Proc. of the 2nd International Conference on Electrical, Communication and Computer Engineering (ICECCE)
12-13 June 2020, Istanbul, Turkey

Reconstruction of Convolutional Neural Network for Sign Language Recognition

Rahib Abiyev
Applied artificial intelligence research centre, department of computer engineering
Near East University
Via mersin-10, Turkey
rahib.abiyev@neu.edu.tr

John Bush Idoko
Applied artificial intelligence research centre, department of computer engineering
Near east university
Via mersin-10, Turkey
john.bush@neu.edu.tr

Murat Arslan
Applied artificial intelligence research centre, department of computer engineering
Near east university
Via mersin-10, Turkey
murat.arslan@neu.edu.tr

***Abstract*— This paper presents a Sign Language translation model using Convolutional Neural Networks (CNN). A sign language is a language which allows mute and hearing-impaired people to communicate. It is a visually oriented, non-verbal communication which facilitates communication through body/facial postures, expressions and a collection of gestures. To contribute to the wellbeing of the affected population, we are motivated to implement a vision-based system to avert their day to day challenges. Our propose model constitutes object detection and classification phases. The first module is made up of single shot multi-box detector (SSD) used for hand detection. The second module constitutes convolutional neural network plus a fully connected network utilized to constructively translate the detected signs into text. The propose model is implemented using American sign language fingerspelling dataset. The propose system outperformed other published results in the comparative analysis, hence recommended for further exploitation in sign language recognition problems.**

Keywords—Convolutional neural networks, Single short multi-box detector, fully connected network, sign language translation

I. INTRODUCTION

Sign language is fundamentally utilized as a communication medium between people who are dumb/deaf. Gesture based communication translation/interpretation problem is an exceptionally significant research topic as a result of its capacity to build the collaboration between the individuals who are hearing-hindered in speech. Sign language is a communication medium that uses body/facial postures, expressions and a set of gestures in human-human communication, as well as through TV and social networks. Sign Language is utilized as the first language by millions of deaf people (hearing impaired people) and people with various speaking challenges. In accordance with the investigation conducted by the British Deaf Association, it is recorded that about 151,000 people utilize Sign Language as a communication medium [1]. There is no all-inclusive gesture based communication and pretty much every nation has its own national communication through signing and fingerspelling letters. They utilized a combination of manual gestures with facial mimics and lips articulation. These sign languages have a special grammar that has fundamental differences to speech-based spoken languages. The American Sign Language (ASL) is one of popular sign language, which has its own rules and grammar. There are sign systems such as Signed English that borrow signs from the ASL, but utilize them in English language order [2]. As sign language involves both reading the signs (receptive skills) and rendering the signs (expressive skills), it is a two-way process. Translation and recognition of sign language is an important research area because it integrates hearing impaired people into the general public and provides equal opportunities for the entire population.

A visual form communication of information between people called sign language is detailed and rapid. Both spelling and accurate translation of thoughts and feelings in a short time are very important. Since some people do not understand sign language, and some people usually find it very difficult to understand, it has become important to design a vision-based sign language translator. The design of such a system allows the communication barrier between people to be significantly reduced. There are two major methods for sign language translation. The first one is vision-based methods which utilize installed camera(s) to capture the target images which are in turn fed into the image analysis module [3–6]. The second approach is a glove-based approach which utilizes sensors and gloves for implementation. Here, the additional hardware (glove) is employed to mitigate the challenges of the traditional vision-based methods. Even though signers/users often find glove-based approaches to be burdening and obtrusive, they give more accurate and precise results [7, 6]. This research proposes a vision-based sign language translation approach which uses a single video camera to dynamically capture hand gestures. In this paper, the proposed sign language translation (SLT) includes three basic modules: object detection, feature extraction and classification. In solving these problems, integration of three powerful models- SSD, CNN and fully connected network is proposed. These algorithms are applied for object detection, feature extraction and classification purpose. The criteria presented for the system were robustness, accuracy, high speed. The classical approach used for Sign Language recognition is fundamentally based on feature extraction and classification. In the paper, these two stages are combined in a convolutional neural network (CNN) based structure to design the sign language translation system. The presented approach simplifies the implementation of the sign language translation system. Nowadays, CNN is actively used for solving different problems. These are human activity recognition [8], vehicle detection in aerial images [9],

