Computers, Materials & Continua
DOI:10.32604/cmc.2022.022471
Article

Tech Science Press

Evaluating the Efficiency of CBAM-Resnet Using Malaysian Sign Language

Rehman Ullah Khan[1,*], Woei Sheng Wong[1], Insaf Ullah[2], Fahad Algarni[3], Muhammad Inam Ul Haq[4], Mohamad Hardyman bin Barawi[1] and Muhammad Asghar Khan[2]

[1]Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, Kuching, 94300, Malaysia
[2]Hamdard Institute of Engineering and Technology, Islamabad, 44000, Pakistan
[3]College of Computing and Information Technology, The University of Bisha, Bisha, 61922, Saudi Arabia
[4]Department of Computer Sciences, Khushal Khan Khattak University, Karak, 27200, Pakistan
*Corresponding Author: Rehman Ullah Khan. Email: rehmanphdar@gmail.com
Received: 09 August 2021; Accepted: 27 September 2021

Abstract: The deaf-mutes population is constantly feeling helpless when others do not understand them and vice versa. To fill this gap, this study implements a CNN-based neural network, Convolutional Based Attention Module (CBAM), to recognise Malaysian Sign Language (MSL) in videos recognition. This study has created 2071 videos for 19 dynamic signs. Two different experiments were conducted for dynamic signs, using CBAM-3DResNet implementing 'Within Blocks' and 'Before Classifier' methods. Various metrics such as the accuracy, loss, precision, recall, F1-score, confusion matrix, and training time were recorded to evaluate the models' efficiency. Results showed that CBAM-ResNet models had good performances in videos recognition tasks, with recognition rates of over 90% with little variations. CBAM-ResNet 'Before Classifier' is more efficient than 'Within Blocks' models of CBAM-ResNet. All experiment results indicated the CBAM-ResNet 'Before Classifier' efficiency in recognising Malaysian Sign Language and its worth of future research.

Keywords: CBAM-ResNet; malaysian sign language; within blocks; before classifier; efficiency evaluation

1 Introduction

Malaysia Sign Language or Bahasa Isyarat Malaysia in Malay was founded in 1998 when the Malaysia Federation of the Deaf (MFD) was established [1]. It is the primary sign language in Malaysia. It is used for daily communication for the deaf-mute community, including deaf people, people with hearing impairments, and physically unable to speak. The American Sign Language (ASL) has a significant influence on MSL. Although there are a few similarities between the MSL and Indonesian Sign language, both are perceived as different. Otherwise, the foundation of Indonesian Sign language was based on MSL. The communication is accomplished by interpreting the meaning

of the signer's hand gestures and, on occasion, by using appropriate facial expressions. In 2013, about 58700 people from the Malaysian population used the MSL [2].

Most of the time, the intended meaning that deaf-mutes wish to deliver throughout interaction was often misunderstood or hard to comprehend by others. Many ordinary people are not familiar with and cannot understand MSL. To master sign language is very challenging and highly depends on a person's willingness to learn. If people do not understand the MSL, they will confront problems communicating with the deaf-mutes in the nation. The implications of in effective communication had affected the deaf-mutes in their psychological, educational, employment, and social dimensions. Every one of us will need a good listener to share our feelings and thoughts. The students with hard of hearing resulted in more mental health issues than their peers who can hear [3]. Deaf people at different age groups were often related to higher distress and somatisation and felt lonely and depressed [4]. This depression may be due to their failure in interpersonal communication. In higher education, the chances of deaf-mutes interacting with teachers or lecturers may be less than others because of incommunicability. It can suppress and affect their learning experience in class. It was difficult for deaf-mutes to seek employment in Malaysia due to their inabilities in hearing and speech. A study carried out by Khoo, Tiun, and Lee revealed few cases of job discrimination, bullying, and exploitation on individuals with hearing impairment in Malaysia [5]. Sometimes, negative emotions experience by deaf-mutes made them felt ostracised by society because they were not being understood.

Before emerging sign language translation software or applications, human interpreters have relied on it as the communication bridge between deaf-mutes and people in different fields. The availability of a professional human interpreter to aid in translating sign language has become an issue as it involved the consideration of cost and time to users. Until now, there is no efficient machine learning-based sign language translation software for MSL to convert into voice or sentences into the market for public usage. These issues had caused the communication problems of some people who cannot speak or deaf individuals with talents to present their ideas to others. It was causing the loss of valuable talent. Although much research on MSL recognition that has been conducted in the past had achieved high accuracy of recognition, most of them were only focus on static-type sign language that tested with limited vocabulary. It was insufficient for a daily-use language to sustain deaf-mutes in their communication with others if the mechanism could only translate specific words or phrases. The findings from these studies lacked robustness in developing the efficient MSL translating tool. There is an open research area for the development of MSL recognition related technologies.

Hence, the studies on sign language recognition contribute to the technology used to remove communication barriers between these populations and other people. The introduction of the CBAM-ResNet method in this paper fills up the research gap in MSL recognition technology. Convolutional Based Attention Module (CBAM) [6] consists of a channel and spatial attention submodules, which are used to extend the structure and enhance the performance of Residual Network (ResNet) in videos recognition. This study emphasised this method's performance, efficiency, and practicability to produce a robust sign language translation system that benefits Malaysian deaf-mutes. Two experiments are conducted for this study, where CBAM-3DResNet is used for dynamic signs recognition. CBAM-ResNet using 3D convolution is implemented with two methods known as 'Within Blocks' and 'Before Classifier.' Efficiency evaluation of CBAM-ResNet was completed using multiple metrics such as classification accuracy, loss, precision, recall, F1-score, confusion matrix, and training time.

1.1 Significance and Contribution

The main objective of this research is to test and evaluate the efficiency of the CBAM-ResNet method using MSL. The School of Automation and Electrical Engineering (USTB) of Beijing, China, initially implement CBAM-ResNet neural network on Chinese Sign Language Recognition, which has a different network architecture from this study [7]. As sign languages vary from place to place, it is essential to determine the diversification and performance of CBAM-ResNet in terms of multi-metric before implementing it on MSL. The sub-objectives of this research are as outlined below:

- To implement the new method, which is CBAM-ResNet on Malaysian Sign Language recognition to increase the efficiency of the sign language recognising mechanism.
- To further investigate the differences between CBAM-ResNet 'Within Blocks' and 'Before Classifier' in terms of efficiency of recognising Malaysian Sign Language.
- To develop a real-time Malaysian Sign Language Recognition System through human gestures recognition using the CBAM-ResNet method.

This is the first study that adopts the CBAM-ResNet method in the context of MSL. This study introduced the CBAM-ResNet neural network to resolve the problems such as accuracy and applicability in the previous MSL recognition technology. This study is also crucial to enhance communication used by deaf-mutes and ordinary people in Malaysia to understand each other throughout their conversations. Once the efficiency of CBAM-ResNet on MSL recognition is proven, it can develop a robust MSL translating system.

1.2 Organisation

The paper is organised as follows. Section 2 briefly discusses related past studies on other sign languages recognition. The methodology of CBAM-ResNet implementation on MSL is explained in Section 3. Section 4 presents the experimental settings, results, and discussion to compare CBAM-ResNet 'Within Blocks' and 'Before Classifier' in dynamic signs recognition. Finally, Section 5 provides the conclusion for this paper.

