presented by Mohanty et al. [28] to recognize the static hand gestures in the presence of complex background and varying illumination conditions. Their proposed model consisting of two convolution and pooling layers with ReLu activation function has been tested on three publicly available benchmark datasets namely, NUS hand posture dataset with cluttered background, Triesh hand posture dataset with uniform dark background and Marcel hand posture dataset, and obtained good recognition results.

This paper proposes a method for the automatic recognition of hand postures using convolutional neural networks with deep parallel architectures. The proposed model avoids the need for hand segmentation, which is a very difficult task in images with cluttered backgrounds. Eventhough many different segmentation techniques are possible based on skin color, hand shape etc., they all fail in giving proper results when applying to images with other background objects. This approach also avoids the hectic job of deriving potential feature descriptor capable of recognizing the various gesture classes. The method has been tested on the datasets with very large numbers of gesture classes with images having both uniform as well as complex backgrounds and the results are promising. The paper is organized as follows: Section 2 explains the proposed hand gesture recognition approach, the experimental results are discussed in section 3 and finally section 4 concludes the paper with proper remarks.

2. Proposed Method

Deep learning is an extension of neural network architecture that automate the process of feature extraction from meta data by processing it through a number of hierarchical layers. This paper aims to utilize such a deep learning architecture using convolutional neural network to recognize the static hand gestures. This model avoids the tedious and computationally complex feature extraction phase of the traditional pattern recognition approach. Fig. 1 depicts the architecture of a typical CNN proposed by LeCun et al.[29].

The convolution layers contains units called feature maps and each of them is connected to the local patches in the previous layer through filter banks [30]. Same filter bank is used in all the units of a feature map, and different filter banks are used in different feature maps in a layer. This architecture enables to easily identify the distinctive local patterns from images, even it is located at different parts of the image [30]. The local weighted sum obtained

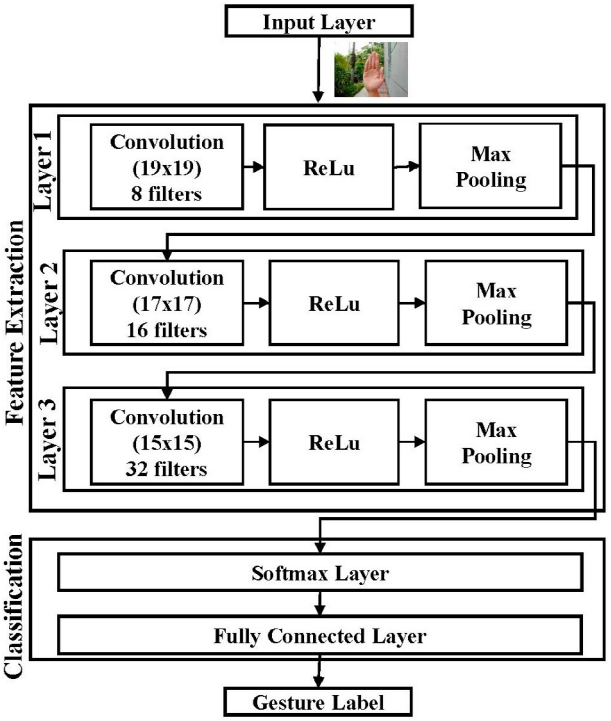

Fig. 2: The schematic representation of the proposed CNN model for hand posture recognition.

through filtering operation is passed through a non-linear function called ReLu(Rectified Linear Unit) to stabilize the convolved results. The pooling operation is incorporated in the CNN structure to group the semantically similar features from the convolution layer. Thus the architecture of a CNN contains two or three convolution layers with the non linear activation and pooling layers, followed by more convolutional layers with pooling and activation, and a final fully connected layer that performs the classification.

2.1. System Architecture

The CNN architecture for the gesture classes considered in our study is shown in Fig. 2. The model is constructed with an input layer, three convolution layers along with ReLu and maxpooling layers for feature extraction, one softmax output layer and a final fully connected output layer for classification of gestures.

