Dynamic Sign Language Recognition Based on Real-Time Videos[1]

https://doi.org/10.3991/ijoe.v18i01.27581

Bushra A. Al-Mohimeed(✉), Hessa O. Al-Harbi, Ghadah S. Al-Dubayan,
Amal A. Al-Shargabi
College of Computer, Qassim University, Buraydah, Saudi Arabia
371204585@qu.edu.sa

Abstract—Sign language is the main communication tool for the deaf and hard of hearing. Deaf people cannot interact with others without a sign language interpreter. Accordingly, sign language recognition automation has become an important application in artificial intelligence and deep learning. Specifically, the recognition of Arabic sign language has been studied using many smart and traditional methods. This research provides a system to recognize dynamic Saudi sign language based on real time videos to solve this problem. We constructed a dataset for Saudi sign language in terms of videos in the proposed system. The dataset was then used to train a deep learning model using convolutional long short-term memory (convLSTM) to recognize the dynamic signs. Implementing such a system provides a platform for deaf people to interact with the rest of the world without an interpreter to reduce deaf isolation in society.

Keywords—dynamic sign language, recognition, deep learning, convLSTM

1 Introduction

Sign language is the way of communication for deaf and hard of hearing people. There is no problem if two deaf persons use sign language to communicate with each other, but the main problem when a non-deaf person communicates with a deaf person. In 2019 the number of people with hearing disabilities was around 466 million, and this number is expected to double in the next 30 years, according to the world health organization [1]. For these reasons, the sign language field focused on research and many attempts to create smart devices that can use it as an interpreter for sign language. These devices are classified as human-computer interaction (HCI) systems. There are two types of HCI devices for hand gesture recognition: sensor-based systems and vision-based systems. The sensor-based systems depend on devices or wearable tools such as gloves, Leap Motion Sensors, and Microsoft Kinect that can be used to collect data and images about gestures (signs) [2], [3], [4]. Sensor-based systems use cables and sensors that restrict user movement. Vision-based systems have

[1] The authors hereby confirm that they have obtained the consent of the persons depicted in the photographs for publication.

overcome this problem. Vision-based systems use video cameras and artificial intelligence to interpret and recognize hand gestures. These systems are more comfortable and flexible for deaf people because they do not have any devices or sensors. After capturing images, the images can be static and dynamic. In this paper, we proposed a Sign Language Recognition model based on Saudi sign language. It is a vision-based system in which a camera is used to record dynamic deaf signs and translates it into text. The remaining of the paper is organized as follows: Section 2 introduces the related works about the datasets and the recognition methods used in previous studies. Section 3 describes the Saudi Sign recognition model and the dataset construction. In Section 4, the details of the results are presented. Finally, Section 5 conclude the paper.

2 Related work

With the increasing need for sign language [1], recent studies have been accelerated to facilitate communication with disabled people. There are many studies have been conducting in this field to cover many aspects such as various languages and the diversity of recognition methods.

In this section, related works are presented in terms of the datasets and the recognition methods used in previous studies.

2.1 Datasets

Table 1 shows a summary of the datasets used in previous studies. Mostly used letters, 0-9 numbers, and a limited number of words in different languages.

Table 1. A summary of datasets used in previous studies

Description	Size	Language	Ref
Letters (32)	25,600 image	Arabic	[5]
Numbers (0-9)	20000 image	Bhutanese	[6]
Letters (26)	61614 image	American	[7]
Letters (23), numbers (0–10), words (67)	35,000 image	English	[8]
Word (15)	13,500 image	American	[9]
Word(20)	6600 video	Italian	[2]
Letters (30) and numbers (1-5)	1400 video	German	[10]
Letters (H – J) and words (8)	300 video	Brazilian	[11]
Words (25)	200 video	Arabic	[12]
Word(40)	8000 video	Arabic	[13]
Words (30)	1500 video	Argentinian	[14]
Words (10)	1080 video	Indian	[15]
Letters (26)	34627 image	American	[16]
Words (5)	500 video	Thai	[17]

