Sign Language Recognition from Digital Videos Using Deep Learning Methods

Jia Lu, Minh Nguyen, and Wei Qi Yan
Auckland University of Technology, Auckland 1010, New Zealand
{jia.lu, minh.nguyen, weiqi.yan}@aut.ac.nz

Abstract. In this paper, we investigate the state-of-the-art deep learning methods for sign language recognition. In order to achieve this goal, Capsule Network (CapsNet) is proposed in this paper, which shows positive result. We also propose a Selective Kernel Network (SKNet) with attention mechanism in order to extract spatial information. Sign language as an important means of communications, the problems of recognizing sign language from digital videos in real time have become the new challenge of this research field. The contributions of this paper are: (1) The CapsNet attains the accuracy of overall recognition up to 98.72% based on our own dataset. (2) SKNet with attention mechanism is able to achieve the best recognition accuracy 98.88%.

Keywords: Sign language recognition · Deep learning · Selective Kernel Network (SKNet) · Attention mechanism · Capsule network (CapsNet) · Long short-term memory (LSTM)

1 Introduction

Gesture plays an important role in our daily conversation. Sign language is an organized form of gestures, including lip movements and gestures. Moreover, it uses symbols in visual space instead of oral communications and voice patterns.

Sign language recognition is regarded as a part of behavior recognition, it is a cooperative research field which is related to pattern recognition, computer vision, etc. The term of sign language recognition refers to the whole process of tracking human gestures, recognizing the representations, and converting them into semantically meaningful commands [1]. For the traditional approaches in sign language/gesture recognition, it includes three stages: Detection, tracking, and recognition to complete the most of hand talks [1–5]. Moreover, hand detection and segmentation of the corresponding image area is the primary task of the gesture recognition system. This segmentation is critical because it separates the task-related data from the image background and then passes it to the subsequent tracking and recognition stage. Therefore, the traditional approaches are time consuming, required multiple pre-processing steps.

In this paper, we implement and investigate multiple sign language recognition models, which are time efficiency and achieve high performance in model training and testing.
Moreover, the model has shown the capabilities to implement sign language recognition, how to make the model more stable and robust for sign language recognition has become a new challenge of this study. In summary, this paper aims to develop an accurate deep learning method to conduct sign language recognition without any interruptions. As the outcome of this paper, we hope to achieve over 90% accuracy for sign language recognition in real time.

Our contribution of this paper is to adopt the deep learning-based model to achieve sign language recognition by using capsule network (CapsNet). Moreover, an RTX 2080Ti GPU is used to accelerate the training process so as to achieve the efficiency.

The remaining parts of this paper are organized as follows: The related work will be described in Sect. 2. Our method will be explicated in Sect. 3. In Sect. 4, we showcase our experimental results. Finally, the conclusion and future work of this paper will be delineated and envisioned in Sect. 5.

2 Related Work

In the past decades, with the growth of computing capacity and computational speed, deep learning methods with its derived neural networks have attracted wide attention in visual object detection, which have opened up a new era of computer vision [6–9]. As the state-of-the-art technology, deep learning (DL) becomes more and more popular because of its superiority to conventional machine learning. Deep learning methods [10, 11] were implemented not only for vision-based object detection, but also for text-based natural language processing (NLP) as well as speech recognition. Moreover, deep learning as an end-to-end model normally does not require low-level processing which is able to cut off human labor and gain time efficiency though the training is costly.

The deep neural networks (DNNs) [11, 12] encapsulate hidden layers, the pretraining method is utilized to resolve the problems of optimal local solution. In the neural network model, the number of hidden layers increases with the “depth”. The most advanced methods rely heavily on artificial neural networks, such as convolutional neural networks (CNNs), single shot multibox detector (SSD), you only look once (YOLO) etc. In addition, deep learning includes both supervised and unsupervised learning [13]. Apparently, the work clearly unfolds the differences between deep neural networks and shallow neural networks in various aspects [14].

CNN as a type of DNNs [15] is derived by combining digital image processing and artificial neural networks. The traditional CNN includes multiple convolutional layers and pooling operations, the outputs of convolutional layers are extracted as the feature maps which are flattened and fed to fully connected layer. As an active research area in computer vision, sign language recognition has successfully exploited by adopting CNN and achieved the outstanding results [7, 16].

