Gesture Recognition of RGB and RGB-D static Images using Convolutional Neural Networks

Article *in* International Journal of Interactive Multimedia and Artificial Intelligence · September 2019

DOI: 10.9781/ijimai.2019.09.002

CITATIONS
52

READS
1,350

4 authors, including:

Manju Khari
Jawaharlal Nehru University
125 PUBLICATIONS **1,044** CITATIONS
SEE PROFILE

Ruben Gonzalez Crespo
Universidad Internacional de La Rioja
322 PUBLICATIONS **2,118** CITATIONS
SEE PROFILE

Elena Verdú
Universidad Internacional de La Rioja
75 PUBLICATIONS **657** CITATIONS
SEE PROFILE

Article in Press IJIMAI journal

Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks

Manju Khari[1] *, Aditya Kumar Garg[1] *, Rubén González Crespo[2], Elena Verdú[2]

[1] Computer Science and Engineering, Ambedkar Institute of Advanced Communication Technologies and Research, Geeta Colony, Delhi (India)
[2] Universidad Internacional de La Rioja, Logroño (Spain)

Received 28 June 2019 | Accepted 10 September 2019 | Published 16 September 2019

Abstract

In this era, the interaction between Human and Computers has always been a fascinating field. With the rapid development in the field of Computer Vision, gesture based recognition systems have always been an interesting and diverse topic. Though recognizing human gestures in the form of sign language is a very complex and challenging task. Recently various traditional methods were used for performing sign language recognition but achieving high accuracy is still a challenging task. This paper proposes a RGB and RGB-D static gesture recognition method by using a fine-tuned VGG19 model. The fine-tuned VGG19 model uses a feature concatenate layer of RGB and RGB-D images for increasing the accuracy of the neural network. Finally, on an American Sign Language (ASL) Recognition dataset, the authors implemented the proposed model. The authors achieved 94.8% recognition rate and compared the model with other CNN and traditional algorithms on the same dataset.

Keywords

American Sign Language, Image Processing, CNN, Gesture Recognition.

DOI: 10.9781/ijimai.2019.09.002

I. Introduction

In Computer Accessibility, researchers keep on investigating new and assistive technologies for the people who suffer from disabilities. These technologies consist of devices and software that are intended to benefit the people who suffer from disabilities. Moreover these technologies may help users in performing tasks with much larger efficiency which in turn will help in increasing their quality of life. According to an estimate of World Health Organization (WHO) in 2019, there are about 466 million people who suffer from hearing loss [1]. In 2005, the number of deaf people was 278 million [2]. In around 15 years, there is a significant rise in the number of people having hearing problems.

Deaf people use numerous methods for communication purpose. One of these methods include Sign Languages that are made up of movements of the hands, torso, arms, head, facial expressions and eyes. Sign Language is a hand gesture language which is most common among the deaf people to be used as way of expressing their feelings, thoughts, and knowledge in the place of verbal communication. According to [3], there are more than 100 sign languages. Moreover, there is no a unified way of writing sign languages, being SignWriting a proposal on which some research work has been developed [4]. American Sign Language (ASL) is one of the most commonly used sign languages throughout the U.S. and Canada and also including Southeast Asia and West Africa. According to an estimate in [5], there are about 250,000–500,000 deaf people who rely on using American Sign Language (ASL). ASL fingerspelling consists of 36 signs and is also the sixth most used sign language in US.

There are many works on gesture recognition for different purposes, some focusing on the whole body [6] while others focusing on a specific part as eyes [7] or hands [8]. Despite of the vast number of research works that have been published, there have been several limitations. Some of the limitations include: (1) many of the previous methods make use of add on devices, (2) most of the previous methods are based on getting more speed than getting a higher recognition rate, (3) most of the previous methods still make use of traditional learning algorithms that are based on feature extraction which requires high computation [9]. Gesture recognition can be classified into two different categories.

* Corresponding author.
E-mail addresses: manjukhari@yahoo.co.in (M. Khari), adityagarg2607@gmail.com (A. Kumar Garg).

