Classification of American Sign Language by Applying a Transfer Learned Deep Convolutional Neural Network

Md. Mehedi Hasan*, Azmain Yakin Srizon†, Abu Sayeed‡ and Md. Al Mehedi Hasan§
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh
Email: *mmehedihasann@gmail.com, †azmainsrizon@gmail.com, ‡abusayeed.cse@gmail.com, §mehedi_ru@yahoo.com

2020 23rd International Conference on Computer and Information Technology (ICCIT) | 978-1-6654-2244-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCIT51783.2020.9392703

***Abstract*—Having the community with more than 500,000 deaf and mute English speaking people, sign alphabet detection has become a domain of interest among the researchers for the last decade. Previously, a lot of signs of progress have been made in accurate recognition of the American Sign Language. Both convolutional neural networks and traditional machine learning classifiers have been applied formerly for the recognition process. In this study, we've considered an American Sign Language dataset having 36 classes of English characters and digits. The latest research on this real-world dataset achieved an accuracy of 90%. However, in our research, we introduced modified inceptionV3 architecture for the detection of American Sign characters and obtained overall correctness of 98.81% which outperformed all the previous studies by a notable margin.**

***Index Terms*—American Sign Language, Alphabets and Digits Recognition, Deep Convolutional Neural Network, InceptionV3, Augmentation**

I. Introduction

An incomplete or absolute inability to hear is recognized as hearing impairment or hearing injury which can appear to the individual ear or two ears [1], [2], [3]. Hearing impairment can be advised through several agents i.e., genetical structure, aging, vulnerability to loud sound, various germs, birth complexities, damage of ear, and unusual medications or toxins [2]. Till 2013, listening impairment affected almost 1100 million individuals to a remarkable extent [4]. This has provoked a weakness to almost 538 million persons and led to severe disabilities in roughly 124 million persons [2], [5], [6].

To defeat the connection passage among deaf-mute people, sign literature is employed that is the common exercised literature among the deaf-mute community to interact among people and bestow opinions[7], [8]. Deaf-mute is an expression that is exercised historically to identify a person who is either deaf or both deaf and cannot speak at the same time, and in both cases, sign language is the process of interaction for them. Sign language is a kind of literature that employs visual-manual procedures to interact or carry meaning. Like natural languages, sign languages possess their grammar and vocabulary [9]. Despite holding notable associations between sign languages, they are not extensively the same and mutually acknowledged [9].

Being stated that, many researchers have contributed significantly for the accurate recognition of sign language characters. However, in this research, we specifically focused on American sign language character recognition. Having 250,000 to 500,000 persons in the deaf community of Americans and some Canadians who utilize American sign language, this was an obvious opportunity for research [10].

II. Literature Review

Previously, many kinds of research have been conveyed to successfully recognize sign languages of various communities all over the world [11], [12]. For example, a study of Indian sign language recognition gained 93% overall recognition accuracy by using enhanced skin and wrist discovery algorithms [13]. Researches toward Spanish sign literature proposed overall correctness of 96% [14]. Various researches have contributed significant discoveries within the field of American sign gesture detection. One of the researches suggested the recognition of characters by using the Support Vector Machine classifier on the histogram of gradients (HOG) features extracted from real-time hand gesture images [15]. However, the problem arises with the increase of dataset as it also increases training time. Another research discussed the variation in performance for hand gesture recognition by applying multiple techniques i.e., Random Tree, C4.5 (J48), Naïve Bayes, NNge, ANN (Multiplayer Perceptron) and Support Vector Machine (Linear and RBF Kernels)[16]. However, a study suggested a deep convolutional neural network for recognizing American sign characters and outperformed previously discovered high-performance approaches like HOG + SVM (Liner Kernel), Random Forest, SVM (Linear Kernel) and Multilayer Perceptron (2 Hidden Layers) [17].

