Pakistan sign language to Urdu translator using Kinect

Saad Ahmed, Hasnain Shafiq, Yamna Raheel, Noor Chishti, Syed Muhammad Asad
Department of Computer Science, IQRA University, Karachi, Pakistan

Article Info

Article history:

Received Mar 10, 2022
Revised Oct 27, 2022
Accepted Nov 4, 2022

Keywords:

Kinect
Long short-term memory
Model garden
Object detection
OpenCV
Sign language translator

ABSTRACT

The lack of a standardized sign language, and the inability to communicate with the hearing community through sign language, are the two major issues confronting Pakistan's deaf and dumb society. In this research, we have proposed an approach to help eradicate one of the issues. Now, using the proposed framework, the deaf community can communicate with normal people. The purpose of this work is to reduce the struggles of hearing-impaired people in Pakistan. A Kinect-based Pakistan sign language (PSL) to Urdu language translator is being developed to accomplish this. The system's dynamic sign language segment works in three phases: acquiring key points from the dataset, training a long short-term memory (LSTM) model, and making real-time predictions using sequences through openCV integrated with the Kinect device. The system's static sign language segment works in three phases: acquiring an image-based dataset, training a model garden, and making real-time predictions using openCV integrated with the Kinect device. It also allows the hearing user to input Urdu audio to the Kinect microphone. The proposed sign language translator can detect and predict the PSL performed in front of the Kinect device and produce translations in Urdu.

Corresponding Author:

Saad Ahmed
Department of Computer Science, IQRA University
Karachi, Pakistan
Email: saadahmed@iqra.edu.pk

1. INTRODUCTION

Communication is the foundation upon which people understand each other. There are different types of communication, such as verbal communication where people engage with each other face-to-face, using their devices, and using applications such as Zoom. Another type of communication is nonverbal communication, which includes facial expressions, body poses, eye contact, hand movements, and touch. Sign language is a nonverbal type of communication. They are languages used to communicate using simultaneous hand motions, the orientation of the fingers and hands, arm or body movements, and facial expressions. It is mainly used by the hearing-impaired.

There are many kinds of sign languages. Sign language is not universal like spoken languages; it is unique and different in every country, even in countries with the same spoken language. Sign language is not an interpretation of a spoken language; it is a deaf person's native or local language. It is a natural and complete language with its own grammatical structure. In Pakistan, there are many different sign languages in the different provinces, cities, and villages, and due to this, people from different cities or villages can't communicate with each other through sign language. This proposed method focuses on Pakistan sign language (PSL) and Urdu language [1].

This sign-based communication can't be perceived by everybody; therefore, we have developed a system that will act as a bridge between the deaf and hearing people to fill the communication gap that lies between the hearing impaired and hearing people of Pakistan [2]. The majority of the working community currently cannot use sign language. It is essential for the growth of the country and workplace that deaf adults are employed and given the chance to work among people. This will help both communities come through and improve the current standards [3].

We have designed a PSL interpreter that will perform immediate sign language translations and audio-to-text translations [4]. The system's sign language module is trained upon key points that have been extracted from multiple frames using media pipe holistic. The key points are collected from a video that is captured through a Kinect device [5]. Then an LSTM model is built which is trained using the key points that have been gathered for dynamic sign language [6]. An image-based dataset is created which is trained using an object detection model garden for static sign language. After successful training PSL is then detected in real-time to perform Urdu translations. The audio to Urdu module uses the Kinect’s microphone to input Urdu audio which is then translated to Urdu text.

The design of the Chinese sign language recognition system incorporates a Specific Hand (SHS) descriptor and encoder-decoder long short-term memory (LSTM) structure for recognizing isolated Chinese sign words. The Microsoft Kinect 2.0 device is used for data input. The database is designed based on a Specific Hand Shape (SHS) descriptor utilizing a convolutional neural network (CNN). The recognition system captures the color image, depth map, and skeletal image to begin with. The hand regions and skeletal joint locations of every word of the isolated Chinese sign language are extracted from the database that was designed; this occurs after the data pre-processing process. After this, the system extracts both the features, i.e., the specific hand shape (SHS), and the trajectory. The final stage includes an encoder-decoder LSTM network that is then trained using the SHS and trajectory features and then applied for the recognition of signs [7].