Please cite this article in press as:
M. Khari, A. Kumar Garg, R. González Crespo, E. Verdú. Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks, International Journal of Interactive Multimedia and Artificial Intelligence, (2019), http://dx.doi.org/10.9781/ijimai.2019.09.002

The first one is electromagnetic gloves and sensors based detection and the other is Computer Vision based. The electromagnetic gloves and sensor based technique is very expensive and is not suitable for real life purposes [10]. On the other hand, the Computer Vision based technique can be further divided into Static Gesture Recognition and Dynamic Gesture Recognition. There are many challenges in the area of hand gesture recognition such as (1) feature extraction (2) variation in hand size, (3) hand partial occlusion.

In recent years, the field of deep learning is under rapid development. Particularly, Convolutional Neural Networks (CNN) are able to achieve far more effective results related to the field of Image Classification, Natural Language Processing, etc. With increasing popularity of CNN, many new CNN Models such as GoogLeNet [11], VGG16, VGG19 [12], and Inception V3 [13] have emerged and were able to achieve significant results in ImageNet Large-scale Visual Recognition Challenge (ILSVRC).

As these CNN models became more popular, a concept termed as Transfer Learning gained reputation. Transfer Learning is the transfer of parameters of a previously trained model to help the training of another model. The main advantage of Transfer Learning is to avoid the overfitting of models and also to reduce the time taken to train a model. With the help of Transfer Learning, the model can be more easily converged. There are a huge number of research results that helped in confirming the performance of models trained with the help of Transfer Learning. For example, GoogLeNet Model which was pre-trained with ImageNet image library can be used to predict diabetic retinopathy [14]

The main contribution of this paper is to propose a Neural Network that will help in increasing the recognition rate on American Sign Language Dataset. The main focus of this paper is on Static Gesture Recognition based techniques. Presently, Static Gesture Recognition disadvantages include non-robust and inaccurate recognition under abrupt lightning changes and complex background. Inaccurate Gesture Segmentation in turn affects the accuracy of gesture classifications [15]. The proposed algorithm uses Transfer Learning, which not only helps in eliminating the need of feature extraction but also helps in reducing the computational power required to obtain a higher accuracy for Sign Recognition.

The rest of the paper is organized as follows. Section II describes the Related Work. Section III explains the proposed Work. Section IV describes the research methodology. Section V explains the results and comparative analysis. Section VI includes the conclusion.

II. Related Work

Pugeault and Bowden (2011) [16] used a Microsoft Kinect for the collection of Intensity and Depth Images of American Sign Language (ASL) 24 letters (except J and Z). On this dataset the authors used OpenNI + NITE framework for gesture detection and tracking. For extraction of features, the authors used a set of Gabor filters and then the classification was done using Random Decision Forest. The paper concluded that the average recognition rate is 73% when only Intensity Images were considered and 69% when only Depth Images were considered. The combined recognition rate is 75%.

Estrela et al. (2013) [17] proposed a framework on the basis of bag of features in combination with Partial Least squares (PLS).The authors split the experiment into two groups and computed results on the basis of ASL dataset. In the first group, the experiments compared Support vector Machine (SVM) classifiers and Partial Least Squares (PLS) classifiers. The accuracy achieved in the first experiment of the PLS and SVM classifiers is 66.27% and 62.85%, respectively. In the second experiment, BASE and SIFT feature descriptors are appraised. BASE feature extractor runs faster and consumed less memory while the SIFT feature extractor achieves a better accuracy. The accuracy with PLS is 71.51% and with SVM is 65.55%, respectively.

Chuan et al. (2014) [18] used a 3D motion sensor based system for implementing American Sign Language Recognition (ASL). The authors applied KNN and SVM for classification of 26 letters on the basis of features derived from the sensory data. According to the experiments, the result shows that the highest rate for average classification is 72.78% and 79.83% which was achieved by k-Nearest Neighbour (KNN) and Support Vector Machine (SVM) respectively. Rioux-Maldague and Giguère (2014) [19] presented a novel feature extraction technique which uses both the depth and intensity image that were captured from a Microsoft KnectTMsensor. Then, the authors used a Deep Belief Network on an American Sign Language dataset.

Ameen and Vadera (2017) [20] developed a CNN aimed for the classification of both the colour and depth Images. The CNN was applied to American Sign Language (ASL) database. The authors were able to achieve a precision equivalent to 82% and recall of 80%. Xie et al. (2018) [15] proposed a RGB-D Static Gesture Recognition based method using a fine-tuned Inception V3. In comparison to a traditional CNN, the authors adapted a two stage training strategy. The authors compared the proposed model with traditional methods and other CNN model. The highest accuracy authors were able to reach was 91.35%.

