Fine-Tuning Frame Range in mmWave Radar Point Clouds Generation for Sign Language Gesture Recognition with Bi-LSTM Attention

KOUASSI, Rene Raymond Anocky Eric

(出版者 / Publisher)
法政大学大学院情報科学研究科

(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編

(巻 / Volume)
19

(開始ページ / Start Page)
1

(終了ページ / End Page)
6

(発行年 / Year)
2024-03-24

(URL)
https://doi.org/10.15002/00030603

Fine-Tuning Frame Range in mmWave Radar Point Clouds Generation for Sign Language Gesture Recognition with Bi-LSTM Attention

Anocky Eric Rene Raymond KOUASSI
Faculty of Computer and Information Sciences
Hosei University Tokyo, Japan
anocky.kouassi.4p@stu.hosei.ac.jp

Abstract— **This research focuses on fine-tuning the frame range for point clouds generated by a millimeter wave radar in the context of American Sign Language (ASL) gesture recognition. The study employs an IWR6843AOP millimeter wave radar, a non-contact device, to capture fifteen ASL gestures, obtaining up to 800 samples per sign gesture. Initially, the frame range varies from 11 to 120. Training an LSTM model with this initial frame range yields an overall accuracy of 88.3% for 15 gestures. By fine-tuning the frame range to [30-54] using a histogram distribution, the accuracy of sign gesture recognition improves to 89.2%. Further enhancement is achieved by adopting a Bi-LSTM learning model, resulting in an accuracy of 91.7%, utilizing approximately 700 samples per gesture as the optimal sample size. The research employs two algorithms for fine-tuning the frame range and explores three learning models: LSTM, Bi-LSTM, and Bi-LSTM with an Attention mechanism. The obtained accuracies for sign gesture recognition are 89.2%, 91.7%, and 92.30%, respectively. This study demonstrates the efficacy of optimizing the frame range and leveraging Bi-LSTM models with attention mechanisms for improved ASL gesture recognition accuracy.**

Keywords— millimeter wave radar, fine-tuning, gesture recognition, sign language, quality data, point cloud.

I. INTRODUCTION

According to the World Health Organization, over 1.5 billion people globally live with hearing loss. This number could rise to over 2.5 billion by 2050 [1]. The hearing loss problem prevents people from effective communication, which sometimes leads to social isolation. The most common means of communication is the use of hand gestures. However, some of them may not be able to access the education of sign language but for sure they can read texts.

To provide such solutions based on transcriptions, there is a growing interest in using deep learning and artificial intelligence (AI) techniques to recognize sign language gestures. The core challenge in implementing such systems lies in the effective approach to improving the recognition accuracy of the sign gestures captured in the form of sequential frame data. These sign gestures can vary in motion, speed, and orientation [2], making it difficult to determine the appropriate number of frames for training the models to better recognize sign gestures. To address this problem, this research proposes a way of fine-tuning the frame range in millimeter wave radar point cloud generation for sign language gesture recognition with Bi-LSTM Attention. The idea is to achieve an optimal frame range and training gesture sample size through the analysis of a histogram distribution before feeding those samples to the learning models. In this research, a mmWave radar sensor is adopted due to the privacy protection of users.

The roadmap of this study is the brief outline of the related work in Section II; Section III introduces the approach to recognizing the sign gesture through frame range fine-tuning, while Section IV describes the American Sign Language gesture point cloud data collection with a millimeter wave radar sensor. Section V gives the details of the frame range fine-tuning with the optimal training and validation sample size. Section VI explains the experiment results with remarks. Finally, in section VII the conclusion is drawn, and the remaining future work is addressed.

II. RELATED WORK

There exist various explorations of sensors and frame selection aiming to make effective recognition of sign language gestures. To begin with, the MIT students Pryor and Azodi in [3] used a pair of electronic gloves to capture and translate American Sign Language (ASL) into texts using statistical regression. Despite the promise of SignAloud which earned the authors the $10,000 MIT-Lemelson prize, the employed sensor was a wearable device that is intrusive and requires a continuous power supply.

As alternative solutions, Mathieu and Mieke [4] chose only RGB camera data. Their data were trained with a Visual Transformer Network (VTN) model to recognize and translate sign gestures with an achieved accuracy of 92.92%. In preprocessing data, after some visual inspection, 16 frames were selected in the middle of the initial video of each sign gesture with a temporal stride of 2 frames, for an effective temporal receptive field of 32 frames. However, frame skipping can result in sequential information loss. This is more problematic in real-time systems. What is more, data from RGB cameras reveal privacy, which is a concern.

