Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 3685614, 9 pages
https://doi.org/10.1155/2020/3685614

Research Article

Arabic Sign Language Recognition and Generating Arabic Speech Using Convolutional Neural Network

M. M. Kamruzzaman

Department of Computer and Information Science, Jouf University, Sakaka, Al-Jouf, Saudi Arabia

Correspondence should be addressed to M. M. Kamruzzaman; mmkamruzzaman@ju.edu.sa

Received 28 January 2020; Revised 15 February 2020; Accepted 25 February 2020; Published 23 May 2020

Guest Editor: Yin Zhang

Sign language encompasses the movement of the arms and hands as a means of communication for people with hearing disabilities. An automated sign recognition system requires two main courses of action: the detection of particular features and the categorization of particular input data. In the past, many approaches for classifying and detecting sign languages have been put forward for improving system performance. However, the recent progress in the computer vision field has geared us towards the further exploration of hand signs/gestures' recognition with the aid of deep neural networks. The Arabic sign language has witnessed unprecedented research activities to recognize hand signs and gestures using the deep learning model. A vision-based system by applying CNN for the recognition of Arabic hand sign-based letters and translating them into Arabic speech is proposed in this paper. The proposed system will automatically detect hand sign letters and speaks out the result with the Arabic language with a deep learning model. This system gives 90% accuracy to recognize the Arabic hand sign-based letters which assures it as a highly dependable system. The accuracy can be further improved by using more advanced hand gestures recognizing devices such as Leap Motion or Xbox Kinect. After recognizing the Arabic hand sign-based letters, the outcome will be fed to the text into the speech engine which produces the audio of the Arabic language as an output.

1. Introduction

Language is perceived as a system that comprises of formal signs, symbols, sounds, or gestures that are used for daily communication. Communication can be broadly categorized into four forms; verbal, nonverbal, visual, and written communication. Verbal communication means transferring information either by speaking or through sign language. However, nonverbal communication is the opposite of this, as it involves the usage of language in transferring information using body language, facial expressions, and gestures. Written communication, however, involves conveying information through writing, printing, or typing symbols such as numbers and letters, while visual communication entails conveying information through means such as art, photographs, drawings, charts, sketches, and graphs.

The movement of the arms and hands to communicate, especially with people hearing disability, is referred to as sign language. However, this differs according to people and the region they come from. Therefore, there is no standardization concerning the sign language to follow; for instance, the American, British, Chinese, and Saudi have different sign languages. Since the sign language has become a potential communicating language for the people who are deaf and mute, it is possible to develop an automated system for them to communicate with people who are not deaf and mute.

Sign language is made up of four major manual components that comprise of hands' figure configuration, hands' movement, hands' orientation, and hands' location in relation to the body [1]. There are mainly two procedures that an automated sign-recognition system has, vis-a-vis detecting the features and classifying input data. Many approaches have been put forward for the classification and detection of sign languages for the improvement of the performance of the automated sign language system. The American Sign Language (ASL) is regarded as the sign language that is widely used in many countries such as the USA, Canada, some parts of Mexico, with little modification it is also used

in few other countries in Asia, Africa, and Central America. The research activities on sign languages have also been extensively conducted on English, Asian, and Latin sign languages, while little attention is paid on the Arabic language. This may be because of the nonavailability of a generally accepted database for the Arabic sign language to researchers. So, researchers had to resort to develop datasets themselves which is a tedious task. Specially, there is no Arabic sign language reorganization system that uses comparatively new techniques such as Cognitive Computing, Convolutional Neural Network (CNN), IoT, and Cyberphysical system that are extensively used in many automated systems [2–7]. The cognitive process enables systems to think the same way a human brain thinks without any human operational assistance. The human brain inspires the cognitive ability [8–10]. On the other hand, deep learning is a subset of machine learning in artificial intelligence (AI) that has networks capable of learning unsupervised from data that is unstructured or unlabeled which is also known as deep neural learning or deep neural network [11–15]. In deep learning, CNN is a class of deep neural networks, most commonly applied in the field of computer vision. The vision-based approaches mainly focus on the captured image of gesture and get the primary feature to identify it. This method has been applied in many tasks including super resolution, image classification and semantic segmentation, multimedia systems, and emotion recognition [16–20]. One of the few well-known researchers who have applied CNN is K. Oyedotun and Khashman [21] who used CNN along with Stacked Denoising Autoencoder (SDAE) for recognizing 24 hand gestures of the American Sign Language (ASL) gotten through a public database. On the other hand, the proposal to use Convolutional Neural Network (CNN) for recognizing the Italian sign language was made by Pigou et al. [22]. Whereas Hu et al. had made a proposal for the architecture of hybrid CNN and RNN to capture the temporal properties perfectly for the electromyogram signal which solves the problem of gesture recognition [23]. An incredible CNN model that automatically recognizes the digits based on hand signs and speaks the particular result in Bangla language is explained in [24], which is followed in this work. In [25] as well, there is a proposal of using transfer learning on data collected from several users, while exploiting the use of deep-learning algorithm to learn discriminant characteristics found from large datasets.

