Turkish Sign Language Recognition Using a Fine-Tuned Pretrained Model

Gizem Ozgul[1], Şeyma Derdiyok[1](✉), and Fatma Patlar Akbulut[2]

[1] Department of Computer Engineering, Istanbul Kültür University,
Istanbul, Turkey
s.derdiyok@iku.edu.tr

[2] Department of Software Engineering, Istanbul Kültür University, Istanbul, Turkey
f.patlar@iku.edu.tr

Abstract. Many members of society rely on sign language because it provides them with an alternative means of communication. Hand shape, motion profile, and the relative positioning of the hand, face, and other body components all contribute to the uniqueness of each sign throughout sign languages. Therefore, the field of computer vision dedicated to the study of visual sign language identification is a particularly challenging one. In recent years, many models have been suggested by various researchers, with deep learning approaches greatly improving upon them. In this study, we employ a fine-tuned CNN that has been presented for sign language recognition based on visual input, and it was trained using a dataset that included 2062 images. When it comes to sign language recognition, it might be difficult to achieve the levels of high accuracy that are sought when using systems that are based on machine learning. This is due to the fact that there are not enough datasets that have been annotated. Therefore, the goal of the study is to improve the performance of the model by transferring knowledge. In the dataset that was utilized for the research, there are images of 10 different numbers ranging from 0 to 9, and as a result of the testing, the sign was detected with a level of accuracy that was equal to 98% using the VGG16 pre-trained model.

Keywords: Sign language · Convolutional Neural Networks (CNN) · Transfer learning

1 Introduction

Communication is the process of conveying the most fundamental information, such as emotions and thoughts, to the other party using a variety of means. Although communication is a multifaceted process, language is the most effective component. Because of language, humans can execute their daily tasks with relative ease. While language speeds up communication, it is inaccessible to many individuals with hearing impairments. Every country has a sign language that is unique to its language structure. However, the fact that sign language and current grammar are often dissimilar makes it challenging for hearing-impaired

A. Ortis et al. (Eds.): ICAETA 2023, CCIS 1983, pp. 73–82, 2024.
https://doi.org/10.1007/978-3-031-50920-9_6

individuals to become literate. According to the 2018 data of the World Health Organization, there were 34 million hearing-impaired individuals in Europe, and this number is projected to increase by around 12 million by 2050[1]. It has been observed that the fact that people with hearing impairments experience communication challenges has led to a rise in the number of studies aimed at resolving this issue. The advancement of artificial intelligence research, which has gained momentum in recent years, has led to a rise in sign language research [1,2]. Serious research has been undertaken [3,4] (particularly in the fields of machine learning and deep learning). Convolutional neural networks (CNN), one of the deep learning methods, are commonly employed in domains such as image classification, similarity-based grouping, and object recognition.

Communication is essential for the continued existence of humans on earth. There are two main components in any communication: the recipient and the sender [5]. During communication, a channel is formed between the transmitter and the receiver; through this channel, many acts, such as emotions and thoughts, can be transmitted to the other side. Sign language is a visual language, that is a collection of gestures, mimics, and hand and facial movements intended for hearing-impaired people to communicate. According to the Turkish Statistical Institute's (TUIK) 2015 figures[2], there are 406 thousand disabled men and 429 thousand disabled women in Turkey.

Hearing-impaired individuals can communicate effectively with the norms they have established among themselves, but they cannot interact efficiently with other individuals or institutions. This extremely difficult-to-express mechanism generates social dysfunction. They cannot communicate themselves clearly and cannot even comprehend the other party's posts. As a result, individuals with hearing loss tend to withdraw themselves from society [6]. In 2018, there were 34 million hearing-impaired people in Europe alone, according to data released by prestigious health agencies such as the World Health Organization [7]. In 32 years, or in 2050, it is expected that this data would expand by 35.29%. Even in sports, it is quite difficult to interact with hearing-impaired individuals from diverse groups when several studies are analyzed [8]. There are more than 120 sign languages in the world [9], and although they are closely related, there are still communication gaps between them. The statistics indicate that the development of digital solutions to enhance the communication of individuals with impairments is necessary [10,11]. This study proposes a model in which the numbers in Turkish sign language will be developed with the assistance of CNN in order to contribute to the stated challenge.

