Privacy-Preserving British Sign Language Recognition Using Deep Learning

Hira Hameed*, Muhammad Usman*, Muhammad Zakir Khan*, Amir Hussain†,
, Hasan Abbas*, Muhammad Ali Imran*, and Qammer H. Abbasi*
*James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ UK
†School of computing, Edinburgh Napier University, Scotland, UK
Email:{2683961H, m.khan.6}@student.gla.ac.uk, a.Hussain@napier.ac.uk
{muhammad.usman, hasan.abbas, qammer.abbasi, muhammad.imran}@glasgow.ac.uk

***Abstract*—Sign language is a mean of communication between the deaf community and hearing people, who use hand gestures, facial expressions, and body language to communicate. It has the same level of complexity as spoken language, but it does not employ the same sentence structure as English. The motions in sign language are made up of a range of distinct hand and finger articulations that are occasionally synchronized with the head, face, and body. Existing sign language recognition systems are mainly camera-based, which have fundamental limitations of poor lighting conditions, potential training challenges with longer video sequence data, and serious privacy concerns. This study presents a contact-less and privacy-preserving British sign language (BSL) Recognition system using Radar and deep learning algorithms, namely Inceptionv3, VGG16, and VGG19. The six most common emotions are considered, namely confused, depressed, happy, hate, lonely, and sad. The collected data is represented in the form of spectrograms. The deep learning models, InceptionV3, VGG19, and VGG16 then extract spatiotemporal features from the Spectrogram. Finally, the BSL emotions are accurately identified by classifying the Spectrograms, into the considered emotions signs. The simulation results demonstrate that a maximum classifying accuracy of 93.33% is obtained using VGG16.**

***Index Terms*—RF sensing, micro-Doppler signatures, British sign language, deep learning**

I. INTRODUCTION

Over 430 million people, which represents 5% of the world's population, live with some form of hearing impairment [1]. It is estimated that hearing loss will affect nearly 700 million people by 2050 [1]. Depending on the level of impairment, people who are hard of hearing generally rely on a sign language for communication. Sign language is a language of millions of deaf people all over the world, who use it every day to communicate and express themselves. This facilitates the integration of deaf people into society. Similar to spoken languages, different versions of sign language are used in different parts of the world. For instance, American, Japanese, Chinese, and Arabic sign languages are a few to name [2] [3]. The United Kingdom has its own sign language, known as British Sign Language (BSL). Over the years, sign language has gained a lot of attention, and its maturity has come to a level of spoken languages humans normally use [1]. However, systems to recognise a sign language remain far behind in terms of development and attention, putting deaf individuals at a distinct disadvantage, when using modern technologies. Some of the common systems to recognise sign language use coloured hands, coloured markers and gloves, Microsoft Kinect devices, gyroscopes, and cameras. For instance, the work presented in [4, 5] uses Kinect gadget and coloured gloves to automatically recognise sign language. Similarly, flex sensors and gyroscopes were employed to recognise Thai sign language by Rujira *et al.* in [6].

Camera-based systems are the most common systems to recognise sign language. This is because cameras are widely available and video images are easier to comprehend. Many existing sign language recognition (SLR) systems use two-dimensional (2D) video cameras [7]. For instance, Pigou et al. [8] investigate a deep end-to-end neural network including temporal convolutions and bidirectional recurrence for gesture identification and show that it considerably enhances frame-wise gesture recognition in video. Neverova et al. [9] present a convolutional neural network (CNN)-based architecture that learns and integrates discriminative data representations from individual channels, such as grayscale video, depth information, and skeletal joints. In [10], histogram of oriented gradients (HOG) was used to examine the photos, and an artificial neural network (ANN) was used as a BSL classifier.

However, all camera-based techniques requires recording the target, which raises serious privacy concerns. Further, poor lighting has an impact on the quality of the photographs captured and hence, BSL classification. In order to mitigate the disadvantages of camera-based system, Radar based sensing systems are proposed, which are immune to environmental lights while protecting the user's privacy. Radar-based sensing works by exploiting the Doppler signatures created on radar due to unique hand's movements. Working on these lines, [11] apply a deep CNN to classify four BSL signs using Doppler Radar. Similarly, the work presented in [12] and [13] apply CNN to design a Radar-based SLR system.