Dai et al. (2017) [21] proposed a SmartWatch-based American Sign Language (ASL) recognition system. The purpose of this system was to be more portable, comfortable and user friendly. The proposed system was designed such that each individual Sign having its own motion pattern can be transformed into accelerometer and gyroscope signals. In the next steps, these signals are analysed with the help of a Long-Short Term Memory Recurrent Neural Network (LSTM-RNN) trained with Connectionist Temporal Classification (CTC). Islam et al. [10] proposed a novel “K convex hull” method which is the combination of K curvature and convex hull algorithms. The “K convex hull” method developed is able to detect fingertip with high accuracy. In this paper, the system gathers ASL gesture images with black background and extracts mainly five features that are fingertip finder, pixel segmentation, elongatedness, eccentricity and rotation.

Tao et al. (2018) [22] proposed a Convolutional Neural Network (CNN) along with inference fusion and multiview augmentation. This method uses depth images captured by Microsoft Kinect. Chong and Lee (2018) [23] developed a prototype with the help of a Leap Motion controller (LMC). This study focussed on full ASL recognition i.e. all the 26 letters and 10 digits. The recognition rate for 26 letters with the help of Support Vector Machine (SVM) and Deep Neural Network (DNN) was 80.3% and 93.81% and for the Combination of 26 letters and 10 digits was 72.79% and 88.79%, respectively.

Lim et al. (2019) [24] proposed a two phase recognition system. The two main phases comprises of hand tracking and hand representation. In the first phase, the hand tracing is performed with the help of particle filter. And in the second phase, a compact hand representation is computed by averaging the segmented hand regions.

Hou et al. (2019) [25] proposed a smartwatch based Sign recognition system termed as SignSpeaker. SignSpeaker was developed by using a smartwatch and a smartphone. The average recognition rate achieved was 99.2% and 99.5%.

III. Proposed Work

A. VGG19 Model

The VGG Network was introduced by Simonyan and Zisserman in [12]. This network only uses 3x3 convolutional layers stacked on the top of each other in increasing depth. On the top of that a max pooling layer is introduced which handles the reducing of volume size. Max-

pooling is performed over a 2 × 2 pixel window. After Max-Pooling, the Model consists of three Fully-Connected Layers (FC Layers): the initial two layers both consist of 4,096 Nodes while the third layer is used to perform 1000-way ILSVRC classification and therefore consists of 1000 channels (one for each class). Finally a soft-max layer is introduced.

All the hidden layer in the VGG19 Model are equipped with rectification (ReLU) [26]. The general architecture of a VGG19 Model is shown in Fig. 1.

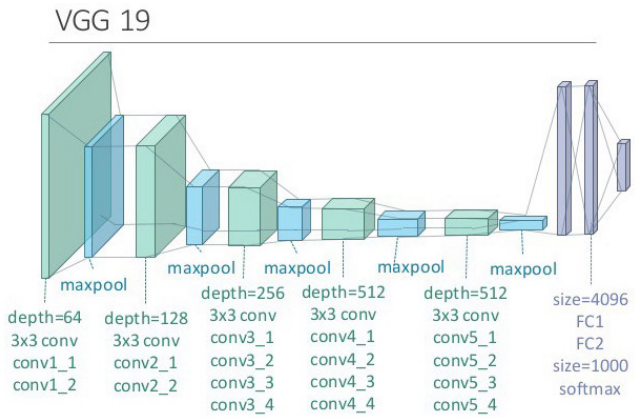


Fig. 1. VGG19 Model Architecture.

B. VGG19 Model Fine Tuning

The original model is pre-trained for the classification of 1000 classes. Since the number of classification classes are inconsistent for the experiment, the authors removed the topmost layer of the model and re-established a new fully connection layer of 24 classes for carrying out the experiment. For Fine-Tuning the model, i.e. for obtaining the appropriate top layer weights, all the other layer of the model are frozen and the model is trained for multiple rounds on the ASL Dataset. After training the model for a significant number of epochs, the model weights were saved. In the second stage, the authors did not trained the complete model due to the presence of relatively less data. Training the complete model will lead to overfitting. Hence, the authors adapt a strategy by freezing the first 16 layers of the model and training the rest of the model. For this stage, the authors used a low learning rate Stochastic Gradient Descent (SGD) and the model is trained.

