Informatics and Control Problems 42 Issue 2 (2022)
journal homepage: www.icp.az

Transfer learning for Azerbaijani Sign Language Recognition

G.G. Abdullayeva[*], **N.O. Alishzade**

Institute of Control Systems under the Ministry of Science and Education, Baku, Azerbaijan

ARTICLE INFO

Article history:
Received 12.04.2022
Received in revised form 28.04.2022
Accepted 18.05.2022
Available online 18.11.2022

Keywords:
Sign Language Recognition
Azerbaijani Sign Language
Transfer Learning
Deep Learning
Computer Vision

ABSTRACT

The goal of sign language technologies is to develop a bridging solution for the communication gap between the hearing-impaired community and the rest of society. Real-time Sign Language Recognition (SLR) is a state-of-the-art subject that promises to facilitate communication between the hearing-impaired community and others. Our research uses transfer learning to provide vision-based sign language recognition. We investigated recent works that use CNN-based methods and provided a literature review on deep learning systems for the sign language recognition (SLR) problem. This paper discusses the architecture of deep learning methods for SLR systems and explains a transfer learning application for fingerspelling sign classification. For the experiments, we used the Azerbaijani Sign Language Fingerspelling dataset and got 88.0% accuracy.

1. Introduction

Complementarily to the healthcare domain, artificial intelligence gets much closer to resolving the problems of people with disabilities. For visually impaired people, it appears as smart walking sticks [1]. For people with speaking/hearing impairment, sign language recognition comes into play. SLR is a subject making a shift in signed-based communication and brings computer vision as a bridge between conventional society and people with hearing/speaking limitations. This subject has been researched for many years, and different methods have been developed and successfully applied in various countries. Each method has its strength and limitations compared to others, and researchers are still using different methods for their sign language. Sign language recognition can be classified into isolated and continuous, depending on whether the video streams are annotated by an isolated gloss or a gloss sequence corresponding to a sentence [2].

Sign languages have their own rule set and components, both manual and nonmanual. Contrary to common belief, there is no one universal sign language; like natural languages, every community has developed its sign language as American Sign Language (ASL), British Sign Language (BSL), Brazilian Sign Language (LIBRAS), Japanese Sign Language (JSL), Arabic Sign Language (ArSL), Indian Sign Language (ISL), and so on. Those who can talk and hear frequently do not understand sign languages. Due to a lack of proficiency in spoken language, written language plays a minor role in establishing communication between hearing and speech-impaired societies and others. However, this technique is hugely sluggish in immediate and emergency face-to-face conversations. According

[*] Corresponding author.
E-mail addresses: ag_gulchin@rambler.ru (G.G. Abdullayeva), nigar.alish@isi.az (N.O. Alishzade).

www.icp.az/2022/2-08.pdf https://doi.org/10.54381/icp.2022.2.08

to the 2017 reports, there are over 30 000 hearing-impaired people in Azerbaijan [3]. They use Azerbaijani Sign Language to express their emotions, which most people do not understand. Interaction between signers and members of the general public necessitates translating sign language into a language that members of the general public can understand. The goal of our work is to establish the development of the AzSL translation system.

Deep learning is a type of learning algorithm developed to describe complex structures by combining a large number of nonlinear adjustments. The neural networks that are linked to the construction of deep neural networks are the fundamental building blocks of deep learning [4]. These methods have enabled significant advancements in sound and image processing, including face recognition, computer vision, voice recognition, automated language processing, text categorization, and a variety of other tasks.

Deep learning allows computational algorithms with multiple processing layers to learn to represent various abstracted dimensions. Deep learning detects unpredictability in large datasets by expressing how a system should modify its inner parameters, which have been used to perform a presentation in each level from the symbolization in the preceding layer, using the backpropagation technique. Deep Convolution Network (DCN) has made significant advances in processing video, pictures, audio, and speech, whereas recurrent networks have focused on sequential information such as voice and text [5]. Additionally, deep learning excels in terms of precision.

Deep learning algorithms have been improved by modern tools and tactics to the point where they can surpass human performance for several problems. Transfer learning is a popular approach in deep learning where pre-trained models are used as the starting point for computer vision [6]. This is a method where a model developed for a task is reused as the starting point for a model on another related task. This approach is effective for our problem because the selected model trained on a large image dataset and can make predictions on a similar set of classes, ensuring that the model efficiently learns to extract features from the next dataset to perform well on the problem.