In this work, images are first rescaled to 100x100 pixels and the dataset is divided into training and test sets prior to be fed as input to the CNN. Input layer takes the RGB images of hand postures to further layers for feature extraction and classification. The real strength of deep learning with CNN lies in the convolution layer. CNN follows cascaded discrete convolution of the kernel with the whole image as well as intermediate feature maps to extract the most potential feature for characterizing the hand shapes. Equation (1) gives the convolution of an image or feature map f, with a kernel k (square matrix). The number of filters in each convolutional layer is an empirical parameter to be determined through experiments.

$$f * k = \sum_{p,q=0}^{r-1} (f_{i+p,j+q})(k_{r-pr-q}) \tag{1}$$

Fig. 3: Sample images from the NUS Hand Posture Dataset.

Three convolutional layers are incorporated in the proposed architecture with eight filters in the first layer each with a size of 19x19, 16 filters in the second layer each having a size of 17x17, and 32 filters in the third layer each with the size of 15x15 respectively.Padding is applied in each convolutional layer to keep the size of the output same as that of the input. The output from the convolution operation is passed through a number of neurons with ReLu activation function. It applies the non-saturating and non-linearity function as in equation (2), to replace the negative values of the pooling layer with zero. The computational simplicity and representational sparsity of ReLu make it a preferred choice as activation function in deep neural networks. Pooling layer that comes after each ReLu layer reduces the dimensions of the feature maps without losing the most important information in them. Among the various pooling functions, maxpooling outperforms the others with its high speed and improved convergence property and it is adopted in this work. The maxpooling operation with a filter size of (2,2) and stride of size (3,3) is applied after each convolution layer to pick up the maximum value from each local patch of the feature maps.

$$y = max(0, x) \tag{2}$$

The most discriminative feature values extracted by the multiple stacked structure composed of the convolution layers, ReLu layers and pooling layers are fed as input of the classification layer composed of softmax layer and fully connected layer/output layer. The number of neurons in the softmax layer is same as that of the output layer and it transforms the feature values into the range 0 to 1 using a multiclass sigmoid function. Actually, the feature vector obtained from this layer very well predicts the chance of occurrence of patterns in an image. The final fully connected layer aims to classify the input images into the corresponding gesture classes based on the feature vector generated from the softmax layer. The number of neurons in this layer correspond to the number hand posture classes.

3. Experimental Results

The proposed methodology has been tested on two different publicly available hand posture datasets. First is the National University of Singapore (NUS) hand posture dataset with cluttered background, containing 10 different hand posture classes with 200 images of each class. The postures are performed by 40 individuals of different ethnicities in complex natural backgrounds [17]. The sample images are shown in Fig. 3.

In the training phase, the images are passed through a three layered convolutional operation with filter sizes of 19x19, 17x17 and 15x15 respectively to extract the features. The numbers of filters in each layer are 8,16 and 32 respectively and proper zero padding is applied in each layer with a stride size of [1,1] to maintain the size of input equal to the size of output. The maxpooling layer with filter size (2,2) and stride size of three reduces the dimensionality of feature maps obtained from each of the convolutional layers. CNN training with stochastic gradient descent with momentum (SGDM) optimization function, having a momentum rate of 0.90 and learning rate of 0.01 is utilized in