For detection of the American sign gestures, a modern approach applying a low-budget base Microsoft's Kinect camera produced overall correctness of 90% where twenty-four classes of English characters were considered [18]. Besides, convolutional neural network was applied on both English characters and digits with a recognition accuracy of 82% and 97% respectively [19]. A research proposed a convolutional neural network architecture to achieve an overall accuracy of 72% [20]. Another research suggested an accuracy of

978-1-6654-2244-4/20/$31.00 ©2020 IEEE

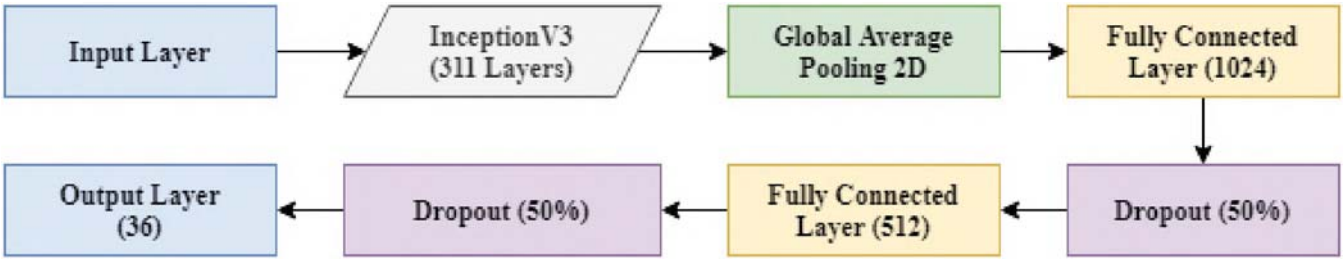

Fig. 1: Modified InceptionV3 Architecture

TABLE I: Class-wise accuracy, precision, recall, f-score, and support of each of the classes

Classes	Class-specific Accuracy	Class-specific Precision	Class-specific Recall	Class-specific F1-Score	Support
A	1.00	1.00	1.00	1.00	14
B	1.00	1.00	1.00	1.00	14
C	1.00	1.00	1.00	1.00	14
D	1.00	1.00	1.00	1.00	14
E	1.00	1.00	1.00	1.00	14
F	1.00	1.00	1.00	1.00	14
G	0.86	1.00	0.86	0.92	14
H	1.00	1.00	1.00	1.00	14
I	1.00	1.00	1.00	1.00	14
J	1.00	1.00	1.00	1.00	14
K	1.00	1.00	1.00	1.00	14
L	0.93	1.00	0.93	0.96	14
M	1.00	1.00	1.00	1.00	13
N	1.00	1.00	1.00	1.00	14
O	1.00	0.93	1.00	0.97	14
P	1.00	0.93	1.00	0.97	14
Q	1.00	1.00	1.00	1.00	14
R	1.00	1.00	1.00	1.00	14
S	1.00	1.00	1.00	1.00	14
T	1.00	0.82	1.00	0.90	14
U	1.00	1.00	1.00	1.00	14
V	0.93	1.00	0.93	0.96	14
W	1.00	1.00	1.00	1.00	14
X	1.00	1.00	1.00	1.00	14
Y	1.00	1.00	1.00	1.00	14
Z	1.00	1.00	1.00	1.00	14
0	0.93	1.00	0.93	0.96	14
1	1.00	0.93	1.00	0.97	14
2	0.93	1.00	0.93	0.96	14
3	1.00	1.00	1.00	1.00	14
4	1.00	1.00	1.00	1.00	14
5	1.00	1.00	1.00	1.00	14
6	1.00	1.00	1.00	1.00	14
7	1.00	1.00	1.00	1.00	14
8	1.00	1.00	1.00	1.00	14
9	1.00	1.00	1.00	1.00	14

87% while recognizing the sign language characters with the assistance of a consumer depth camera [21].

III. Materials and Methods

A. Dataset Description

In this research, we considered an American sign language dataset containing both digits and characters samples [22]. There were 70 samples per class. The dataset contained a total of 2525 images. The pictures were fully PNG type in RGB style having gestures seperated by color and the backdrop was set to no color or black (0, 0, 0).

B. Convolutional Neural Network

CNN, also known as convolutional neural network, is among the most common classes of deep neural networks which is generally practiced for analyzing visual images of different

