MEKATRONIKA
e-ISSN: 2637-0883
VOL. 5, ISSUE 1, 67 – 71
DOI: https://doi.org/10.15282/mekatronika.v5i1.9410

ORIGINAL ARTICLE

Sign Language Recognition using Deep Learning through LSTM and CNN

Kiran Pandian[1], Mohd Azraai Mohd Razman[1,*], Ismail Mohd Khairuddin[1], Muhammad Amirul Abdullah[1], Ahmad Fakhri Ab Nasir[2] and Wan Hasbullah Mat Isa[1]

[1]Faculty of Manufacturing and Mechatronic Engineering Technolgy, Universiti Malaysia Pahang, 26600 Pahang, Malaysia.
[2]Faculty of Computing, Universiti Malaysia Pahang, 26600 Pekan, Pahang.

ABSTRACT – This study presents the application of using deep learning to detect, recognize and translate sign language. Understanding sign language is crucial for communication between the deaf and mute people and the general society. This helps sign language users to easily communicate with others, thus eliminating the differences between both parties. The objectives of this thesis are to extract features from the dataset for sign language recognition model and the formulation of deep learning models and the classification performance to carry out the sign language recognition. First, we develop methodology for an efficient recognition of sign language. Next is to develop multiple system using three different model which is LSTM, CNN and YOLOv5 and compare the real time test result to choose the best model with the highest accuracy. We used same datasets for all algorithms to determine the best algorithm. The YOLOv5 has achieved the highest accuracy of 97% followed by LSTM and CNN with 94% and 66.67%.

ARTICLE HISTORY
Received: 28th Feb 2023
Revised: 4th April 2023
Accepted: 10th April 2023
Published: 21st April 2023

KEYWORDS
Sign Language
Deaf
CNN
LSTM

INTRODUCTION

When spoken communication is impossible or undesirable, sign language is any method of communication based on body gestures, particularly those of the hands and arms. It is likely that the practice predates speech. Sign language can be as simple as grimaces, shrugs, and pointing, or it can be a carefully nuanced blend of coded manual signals complemented by facial expression and possibly words typed out in a manual alphabet. When verbal communication is impossible, such as between speakers of mutually incomprehensible languages or when one or more potential communicators are deaf, sign language can be employed to bridge the gap. That said, how many speakers can communicate or at least understand sign language? There are approximately 40,000 deaf populations registered with Social Welfare Department of Malaysia as of 2023 [1]. Meanwhile, only 1% of our country's people can actually communicate with the deaf and mute populations.

To increase the percentage of normal people to easily communicate with the deaf populations, deep learning was developed to overcome communication barrier in between normal and hearing-impaired people. Deep learnings were developed using models such as Convolution Neural Network (CNN), Keypoint Detection, VGG-16 and Yolov3 for various applications such as medical, sports and rehabilitation [2]–[6].

Based on previous researches regarding sign language recognition, these particular models have proven to have high accuracy and also efficient. With these available models we can work on developing a fully functional sign language recognition system and use it to help the society to understand sign languages and communicate with the deaf populations.

One of the most important factors to be considered is the dataset required for this model to succeed. Various dataset can be acquired from multiple country's sign language vocabulary. For example, American sign language, Indian sign language and Bahasa Isyarat Malaysia all have common vocabularies when it comes to alphabets and words.

RELATED WORK

A study suggested a deep learning-based dynamic HGR for a set of ISL terms regularly used by deaf. With realistic background and illumination conditions, a new dataset of hand gesture movies for ISL phrases has been collected. For the suggested hand gesture categorization, a hybrid model of GoogleNet network and BiLSTM sequence classifier was used, with an average accuracy of 76.21%.

A study was done where researcher created a dataset and a sign language interface system based on convolutional neural networks to translate hand movements and sign language into normal language [7]. The convolutional neural network (CNN) used in this study improves the predictability of the American Sign Language alphabet (ASLA). A novel contribution to the field of sign language recognition (SLR) is the dataset produced in this work. SLR systems might be developed using this dataset. For all of the evaluated datasets, the suggested CNN model showed good accuracy. Despite the new dataset's volume and various circumstances, it was accurate to 99.38% with great prediction and a minimal loss (0.0250).

***CORRESPONDING AUTHOR** | Mohd Azraai Mohd Razman | ✉ mohdazraai@ump.edu.my

© The Authors 2023. Published by Penerbit UMP. This is an open access article under the CC BY-NC 4.0 license.

