2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)
Military Institute of Science and Technology (MIST), Dhaka-1216, Bangladesh

2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT) | 978-1-6654-9522-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICEEICT53905.2021.9667804

Improving Automatic Sign Language Translation with Image Binarisation and Deep Learning

Mahmudul Haque*, Syma Afsha†, Tareque Bashar Ovi‡ and Hussain Nyeem§
Department of Electrical, Electronic and Communication Engineering (EECE)
Military Institute of Science and Technology (MIST)
Mirpur Cantonment, Dhaka–1216, Bangladesh
*mahmud.eece@gmail.com †symaafsha.eece@gmail.com
‡ovitareque@gmail.com §h.nyeem@eece.mist.ac.bd

***Abstract*—Sign Language Translation (SLT) has been widely investigated to provide a futuristic solution to tackle human speech and hearing disability. Recent deep learning-based SLT models have redefined computer vision-based detection and classification to automatically translate the hand-gestured based sign language (SL) into natural language (NL) with higher accuracy. Unlike the existing models that directly learn from the natural image-sets, in this paper, we propose a 2D Convolutional Neural Network (CNN) model with customised hyper-parameters to be trained with binary SL image-sets. We thus introduce a binarisation step to preprocess the images of size 28×28 to feed the model. Preliminary results of our model trained with binarised image-set demonstrate its potential with an impressive classification accuracy of 99.99% on the NVIDIA Tesla K80 GPU environment (Google Colab) for an automatic SLT system.**

***Index Terms*—Sign language, SLT, CNN, binarisation, speaking disability**

I. Introduction

A sign language translation (SLT) system can tackle human speech and hearing impairments capturing more than 5% of the world's population [1]. This system converts the sign language (SL) to natural language (NL). As an alternative to NL, the SL communicates primarily via hands, face, and eyes with their different orientations and movements [2]. As the NL varies with the nation, and so does the SL. Besides, dominant variations of SL also set additional criteria of for using single or double-handed gestures.

Irrespective of the dominant and locality variants, an automatic SLT system requires recognising individual gestures and human activities, which is recently investigated using vision-based models with deep learning. These models having several computational layers can process data and identify high-level features, and thus, have refined the computer-vision based classification of SL's gestures [3]. Among their classes, Convolutional Neural Network (CNN) is mainly used for image-based recognition and classification of objects that deterministically sets network weights to extract information from the fed images [4].

Recent SLT models with CNN have seen considerable performance improvements [5]. As an early effort, Krishnan *et al*. [6] proposed a 2D-CNN model to classify hand-gestures for English language alphabets with an accuracy of 82%. Shareef *et al*. [7] reported a mean accuracy of 93.4% for their vision-based CNN model to automatically identify hand motions for gesture translation. Similarly, Hasan *et al*. [8] analysed the SL MNIST dataset for their shallow CNN and Capsule Networks with an overall accuracy of 97.62% surpassing all the earlier models to classify and recognise the English sign alphabets.

Despite the improvement of the hand-gesture classification, the above SLT models largely overlooked the potential of the preprocessing of the training image set and hyper-parameter for better optimisation. A pre-processed image set with binarisation to individually create a bi-level document image is conventionally used for more accurate Optical Character Recognition (OCR) and layout analysis [9]. Unlike the characters, similar improvements were also reported for classification of the hand-vein patterns [10] and tumours [11].

Besides, the Learning Rate (LR) hyper-parameter can control the converging rate to local minima in a neural network. A dynamic LR-based CNN model [12] was reported to reduce the detection error cost using a high-pass filtering layer on the training dataset. Additionally, selective LR rules were found useful to control the classification confidence, variance, cost, and robustness, and thereby, the best LR update schedules were suggested in an appropriate LR value range [13].

We, therefore, present a 2D-CNN model with binarisation and LR update scheduling for a more accurate SLT system (Sec. II). The model is primarily developed to convert the SL to text. The performance of the model is investigated for the American sign language (ASL), and thus, trained using the SL MNIST dataset [14]. Binarisation of the images fed to model with selective LR, which is introduced to the model to improve its automatic classification accuracy promising for an SLT system (Sec. III).

II. Proposed CNN with Image Binarisation

We now present the proposed CNN model with binarisation for an SLT system. For the demonstrated promises, we considered CNN with TensorFlow (TF) v2.6 framework for the image-based multi-class classification. The overall construction and workflow of the proposed SLT system is illustrated in Fig. 1. Technical details of the development and implementation of the system is briefly presented below.