There are several other techniques, which are used to recognize the Arabic Sign Language such as a continuous recognition system using the K-nearest neighbor classifier and statistical feature extraction method for the Arabic sign language was proposed by Tubaiz et al. [26]. Unfortunately, the main drawback of the Tubaiz's approach is that the users are required to use an instrumented hand gloves to obtain the particular gesture's information that often causes immense distress to the user. Following this, [27] also proposes an instrumented glove for the development of the Arabic sign language recognition system. The continuous recognition of the Arabic sign language, using the hidden Markov models and spatiotemporal features, was proposed by [28]. Research on translation from the Arabic sign language to text was done by Halawani [29], which can be used on mobile devices. In [30], the automatic recognition using sensor and image approaches are presented for Arabic sign language. [31] also uses two depth sensors to recognize the hand gestures of the Arabic Sign Language (ArSL) words. [32] introduces a dynamic Arabic Sign Language recognition system using Microsoft Kinect which depends on two machine learning algorithms. However, Arabic sign language with this recent CNN approach has been unprecedented in the research domain of sign language. Therefore, this work aims at developing a vision-based system by applying CNN for the recognition of Arabic hand sign-based letters and translating them into Arabic speech. A dataset with 100 images in the training set and 25 images in the test set for each hand sign is also created for 31 letters of Arabic sign language. The suggested system is tested by combining hyperparameters differently to obtain the optimal outcomes with the least training time.

2. Data Preprocessing

Data preprocessing is the first step toward building a working deep learning model. It is used to transform the raw data in a useful and efficient format. Figure 1 shows the flow diagram of data preprocessing.

2.1. Raw Images. Hand sign images are called raw images that are captured using a camera for implementing the proposed system. The images are taken in the following environment:

(i) From different angles

(ii) By changing lighting conditions

(iii) With good quality and in focus

(iv) By changing object size and distance

The objective of creating raw images is to create the dataset for training and testing. Figure 2 shows 31 images for 31 letters of the Arabic Alphabet from the dataset of the proposed system.

2.2. Classifying Images. The proposed system classifies the images into 31 categories for 31 letters of the Arabic Alphabet. One subfolder is used for storing images of one category to implement the system. All subfolders which represent classes are kept together in one main folder named "dataset" in the proposed system.

2.3. Formatting Image. Usually, the hand sign images are unequal and having different background. So, it is required to delete the unnecessary element from the images for getting the hand part. The extracted images are resized to 128×128 pixels and converted to RGB. Figure 3 shows the formatted image of 31 letters of the Arabic Alphabet.

2.4. Dividing Dataset for Training and Testing. For each of the 31 alphabets, there are 125 pictures for each letter. The dataset is broken down into two sets, one for learning set and one for the testing set. A ratio of 80 : 20 is used for dividing the dataset into learning and testing set. There are 100

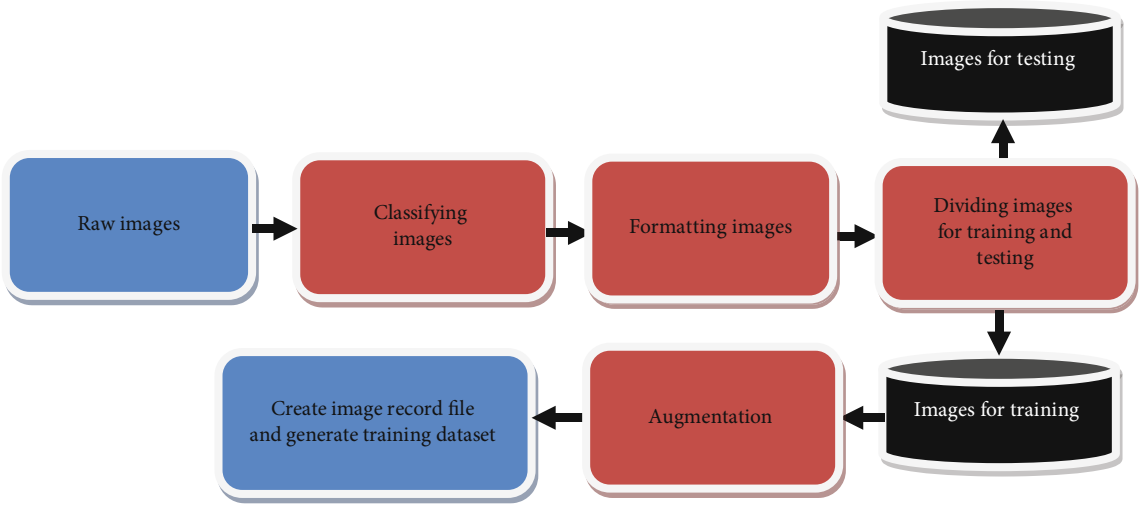

Figure 1: Flow diagram of data preprocessing.