2 Related Work

The scientific field of sign language recognition is expanding in the field of gesture recognition. Research on the recognition of sign language has been carried out all around the world utilizing a variety of sign languages. These sign languages

[1] Available at http://www.who.int/en/data-and-evidence.

[2] Available at https://data.tuik.gov.tr.

include American Sign Language [29], Chinese Sign Language [28], Japanese Sign Language [27], Turkish Sign Language [26], etc. Numerous systems for sign language recognition employ machine learning due of its capacity to train useful models using limited and sometimes noisy sensor input. There are a variety of sensor options, including data gloves and other tracker systems, computer vision approaches employing a single camera, numerous cameras, and motion capture systems, and handcrafted sensor networks.

Approaches to representing basic units of signed languages vary significantly across researchers. The simultaneous nature of significant left-hand, right-hand, and head gestures in sign languages presents a barrier for many sequential approaches [12]. Others attempt to develop models with a structure resembling phonemes, whereas the majority of studies opt to use the sign as their modeling unit of origin. Utilizing technology means allows for the possibility of locating a solution to the problem that will remove the bottlenecks that are now in place.

Examining the studies in the scientific literature reveals that image processing [13] technologies are commonly utilized to detect human limb motions. Numerous models have been developed in this direction with the contribution of deep learning models [14,15], which have recently acquired prominence in this field. Attractiveness has been drawn to the success of deep learning systems in image processing and classification. In the study of Kemalolu and Sevli [16], for instance, convolutional neural networks (CNN), one of the deep learning techniques, are utilized to train and process an image set including Turkish sign language numbers. In the process of classifying sign languages, a considerable amount of strategies and methodologies have been presented. Pigou et al. [15] completed a deep learning investigation to describe 20 Italian sign language hand movements. In the study, the results of an artificial neural network were mixed with the CNN model. As a result of this combination structure, they attained a 91.7% success rate. Bheda et al. [17] completed another investigation utilizing the deep learning model on American sign language. They utilized a small-scale dataset that they previously developed. Consequently, while expanding the datasets and utilizing them with the CNN model, a 97% success rate has developed. Kalam et al. [18] generated a total of 7000 images by rotating 700 numerals (images) in American sign language from ten different angles, yielding a total of 7000 images. By training the dataset they created using the CNN architecture, they attained a success rate of 97.28%.

3 Material and Method

This section discusses the dataset that was used, the preprocessing techniques that were implemented, as well as the CNN and pretrained models that were developed to train using the dataset.

3.1 Dataset

In this study, Turkish sign language images obtained with the participation of 218 students studying at Ankara Ayrancı Anatolian High School were used as a dataset [19]. The dataset was created in jpeg (rgb) format to represent numbers between 0 and 9 at 100 × 100-pixel resolution. Each student was asked to create 10 different numbers from 0 to 9. In this manner, 2180 image data were obtained. Figure 1 depicts a sub-example of sign images ranging from 0 to 9.

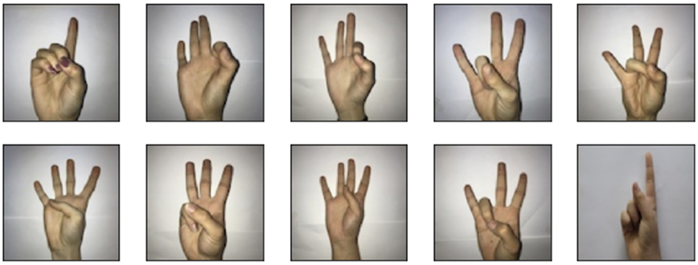

Fig. 1. Samples from Sign Language Dataset

3.2 Data Preprocessing

Red-green-blue (RGB) is the format of the study's data set. RGB (red-green-blue) channels allow for the coloring of images, although working with colored images can be challenging at times. Thus, images will be examined and analyzed in grayscale. The grayscale nature of the images renders them two-dimensional. In this scenario, image colors can have values between 0 and 255 in a single dimension. The images are normalized because the findings of investigations on one-dimensional values between 0 and 255 are often unsuccessful. Normalization identifies the minimum and maximum values of all existing numeric values in a column and reduces these values to 1. As this circumstance falls between 0 to 255 in the present investigation, the image values have been lowered from 0 to 1.