978-1-7281-7116-6/20/$31.00 ©2020 IEEE

detection of smoke as a moving object [10]. Recently, several approaches have been presented for the purpose of sign language gesture identification. In the early 2000s, sensor-based approaches with neural networks and Bayesian networks were explored [11–28]. Cheap wearable technologies such as wearable sensor gloves are utilized to get the relative gesture of hands and fingers in order to predict sign language [28]. The use of constrained grammars and colored gloves produced low error rates on both training and test data [29]. Using sensor devices, a multimodal framework is applied for isolated sign language translation [30]. The sensors are used to capture finger, palm positions and then the bidirectional long short-term memory NN (BLSTM-NN) and hidden markov model (HMM) are used for classification purpose. In-depth knowledge of sign language can lead to a better understanding of the gesture classification problem. In this regard, Bheda et al. [31] addressed the gesture classification problem using a deep CNN. In some studies, the color and depth of images are utilized for recognition purposes. Here, Ameen et al. classified ASL utilizing CNN with the color and depth of images, and in their experiments, they obtained 80% recall and 82% precision rates [32].

The paper is organized as follows: Section two presents the proposed method utilized during design and implementation. Section three depicts simulation results and discussions. Comparative results analysis is presented in section four. Conclusively, conclusion and further thoughts are discussed in Section five.

II. PROPOSED METHODS

A. Single Shot Multi-Box Detector (SSD)

SSD modul is used for hand capturing, in which the captured hand serves as an input to the CNN module. SSD is a machine learning model developed in 2016 [33], in which high precision and efficiency were reported for object detection challenges. SSD operates by adding a collection of convolutional layers sequentially. By doing so, it extracts important features on various dimensions and progressively reduces the input size of each successive layer. The predefined prerogatives that closely fit the distribution of the original ground truth boxes are checked in Multi Box. In addition, such priorities are chosen in such a way that the ratio of intersection over union rate is 0.5 or narrowly above 0.5. In Fig. 1, 0.5 IOU isn't good enough, but the limiting box provides a strong starting point for the regression algorithm. Here, the multi box begins as estimators with its goals and seeks to move closer to the boundary boxes of ground reality. Conclusively, the SSD's multi box retains the top K prediction that minimizes loss of confidence and position.

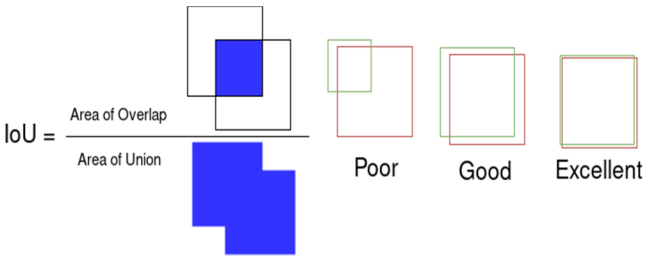

Fig. 1: Box overlapping generated by SSD structure

B. Convolutional Neural Network (CNN)

The CNNs are deep learning structure with one or several convolution layers, pooling layers as well as feedforward layers. Convolutional neural networks comprise of convolutional layers charaterized by input map(*I),* biases(*b)* and filters(*K)*. For an input image having height(*H)*, width(*W),* and channels(*C)* ((red, blue, green) $I \in R^{HxWxC}$), filters $K \in R^{k_1 x k_2 x C x D}$, biases $b \in R^D$ one for each filter, the convolutional output is computed thus:

$$(I * K)_{ij} = \sum_{m=0}^{k_1-1} \sum_{n=0}^{k_2-1} \sum_{c=1}^{C} K_{m,n,c} \cdot I_{i+m,j+n,c} + b \quad (1)$$

The convolutional and pooling layers output is calculated using:

$$O^{l}_{i,j} = pool(f(\sum_m \sum_n \omega^l_{m,n} o^{l-1}_{i+m,j+n} + b^l)) \quad (2)$$

At the end of *pooling* operation, the *flatten* is performed using $o^L_{flatten} = flatten(o^L_{i,j})$. Using the fully-connected layer $y = F(o^L_{flatten})$, the obtained feature vector automatically becomes the output of the model.