Another implementation of a sign language study a Kinect-based Taiwanese sign-language recognition system has presented a solution using hidden Markov models to recognize the direction of the hands and an SVM to recognize the hand shape. Hand information is extracted that provides skeletal data from 20 joints: hip center, spine, shoulder center, head, left shoulder, left elbow, left wrist, left hand, right shoulder, right elbow, right wrist, right hand, left hip, left knee, left ankle, left foot, right hip, right knee, right ankle, and right foot. Each joint's data includes the X, Y, and Z position values. The positions of the wrist, shoulder, spine, and hip are used to localize the positions of the hands. Then the positions of the wrists are recorded as a gesture trajectory over a certain time interval. Velocity, angle, distance, and distance between the two hands of the gesture trajectory are extracted as features. HMMs are then used to recognize the hand directions from the extracted features. The position of the hand is classified into six areas as X and Y, i.e., spine position, trajectory, palm segmentation, direction recognition, hand position, and handshape. The experiment is running and yielding results of around 84% [8].

Another approach is hierarchical LSTM (HLSTM) for sign language targets to interpret video into understandable text and language to help work out vision-based sign language translation (SLT). To solve the issue of continuous sign language translation (CSLT), a hierarchical LSTM encoder-decoder model with visual content and word embedding was developed for SLT. It tackles different granularities by conveying spatio-temporal transitions among frames, clips, and viseme units. First, it uses 3D CNN to investigate the spatiotemporal cues of video clips and then packs appropriate visual themes using online key clip mining with adaptive variable length. After pooling the recurrent outputs of the top layer of HLSTM, a temporal attention-aware weighting mechanism is proposed to balance the intrinsic relationship among viseme source positions. Lastly, another two LSTM layers are used to separately retrieve verb vectors and translate semantics. After preserving original visual content with 3D CNN and the top layer of HLSTM, it shortens the encoding time step of the bottom two LSTM layers with less computational complexity while attaining more nonlinearity. The model performs well, particularly in independent tests for seen sentences with discriminative capability [9].

Another proposed technique is hybrid deep architecture, which consists of a temporal convolution module (TCOV), a bidirectional gated recurrent unit module (BGRU), and a fusion layer module (FL) to address the CSLT problem. The design is based on an end-to-end trainable network that benefits from both TCOV and BGRU modules. BGRU keeps the long-term temporal context transition pattern (global pattern), while TCOV focuses on the short-term temporal pattern (local pattern) on adjacent clip features. A fusion layer with MLP that integrates different feature embedding representations to learn the complementary relationship is proposed. It measures the mutual accommodation extent of TCOV and BGRU. The performance of the model with CTC constraints is about the same as that of other methods with multiple iterations [10].

An overview of sign language and hand gesture recognition techniques describes how they are recognized. Image processing, computer vision, and machine learning are used in many methods. Sign language covers mostly the upper body, from the waist up. The gesture approach initially yields 94%, but if the individual changes, the percentage drops to 40%, thus it is abandoned and work on alternative ways begins. 3D-modeled appearance-based hand gestures Hand gesture recognition is essential for feature extraction and categorization. Dynamic sign languages use video, while static gesture recognition uses single frames of graphics. Vision-based methods differ in data gathering. Camera frames are data. Kinect and LMC are depth-sensitive 3D cameras. Image and video inputs are modified during image preprocessing to increase performance. Segmentation depends on the image's backdrop and skin tone, making it unreliable. To increase performance, active approaches in image pre-processing change image or video inputs. IMU sensors like gyroscopes and accelerometers are utilized in data gloves for gesture and sign language detection. Wi-Fi-based gesture control is also utilized for gesture recognition. Many new works are being made utilizing these methods [11]. In the approach of isolated sign language recognition with Depth Cameras, they used a depth camera sensor with data provided by a depth camera is presented. In the introduced method, sequences of depth maps of dynamic sign language gestures are divided into smaller regions (cells). Then, statistical information is used to describe the cells. Since gesture executions have different lengths, the dynamic time warping (DTW) technique with the nearest neighbour rule is often employed for their comparison. However, due to time-consuming computations, The DTW limits the usability of the classifier [12]–[15].