3.3 Methods

In deep learning applications, a learning model may be developed from scratch. However, transfer learning has increased the performance of models. The weights of a previously trained network can be utilized to train the initial model. When comparing these two methods in terms of performance evaluation, it has been discovered that transfer learning is quicker and more efficient. In this investigation, a model was constructed, and transfer learning techniques were utilized to train the sign language visuals. In the study, a 2D-CNN was developed, and it was fine-tuned for different pre-train models, including VGG16, ResNet50V2, EfficientNetB7, InceptionV3, and MobileNetV2, as depicted in Fig. 2. The Adam optimizer, a 0.001 learning rate (lr), and categorical cross-entropy were chosen for the optimization of the specified models. All models employ the same structure since the Adam optimizer and learning rate selections are the metrics that yield the greatest outcomes.

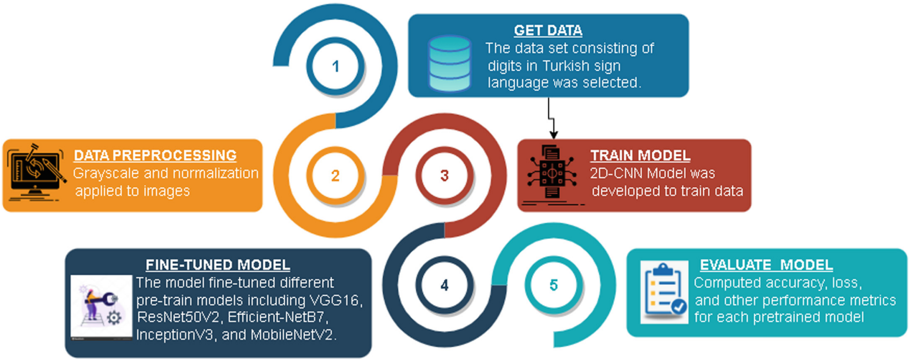
GET DATA
The data set consisting of digits in Turkish sign language was selected.
1
DATA PREPROCESSING
Grayscale and normalization applied to images
2
3
TRAIN MODEL
2D-CNN Model was developed to train data
FINE-TUNED MODEL
The model fine-tuned different pre-train models including VGG16, ResNet50V2, Efficient-NetB7, InceptionV3, and MobileNetV2.
4
5
EVALUATE MODEL
Computed accuracy, loss, and other performance metrics for each pretrained model

Fig. 2. A graphical representation of the concept of the research methodology.

4 Experimental Results

The aim of this study is to classify sign language gestures consisting of numbers and to further increase the performance of training with transfer learning. In the initial phase of our classification efforts, a two-dimensional CNN model was developed. After the initial 2 convolution layers, the max pooling layer was added, followed by 2 further convolution layers. It was then leveled by going through a layer of maximum pooling. Three dense layers have been traversed to reach the final layer. Ten different classes are predicted by training the last layer using the softmax activation function. Following model training, 86% accuracy was determined. The model's confusion matrix is depicted in Fig. 3. According to the basic model, the distortion rate in the images between the layers was not significant and was found to be normal.

In the second part of our experiment, the model was fine-tuned using the most prominently pretrained models from the literature [25]. We applied the several CNN architectures such as VGG16, ResNet50V2, EfficientNetB7, InceptionV3, and MobileNetV2, each of which offered distinct capabilities. *VGG16* is a convolutional neural network model developed in 2014 by a University of Oxford working group with the same name [20]. As the name suggests, there are sixteen distinct layers. When training our own dataset using the VGG16 architecture's weights, 98% accuracy is attained. In the VGG16 model, the rate of distortion and loss in the images between the layers was quite high without fine-tuning. When the fine-tuning is applied to the model the rate of distortion in images decreased. It is depicted in Fig. 4.