Another research suggested using a deep learning method called Convolutional Neural Network (CNN) to identify hand motions from video or picture data. The ResNeXt-101 model is employed to categorise hand motions. The dataset was assembled using TEDx videos. In both training and testing, the proposed approach has a high accuracy of 95% to 99% for recognising gestures. Each motion in a frame is recognised and categorised. This method also keeps track of how frequently the gesture was used and aids in a more thorough analysis of the conversations. The influence of appropriate gestures on viewer count and viewer emotion were the two experimental studies that were conducted to examine audience engagement. Interesting findings show that talkers' appropriate gestures might increase the number of viewers and favourable reviews. [8]

A study by Suneetha et al., (2021) proposes a deep learning approach for multi view sign language recognition. The 8-stream convolutional neural network has 4 spatial and 4 motion streams working on 4 views [9]. The motion streams induce attention into the spatial features in each of the layers. To generate view invariant features, this work proposes VPPN, which is a multi-view feature learning network that learns the pooling process. The learned pooling process had shown the ability to draw view specific features for maximizing recognition accuracy. The proposed M2DA-Net has pushed the performance upwards on our multi view sign language dataset, KL_MV2DSL. Consequently, to investigate its effectiveness in different circumstances, benchmark action datasets were trained and tested on M2DA. The results were encouraging on some challenging datasets when compared baseline models on multi view learning. Finally, M2DA-Net with VPPN has shown capabilities to learn multi view features for video-based sign language recognition. In future, more views with small intra view variations can be approached using the same architecture.

In addition, a study that focused on the detection of gesture-based sign language offered a deep learning-based convolutional neural network (CNN) model. Compared to other CNN designs now in use, this model's compact representation delivers superior classification accuracy with fewer model parameters. In this study, VGG-11 and VGG-16 have also been trained and tested in order to assess the efficacy of this model. Two datasets have been taken into consideration for assessing performance. Both a publicly accessible American sign language (ASL) dataset and a sizable collection of Indian sign language (ISL) motions totaling 2150 pictures each are used in this study. The suggested model achieves the greatest accuracy of 99.96% and 100% for the ISL and ASL datasets, respectively. Experimental evaluation and comparison of the performance of the proposed system, VGG-11, and VGG-16 with current state-of-the-art methods are conducted [10].

ALGORITHMS IN DEEP LEARNING

Convolutional Neural Networks (CNN)

A Convolutional Neural Network (CNN) is a Deep Learning system that can take an input image, assign relevance (learnable weights and biases) to various aspects/objects in the image, and distinguish between them depicted in Figure 1. In comparison to other classification algorithms, CNN requires substantially less pre-processing. While basic approaches require hand-engineering of filters, CNN can learn these filters/characteristics with enough training.

The organization of the Visual Cortex influenced the design of a CNN, which is similar to the connectivity network of Neurons in the Human Brain. Individual neurons respond to stimuli exclusively in the Receptive Field, a small area of the visual field. A group of such fields can be stacked to encompass the full visual field.

Currently, multi view CNNs are used exclusively for either human action recognition with multi camera video data or 3D object recognition with 360-degree models. CNN have generated considerable recognition accuracies on 2D video and 3D skeletal data models. In this work, we propose to apply motion modelled attention mechanism for 2D sign language videos to enhance opportunities and apply knowledge gained through this study in real time sign language machine translator [11].

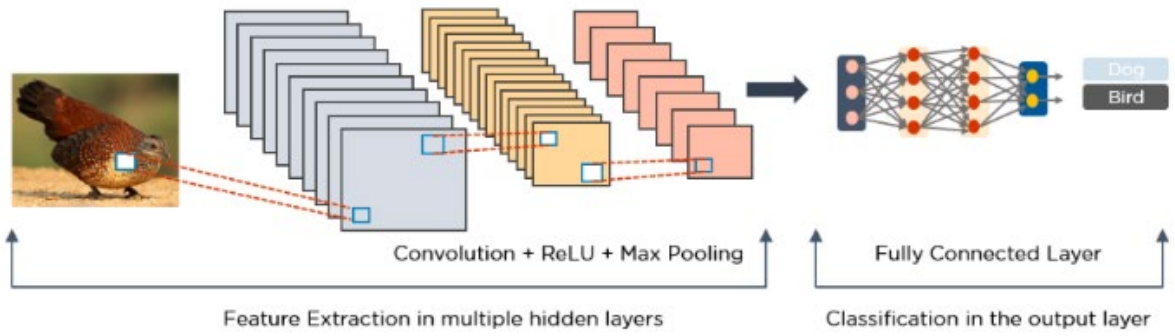

Figure 1. The working principles of Convolutional Neural Network.