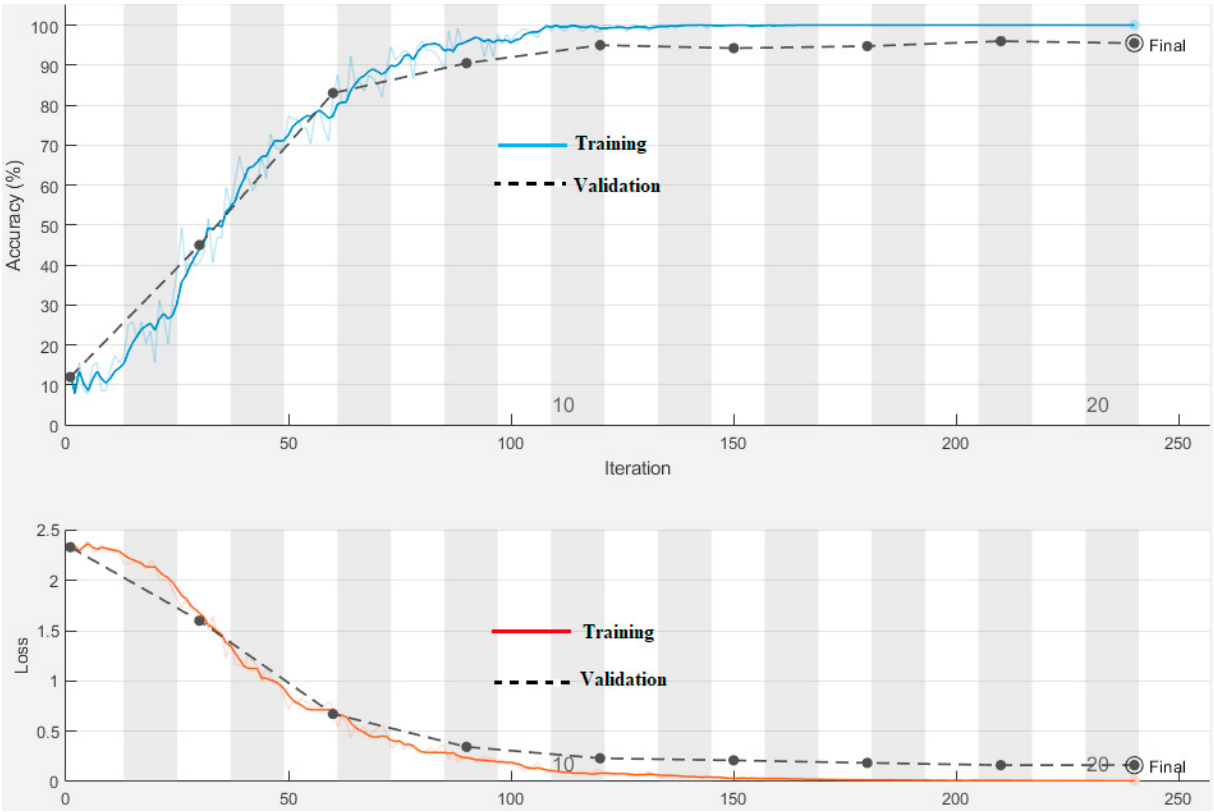

Fig. 4: Simulation graph showing the accuracy and loss function for the proposed CNN training on NUS hand posture dataset.

this study. The training process took 20 epochs to converge to an optimal CNN model to recognize the hand postures. An example for the simulation graph is given in Fig. 4.

The performance of the classification is evaluated with a five fold cross validation. The dataset is divided into five subsets including 40 sample images of each gesture class. The classifier is trained with any four subsets and the remaining one subset is used for testing. The experiment is repeated five times in a similar manner until each of the subset is used for development and testing. The classification result evaluated using the average accuracy, precision, recall and F1-score values is given in Table 1. The macro averaging approach has been adopted here for calculating the values of the performance metrics. The table illustrates the greater recognition capability of the proposed CNN model. Table 2 gives a comparison of the average accuracy (calculated over the five runs) with the results from the works [17][28].

Table 1: Interpretation of the Classification Performance of the Proposed CNN Model on NUS Hand Posture Dataset using Statistical Measures

Accuracy	Precision	Recall	F1-Score
94.7 ± 0.80 %	94.96 ± 1.20 %	94.85 ± 1.30 %	94.26 ± 1.70 %

Second dataset is the American fingerspelling A dataset which contains 24 letters of the ASL alphabet excluding the letters 'j' and 'z' (since they involve motion) [16] [27]. The images of this set are captured in five different sessions, with five different users in similar lighting conditions in presence of complex background objects. Since the number of samples of each gesture classes are different, 470 samples of each class is selected for this experiment. The sample images from the dataset is shown in Fig. 5.