C. Concatenation

The authors used the above strategy to develop two different VGG19 models namely VGG19-v1 and VGG19-v2. VGG19-v1 was trained by using only the RGB Images and similarly VGG19-v2 was trained only using the Depth Images.

On analysing the past researches on Static Gesture Recognition, there was a need to merge depth-information and colour information together for obtaining high accuracy and recognition rates. Thus, the authors combined the results of both VGG19-v1 and VGG19-v2, as shown in Fig. 2.

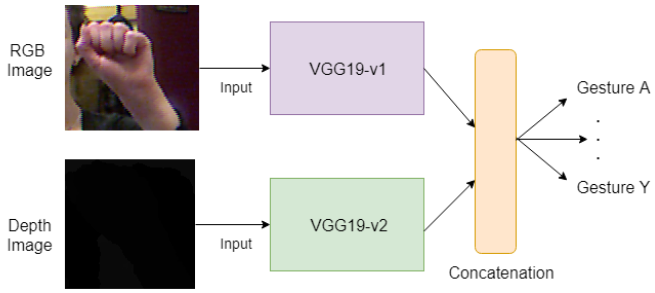
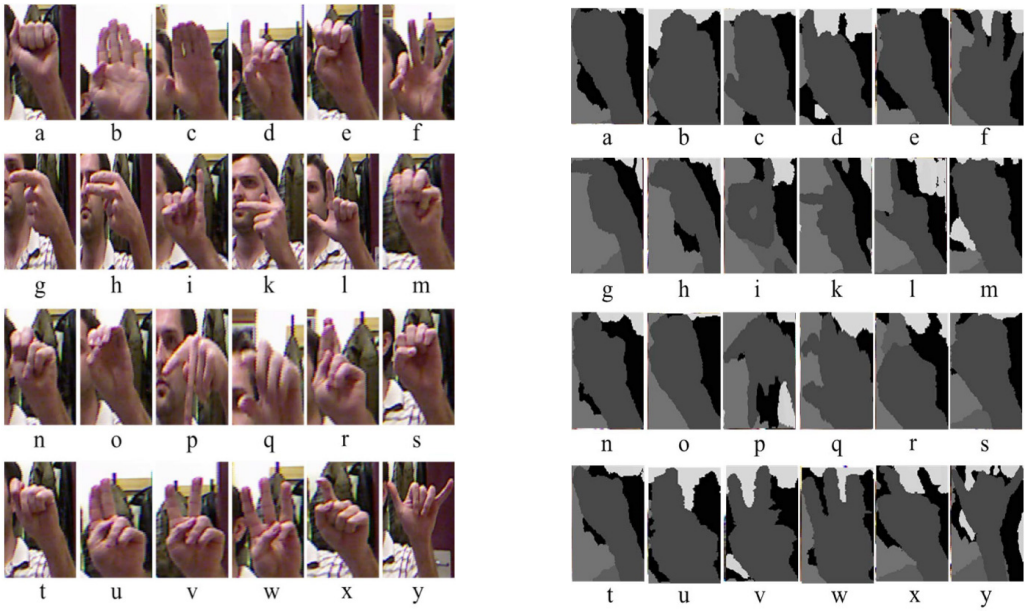


Fig. 2. Proposed Model.

IV. Methodology

A. Dataset

In this research, the ASL dataset published by Pugeault and Bowden in 2011 [16] is used. The dataset provides 24 (except y and z) English letter images in the form of gesture expressions. The ASL dataset is recorded by 5 different persons with the help of Kinect, with non-identical lightning conditions and background conditions. In the ASL dataset, there are approximately ~500 non identical Hand Gesture Images which correspond to each alphabet. Hence the dataset contains approximately 60,000 images for colour and depth. Fig. 3 shows some images from the dataset.



(a) (b)

Fig. 3. ASL Dataset (a) RGB Images (b) Depth Images.

B. Data Augmentation

For Data Augmentation, the authors used a data enhancement tool known as ImageDataGenerator provided by Keras framework. In this tool, the authors set different parameters like roation_range, height_shift_range, Width_shift_range, ZCA_whitening etc. for implementing data augmentation. The transformation helped in increasing the amount of data and also in preventing overfitting.