2.2 Recognition methods

CNN is one of the most efficient deep learning tools for classifying and recognizing images. In the study of Saleh and Issa [5], the authors proposed to use fine-tuning with CNN architecture to recognize images and chose the ResNet152 model for high-performance capability. Fine-tuning will provide the ability to use a smaller size dataset in networks. In the study of Wangchuk et al. [6], the authors proposed to use CNN to extract features from images and classify digits with the trained model in real-time using the webcam. Also, Kadhim and Khamees [7] built A Real-Time American Sign Language Recognition System by a multi-classification system. A multi classification system was built based on VGG_Net, a profound CNN architecture for large-scale image recognition.

The work of Wadhawan and Kumar [8] used CNN architecture that is composed of convolutional layers with ReLU and max-pooling layers. Each convolutional layer is composed of different filtering window sizes to increase the speed and accuracy of recognition. The architecture tested on 50 deep learning models using different optimizers such as Adam, SGD, and RMSProp. In the study of Rahim et al. [9], the authors proposed a CNN, which has two channels for the input: one for gesture images and another for segmented images. Then the fully connected layer integrates all the features and provides the softmax classifier. Feature fusion of CNN extracts efficient features that improve the classification accuracy. It was live video frames using input from a webcam. In the work of Kumar et al. [16], used CNN architecture with Batch Normalization, and data augmentation to recognize the dataset. Also proposed a solution to overcome the issue of overfitting in the dataset.

The study of Pigou et al. [2], the authors proposed an architecture that uses two CNNs, one for extracting the hand and another for upper body features. Each CNN consists of three deep layers and then uses ANN with one hidden layer to classify both outcomes CNNs. Also Imran and Raman [10] proposed an approach that used three-stream CNN. In the first step, the input video was used to generate multiple types of motion templates. In the second step used pre-trained CNNs to extract features. In three-step, use fusion for classification. The study of maral et al. [11], the authors compared between 3DCNN and Long-term Recurrent Convolutional Network (LRCN). LRCN use 2DCNN to extract spatial features, with LSTM used to learn long-term temporal dependencies.

The work of Elbadawy et al. [12] used 3DCNN to extract spatial-temporal features and then in the last layer of 3DCNN, use Softmax to classify these features. The study of Al-Hammadi et al. [13] used 3DCNN architecture for spatiotemporal feature learning using two approaches. In the first approach, a single 3DCNN structure was used to extract the video sample features and then used the SoftMax layer for classification. In the second approach, parallel 3DCNN structures were trained to extract the features from different video sample regions and then use different feature fusion techniques. The approaches were evaluated by using different datasets. The datasets were taken by different devices such as RGB cameras and Microsoft Kinect.

The work of Siriak et al. [14] used CNN with LSTM for recognizing hand gestures from a video stream in real-time. CNN trained with Categorical Cross-Entropy (CLE) loss function is applied to address the gesture recognition problem. Also, use the convolutional layers to the identification of spatial features and use LSTM to the identification of temporal features. In the study of Chaikaew et al. [17], the authors used MediaPipe that is framework for building a pipeline to extract hand key point from video. Then they built hand gestures recognition model with various RNN algorithms such as LSTM, BLSTM and GRU.

Also, Bhagat et al. [15] used a convolutional kernel with LSTMs for training the videos over a 3DCNN architecture. They proposed two Convolutional LSTM based architectures, one to train the depth and RGB videos separately and another to train them concurrently using dual-channel architecture. In the dual-channel Convolutional LSTM architecture there are two convLSTM layers, and the outcome from these layers are passed through a fully connected layer and ReLU activation. In the last SoftMax layer classified the videos.