This work focuses on identifying different emotions in BSL using micro-Doppler signatures of the data collected using a Radar sensor. Six different emotions are considered namely, sad, depressed, happy, confused, hate, and lonely. These emotions are represented through dynamic sign language, which is kind of sign language where signs are identified by the mobility or movement of the hands. A ultra wide band (UWB) Radar, XeThru X4M03 was used in our experiments. The received data was represented in the form of specrtograms wherein spatiotemporal features were extracted using InceptionV3, VGG19, and VGG16 CNNs.

(a) *Sad* (b) *Depressed* (c) *Happy* (d) *Confused* (e) *Hate* (f) *Lonely*

Fig. 1: Gestures of the six considered emotions

The remainder of the paper is organised as follows: Section II discusses the adopted methodology in terms of experimental setup, hardware specification, data collection and deep learning algorithms. Simulation results are discussed in Section III. Section IV concludes our work with some future insights.

II. Methodology

This section presents the adopted methodology in terms of considered steps to design proposed BSL recognition system.

A. The Intricacy of sign language

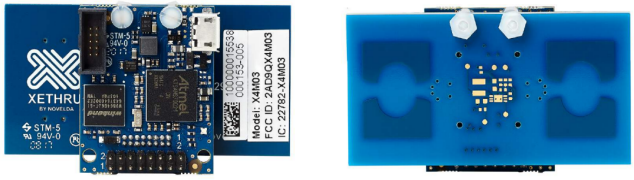
Sign languages are generally complicated requiring motions of either one hand or both. The position and the movements of hands and figures are very important to make different gestures in a sign language. The six signs chosen for this work represent a wide range of positions of hands and fingers, which guarantee that the system being developed is capable of dealing with a variety of hands' movements.

Fig1 represents BSL signs of considered emotions. It can be observed from the figure that some signs, such as depressed, happy, hate, and confused are performed double-handedly, with head and body action, while other like sad and lonely are performed one-handedly.

B. Experimental setup and data collection

The BSL recognition system proposed in this work is based on UWB Radar sensor (XeThru X4M03). The Radar is based on Novelda's X4 system-on-chip (SoC) with integrated antennas and transceiver, which provides extremely accurate measurements of an individual's distance and movement details. Fig2 depicts the front and rear view of the UWB, XeThru X4M03, Radar sensor used in this work.

For data collection, the target was standing at a distance of 1.5 meters from the Radar. Each activity presented in Fig. 1 was executed for 6 seconds. For each activity, the RF signal was transmitted and received within the specified range. The received data was stored in the form of spectrograms, with time on x-axis and Doppler[Hz] on y-axis. Spectrograms stores information about the dynamic movement of hands and head during BSL communication. The target was asked to repeat each sign pattern multiple times in order to collect significant amount of data samples. A total of 300 spectrograms were generated for six BSL signs, with 240 being utilised for training and 60 for testing. The spectrograms of all activities are illustrated in Fig. 3. The details of experimental setup and system parameters are presented in Table I.



(a) Front view (b) Rear view

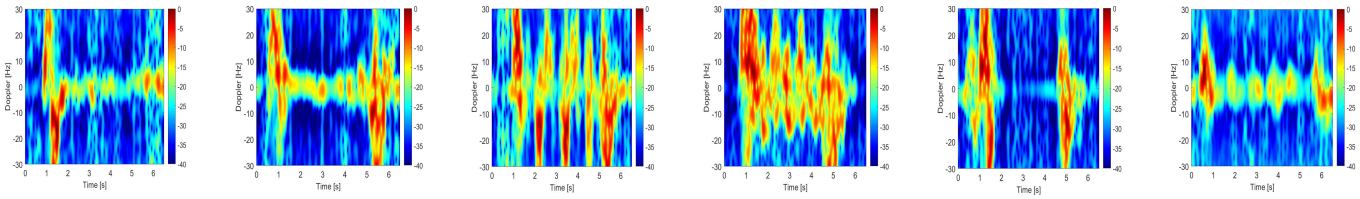
Fig. 2: XeThru X4M03 UWB RADAR sensor

Parameter	Value
Platform	Xetru Radar X4MO3
Instrumental Range	9.6 meters
Target's distance from Radar	1.5 meters
Operating Frequency	7.29 or 8.748 GHz
Receiver Gain(dB)	ETSI(14.1),KCC(12.7)
Transmitter Power	6.3 dBm
Activity duration	6 seconds
Collected samples in each class	50