At the end of output signals computation, modelling of the weight coefficients (unknown parameters) of the convolutional neural networks is started. Let the CNN unknown parameters be θ, a precise Loss Function is designed to compute the accuracy of parameters θ. This is implemented by keeping the Loss Function at minimal using input-output training pairs$\{(x^{(i)}, y^{(i)}); i \in [1, \ldots, N]\}$,. Here $x^{(i)}$ is the *i-th* input data, $y^{(i)}$ equals the corresponding output data. Furthermore, let's consider the CNN current output to be $o^{(i)}$, we compute Loss of the CNN using:

$$L = \frac{1}{N} \sum_{i=1}^{N} l(\theta; y^{(i)}, o^{(i)}) \quad (3)$$

C. Dataset Analysis

To evaluate the proposed hybrid model, we utilized the 'Kaggle' American Sign Language fingerspelling dataset. This database is comprised of 24 classes of signs or letters. All the English alphabets are in the database except *Z* and *J*. This is because *Z* and *J* do not have static postures. Training and test sets consist of 27,455 and 7,172 items respectively. The data is provided in XLS files, and to achieve high performance, each of the images is converted into a 28x28 grayscale image. The images were updated with at least 50 + variations, as shown in the description file of the dataset. For instance, three degrees rotation, +/- 15% contrast/brightness/, 5% random pixelation, and so on. When exploring this domain, researchers face numerous difficulties as a result of these modifications, which in turn alter the resolutions of the images. Fragments of the explored data are demonstrated in Fig. 2.

Our focus during the training of this proposed hybrid model was on minimizing the loss function. We determine precise values of the parameters in the course of the

Fig. 2: Fragment of ASL fingerspelling dataset

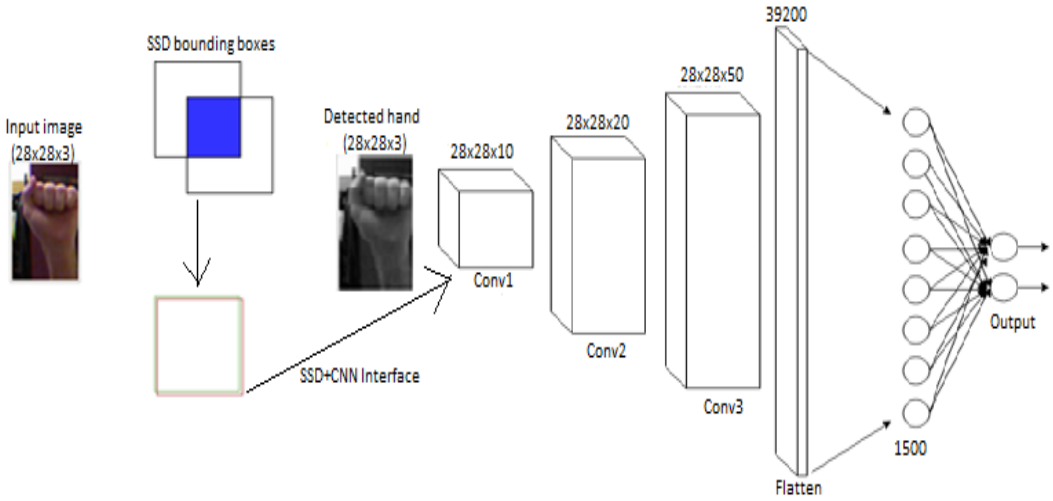

Fig. 3: Structure of the proposed hybrid network

training. In this regard, Adam optimizer is used for finding and updating network parameters [34]. Fig. 3 illustrates the simple structure of the proposed systems.

III. SIMULATION AND RESULTS

The propose network depicted in Fig. 3 is designed using the American Sign Language (ASL) fingerspelling dataset, where each of the hand images in the dataset represents a sign in ASL. The constructed hybrid model will translate a real-time finger sign into one of the 24 signs in the ASL. The system is implemented using SSD plus CNN plus FCN classifier by referencing the CNN learning technique. The input of the proposed framework is images from the camera mounted in front of the user. Here, the user's hand images are fed into the proposed hybrid model via the SSD module. The cropped hand image is fed into the input layer of the CNN module, where hand gesture features are extracted. The user's extracted hand gesture features are used as an input to the FCN. The incorporated FCN classifier uses the user's extracted hand gesture features to find the state/position of the hand of the signer. The determined state or position corresponds to a sign language, which in this case is ASL.

At the first stage, SSD module detects the hand. In the next phase, the CNN structure with a fully connected network is applied to perform ASL translation. Table 1 presents the properties of the CNN utilized for American Sign Language translation. These properties include two convolutional layers, max pooling and the fully connected layers.

During training, the data is split into two varying portions: 80% and 20%. 80% of the domain is utilized for training while the remaining 20% of the data is used for testing. From the 80% of the data assigned for training, 60% is used for training while 40% is used for validation. We utilized equations (1-3) to determine the output signals of the CNN.