3 The Saudi sign recognition model

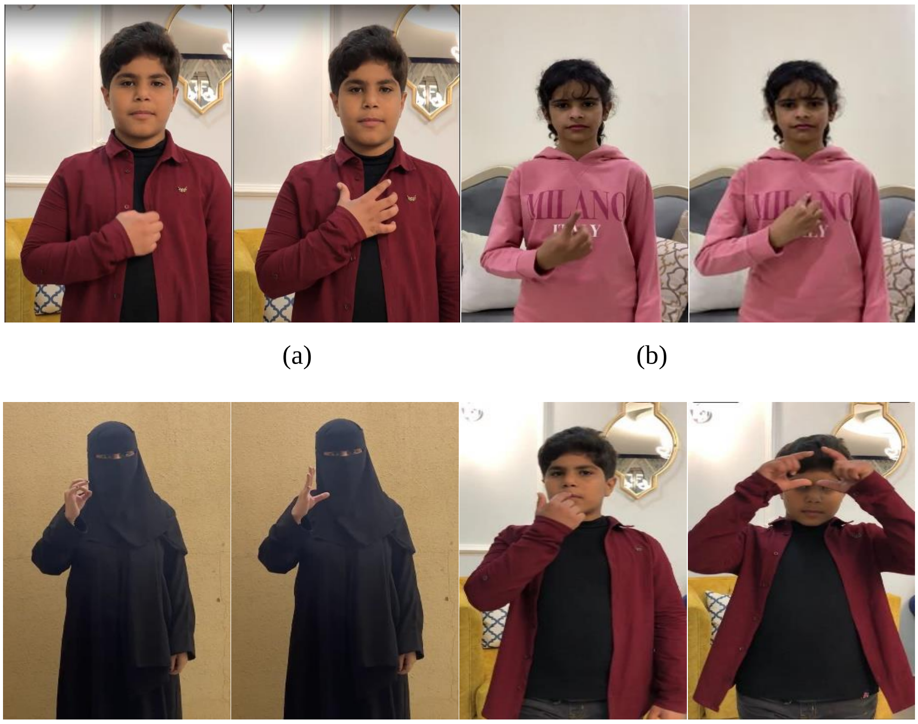
3.1 Dataset construction

The dataset is based on the Saudi dictionary, which was published by the Saudi Association for Hearing Disability in 2018. This study focused on the health and disease field due to it being one of the most important fields in which the deaf need to communicate. So our dataset comprises selected thirty-five dynamic gestures from the common words and health queries. Table 2 listed the selected gestures. These gestures contain single-handed actions and two-handed actions. Our dataset was taken based on the Saudi dictionary using a smartphone camera and with the help of some deaf people and non-deaf people as shown in Figure 1, some dynamic gestures. The videos were captured with different lights, backgrounds, places, and ages of people without any restrictions. This makes the proposed convLSTM model more general as it is trained on series images at different conditions. In total, we took around 98 videos for each class, resulting in 3454 videos.

Table 2. The selected Saudi signs

No.	Arabic gesture	English meaning	No.	Arabic gesture	English meaning
1	أنا	I am	19	منذ	Since
2	الشعور	Feeling	20	يوم	Day
3	تعبان	Tired	21	حساسية	Allergy
4	مستشفى	Hospital	22	كيف حالك؟	How are you
5	صيدلية	Pharmacy	23	كم؟	How
6	طبيب	Physician	24	أين؟	Where
7	دواء	Medicine	25	هو	He
8	صداع	Headache	26	هي	She
9	تحليل دم	Blood analysis	27	لماذا؟	Why

10	وصفة طبية	Prescription	28	أعد ما قلت	Repeat what you said
11	ألم	Pain	29	مصعد	Elevator
12	حمى	Fever	30	السلام عليكم	Hi
13	حمل	Pregnancy	31	التهاب	Inflammation
14	دوار – دوخة	Dizziness	32	إسعاف	Ambulance
15	مختبر	Laboratory	33	ممرضة – ممرض	Nurse
16	فقر دم	Anemia	34	شكراً	Thank you
17	سن – أسنان	Tooth	35	جرعة دواء	potion Medicine
18	الفحص الطبي قبل الزواج	Medical examination before marriage			