TABLE I: Experimental setup and system's parameters

C. Deep Learning Architecture

The spectrograms generated from the previous step are fed into deep learning models for classification purposes. For this purpose, three different pre-trained models were considered, namely VGG16,VGG19, and InceptionV3. The high-level signal flow diagram of the proposed BSL recognition system is illustrated in Fig. 4. It can be observed from the figure that the activities were performed in standing position at 1.5 meters away from the Radar. The collected data samples were represented in the form of spectrograms. Then DL models, InceptionV3, VGG16, and VGG19 were applied to classify the collected data into six considered classes, namely sad, depressed, happy, confused, hate, and lonely. In what follows, a brief description of the considered pre-trained deep learning models is presented.



(a) *Sad* (b) *Depressed* (c) *Happy* (d) *Confused* (e) *Hate* (f) *Lonely*

Fig. 3: Obtained spectrum's sample of (a) sad, (b) depressed, (c) happy, (d) confused, (e) hate, and (f) lonely signs

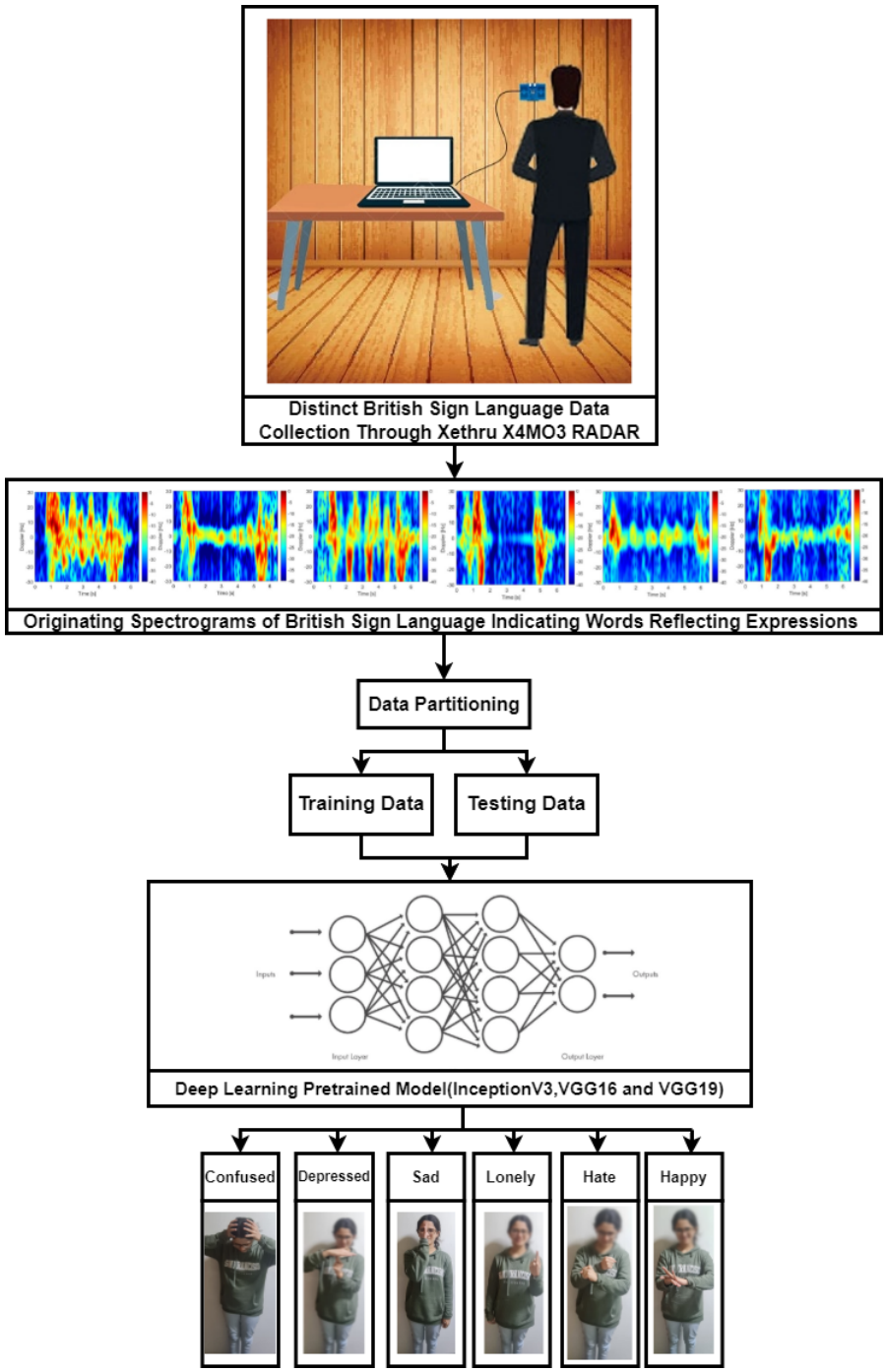


Fig. 4: Flow diagram of the proposed BSL recognition system.

InceptionV3: InceptionV3 is a deep network with 48 layers. The network starts with three convolution layers, followed by a max pooling layer, two additional convolution layers, and another max pooling layer. The image is passed to different convolutions, which are simultaneously convoluting the input images by using different filters, stacking the extracted information together, and passing it forward; this is known as inception convolution, and it is repeated multiple times throughout the network[14]. Instead of setting the filter size manually for each layer, the Inception network chooses the best filter that meets the requirement for that particular image at each time step.

VGG16: VGG16 has 16 convolution layers with the rectified linear unit (ReLU) activation function, while all kernel sizes are 3x3. After each convolution layer is followed by a max-pooling layer with all 2x2 kernel sizes. Convolution layers are used to retain training weights and act as an automatic feature extraction system. The final layer as a classifier is made up of three fully connected layers (FC). The weight of the training results can be stored by the convolution layer and FC, allowing them to calculate the number of parameters.