2 Literature Review

Researchers have applied various machine learning methods in different sign language studies. These methods have their respective strengths and limitations in recognising sign languages. Generally, there are two main streams of sign languages recognition methods: vision-based and glove-based techniques. The vision-based method was relatively more convenient than the glove-based method as it does not require any wearable device, making it a hassle-free solution. However, the vision-based method still has limitations, such as the quality of camera and image used, capturing distance and direction from the camera, lighting of surroundings, accessories worn by signers, and overlapping hands in presenting sign language. These factors may affect the performance of the model. The critical evaluation parameters such as accuracy, speed of recognition, time of response, applicability, and accessibility are used to measure the efficiency of the sign language recognition algorithm.

2.1 Relevant Past Studies on Different Sign Languages Around the World

As the world is more concerned with deaf-mute's welfare, it shows positive development and a gradual increase in research associated with sign languages in recent times. Researchers worldwide had proposed different machine learning algorithms in sign languages recognition. Meanwhile, methods

implemented on sign languages recognition also change with advancements in technology, which can boost the performance of those machine learning algorithms.

2.1.1 Artificial Neural Network (ANN)

An artificial neural network (ANN) consists of nodes, simulating the neurons interconnections in biological life's brain [8]. It was usually applied to solve the problems that required data processing and knowledge representation. For example, Tangsuksant, Adhan, and Pintavirooj [9] researched American sign language static alphabets recognition by using feed forward back propagation of ANN. Their research returned an average accuracy of 95% throughout repeating experiments. Another study used the same method and achieved a higher average accuracy of 96.19%, with 42 letters of Thai Sign Language examined [10].

López-Noriega, Fernández-Valladares, and Uc-Cetina selected gloves with built-in sensors for sign alphabets recognition using ANN with Back propagation, Quick propagation, and Manhattan propagation [11]. Mehdi and Khan [12] carried out a study with seven sensors equipped with gloves and ANN architecture, which had achieved an accuracy rate of 88%. Finally, Allen, Asselin, and Foulds [13] developed a finger spelling recognition system using MATLAB for American sign language alphabets. The chosen neural network was perceptron which received amatrix with 18 rows and 24 columns as input from the 18 sensors on Cyber Glove through training. Their model got an accuracy of 90%.

2.1.2 Convolutional Neural Network (CNN)

Convolutional Neural Network is a subtype of the Feed-Forward Neural Network (FNN) suitable for images and videos processing [14]. Jalal, Chen, Moore, and Mihaylova [15] proposed an American sign language translator that did notrely on pre-trained models. They proved that their model has up to 99% recognition accuracy. It was higher than the modified CNN model from Alexnet [16]. Another study that employed CNN on an American sign language dataset with around 35000 images was carried out in India [17]. This study adapted a CNN with the topology of three convolutional layers with 32, 64, and 128 filters, max-pooling layers, and Rectified Linear Unit (ReLu) activation function [18]. Through experimental testing, their proposed system was able to achieve 96% recognition accuracy.

2.1.3 Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM)

A Recurrent Neural Network (RNN) is one of the neural networks equipped with internal memory, where its output will be mapped again into RNN for duplication. As RNN depends on inputs from previous sessions in the sequence, the duplicated antecedent elements will merge with the new input for completing decision-making tasks [19]. However, RNN usually has the problem of vanishing gradient in training. Therefore, Long Short-Term Memory (LSTM) is introduced as arefined version of RNN that can deal with this problem.

Liu, Zhou, and Li [20] suggested an LSTM-based Chinese sign language system with their self-build sign language vocabulary datasets using Microsoft Kinect 2.0. Their study returned an accuracy rate of 63.3%. Besides, RNN and LSTM are also applied in the sign language of Bahasa Indonesia with the use of Tensor Flow [21]. They extend to recognising root words attached with affixes, which vary from the original meaning and parts of speech such as 'noun' or 'verb'. A study from Indonesia implemented the 1-layer, 2-layers, and bidirectional LSTM. It achieved 78.38% and 96.15% of recognition accuracy on inflectional words and root words.

All efforts contributed by researchers in previous studies on exploring robust sign language recognition mechanisms are much appreciated.

2.2 Relevant Past Studies on Malaysian Sign Language

Fig. 1 depicted a timeline diagram for some published studies that have been done on MSL in the past 13 years, which shows future directions and trends. For example, Akmeliawati, Ooi, and Kuang [22] proposed an automatic sign language translator to recognise finger spelling and sign gestures. Another gesture recognition system for Malay sign language collected inputs from 24 sensors, consisting of accelerometers and flexure sensors connected via Bluetooth module wirelessly [23].

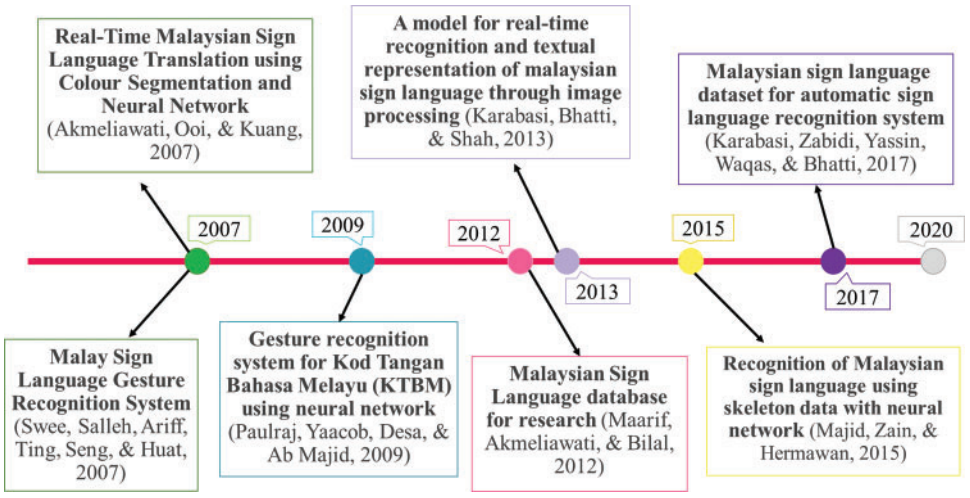

Figure 1: The timeline diagram for past studies related to Malaysian Sign Language

A gesture recognition system was developed for Kod Tangan Bahasa Melayu (KTBM) in 2009. It captured images through a webcam and then processed with Discrete Cosine Transform (DCT) to produce feature vectors [24]. This system obtained 81.07% for classification rate using an ANN model. In 2012, researchers for MSL built a well-organised database with different classifications [25]. In the consequent year, Karabasi, Bhatti, and Shah [26] proposed a model for a signs recognition system through a mobile device in real-time. Majid, Zain, and Hermawan [1] implemented ANN with Back propagation to classify skeleton data of signs obtained from Kinect sensors. They trained the network with a learning rate of 0.05 using 225 samples and achieved 80.54% accuracy on 15 dynamic signs. In 2017, Karbasi, Zabidi, Yassin, Waqas, and Bhatti [27] demonstrated a dataset development for MSL consisting of alphabets and ten (10) dynamic signs using Microsoft Kinect.