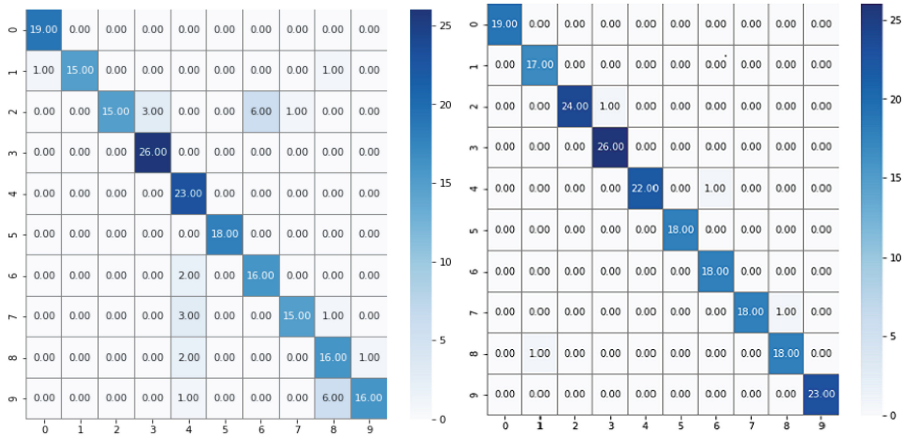

Fig. 3. The Confusion Matrix of a) base and b) fine-tuned VGG16 models

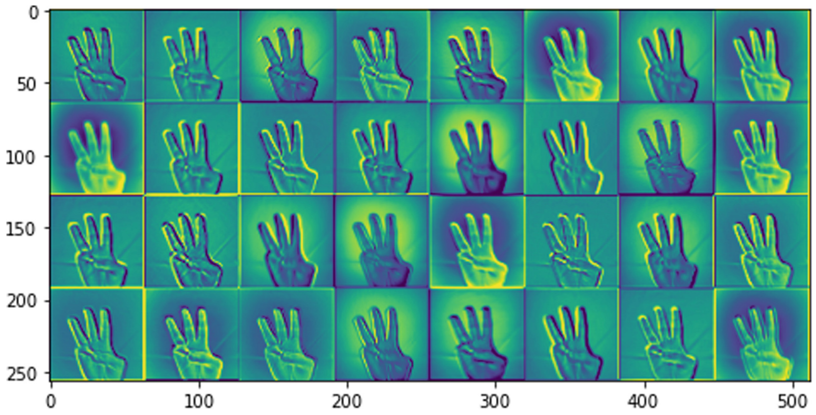

Fig. 4. Visualization of high-level feature map from *conv*2$d_5$7 layer of the fine-tuned VGG16 model using samples from the dataset.

ResNet50V2 is another pretrained model created for imagenet database classification that is designed by Microsoft [21]. There are fifty layers. Its success percentage on the dataset of sign language remained at 89%. *EfficientNetB7*, is also a pretrained model built by Google [22], which can be classified into eight different architectures. It has evolved consistently from B0 to B7. In these cases, the EfficientNetB7 architecture enables more effective training. The proper classification success percentage for the dataset of sign language was determined to be 90%. *InceptionV3* is a model developed by Google with 50 deep layers [23]. It has the ability to classify nearly 1,000 objects using ImageNet weights. The first input size of this network is 299 × 299 pixels. When we pre-train our sign language dataset with the InceptionV3 model, the obtained accuracy value is 97%.

MobileNetV2 is a kind of convolutional neural network developed by Google for mobile display applications [24]. It is a model that uses a limited number of resources. It offers precise validation for tiny datasets. However, it has become clear in our experience that it is not suitable for a sign language dataset. The obtained accuracy value could not exceed 21%.

We adjusted models for our problem by adding a new fully connected layer for each of the 10 classes in our dataset. Backpropagation was then used to fine-tune the original CNN filter weights acquired from natural images such that they more accurately mirrored the modalities in the sign language dataset. It was decided that the VGG16 had the best performance out of all the models. The training and validation error for the 10 epochs of fine-tuned and base models are depicted in Fig. 5. The training error rates for both CNNs follow a consistent trend of a steady decline followed by a plateau. The similarity between the training and validation curves indicates that the proposed fine-tuned model did not overfit the training data.