During the simulation, the system specifications and size of available training data used to construct the CNN model were considered. Thus, during simulation, Z-score normalization was applied to scale each input signal and this technique enhanced the generalization of the model. Note that an RMSprop learning algorithm is employed for training. In addition, we trained the model using 150 epochs.

TABLE I. MODEL PROPERTIES

Layer type	*Output shape*	*Parameters*
conv2d_1 (Conv2D)	(None, 26, 26,16)	160
Max_pooling2d_1	(None, 13, 13,16)	0
conv2d_2 (Conv2D)	(None, 11, 11,32)	4640
max_pooling2d_2	(None, 5, 5, 32)	0
conv2d_3 (Conv2D)	(None, 3, 3, 64)	18496
max_pooling2d_3	(None, 1, 1, 64)	0
flatten_1 (Flatten)	(None, 64)	0
dense_1 (Dense)	(None, 768)	49920
dense_2 (Dense)	(None, 128)	98432
dense_3 (Dense)	(None, 24)	3096

The model includes two convolutional layers. The input size is 4096 and the kernel size is 3. Accordingly, the fully connected network is applied for American Sign Language classification purpose. As earlier stated, model was trained using 150 epochs, at every iteration of each epoch, 60% of the training set was utilized for training and 40% for validation. Fig. 4 depicts the simulation results realized for the loss function and accuracy, while Table 2 depicts the simulation results of the model. During training, the value of the loss function obtained is 1.5676e-08. The value of the loss function for validation data obtained is 0.0054, and that of the test data is 0.0054. For test data, the value of accuracy is 92.21% and the error is as low as 0.0234.

Fig. 4: Loss function and accuracy of the propose model

TABLE II. SIMULATION RESULTS OF THE MODEL

	Loss Function	**AUC (%)**	**RMSE**	**Accuracy (%)**
Training	105676e-08	100	2.2019e-05	100
Validation	0.0054	97.72	0.0234	92.22
Testing	0.0054	97.71	0.0234	92.21

IV. COMPARATIVE RESULTS

Table 3 illustrates the results and performances of renowned models used for sign language translation. Precisely, we referenced papers that presented results of accuracy. The simulations were carried out using various databases chosen by the authors. Some of the authors used American sign language dataset, some explored Indian sign language dataset, and others utilized Indonesian sign language dataset. In the introduction section, the analyzes of most of these studies are provided. In this paper, using feature extraction and classification techniques we designed a hybrid model capable of translating American sign language into text. Table 3 depicts the performances of some models that treated ASL translation problem. In the propose model, we performed transfer learning where we reused pretrained SSD and CNN models, and concatenate them with FCN classifier to perform classification on ASL dataset. As demonstrated, the proposed system (SSD + CNN + FCN) has greater performance compared with other models in Table 3. Inclusion of SSD in the second module for hand detection makes extraction of the features faster and easy. The proposed system was tested in real time and is capable of converting ASL into text in real time. The results obtained characterize the strong convergence in learning and high performance of the proposed system. Our proposed system combines detection of object, extraction of feature and classification modules which simplify the ASL translation model structure. Surely these comparative findings indicate the efficiency of the proposed hybrid system over other models aimed at solving the same problem.

TABLE III. COMPARATIVE RESULTS

Author (year)	**Method**	**Accuracy (%)**
Vaitkevičius et al. (2019) [35]	Hidden Markov classification	86.10
Bheda & Radpour (2017) [36]	deep CNN	82.50
Dong et al. (2015) [37]	Microsoft Kinect	90.00
Propose model	SSD+CNN+FCN	92.21

V. CONCLUSION

A vision-based American sign language translator that incorporates three concatenated modules- SSD, CNN and FCN is presented. The translator is trained using cross validation technique. Several experiments were performed but the results of the experiment with the best setting are presented in this paper. The averaged value of accuracy rate of the model was 92.21%, and RMSE was 0.0234. As depicted in section 3, experiments demonstrate that the presented sign language translation system performed well when classifying ASL fingerspelling texts. Experiments also showed that the sign labeling algorithm was capable of automatically detecting and differentiating the different signs within a short period of time with very high accuracy. The high performance of the proposed hybrid SLT demonstrates the robustness of the different concatenated modules. Future thought would feature application of several deep structures to same domain, to discover better and higher performing models.