Long Short Term Memory Networks (LSTMs)

Long-term dependencies can be learned and remembered using LSTMs, which are a form of Recurrent Neural Network (RNN) shown in Figure 2. The default behaviour is to recall past information over long periods of time.

LSTMs keep track of data throughout time. Because they remember prior inputs, they are valuable in time-series prediction. Four interacting layers communicate in a unique way in LSTMs, which have a chain-like structure. LSTMs are commonly employed for speech recognition, music creation, and pharmaceutical research, in addition to time-series predictions.

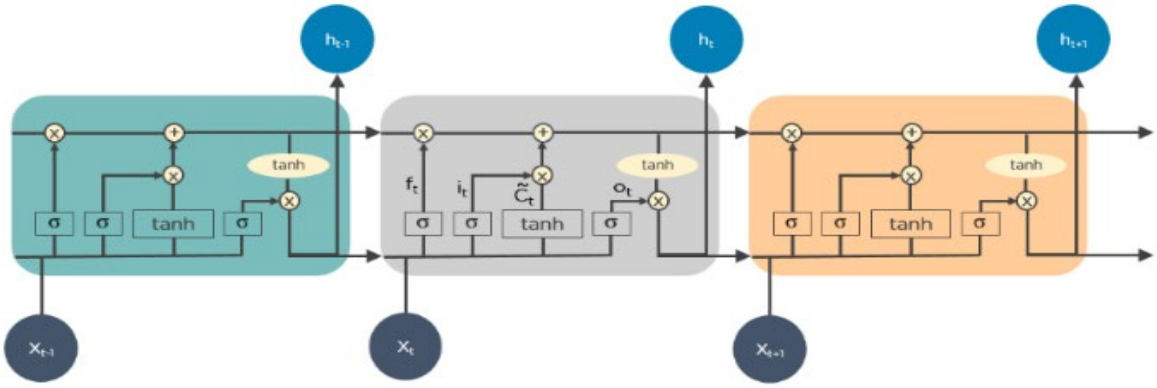

Figure 2. The working principles of Long Short Term Memory Networks (LSTMs).

You Only Look Once (YOLO)

YOLO is a neural network for performing object detection in real-time. CNNs are classifier-based systems that can process input images as structured arrays of data and identify patterns between them as in Figure 3. YOLO has the advantage of being much faster than other networks and still maintains accuracy.

It allows the model to look at the whole image at test time, so its predictions are informed by the global context in the image. YOLO and other convolutional neural network algorithms "score" regions based on their similarities to predefined classes.

High-scoring regions are noted as positive detections of whatever class they most closely identify with. For example, in a live feed of traffic, YOLO can be used to detect different kinds of vehicles depending on which regions of the video score highly in comparison to predefined classes of vehicles. The principle of YOLOV3 is to artificially divide the input image into T2 squares. Each small square generates a bounding box. If the center of the object to be detected is in a small square, then the small square will be responsible for the prediction of the object [12].

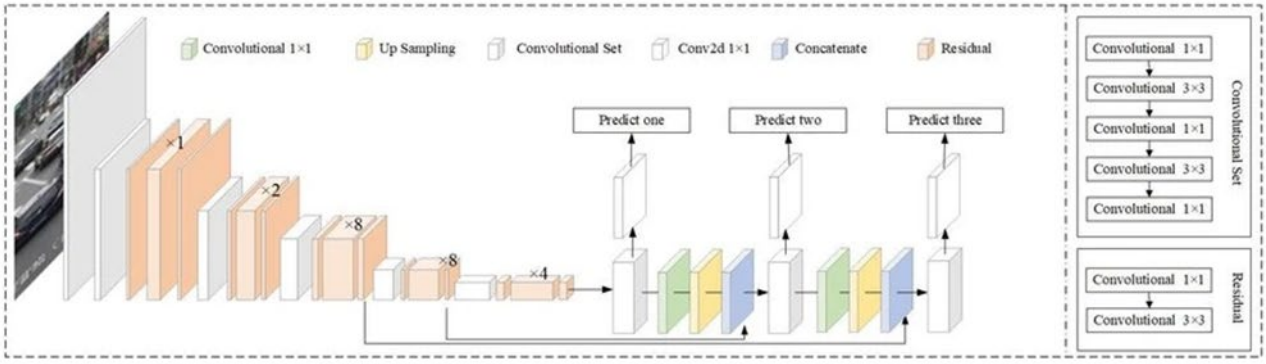

Figure 3. The working principles of You Only Look Once (YOLO).