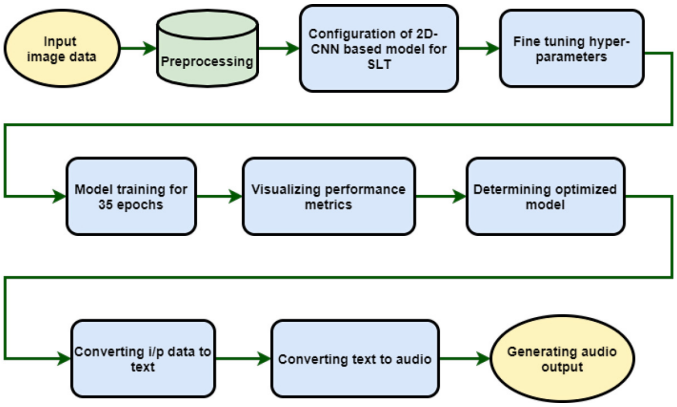


Fig. 1. Demonstration and overall workflow of SLT system

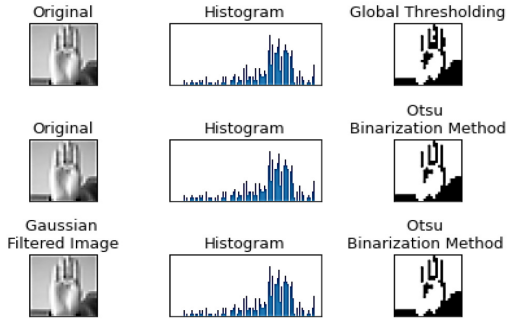


Fig. 2. Binarisation of an input image

A. Image Binarisation and Augmentation

A popular ASL dataset MNIST [14] is considered for our work that contains total 27,455 images for the training set and 7,172 images for the test set, all are greyscale with the dimension 28×28. Unlike the standard image formats, the training and testing images are given in individual excel files, where the pixel values of each image in the dataset are given in a row with the label (*i.e.*, corresponding English numeric or alphanumeric symbol) in followed-by column. The dataset contains images for all English alphabets except J and Z, since motion is involved in their representation in ASL.

The 2D form (*i.e.*, 28×28) of all the images were reconstructed from the dataset followed by their binarisation. We investigated several binarisation techniques for our work, and observed that the *Global Thresholding* gave better outcomes in terms of binarisation, as reported in [9] and seen in Fig. 2.

We categorised the alphabets in their serials, starting from 0 in alphanumeric arrangement, and labelled the corresponding sign. As illustrated in Fig. 3, the samples representing each alphabet were labelled from 0 to 24, with no sample being labelled 9 and 25 with the images for J and Z being absent in the dataset.

Once the labelling is done, and it is ensured that there is approximately equal distribution of data for each class, *i.e.*, 10% or 2,745 samples were separated for the validation purpose leaving the training set with 24,710 samples. Using data augmentation techniques by Keras available with TF v2.6 framework, the samples were pre-processed for the proposed model.

B. Modelling SLT with Hyper-parameters

After binarisation and augmentation of the training images, a model was designed to translate SL to NL as English text using the proposed 2D-CNN model. The first layer is for the input greyscale images of size 28×28. As shown in Fig. 4, the model comprises three 2D-Convolutional layers and three 2D-MaxPool layers. These layers mainly work to identify unique patterns within an image. The Flatten layer is then used to vectorise the tensor followed by the successive dense layers to extract complex features of the previously recognised unique patterns. A dropout layer is also used to prevent the model from over-fitting.

The kernels of size (3, 3) are used for both the 2D convolutional layers and MaxPool 2D layers, while only the convolutional kernels are learning-based. We defined 32, 128, and 512 filters, respectively, for the three convolutional layers. For each layer except for the output layer, relu activation functions are used. For the activation of the output layer, learning-based *softmax* is used.

III. Results

The proposed model is trained for 35 epochs rendering the preprocessed training data comprising 256 samples at a time. We have used a Google Colaboratory environment for the training, which includes a 2.30GHz Intel(R) Xeon(R) CPU, 12.63 GB RAM, and a 12 GB NVIDIA Tesla K80 GPU for computational purposes. After each epoch, we tested the model using the validation dataset. Based on the accuracy and value of the loss function, we have optimised the model's parameters. The Adam's optimiser is used for optimisation, and sparse categorical cross-entropy is used as a loss function.

The model is thus trained with a dynamic LR, where the minimum LR was set at 0.00001. A total of 2,994,649 parameters are trained and optimised. A checkpoint is created at the end of each epoch, so the best model can be retrieved later with the evaluation of performance metrics. Once the model is trained, we test the best fit of the model using the test dataset separated prior to training. Upon evaluation of the accuracy, categorical precision, categorical f1 score and other metrics, the model's performance is evaluated, which are now presented below.