The primary goal is to build a social-oriented AI technology for easier communication between these hindered people and us. The contributions are as follows: a) several preprocessing techniques have been applied to collected images to make the training process faster and simpler; b) The proposed work based on two-phase transfer learning using several architectures pre-trained on the 1) ImageNet Dataset [7]; 2) Kaggle ASL Dataset [8]; 3) After the selected model was trained on both datasets, second phase of transfer learning applied on with AzSL Dataset [9]; 4) Experiments carried out to test the trained model on the unseen data.

The rest of the paper organized as follows:

In Section 2, we talk about the main challenges of sign language technologies and refer to the related work. For the reasoning of the utilized idea, we provide a brief review of recent transfer learning work for SLR. In Section 3, we present our methodology and interpret the process and results of the training. After the evaluation of the method, a conclusion, and discussion on future work are provided.

2. Related work

The increasing amount of visual data leads to superior performance in computer vision (CV) tasks. In the case of the Sign Language Recognition (SLR) task, different methods of machine learning have been developed. Real-time SLR is a state-of-the-art subject that promises to facilitate communication between the hearing-impaired community and others. In our study, we exclude sensor-based methods and consider only appearance-based recognition.

Besides growing data volume, several factors have led many researchers to adopt deep learning techniques to improve computer vision models' performance. These factors include large volumes of widely usable multimodal datasets; more powerful computers with fast graphical processing units and tensor processing units, and high-quality feature representation at multiple scales [10]. This is also true for sign language technologies. The more computing power is accessible, the more complex methods are developing.

SLR is a tool to translate the sign language created by the disabled into a textual form that non-signers can understand. Due to the latest advancement in classification methods, several currently proposed works specifically contribute to classification methods, together with hybrid techniques and deep learning. Many reviews and survey studies provide knowledge about the overall picture of SLR as a research subject. The authors of [11] focus on classification approaches utilizing ML algorithms. Many works use SVM for classification, and some studies use Hidden Markov Models.

Authors of [12] use Random Forest, Support Vector Machine, and Gradient Boosting algorithms for the classification of the Kazakh fingerspelling alphabet. They achieved 98.86 %, 98.68 %, and 98.54 % overall test accuracy respectively for each algorithm.

In [13] authors use DeepCNN for ASL classification and emphasize hyperparameter tuning to enhance accuracy. They fine-tuned CNN that includes Conv2D layers and got the highest validation accuracy is 99.96%.

In [14] authors provided CNN-based feature extraction and classification for sign language; they used ASL letters and numerals dataset and achieved 99.82% maximum accuracy.

In [15] authors worked on the Macedonian fingerspelling alphabet and applied transfer learning. They used pre-trained architectures like MobileNet and a state-of-the-art prediction accuracy on the Macedonian sign language alphabet classification task.

Authors of [16] apply transfer learning for end-to-end sign language translation on two different datasets. They used GPT2 pre-trained model on the German Sign Language corpus and got remarkable results.

In [17] authors provide transfer learning using 3DCNN for isolated SLR. They also propose a random knowledge distillation strategy to transfer knowledge from larger models to smaller ones.

Another work with 3DCNN is a master thesis [18] where American and Chinese sign languages were used. The authors trained 3DCNN with different modalities and fused the predictions of those models using Bayesian fusion for more accurate sign classification results. They also tested how the number of signers in the dataset affects the prediction score.

[19] and [20] provide a Transfer Learning approach for the Arabic Sign Language fingerspelling alphabet. Authors of [19] used RNN architecture to boost accuracy. In [20], Keras pre-trained models with EfficientNet architecture were used.

These papers demonstrate the typical workflow of the current SLR research. As all the sign languages are similar in the core, every remarkable work towards sign language technologies is useful to consider in building a new one. We can see that several cutting-edge approaches use pre-trained models to deal with data shortage issues while also benefiting from the simplicity of the training process. In this work, we take a transfer learning approach for AzSL.

3. Methodology

Unlike other natural languages, sign languages use significant bodily motions to communicate messages, known as gestures or signs. Each gesture (symbol) represents a different letter, word, or verdict. A phrase is formed by the combination of signals, much as the string of words includes words in spoken languages. Each sign language has a subset of alphabet signs, also known as the fingerspelling alphabet. Those signs represent letters of the assigned language and help signers to express names, abbreviations, and forgotten/unknown words. We have created our approach for Azerbaijani Sign Language (AzSL) fingerspelling alphabet. We used the CNN-based training method and applied two-phase transfer learning to overcome the data shortage problem. First, we have gripped CNN-based Inception V3 model that has been trained on ImageNet Dataset. Then, we took that model trained via transfer learning with Kaggle ASL Dataset. Then, we trained this model with our AzSL Dataset. After two-phase transfer learning, we evaluated the model with test data.