2D CNN based on the single frame has been employed to extract feature maps and conduct temporal information to recognize gestures [17, 18]. Moreover, 2D CNN is expanded to 3D CNN [19–21] so as to learn the motion features by adopting 3D filters in the convolutional layers, which show the positive results for recognizing hand gestures. CNN model was proposed to detect and segment hands in both unlabeled and synthetic dataset, which achieved 82% accuracy based on segmentation and detection [22].
The CNN network has been well investigated to solve the image classification and recognition tasks. Moreover, it also has been investigated and implemented for sign language recognition in recent years. A CNN-based method was proposed with Gaussian skin color model and background subtraction to achieve gestures recognition from the camera images. The Gaussian skin color model controlled the influence of light on skin color, and the non-skin color of image is filtered out directly, which has 93.80% accuracy from a given dataset [23]. A two-stage CNN architecture (HGR-Net) was given, where the first stage was proposed to determine the region of interest by performing pixel-level semantic segmentation, the second stage is to recognize hand gesture [24]. Moreover, the combination of fully convolutional residual network with spatial pyramid pooling was adopted at the first stage, the result shows that proposed architecture improves 1.6% accuracy for recognition by using OUHands dataset.

A deep convolutional network was proposed with multidimensional feature learning approach (MultiD-CNN) to recognize the gestures from the RGB-D videos [25]. The method took use of 3D ResNet for training a model with both spatiotemporal features, the long short-term memory (LSTM) for processing temporal dependencies and the proposed method is outperformed compared with the previous methods based on different datasets. Chen et al. implemented the spatiotemporal attention with dynamic graph constructed (DG-STA) method to achieve hand gesture recognition. It took advantage of fully connected graph and self-attention mechanism to learn the node features and edges from the hand skeleton, a novel spatiotemporal mask is applied to reduce the computational cost. According to the experimental results, DG-STA method achieved the superior performance compared with others for recognizing hand gestures [26].

A deep-learning-based method was proposed by adopting two ResNet CNNs and soft attention with fully connected layer to recognize dynamic gestures. Moreover, a method was proposed to condense a digital video into a single RGB image and passed to the model for the final classification. The experimental result based on public datasets shows that the proposed method is able to improve the accuracy compared with other methods [27]. Three representations of depth sequences are constructed, which includes dynamic depth images (DDI), dynamic depth normal images (DDNI), and dynamic depth motion normal images (DDMNI) from the depth maps to capture the spatiotemporal information by adopting the bidirectional rank pooling, the CNNs-based model is considered to achieve gesture recognition. The proposed model was evaluated based on large-scale isolated gesture recognition at the ChaLearn LAP challenge 2016 and the model was achieved the growth of 16.34% accuracy on the IsoGD dataset [28].

Two different deep learning methods were fused to achieve gesture recognition. The convolutional two-stream consensus voting network (2SCVN) to explicitly simulate the short-term and long-term structures of RGB sequences, and 3D Depth-Saliency CNN stream (3DDSN) was used to present the motion features. The proposed methods have been evaluated based on ChaLearn IsoGD dataset with 4.47% growth of accuracy compared with other models in 2016 [29]. Molchanov et al. designed a dynamic hand gesture recognition method by adopting a recurrent 3D CNN model. Four kinds of visual data were fused to boost the recognition rate, which includes RGB, depth, optical flow and stereo IR. The proposed model achieved the positive accuracy rate based on ChaLearn dataset, which has 1% growth compared with other models [20]. A hand
gesture recognition and identification model was proposed based on the two-stream CNNs, the depth map and optical flow as the inputs were utilized in this method. The proposed model has 18.91% accuracy improvement based on MSR Action3D dataset compared with the relevant models [18].