Experimental Results

This study is able to identify the best model for Sign Language Detection through comparing a few conditions Hello, Thank You and Goodbye as in Figure 4. First is the accuracy. When compared YOLOv5 model has the highest accuracy of 97% while LSTM has 94% and Computer Vision has 66.67%. The accuracy for both LSTM model and YOLOv5 model is very close but YOLOv5 proves to have the highest accuracy. Next is the training time. The fastest training time

acquired using these models is 10 minutes which is done by YOLOv5 model. LSTM model takes 30 minutes to complete the training while Computer Vision takes 15 minutes to complete the training. Training time varies because of the number of epoch set for each models. As we can see the number of epochs set for Computer Vision is only 50 yet it takes more time to complete the training. This is because we are using website instead of our own code. While for LSTM model we use 2000 epochs and for YOLOv5 we use 1000 epochs. We can also see that YOLOv5 takes a short number of epochs to get a stable accuracy which is 75 epochs while LSTM takes 500 epochs to get a stable accuracy. Finally, is the user friendliness. User friendliness can be compared using data collection, data training and real time testing. For LSTM model we have to start with data collection every time we run the code while for Computer Vision and YOLOv5 the image is collected separately before labeling and training. As for dataset training all models use the same methods of training except for Computer Vision. LSTM model training is included in the code, Computer Vision uses Teachable Machine as a medium for training and YOLOv5 has code included for training image quality and clarity are important to ensure the accuracy of the sign language detection. It is easier to train dataset using our own code as we can make adjustments in the hyper parameter optimization compared to doing it in the website. The training accuracy obtained from the website is not accurate thus the accuracy during real time testing also varies. Lastly, the real time testing. LSTM model has training included in the code. So when we want to run the test we have start from data collection to training and after that can we only do real time testing. As for Computer Vision and YOLOv5 model the training code is separated. All we have to do is upload the weight to the code and run it to start testing. It is much easier to have it separately than all combined together. By comparing all the conditions, we can deduce that YOLOv5 is the best model for our system while LSTM can be used as backup model for this system. Figure 5 demonstrates the comparison of the accuracy obtained for each model thus finalizing the best model for this study.