3.1 Dataset

We use AzSL fingerspelling data [9] that consists of 32 classes representing 32 letters of the Azerbaijani alphabet. Each class contains 500 images. Originally, 24 letters are static, and 8 letters expressed by moving gestures, i.e. they are dynamic. We tackle this problem by using only the last frames of those letters in training.

3.2 Preprocessing

Azerbaijani Sign Language (AzSL) is a native visual communication tool for the Deaf community living in Azerbaijan. It is based on the Russian Sign language, which belongs to French Sign Languages Cluster. Those sign languages possess many common elements and have been developed their particular elements and structure as well. In our study, we use a dataset of AzSL's fingerspelling alphabet that contains 32 letters.

We used the Mediapipe library for feature extraction. Using pictorial models for signing images is very useful as it helps the model learn necessary features in hands during the signing. It retrieves keypoints for hands' positions, emphasizes slight differences between two signs, and makes the system focus on a signer rather than the background.

In each frame, the P number of points detected. The number of frames is N. Each detected point is denoted by $p^{[i,j]} = \left(p_x^{[i,j]}, p_y^{[i,j]}\right)$, where i = 1, 2, 3, …, P and j = 1, 2, 3, …, N [21]

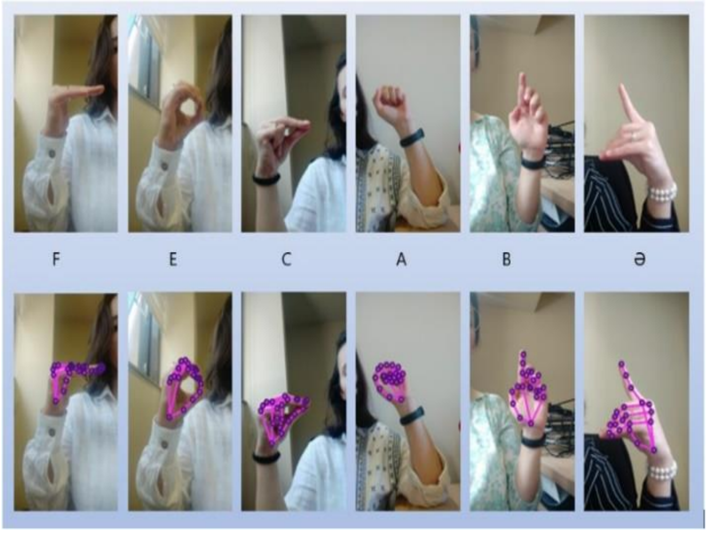


Fig. 1. Results out of Mediapipe

Fig. 1: AzSL letters with Mediapipe

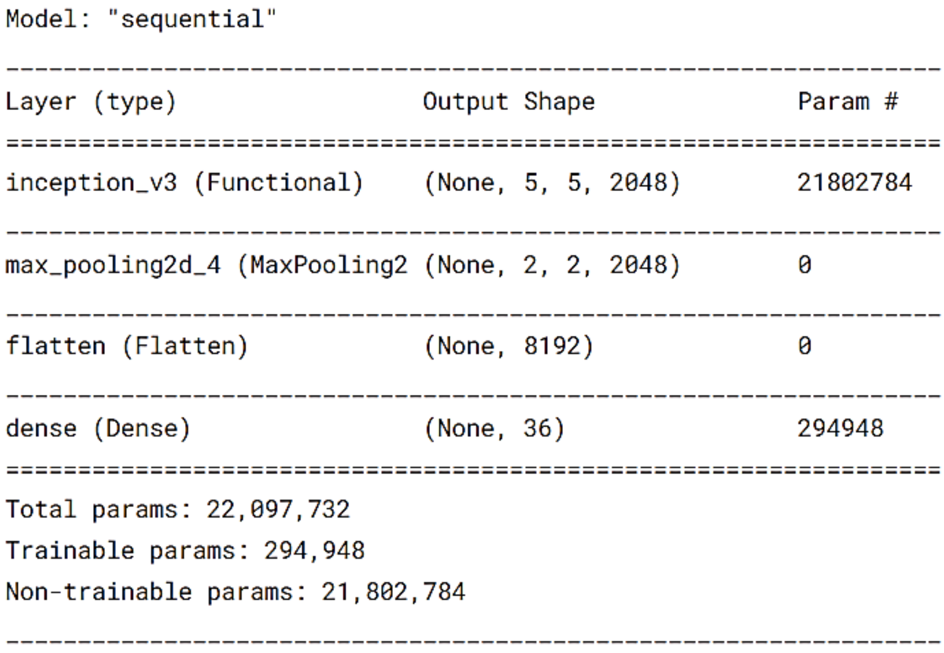
All detected points for static gestures should not change their coordinate in time. This is true for dynamic eight letters also, as we handle them as static, taking last frames in the collected videos of them. The estimator is a point $\acute{p}^{[i,j]} = \left(\acute{p}_x^{[i,j]}, p_y^{[i,j]}\right)$, where $\acute{p}_x^{[i,j]} = \frac{1}{N}\sum_{j=1}^{N} p_x^{[i,j]}$, $\acute{p}_y^{[i,j]} = \frac{1}{N}\sum_{j=1}^{N} p_y^{[i,j]}$. In our case, in every hand, 21 landmarks are pointed. The result out of our application Mediapipe is presented on Fig. 1. For several letters.