C. Model Training

The Model proposed in Section III is trained using the ASL Dataset. Both the models i.e. VGG19-v1 and VGG-v2 are trained using the RGB and RGB-D images as proposed. Both the models are trained using the two stage strategy adopted by the authors and the final weights of the model were saved. For avoiding the overfitting of the models, the authors used Early Stopping method in the training process. Early Stopping is the method that monitors model for stopping the training process. With the help of Early Stopping, the authors monitor the accuracy of Validation Set i.e. Validation Accuracy. If the validation accuracy falls to a certain level, the Early Stopping process stops the training of the model after some consecutive epochs.

Now, as the training data consist of both Intensity and Depth images, the authors implemented feature concatenation. Both the Intensity and Depth images are provided as input to the fine-tuned VGG-19 model and then a concatenate function is set just before the topmost soft-max layer and then the soft-max layer is used for classification.

V. Result and Discussion

A. Dataset Processing

1. Data Preprocessing

After carefully studying the dataset, the researchers found out that there are some unusual patterns in the dataset. In some folder there are unequal distribution of colour and depth images. To solve the issue, the researchers carefully removed the depth images that are different and do not correspond to any colour image. Since this research uses the VGG19 Model with an input image dimensions of 299 x 299 x 3, the researchers converted the single channel depth images to 3-channel images in which 1 channel preserves the original depth information and the remaining two channels are set to 0.

2. Augmentation

As the amount of data present in the original dataset is not very large, there was a need to perform Data Augmentation for the Dataset. Data Augmentation is defined as the way of creating new data which will have different orientations. Data Augmentation has two benefits which are as follows:

1. Data Augmentation helps in preventing overfitting.
2. Data Augmentation gives the ability to generate new data from limited data.

3. Final Dataset

After all the phases of cleaning and pre-processing through which the Dataset is passed, the Dataset was converted into two different subsets: Intensity and Depth. Both of these Subsets are divided into 3 categories namely Training Set, Validation Set and Test Set.

Training Set is the Set on which the model will be trained to recognize the ASL. Training Set contains approximately ~46000 Images. Validation Set is defined as the Set which will validate that whether the model trained using the Training Set is accurate or not. Test Set is the unseen Data that is used on the well trained model to predict the accuracy and performance of the model. Both the Validation Set and the Test Set contain approximately ~9000 Images. Fig. 4 shows the distribution of Images among the different sets.

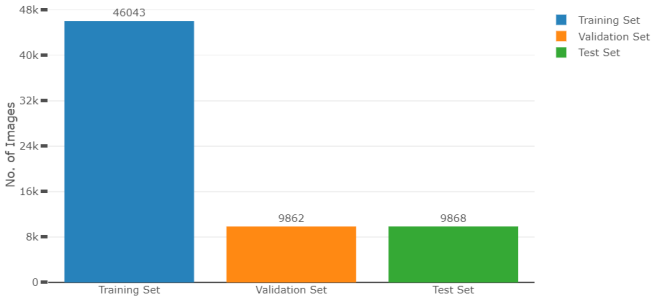


Fig. 4. Train, Test and Validation Split.

B. Tools

In this research, the authors performed comparative experiments. The proposed model is compared with deep learning algorithms and advanced machine learning algorithms. For ensuring the fairness of the experiment, all the models used the same dataset proposed in section III.C. The operating system which was used by the authors for the experiment was Ubuntu 18 and the GPU is Tesla K80. For CNN implementation, Keras framework is used with tensor flow as backend. Keras is an open-source library on neural network which is written in python.

C. Experimental Analysis

The experiments were performed on the ASL dataset. The Model was trained with the help of approximately ~46000 Intensity and ~46000 Depth Images and is tested on approximately ~9000 Intensity and ~9000 Depth Images. The average accuracy obtained by the authors is about 95.29% on the Training Set and about 94.80% on the Test Set. For evaluation of the performance of the proposed model, the authors proposed a Confusion Matrix on the basis of the Test Set Results for the combined Model. Fig. 5 shows the proposed Confusion Matrix.