Researchers favoured ANN in recognising MSL signs. Therefore, this study implemented a CNN-based neural network called CBAM-ResNet, introducing a new classification method in MSL recognition.

3 Methodology

3.1 Convolutional-Based Attention Module (CBAM)

The strength of CNN in images and videos recognition is the availability of different convolution alkernels capable of extracting variations of features in the image. In this research, CBAM is adopted, which has two sub-modules: channel and spatial focused on detection tasks. Both attention

sub modules presented different functions.The channel attention sub-module gives prominence to representative information provided by an input image. The spatial attention sub-module focuses on the representative region that contributes to the meaningfulness of the image. In addition, both sub-modules emphasised the concept of 'What' and 'Where'. Fig. 2 shows the sequential order of these two sub-modules when processing information flow in the convolution block of the neural network.

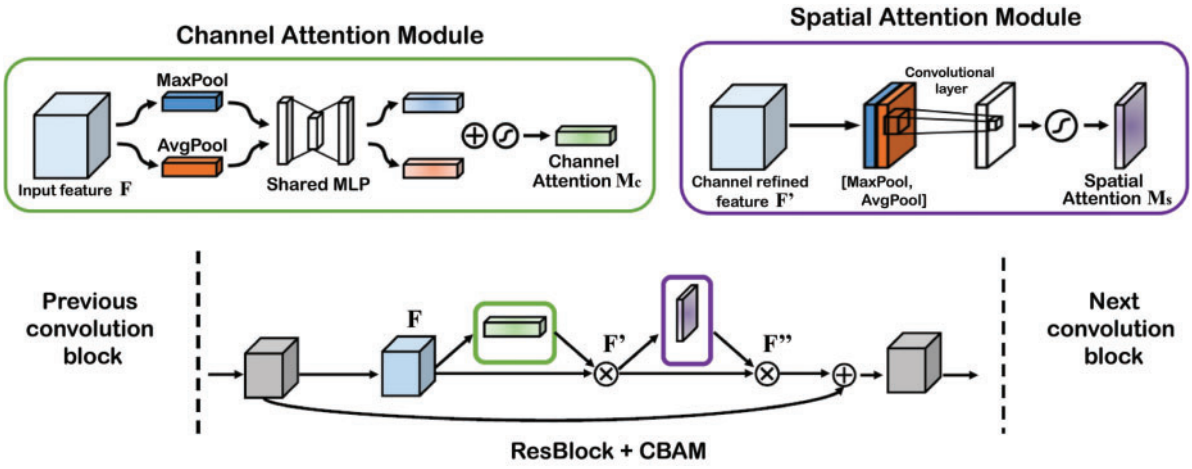

Figure 2: The sequential arrangement of CBAM channel and spatial attention submodules

The sequential order is chosen before parallel structure for both sub-modules, where input features are directed to channel attention followed by spatial attention. It was proven that sequential order generated better results [6].

In channel attention, average pooling and maximum pooling are applied separately. The input features will be directed into a multi-layer perceptron (MLP) with only one hidden layer to generate channel attention maps. The element-wise summation will combine the two output maps to compute the channel attention sub-module. Eq. (1) shows the representation of channel attention, *Mc* in symbols:

$$Mc(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{1}$$

where σ refers to the sigmoid function applied, *MLP* is the multi-layer perceptron, *Avg Pool* and *Max Pool* represented average pooling and maximum pooling, respectively.

Unlike channel attention, spatial attention sub modules apply both average pooling and maximum pooling processes along the channel axis with a convolution layer to produce a spatial attention map. At this time, MLP is not implemented. The spatial attention, *Ms* is shown in Eq. (2) below:

$$Ms(F') = \sigma(f([AvgPool(F'), MaxPool(F')])) \tag{2}$$

where f implies the convolution layer computation.

3.2 Malaysian Sign Language Videos Dataset

To acquire a dataset for CBAM-3DResNet, signers were required to present MSL in videos recording. The scope of sign language covered in this study data set is the vocabulary provided by the MSL and Deaf Studies Association or MyBIM [28]. Targeted signers are selected from approachable volunteers who practised MSL. This study adapted the videos specifications and recording techniques

of the American Sign Language Lexicon Videos Dataset [29] and Chinese Sign Language Recognition Dataset [30]. Both are open-source and reliable datasets. A Samsung Galaxy S9 Plus with dual 12 megapixels cameras was used to capture videos for this dataset collection.

The RGB videos were captured with a resolution of 1920 × 1080 pixels at 30 frames per second, focusing on the signer from frontal view at about 1 meter. This study selected neutral background with optimal lighting for the videos recording scene to prevent unnecessary elements. Sign language interpretation usually requires the facial expression and hand gestures of the signer. Therefore, the region of focus for the videos dataset is set above the signer's waist. A scene simulation is created and shown in Fig. 3 to demonstrate the videos recording process. Tab. 1 shows the words included in the videos dataset, consisting of 19 classes of MSL commonly used vocabulary. A total of 2071 verified sign videos clips were collected according to these words. In order to maintain the natural communication speed, all video shave different lengths ranged from 2 to 5 s. The videos were stored according to standard data-folder structure [31]. All videos in MPEG-4 file format had an original frame rate of 30 fps and 352 × 288 pixels resolution. Annotation files generated from CSV files bundled can be used along with utility codes to indicate the training and validation subset from this videos dataset.

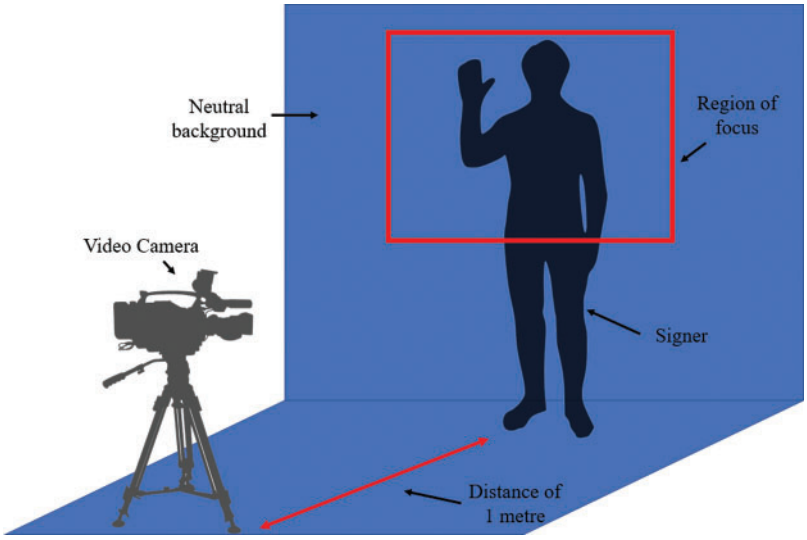

Figure 3: The simulation of the scene on Malaysian sign language dataset acquisition

Table 1: List of 19 words included in Malaysian Sign Language videos dataset

The vocabulary of the Malaysian Sign Language videos dataset				
Hello	Help	Welcome	Father	Sleep
Mother	Thank you	What	Time	Happy
Why	Water	Sit	Sick	Love
Eat	Hospital	Drink	Toilet	

3.3 Integration of CBAM into ResNet-18 Architecture

3.3.1 Within Blocks

A residual block in ResNet-18 has a depth of 2 convolutional layers. 'Within Blocks' refers to a method that plugged CBAM atevery ResNet residual block in the neural network architecture [6]. The middle 16 convolutional layers in ResNet-18 will form 8 residual block structures. This structure inferred that the 'Within Blocks' method integrated CBAM eight (8) times between these consequent residual blocks. This CBAM the residual network can refine intermediate feature maps to vital information that better represents the input. Fig. 4 shows a single residual block in CBAM-ResNet, which visualise the exact place of integrated attention module in residual network architecture using the 'Within blocks' method. CBAM is implemented at the end of the residual function, **F** in its block.