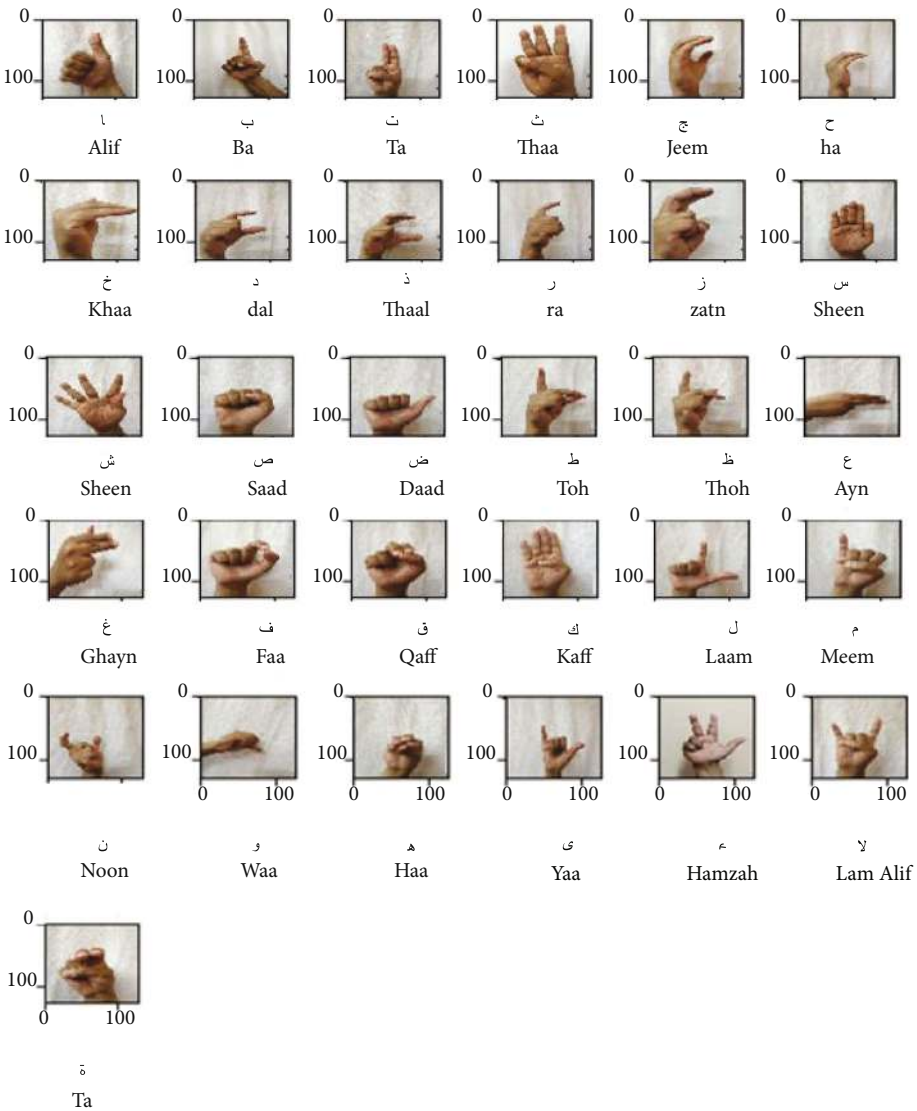

Figure 2: Raw images of 31 letters of the Arabic Alphabet for the proposed system.