VGG19: VGG19 network is a 19-layer VGG network. To capture image details, VGG19 employs a 3x3 filter, which consists of five stages of convolution layers, five pooling layers, and three fully linked layers. The depth of the convolution kernel in the VGG19 network has been raised from 64 to 512, allowing for improved image feature vector extraction. A pooling layer is applied after each stage of convolutional layers. Each pooling layer has the same size and step size, which is 2x2.

III. System Evaluation and Results

This section elaborates on the system evaluation and the obtained classification results from all considered pre-trained models.

A. Evaluation criteria

The results are measured in the form of average test accuracy in classifying six different emotions using different deep learning models. Further, we calculate F1 Score, which is a combination of precision and recall. F1 score is calculated by the following equation.

$$F1 = 2\frac{Precision.Recall}{Precision + Recall} \tag{1}$$

$$Precision = \frac{\sum truepositive}{\sum truepositive + \sum falsepositive} \tag{2}$$

$$Recall = \frac{\sum truepositive}{\sum truepositive + \sum falsenegative} \tag{3}$$

B. Results and discussions

Experiments were run using test and train split techniques in which 80% data serves as training data and 20% serves as testing data. . The pre-trained models VGG16, VGG19, and IncetionV3 have 25 epochs using Adamax as the optimizer with a learning rate of 0.001. The experimental results are shown in Figs. 5, 6, and 7, where Fig. 5 illustrates the confusion matrix of InceptionV3 model in classifying the the considered classes. It can be observed from the figure that most of the classes are correctly classified having a lowest classification of accuracy of 70% for depressed class, which shows some resemblance to sad.

Similarly, the confusion matrix of VGG16 is presented in Fig. 6, where the classification accuracy is 100% for all classes except depressed, which shows resemblance to sad. This may be because of the reason that in both cases the right hand is moved closer to head. Likewise, the confusion matrix of classifying the considered emotions using VGG19 is presented in Fig. 7. Here again, most of the classes are rightly classified with an exception of depressed and sad, which shows similarities. However, 80% of test samples are correctly classified for sad class with only 20% matching with depressed class.

Fig. 5: The Confusion matrix of InceptionV3 Model.

Table II enlists the overall accuracy, precision, recall and F1 score of the considered deep learning models. It can

Fig. 6: The Confusion Matrix of VGG16 Model.