During training and validation, the accuracy and loss function values have been inspected to understand the effectiveness of the model. Training and validation accuracies and losses per epoch are illustrated in Fig. 5a and Fig. 5b, respectively. With these accuracy and loss values per epoch, an optimum performance of the model is observed after 25 epochs with an training accuracy of 0.9992 (99.92%) and a loss value of 0.0044. In contrast, the validation accuracy and loss values were 1.00 (100%) and 4.2464×10^{-4}, respectively. This is because, the dynamic LR (see Fig. 5c) helps the model to converge with a minimum loss-function value throughout the training process.

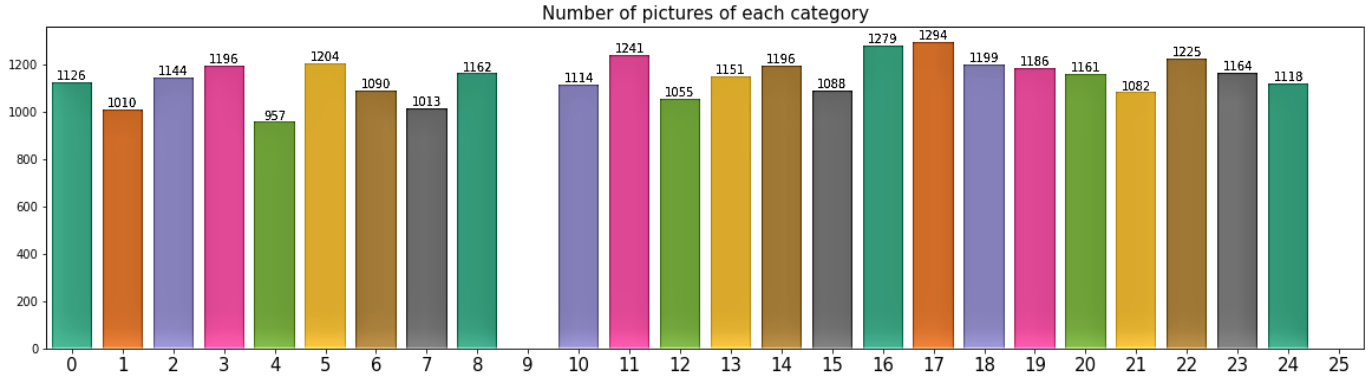


Fig. 3. Class-wise distribution of dataset for training.