(a) (b)

(c) (d)

Fig. 1. Samples of the created dataset (a) Feel: الشعور, (b) Me : أنا, (c) Headache :صداع, (d) Pharmacy : صيدلية

3.2 The proposed model

ConvLSTM have proven to be successful in sequential images recognition [15]. The ConvLSTM is a gathering Convolution and LSTM. convLSTM is an extension of LSTM RNN. In this, the fully-connected gates of the LSTM are completely replaced by the convolutional gates. convLSTM replaces the matrix multiplication with the convolution operation at each gate in the LSTM cell. This makes it able to encode spatial and temporal features. Figure 2 shows the convLSTM architecture that is used to build the recognition model.

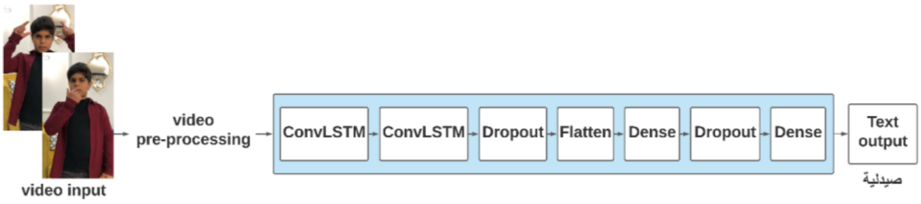


Fig. 2. The convLSTM model

In the model, there are two convolutional LSTM layers with kernel sizes of (3,3) with number of filters being 64, dropout layer, flatten layer, dense layer, dropout layer and the final dense layer use SoftMax activation function to classifies the videos into one of the classes.

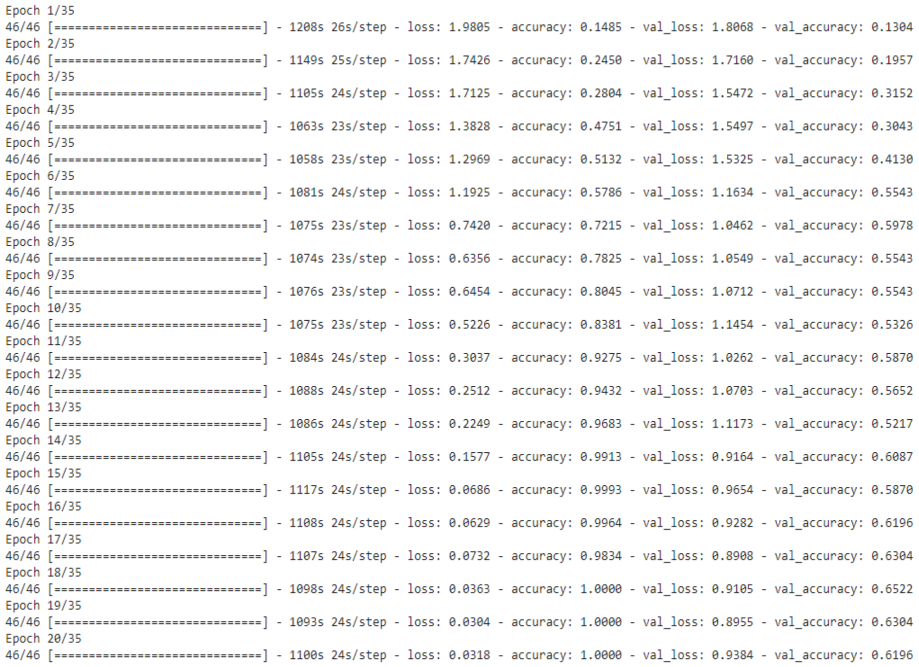
4 Results

The model was implemented and tested for signs of some health and general words as we mentioned in Table 2. We chose six classes from the 35 classes mentioned in the Table. The dataset reached 585 videos divided into training data and test data, where the training was 468 videos, and the test was 117 videos. Through experience, we tested our model by adding and removing layers of different types, also tested by increasing and decreasing the value of batch size, epoch, learning rate. The highest accuracy was when the training values were 468 samples, the test was 117 samples, epochs are 35 with early stop, the batch size is 8, and learning rate = 0.001 (See Table 3).