Table 2: Comparison of the Recognition Accuracy of the Proposed Method on NUS Hand Posture Dataset

Author	Method	Number of samples Training/Testing	Accuracy
Pramod et al. [17]	C2SMF (color,shape and texture)/ multiclass SVM	1800/200	94.36 %
Mohanty et al. [28]	Deep Learning with CNN	1200/800	89.10 %
Proposed	Deep learning with CNN	1600/400	94.7 ± 0.8 %

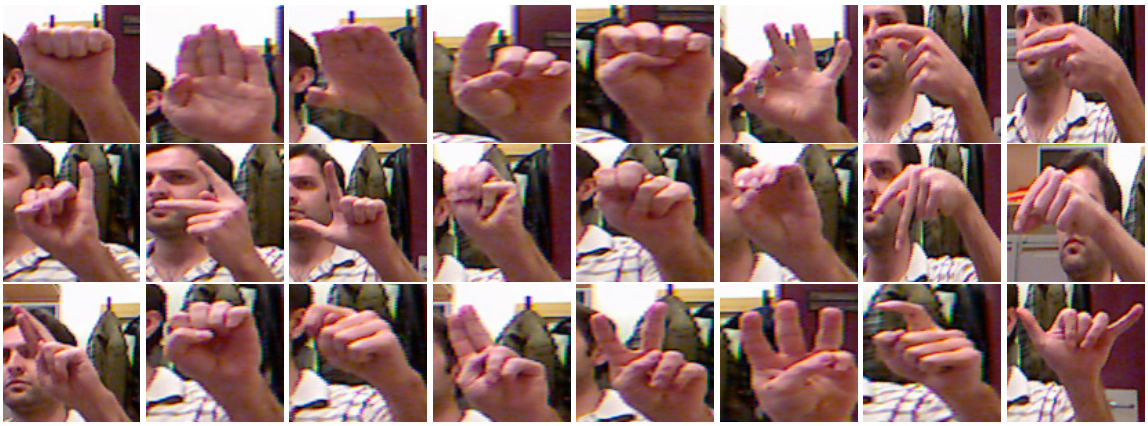

Fig. 5: Sample images from the American Finger Spelling A Dataset.

The dataset is divided into five subsets including 94 sample images of each gesture class. The classifier is trained with any four subsets and the remaining one subset is used for testing. The number of convolutional layers and the other training parameters are same as that of the experiment conducted for NUS dataset. The experiment is repeated five times in a similar manner until each of the subset is used for development and testing. Table 3 shows the values of the statistical performance measures which indicate the higher recognition ability of the proposed CNN architecture and Table 4 gives the comparative analysis of the average accuracy (calculated over the five runs) with the results from the work [27].

Table 3: Interpretation of the Classification Performance of the Proposed CNN Model on ASL Fingerspelling Dataset using Statistical Measures

Accuracy	Precision	Recall	F1-Score
99.96 ± 0.04 %	99.96 ± 0.04 %	99.96 ± 0.04 %	99.96 ± 0.04%

4. Conclusion

The power of deep learning tool for the recognition of hand postures from raw images (RGB) has been utilized in this work. The proposed CNN architecture eliminates the need of detection and segmentation of hands from the captured images, thus reducing the computational burden faced during hand posture recognition with classical ap-

Table 4: Comparison of the Recognition Accuracy of the Proposed Method on ASL Fingerspelling A Dataset

Author	Method	Number of samples Training/Testing	Accuracy
Rastagoo et al. [27]	Deep Learning with RBM	–	98.13 %
Proposed	Deep learning with CNN	9024/2256	99.96 ± 0.04 %

proaches. Also the model can automatically derive the potential features that discriminates the hand postures even having very small interclass variations. The performance of the proposed method has been evaluated on two publicly available datasets through five fold cross validation. The performance analysis using the statistical metrics such as accuracy, precision, recall and F1-score shows greater recognition power of the proposed CNN model.