2. METHOD

Pakistan sign language (PSL) to the Urdu language consists of letters, words, and sentence-level translation which is then distributed into static and dynamic sign language. Static sign language translation is achieved using tensor flow object detection model garden. We collected an image-based dataset using Kinect which was distributed into 34 classes that include Urdu letters. This data was labelled and then distributed into a set of test and train data. The model garden [16] was used for the training purpose and real-time sign language translation was performed using openCv2 which was integrated with Kinect as shown in Figure 1.

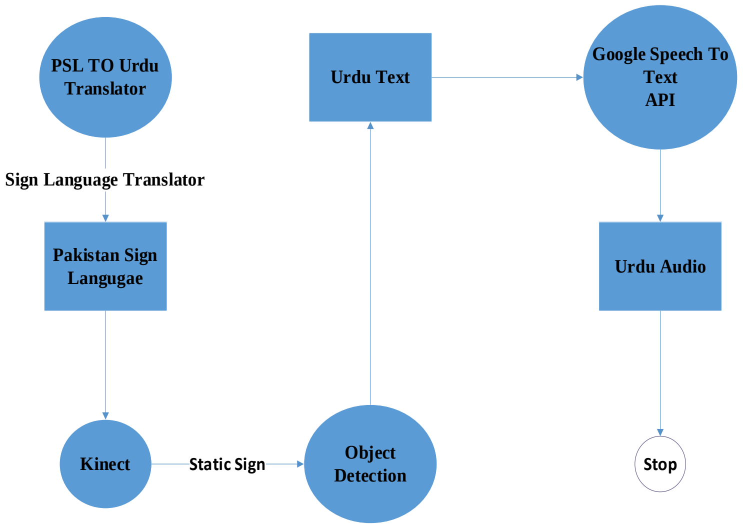
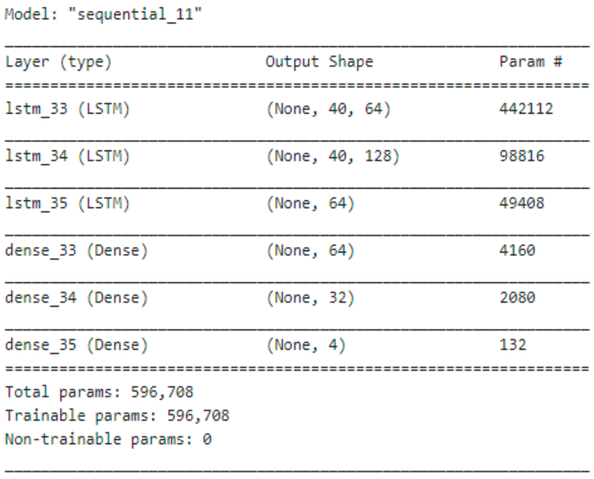


Figure 1. Static sign language translation

Dynamic sign language is achieved using googles media pipe holistic library through which we extracted the key points of both hands faces and shoulders. We created a dataset of 4 dynamic signs using Kinect in which we extracted key points through the video of 30 frames which consisted of 60 sequences.

This dataset was distributed into a set of test and train data. This data set was trained using recurrent neural network (RNN) architecture called LSTM which consisted of 3 LSTM layers and 3 dense layers as shown in Figure 2 [17], [18]. The categorical Accuracy of the model is shown in the form of a graph in Figure 3 and the model was trained for around 2000 Epochs. Epoch Loss is shown as a graph in Figure 4.

The trained dataset was saved and evaluated using confusion matrix and we achieved an accuracy of 1.0 as shown in Figure 5. Real-time dynamic sign language translation shown in Figure 6 was performed using open CV2 and ML that was integrated with Kinect. [16], [18]–[22] Urdu audio to Urdu text translation is performed using the google speech App which performs complete translation of all the letters, Words, and Sentences as shown in Figure 7.



```
Model: "sequential_11"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_33 (LSTM)               (None, 40, 64)            442112
_________________________________________________________________
lstm_34 (LSTM)               (None, 40, 128)           98816
_________________________________________________________________
lstm_35 (LSTM)               (None, 64)                49408
_________________________________________________________________
dense_33 (Dense)             (None, 64)                4160
_________________________________________________________________
dense_34 (Dense)             (None, 32)                2080
_________________________________________________________________
dense_35 (Dense)             (None, 4)                 132
=================================================================
Total params: 596,708
Trainable params: 596,708
Non-trainable params: 0
_________________________________________________________________
```