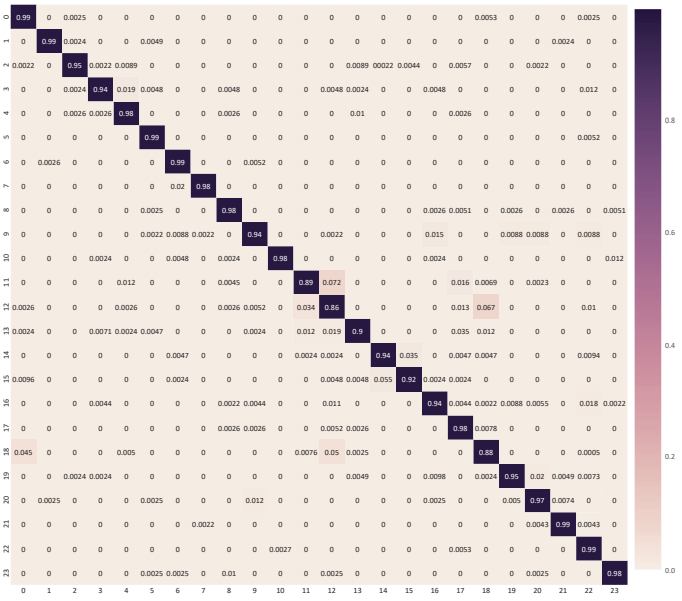


Fig. 5. Confusion Matrix.

D. Comparitive Analysis

The authors compared the results of the proposed model with results of the methods SIFT+PLS, H3DF+SVM and Gabor + RDF that were introduced in Section II. Table I and Fig. 6 represent the

TABLE I. Comparison between Traditional Models and Proposed Model

Recognition Methods	Gabor+RDF	SIFT+PLS	H3DF+SVM	Our Model
Recognition Rate	75%	71.51%	73.3%	94.8%

TABLE II. Comparison between Other CNN Models and Proposed Model

Recognition Methods	CaffeNet	VGG16	VGG19	Inception V3	Our Model
Recognition Rate	73.75%	83.44%	87.37%	88.15%	94.8%

Accuracy Comparison. All these algorithms are not implemented by the author, the accuracy of these models is compared with the proposed model's accuracy. Apart from these traditional methods, the authors also directly tested other CNN algorithms such as VGG16, VGG19, CaffeNet and Inception V3 without performing any Fine Tuning. Table II and Fig. 7 represent the Comparison between their Accuracy. It can be seen from the results that the model proposed in this paper has the highest accuracy among all the models.

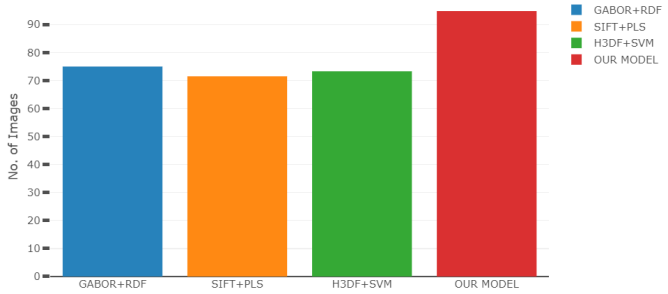


Fig. 6. Traditional Models vs. Proposed Model.

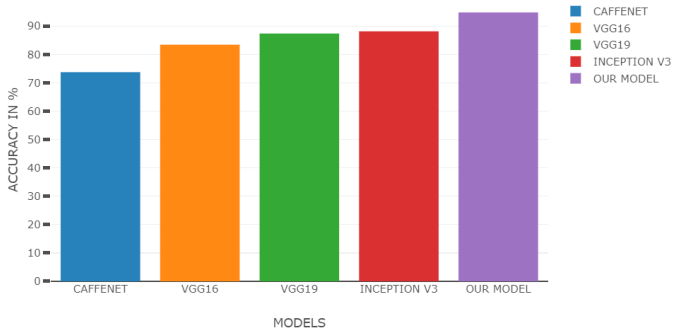


Fig. 7. CNN Models vs Proposed Model.

VI. Conclusion

CNN is currently a powerful artificial intelligence tool that can recognise patterns with high accuracy. In this paper, the authors proposed a fine-tuned VGG19 Model for implementing static gesture recognition. The VGG19 Model was fine-tuned using a two-stage process. In the first stage, all the layers of the model were frozen except the last layer and the model was trained on multiple rounds of ASL Dataset. In the Second stage, only the initial 16 layers of the model were frozen and rest of layers were trained on multiple rounds with low rate SGD due to presence of relatively less data. In comparison with other methods, many of which rely on Features Extraction, the proposed method can easily automate that task for classification. The proposed model is tested on ASL dataset and the recognition rate attained is 94.8%.