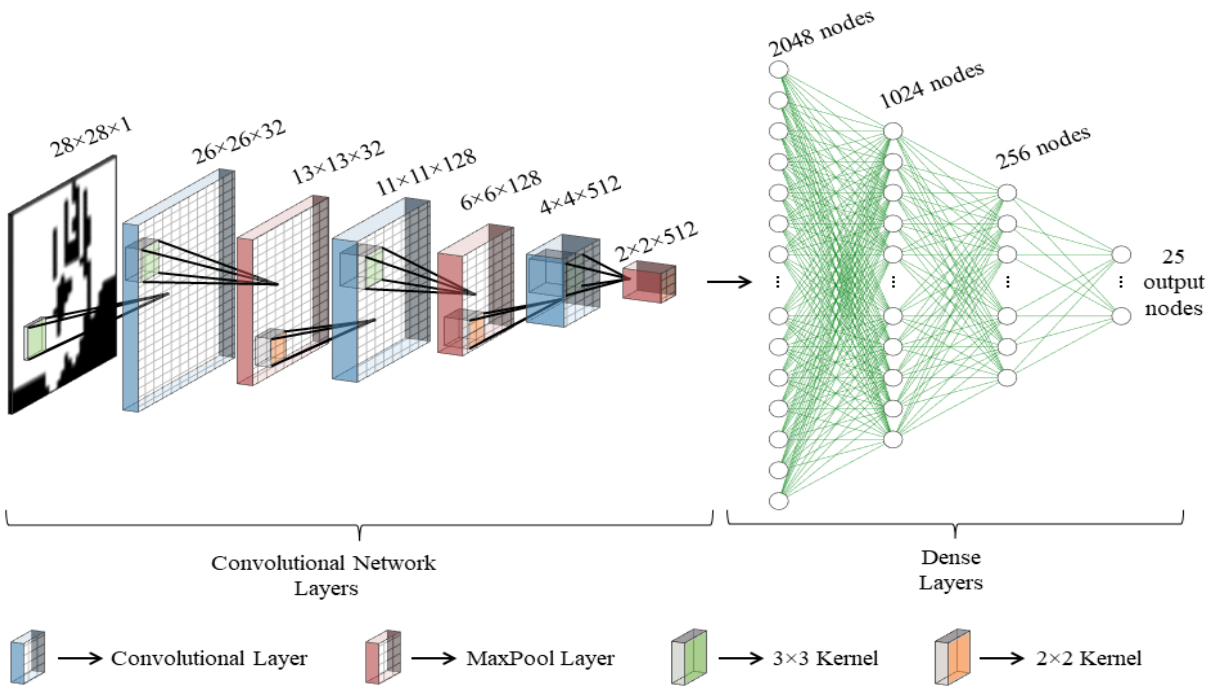


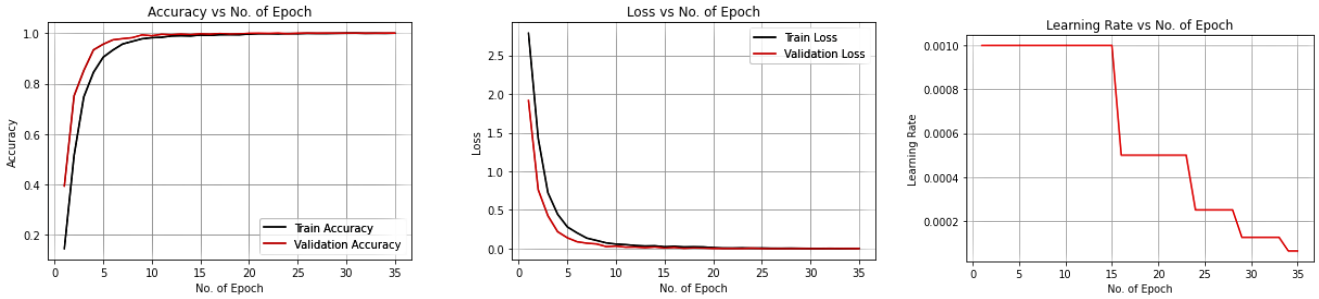
Fig. 4. Architecture for SLT, based on 2D-CNN.

Besides, classification performance of our model in terms of normalised confusion matrix is given in Fig. 6. Categorical precision, recall, f1 score and support values of our model along with their macro and weighted averages are also given in Table. I. From the metrics shown further in Table I and Fig. 6, it is seen that the individual success rate of true predictions among all classes is very high. The average precision, recall and f1 score of each class are 99.97%, 99.99% and 99.98%, respectively. Considering their weighted average, all these values have reached to an equal rate of 99.99%.

The performance observed with the above results demonstrates that classification accuracy (99.99%) of our model is nearly perfect. This rate is also higher than the recent models developed for SLT, as shown in Table II. Particularly, the highest accuracy of 97.62% for the ASL dataset can be tracked in the literature. Also note that the earlier 2D-CNN [6] model developed for SLT offered an accuracy of 82%, while with binarisation and dynamic LR, the accuracy of the proposed model sees an impressive improvement of 21.94%. All these improvements indicates that all the ASL alphabets can be classified by the proposed model with higher accuracy and greater confidence.

IV. Conclusion

An automated SLT system has the potential of opportune and expedient communication means for people with speech and hearing disabilities. Existing deep learning based models were found promising for their better classification accuracy,



(a) (b) (c)

Fig. 5. Performance per epoch: (a) accuracy, (b) loss and (c) learning rate.

TABLE I
EVALUATION METRICS FOR THE SLT MODEL

Class	precision	recall	f1-score	support
a	1	1	1	331
b	1	1	1	432
c	1	1	1	310
d	1	1	1	245
e	1	1	1	498
f	1	1	1	247
g	1	1	1	348
h	1	1	1	436
i	1	1	1	288
k	1	1	1	331
l	1	1	1	209
m	1	1	1	394
n	1	1	1	291
o	1	1	1	246
p	1	0.9971	0.9986	347
q	0.9939	1	0.997	164
r	1	1	1	144
s	1	1	1	246
t	1	1	1	248
u	1	1	1	266
v	1	1	1	346
w	1	1	1	206
x	1	1	1	267
y	1	1	1	332
macro avg	0.9997	0.9999	0.9998	7172
weighted avg	0.9999	0.9999	0.9999	7172

leaving room for their further development. Particularly, the conventional preprocessing steps used to consider in the vision-based character or pattern recognition was overlooked in those models. Considering that preprocessing, we have developed and presented a 2D-CNN model with image binarisation. Experimental results demonstrated a high accuracy of 99.99% indicating its promises for accurate classification among all the classes of the ASL dataset.