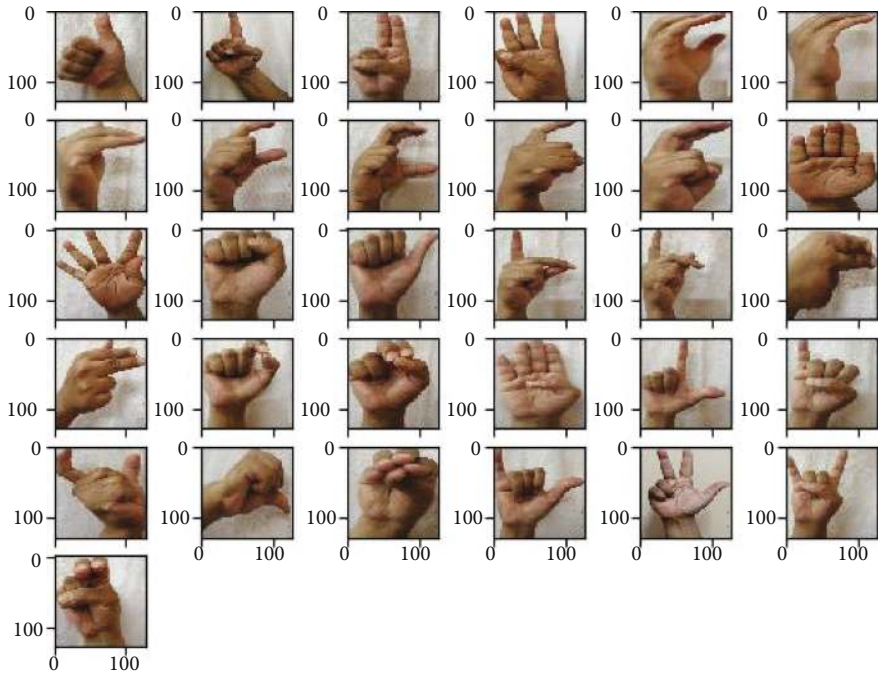

Figure 3: Formatted image of 31 letters of the Arabic Alphabet.

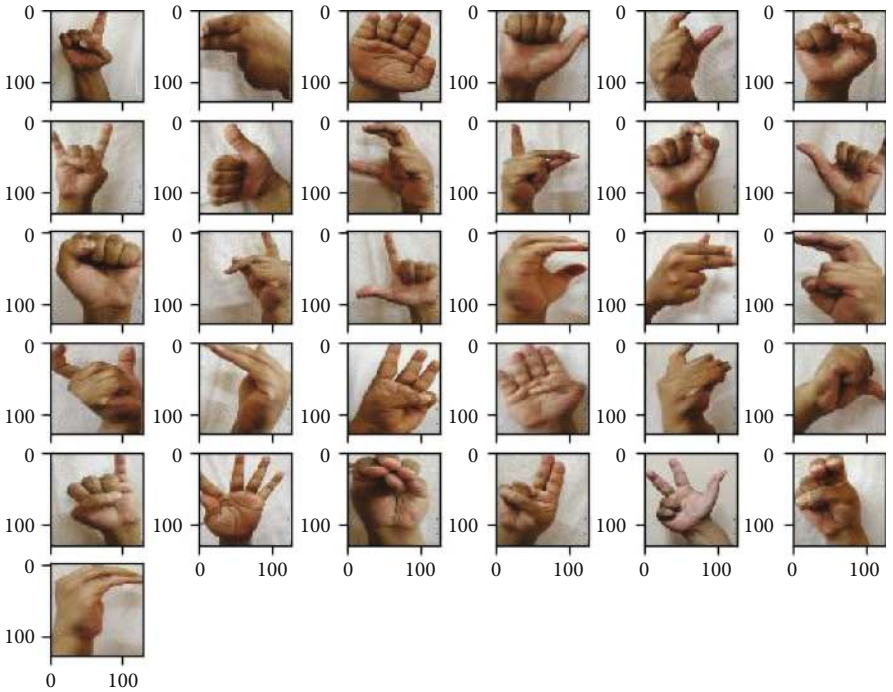

Figure 4: Snapshot of the augmented images of the proposed system.

images in the training set and 25 images in the test set for each hand sign.

2.5. Augmentation. Real-time data is always inconsistent and unpredictable due to a lot of transformations (rotating, moving, and so on). Image augmentation is used to improve deep network performance. It creates images artificially through various processing methods, such as shifts, flips, shear, and rotation. The images of the proposed system are rotated randomly from 0 to 360 degrees using this image augmentation technique. Few images were also sheared randomly with 0.2-degree range and few images were flipped horizontally. Figure 4 shows a snapshot of the augmented images of the proposed system.