In addition, the authors did a comparative study with other models and on comparison it was determined that the proposed model outperforms certain traditional machine learning methods namely Gabor+RDF, SIFT+PLS and H3DF+SVM. Moreover, the model was compared with different CNN models such as VGG16, CaffeNet, VGG19 and Inception V3 without fine-tuning. The maximum recognition rate among these four models was 88.15% with is much lower than the recognition rate 94.8% of the proposed model. For future work, the authors will continue the research in the field of Computer Vision and for optimizing Neural Networks for complex gesture Recognition.

References

[1] W. H. O. (WHO), "Deafness and hearing loss," 2019. Available at: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (last visited on 13 June 2019).

[2] L. Kin, T. Tian, R. Anuar, Z. Yahya, and A. Yahya, "Sign Language Recognition System using SEMG and Hidden Markov Model," *Conference on Recent Advances in Mathematical Methods, Intelligent Systems and Materials*, 2013, pp. 50–53.

[3] M. P. Lewis, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the World*, 17 edn. Dallas: Sil International, 2013.

[4] E. Verdú, C. Pelayo G-Bustelo, M. A. Martínez and R. Gonzalez-Crespo, "A System to Generate SignWriting for Video Tracks Enhancing Accessibility of Deaf People," *International Journal of Interactive Multimedia and Artificial Intelligence,* vol. 4, no. 6, pp. 109-115, 2017. doi: 10.9781/ijimai.2017.09.002

[5] R.E. Mitchell, T.A. Young, B. Bachleda, M.A. Karchmer, "How many people use ASL in the United States? Why estimates need updating," *Sign Language Studies*, vol. 6, no. 3, pp. 306–335, 2006.

[6] A. Kumar, A. Kumar, S. K. Singh and R. Kala, "Human Activity Recognition in Real-Times Environments using Skeleton Joints," *International Journal of Interactive Multimedia and Artificial Intelligence,* vol. 3, no. 7, pp. 61-69, 2016. doi: 10.9781/ijimai.2016.379

[7] M. Raees and S. Ullah, "EVEN-VE: Eyes Visibility Based Egocentric Navigation for Virtual Environments," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 3, pp. 141-151, 2018. doi: 10.9781/ijimai.2018.08.002

[8] I. Rehman, S. Ullah and M. Raees, "Two Hand Gesture Based 3D Navigation in Virtual Environments," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 4, pp. 128-140, 2019. doi: 10.9781/ijimai.2018.07.001D.

[9] Aryanie and Y. Heryadi, "American sign language-based finger-spelling recognition using k-Nearest Neighbors classifier," *2015 3rd International Conference on Information and Communication Technology (ICoICT),* Nusa Dua, 2015, pp. 533-536. doi: 10.1109/ICoICT.2015.7231481

[10] M. M. Islam, S. Siddiqua and J. Afnan, "Real time Hand Gesture Recognition using different algorithms based on American Sign Language," *2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR),* Dhaka, 2017, pp. 1-6. doi: 10.1109/ICIVPR.2017.7890854

[11] C. Szegedy, W. Liu, Y. Jia, Y. *et al.*, "Going deeper with convolutions," In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 1-9. doi: 10.1109/CVPR.2015.7298594.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for largescale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe *et al.*, "Rethinking the inception architecture for computer vision," *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, USA, June 2016, pp. 2818–2826.

[14] V. Gulshan, L. Peng, M. Coram, *et al.,* "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal

fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016. doi:10.1001/jama.2016.17216

[15] B. Xie, X. He, and Y. Li, "RGB-D static gesture recognition based on convolutional neural network," *The Journal of Engineering*, vol. 2018, no. 16, pp. 1515-1520, 2018, doi: 10.1049/joe.2018.8327

[16] N. Pugeault and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition," *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Barcelona, 2011, pp. 1114-1119. doi: 10.1109/ICCVW.2011.6130290

[17] B. Estrela, G. Cámara-Chávez, M.F. Campos, W.R. Schwartz and E.R. Nascimento, "Sign language recognition using partial least squares and RGB-D information," In *Proceedings of the IX Workshop de Visao Computacional, WVC*, 2013.