Acknowledgements

The first author thanks Kerala State Council for Science Technology and Environment (KSCSTE) for the research fellowship. The authors also express their thanks to Central University of Kerala for the research support.

References

[1] Ashok K Sahoo, Gouri Shankar Mishra and Kiran Kumar Ravulakollu (2014), "Sign Language Recognition: State of the Art." *ARPN Journal of Engineering and Applied Sciences* **9** (2).
[2] Alice Caplier, Sebastien Stillittano, Oya Aran, Lale Akarun, Gerard Bailly, Denis Beautemps, Nouredine Aboutabit and Thomas Burger (2007). "'Image and Video for Hearing Impaired People."' *EURASIP Journal on Image and Video Processing*.
[3] Sushmita Mitra, Tinku Acharya (2007), "Gesture Recognition: A Survey." *IEEE Transactions on Systems, Man and Cybernetics- Part C: Applications and Reviews* **37**(3).
[4] Siddharth S. Rautaray and Anupam Agrawal (2015), "Vision based Hand Gesture Recognition for Human Computer Interaction: A Survey." *Artificial Intelligence Review, Springer* **43** : 1–54.
[5] V. I. Pavlovic, R. Sharma, and T. S. Huang (1997), "Visual interpretation of hand gestures for human computer interaction." *IEEE Trans. Pattern Analysis and Machine Intelligence* **19**(7) : 677–695.
[6] Hongyi Liu, Lihui Wang (2017), "Gesture recognition for human-robot collaboration: A review." *International Journal of Industrial Ergonomics*(in press).
[7] Martin Sagayam, Jude Hemanth (2017), "Hand Posture and Gesture Recognition Techniques for virtual reality applications: a survey." *Virtual Reality*.
[8] Mithun G. J., Juan P. W. and Rebecca A. P. (2012), "Hand Gesture based Sterile Interface for the Operating Room using Contextual Cues for the Navigation of Radiological Images." *Journal of American Medical Informatics Association*: 183–186.
[9] Cheok, M.J., Omar, Z. and Jaward, M.H. (2019), "A review of hand gesture and sign language recognition techniques." *International Journal of Machine Learning and Cybernetics* **10** : 131–153.
[10] Pramod Kumar Pisharady and Martin Saer Beck (2015), "Recent Methods and Databases in Vision based Hand Gesture Recognition: A Review." *Computer Vision and Image Understanding* **141** : 152–165.
[11] Orazio, Marani, Reno and Cicirelli (2016), "Recent trends in gesture recognition: how depth data has improved classical approaches." *Image and Vision Computing* **52** : 56–72.
[12] Ali Karami, Bahman Zanj and Azadeh Kiani Sarkaleh (2011), "Persian sign language (PSL) recognition using wavelet transform and neural networks." *Expert Systems with Applications* **38** : 2661–2667.
[13] AL-Rousan, K. Assaleh and A. Talaa (2009), "Video-based signer-independent Arabic sign language recognition using hidden Markov models." *Applied Soft Computing* **9** : 990–999.
[14] Jiangtao Cao, Siquan Yu, Honghai Liu and Ping Li (2016), "Hand Posture Recognition based on Heterogeneous Features Fusion of Multiple Kernels Learning." *Multimedia Tools Applications, Springer* **75** : 11909–11928.
[15] Nasser H.Dardas and Nicolas D.Georganas (2011), "Real Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques", *IEEE Transactions on Instrumentation and Measurement* **60** : 3592–3607.
[16] Pugeault N.,Bowden R. (2011), "Spelling It Out: Real-Time ASL Fingerspelling Recognition." *In Proceedings of the 1st IEEE Workshop on Consumer Depth Cameras for Computer Vision, jointly with ICCV'2011*.