Owoyemi and Hashimoto [5] converted point cloud frames into 3D voxels before feeding them into 3DCNN for gesture prediction. Employing such a data representation has a few drawbacks. First, 3D voxels are memory-consuming due to the need for storing 3D grids. Second, the models require much computational cost because the conversion from point clouds to 3D voxels is required before a prediction.

On the other hand, Palipana. S, Salami. D, Leiva. L.A. and Sigg. S [6] opted to improve the resolution of the point cloud through the frame of point cloud fusions, which consists of assembling (clubbing together) point clouds from adjacent frames into fewer subframes based on their time of reception. This technique reduces the total number of frames into a fixed and smaller number of frames. Following this, a combination of LSTM and PointNet++ was leveraged to recognize gestures. Although the frame fusions have the potential to improve spatial resolution and reduce the sequence length to be processed during model training, fixing the frame length can result in a decrease of temporal information or their complete loss.

Even though millimeter-wave radar point clouds have been utilized for gesture recognition and broadly in human activity recognition, it must be emphasized that less is known about extensive work aiming to utilize such data for sign language recognition. Compared to other gestures, sign language gestures are dialect and expressive which consist of complicated grammar as well as huge and different etymology (vocabularies). So, they require subtle finger coordination and body motions to effectively convey a message. A slight change in performing sign gestures can result in a wrong interpretation which can decrease the recognition accuracy of sign gestures.

To improve accuracy, this research proposes to fine-tune the frame range in point cloud generation for sign language gesture recognition with LSTM, Bi-LSTM, and Bi-LSTM Attention models. The contribution of this work is summarized as follows:

- Introducing a novel approach to improve accuracy while fine-tuning the frame range to a narrower range and at the same time obtaining the optimal sample size for efficient model learning.
- Show the potential for recognizing sign language gestures using millimeter-wave radar point clouds which is crucial for privacy preservation.
- Make the dataset publicly available for researcher communities.

III. Methodology for Improving the Accuracy of Sign Language Geture Recognition

In this research, the core methodology employs a non-contact device, a millimeter wave radar sensor to capture a sequence of sign gestures performed by different participants. Then a histogram distribution is used to fuse all gesture samples across the different signs versus the number of frames. Then two algorithms are employed to fine-tune the frame range with optimal training and validation sample sizes. Following this, those samples obtained within the fined-tuned frame range are used for training the LSTM, Bi-LSTM, and Bi-LSTM Attention mechanisms for learning. The millimeter wave radar is utilized in this work because it is non-intrusive and ensures users' privacy. mmWave radars are resilient to environmental factors such as fog and illumination, which guarantee consistent performance under diverse conditions. Moreover, the device consumes low power. Also, its compactness and portability make it a better choice for many applications. Fig. 1 shows the diagram of the overview of our approach.

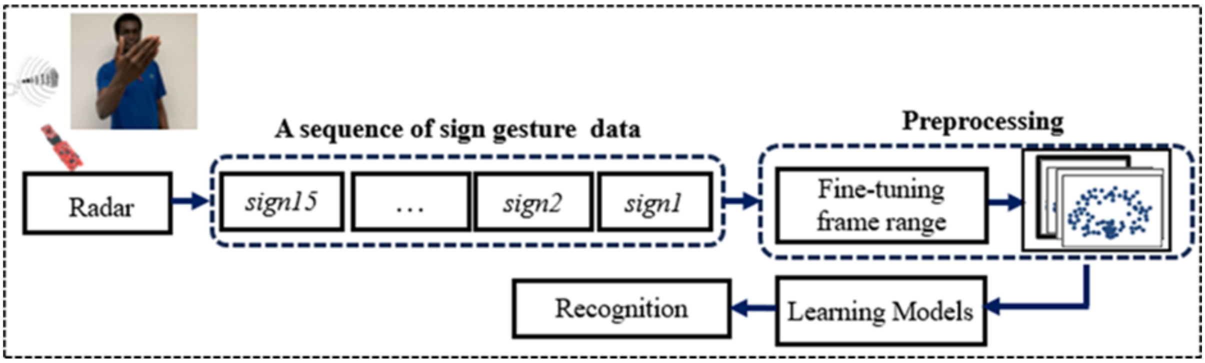


Fig. 1. Fine-tuning frame range in millimeter wave radar point clouds generation for improving sign language gesture recognition.

The above system consists of three major steps which can be summarized as data collection, frame range fine-tuning, and their evaluations through learning models.