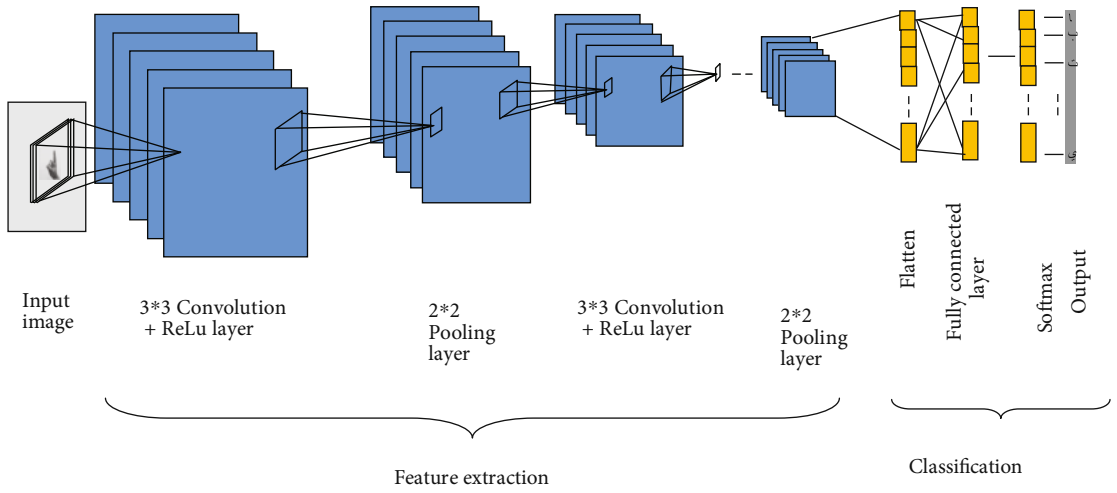

Figure 5: Architecture of Arabic Sign Language Recognition using CNN.

Table 1: Loss and Accuracy with and without Augmentation.

Batch size	Augmentation	Loss	Accuracy
64	False	0.84	83%
	True	0.57	85%
128	False	0.53	86%
	True	0.50	90%

2.6. Create Image Record File and Generate Training Dataset. It is required to create a list of all images which are kept in a different folder to get label and filename information.

3. Architecture

Figure 5 shows the architecture of the Arabic sign language recognition system using CNN. CNN is a system that utilizes perceptron, algorithms in machine learning (ML) in the execution of its functions for analyzing the data. This system falls in the category of artificial neural network (ANN). CNN is mostly applicable in the field of computer vision. It mainly helps in image classification and recognition.

The two components of CNN are feature extraction and classification. Each component has its characteristics that need to be explored. The following sections will explain these components.

3.1. Feature Extraction Part. CNN has various building blocks. However, the major building block of the CNN is the Convolution layer. Convolution layer refers to the mathematical combination of a pair of functions to yield a third function. It is required to do convolution on the input by using a filter or kernel for producing a feature map. The execution of a convolution involves sliding each filter over particular input. At each place, a matrix multiplication is conducted and adds the output onto a particular feature map.

Every image is converted as a 3D matrix by specified width, specified height, and specified depth. The depth is included as a dimension since image (RGB) contains color channels. Numerous convolutions can be performed on input data with different filters, which generate different feature maps. The different feature maps are combined to get the output of the convolution layer. The output is then going through the activation function to generate nonlinear output.

One of the most popular activation function is the Rectified Linear Unit (ReLU) which operates with the computing the function $f(\kappa) = \max(0,\kappa)$. The function shows that the activation is threshold at zero. The ReLU is more reliable and speeds up convergence six times compared to sigmoid and tanh, but it is much fragile during operations. This disadvantage can, however, be overcome by fixing the appropriate learning rate.

Stride refers to the size of a particular step that the convolution filter functions each time. The size of a stride usually considered as 1; it means that the convolution filter moves pixel by pixel. If we increase the size of the particular stride, the filter will slide over the input by a higher interval and therefore has a smaller overlap within the cells.

Because the feature map size is always lesser than the size of the input, we must do something to stop shrinking our feature map. Here, we are intended to use padding.

Now it is required to add zero-value pixels layer to gird particular input by zeros to prevent the feature map from shrinking. Padding also helps in maintaining the spatial dimension constant after doing convolution so that the kernel and stride size matches with the input. So it enhances the performance of the system.

There are three main parameters that need to be adjusted in a convolutional neural network to modify the behavior of a convolutional layer. These parameters are filter size, stride, and padding. It is possible to calculate the output size for any given convolution layer as:

$$\text{Output}_{\text{size}} = \frac{\text{input}_{\text{size}} - \text{filter}_{\text{size}} + 2 * \text{padding}_{\text{size}}}{\text{Stride}_{\text{size}}} + 1, \quad (1)$$

where $\text{output}_{\text{size}}$= the size of the output Convolution layer. $\text{input}_{\text{size}}$= the size of input image. $\text{filter}_{\text{size}}$= the size of filter.

