A Saudi Sign Language Recognition System based on Convolutional Neural Networks

Alaa H Al-Obodi, Ameerh M Al-Hanine, Khalda N Al-Harbi, Maryam S Al-Dawas, and Amal A. Al-Shargabi

Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia

Abstract

Sign language is the main communication method for deaf people. It is a collection of signs that deaf people use to deal with each other. Deaf people find it difficult to communicate with normal people, as most of them do not understand the signs of the sign language. Sign language recognition systems translate the signs into natural languages and thus shorten the gap between deaf and normal people. Many studies have been done on different sign languages. There is a considerable number of studies on the standard Arabic sign language. In Saudi Arabia, deaf people use the Saudi language, which is different from standard Arabic. This study proposes a smart recognition system for Saudi sign language based on convolutional neural networks. The system is based on the Saudi Sign language dictionary, which was published recently in 2018. In this study, we constructed a dataset of 40 Saudi signs with about 700 images for each sign. We then developed a deep convolutional neural network and trained it on the constructed dataset. To have better recognition, we took images of the signs with different hand sizes, skin colors, lighting, backgrounds, and with/without accessories. The results showed that the recognition model achieved an accuracy of 97.69% for training data and 99.47% for testing data. The model was implemented in two versions: mobile and desktop.

Keywords: Sign language, recognition, convolutional neural networks.

I. INTRODUCTION

Sign language is a visual language used for hand movements and facial expressions by people with speech and hearing disabilities to communicate, and each gesture refers to a specific meaning [1]. As in oral language, each country or even region has its own sign language. There are two main types of sign languages: Alphabet notational hand gesture and Ideographic notational hand gesture. Alphabet's notational hand gesture is translated word letter by letter. The ideographic notational hand gesture expresses each meaningful word by a specific hand gesture. The second type is general and commonly used today [2].

Sign language recognition systems are classified into two categories: Device-based systems and Vision-based systems. Device-based systems use wearable tools for gesture tracking such as Microsoft Kinect, Leap Motion Sensors, and gloves [3]. Vision-based systems are techniques processing and analysis using artificial intelligence of images that are captured by the camera to recognize gestures; this is easier for the deaf because it does not need any device to sensor [4]. These systems are important for the development of sign language to make information accessible to the deaf public, and these systems have varying degrees of success.

Deaf people face problems in social life and depend on other understanding environments. Most people might not understand sign language clearly, even need human experts in sign language translation; this way is very expensive, uncomfortable, and may lead to the isolation of deaf people.

In this paper, we proposed a sign language recognition system designed for Saudi deaf people. It is a vision-based system in which a camera is used to shot the deaf sign and translates it into text. The system was based on a convolutional neural networks model. The model is trained on a dataset that contains 40 Saudi signs. We constructed the dataset so that every class includes around 700 images of different backgrounds and conditions.

The remaining of the paper is organized as follows: Section II presents the related works. Section III presents the proposed model, Section IV presents the system prototype, and Section V concludes the papers.

II. RELATED WORK

The related works are presented in terms of the datasets and the recognition methods used in previous studies.

II.I Datasets

Table 1 shows a summary of the datasets used in the state of the art studies. As shown in the Table, many studies have been done based on the standard Arabic, and most of them were to recognize alphabets, numbers, and a limited number of words.

II.II Recognition Methods

CNN is a powerful artificial intelligence tool in pattern classification. In the study of ElBadawy et al. [3], the authors used a 3D CNN to extract the Spatio-temporal features and then classify 25 classes that exist in the dataset.

In the study of Adithya et al.[4], the authors designed a system based on artificial neural networks for the automatic recognition of fingerspelling for Indian sign language by using digital image processing techniques and artificial neural networks.

The study of Hayani et al.[5] developed a system using convolutional neural networks (CNN), which are multi-layer neural networks that make use of deep learning to analyze images. The proposed model used CNN inspired fbyLeNet-5. They used seven adjacent layers, four layers to extract deep features, and three layers to classify them.