Figure 2. Model summary

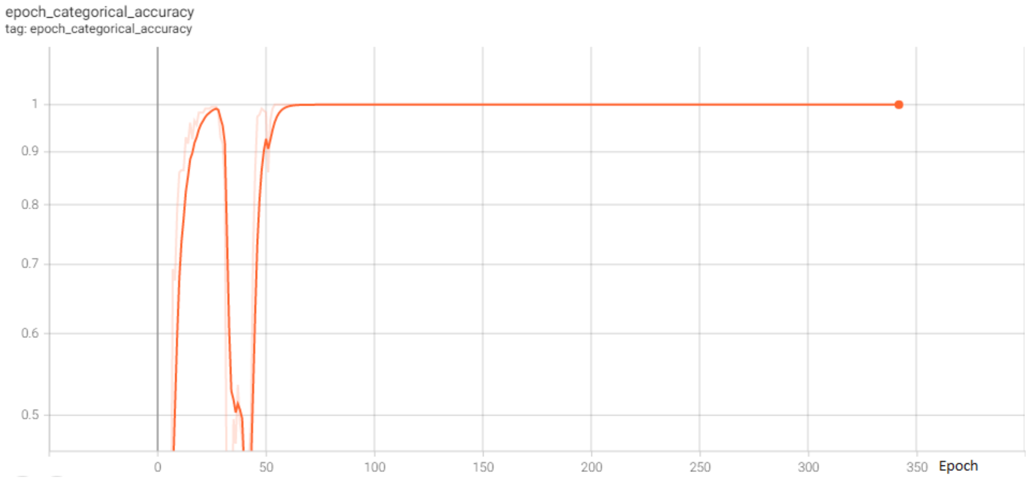


Figure 3. Categorical accuracy

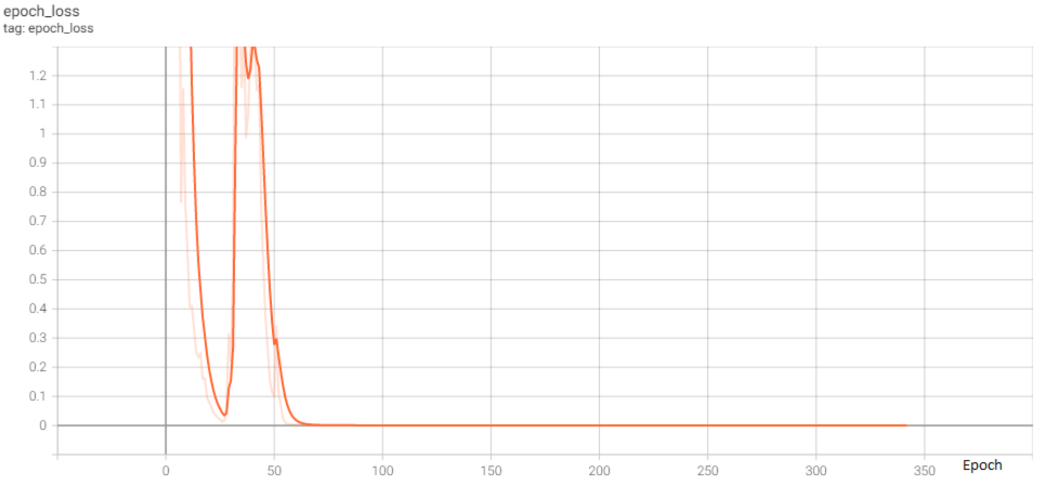
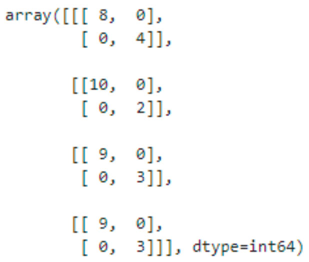


Figure 4. Epoch loss



```
array([[[ 8,  0],
        [ 0,  4]],

       [[10,  0],
        [ 0,  2]],

       [[ 9,  0],
        [ 0,  3]],

       [[ 9,  0],
        [ 0,  3]]], dtype=int64)
```