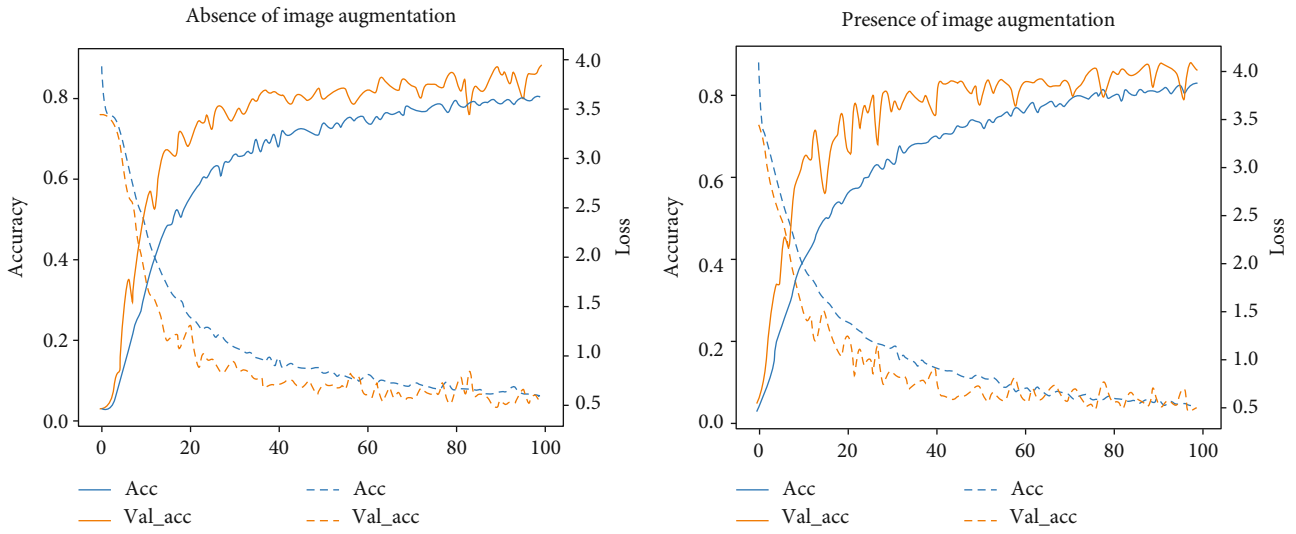

Figure 6: Loss and accuracy graph of training and validation in the absence and presence of image augmentation for batch size 128.

$\text{padding}_{\text{amount}}$ = the amount of padding. $\text{stride}_{\text{size}}$= the size of stride.

3.2. Pooling Layer. Naturally, a pooling layer is added in between Convolution layers. However, its main purpose is to constantly decrease the dimensionality and lessen computation with less number of parameters. It also regulates overfitting and reduces the training time. There are several forms of pooling; the most common type is called the max pooling. It uses the highest value in all windows and hence reduces the size of the feature map but keeps the vital information. It is required to specify the window sizes in advance to determine the size of the output volume of the pooling layer; the following formula can be applied.

$$\text{output}_{\text{size}} = \frac{\text{input}_{\text{size}} - \text{filter}_{\text{size}}}{\text{stride}_{\text{size}}} + 1. \quad (2)$$

In all situations, some translation invariance is provided by the pooling layer which indicates that a particular object would be identifiable without regard to where it becomes visible on the frame.

3.3. Classification. The second important component of CNN is classification. The classification consists of a few layers which are fully connected (FC). Neurons in an FC layer own comprehensive connections to each of the activations of the previous layer. The FC layer assists in mapping the representation between the particular input and output. The layer executes its functions by applying the same principles of a regular Neural Network. However, One Dimensional data can only be accepted by an FC layer. For transforming three Dimensional data to one Dimensional data, the flatten function of Python is used to implement the proposed system.

4. Experimental Result and Discussion

The proposed system is tested with 2 convolution layers. Then 2×2 maximum pooling layers follow each convolution layer. The convolution layers have a different structure in the first layer; there are 32 kernels while the second layer has 64 kernels; however, the size of the kernel in both layers is similar 3×3. Each pair of convolution and pooling layer was checked with two different dropout regularization values which were 25% and 50%, respectively. So, this setting allows eliminating one input in every four inputs (25%) and two inputs (50%) from each pair of convolution and pooling layer. The activation function of the fully connected layer uses ReLu and Softmax to decide whether the neuron fire or not. The experimental setting of the proposed model is given in Figure 5.

The system was trained for hundred epochs by RMSProp optimizer with a cost function based on Categorical Cross Entropy because it converged well before 100 epochs so the weights were stored with the system for using in the next phase.

The system presents optimistic test accuracy with minimal loss rates in the next phase (testing phase). The loss rate was further decreased after using augmented images keeping the accuracy almost the same. Each new image in the testing phase was processed before being used in this model. The size of the vector generated from the proposed system is 10, where 1/10 of these values are 1, and all other values are 0 to denote the predicted class value of the given data. Then, the system is linked with its signature step where a hand sign was converted to Arabic speech. This process was completed into two phases. The first phase is the translation from hand sign to Arabic letter with the help of translation API (Google Translator). The generated Arabic Texts will be converted into Arabic speech. In this stage, Google Text To Speech (GTTS) was used.

The system was constructed by different combinations of hyperparameters in order to achieve the best results. The

Table 2: Confusion Matrices with the presence of image augmentation—Ac: Actual Class and Pr: Predicted Class.