Table 1. A summary of the dataset used in previous studies

Description	Size	Language	Ref
Words (25)	200 Images	Arabic	[3]
Alphabet and Numbers from 0 to 10	7869 Images	Arabic	[5]
Words (50)	450 Videos	Arabic	[6]
Alphabets and Numbers from 0 - 10	900 Images	American	[7]
Alphabets (26 letters)	1100 images	Arabic	[8]
Alphabets	900 Images	Arabic	[9]
Six Alphabets	200 Images	Arabic	[10]
Alphabets (26 letters)	1100 images	Arabic	[11]
Alphabets (37 letters)	1147 Images	Bengali	[12]

The work of Ibrahim et al. [6] used Euclidean distance as a feature extraction method for the Arabic sign language. Hand segmentation was applied to the different frames of a video-based dataset. This method tracks the trajectory of an object in the image plane as it moves around a scene, detecting motion with an active camera.

Also, Maraqa and Abu-Zaiter [8] developed a system using a recurrent neural network (RNN) for static images. The algorithm is applied to all frames, and then the segmented skin blobs are used in identifying and tracking the hands with the help of the head.

In the study of Albelwi and Alginahi [10], the authors assessed the signs' classifications by the K-Nearest Neighbor (KNN) algorithm, which is one of the most commonly used methods in sign language recognition systems.

In the work of Hossen et al. [12], the CNN network consists of the following: convolution layer, max-pooling layer, ReLU layer, dropout layer, fully connected layer, softmax layer. The dataset was small, but the authors expanded the dataset by including more samples by scaling and rotating images.

In the study of Sawant and Kumbhar [13], the authors used Euclidean distance to calculate the difference between testing and training images, and then gestures can be recognized.

The study of Rao et al. [14] used the CNN architecture for classifying selfie sign language gestures. The CNN architecture was designed with four convolutional layers, and different filtering window sizes were considered. This improved the speed and accuracy of recognition.

The study of Hartanto and Kartikasari performed feature extraction for each hand gesture image in real-time, which was efficient in terms of computation time and good performance [15].

III. The Saudi Sign Recognition Model

III.I Dataset Construction

The database is based on Saudi Dictionary that was prepared and published by the Saudi Association for Hearing Disability in 2018. The dictionary contains thousands of signs from 28 fields such as medical, social, religious, and others. Fig. 1 shows a picture of the dictionary.

Fig. 1. The standard Saudi sign dictionary

In this study, we scope our dataset to the signs that contain only one gesture, and thus static images are making up the dataset. The signs which include multiple gestures were avoided. The images of the alphabet, numbers, and one word which is 'أنا' which means 'I' in the English language. Sample images of the alphabets' and numbers' signs are shown in Fig. 2 and 3.

It is essential to mention that the images were taken by a mobile camera so that the system can recognize late any images taken by primary cameras. Also, the images were taken by different people and with different backgrounds, lighting, and indoors and outdoors. This makes the proposed CNN model more general as it is trained on images at different conditions. In total, we took around 700 images for each class, resulting in 27,301 images.