Figure 5. Confusion matrix

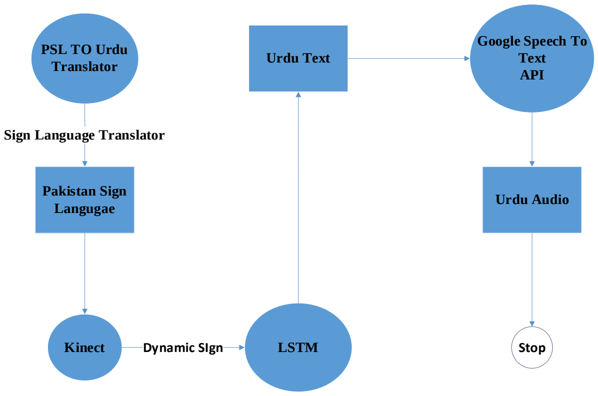


Figure 6. Dynamic sign language translation

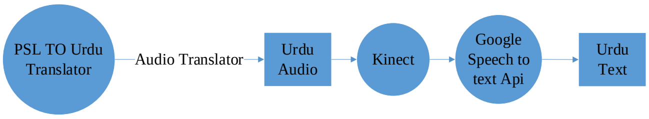


Figure 7. Audio to Urdu Translator

The text gathered from sign language is then converted into audio using Google Text to Speech App which helps in converting Urdu text into Urdu audio and a complete system of PSL to Urdu Translator is formed as shown in Figure 8. This is a very user friend system [23] which can facilitate both hearing impaired person [24] and the normal person to communicate with each other without facing any challenges [25].

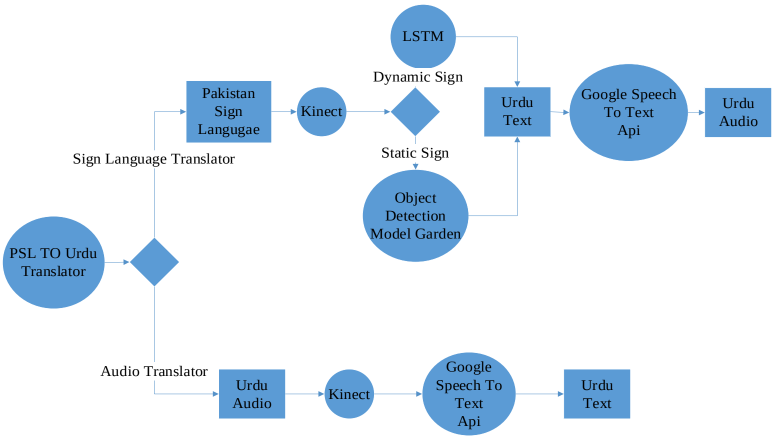


Figure 8. PSL to Urdu Translator

3. RESULTS AND DISCUSSION

This system has attempted to give a solution to the barrier of communication faced by the Pakistani deaf and dumb society by developing a sign language translator application with a user-friendly GUI and better functionalities. It uses a novel approach for PSL recognition using Kinect sensors. A vast amount of videos or sequences and frames for acquiring the key points is used to develop the dataset for training and testing. The dataset can be made in real-time and existing videos or datasets can be used. The dataset consisting of the key points for the training is then further transformed into NumPy arrays.

The dataset is then split into test and train sets. The data is then fed to an LSTM network to train upon. After successful training sign language prediction can be made. Some of the real-time inputs and results achieved are shown in Figure 9 and Table 1. The same technique of key point extraction has been used to make real-time sign language predictions afterward as well. The key points of the user are collected by the Kinect sensors which capture the sequences of the facial, hands, and pose landmarks to be processed frame by frame to match with the training datasets. After the user has successfully performed signs, real-time sign language predictions are made. These predictions are then viewed by the user in the form of Urdu text and Urdu audio.