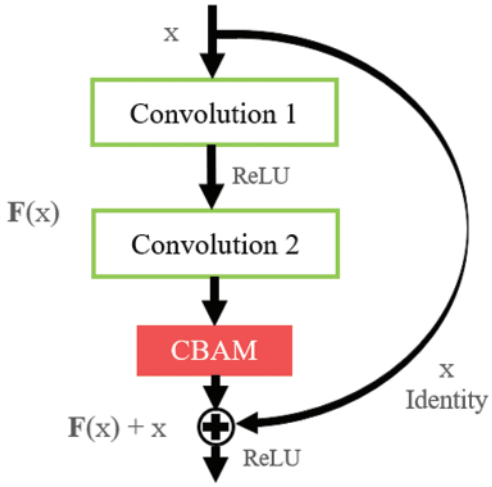

Figure 4: CBAM integrated into a residual block using the 'Within Blocks' method

3.3.2 Before Classifier

Unlike the previous method, this 'Before Classifier' technique integrated CBAM at the end part of the whole residual network, right before the average pool layer and fully connected (FC) layer. Through this implementation, CBAM will be used only once for every epoch of training, which has lower network complexity and consumes less computational cost compared to the 'Within Blocks' method. After a given input in tensor format passes along all the convolutions in a residual block of CBAM-ResNet, transforming the final feature map into the average pool and FC layers. At this stage, only the last feature map will undergo refinement by CBAM. The refined outcome will then be classified to predict the label of input. Fig. 5 shows the exact location of CBAM, which is the bottom part of CBAM-ResNet architecture using the 'Before Classifier' method.

3.4 Technical Detail of CBAM Integration into 3D-ResNet for Videos Recognition

A tensor conversion method for Mixed 3D/2D Convolutional Tube [32] was adapted into this CBAM-3DResNet implementation. By stacking batch size and depth of the tensor, a converted 5D tensor will have 4D tensor dimension in the arrangement of [(Batch size $\times$ depth), a number of channels, height, width]. Tensor transformations were performed repeatedly throughout 2D convolutions and 3D convolutions in CBAM-3DResNet.

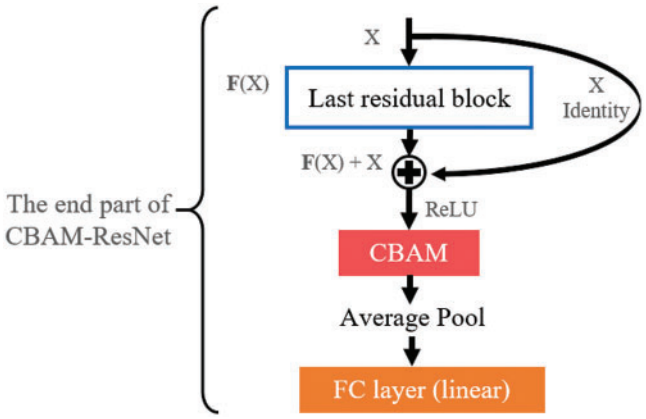

Figure 5: CBAM integrated at the end part of CBAM-ResNet using the 'Before Classifier' method

4 Experiment Settings, Results, and Discussion

4.1 Experiment Settings on Malaysian Dynamic Signs Videos Recognition

This study conducted experiments for Malaysian dynamic sign videos recognition by implementing CBAM-3DResNet. The experiment performed an efficiency evaluation and compared CBAM-3DResNet implemented using 'Within Blocks' and 'Before Classifier' methods in classifying dynamic sign videos.

The development phase used Python programming language version 3.6 with Anaconda Spyder integrated development environment and used essential Python deep learning libraries such as Pytorch, Torch vision, and Cuda Toolkit. This experiment was conducted in Google Colab with Tesla K80 GPU for CBAM-3DResNet 'Within Blocks' and 'Before Classifier'. Fig. 6 shows the summarised flow diagram of experimental procedures prepared for MSL dynamic signs videos recognition.

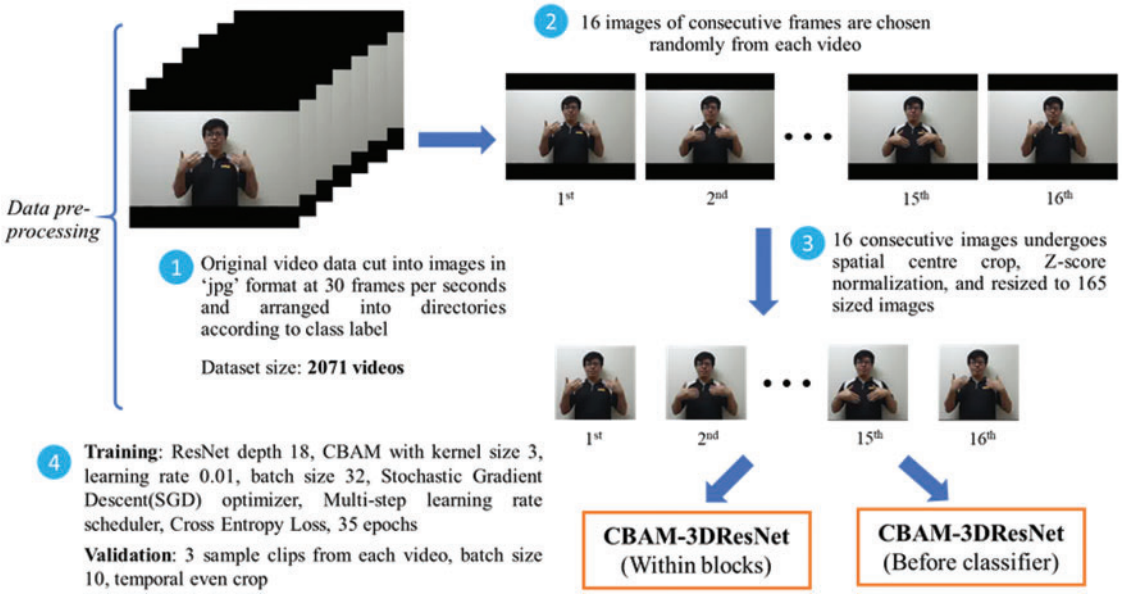

Figure 6: Summarised experimental procedures flow diagram on signs videos recognition

In the pre-processing data phase, each sign videos in the videos dataset was cut into frames in JPG format. To capture the partial pattern of the signer's hand motion in performing complete sign and increase the CBAM-3DResNet classification model's generalization ability to classify any temporal segments in videos, 16 consecutive frames were randomly selected from each video in the training subset. For the validation subset, the random temporal videos frames selection was done on alternate frames and repeated3 times, which a single sign video can provide 3 clip instances for validation. All the videos frames extracted were centre cropped and resized to 165 × 165 resolution images to remove useless data. Z-score normalisation is also applied to these images data to ensure that every image feature is weighted equally during the training process.