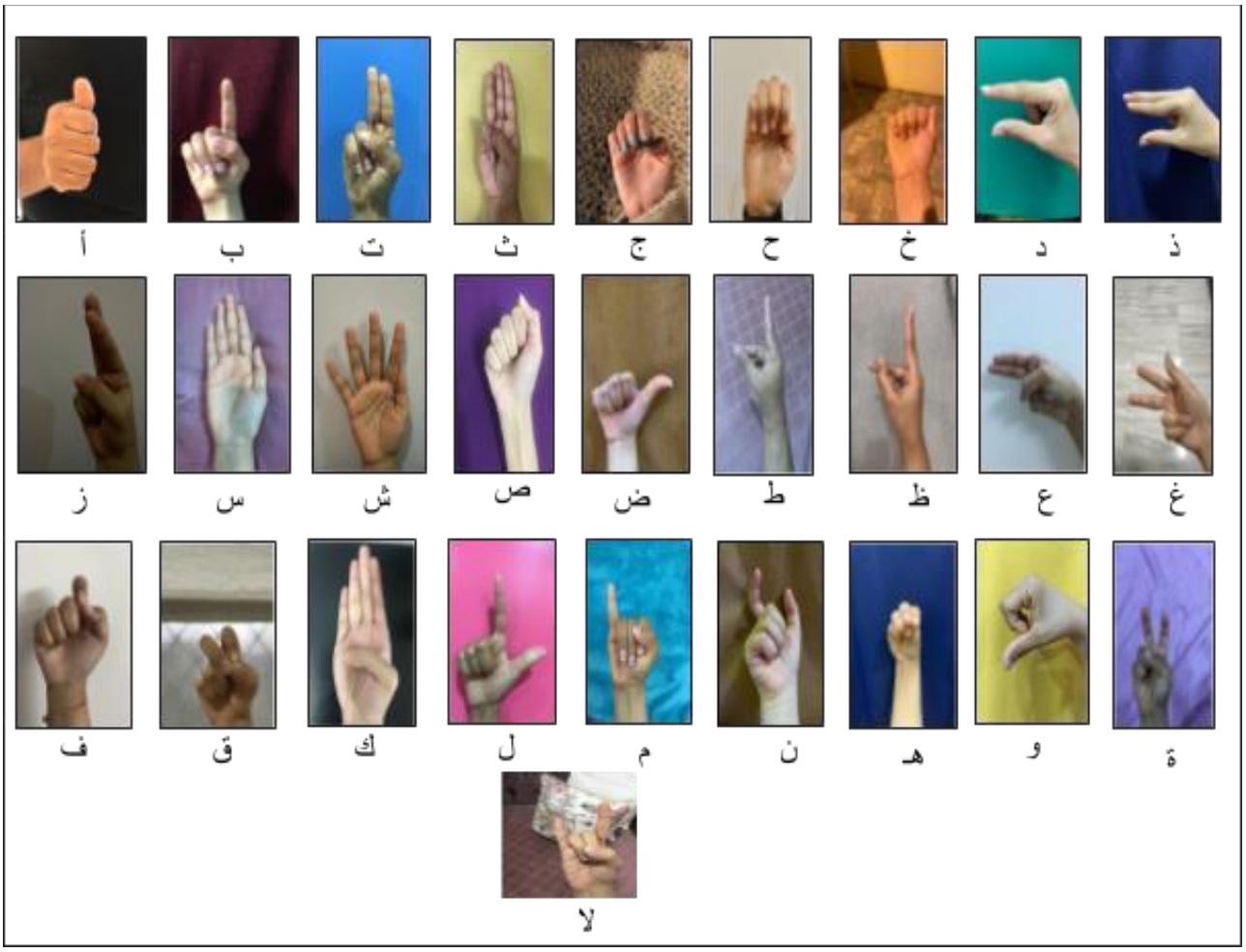

Fig. 2. Sample images of Alphabet Signs

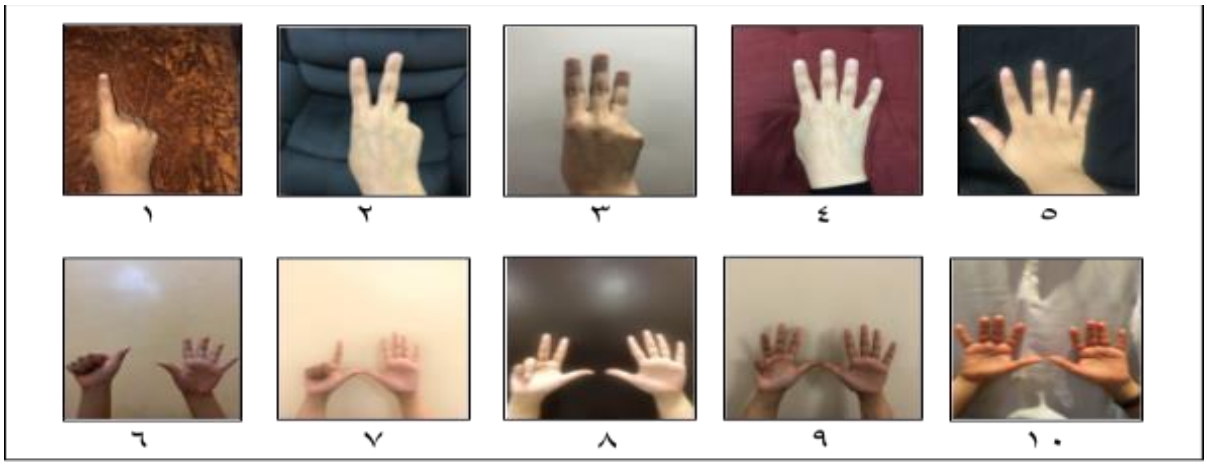

Fig. 3. Sample images of number Signs