Rastgoo et al. set forth the model for hand sign language recognition by utilizing the restricted Boltzmann machine (RBM) for visual data. The model took use of RGB and depth as the input: Original image, cropped image, and noisy cropped image. The CNN is used to detect the hand in each image, three forms of the detected hand images are generated to the RGB and depth will be inputted to the RBM. The output of the RBM will be fused to recognize the sign label. As the result, the proposed model has been able to achieve significant improvement based on four different public datasets compared with the state-of-the-art models [30]. After the RBM model, Rastgoo et al. proposed a deep cascaded model for sign language recognition from the videos in 2020. The model employed three spatial features: Hand features, extra spatial hand relation (ESHHR), and hand pose (HP) features which were fused in the model and feed into the LSTM for temporal feature extraction. The SSD model was also adopted for hand detection. The proposed model was evaluated based on IsoGD dataset, which achieved 4.25% accuracy improvement compared with others [31].

3 Our Method

CNN is good at capturing the existence of features, because the convolution structure was designed for this purpose. However, CNN is unable to explore the relationship between each feature attributes, such as relative positions, size, and direction of the feature etc. Thus, Sabour et al. firstly proposed a new deep learning network that is much effective for image processing, which called capsule network (CapsNet) [32]. It combines the advantages of CNN structure and takes into account the relative position, angle and other information that are missing from the CNN, thereby improving the recognition effect. The CapsNet structure contains two main components, which are primary capsule and digital capsule.

Capsule is a group of neurons whose input and output vectors represent the parameters of a specific type like the probability of occurrence of objects, conceptual entities, etc. It took use of the length of input and output vectors to represent the probability of the existence of the entity, the direction of a vector represents the instantiation parameters. Capsules of the same level predict the instantiation parameters of higher-level capsules through the transformation matrix. The higher-level capsules become active, if the multiple predictions are consistent. The activity of the neurons in an active capsule represents the various attributes of the specific entities that are appearing in the image, which contains the parameters, such as posture (position, size, direction), deformation, speed, reflectivity, color, texture, etc. The lengths of the output vectors represent the probability of an entity, its range is between [0, 1]. A nonlinear function called “squashing” was proposed to ensure that the length of the short vector is reduced to almost zero, while the length of the long vector is compressed. The following equation is the
expression of the non-linear function:
\[
V_j = \frac{\|S_j\|^2}{1 + \|S_j\|^2} S_j
\]

where \(V_j\) represents the vector output of capsule \(j\), \(S_j\) denotes the total input. Moreover, the total input \(S_j\) is a weighted sum over all predicted vectors from the capsules, which acquired by multiplying the vector output \(u_i\) from capsule in the layer below by the weight matrix \(W_{ij}\).

\[
S_j = \sum_i c_{ij} \hat{u}_{j|i}
\]

where \(c_{ij}\) represents coupling coefficients, which are updated and determined iteratively by the dynamic routing process.

A spatial attention-based model SKNet is put forward \cite{33} in this paper to recognize sign language, which explicates the positive results than previous deep learning models. Regarding the spatial attention, its focus is on where the information is. The Eq. (3) shows how to calculate the spatial attention:

\[
M_s(F) = \sigma \left( f^{7\times7} \left( [\text{AvgPool}(F); \text{MaxPool}(F)] \right) \right)
\]

where \(\sigma(\cdot)\) is the sigmoid function, \(f^{7\times7}\) is a convolution operation with the filter size of \(7 \times 7\). The spatial attention module utilizes average-pooling and max-pooling operations along the channel axis and concatenates them to generate an efficient feature descriptor. Consequently, a convolution operation with a \(7 \times 7\) filter is applied to produce the feature maps, a sigmoid function for normalization is offered to yield the final feature maps.

A highly modularized deep learning network (ResNeXt) was recommended for image classification. It raises up accuracy without ramping up the complexity of the deep learning method, meanwhile it effectively cuts off the number of hyperparameters \cite{34}. The ResNeXt was motivated from the idea of VGG stacking blocks of the same shape and the split-transform-merge idea of the Inception model, which holds robust scalability and is able to meliorate the accuracy without substantially altering the complexity of the model. ResNeXt is similar to Inception, the both follow the “split-transform-merge” paradigm. However, in ResNeXt, the outputs of the paths are combined by addition operations, in Inception module, they are deeply concatenated. Each path in Inception module has bias from each other, while in the ResNeXt module, all paths follow the same topology. It replaces the three-layer convolution block of original ResNet model by using a parallel stack of blocks with the same topology, which uplifts the accuracy without significantly raising the number of the parameters.