[18] C. Chuan, E. Regina and C. Guardino, "American Sign Language Recognition Using Leap Motion Sensor," *2014 13th International Conference on Machine Learning and Applications*, Detroit, MI, 2014, pp. 541-544. doi: 10.1109/ICMLA.2014.110

[19] L. Rioux-Maldague and P. Giguère, "Sign Language Fingerspelling Classification from Depth and Color Images Using a Deep Belief Network," *2014 Canadian Conference on Computer and Robot Vision*, Montreal, QC, 2014, pp. 92-97. doi: 10.1109/CRV.2014.20

[20] S. Ameen and S. Vadera, "A convolutional neural network to classify American Sign Language fingerspelling from depth and colour images," *Expert Systems*, vol. 34, no. 3, 2017, e12197.

[21] Q. Dai, J. Hou, P. Yang, X. Li, F. Wang, and X. Zhang, "The Sound of Silence: End-to-End Sign Language Recognition Using SmartWatch," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking,* 2017, pp. 462-464.

[22] W. Tao, M.C. Leu and Z. Yin, "American Sign Language alphabet recognition using Convolutional Neural Networks with multiview augmentation and inference fusion," *Engineering Applications of Artificial Intelligence*, vol. 76, pp. 202-213, 2018.

[23] T. W. Chong and B.G. Lee, "American sign language recognition using leap motion controller with machine learning approach," *Sensors*, vol. 18, no. 10, 3554, 2018.

[24] K. M. Lim, A. W. C. Tan, C. P. Lee and S. C. Tan, "Isolated sign language recognition using Convolutional Neural Network hand modelling and Hand Energy Image," *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 19917-19944, 2019.

[25] J. Hou, X. Y. Li, P. Zhu, Z. Wang, Y. Wang, J. Qian, J. and P. Yang, "SignSpeaker: A Real-time, High-Precision SmartWatch-based Sign Language Translator," in *Proceedings of the 25th Annual International Conference on Mobile Computing and Networking (MobiCom '19),* Los Cabos, Mexico, 2019, article no. 24.

[26] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, vol. 1, Lake Tahoe, Nevada, 2012, pp. 1097-1105.

Manju Khari

Dr. Manju Khari is an Assistant Professor in Ambedkar Institute of Advanced Communication Technology and Research, Under Govt. Of NCT Delhi affiliated with Guru Gobind Singh Indraprastha University, Delhi, India. She is also the Professor- In-charge of the IT Services of the Institute and has experience of more than twelve years in Network Planning & Management. She holds a Ph.D. in Computer Science & Engineering from National Institute Of Technology Patna and She received her master's degree in Information Security from Ambedkar Institute of Advanced Communication Technology and Research, formally this institute is known as Ambedkar Institute Of Technology affiliated with Guru Gobind Singh Indraprastha University, Delhi, India. Her research interests are software testing, software quality, software metrics, information security, optimization, Artificial Intelligence and nature-inspired algorithms. She has 70 published papers in refereed National/International Journals & Conferences (viz. IEEE, ACM, Springer, Inderscience, and Elsevier), Besides this, she associated with many International research organizations as Editor of Springer , Wiley and Elsevier books and Guest editor of International Journal of Advanced Intelligence paradigms, reviewer for International Journal of Forensic Engineering, Inder Science, and editorial board member of International Journal of Software Engineering and Knowledge Engineering, World Scientific.

Aditya Kumar Garg

Aditya Kumar Garg is Associate Software Developer at Kuliza Technologies, Bangalore. He is currently working in backend Software development on latest technologies such as Spring Boot. He received his bachelor's degree from Ambedkar Institute of Advanced Communication Technologies and Research. His research interests include machine learning, Internet of Things, information security.

Rubén González Crespo

Dr. Rubén González Crespo has a PhD in Computer Science Engineering. Currently he is Vice Chancellor of Academic Affairs and Faculty from UNIR and Global Director of Engineering Schools from PROEDUCA Group. He is advisory board member for the Ministry of Education at Colombia and evaluator from the National Agency for Quality Evaluation and Accreditation of Spain (ANECA). He is member from different committees at ISO Organization. Finally he has published more than 200 paper in indexed journals and congresses.

Elena Verdú

Elena Verdú received her master's and Ph.D. degrees in telecommunications engineering from the University of Valladolid, Spain, in 1999 and 2010, respectively. She is currently an Associate Professor at Universidad Internacional de La Rioja (UNIR) and member of the Research Group "Data Driven Science" of UNIR. For more than 15 years, she has worked on research projects at both national and European levels. Her research has focused on e-learning technologies, intelligent tutoring systems, competitive learning systems, accessibility, data mining and expert systems.