IV. ASL Gesture Data Collection using Millimeter Wave Radar

A. Experiment Setting

In data collection for this research, the device employed was an IWR6843AOP millimeter wave radar sensor from Texas Instrument operating in the 60-64 GHz frequency range. It consists of 3 transmitting antennas(TX) and 4 receiving antennas(RX). The configuration parameters of the device are shown in Table 1. The experiment was conducted indoors, in an area of dimensions 6m*5m. The device was mounted on a stand at a height of 1.5 m with a tilt angle of 0°. This altitude and tilt angle were appropriate for capturing all the sign gesture data. This angle also canceled the need to later rotate the point cloud. On the other hand, the Field of View (FoV) was set to 110° to ensure that all the gestures performed remained in the waves transmitted by the radar. Furthermore, the width of the bounding box was set along the x-axis between -3 and 3 to prevent any interference from the walls on the signer's right and left sides. Similarly, the length of the box describing the y-axis was set between 0 and 6 to avoid radar waves reaching the walls behind the signer to reflect clutter or noise. Therefore, any value out of these range sets is considered as noise and cut off immediately. During the data collection process, it was observed that the proximity between the radar location and the signer remained within a controlled range. The minimum distance maintained between the signers and the radar was 1.5 meters, and the maximum was 6 meters.

Table 1. The radar configuration parameters

Parameters	Value
Start Frequency	60-64
Number of transmitters	3
Number of receivers	4
ADC sampling rate (MSPS)	25
Range Resolution(m)	0.084
Maximum unambiguous range(m)	7.279
Maximum Radial Velocity(m/s)	4.618
Radial velocity resolution(m/s)	0.032

B. Radar Point Cloud Generation Techniques

The point clouds are obtained through a series of Fast Fourier Transformers of the transmitted and received signal from the radar [7].

(1) Range-FFT: The radar sends a chirp signal whose frequency increases linearly over time and the mixing operation between the transmitted and received chirps produces an intermediate frequency (IF) signal. The range of the detected object is linearly proportional to

the frequency of the IF signal, which is computed using the Fast-Fourier Transform (FFT) operation.

(2) Doppler -FFT: At least two chirps are used to estimate the radial velocity of an object. The phase difference of two chirps at the range-FFT peak is proportional to the radial velocity of the detected object. An FFT operation on the range signal produces a peak at the velocity of the object.

(3) CFAR: The sum of the Doppler-FFT matrices creates a pre-detection matrix that corresponds to the detected objects.

(4) Angle-FFT: For each object, an FFT of the angle is performed on the corresponding CFAR peaks across multiple Doppler-FFTs. Velocity-induced phase changes are Doppler-corrected before computing the angle-FFT.

The points clouds are generated within sequential frames. The above steps of point cloud generations were not performed in this work. It has already been implemented in the TI module.

C. The Structure of the Frames

To capture the frames of point clouds, the Texas Instrument demo (TI) module was used for parsing data that is received at the serial port. The demo generates the point cloud and tracking information using a TLV (type-length-value) data structure scheme. For every frame, a packet is sent consisting of a fixed-sized Frame Header and then a variable number of TLVs based on what was detected in that scene. The TLVs are encoded in a data structure forming three parts which are 3D point clouds, target list objects detected in the scene, and the associated points.

D. The American Sign Language Data Collection

In this study, the American Sign Language (ASL) gestures were adopted because it is widely used by deaf communities in over 20 countries. Fifteen distinct ASL words were selected for the study and their reference videos were identified in Sign_lex [8]. All these sign words are different in meaning and they express a meaningful idea whenever they are arranged in a certain order. For example,

"Good afternoon, everyone, my name is Eric. Welcome to this research. We learn sign language with radar."

These words are finally counted as 15 sign gestures because "name" and "is" form a compound word (e.g., name_is), and "sign*" and "*language" form another compound word (e.g., sign_language*)*. The images of the different gestures are shown in Fig.2

E. The Studies of the ASL before Data Collection

The dynamics of the phonological structures such as shape, movement, location of hands, and the coordination of fingers were thoroughly studied from [8] considering the professional credibility of these sources. This primary study aimed to obtain quality data. Due to the impact related to the phonology complexity on the sign gestures, a subtle change in the to-and-from motion, repetition frequency, and movement direction can result in a wrong interpretation [8] of the gestures. Each sign has an onset offset (the start and end of the sign times respectively) which is crucial for defining the temporal length of the sign gesture. On the completion of the sign gesture studies, the participants were trained to ensure proficiency like native signers.