Simultaneously, the hyperparameters are reduced because of topological structure of ResNeXt model. Moreover, cardinality in ResNeXt is the size of the set of transformation and an essential factor to the dimensions of depth and width. By adopting ResNeXt model, the training error is much lower compared with the ResNet parameters. Moreover, by extending the cardinality, the model is much efficient compared with only extending the network depth or width on the ResNet model, thus dips 1.6% error rate.
In sign language recognition, most of researchers investigated both spatiotemporal information to extract the motion features in time sequences. Thus, LSTM is considered in this paper. We take use of LSTM to acquire the temporal information from video frames and predict sign language. Moreover, CapsNet and LSTM are combined together in the end by adopting the class score fusion to achieve sign language recognition in this paper.

4 Experimental Results

In this paper, we create our own sign language dataset for the purpose of model testing and validation in this project. Our dataset contains nine video footages of four classes with the tags: Hello, Nice, Meet, You, which includes the sign language data captured by ourself with the static camera. The resolution of this dataset is $960 \times 564$. The dataset contains 3,596 frames in total, there are 2,500 frames chosen for model training, 1,096 frames were selected for model testing. Figure 1 shows the example of our own dataset.

Our focus of this paper is mainly on the proposed deep learning methods and its impact on our result outcomes. We chiefly took use of three different state-of-the-art deep learning methods to fulfil the sign language recognition. Moreover, an attention mechanism based on the SKNet and ResNeXt was verified in this paper. In Fig. 2, we demonstrate the result of sign language recognition from the video frames.

Figure 3 shows the training and validation losses by using adopted deep learning methods, especially SKNet, ResNeXt and its attention mechanism. From Fig. 3, we see that SKNet models are able to achieve 97.95% accuracy. Moreover, by combining the attention model with the SKNet net, the accuracy reaches to 98.88%. Compared with these two models, the accuracy grows 0.93%. ResNeXt individually achieved 97.82% accuracy with the assistance of our dataset; ResNeXt after combined with attention mechanism is able to earn 98.19% accuracy. In these experiments, it required large amount of computations, we chose the batch size 8 and learning rate 0.001. Moreover, the number of the epoch is set to 60. In Fig. 3, the green dots represent the training and validation accuracy, the red dots stand for the training and validation loss, and the x-axis denotes the number of epochs, the y-axis represents the accuracy/loss values.
Figure 4 shows the training/testing accuracy and loss by using CapsNet for sign language recognition. The CapsNet individually achieved 98.72% accuracy with the assistance of our dataset. Compared with the previous SKNet method, the accuracy grows 0.77%. Moreover, the orange line represents the training set; the blue line stands for the testing set; where the x-axis represents the training/test steps, the y-axis represents the training/test values. In our experiment, the number of iterations is set to 10,000, the batch size is 8, and the learning rate is 0.001.
Throughout our experiments, we took use of various deep learning methods to compare our experimental results. The deep learning models with attention mechanism are much stable and robust in sign language recognition. Table 1. shows the comparison of our deep learning models on sign language recognition by adopting our own dataset.

Figure 5 exhibits the sign language recognition of training/validation accuracy and loss by using LSTM, where the blue line denotes the training accuracy; the orange line denotes the training loss; the black dots denotes the training/validation accuracy and loss; the x-axis represents the number of iterations and the y-axis represents the values of the accuracy and loss. The LSTM attained 99.56% accuracy with the assistance of our dataset. In this experiment, due to the small dataset we use, the number of iterations is set to 380, the batch size is 4, and the learning rate is 0.001. We adopted CapsNet + LSTM model with class score fusion to fulfil this study.