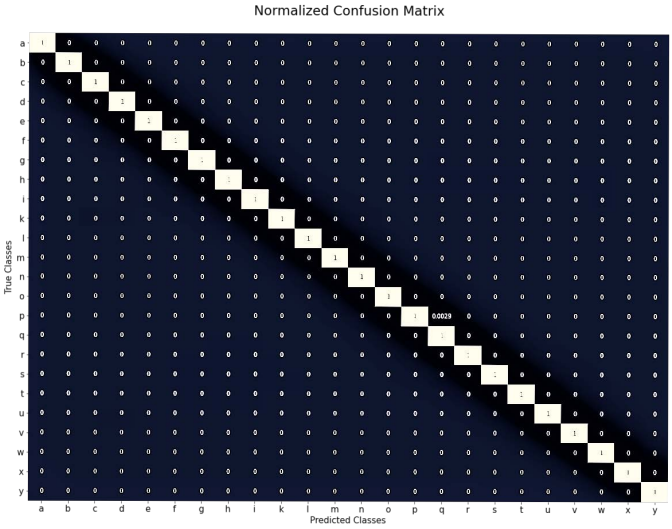


Fig. 6. Normalised confusion matrix for test data.

TABLE II
COMPARISON OF METHOD AND ACCURACY

Work	Method	Dataset	Accuracy
Krishnan *et al.* [6]	2D-CNN	ASL	82%
Shareef *et al.* [7]	Image processing + DL	ASL	93.4%
Hasan *et al.* [8]	CNN and Capsule Networks	ASL	97.62%,
Ours	Image Binarisation + 2D-CNN	ASL	99.99%

A dataset including motion-based gestures and sequential model-based architecture may further increase the efficiency of SLT, which are open to future research.

REFERENCES

[1] K. Tiku, J. Maloo, A. Ramesh, and R. Indra, "Real-time conversion of sign language to text and speech," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2020, pp. 346–351.

[2] U. Farooq, M. S. M. Rahim, N. Sabir, A. Hussain, and A. Abid, "Advances in machine translation for sign language: approaches, limitations, and challenges," *Neural Computing and Applications*, pp. 1–43, 2021.

[3] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif, T. S. Alrayes, H. Mathkour, and M. A. Mekhtiche, "Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation," *IEEE Access*, vol. 8, pp. 192 527–192 542, 2020.

[4] A. Orbay and L. Akarun, "Neural sign language translation by learning tokenization," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 222–228.

[5] K. Nimisha and A. Jacob, "A brief review of the recent trends in sign language recognition," in *2020 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2020, pp. 186–190.

[6] P. T. Krishnan and P. Balasubramanian, "Detection of alphabets for machine translation of sign language using deep neural net," in *2019 International Conference on Data Science and Communication (IconDSC)*. IEEE, 2019, pp. 1–3.

[7] S. K. Shareef, I. S. L. Haritha, Y. L. Prasanna, and G. K. Kumar, "Deep learning based hand gesture translation system," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2021, pp. 1531–1534.

[8] M. M. Hasan, A. Y. Srizon, A. Sayeed, and M. A. M. Hasan, "Classification of sign language characters by applying a deep convolutional neural network," in *2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT)*. IEEE, 2020, pp. 434–438.

[9] Puneet and N. Garg, "Binarization techniques used for grey scale images," *International Journal of Computer Applications*, vol. 71, no. 1, pp. 8–11, 2013.

[10] R. J. Tazim, M. M. M. Miah, S. S. Surma, M. T. Islam, C. Shahnaz, and S. A. Fattah, "Biometric authentication using cnn features of dorsal vein pattern extracted from nir image," in *TENCON 2018-2018 IEEE Region 10 Conference*. IEEE, 2018, pp. 1923–1927.

[11] S. Burje, S. Rungta, and A. Shukla, "Two stage adaptive histogram valley based thresholding for tumor extraction in brain mri images," *Psychology and Education Journal*, vol. 58, no. 2, pp. 10 462–10 471, 2021.

[12] E. M. Mustafa, M. A. Elshafey, and M. M. Fouad, "Accuracy enhancement of a blind image steganalysis approach using dynamic learning rate-based cnn on gpus," in *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 1. IEEE, 2019, pp. 28–33.

[13] Y. Wu, L. Liu, J. Bae, K.-H. Chow, A. Iyengar, C. Pu, W. Wei, L. Yu, and Q. Zhang, "Demystifying learning rate policies for high accuracy training of deep neural networks," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 1971–1980.

[14] Kaggle. (2017) Sign language mnist: Drop-in replacement for mnist for hand gesture recognition tasks. Accessed on: 16 August, 2021. [Online]. Available: https://www.kaggle.com/datamunge/sign-language-mnist