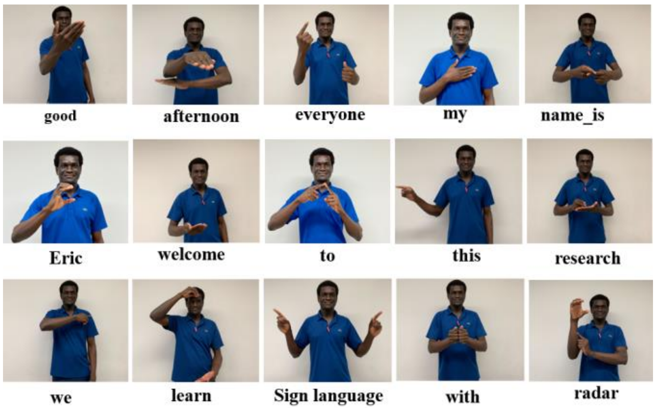


Fig. 2. Pictures of the fifteen words.

Four volunteers, one female and 3 males, aged between 21 to 36 with heights between 150 and 186 cm participated in the signing experiment. These varying gender and physical attributes make a good diversity enriching the dataset quality. During data collection, only one participant at the time stood upright in front of the radar to sign. His/her hands remained in the upper body environment with reference to the prescription [8]. What is more, to increase the spatial diversity of the dataset, the distance between the radar and the signer was often varied between 1 to 5 meters, with varying angles of 0 to 45° from the axe of the radar to the signer.

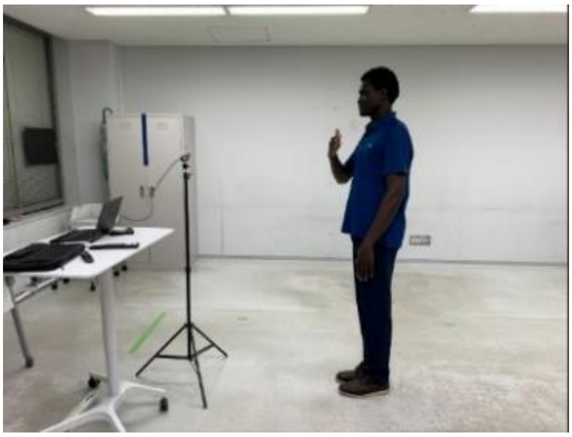


Fig. 3. A person performing a sign gesture during data collection

A software developed by Professor Huang Laboratory was used to initiate the frame capture and stop capturing whenever the gesture starts and ends respectively. After the stop, the varying frames of point clouds were saved into JSON files according to their time of arrival. Each sign's overall data contains on average 200 samples from each person. Figure 3 shows the sign language gesture data collection with a signer in front of the radar standing upright.

F. Description of the Point Cloud Generated in the Frames

The sign gesture samples consist of varying numbers of frames of point clouds. This initial number of frames ranges between 11 and 120. Each frame contains an unordered set of point cloud data represented by x, y, and z coordinates in 3D space. Each point is associated with a Doppler information value. The Doppler represents a micro-motion signature of the sign performed. The change in the coordinates values indicates the point motion in space. The coordinates and the

Doppler represent the features for training the models. The initial varying frame numbers are later fine-tuned to another range before using the related data for the learning models.

V. Fine-Tuning Frame Range with Optimal Sample Size

After data collection, comes the fine-tuning of the varying frame numbers to a range. The approach consists of three steps as follows:

- The representation of the histogram of all gesture sample counts versus(vs.) frame numbers.
- Fusion of histograms and identification of the frame range with the optimal sample sizes through Algorithm1&2.
- Applying the algorithms to the fused histograms

A. Histogram of Gestures Counts vs. Frames Numbers.

In this section, the histogram of the distribution of all sign gesture samples versus the number of frames was represented and visualized as shown in Fig. 4. The histogram data structures are dictionaries where the key represents the frame number, and the values represent the sign gesture sample counts associated with the keys. These histograms were later fused into one histogram as described in the next section *B.*