3.3 Training

Transfer learning technique provides good results for computer vision problems. This technique has two main pros: skipping the problem of training from scratch big datasets and high demand for computational resources. However, it is not only useful to overcome these shortcomings but also to transfer the knowledge gathered from other datasets to the current task. Instead of training the entire neural network from scratch, we can use the pre-trained neural model. The convolutional neural network is both a feature extractor and a classifier. Transfer Learning is an approach to training a model on one problem and re-using the obtained weights on a second related, but different problem. There are three major transfer learning methods:

- ConvNet as a feature extractor
- ConvNet fine-tuning
- Pre-trained models

The model prepared for transfer learning with ASL is shown in Fig. 2.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
inception_v3 (Functional)    (None, 5, 5, 2048)        21802784
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 2, 2, 2048)        0
_________________________________________________________________
flatten (Flatten)            (None, 8192)              0
_________________________________________________________________
dense (Dense)                (None, 36)                294948
=================================================================
Total params: 22,097,732
Trainable params: 294,948
Non-trainable params: 21,802,784
_________________________________________________________________
```

Fig. 2. The model architecture and number of parameters

Now the model is trained with two different datasets [22]. After this transfer training, we carry another transfer training process on the gathered weights of the model. Total schema of our work is shown in Fig. 3.

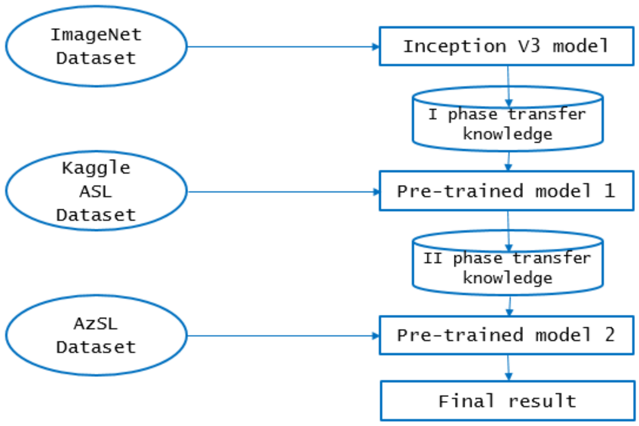


Fig. 3. Diagram of carried work

In the first two training processes, models fed with RGB image pixels. We extracted hands keypoints out of collected images, and fed the pre-trained model with their matric files (numpy fies). For this, we had to modify input shapes of the model.

3.4 Early stopping

Early Stopping is done to make sure the model fitting stops at the most optimized accuracy point. After the early stopping point, the model might start overfitting [23]. For testing purposes, this step can be skipped and complete training can be done. We used early stopping in the second phase of transfer learning as well.

3.4 Model evaluation

For the first stage of transfer learning, the model was evaluated and 99% accuracy for Kaggle ASL Dataset was obtained. In the Fig. 4. and Fig. 5. loss plot and accuracy plot is shown.

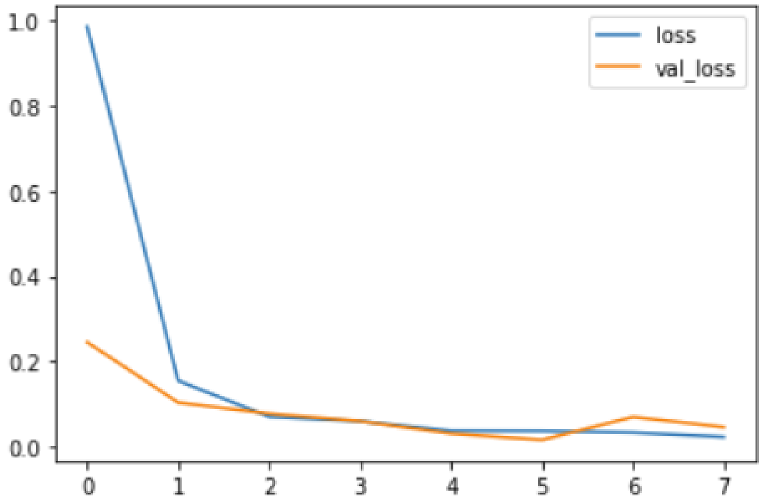


Fig. 4. Loss plot of first phase of transfer learning

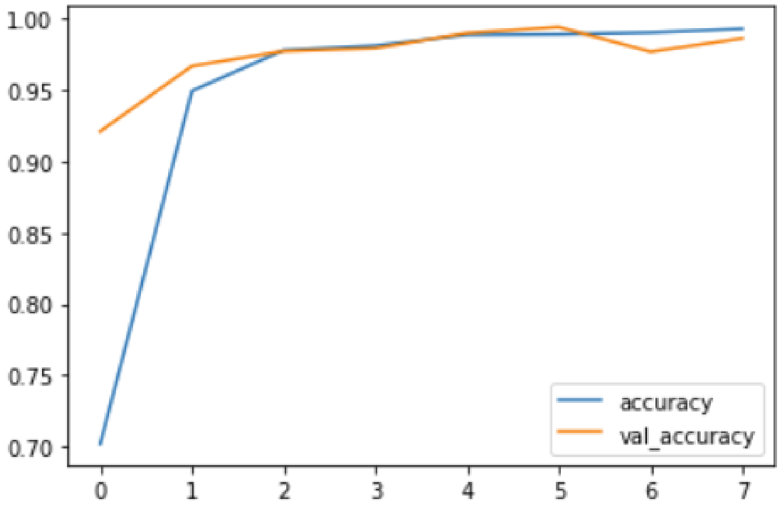


Fig. 5. Accuracy plot of first phase of transfer learning