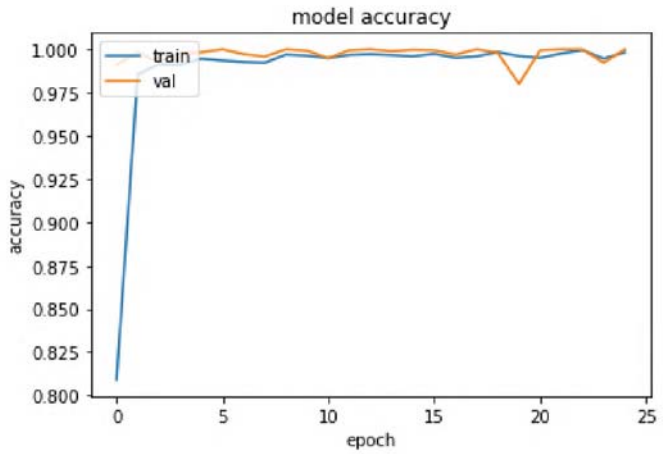

Fig. 2: Training accuracy and validation accuracy of our proposed architecture

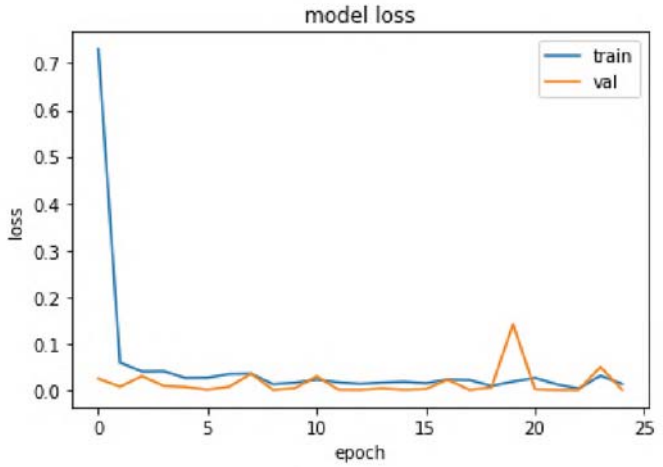

Fig. 3: Training loss and validation loss of our proposed architecture

research areas [23]. CNN models are utilized for the detection in images and videos, image identification, recommender operations, NLP, medical illustration study, and financial analysis [24], [25], [26]. If we regularize the multilayer perceptrons, we will get the CNNs. In multilayer perceptrons, neurons from a layer are attached to all other neurons of the preceding stage, hence, CNN forms one fully connected sequence between two layers. But the term fully-connectedness allows us to introduce overfitting to the network. On the other hand, CNN also practices regularization but by assembling patterns with the help of simpler and smaller patterns instead of applying fully-connectedness. Because of taking advantage of hierarchical patterns of data, CNNs don't need huge pre-processing models. CNNs are capable of finding out the valuable information from the input and learn the filters which are not general in traditional algorithms. In traditional algorithms, the inputs are hand-engineered or highly preprocessed in a different step. CNNs were inspired by the connectivity among neurons which can be discovered in the animal cortex and are commonly practiced for image recognition or identification in various fields of research.

Typically a CNN model is built with one input and one output layer along with various private layers. The private layers of a CNN model may contain a chain of multiple convolutional courses which are combined with a dot product or multiplication. An activation function is practiced and it is accompanied by supplementary convolutions i.e., pooling, fully connected, and hidden or normalization layers. These layers are applied in sequence to obtain more information from the input data. Although these layers are called convolutions, it is only a convention. In terms of mathematics, it is actually a cross-correlation or a sliding dot product. It possesses importance for the contents in the matrix, in which it concerns by which way weight is arranged at a particular index position.

C. Transfer Learning

Transfer learning concentrates toward collecting information obtained during resolving individual obstacle and implementing that toward another similar dilemma [27]. For instance, the information obtained while discovering to identify cars could utilize during the recognition of trucks. This field of investigation shows a remarkable relationship over an enduring chronicle of cerebral research about the transfer of knowledge, though confirmed relations among the two areas are inadequate. From a pragmatic viewpoint, transferring or conveying knowledge from earlier accomplished assignments for the training of new jobs has the potentiality to dramatically enhance the individual performance of an agent [28].

D. Modified InceptionV3 Architecture

In this research, we've considered and applied inceptionV3 architecture [29]. Inception version 3 is a CNN model to support picture interpretation and target recognition that made the journey as a part of Googlenet. This was the 3rd version of Google's Inception CNN architecture, formerly proposed for the ImageNet Contest. We started with the inceptionV3 of 311 layers. After that, we applied a fully connected layer having volume 1024 supported by a dropout of 50%. Then another fully connected layer of volume 512 was added supported by another dropout of 50%. Finally, an output layer is added having a size of 36. Figure-1 illustrates the architecture of the modified inceptionV3 architecture. While training by the architecture, no layer was kept frozen.

E. Augmentation

Inadequate data has always been a noteworthy obstacle while performing deep architectures i.e., CNNs. Moreover, not balanced data in terms of labels can be an additional barrier. There can be sufficient samples for many labels, uniformly meaningful, but the lower-sampled labels will experience ineffectual class-wise performance or correctness. That phenomenon is consistent. If the representation learns of several samples or occurrences from a presented label, it's less probable in predicting the group label or test label.