Table 3. Configurations of model training and testing

Training	Testing	Epoch	Batch size
468 of video (80%)	117 of videos (20%)	35	8

To get an idea about the training process, Figure 3 shows the training status.



```
Epoch 1/35
46/46 [==============================] - 1208s 26s/step - loss: 1.9805 - accuracy: 0.1485 - val_loss: 1.8068 - val_accuracy: 0.1304
Epoch 2/35
46/46 [==============================] - 1149s 25s/step - loss: 1.7426 - accuracy: 0.2450 - val_loss: 1.7160 - val_accuracy: 0.1957
Epoch 3/35
46/46 [==============================] - 1105s 24s/step - loss: 1.7125 - accuracy: 0.2804 - val_loss: 1.5472 - val_accuracy: 0.3152
Epoch 4/35
46/46 [==============================] - 1063s 23s/step - loss: 1.3828 - accuracy: 0.4751 - val_loss: 1.5497 - val_accuracy: 0.3043
Epoch 5/35
46/46 [==============================] - 1058s 23s/step - loss: 1.2969 - accuracy: 0.5132 - val_loss: 1.5325 - val_accuracy: 0.4130
Epoch 6/35
46/46 [==============================] - 1081s 24s/step - loss: 1.1925 - accuracy: 0.5786 - val_loss: 1.1634 - val_accuracy: 0.5543
Epoch 7/35
46/46 [==============================] - 1075s 23s/step - loss: 0.7420 - accuracy: 0.7215 - val_loss: 1.0462 - val_accuracy: 0.5978
Epoch 8/35
46/46 [==============================] - 1074s 23s/step - loss: 0.6356 - accuracy: 0.7825 - val_loss: 1.0549 - val_accuracy: 0.5543
Epoch 9/35
46/46 [==============================] - 1076s 23s/step - loss: 0.6454 - accuracy: 0.8045 - val_loss: 1.0712 - val_accuracy: 0.5543
Epoch 10/35
46/46 [==============================] - 1075s 23s/step - loss: 0.5226 - accuracy: 0.8381 - val_loss: 1.1454 - val_accuracy: 0.5326
Epoch 11/35
46/46 [==============================] - 1084s 24s/step - loss: 0.3037 - accuracy: 0.9275 - val_loss: 1.0262 - val_accuracy: 0.5870
Epoch 12/35
46/46 [==============================] - 1088s 24s/step - loss: 0.2512 - accuracy: 0.9432 - val_loss: 1.0703 - val_accuracy: 0.5652
Epoch 13/35
46/46 [==============================] - 1086s 24s/step - loss: 0.2249 - accuracy: 0.9683 - val_loss: 1.1173 - val_accuracy: 0.5217
Epoch 14/35
46/46 [==============================] - 1105s 24s/step - loss: 0.1577 - accuracy: 0.9913 - val_loss: 0.9164 - val_accuracy: 0.6087
Epoch 15/35
46/46 [==============================] - 1117s 24s/step - loss: 0.0686 - accuracy: 0.9993 - val_loss: 0.9654 - val_accuracy: 0.5870
Epoch 16/35
46/46 [==============================] - 1108s 24s/step - loss: 0.0629 - accuracy: 0.9964 - val_loss: 0.9282 - val_accuracy: 0.6196
Epoch 17/35
46/46 [==============================] - 1107s 24s/step - loss: 0.0732 - accuracy: 0.9834 - val_loss: 0.8908 - val_accuracy: 0.6304
Epoch 18/35
46/46 [==============================] - 1098s 24s/step - loss: 0.0363 - accuracy: 1.0000 - val_loss: 0.9105 - val_accuracy: 0.6522
Epoch 19/35
46/46 [==============================] - 1093s 24s/step - loss: 0.0304 - accuracy: 1.0000 - val_loss: 0.8955 - val_accuracy: 0.6304
Epoch 20/35
46/46 [==============================] - 1100s 24s/step - loss: 0.0318 - accuracy: 1.0000 - val_loss: 0.9384 - val_accuracy: 0.6196
```