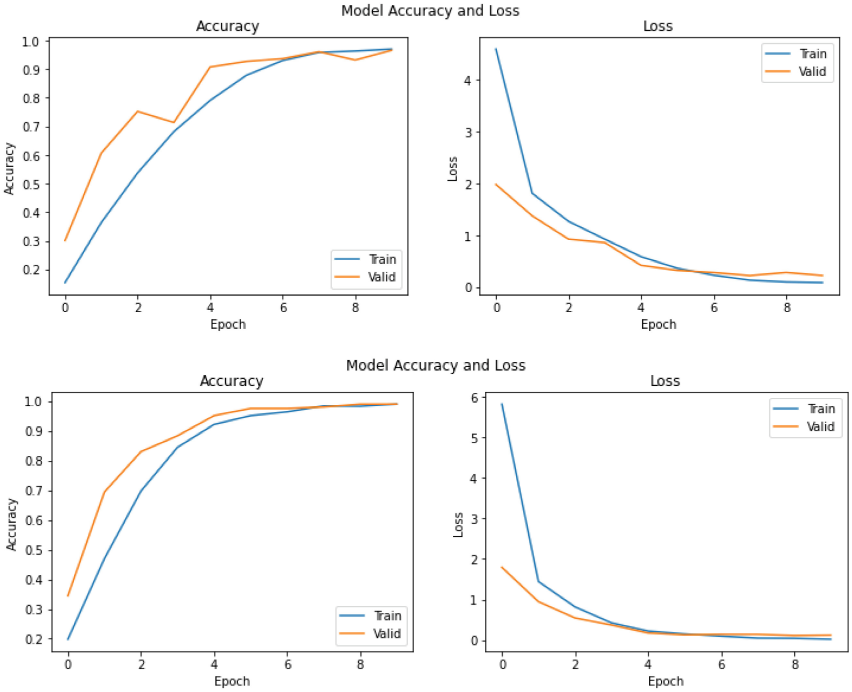

Fig. 5. The accuracy and loss scores of a) base and b) fine-tuned VGG16 models

In Table 1, all results are presented in a comparable manner.

Table 1. Classes-based classification performance of base and fine-tunes models.

Classes	VGG16	ResNet50V2	InceptionV3	MobileNetV2	EfficientNetB7	2DCNN
0	**1.0**	**1.0**	**1.0**	**0.0**	**1.0**	**1.0**
1	**1.0**	**1.0**	**0.82**	**0.00**	**0.94**	**0.88**
2	**0.96**	**0.84**	**0.92**	**0.48**	**0.88**	**0.60**
3	**1.0**	**1.0**	**1.0**	**0.00**	**0.96**	**1.0**
4	**0.95**	**0.95**	**0.91**	**0.04**	**0.86**	**1.0**
5	**1.0**	**0.94**	**1.0**	**0.05**	**0.72**	**1.0**
6	**1.0**	**1.0**	**0.83**	**0.72**	**0.83**	**0.88**
7	**0.94**	**0.73**	**0.68**	**0.00**	**0.63**	**0.78**
8	**0.94**	**0.84**	**0.84**	**0.00**	**0.57**	**0.84**
9	**1.0**	**0.82**	**0.86**	**0.00**	**0.86**	**0.69**
Avg	**0.98**	**0.91**	**0.89**	**0.13**	**0.84**	**0.86**

5 Conclusions

Despite the fact that sign language was developed to aid hearing-impaired individuals in talking with others, it is obvious that they continue to struggle with communication in society. To cover all aspects of sign languages, powerful algorithms that reliably extract characteristic features in uncontrolled contexts were developed. In this research, we present a CNN-based architecture for the classification of sign language gestures. The CNN model has a two-dimensional structure. VGG16, ResNet50V2, EfficientNetB7, InceptionV3, and MobileNetV2 were also trained using a pre-trained model to improve performance and decrease training time. We have observed that transfer learning allows for the creation of more reliable systems. The suggested model outperforms prior state-of-the-art classifiers on average with a recognition rate of 98%. The intriguing results of this study can be used as a starting point for further research into how to recognize complex hand and face movements.