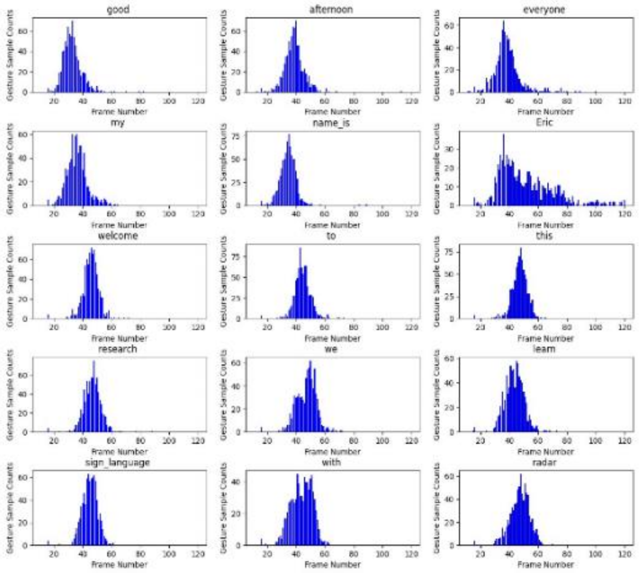


Fig. 4. The histogram of gesture samples vs the number of frames.

Fig. 4 presents histograms for 15 sign gestures, depicting the gesture sample sizes (Vs) frame numbers, with frame numbers on the x-axis ranging from 11 to 120 frames, and the number of gestures on the y-axis with the peak reaching 86 samples.

B. Fusion of Histogram

To fine-tune the frame range, all the fifteen sign gestures' histogram distributions were fused into one and visualized as shown in Fig. 5, and ***Algorithms 1&2*** were applied. The reason for this fusion was to get insights into the concentration of all samples united. The fusion process begins with the selection of the histogram containing a greater number of frames within its frame range axis. Once this base histogram has been identified, the next step involves the integration of the remaining histograms one by one. So, all the remaining samples are distributed across their equivalent frame numbers. This makes the base histogram grow in height continuously until the fusion is completed.

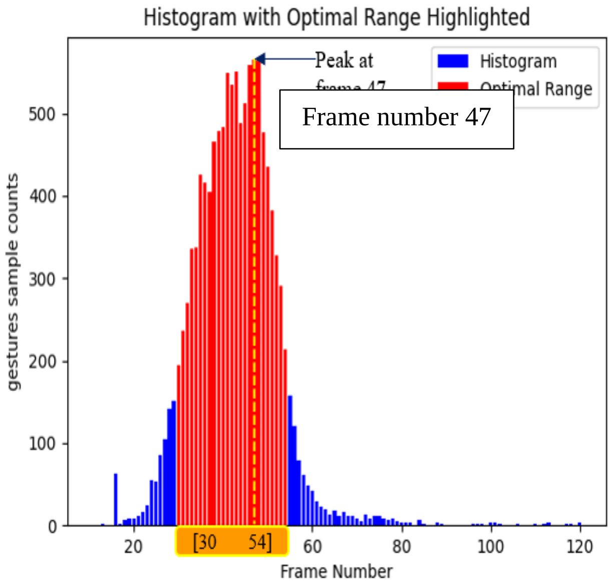


Fig. 5. The fused histogram with the fine-tuned frame range.

Fig. 5 shows the fused histogram of all sign data with a peak value of 565, occurring at frame number 47. This is an important starting point for the frame range fine-tuning with the algorithms.

C. Algorithm 1 for Fine-Tuning Frame Range

The input of Algorithm1 is the fused histogram in Fig. 5 and the output is a fine-tuned frame range. Algorithm 1 starts on the peak, at frame number 47 expanding one frame left and another right simultaneously and computes the sum of samples in the window until the optimal training and validation target sample sizes are reached. For example, the calculation is given below.

700 classes * 15 samples/class

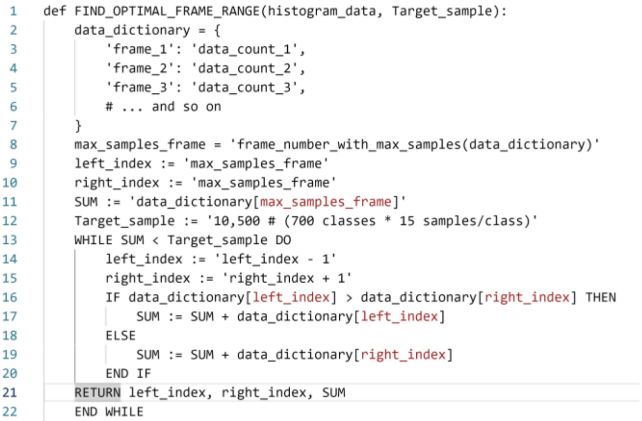
The 700 samples were decided empirically through experiments. If the training and validation sample in summation is equal to the optimal target, then the process stops otherwise the frames are further expanded and the same summation is recomputed. The summation formula is as follows:

$$\text{Optimal training data} = \sum_{\text{k=(left index)}}^{\text{right index}} values(i) \qquad (1)$$

Where values(i) represent the number of data samples at the frame number *i* and left index and right index are the expanded frames on the left- and right-hand side respectively. The resulting frame range is the range, [30-60] with an accuracy of 89%. The problem with Algorithm 1 is an over-selection due to the asymmetry of histogram distribution.