PR ⟶ ↓ AC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	2	1	0	12	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0
5	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	12	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	10	4	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	19	0	0	0	0	0	0	0	0	0	0	0	0
20	0	2	0	0	0	1	0	0	0	0	0	0	0	1	5	0	1	0	1	9	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	18	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	18	0	0	0	0	0	1	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0	0	0	10	0	0	1	0	4	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	19	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	18	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	4	0	15	0
31	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	16

results indicated 83 percent accuracy and only 0.84 validation loss for convolution layers of 32 and 64 kernels with 0.25 and 0.5 dropout rate. The system is also tested for convolution layers with batch size 64 and 128. Furthermore, in the presence of Image Augmentation (IA), the accuracy was increased 86 to 90 percent for batch size 128 while the validation loss was decreased 0.53 to 0.50. Table 1 represents these results. It was also found that further addition of the convolution layer was not suitable and hence avoided. Figure 6 presents the graph of loss and accuracy of training and validation in the absence and presence of image augmentation for batch size 128. It is indicated that prior to augmentation, the validation accuracy curve was below the training accuracy and the accuracy for training and loss of validation both are decreased after the implementation of augmentation. The graph is showing that our model is not overfitted or underfitted.

The confusion matrix (CM) presents the performance of the system in terms of correct and wrong classification developed. Therefore, CM of the test predictions in absence and presence of IA is shown in Table 2 and Table 3, respectively.

5. Conclusion

The main objective of this work was to propose a model for the people who have speech disorders to enhance their communication using Arabic sign language and to minimize the implications of signs languages. This model can also be used in hand gesture recognition for human-computer interaction effectively. However, the model is in initial stages but it is still efficient in the correct identification of the hand digits and transferred them into Arabic speech with higher 90% accuracy. In order to further increase the accuracy and quality of the model, more advanced hand gestures recognizing

Table 3: Confusion Matrices in absence of image augmentation—Ac: Actual Class and Pr: Predicted Class.

PR ⟶ ↓ AC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	17	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
4	0	0	0	13	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	19	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	17	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	1	0	1	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	6	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	18	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0	0	0	0	0	0	0	0	0	0	0	0	1	4	0	1	0	0	10	0	0	0	0	0	2	1	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	3	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	17	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	17	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	14	0	0	0	0	2	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	19	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0
31	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	3	0	0	0	0	0	2	0	0	0	12

devices can be considered such as Leap Motion or Xbox Kinect and also considering to increase the size of the dataset and publish in future work. The proposed system also produces the audio of the Arabic language as an output after recognizing the Arabic hand sign based letters. In spite of this, the proposed tool is found to be successful in addressing the very essential and undervalued social issues and presents an efficient solution for people with hearing disability.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Jouf University, Sakaka, Saudi Arabia, under Grant 40/140.

References

[1] E. Costello, *American Sign Language Dictionary*, Random House, New York, NY, USA, 2008.

[2] Y. Zhang, X. Ma, S. Wan, H. Abbas, and M. Guizani, "Cross-Rec: cross-domain recommendations based on social big data and cognitive computing," *Mobile Networks & Applications*, vol. 23, no. 6, pp. 1610–1623, 2018.

[3] Y. Hao, J. Yang, M. Chen, M. S. Hossain, and M. F. Alhamid, "Emotion-aware video QoE assessment via transfer learning," *IEEE Multimedia*, vol. 26, no. 1, pp. 31–40, 2019.

[4] Y. Qian, M. Chen, J. Chen, M. S. Hossain, and A. Alamri, "Secure enforcement in cognitive internet of vehicles," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1242–1250, 2018.

[5] A. Yassine, S. Singh, M. S. Hossain, and G. Muhammad, "IoT big data analytics for smart homes with fog and cloud computing," *Future Generation Computer Systems*, vol. 91, pp. 563–573, 2019.

[6] X. Ma, R. Wang, Y. Zhang, C. Jiang, and H. Abbas, "A name disambiguation module for intelligent robotic consultant in industrial internet of things," *Mechanical Systems and Signal Processing*, vol. 136, article 106413, 2020.

[7] M. S. Hossain, M. A. Rahman, and G. Muhammad, "Cyber–physical cloud-oriented multi-sensory smart home framework for elderly people: an energy efficiency perspective," *Journal of Parallel and Distributed Computing*, vol. 103, no. 2017, pp. 11–21, 2017.

[8] K. Lin, C. Li, D. Tian, A. Ghoneim, M. S. Hossain, and S. U. Amin, "Artificial-intelligence-based data analytics for cognitive communication in heterogeneous wireless networks," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 83–89, 2019.

[9] M. S. Hossain and G. Muhammad, "An audio-visual emotion recognition system using deep learning fusion for a cognitive wireless framework," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 62–68, 2019.