REFERENCES

[1] M.A. Jalal, R.Chen, R.Moore, et al. "American sign language posture understanding with deep neural networks," In *Proc. 21st International Conference on Information Fusion* (FUSION), pp. 573-579, 10-13 Jul 2018, DOI:10.23919/ICIF.2018.8455725

[2] S.P. Becky, "Sign Language Recognition and Translation: A Multidisciplined Approach From the Field of Artificial Intelligence," *Journal of Deaf Studies and Deaf Education Advance Access*. Vol. 11, no.1, pp.94-101, 2006, DOI:10.1093/deafed/enj003

[3] Rahib H. Abiyev, "Facial Feature Extraction Techniques for Face Recognition," *Journal of Computer Science*, Vol.10, no.12, pp.2360-2365, 2014, DOI :10.3844/jcssp.2014.2360.2365

[4] C. Jennings, "Robust finger tracking with multiple cameras," In Proc. of Int. Workshop on Recognition, Analysis, and tracking of Faces and Gestures in Real-Time Systems, 1999. DOI: 10.1109/RATFG.1999.799238

[5] S. Malassiotis, N. Aifanti, and M.G. Strintzis: A Gesture Recognition System Using 3D Data, In: *Proc. of IEEE 1st International Symposium on 3D Data Processing Visualization and Transmission*, June 2002. DOI: 10.1109/TDPVT.2002.1024061

[6] B. S.Parton , "Sign Language Recognition and Translation: A Multidisciplined Approach From the Field of Artificial Intelligence," *Journal of Deaf Studies and Deaf Education*, Vol.11, no.1, pp.94-101, 2005, http://jdsde.oxfordjournals.org/cgi/content/full/11/1/94

[7] G. Fang, W. Gao, X. Chen, C. Wang, and J. Ma, "Signer independent continuous sign language recognition based on SRN/HMM," In *Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time*, pp. 90–95, July 2001, DOI: 10.1007/3-540-47873-6_8

[8] Md. Zia Uddin and Jaehyoun Kim, "A Robust Approach for Human Activity Recognition Using 3-D Body Joint Motion Features with Deep Belief Network," *KSII Transactıons on Internet and Informatıon Systems* vol. 11, no. 2, pp.1118-1133, Feb. 2017, DOI: 10.3837/tiis.2017.02.028

[9] Jiaquan Shen, Ningzhong Liu, Han Sun, Xiaoli Tao, Qiangyi Li, "Vehicle Detection in Aerial Images Based on Hyper Feature Map in Deep Convolutional Network," *KSII Transactıons on Internet and Informatıon Systems* vol. 13, No. 4, pp.1989-2011, Apr. 2019 DOI: 10.1007/s11042-016-4043-5

[10] Nguyen Manh Dung , Dongkeun Kim and Soonghwan Ro, "A Video Smoke Detection Algorithm Based on Cascade Classification and Deep Learning," *KSII Transactıons on Internet and Informatıon Systems,* vol. 12, no. 12, pp.6018-6033, Dec. 2018, DOI: 10.3837/tiis.2018.12.022

[11] S. A. K. Mehdi Y. N., "Sign language recognition using sensor gloves," *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02.*, vol. 5, pp. 2204–2206, 2002. DOI: 10.1109/ICONIP.2002.1201884

[12] S. Sidney Fels and G. E. Hinton, "Glove-Talk: A Neural Network Interface Between a Data-Glove and a Speech Synthesizer," *IEEE Transactions on Neural Networks*, vol. 4, no. 1, pp. 2–8, 1993. DOI: 10.1109/72.182690

[13] A. K. Singh, B. P. John, S. R. Venkata Subramanian, A. Sathish Kumar, and B. B. Nair, "A low-cost wearable Indian sign language interpretation system," in *International Conference on Robotics and Automation for Humanitarian Applications, RAHA 2016 - Conference Proceedings*, 2017. DOI: 10.1109/RAHA.2016.7931873

[14] Bush, I. J., Abiyev, R., & Arslan, M. (2019). Impact of machine learning techniques on hand gesture recognition. *Journal of Intelligent & Fuzzy Systems*, *37*(3), 4241-4252.