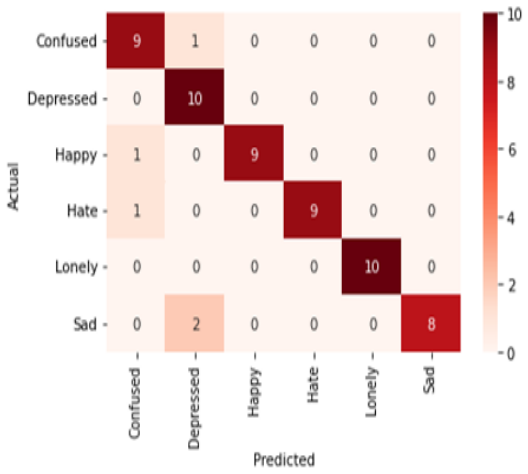


Fig. 7: The Confusion Matrix of VGG19 Model.

be observed from the Table that VGG16 outperforms other models giving overall test accuracy of 93.33%. Similarly, the same model yields the best results in terms of precision, recall and F1-score.

DL Model	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
InceptionV3	88.33	0.88	0.88	0.88
VGG16	93.33	0.95	0.93	0.93
VGG16	91.67	0.93	0.92	0.92

TABLE II: A comparison of accuracy, macro-recall, macro precision and macro-F1-score between InceptionV3, VGG16 and VGG19.

IV. CONCLUSION AND FUTURE WORK

In this work, a privacy-preserving BSL recognition system is presented using a XeThru X4M03 UWB RADAR sensor and deep learning algorithms. For BSL, six different emotion were considered, namely, sad, depressed, happy, confused, hate, and lonely. For each class micro-Doppler unique features were stored in the form of spectrograms, which were used to train three deep learning models, namely InceptionV3, VGG16, and VGG19. The classification accuracy for most of the classes was close to 100% with VGG16 outperformed others, giving overall accuracy of 93.33% on all six classes. This work produced a six-class BSL emotions dataset, which is aimed to be enhanced for future studies. The long-term goal is create a real-time version of BSL that is suitable for deaf children.

ACKNOWLEDGEMENTS

This work was supported in parts by Engineering and Physical Sciences Research Council (EPSRC) grant EP/T021063/1.

REFERENCES

[1] "WHO. (2021). Deafness and Hearing Loss". In: Accessed: 1 April 2021.

[2] G. F. Simons and C. D. Fennig. "Ethnologue: Languages of the world". In: 2017.

[3] U. Zeeshan. "Sign languages of the world," in Encyclopedia of Language and Linguistics". In: 2006, pp. 358–365. DOI: http://clok.uclan.ac.uk/9631/.

[4] Saba Jadooki et al. "Fused features mining for depth-based hand gesture recognition to classify blind human communication". In: *Neural Computing and Applications* 28.11 (2017), pp. 3285–3294.

[5] Britta Bauer and Hermann Hienz. "Relevant features for video-based continuous sign language recognition". In: *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. IEEE. 2000, pp. 440–445.

[6] R Jitcharoenpory et al. "Recognizing words in Thai Sign Language using flex sensors and gyroscopes". In: *i-CREATe2017* 4 (2017).

[7] Mohamed Mohandes, Mohamed Deriche, and Junzhao Liu. "Image-based and sensor-based approaches to Arabic sign language recognition". In: *IEEE transactions on human-machine systems* 44.4 (2014), pp. 551–557.

[8] Lionel Pigou et al. "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video". In: *International Journal of Computer Vision* 126.2 (2018), pp. 430–439.

[9] Natalia Neverova et al. "Multi-scale deep learning for gesture detection and localization". In: *European conference on computer vision*. Springer. 2014, pp. 474–490.

[10] RAHUL D RAJ and ASHISH JASUJA. "British sign language recognition using HOG". In: *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*. IEEE. 2018, pp. 1–4.

[11] James McCleary et al. "Sign Language Recognition using micro-Doppler and Explainable Deep Learning". In: *2021 IEEE Radar Conference (RadarConf21)*. IEEE. 2021, pp. 1–6.

[12] David Leslie. "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector". In: *Available at SSRN 3403301* (2019).

[13] Bruno Lepri et al. "Fair, transparent, and accountable algorithmic decision-making processes". In: *Philosophy & Technology* 31.4 (2018), pp. 611–627.

[14] *MS Windows NT Kernel Description*. http://https://machinelearningmastery.com/transfer-learning-for-deep-learning/.htm. Accessed: 2019-09-16.