After obtaining the pre-trained model 2, we fed it with preprocessed AzSL dataset. We evaluated the final model with unseen test dataset of 32 letters. In the Tab. 1., we show the values of different evaluation methods for each letter of Azerbaijani alphabet. Here, we conclude that overall 88.00% accuracy observed in our study.

Tab. 1.

Accuracy measurment for AzSL

Class	Precision	Recall	F1-score
A	0.50	1.00	0.67
B	1.00	1.00	1.00
C	0.50	1.00	0.50
Ç	0.50	0.44	0.50
D	1.00	1.00	1.00
E	1.00	1.00	1.00
Ə	1.00	1.00	1.00
F	1.00	1.00	1.00
G	0.67	0.50	0.67
Ğ	1.00	1.00	1.00
H	1.00	1.00	1.00
İ	0.67	1.00	0.67
I	1.00	1.00	1.00
J	0.67	1.00	1.00
K	0.67	0.67	0.67
Q	1.00	1.00	1.00
L	0.67	1.00	1.00
M	0.67	1.00	1.00
N	1.00	1.00	1.00
O	1.00	1.00	1.00
Ö	0.67	0.67	0.67
P	0.67	0.67	0.67
Q	1.00	1.00	1.00
R	1.00	1.00	1.00
S	1.00	1.00	1.00
Ş	1.00	1.00	1.00
T	0.67	0.67	0.67
U	1.00	1.00	1.00
Ü	0.67	0.67	0.67
V	1.00	1.00	1.00
Y	0.67	0.67	0.67
Z	1.00	0.67	1.00
accuracy			0.88

4 Conclusion and future directions

In this study, we talked about the sign language recognition task, and proposed two-phase transfer learning for SLR. We worked with Inception V3 pre-trained model. Firstly, it has been trained on ImageNet Dataset. Secondly, it has been trained on Kaggle ASL Dataset. We applied two-phase transfer learning and trained this model with our AzSL data. Table 1 shows that low accuracy values have been reached for the letters Ç, G, K, Ü, and Y. This is because they are actually dynamic letters. But for simplicity we handled them as static letters.

For further course of our study, we consider to develop a method to handle dynamic letters in a different way. As their static frames are very similar to some other letters, it is hard for our model to distinguish inter-class differences.

References

[1] R. Bhavani, S. Ananthakumaran. Development of a smart walking stick for visually impaired people, Turkish Journal of Computer and Mathematics Education, Vol.12 No.2 (2021), DOI:10.17762/TURCOMAT.V12I2.1112

[2] Ilias Papastratis et al., Artificial Intelligence Technologies for Sign Language, Sensors. 21 No.17 (2021) 5843. https://doi.org/10.3390/s21175843

[3] https://www.undp.org/sites/g/files/zskgke326/files/migration/az/AZ_Disability_Report_Eng.pdf

[4] F. Emmert-Streib et al., An Introductory Review of Deep Learning for Prediction Models With Big Data, Front. Artif. Intell., 28 February 2020 Sec. Machine Learning and Artificial Intelligence, https://doi.org/10.3389/frai.2020.00004

[5] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8

[6] Junyi Chai, Hao Zeng, Anming Li, Eric W.T. Ngai, Deep learning in computer vision: A critical review of emerging techniques and application scenarios, Machine Learning with Applications, Volume 6, 2021, https://doi.org/10.1016/j.mlwa.2021.100134

[7] https://www.image-net.org/

[8] https://www.kaggle.com/datasets/grassknoted/asl-alphabet