In Table 1, SKNet net with attention mechanism shows positive results for sign language recognition. The network of SKNet with attention mechanism is able to get 98.88% accuracy which has 5.19% growth of the total accuracy compared with the
Table 1. The different deep learning methods on sign language recognition

| Our dataset       | Hello   | Nice    | Meet    | You     | Accuracy |
|-------------------|---------|---------|---------|---------|----------|
| DenseNet          | 95.23%  | 93.82%  | 95.28%  | 94.11%  | 94.61%   |
| ResNet            | 94.77%  | 93.27%  | 94.89%  | 91.83%  | 93.69%   |
| YOLOv3            | 94.35%  | 95.26%  | 96.71%  | 97.12%  | 95.86%   |
| YOLOv4            | 96.37%  | 97.49%  | 97.55%  | 98.35%  | 97.44%   |
| ResNeXt           | 96.87%  | 98.64%  | 98.25%  | 97.52%  | 97.82%   |
| ResNext + Attention | 97.63% | 98.66%  | 98.36%  | 98.11%  | 98.19%   |
| SKNet             | 97.89%  | 97.02%  | 97.93%  | 98.96%  | 97.95%   |
| SKNet + Attention | 98.91%  | 98.64%  | 98.93%  | 99.04%  | 98.88%   |
| YOLOv3 + LSTM     | 97.28%  | 97.79%  | 95.33%  | 95.28%  | 96.42%   |
| CapsNet           | 98.76%  | 98.96%  | 98.94%  | 98.22%  | 98.72%   |
| CapsNet + LSTM    | 99.24%  | 98.53%  | 99.45%  | 98.62%  | 98.96%   |

traditional deep learning Resnet model. CapsNet for sign language recognition reaches 98.72% accuracy. Moreover, LSTM was adopted to extract the temporal information from the video frames and the CapsNet was employed to extract the spatial information. Finally, we combined these two deep learning models with score fusion which shows the positive result for sign language recognition. In this paper, by combining CapsNet with LSTM we are able to achieve 98.96% total accuracy of recognition rate, which has 2.54% increasing compared with YOLOv3 + LSTM and 0.24% growth compared with only extracting spatial information by adopting CapsNet.

5 Conclusion and Future Work

In this paper, we have proffered CapsNet-based deep learning model to achieve sign language recognition. Through the experiments, deep learning model is well implemented to achieve our goal. The CapsNet model shows positive results for sign language recognition, which has 0.77% growth of accuracy by using the SKNet. The SKNet with attention mechanism attains the highest recognition accuracy 98.88%.

From the outcomes, we see that most of the traditional deep learning models could be extended through either depth or width of the layers to improve the accuracy of the model and make the model more robust and stable. The convolution-based deep learning showcases that without increasing the complexity of the parameters, the model will be efficient and accurate. Moreover, by adopting the attention mechanism in the convolution-based deep learning methods, it shows the positive results in sign language recognition. Meanwhile, CapsNet model is also outperformed in this paper compared with other traditional deep learning methods.

In our future work, we will add the attention module into the proposed CapsNet model in order to get the temporal information, which achieves a better accuracy in sign language recognition.
References

1. Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: a survey. Artif. Intell. Rev. 43(1), 1–54 (2012). https://doi.org/10.1007/s10462-012-9356-9

2. Dardas, N.H., Georganas, N.D.: Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. IEEE Trans.Instrum. Meas. 60(11), 3592–3607 (2011)

3. Tharwat, A., Gaber, T., Hassanien, A.E., Shahin, M.K., Refaat, B.: SIFT-based arabic sign language recognition system. In: Abraham, A., Krömer, P., Snasel, V. (eds.) Afro-European Conference for Industrial Advancement. AISC, vol. 334, pp. 359–370. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-13572-4_30

4. Jasim, M., Hasanuzzaman, M.: Sign language interpretation using linear discriminant analysis and local binary patterns. In: International Conference on Informatics, Electronics & Vision, pp. 1–5 (2014)

5. Cote, M., Payeur, P., Comeau, G.: Comparative study of adaptive segmentation techniques for gesture analysis in unconstrained environments. In: IEEE International Workshop on Imaging Systems and Techniques, pp. 28–33 (2006)

6. Lu, J., Shen, J., Yan, W., Bacic, B.: An empirical study for human behavior analysis. Int. J. Digit. Crime Forensics 9, 11–27 (2017)

7. Asadi-Aghbolagh, M., et al.: A survey on deep learning based approaches for action and gesture recognition in image sequences. In: IEEE International Conference on Automatic Face & Gesture Recognition, pp. 476–483 (2017)