The processed videos frames were transformed and loaded as 5D tensors to CBAM-3DResNet implemented with 'Within Blocks' and 'Before Classifier'. Both classification models adopted similar network parameters, with 3 × 3 CBAM kernels, the learning rate of 0.01, momentum of 0.9, and batch size of 32 to train for 35 epochs using Google Colab. A Pytorch Multi-step learning rate scheduler and Stochastic Gradient Descent (SGD) optimiser were used in the training framework to decay the current learning rate when the number of epochs passed a pre-set benchmark. The best trained CBAM-3DResNet model was used to produce validation results for efficiency evaluation of video recognition experiments at clips and videos levels.

4.2 Result of Signs Videos Recognition Experiment Using CBAM-3DResNet

4.2.1 Comparison of Training and Validation CBAM-3DResNet 'Within Blocks' and Before Classifier'

Fig. 7 represented the comparison graph between training and validation loss accuracys lope of CBAM-3DResNet 'Within Block' over 35 epochs. A steep decrease was observed in the training loss slope starting from 2.876 to 0.294 over epoch 1 to epoch 7 and later changed to a steady and slowly decreasing rate since epoch 8. While the slope of validation loss fluctuated heavily within values ranged between 3.293 to 0.543 at the first 12 epochs and responded to more minor fluctuations on the epochs afterwards. These fluctuations were expected because the model optimised its learning in the earlier epochs, which resulted in slowly diminished loss luctuations at the end of epochs when it was successfully optimised. Both training and validation loss curves showed a distinctive decreasing trend in the graph, where training loss was relatively consistent compared to the zig-zag slope of validation loss. The lowest training loss was recorded at epoch 33 with a value of 0.00983, while for lowest validation loss recorded was 0.324 at epoch 32, which both values contributed a difference of 0.314.The training accuracy increased rapidly from 1.20% to 93.90% over epoch 1 to epoch 9and slowly reached its peak after epoch 9. At the same time, validation accuracy increased from 1.51% to 88.66% over epoch 1 to epoch 13 in a fluctuating pattern and turned to a moderate increasing rate on the epochs afterwards. Both training and validation accuracy plots showed different rising trends in their slopes. The highest training accuracy was achieved at epoch 33 with 99.84%, while for highest validation accuracy attained was 92.84% at epoch 32. By losses and accuracies difference of 7% exist between training and validation, it showed a little overfitting problem on CBAM-3DResNetin this videos recognition experiment.

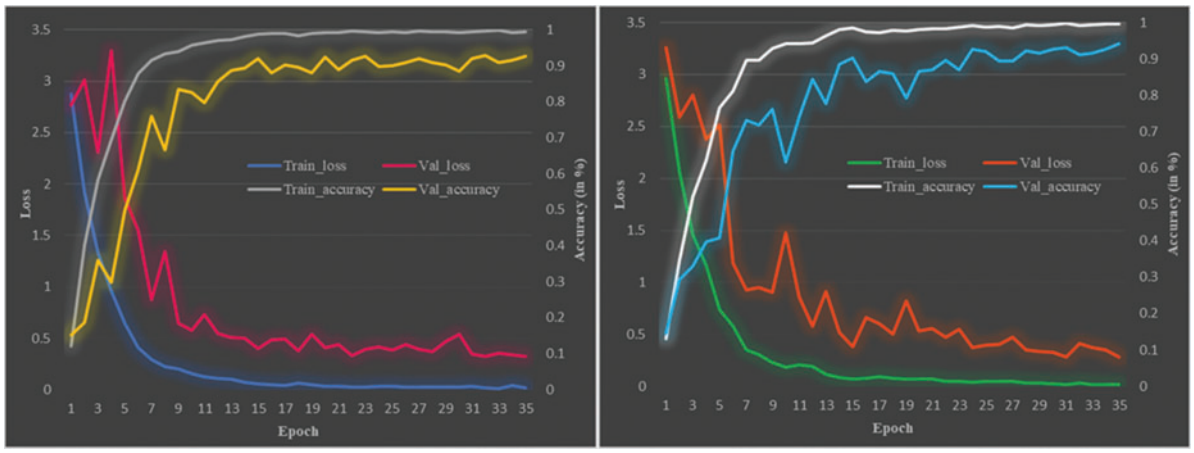

Figure 7: Training *vs.* Validation loss and accuracy of CBAM-3DResNet 'Within Blocks' and 'Before Classifier'

Fig. 7 also showed the training and validation loss accuracy comparison in the plotted line graph for CBAM-3DResNet 'Before Classifier' performed for 35 epochs. A noticeable difference in decreasing trends was found between loss of training and validation throughout epochs scheduled. The training loss decreased rapidly from 2.961 to 0.266 over epoch 1 to epoch 9 and diminished at a lower decreasing rate on later epochs. Meanwhile, the slope of validation loss reflected more fluctuations at earlier epochs. These fluctuations diminished and started converging to the point of stability as epochs continued to end. The loss fluctuations resulted in the effect of SGD optimisation on the model. The lowest loss value recorded for training was 0.152 at epoch 31, while for lowest validation loss was also recorded at the same epoch with a value of 0.278. These two-loss values had a difference of 0.126.Training accuracy increased rapidly from 12.94% to 89.53% over epoch 1 to epoch 7. It continued with a slower increase and remained stable as it was approaching the maximum.

On the other hand, validation accuracy also fluctuated heavily from epoch 1 to epoch 25 as it increased. Training and validation accuracy behaved distinct increasing trends throughout all epochs, which training accuracy only had a slight variation in its increased values. The highest accuracy for training and validation is 99.73% at epoch 31% and 94.06% at the last epoch, respectively. Both accuracies resulted in a difference of 5.67%. Tracking the losses and accuracies differences between training and validation inferred that CBAM-3DResNet 'Before Classifier' also possessed a little overfitting.

4.2.2 Top-1 and Top-5 Accuracy of Validation Result at the Videos Level CBAM-3DResNet 'Within Blocks' and Before Classifier'

Given 247 signs videos for validation, the CBAM-3DResNet 'Within Blocks' achieved 94.74% for the videos level Top-1 accuracy. To evaluate based on the 5 highest probable classes predicted, an increase of 2.02% was attained, which 96.67% achieved for the model's Top-5 accuracy at the levels of the videos. Fig. 8 showed the bar graph for each class Top-1 accuracy at the level of the videos. Among 19 dynamic sign classes, 7 classes reached the perfect video-level Top-1 accuracy. 'Sit' was the class with the lowest score of 84.61%.