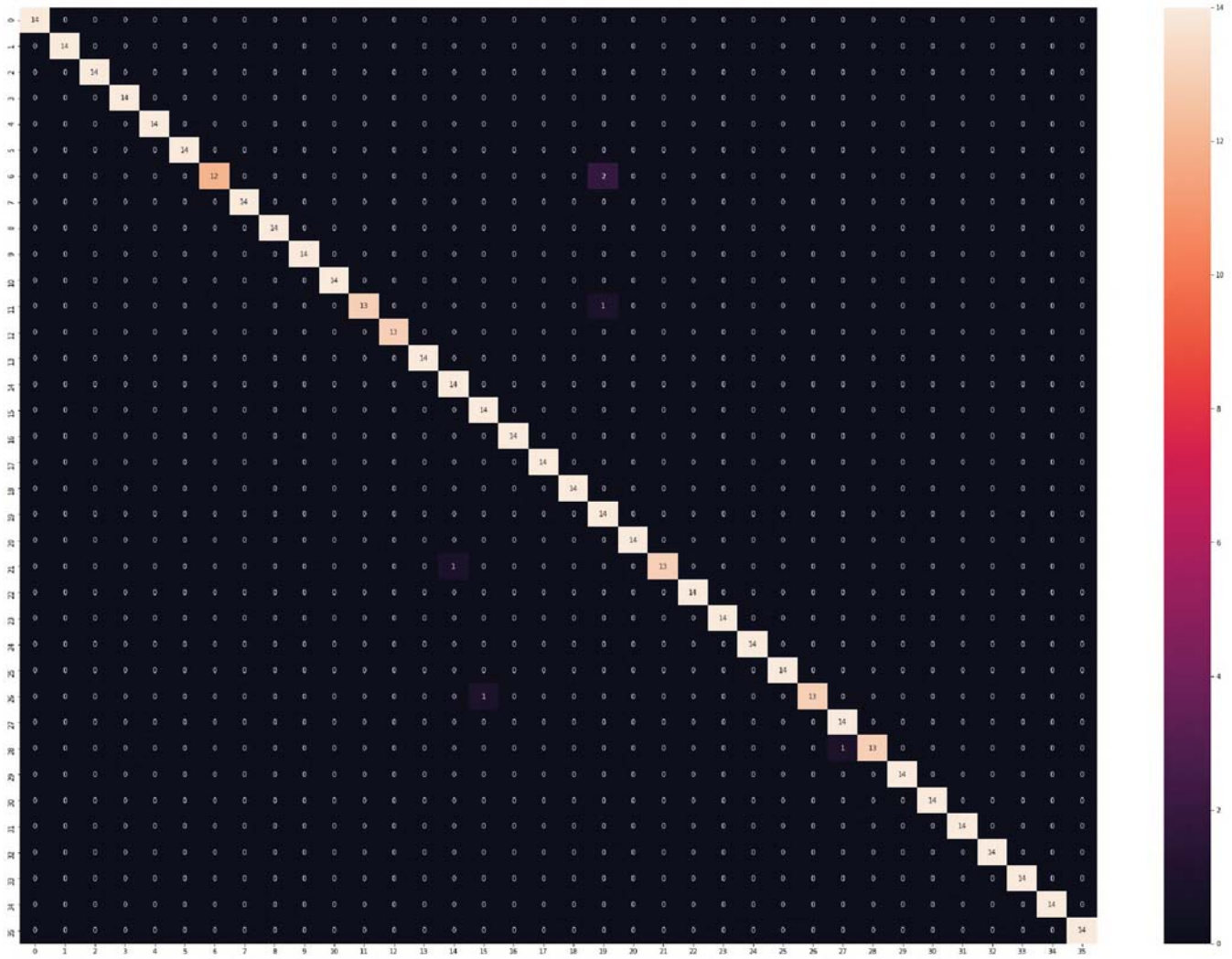

Fig. 4: Confusion matrix of ASL detection

Augmentation of pictures computationally generates training images through numerous procedures of steps or incorporation of multiple steps, such as rotating, transfering, shearing, flipping, etc. [30].

Augmentation has been demonstrated efficient against various obstacles like input augmentation conducted through proficient understanding [31], universal picture augmentation [32], have confirmed useful outcomes in picture recognition [33]. Modest performers in AI business usually require entree toward notable quantities of knowlegde. The impulse regarding our problem domain is both comprehensive and precise which triggers the usage of data augmentation here. Among numerous data augmentation methods some popular ones are flip, rotate, scale, crop, translate, Gaussian noise, etc.