[15] Helwan, A., Uzun, D., Abiyev, R.H., Idoko, J.B.: One-Year Survival Prediction of Myocardial Infarction. International Journal of Advanced Computer Science and Applications. 8, (2017)

[16] Abiyev, R.H., Arslan, M., Gunsel, I., Cagman, A.: Robot Pathfinding Using Vision Based Obstacle Detection. (2017)

[17] Idoko, J.B., Abiyev, R., Maaitah, K.S.M.: Intelligent machine learning algorithms for colour segmentation, WSEAS Transactions on Signal Processing (13) 232-240 (2017)

[18] Idoko, J.B., Abiyev, R.H., Ma'aitah, K.S.M., Altıparmak, H.: Integrated artificial intelligence algorithm for skin detection. ITM Web of Conferences. 16, 02004 (2018)

[19] Maaitah, K.S.M., Abiyev, R.H., Idoko, J.B.: Intelligent Classification of Liver Disorder using Fuzzy Neural System. International Journal of Advanced Computer Science and Applications. 8, (2017)

[20] Kamil, D., Idoko, J.B.: Automated classification of fruits: pawpaw fruit as a case study. International conference on man–machine interactions. Springer, Cham. 365-374 (2017)

[21] Idoko, J.B., Kamil, D.: Static and Dynamic Pedestrian Detection Algorithm for Visual Based Driver Assistive System. ITM Web of Conferences. 9, 03002 (2017)

[22] Rahib H.A., Abdulkader H.: Fuzzy neural networks for identification of breast cancer using images' shape and texture features. Journal of Medical Imaging and Health Informatics, Vol.8(4), 817-825 (2018)

[23] Helwan, A., Idoko, J.B., Abiyev, R.H.: Machine learning techniques for classification of breast tissue. Procedia Computer Science. 120, 402-410 (2017)

[24] Idoko, J.B., Arslan, M., Abiyev, R.H.: Fuzzy Neural System Application to Differential Diagnosis of Erythemato-Squamous Diseases. Cyprus J Med Sci, 3: 90-7 (2018)

[25] Arslan, M., Abiyev, R.H., Idoko, J.B.: Head Movement Mouse Control Using Convolutional Neural Network for People with Disabilities. in Proc. 13th Int. Conf. on Application of Fuzzy Systems and Soft Computing, ICAFS 2018, Varsava , Advances in Intelligent Systems and Computing, 896, XIV, pp.239-248 (2018)

[26] Idoko, J.B., Arslan, M., Abiyev, R.H.: Intensive Investigation in Differential Diagnosis of Erythemato-Squamous Diseases. In Proc. 13th International Conference on Theory and Application of Fuzzy Systems and Soft Computing (ICAFS-2018), 2019, DOI: 10.1007/978-3-030-04164-9_21 (2018)

[27] Abiyev, R.H., Arslan, M.: Head mouse control system for people with disabilities, Expert Systems, DOI: 10.1111/exsy.12398 (2019)

[28] Abiyev, R.H., Maaitah, K.S.M.: Deep Convolu-tional Neural Networks for Chest Diseases Detection. Journal of Healthcare Engineering, (2018)

[29] T. E. Starner and A. Pentland, "Visual Recognition of American Sign Language Using Hidden Markov Models," *Media*, pp. 189–194, 1995.

[30] Pradeep Kumar, Himaanshu Gauba, Partha Pratim Roy, Debi Prosad Dogra, "A multimodal framework for sensor based sign language recognition," *Neurocomputing* Vol.259, pp. 21-38, 11 October 2017, DOI:10.1016/j.neucom.2016.08.132

[31] V. Bheda and D. Radpour, "Using Deep Convolutional Networks for Gesture Recognition in American Sign Language," *CoRR*, vol. abs/1710.06836, 2017. https://arxiv.org/ftp/arxiv/papers/1710/1710.06836.pdf

[32] 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. http://dx.doi.org/10.1111/exsy.12197

[33] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, Y., Berg, A.: Ssd: Single shot multibox detector. In European Conference on Computer Vision. 21–37. Springer (2016).

[34] Diederik, P.K., and Jimmy, L.B. (2015). ADAM: a method for stochastic optımızatıon. ICLR (2015)

[35] Vaitkevičius, A., Taroza, M., Blažauskas, T., Damaševičius, R., Maskeliūnas, R., & Woźniak, M., "Recognition of American Sign Language Gestures in a Virtual Reality Using Leap Motion," *Applied Sciences*, Vol.*9*, no.3, 445, 2019, doi: 10.3390/app9030445

[36] V. Bheda and D. Radpour, "Using Deep Convolutional Networks for Gesture Recognition in American Sign Language," *CoRR*, vol. abs/1710.06836,2017. https://arxiv.org/ftp/arxiv/papers/1710/1710.06836.pdf

[37] Cao Dong, Ming C. Leu and Zhaozheng Yin, "American Sign Language Alphabet Recognition Using Microsoft Kinect," In *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2015, DOI: 10.1109/CVPRW.2015.7301347