D. Algorithm 2 for Fine-Tuning Frame Range

Algorithm 2 is a revised version of ***Algorithm 1.*** They share the same frame expansion concepts. However, the key difference is the selection of the greatest sign gesture sample value by comparing the number of sign gesture samples at the expanded frames and adding that value to the summation representing the future training and validation samples in optimization. The objective is to meet the training sample requirements with a narrower frame range while solving the asymmetric distribution with respect to the peak. The pseudo-code of ***Algorithm 2*** is shown below in Fig. 5.



```
def FIND_OPTIMAL_FRAME_RANGE(histogram_data, Target_sample):
    data_dictionary = {
        'frame_1': 'data_count_1',
        'frame_2': 'data_count_2',
        'frame_3': 'data_count_3',
        # ... and so on
    }
    max_samples_frame = 'frame_number_with_max_samples(data_dictionary)'
    left_index := 'max_samples_frame'
    right_index := 'max_samples_frame'
    SUM := 'data_dictionary[max_samples_frame]'
    Target_sample := '10,500 # (700 classes * 15 samples/class)'
    WHILE SUM < Target_sample DO
        left_index := 'left_index - 1'
        right_index := 'right_index + 1'
        IF data_dictionary[left_index] > data_dictionary[right_index] THEN
            SUM := SUM + data_dictionary[left_index]
        ELSE
            SUM := SUM + data_dictionary[right_index]
        END IF
    RETURN left_index, right_index, SUM
    END WHILE
```

Fig. 5. The pseudo code of ***Algorithm 2***

Fig. 5 shows the revised algorithm for fine-tuning the frame range. The algorithm returns "left_index" and "right_index" which are respectively the lower number of and the highest in the range. Upon implementation of the algorithm, the resulting range was [30,54]. Furthermore, sign gestures whose sequence lay in this fine-tuned range were used for training the learning models.

VI. Experiment Results and Remarks

For the evaluation, 90 % of data was used for training and 10% for validation, and the Long Short-Term Memory (LSTM) network model designed by Piotr Grobelny and Adam Narbudowicz in [9] for recognizing the 12 gestures was adopted with different parameters for the recognition of the gestures within the fine-tuned range. LSTM was used considering its efficiency in sequential data modeling.

Furthermore, this LSTM model was made Bidirectional in the current work, making it possible to train backward and forward directions. The activation function was tanh, and the loss function was cross-entropy. Recognition accuracy was defined as the percentage of the correctly recognized gestures divided by the total number of tested gestures.

$$Accuracy = 100\% \left(\frac{N_{correct}}{N_{total}}\right) \qquad (2)$$

A. Integration of An Attention mechanism

An attention [15] function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. It is useful in sequence-to-sequence modeling, and it enables representing data with better context. The attention mechanism is defined as follows:

$$Attention(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \qquad (3)$$

Where Q, and K are queries and keys of dimension d_k (the dimension of k) and V, the value at the keys. The dot product of the query with all the keys is performed and divided by $\sqrt{d_k}$. Following this, the softmax function is applied to the results to scale the feature values between [0,1]. In this research, before the generation of the final output of the Bi-LSTM, the mapping result of the initial input in the hidden layer was recomputed with Attention mechanism before passing it to the fully connected layer for better interpretation. The dot product of the query and keys was performed without the scaling factor $\frac{1}{\sqrt{d_k}}$.

The model parameters were decided empirically and are summarized in Table 2. The results of the recognition accuracies are summarized in Table 3. A K-fold cross-validation method was used to confirm the consistency of the accuracy. In this case, K was equal to 5.

Table 2. Parameters of the learning models

parameters	**values**
layer	2
neurons	64
Epochs	10
Optimizer	Adam

Table 2. shows the learning model parameters which can be fine-tuned. After experimenting with several parameters, 2 layers of 64 neurons and 10 epochs were the best and the Adam optimizer was used.