Fig. 3. Result of training

Due to the computer device limitations used in implementation, the model was trained on six classes only. Figure 4 shows the confusion matrix for the six classes which are headache, feeling, I, where, pharmacy and pregnancy. Also, Table 4 shows the model performance in terms of precision, recall, F1-score, and support. As shown in the figure, the model achieved 70% of accuracy which is a comparable performance to previous models that classify videos.

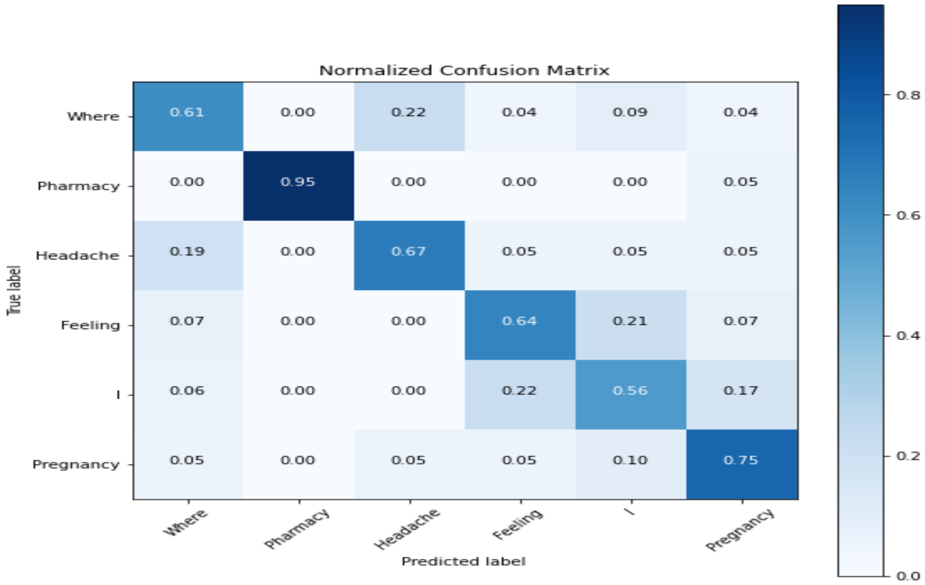
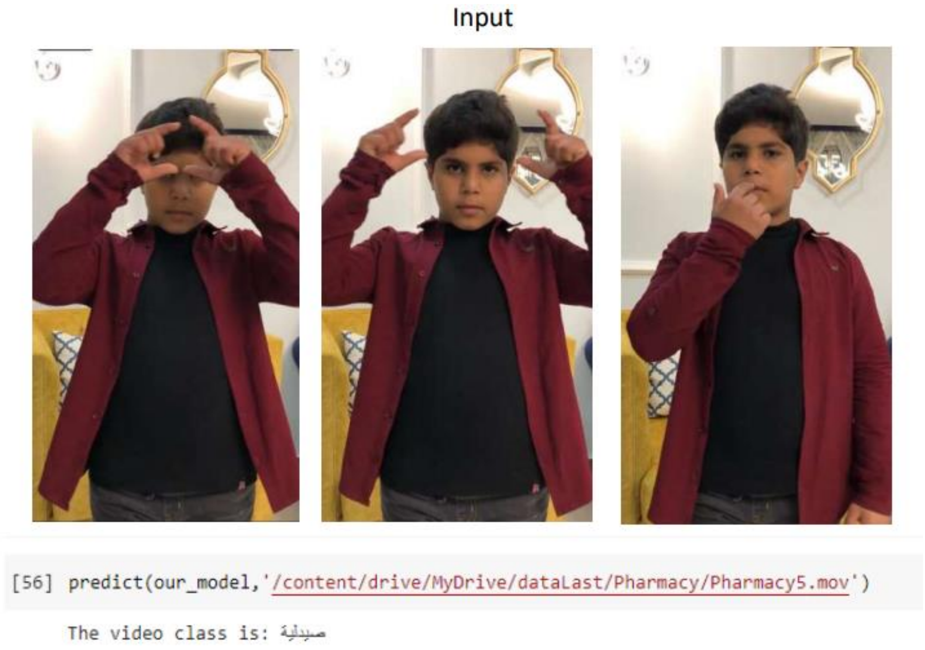