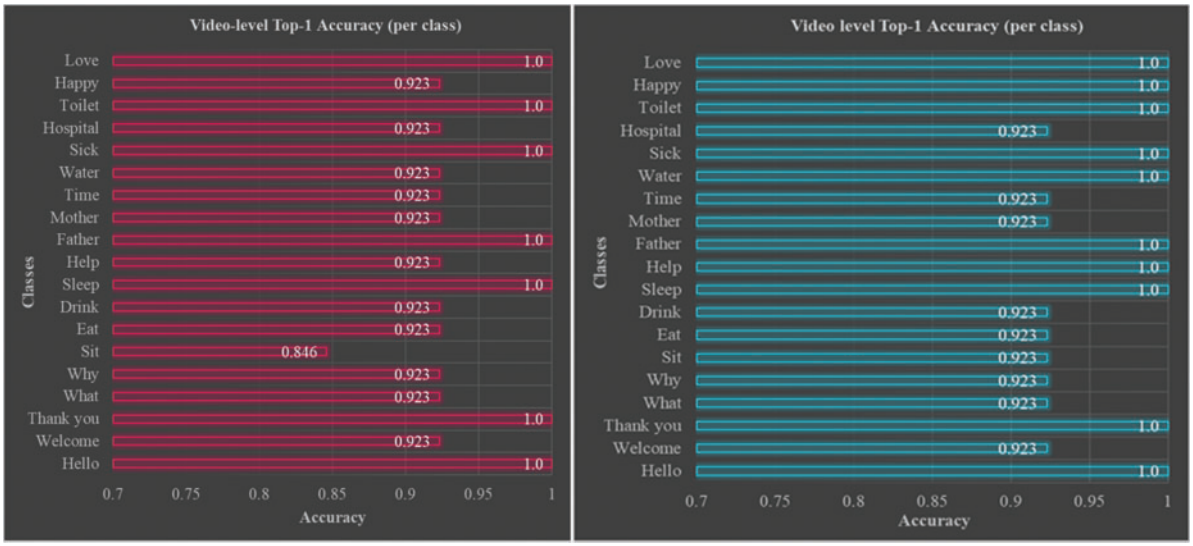

Figure 8: Videos level Top-1 accuracy of CBAM-3DResNet 'Within Blocks' and 'Before Classifier'

CBAM-3DResNet 'Before Classifier' reached the Top-1 accuracy of 96.36% on 247 sign videos. A higher score of 98.79% was achieved for Top-5 accuracy when the model's first 5 most probable predictions were used for evaluation. The bar graph of Top-1 accuracy at video level among classes was also depicted in Fig. 8. Ten (10) out of 19 dynamic sign classes achieved the perfect accuracy score of 100%, while another 9 classes ranked the same Top-1 accuracy of 92.30% at the level of the videos.

4.2.3 Classification Report and Confusion Matrix of Validation Result at Clip Level CBAM-3DResNet' Within Blocks' and Before Classifier'

The precision, recall, and F1-score of CBAM-3DResNet 'Within Blocks' and 'Before Classifier' for each dynamic sign class on validation subset at clip level was tabulated in Tab. 2. The validation subset at clip level took 3 instances from each video attributed to validate a single class. Thus each dynamic sign class had an equally distributed number of instances, with 39 sample clips listed in the column 'support'. For the 'Within Blocks' model, the class 'Hello' had the lowest precision with a value of 0.69, while classes 'Welcome', 'What', 'Eat', 'Help' and 'Hospital' obtained a perfect precision score of 1. For recall computed, two classes, 'Sit' and 'Sick' had the lowest score of 0.87, while the highest recall score of 1 belonged to classes' Thank you', 'Father' and 'Toilet'. At clip-level F1-score, class 'Hello' had the lowest value of 0.81, while class 'Toilet' achieved the highest with 0.97. The precision, recall, and F1-score for both macro average and weighted average were reported with 0.94, 0.93, and 0.93values, respectively, after round-off.

Table 2: Classification report of CBAM-3DResNet 'Within Blocks' at the clip level

	Precision		Recall		F1-score		Support
	Within blocks	Before classifier	Within blocks	Before classifier	Within blocks	Before classifier	
Hello	0.69	0.84	0.97	0.97	0.81	0.90	39
Welcome	1.00	0.97	0.92	0.90	0.96	0.93	39
Thank you	0.80	0.86	1.00	0.97	0.89	0.92	39
What	1.00	1.00	0.90	0.92	0.95	0.96	39
Why	0.97	0.97	0.92	0.92	0.95	0.95	39
Sit	0.97	0.95	0.87	0.95	0.92	0.95	39
Eat	1.00	0.97	0.90	0.92	0.95	0.95	39
Drink	0.95	1.00	0.92	0.90	0.94	0.95	39
Sleep	0.95	1.00	0.92	0.95	0.94	0.97	39
Help	1.00	0.97	0.92	1.00	0.96	0.99	39
Father	0.87	0.86	1.00	0.97	0.93	0.92	39
Mother	1.00	0.97	0.92	0.92	0.96	0.95	39
Time	0.97	0.97	0.92	0.95	0.95	0.96	39
Water	0.88	0.88	0.90	0.95	0.89	0.91	39
Sick	0.97	0.87	0.87	0.87	0.92	0.87	39
Hospital	1.00	1.00	0.92	0.92	0.96	0.96	39
Toilet	0.95	1.00	1.00	0.97	0.97	0.99	39
Happy	0.92	0.88	0.92	0.97	0.92	0.93	39
Love	0.95	0.95	0.92	0.92	0.94	0.94	39
Accuracy					0.93	0.94	741
Macro avg	0.94	0.94	0.93	0.94	0.93	0.94	741
Weighted avg	0.94	0.94	0.93	0.94	0.93	0.94	741

Meanwhile, in 'Before Classifier' model, the lowest precision score with 0.84 was computed for class 'Hello', while another 5 classes 'What', 'Drink', 'Sleep', 'Hospital' and 'Toilet' achieved a perfect precision score of 1. Class 'Sick' reported the lowest recall score of 0.87, while class 'Help' recorded the perfect recall score of 1. For the F1-score, the lowest value of 0.87 was computed for class 'Sick', while class 'Help' and 'Toilet' attained the highestscore at 0.99. The weighted average and a macro average of recall, precision, and F1-score among classes were all reported with a score of 0.94 after round-off.

Fig. 9 depicted the multi-class confusion matrix plotted for CBAM-3DResNet 'Within Blocks' classification result on the videos validation subset at the clip level. Purple-coloured diagonal cells in the confusion matrix represented the true positives of each dynamic sign class, while off-diagonal cells recorded the misclassified instances. The model had the worst prediction by misclassifying 5 clips on class 'Sit' and class 'Sick'. For example, class 'Sit' had 34 correctly classified instances, with a false-negative of 5, which 3 and 2 clips wrongly predicted as 'Hello' and 'Thank you', respectively.

Meanwhile, the best-predicted classes are ‘Thank you’, ‘Father’ and ‘Toilet’ with all instances correctly classified by the trained CBAM-3DResNet ‘Within Blocks’.