[9] https://www.kaggle.com/datasets/aykhannazimzada/azsl-dataset

[10] Junyi Chai, Hao Zeng, Anming Li, Eric W.T. Ngai, Deep learning in computer vision: A critical review of emerging techniques and application scenarios, Machine Learning with Applications, Volume 6, 2021, https://doi.org/10.1016/j.mlwa.2021.100134

[11] S. Subburaj, S. Murugavalli, Survey on sign language recognition in context of vision-based and deep learning, Measurement: Sensors, Volume 23, 2022, https://doi.org/10.1016/j.measen.2022.100385

[12] Kenshimov, C., Buribayev, Z., Amirgaliyev, Y.N., Ataniyazova, A., & Aitimov, A. (2021). Sign language dactyl recognition based on machine learning algorithms. Eastern-European Journal of Enterprise Technologies. https://doi.org/10.15587/1729-4061.2021.239253

[13] Abdul Mannan et al. Hypertuned Deep Convolutional Neural Network for Sign Language Recognition, Natural Language Processing and Human Computer Interaction, 2022. https://doi.org/10.1155/2022/1450822

[14] Barbhuiya, A.A., Karsh, R.K. & Jain, R. CNN based feature extraction and classification for sign language. Multimed Tools Appl 80, 3051–3069 (2021). https://doi.org/10.1007/s11042-020-09829-y

[15] A. Kralevska, R. Trajanov and S. Gievska, "Real-time Macedonian Sign Language Recognition System by using Transfer Learning," 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), 2022, pp. 906-911, doi: 10.23919/MIPRO55190.2022.9803692.

[16] https://doi.org/10.48550/arXiv.2203.04287

[17] Han, X., Lu, F. & Tian, G. Efficient 3D CNNs with knowledge transfer for sign language recognition. Multimed Tools Appl 81, 10071–10090 (2022). https://doi.org/10.1007/s11042-022-12051-7

[18] Sharma, S., Kumar, K. ASL-3DCNN: American sign language recognition technique using 3-D convolutional neural networks. Multimed Tools Appl 80, 26319–26331 (2021). https://doi.org/10.1007/s11042-021-10768-5

[19] Mahmoud, E., Wassif, K., Bayomi, H. (2022). Transfer Learning and Recurrent Neural Networks for Automatic Arabic Sign Language Recognition. In: Hassanien, A.E., Rizk, R.Y., Snášel, V., Abdel-Kader, R.F. (eds) The 8th International Conference on Advanced Machine Learning and Technologies and Applications (AMLTA2022). AMLTA 2022. Lecture Notes on Data Engineering and Communications Technologies, vol 113. Springer, Cham. https://doi.org/10.1007/978-3-031-03918-8_5

[20] Mohammed Zakariah, Yousef Ajmi Alotaibi, Deepika Koundal, Yanhui Guo, Mohammad Mamun Elahi, "Sign Language Recognition for Arabic Alphabets Using Transfer Learning Technique", Computational Intelligence and Neuroscience, vol. 2022, Article ID 4567989, 15 pages, 2022. https://doi.org/10.1155/2022/4567989

[21] https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html

[22] https://www.kaggle.com/code/stpeteishii/sign-language-transfer-learning

[23] https://paperswithcode.com/method/early-stopping