Fig. 4. The Confusion matrix of model

Table 4. Model performance

	Precision	Recall	F1-score	Support
0	0.67	0.61	0.64	23
1	1.00	0.95	9.97	19
2	0.70	0.67	0.68	21
3	0.56	0.64	0.60	14
4	0.56	0.56	0.56	18
5	0.68	0.75	0.71	20
Accuracy			0.70	115
Macro avg	0.69	0.70	0.69	115
Weighted avg	0.70	0.70	0.70	115

Figure 5 shows a sample sign and the results of prediction which is pharmacy.



```
[56] predict(our_model,'/content/drive/MyDrive/dataLast/Pharmacy/Pharmacy5.mov')

     The video class is: صيدلية
```

Fig. 5. A sample of a predicted sign

To evaluate the proposed model, it was compared with the previous studies depending on the recognition method, the image type, and the method of image capturing, as shown in Table 5.

Table 5. Proposed system results compared to previous studies

Dataset	Recognition method	Reference number
Words (40)	3DCNN	[13]
Words (10)	ConvLSTM	[15]
Words (25)	3DCNN	[12]
Words (20)	2DCNN	[2]
Words (30)	LSTM	[14]

Based on comparison with Table 5, our system consists of dataset videos (35), image capturing (vision-based) and type of image (dynamic). With this number of gestures, the model outperforms previous models that are based on convLSTM method.

5 Conclusion

This paper has proposed a system for identifying dynamic signs of the Saudi sign language, based on the Saudi sign language dictionary provided by the Saudi Association for Hearing Impairment. The system is not the first of its kind in the Arabic language based on the dictionary, but the first system that is trained on Saudi dynamic gestures and with a dataset of 35 classes using convLSTM method. In the future, we plan to extend our dataset to include all classes and cover all fields in the Saudi sign language dictionary.

6 References

[1] Organization. W. H. (2021) Deafness and hearing loss. Accessed: 12.10.2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] Pigou, L. Dieleman, S. Kindermans, P.-J. and Schrauwen, B. Sign language recognition using convolutional neural networks. European Conference on Computer Vision. Springer, 2014, pp. 572–578. https://doi.org/10.1007/978-3-319-16178-5_40

[3] Rahagiyanto, A. Basuki, A. Sigit, R. Anwar, A. and Zikky, M. Hand gesture classification for sign language using artificial neural network. 21st International Computer Science and Engineering Conference (ICSEC). IEEE, 2017, pp. 1–5. https://doi.org/10.1109/icsec.2017.8443898

[4] Jani, A. B., Kotak, N. A., & Roy, A. K. (2018, October). Sensor based hand gesture recognition system for English alphabets used in sign language of deaf-mute people. In 2018 IEEE SENSORS IEEE, pp. 1-4. https://doi.org/10.1109/icsens.2018.8589574

[5] Saleh, Y. and Issa, G. (2020). Arabic sign language recognition through deep neural networks fine-tuning. International journal of online and biomedical engineering, 16(5): 71–83, 2020. https://doi.org/10.3991/ijoe.v16i05.13087

[6] Wangchuk, K., Riyamongkol, P., and Waranusast, R. (2021). Real-time bhutanese sign language digits recognition system using convolutional neural network. ICT Express, 7(2): 215-220. https://doi.org/10.1016/j.icte.2020.08.002