8. Herath, S., Harandi, M., Porikli, F.: Going deeper into action recognition: a survey. Image Vis. Comput. 60, 4–21 (2017)

9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)

10. LeCun, Y., Huang, F.J., Bottou, L.: Learning methods for generic object recognition with invariance to pose and lighting. In: IEEE Conference on Computer Vision and Pattern Recognition (2004)

11. Hinton, G.E., Osindero, S., Teh, Y.-W.: A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527–1554 (2006)

12. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: AAAI Conference on Artificial Intelligence (2017)

13. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35, 221–231 (2013)

14. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2

15. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE 86(11), 2278–2324 (1998)

16. Rao, G.A., Syamala, K., Kishore, P.V., Sastry, A.S.C.S.: Deep convolutional neural networks for sign language recognition. In: The Conference on Signal Processing and Communication Engineering Systems, pp. 194–197 (2018)

17. Koller, O., Ney, H., Bowden, R.: Deep hand: how to train a CNN on 1 million hand images when your data is continuous and weakly labelled. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3793–3802 (2016)
18. Wu, J., Ishwar, P., Konrad, J.: Two-stream CNNs for gesture-based verification and identification: Learning user style. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 42–50 (2016)
19. Liu, Z., Zhang, C., Tian, Y.: 3D-based deep convolutional neural network for action recognition with depth sequences. Image Vis. Comput. 55, 93–100 (2016)
20. Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., Kautz, J.: Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4207–4215 (2016)
21. Huang, J., Zhou, W., Li, H., Li, W.: Sign language recognition using 3D convolutional neural networks. In: IEEE International Conference on Multimedia and Expo, pp. 1–6 (2015)
22. Neverova, N., Wolf, C., Taylor, G.W., Nebout, F.: Hand segmentation with structured convolutional learning. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9005, pp. 687–702. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16811-1_45
23. Han, M., Chen, J., Li, L., Chang, Y.: Visual hand gesture recognition with convolution neural network. In: IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp. 287–291 (2016)
24. Dadashzadeh, A., Targhi, A.T., Tahmasbi, M., Mirmehdi, M.: HGR-net: a fusion network for hand gesture segmentation and recognition. IET Comput. Vis. 13(8), 700–707 (2019)
25. Elboushaki, A., Hannane, R., Afdel, K., Koutti, L.: MultiD-CNN: a multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences. Expert Syst. Appl. 139, 112829 (2020)
26. Chen, Y., Zhao, L., Peng, X., Yuan, J., Metaxas, D. N.: Construct dynamic graphs for hand gesture recognition via spatial-temporal attention. In: British Machine Vision Conference, pp. 1–13 (2019)
27. dos Santos, C.C., Samatelo, J.L.A., Vassallo, R.F.: Dynamic gesture recognition by using CNNs and star RGB: a temporal information condensation. Neurocomputing 400, 238–254 (2020)
28. Wang, P., Li, W., Liu, S., Gao, Z., Tang, C., Ogunbona, P.: Large-scale isolated gesture recognition using convolutional neural networks. In: International Conference on Pattern Recognition, pp. 7–12 (2016)
29. Duan, J., Zhou, S., Wan, J., Guo, X., Li, S. Z.: Multi-modality fusion based on consensus-voting and 3D convolution for isolated gesture recognition. arXiv:1611.06689 (2016)
30. Rastgoo, R., Kiani, K., Escalera, S.: Multi-modal deep hand sign language recognition in still images using restricted Boltzmann machine. Entropy 20(11), 809 (2018)
31. Rastgoo, R., Kiani, K., Escalera, S.: Video-based isolated hand sign language recognition using a deep cascaded model. Multimed. Tools Appl. 79, 22965–22987 (2020). https://doi.org/10.1007/s11042-020-09048-5
32. Sabour, S., Frosst, N., Hinton, G. E.: Dynamic routing between capsules. In: Advances in Neural Information Processing Systems, pp. 3856–3866 (2017)
33. Lu, J., Nguyen, M., Yan, W.: Deep learning methods for human behavior recognition. In: IEEE IVCNZ (2020)
34. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)