References

1. Sertkaya, M., Ergen, B., Togacar, M.: Diagnosis of eye retinal diseases based on convolutional neural networks using optical coherence images. In: 2019 23rd International Conference Electronics, pp. 1–5. IEEE (2019)
2. Altuntas, Y., Comert, Z., Kocamaz, A.: Identification of haploid and diploid maize seeds using convolutional neural networks and a transfer learning approach. Comput. Electron. Agric. **163**, 104874 (2019)
3. Koller, O., Ney, H., Bowden, R.: Deep learning of mouth shapes for sign language. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 85–91 (2015)
4. Huang, J., Zhou, W., Li, H., Li, W.: Sign language recognition using 3D convolutional neural networks. In: 2015 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2015)

5. Elmas, N.: Örgütsel iletisimin is tatmini üzerindeki etkisi ve bir uygulama. Master Thsesis, Istanbul Ticaret University (2017)
6. Yildiz, Z., Yildiz, S., Bozyer, S.: Isitme Engelli Turizmi(Sessiz Turizm): Dunya ve Turkiye Potansiyeline Yonelik Bir Degerlendirme. Suleyman Demirel University Vizyoner Dergisi **9**(20), 103–117 (2018)
7. Campbell, L.: Ethnologue: languages of the world. JSTOR (2008)
8. Togacar, M., Comert, Z., Ergen, B.: Siyam Sinir Aglarini Kullanarak Turk Isaret Dilindeki Rakamlarin Tanimlanmasi. Dokuz Eylul University Muhendislik Fakültesi Fen ve Muhendislik Dergisi **23**(68), 349–356 (2021)
9. Haualand, H.: The Two Week Village-The Significance of Sacred Occasions. Disability in Local and Global Worlds. University of California Press, Berkeley (2003)
10. Murray, J.: Coequality and transnational studies: understanding deaf lives. Open Your Eyes Deaf Stud. Talk. **100**, 110 (2008)
11. Wang, H., Leu, M., Oz, C.: American sign language recognition using multidimensional hidden Markov models. J. Inf. Sci. Eng. **22**(5), 1109–1123 (2006)
12. Patlar, F., Akbulut, A.: Triphone based continuous speech recognition system for Turkish language using hidden Markov model. In: 12th IASTED International Conference in Signal and Image Processing, pp. 13–17 (2010). https://doi.org/10.2316/P.2010.710-059
13. Shanableh, T., Assaleh, K.: User-independent recognition of Arabic sign language for facilitating communication with the deaf community. Digit. Signal Process. **21**(4), 535–542 (2011)
14. Cömert, Z., Kocamaz, A.F.: Fetal hypoxia detection based on deep convolutional neural network with transfer learning approach. In: Silhavy, R. (ed.) CSOC2018 2018. AISC, vol. 763, pp. 239–248. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-91186-1_25
15. Pigou, L., Dieleman, S., Kindermans, P.-J., Schrauwen, B.: Sign language recognition using convolutional neural networks. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8925, pp. 572–578. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16178-5_40
16. Kemaloglu, N., Sevli, O.: Evrisimsel Sinir Aglari ile Isaret Dili Tanima. In: Proceedings on 2nd International Conference on Technology and Science, pp. 942–948 (2019)
17. Bheda, V. and Radpour, D.: Using deep convolutional networks for gesture recognition in American sign language. arXiv preprint arXiv:1710.06836 (2017)
18. Kalam, M., Mondal, M., Ahmed, B.: Rotation independent digit recognition in sign language. In: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1–5. IEEE (2019)
19. MAvi, A.: A new dataset and proposed convolutional neural network architecture for classification of American sign language digits. arXiv preprint arXiv:2011.08927 (2020)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
22. Tan, M., Le, Q.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
23. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)

24. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.H.: MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2015)
25. Kocacinar, B., Tas, B., Akbulut, F., Catal, C., Mishra, D.: A real-time CNN-based lightweight mobile masked face recognition system. IEEE Access **10**, 63496–63507 (2022)
26. Yirtici, T., Yurtkan, K.: Regional-CNN-based enhanced Turkish sign language recognition. Signal Image Video Process. 1–7 (2022)
27. Brock, H., Farag, I., Nakadai, K.: Recognition of non-manual content in continuous Japanese sign language. Sensors **20**(19), 5621 (2020)
28. Zhang, J., Zhou, W., Xie, C., Pu, J., Li, H.: Chinese sign language recognition with adaptive HMM. In: 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2016)
29. Bantupalli, K., Xie, Y.: American sign language recognition using deep learning and computer vision. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 4896–4899. IEEE (2018)