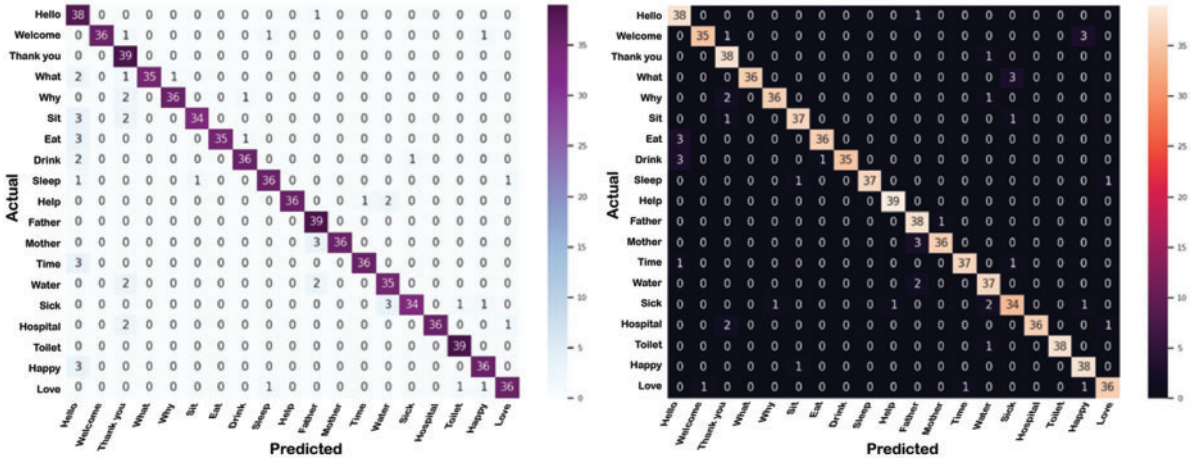

Figure 9: Confusion matrix plotted for CBAM-3DResNet’ Within Blocks’ and Before Classifier’

The multi-class confusion matrix based on the classification result of CBAM-3DResNet ‘Before Classifier’ on videos validation subset at clip level was also shown in Fig. 9. Orange-coloured cells that were diagonally oriented were true positives, while the black-coloured cells represented wrongly predicted instances. Class ‘Sick’ was the worst predicted class with a true positive of 34. Out of 39 sample clips of ‘Help’ and ‘Happy’, 2 instances were misclassified as ‘water’, and 1 instance was misclassified as ‘Why’. The only class that obtained a 100% best prediction result from CBAM-3DResNet ‘Before Classifier’ was class ‘Help’ on all its 39 instances.

5 Discussion

This section will compare the results of CBAM-3DResNet of both ‘Within Blocks’ and ‘Before Classifier’ models. A comparison table was made according to essential evaluation metrics, including training duration, lowest-level validation loss, highest-level validation accuracy, clip-level F1-score, videos-level Top-1 accuracy, videos-levelTop-5 accuracy, and network generalisation performance shown in Tab. 3.

Noticed on the videos recognition result of both CBAM-3DResNet models with different CBAM implementation methods interms of various performance metrics were comparable, except training duration, and the clip level validation accuracy of videos level Top-1, and Top-5 accuracy. Similarly, both models had minor overfitting issues for the signs videos classification tasks. The lowest validation loss that the ‘Within Blocks’ model reached was higher than the ‘Before Classifier model’ by 0.046 and correspondingly differs in their highest validation accuracies by 1.22% at the clip level. A small difference of 0.01 existed between both clip-level F1-score. CBAM-3DResNet ‘Before Classifier’ had the highest Top-1 and Top-5 accuracy than the ‘Within Blocks’ model by the difference of 1.62% and 2.12%, respectively. A comparison line graph of validation loss and accuracy slopes was plotted from both models’ results, as shown in Fig. 10.

Table 3: Comparison table for CBAM-3DResNet 'Within Blocks' and 'Before Classifier'

Metrics	CBAM-3DResNet 'Within blocks'	CBAM-3DResNet 'Before classifier'
Training duration	8141.02 s	3232.29 s
Clip level lowest validation loss	0.324	0.278
Clip level highest validation accuracy	92.84%	94.06%
Clip level F1-score achieved	0.93	0.94
Videos level Top-1 accuracy	94.74%	96.36%
Videos-level Top-5 accuracy	96.67%	98.79%
Generalisation performance	Little overfitting	Little overfitting

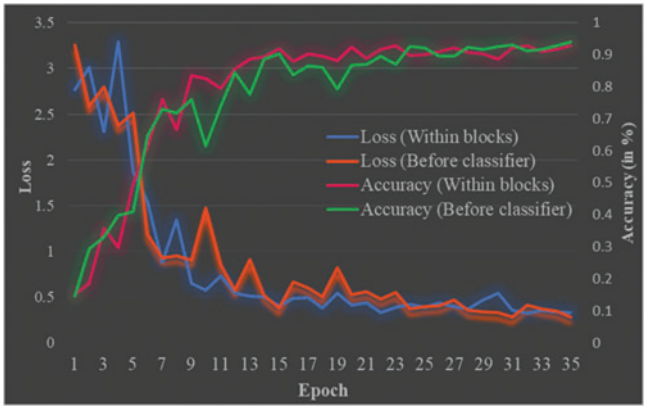

Figure 10: Comparison between CBAM-3DResNet 'Within Blocks' and 'Before Classifier' models in validation loss and accuracy at the clip level

It was observed that both models' losses and accuracy on validation videos subset at clip level showed a different pattern. Validation loss of CBAM-3DResNet 'Within Block' had the most significant fluctuation between epoch 3 and 5 and reached the peak loss value of 3.293 at epoch 4. This significant fluctuation happened because each sub-optimisation step of the SGD optimiser at earlier epochs had a big impact on losses and slowly diminished when the gradient computed become smaller. It can be noticed that both losses slopes had more minor variations in losses values at the later epoch. Both losses curves seem to have the potential to improve when set with more epochs as they still demonstrated a decreasing trend in the graph. On the other hand, the complete training duration of CBAM-3DResNet 'Before Classifier' for 35 epochs recorded only 3232.29 s due to its lower network architecture complexity, which was faster than the 'Within Blocks'model by 4908.73 s under this experiment configurations. In general, CBAM-3DResNet 'Before Classifier' is more capable than CBAM-3DResNet 'Within Blocks' in recognising dynamic signs videos.

6 Conclusion

This Research achieved its objectives and presented a new approach in MSL recognition, the CBAM-ResNet, to help Malaysian signers in their daily communications. All models implemented in this research were achieving an accuracy of more than 90%. The CBAM-ResNet 'Before Classifier' overall excels in recognition tasks on the videos dataset. The CBAM-ResNet 'Before classifier' with its strength of less computational cost, was 2.52 times speed faster in training than CBAM-ResNet 'Within Blocks' to achieve more effective classification performance on videos recognition experiments.

An overfitting issue was observed from the results of dynamic signs videos recognition experiments using CBAM-3DResNet 'Within Blocks' and 'Before Classifier'. The overfitting maybe because of the small data set as Kataoka, Wakamiya, Hara, and Satoh had concluded that a large-scale videos dataset of over 100k samples is required to successfully optimise convolution kernels in 3D CNNs architecture [33].

The concept of transfer learning can be applied in future research in coping with minor overfitting issues of CBAM-3DResNet in signs videos recognition.

Acknowledgement: We want to thank the Research, Innovation and Enterprise Centre (**RIEC**) and Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak (UNIMAS), for their support and funding of the publication of this project.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. B. A. Majid, J. B. M. Zain and A. Hermawan, "Recognition of Malaysian sign language using skeleton data with neural network," in *Proc. 2015 Int. Conf. on Science in Information Technology (ICSITech) IEEE, Yogyakarta*, Indonesia, pp. 231–236, 2015.