Table 3. Recognition accuracy before and after optimization

	Recognition Accuracy in percentage (%)			
	Frame range	***LSTM***	***Bi-LSTM***	***Bi-LSTM Attention***
Before optimization	[11-120] 800 samples	88.3	88.7	89.86
After optimization	**[30-54] 700 samples**	**89.2**	**91.7**	**92.30%**
>=740 per class	[30-54] 740 samples	85.5	87.3	92.22%

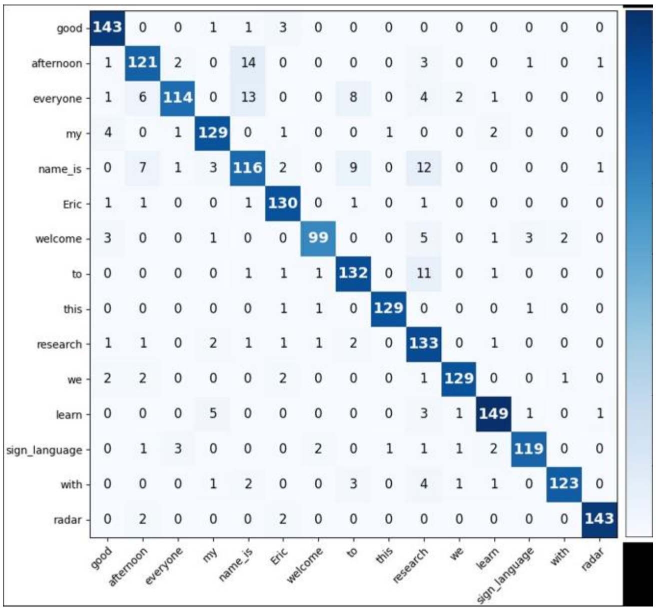
Table 3. shows the overall recognition accuracy of the 15 words ASL sign gestures before and after the fine-tuned frame range with different models. Before fine-tuning, the initial frame range was [11-120] and the recognition accuracy sign of the LSTM, Bi-LSTM, and Bi-LSTM Attention models were 88.3%, 88.7%, and 89.7% respectively. After Fine-tuning, the frame range was reduced to [30-54]. The recognition accuracies of the LSTM, Bi-LSTM, and Bi-LSTM Attention were 89.2% (+1%), 91.7% (+3%), and 92.30% (+2.44%), respectively. Only the Bi-LSTM Attention was K-fold cross-validated. The results show an improvement in the recognition accuracies across all the models after the fine-tuning. The reason for this improvement is the ability of the fine-tuning algorithm2 to extract more informative samples from the initial dataset for appropriately training the models. On the other hand, the higher recognition accuracy of the Bi-LSTM over the LSTM is because the backward and forward direction of the Bi-LSTM made it possible to capture the past and future context data sequence. As for the Bi-LSTM Attention, its superiority in the recognition accuracy over both LSTM and Bi-LSTM Attention is because the Attention mechanism made a better representation of the input sequence which led to better interpretation in the Fully connected layer.

As can be seen in Table 4, after fine-tuning, the training and validation size which initially was 12032 samples was reduced to 10535 through some adjustments. This represents a 12.46% decrease. There is a decrease in 6003 frames out of the overall frames. This significant decrease in the amount of training and validation samples as well as the frame range while improving the recognition accuracy across models proves the effectiveness of the fine-tuning approach.The statistics are shown in Table 4.

Table 4. The statistic parameters before and after optimization

	Frame range	**Training samples**	**Frame number**
Before optimization	[11-120]	12032	37918
After optimization	[30-54]	10535	31915

Figure 6 gives the confusion matrix of the fifteen-sign gesture recognition with the Bi-LSTM Attention model. From this figure, we can see a few pairs of sign gestures are confused to a certain extent observed in the matrix. This is due to the strong similarity between those gestures. However, the models can recognize many sign gestures with less confusion. For instance, the recognition of accuracy of the Bi-LSTM Attention which is 92.30% shows the efficiency of the models in better performance.