[17] Pramod Kumar Pisdharady, Prahlad Vadakkeppat and Loh Ai Poh (2013), "Attention Based Detection and Recognition of Hand Postures Against Complex Backgrounds." *International Journal of Computer Vision* **101** : 403–419.

[18] Yang and Lee (2013), "Robust Sign Language Recognition by Combining Manual and Non-Manual Features based on Conditional Random Field and Support Vector Machine." *Pattern Recognition Letters* **34** : 2051–2056.

[19] Ekaterini Stergiopoulou, Kyriakos Sgouropoulos, Nikos Nikolaou, Nikos Papamarkos and Nikos Mitianoudis (2014), "Real time hand detection in a complex background." *Engineering Applications of Artificial Intelligence* **35** : 54–70.

[20] Neto G.M.R., Junior G.B., de Almeida J.D.S., de Paiva A.C. (2018), "Sign Language Recognition Based on 3D Convolutional Neural Networks." *In: Campilho A., Karray F., ter Haar Romeny B. (eds) Image Analysis and Recognition. ICIAR Lecture Notes in Computer Science* **10882**.

[21] Gongfa Li1, Heng Tang, Ying Sun, Jianyi Kong, Guozhang Jiang, Du Jiang, Bo Tao, Shuang Xu and Honghai Liu (2017), "Hand gesture recognition based on convolution neural network." *Cluster Computing, Springer* doi: https://doi.org/10.1007/s10586-017-1435-x.

[22] Wenjin Tao, Ming C. Leu and Zhaozheng Yin b (2018), "American Sign Language alphabet recognition using Convolutional Neural Networks with multiview augmentation and inference fusion." *Engineering Applications of Artificial Intelligence* **76** : 202–213.

[23] Kai Xing, Zhen Ding, Shuai Jiang, Xueyan Ma, Kai Yang, Chifu Yang, Xiang Li and Feng Jiang (2018), "Hand Gesture Recognition Based on Deep Learning Method." *IEEE Third International Conference on Data Science in Cyberspace (DSC)* doi:10.1109/DSC.2018.00087

[24] Boukaye Boubacar Traore, Bernard Kamsu-Foguem and Fana Tangara (2018), "Deep convolution neural network for image recognition." *Ecoinf.* doi:10.1016/j.ecoinf.2018.10.002.

[25] Carlos Affonso, Andre Luis Debiaso Rossi, Fabio Henrique Antunes Vieira, Andre Carlos Ponce de Leon Ferreira de Carvalho (2017), " Deep learning for biological image classification." *Expert Systems With Applications* **85** : 114–122.

[26] Salem Ameen, Sunil Vadera (2016), "A convolutional neural network to classify American Sign Language fingerspelling from depth and colour images." *Wiley Expert Systems*.

[27] Razieh Rastgoo, Kourosh Kiani and Sergio Escalera (2018), "Multi-Modal Deep Hand Sign Language Recognition in Still Images Using Restricted Boltzmann Machine." *entropy* **20**(11).

[28] Mohanty A., Rambhatla S.S., Sahay R.R. (2017), "Deep Gesture: Static Hand Gesture Recognition Using CNN." *In: Raman B., Kumar S., Roy P., Sen D. (eds) Proceedings of International Conference on Computer Vision and Image Processing. Advances in Intelligent Systems and Computing, Springer, Singapore* **460**.

[29] Lecun, Y., Bottou, L., Orr, G. and Muller, K.R. (1989). "Efficient Back Prop. In G. Orr, and K.R. Muller (Eds.)." *Neural networks: Tricks of the trade* **1524** : 9–50.

[30] Lecun, Y., Bengio, Y. and Lhinton G. (2015). "Deep learning." *Nature* **521** : 436–444.