[2] S. B. Sajap, "Malaysian sign language translator," *Internasssstional Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1.1, pp. 385–394, 2020.

[3] P. M. Brown and A. Cornes, "Mental health of deaf and hard-of-hearing adolescents: What the students say, " *Journal of Deaf Studies and Deaf Education*, vol. 20, no. 1, pp. 75–81, 2015.

[4] R. Saha, A. Sharma and M. Srivastava, "Psychiatric assessment of deaf and mute patients–a case series," *Asian Journal of Psychiatry*, vol. 20, no. 1, pp. 31–35, 2017.

[5] S. L. Khoo, L. T. Tiun and L. W. Lee, "Workplace discrimination against Malaysians with disabilities: Living with it or fighting against it?," *Disability Studies Quarterly*, vol. 33, no. 3, pp. 75–81, 2013.

[6] S. Woo, J. Park, J. Y. Lee and I. S. Kweon, "Cbam: Convolutional block attention module, " in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 3–19, 2018.

[7] H. Chao, W. Fenhua and Z. Ran, "Sign language recognition based on cbam-resnet," in *Proc. of the 2019 Int. Conf. on Artificial Intelligence and Advanced Manufacturing*, Dublin, Ireland, pp. 1–6, 2019.

[8] J. M. Zurada, "*Introduction to Artificial Neural Systems*, " vol. 8, New York, USA, West Publishing Company, pp. 15–64, 1992.

[9] W. Tangsuksant, S. Adhan and C. Pintavirooj, "American sign language recognition by using 3d geometric invariant feature and ann classification," in *Proc.the 7th 2014 Biomedical Engineering Int. Conf., IEEE*, Fukuoka, Japan, 2014.

[10] S. Adhan and C. Pintavirooj, "Thai sign language recognition by using geometric invariant feature and ann classification," in *Proc. 2016 9th Biomedical Engineering Int. Conf. (BMEiCON), IEEE*, Laung Prabang, Laos, 2016.
[11] J. E. López-Noriega, M. I. Fernández-Valladares and V. Uc-Cetina, "Glove-based sign language recognition solution to assist communication for deaf users," in *Proc.2014 11th Int. Conf. on Electrical Engineering, Computing Science and Automatic Control (CCE), IEEE*, Ciudad del Carmen, Mexico, 2014.
[12] S. A. Mehdi and Y. N. Khan, "Sign language recognition using sensor gloves," in *Proc. of the 9th Int. Conf. on Neural Information Processing, ICONIP'02,IEEE*, Singapore, 2002.
[13] J. M. Allen, P. K. Asselin and R. Foulds, "American Sign Language Finger Spelling Recognition System," in *Proc.2003 IEEE 29th Annual Proc. of Bioengineering Conf*., IEEE, Newark, NJ, USA, 2003.
[14] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The Handbook of Brain Theory Neural Networks*, vol. 3361, no. 10, pp. 255–258, 1995.
[15] M. A. Jalal, R. Chen, R. K. Moore and L. Mihaylova, "American sign language posture understanding with deep neural networks,"in *Proc. 21st Int. Conf. on Information Fusion (FUSION)*, IEEE, Cambridge, UK, 2018.
[16] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 60, no. 6, pp. 84–90, 2012.
[17] R. Patel, J. Dhakad, K. Desai, T. Gupta and S. Correia, "Hand gesture recognition system using convolutional neural networks," in *Proc. 4th Int. Conf. on Computing Communication and Automation (ICCCA)*, IEEE, Greater Noida, India, 2018.
[18] A. F. Agarap, "Deep learning using rectified linear units (relu)," *in arXiv preprint arXiv:* 08375, 2018.
[19] A. Graves, "Supervised sequence labelling," *Supervised Sequence Labelling with Recurrent Neural Networks*, vol. 385, no. 1, pp. 5–13, 2012.
[20] T. Liu, W. Zhou and H. Li, "Sign language recognition with long short-term memory," in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, IEEE, Phoenix, AZ, USA, 2016.
[21] K. Halim and E. Rakun, "Sign language system for bahasa Indonesia (known as sibi) recogniser using tensorflow and long short-term memory," in *Proc. 2018 Int. Conf. on Advanced Computer Science and Information Systems (ICACSIS)*, IEEE, Yogyakarta, Indonesia, 2018.
[22] R. Akmeliawati, M. P. L. Ooi and Y. C. Kuang, "Real-time Malaysian sign language translation using colour segmentation and neural network," in *Proc. IEEE Instrumentation & Measurement Technology Conf. IMTC, IEEE*, Warsaw, Poland, 2007.
[23] T. T. Swee, S. H. Salleh, A. K. Ariff, C. M. Ting, S. K. Seng *et al.,* "Malay sign language gesture recognition system," in *Proc. Int. Conf. on Intelligent and Advanced Systems, IEEE*, Kuala Lumpur, Malaysia, 2007.
[24] M. P. Paulraj, S. Yaacob, H. Desa and W. M. R. W. Ab Majid, "Gesture recognition system for kod tangan bahasa melayu (ktbm) using neural network," in *Proc. 2009 5th Int. Colloquium on Signal Processing & Its Applications, IEEE*, Kuala Lumpur, Malaysia, 2009.
[25] H. A. Q. Maarif, R. Akmeliawati and S. Bilal, "Malaysian sign language database for research," in *Proc. Int. Conf. on Computer and Communication Engineering (ICCCE),IEEE*, Kuala Lumpur, Malaysia, 2012.
[26] M. Karabasi, Z. Bhatti and A. Shah, "A model for real-time recognition and textual representation of Malaysian sign language through image processing," in *Proc. Int. Conf. on Advanced Computer Science Applications and Technologies*, IEEE, 2013.
[27] M. Karbasi, A. Zabidi, I. M. Yassin, A. Waqas and Z. Bhatti, "Malaysian sign language dataset for automatic sign language recognition system," *Journal of Fundamental Applied Sciences*, vol. 9, no. 4S, pp. 459–474, 2017.
[28] J. Mak, C. Y. Vee, A. Hamidi, N. A. Venugopal, H. K. Wei *et al.,* "The Malaysian sign language and deaf studies association (mybim)," in *Mybim Official Website*, vol. 1, pp. 1, 2014.
[29] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan *et al.,* "The American sign language lexicon video dataset," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, Anchorage, AK, USA, 2008.

[30] J. Huang, W. Zhou, Q. Zhang, H. Li and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. 32nd AAAI Conf. on Artificial Intelligence (AAAI-18)*, New Orleans, Louisiana, USA, pp. 2257–2264, 2018.

[31] K. Soomro, A. R. Zamir and M. Shah, "A dataset of 101 human action classes from videos in the wild," in *Proc. Center for Research in Computer Vision*, Central Florida, USA, pp. 7, 2012.

[32] Y. Zhou, X. Sun, Z. J. Zha and W. Zeng, "Mict: Mixed 3d/2d convolutional tube for human action recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[33] H. Kataoka, T. Wakamiya, K. Hara and Y. Satoh, "Would mega-scale datasets further enhance spatiotemporal 3d cnns?,"*ArXiv Preprint ArXiv:2004.04968*, vol. 1, pp. 4968, 2020.