IV. Experimental Analysis

In this section, firstly, preprocessing will be discussed. After that design of the experiment and result analysis will be presented.

TABLE II: Comparison between our proposed work and notable previous works

Classifier or Model Name	Overall Accuracy
Microsoft Kinect [18]	90.00%
DCNN [19]	82.00%
CNN [20]	72.00%
Consumer Depth Camera [21]	87.00%
Proposed	**98.81%**

A. Preprocessing

Because of rendering images to a convolutional neural network, a heavy preprocessing of the pictures was leaped as CNN is a powerful network that can identify valuable features from raw photographs. However, some preprocessing steps were required. The input images were in the different shapes, hence, they were reshaped to 224x224x3. For more accurate recognition, augmentation was applied using the assistance of Augmentor Library [34]. During applying the method of

augmentation, max left turn, the max right turn, and the probability of turn of the rotating reception was set to 3, 3, 0.4 respectively. The values of diameter, length, probability, and size of the distortion reception were fixed to 4, 4, 0.4, 4 respectively. Furthermore, the percentage domains and the probability of the zooming reception were fixed to 0.9, 0.2 respectively. After augmenting, we had 500 images per label, a total of 18,000 pictures of 36 characters of English sign characters and digits.

B. Design of Experiment

The architecture was run for 25 times where the batch size was set to 24 because after 25 epochs the training and validation loss shifted to nearly unchanged era. The Adam optimization having a 0.0001 learning speed was utilized to produce minimize errors. A categorical cross-entropy functionality was applied for measuring error. To avoid overtraining, the dropout mechanism was followed.

C. Result Analysis

Firstly, the data was split into the train set and test set. 80% of data was kept in the train set and the rest of the 20% data was kept to test set. Then, the proposed modified inceptionV3 architecture was employed to train set. Training accuracy and validation accuracy of our proposed architecture are illustrated in Figure-2. On the other hand, Figure-3 illustrates the training and validation loss of our proposed architecture. Figure-4 illustrates the confusion matrix of ASL detection. Table-1 illustrates the class-wise accuracy, precision, recall, f-score, and support of each of the classes under consideration. Our proposed architecture produced overall correctness of 98.81% for American Sign Language dataset. Table-2 illustrates the comparison among our proposed work and notable previous works. From Table-2 it can be noticed that our proposed architecture outperformed all the previous approaches by a notable margin, hence, our model is capable of recognizing the considered classes more accurately.

V. Conclusion

In this investigation, we started with the American Sign Language dataset. Many works have already been conveyed toward American Sign Literature. But in most of the cases, only digits or only characters were taken under consideration while applying the recognition process. However, no notable works have been conducted on the dataset used in this research. One of the main reasons is that the dataset consists of 10 digits and 26 characters of the English language which makes it a 36 class dataset. Moreover, the size of the pictures is not the same and the images are in RGB format. After some preprocessing steps, we applied modified inceptionV3 architecture on the dataset and obtained an overall accuracy of 98.81% which outperformed all the previous studies by a fine margin.

References

[1] E. Britannica and E. Britannica, "Inc., 2012," *Encyclopædia Britannica Online. Website*, 2012.

[2] W. H. Organization *et al.*, "Deafness and hearing loss. fact sheet n 300. updated march 2015," 2015.

[3] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 1, pp. 131–153, 2019.

[4] T. Vos, R. M. Barber, B. Bell, A. Bertozzi-Villa, S. Biryukov, I. Bolliger, F. Charlson, A. Davis, L. Degenhardt, D. Dicker *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the global burden of disease study 2013," *The Lancet*, vol. 386, no. 9995, pp. 743–800, 2015.

[5] W. H. Organization *et al.*, *The global burden of disease: 2004 update*. World Health Organization, 2008.

[6] B. O. Olusanya, K. J. Neumann, and J. E. Saunders, "The global burden of disabling hearing impairment: a call to action," *Bulletin of the World Health Organization*, vol. 92, pp. 367–373, 2014.

[7] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef *et al.*, "Sign language recognition, generation, and translation: An interdisciplinary perspective," in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 2019, pp. 16–31.

[8] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4165–4174.

[9] W. Sandler and D. Lillo-Martin, *Sign language and linguistic universals*. Cambridge University Press, 2006.