[7] Kadhim, R. A., and Khamees, M. (2020). A Real-Time American Sign Language Recognition System using Convolutional Neural Network for Real Datasets. TEM Journal, 9(3): 937. https://doi.org/10.18421/tem93-14

[8] Wadhawan, A., and Kumar, P. (2020). Deep learning-based sign language recognition system for static signs. Neural Computing and Applications, 32(12): 7957-7968. https://doi.org/10.1007/s00521-019-04691-y

[9] Rahim, M. A. Shin, J. and Islam, M. R. Dynamic hand gesture-based sign word recognition using convolutional neural network with feature fusion. 2nd International Conference on Knowledge Innovation and Invention (ICKII). IEEE, 2019, pp. 221–224. https://doi.org/10.1109/ickii46306.2019.9042600

[10] Imran, J., & Raman, B. (2020). Deep motion templates and extreme learning machine for sign language recognition. The Visual Computer, 36(6): 1233-1246. https://doi.org/10.1007/s00371-019-01725-3

[11] Amaral, L. Júnior, G. L. Vieira, T. and Vieira, T. Evaluating deep models for dynamic brazilian sign language recognition. in Iberoamerican Congress on Pattern Recognition. Springer, 2018, pp. 930–937. https://doi.org/10.1007/978-3-030-13469-3_107

[12] ElBadawy, M. Elons, A. S. Shedeed, H. A. and Tolba, M. F. Arabic sign language recognition with 3d convolutional neural networks. 8th International Conference on Intelligent Computing and Information Systems (ICICIS), 2017, pp. 66–71. https://doi.org/10.1109/intelcis.2017.8260028

[13] Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M. A., & Mekhtiche, M. A. (2020). Hand gesture recognition for sign language using 3DCNN. IEEE Access, 8: 79491-79509. https://doi.org/10.1109/access.2020.2990434

[14] Siriak, R. Skarga-Bandurova, I. and Boltov, Y. Deep convolutional network with long short-term memory layers for dynamic gesture recognition. 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), vol. 1. 2019, IEEE, pp. 158–162. https://doi.org/10.1109/idaacs.2019.8924381

[15] N. K. Bhagat, Y. Vishnusai, and G. Rathna, Indian sign language gesture recognition using image processing and deep learning. Digital Image Computing: Techniques and Applications (DICTA). 2019, IEEE, pp. 1–8. https://doi.org/10.1109/dicta47822.2019.8945850

[16] Kumar, M. Gupta, P. Jha, R. K. Bhatia, A. Jha, K. and Shah, B. K. Sign language alphabet recognition using convolution neural network. 5th International Conference on Intelligent Computing and Control Systems (ICICCS). 2021, IEEE, pp. 1859–1865. https://doi.org/10.1109/iciccs51141.2021.9432296

[17] Chaikaew, A. Somkuan, K. and Yuyen, T. Thai sign language recognition: an application of deep neural network, International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering. 2021, IEEE, pp. 128–131. https://doi.org/10.1109/ectidamtncon51128.2021.9425711

7 Authors

Bushra A. Al-Mohimeed is a graduate of the Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia.

Hessa O. Al-Harbi is a graduate of the Department of Information Technology, College of Computer, Qassim University , Buraydah, Saudi Arabia (Email: 371205468@qu.edu.sa).

Ghadah S. Al-Dubayan is a graduate of the Department of Information Technology, College of Computer, Qassim University , Buraydah, Saudi Arabia (Email: 371201914@qu.edu.sa).

Amal A. Al-Shargabi received the master's and Ph.D. degrees from Universiti Teknologi Mara (UiTM), Malaysia. She is currently an Assistant Professor with the College of Computer, Qassim University. Her research interests include program comprehension, empirical software engineering, and machine learning. She was a recipient of a number of Qassim University's research grants, since 2018 (Email: a.alshargabi@qu.edu.sa).

Article submitted 2021-10-15. Resubmitted 2021-12-03. Final acceptance 2021-12-07. Final version published as submitted by the authors.