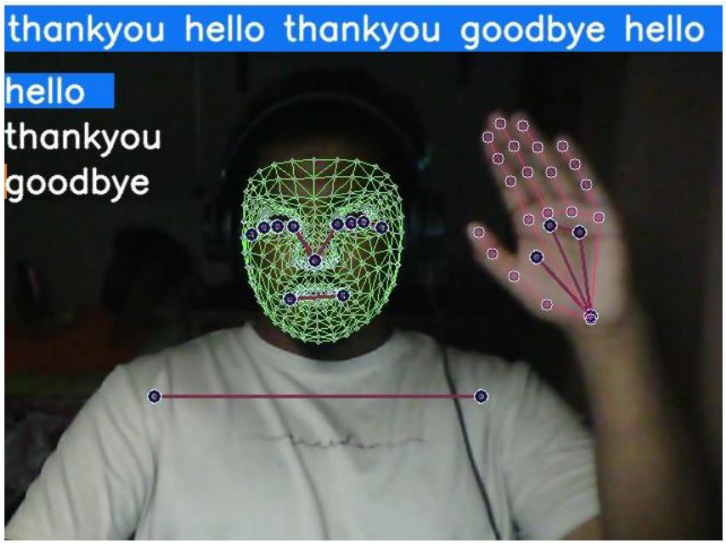

Figure 4. Real Time Test Accuracy Comparison

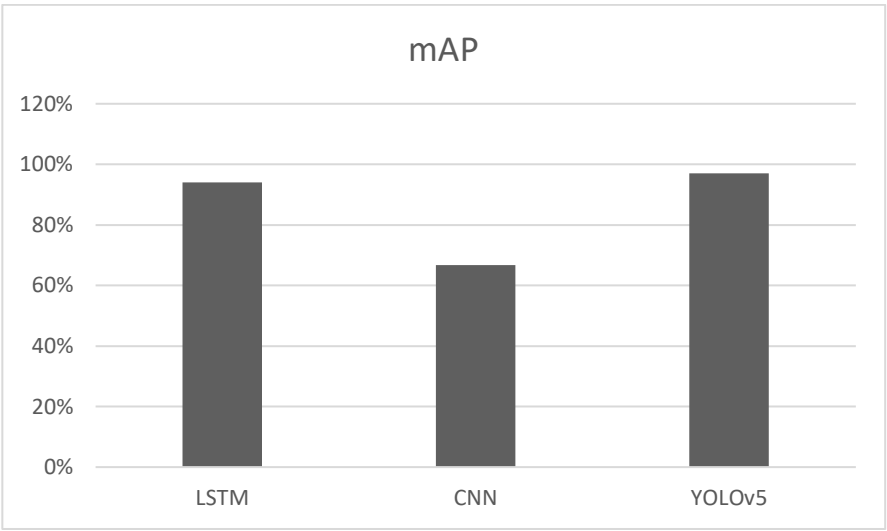

Figure 5. Real Time Test Accuracy Comparison

CONCLUSION

This paper reports the systems are successfully trained using the datase collected. Every dataset is classified and trained properly using the designated models. The real time testing results of the selected models are compared and the best model was chosen for this system using the mean average precision (mAP) obtained. YOLOv5 has the highest accuracy of 97% thus chsen as the best suitable model followed by LSTM and CNN with 94% and 66.67%

REFERENCES

[1] "Jabatan Kebajikan Masyarakat." https://www.jkm.gov.my/jkm/index.php?r=portal/full&id=ZUFHVTB1NnJWM0EreGtwNC9Vb1hvdz09 (accessed Jun. 09, 2023).

[2] Y. Wang, J. Wu, and H. Li, "Human Detection Based on Improved Mask R-CNN," 2020, doi: 10.1088/1742-6596/1575/1/012067.

[3] M. Buric, M. Pobar, and M. Ivasic-Kos, "Ball detection using yolo and mask R-CNN," *Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018*, pp. 319–323, 2018, doi: 10.1109/CSCI46756.2018.00068.

[4] A. A. Salman, I. M. Khairuddin, A. P. P. A. Majeed, and M. A. M. Razman, "The Diagnosis Of Diabetic Retinopathy By Means Of Transfer Learning And Fine-Tuned Dense Layer Pipeline," *MEKATRONIKA*, vol. 2, no. 1, pp. 68–72, Jun. 2020, doi: 10.15282/MEKATRONIKA.V2I1.6741.

[5] J. L. Mahendra Kumar *et al.*, "The classification of EEG-based winking signals: A transfer learning and random forest pipeline," *PeerJ*, vol. 9, p. e11182, Mar. 2021, doi: 10.7717/PEERJ.11182/SUPP-18.

[6] F. N. M. Noor, W. H. M. Isa, and A. P. P. A. Majeed, "The Diagnosis Of Diabetic Retinopathy By Means Of Transfer Learning With Conventional Machine Learning Pipeline," *MEKATRONIKA*, vol. 2, no. 2, pp. 62–67, Dec. 2020, doi: 10.15282/MEKATRONIKA.V2I2.6769.

[7] A. KASAPBAŞI, A. E. A. ELBUSHRA, O. AL-HARDANEE, and A. YILMAZ, "DeepASLR: A CNN based human computer interface for American Sign Language recognition for hearing-impaired individuals," *Computer Methods and Programs in Biomedicine Update*, vol. 2, p. 100048, Jan. 2022, doi: 10.1016/J.CMPBUP.2021.100048.

[8] K. Anand, S. Urolagin, and R. K. Mishra, "How does hand gestures in videos impact social media engagement - Insights based on deep learning," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100036, Nov. 2021, doi: 10.1016/j.jjimei.2021.100036.

[9] M. Suneetha, P. M.V.D., and K. P.V.V., "Multi-view motion modelled deep attention networks (M2DA-Net) for video based sign language recognition," *J Vis Commun Image Represent*, vol. 78, p. 103161, Jul. 2021, doi: 10.1016/J.JVCIR.2021.103161.

[10] S. Sharma and S. Singh, "Vision-based hand gesture recognition using deep learning for the interpretation of sign language," *Expert Syst Appl*, vol. 182, Nov. 2021, doi: 10.1016/j.eswa.2021.115657.

[11] M. Suneetha, P. M.V.D., and K. P.V.V., "Multi-view motion modelled deep attention networks (M2DA-Net) for video based sign language recognition," *J Vis Commun Image Represent*, vol. 78, Jul. 2021, doi: 10.1016/j.jvcir.2021.103161.

[12] X. Cui and R. Hu, "Application of intelligent edge computing technology for video surveillance in human movement recognition and Taekwondo training," *Alexandria Engineering Journal*, vol. 61, no. 4, pp. 2899–2908, Apr. 2022, doi: 10.1016/j.aej.2021.08.020.