	good	afternoon	everyone	my	name_is	Eric	welcome	to	this	research	we	learn	sign_language	with	radar
good	143	0	0	1	1	3	0	0	0	0	0	0	0	0	0
afternoon	1	121	2	0	14	0	0	0	0	3	0	0	1	0	1
everyone	1	6	114	0	13	0	0	8	0	4	2	1	0	0	0
my	4	0	1	129	0	1	0	0	1	0	0	2	0	0	0
name_is	0	7	1	3	116	2	0	9	0	12	0	0	0	0	1
Eric	1	1	0	0	1	130	0	1	0	1	0	0	0	0	0
welcome	3	0	0	1	0	0	99	0	0	5	0	1	3	2	0
to	0	0	0	0	1	1	1	132	0	11	0	1	0	0	0
this	0	0	0	0	0	1	1	0	129	0	0	0	1	0	0
research	1	1	0	2	1	1	1	2	0	133	0	1	0	0	0
we	2	2	0	0	0	2	0	0	0	1	129	0	0	1	0
learn	0	0	0	5	0	0	0	0	0	3	1	149	1	0	1
sign_language	0	1	3	0	0	0	2	0	1	1	1	2	119	0	0
with	0	0	0	1	2	0	0	3	0	4	1	1	0	123	0
radar	0	2	0	0	0	2	0	0	0	0	0	0	0	0	143

Fig. 6. Confusion matrix using Bi-LSTM Attention without cross-validation.

VII. Conclusion and Future Work

This research fine-tuned the frame range from [11-120] to [30-54] of point clouds of 15 American sign language gestures captured with an IWR6843AOP millimeter. The work fuses the histogram distribution of all samples across the 15 sign gestures versus the number of frames and uses a revised algorithm to determine the area in the fused histogram where the quality and optimal data size for training and validation are concentrated. As a result, this contributed to improving the recognition accuracy of LSTM, Bi-LSTM, and Bi-LSTM Attention mechanism models with data whose sequence lies in this fine-tuned range. This significant recognition accuracy increase across the models shows the potential of recognizing sign language with mmWave radar point cloud.

For further use, this frame range fine-tuning can be applied to many other activity recognitions using sequential point cloud data as well as in other fields. For instance, in the fields where Real-time decisions are critical and where memory constraints are prominent to optimize processing time and memory. However, the fact that, after the fine-tuning, some sign gesture samples may be slightly imbalanced across classes may require additional data in the fine-tuned range for adjustments. In future work, this can be addressed while refining the current algorithm for real-time application.

Acknowledgment

I am deeply thankful to Professor Runhe Huang, my academic supervisor, for her valuable guidance and advice and many other supports without which I would not be able to complete my studies and make this achievement. Similarly, my thanks are addressed to all professors of the computer and information sciences faculty for their support in my studies and research. I am grateful to my lab colleagues Kong Lin, Watanabe Taku Guochan, and Zhou Hopu as Professor Sue for their help with data and discussions. I would also like to express my thanks to the Japan International Cooperation Agency (JICA) for sponsoring my studies and this research.

Reference

[1] World Health Organization, "Deafness and hearing loss," WHO, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (accessed in November 2023).

[2] Stokoe. Jr, W.C, "Sign language structure: An outline of the visual communication systems of the American deaf. Journal of deaf studies and deaf education", Volume 10, Issue 1, pp. 3-37, 2005

[3] Pryor. T and Azodi. N, " SignAloud: Gloves that Transliterate Sign Language into Text and Speech ", LEMELSON MIT, https://lemelson.mit.edu/sites/default/files/2020-09/SignAloud_Fact_Sheet.pdf, (Accessed: November 2023).

[4] De Coster.M, Van Herreweghe. M and Dambre. J, "Isolated sign recognition from rgb video using pose flow and self-attention", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3441-3450, 2021.

[5] Owoyemi. J and Hashimoto. K. May, " Spatiotemporal learning of dynamic gestures from 3d point cloud data", In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 5929-5934). IEEE, 2018.

[6] Palipana. S, Salami. D, Leiva. L.A. and Sigg. S, "Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds",Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, Volume 5, Issue 1, pp. 1-27, 2021.

[7] Texas Instrument, "3D people counting demo software implementation guide," Rev 2.0, March 2022.

[8] Caselli. N. K, Sevcikova Sehyr. Z. Cohen-Goldberg. A. M an Emmorey. K, "ASL-LEX: A lexical database of American Sign Language", Behavior Research Methods, Volume 4, Issue 2, https://doi.org/10.3758/s13428-016-0742-0, pp.549-556, 2016.

[9] Grobelny. P, and Narbudowicz. A, "MM-Wave radar-based recognition of multiple hand gestures using long short-term memory (LSTM) neural network". Electronics, Volume 11, Issue 5, p.787, 2022.

[10] Vaswani. A, Shazeer. N, Parmar. N, Uszkoreit. J, Jones. L, Gomez. A.N, Kaiser. Ł and Polosukhin. I, "Attention is all you need", Advances in neural information processing systems, 30, 2017.