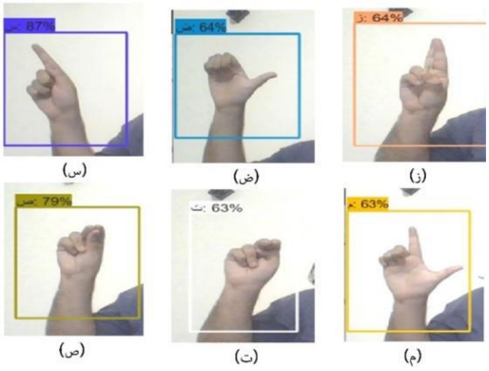


Figure 9. Real-time input sign language to the model

Table 1. Accuracy achieved

Sign language alphabet	Result in %
ا	80%
ب	85%
ت	63%
خ	65%
س	87%
ص	79%
ز	64%
ض	64%
م	63%
ل	85%

4. CONCLUSION

In this research work, we have proposed a methodology to help the deaf community in Pakistan, we have designed and developed a framework to solve the problem hearing impaired people face to communicate with normal people. The purpose behind this is to help reduce the struggle of the hearing-impaired people of Pakistan and make them a more useful part of our society. The solution is simple, effective, and affordable. This proposed system was tested in the Computer Science laboratory of IQRA University. Experimental results have shown that this KINECT-based system has shown promising results and is reliably meeting the requirements to solve the communication problem faced by hearing-impaired people.

REFERENCES

[1] H. Tahir Jameel and S. Bibi, “Benefits of sign language for the deaf students in classroom learning,” *Article in International Journal of ADVANCED AND APPLIED SCIENCES*, vol. 3, no. 6, pp. 24–26, 2016.

[2] M. Burton, “Evaluation of sign language learning tools: Understanding features for improved collaboration and communication between a parent and a child,” *Iowa State University*, 2013, doi: 10.31274/etd-180810-3593.

[3] A. Van Staden, G. Badenhorst, and E. Ridge, "The benefits of sign language for deaf learners with language challenges," *Per Linguam*, vol. 25, no. 1, 2011, doi: 10.5785/25-1-28.

[4] Suharjito, R. Anderson, F. Wiryana, M. C. Ariesta, and G. P. Kusuma, "Sign language recognition application systems for deaf-mute people: A review based on input-process-output," *Procedia Computer Science*, vol. 116, pp. 441–448, 2017, doi: 10.1016/j.procs.2017.10.028.

[5] Hee-Deok Yang, "Sign language recognition using kinect," *Journal of Advanced Engineering and Technology*, vol. 8, no. 4, pp. 299–303, 2015, doi: 10.35272/jaet.2015.8.4.299.

[6] V. Hernandez, T. Suzuki, and G. Venture, "Convolutional and recurrent neural network for human activity recognition: Application on American sign language," *PLoS ONE*, vol. 15, no. 2, 2020, doi: 10.1371/journal.pone.0228869.

[7] X. Li, C. Mao, S. Huang, and Z. Ye, "Chinese sign language recognition based on SHS descriptor and encoder-decoder LSTM model," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10568 LNCS, pp. 719–728, 2017, doi: 10.1007/978-3-319-69923-3_77.

[8] G. C. Lee, F. H. Yeh, and Y. H. Hsiao, "Kinect-based Taiwanese sign-language recognition system," *Multimedia Tools and Applications*, vol. 75, no. 1, pp. 261–279, 2016, doi: 10.1007/s11042-014-2290-x.

[9] D. Guo, W. Zhou, H. Li, and M. Wang, "Hierarchical LSTM for sign language translation," *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 6845–6852, 2018, doi: 10.1609/aaai.v32i1.12235.

[10] S. Wang, D. Guo, W. G. Zhou, Z. J. Zha, and M. Wang, "Connectionist temporal fusion for sign language translation," *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, pp. 1483–1491, 2018, doi: 10.1145/3240508.3240671.

[11] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 1, pp. 131–153, 2019, doi: 10.1007/s13042-017-0705-5.

[12] M. Oszust and J. Krupski, "Isolated sign language recognition with depth cameras," *Procedia Computer Science*, vol. 192, pp. 2085–2094, 2021, doi: 10.1016/j.procs.2021.08.216.

[13] B. S. Parton, "Sign language recognition and translation: A multidisciplined approach from the field of artificial intelligence," *Journal of Deaf Studies and Deaf Education*, vol. 11, no. 1, pp. 94–101, 2006, doi: 10.1093/deafed/enj003.

[14] N. Adaloglou *et al.*, "A comprehensive study on deep learning-based methods for sign language recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 1750–1762, 2022, doi: 10.1109/TMM.2021.3070438.

[15] D. Bragg *et al.*, "Sign language recognition, generation, and translation: An interdisciplinary perspective," *ASSETS 2019 - 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 16–31, 2019, doi: 10.1145/3308561.3353774.