[10] Y. Zhang, X. Ma, J. Zhang, M. S. Hossain, G. Muhammad, and S. U. Amin, "Edge intelligence in the cognitive internet of things: improving sensitivity and interactivity," *IEEE Network*, vol. 33, no. 3, pp. 58–64, 2019.

[11] X. Chen, L. Zhang, T. Liu, and M. M. Kamruzzaman, "Research on deep learning in the field of mechanical equipment fault diagnosis image quality," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 402–409, 2019.

[12] G. B. Chen, X. Sui, and M. M. Kamruzzaman, "Agricultural remote sensing image cultivated land extraction technology based on deep learning," *Revista de la Facultad de Agronomia de la Universidad del Zulia*, vol. 36, no. 6, pp. 2199–2209, 2019.

[13] P. Yin and M. M. Kamruzzaman, "Animal image retrieval algorithms based on deep neural network," *Revista Cientifica-Facultad de Ciencias Veterinarias*, vol. 29, pp. 188–199, 2019.

[14] G. Chen, Q. Pei, and M. M. Kamruzzaman, "Remote sensing image quality evaluation based on deep support value learning networks," *Signal Processing: Image Communication*, vol. 83, article 115783, 2020.

[15] G. Chen, L. Wang, and M. M. Kamruzzaman, "Spectral classification of ecological spatial polarization SAR image based on target decomposition algorithm and machine learning," *Neural Computing and Applications*, vol. 32, no. 10, pp. 5449–5460, 2020.

[16] B. Kayalibay, G. Jensen, and P. van der Smagt, "CNN-based segmentation of medical imaging data," 2017, http://arxiv.org/abs/1701.03056.

[17] M. S. Hossain and G. Muhammad, "Emotion recognition using secure edge and cloud computing," *Information Sciences*, vol. 504, no. 2019, pp. 589–601, 2019.

[18] M. M. Kamruzzaman, "E-crime management system for future smart city," in *Data Processing Techniques and Applications for Cyber-Physical Systems (DPTA 2019)*, C. Huang, Y. W. Chan, and N. Yen, Eds., vol. 1088 of Advances in Intelligent Systems and Computing, Springer, Singapore, 2020.

[19] Y. Zhang, Y. Qian, D. Wu, M. S. Hossain, A. Ghoneim, and M. Chen, "Emotion-aware multimedia systems security," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 617–624, 2019.

[20] M. S. Hossain, G. Muhammad, W. Abdul, B. Song, and B. B. Gupta, "Cloud-assisted secure video transmission and sharing framework for smart cities," *Future Generation Computer Systems*, vol. 83, pp. 596–606, 2018.

[21] O. K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition," *Neural Computing and Applications*, vol. 28, no. 12, pp. 3941–3951, 2017.

[22] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *European Conference on Computer Vision*, pp. 572–578, 2015.

[23] Y. Hu, Y. Wong, W. Wei, Y. Du, M. Kankanhalli, and W. Geng, "A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition," *PLoS One*, vol. 13, no. 10, article e0206049, 2018.

[24] S. Ahmed, M. Islam, J. Hassan et al., "Hand sign to Bangla speech: a deep learning in vision based system for recognizing hand sign digits and generating Bangla speech," 2019, http://arxiv.org/abs/1901.05613.

[25] U. Cote-Allard, C. L. Fall, A. Drouin et al., "Deep learning for electromyographic hand gesture signal classification using transfer learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 760–771, 2019.

[26] N. Tubaiz, T. Shanableh, and K. Assaleh, "Glove-based continuous Arabic sign language recognition in user-dependent mode," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 4, pp. 526–533, 2015.

[27] S. Ai-Buraiky, *Arabic Sign Language Recognition Using an Instrumented Glove, [M.S. thesis]*, King Fahd University of Petroleum & Minerals, Saudi Arabia, 2004.

[28] K. Assaleh, T. Shanableh, M. Fanaswala, F. Amin, and H. Bajaj, "Continuous Arabic sign language recognition in user dependent mode," *Journal of Intelligent Learning Systems and Applications*, vol. 2, no. 1, pp. 19–27, 2010.

[29] S. Halawani, "Arabic sign language translation system on mobile devices," *IJCSNS International Journal of Computer Science and Network Security*, vol. 8, no. 1, 2008.

[30] M. Mohandes, M. Deriche, and J. Liu, "Image-based and sensor-based approaches to Arabic sign language recognition," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 4, pp. 551–557, 2014.

[31] M. Almasre and H. Al-Nuaim, "Comparison of four SVM classifiers used with depth sensors to recognize Arabic sign language words," *Computers*, vol. 6, no. 2, p. 20, 2017.

[32] B. Hisham and A. Hamouda, "Supervised learning classifiers for Arabic gestures recognition using Kinect V2," *SN Applied Sciences*, vol. 1, no. 7, 2019.