[10] R. E. Mitchell, T. A. Young, B. Bachelda, and M. A. Karchmer, "How many people use asl in the united states? why estimates need updating," *Sign Language Studies*, vol. 6, no. 3, pp. 306–335, 2006.

[11] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1880–1891, 2019.

[12] Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, "Dynamic sign language recognition based on video sequence with blstm-3d residual networks," *IEEE Access*, vol. 7, pp. 38 044–38 054, 2019.

[13] K. Yadav, L. P. Saxena, B. Ahmed, and Y. K. Krishnan, "Hand gesture recognition using improved skin and wrist detection algorithms for indian sign," *Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org*, vol. 9, no. 2, 2019.

[14] G. Saldaña González, J. Cerezo Sánchez, M. M. Bustillo Díaz, and A. Ata Pérez, "Recognition and classification of sign language for spanish," *Computación y Sistemas*, vol. 22, no. 1, pp. 271–277, 2018.

[15] E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE transactions on intelligent transportation systems*, vol. 15, no. 6, pp. 2368–2377, 2014.

[16] M. Lech, B. Kostek, and A. Czyżewski, "Examining classifiers applied to static hand gesture recognition in novel sound mixing system," in *Multimedia and Internet Systems: Theory and Practice*. Springer, 2013, pp. 77–86.

[17] D. Chakraborty, D. Garg, A. Ghosh, and J. H. Chan, "Trigger detection system for american sign language using deep convolutional neural networks," in *Proceedings of the 10th International Conference on Advances in Information Technology*, 2018, pp. 1–6.

[18] C. Dong, M. C. Leu, and Z. Yin, "American sign language alphabet recognition using microsoft kinect," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 44–52.

[19] V. Bheda and D. Radpour, "Using deep convolutional networks for gesture recognition in american sign language," *arXiv preprint arXiv:1710.06836*, 2017.

[20] B. Garcia and S. A. Viesca, "Real-time american sign language recognition with convolutional neural networks," *Convolutional Neural Networks for Visual Recognition*, vol. 2, pp. 225–232, 2016.

[21] A. Kuznetsova, L. Leal-Taixé, and B. Rosenhahn, "Real-time sign language recognition using a consumer depth camera," in *2013 IEEE International Conference on Computer Vision Workshops*, 2013, pp. 83–90.

[22] A. Barczak, N. Reyes, M. Abastillas, A. Piccio, and T. Susnjak, "A new 2d static hand gesture colour image dataset for asl gestures," 2011.

[23] M. Valueva, N. Nagornov, P. Lyakhov, G. Valuev, and N. Chervyakov, "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation," *Mathematics and Computers in Simulation*, 2020.
[24] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Advances in neural information processing systems*, 2013, pp. 2643–2651.
[25] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 160–167.
[26] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Forecasting stock prices from the limit order book using convolutional neural networks," in *2017 IEEE 19th Conference on Business Informatics (CBI)*, vol. 1. IEEE, 2017, pp. 7–12.
[27] J. West, D. Ventura, and S. Warnick, "Spring research presentation: A theoretical foundation for inductive transfer," *Brigham Young University, College of Physical and Mathematical Sciences*, vol. 1, no. 08, 2007.
[28] T. George Karimpanal and R. Bouffanais, "Self-organizing maps for storage and transfer of knowledge in reinforcement learning," *Adaptive Behavior*, vol. 27, no. 2, pp. 111–126, 2019.
[29] J. Tang, *Intelligent Mobile Projects with TensorFlow: Build 10+ Artificial Intelligence Apps Using TensorFlow Mobile and Lite for IOS, Android, and Raspberry Pi*. Packt Publishing Ltd, 2018.
[30] M. D. Bloice, C. Stocker, and A. Holzinger, "Augmentor: an image augmentation library for machine learning," *arXiv preprint arXiv:1708.04680*, 2017.
[31] C. N. Vasconcelos and B. N. Vasconcelos, "Increasing deep learning melanoma classification by classical and expert knowledge based image transforms," *CoRR, abs/1702.07025*, vol. 1, 2017.
[32] Y. Xu, R. Jia, L. Mou, G. Li, Y. Chen, Y. Lu, and Z. Jin, "Improved relation classification by deep recurrent neural networks with data augmentation," *arXiv preprint arXiv:1601.03651*, 2016.
[33] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?" in *2016 international conference on digital image computing: techniques and applications (DICTA)*. IEEE, 2016, pp. 1–6.