[16] R. C. Dalawis, K. D. R. Olayao, E. G. I. Ramos, and M. J. C. Samonte, "Kinect-based sign language recognition of static and dynamic hand movements," *Eighth International Conference on Graphic and Image Processing (ICGIP 2016)*, vol. 10225, p. 102250I, 2017, doi: 10.1117/12.2266729.

[17] C. K. M. Lee, K. K. H. Ng, C. H. Chen, H. C. W. Lau, S. Y. Chung, and T. Tsoi, "American sign language recognition and training method with recurrent neural network," *Expert Systems with Applications*, vol. 167, 2021, doi: 10.1016/j.eswa.2020.114403.

[18] Manisha U. Kakde, Mahender G. Nakrani, and Amit M. Rawate, "A review paper on sign language recognition system for deaf and dumb people using image processing," *International Journal of Engineering Research and*, vol. V5, no. 03, 2016, doi: 10.17577/ijertv5is031036.

[19] K. Amrutha and P. Prabu, "ML based sign language recognition system," *2021 International Conference on Innovative Trends in Information Technology, ICITIIT 2021*, 2021, doi: 10.1109/ICITIIT51526.2021.9399594.

[20] R. Elakkiya, "Machine learning based sign language recognition: A review and its research frontier," *Journal of Ambient Intelligence and Humanized Computing*, 2020, doi: 10.1007/s12652-020-02396-y.

[21] M. Al-Qurishi, T. Khalid, and R. Souissi, "Deep learning for sign language recognition: Current techniques, benchmarks, and open issues," *IEEE Access*, vol. 9, pp. 126917–126951, 2021, doi: 10.1109/ACCESS.2021.3110912.

[22] I. Papastratis, C. Chatzikonstantinou, D. Konstantinidis, K. Dimitropoulos, and P. Daras, "Artificial intelligence technologies for sign language," *Sensors*, vol. 21, no. 17, 2021, doi: 10.3390/s21175843.

[23] U. Zeshan, "Aspects of Pakistan sign language," *Sign Language Studies*, vol. 1092, no. 1, pp. 253–296, 1996, doi: 10.1353/sls.1996.0015.

[24] N. S. Khan, A. Abid, K. Abid, U. Farooq, M. S. Farooq, and H. Jameel, "Speak Pakistan: Challenges in developing Pakistan sign language using information technology," *South Asian Studies: A research journal of South Asian Studies*, vol. 30, no. 2, pp. 367–379, 2015, [Online]. Available: http://journals.pu.edu.pk/journals/index.php/IJSAS/article/view/3027.

[25] F. Shah, M. S. Shah, W. Akram, A. Manzoor, R. O. Mahmoud, and D. S. Abdelminaam, "Sign language recognition using multiple Kernel learning: A case study of Pakistan sign language," *IEEE Access*, vol. 9, pp. 67548–67558, 2021, doi: 10.1109/ACCESS.2021.3077386.

BIOGRAPHIES OF AUTHORS

Saad Ahmed received MS. degree in Computer Science from Hamdard University, Karachi Pakistan, in 2012 and a Ph.D. degree in Computer Science from the NED University of Engineering and Technology Karachi Pakistan 2019. He currently works as an assistant professor at the Department of Computer Science, IQRA University Karachi Pakistan. His current research interests include Natural Language Processing, Data mining, and Big Data Analysis and its applications in interdisciplinary domains. He can be contacted at email: saadahmed@iqra.edu.pk.

Hasnain Shafiq is a BS holder in Computer Science from IQRA University. He is currently working in the software industry and pursuing different certifications from platforms such as Datacamp, Coursera in the domain of Data Science. He can be contacted at email: hasnainshafeeq@gmail.com.

Yamna Raheel is a BS degree holder in Computer Science from IQRA University. She is pursuing teaching as a profession and currently enrolled in different courses related to technology. She can be contacted at email: yamnaraheel02@gmail.com.

Noor Chishti is a BS degree holder in Computer Science from IQRA University. She is currently pursuing certification in the domain of cloud computing from AWS. She can be contacted at email: noorrchishti@gmail.com.

Syed Muhammad Asad is a BS degree holder in Computer Science from IQRA University. He is currently working in software industry as an Apex Developer. He can be contacted at email: asadsyed924@gmail.com.