https://journal.utripoli.edu.ly/index.php/Alqalam/index eISSN 2707-7179

Original article

Arabic Sign Language Recognition in Real Time Using Transfer Deep Learning

Noura Alshareef, Rema Abobake*, Asma Abd Aljalil

Department of Computer Science, Faculty of Science, Omar Al-Mukhtar University, Al-Bayda, Libya

ARTICLE INFO

Corresponding Email. *Reema.jawad@omu.edu.ly*

***Received**: 15-06-2024*
***Accepted**: 01-08-2024*
***Published**: 05-08-2024*

***Keywords. Keywords:** ArSL Recognition; Convolutional Neural Network (CNN), Hand Gesture Recognition, Mobilenet, Teachable Machine, Transfer Learning.*

ABSTRACT

People with hearing and speech disabilities in Arab society have obstacles to communicate since Arabic sign language (ArSL) has not been widely understood among society's members. Using technology to translate hand gestures from (ArSL) to written Arabic language can help bridge communication gaps and break down disability barriers in Arab society. latest research utilizes the camera to record the user's hand features to recognize its gestures. In this research a software was developed to recognize ArSL hand gestures in real time and immediately translate them to written alphabet Arabic letters utilizing Convolutional Neural Networks (CNNs), Teachable Machine. Our methodology involves data collection and preparation, Therefore an ArSL alphabet database has been created, with 28 categories representing the Arabic letters. For each letter, 400 images were captured. Then 87.5% of the dataset was passed to Google's Teachable Machine for training process, after that the dataset was tested and evaluated. Finally, the remaining 12.5% of dataset were used on local host for testing the model generated. The accuracy value for the model on local host was 92%. The experimental output shows the proposed model has good performance results in real time where the average recognition accuracy was 93.8%.

***Cite this article.** Alshareef N, Abobake R, Abd Aljalil A. Arabic Sign Language Recognition in Real Time Using Transfer Deep Learning. Alq J Med App Sci. 2024;7(3):730-739. https://doi.org/10.54361/ajmas.247338*

INTRODUCTION

One of life's most important elements is communication since it helps us express ourselves and our needs more easily, quickly, and effectively. Language is considered as crucial communication tools in which we interact with others, and it must be learned in order for efficient communication to take place; otherwise, we would have difficulties transmitting our intentions. As it is known, there are many different languages on the globe, both spoken and unspoken languages. And learning the appropriate spoken languages is the most common. However, there are some groups in our society that require nonverbal communication, such as deaf and mute who communicate using Sign Language (SL). Therefore, this kind of language needs to be taught to those in need and their relatives in order for them to communicate efficiently .SL is systematics approach of hand motion and hand gestures It is considered as a vital tool for daily interaction for the hearing and speech disabilities communities. Actually. Hand gestures are a fundamental part of human communication and may also serve as an essential form of human computer interaction [1]. However, SL is not widely used within the hearing community, and fewer are able to understand it. This creates a serious communication barrier between the deaf population and the rest of society, a problem that has yet to be entirely resolved.

Recognizing hand gestures is essential for overcoming numerous obstacles and simplifying people's lives. Countless applications can make use of machines' ability to comprehend human activity and its significance. Sign language recognition is one area of interest in particular [2]. However, SL is not standardized. And that is one of the most difficult

problems researchers faces. Every country around the world has its own sign language. Even though the Arabic sign language (ArSL) differs from country to another and that due to the lack of coordination between deaf communities and the institutions providing care for them in the Arab nation. Researchers and engineers have begun to adopt vision-based systems, which are quite affordable due to the utilization of cameras. Many sign language research efforts have been conducted extensively in English, Asian, and Latin sign languages, with little attention dedicated to Arabic due to the fact that there is no generally recognized database among researchers of Arabic sign language. As a result, the researchers were forced to create the datasets themselves, which was a difficult process. Despite the efforts of numerous academics, a comprehensive solution to this problem has yet to be found.

Sign language has become a method of oral teaching for deaf people, especially children, using hand gestures to mimic the forms of the letters of the alphabet. Figure 1 depicts the form of Arabic language signs (ArSL) in order to enhance communication with others. There are three layers for learning ArSL, Arabic alphabet gestures, Isolated Gestures (Word level), Continuous gesture (Sentence level) [3].

The greatest challenge that peoples with hearing and speaking disabilities face is a lack of communication with their communities. They need to pay translators and other services to communicate with others in their daily lives, but the costs are immense. Thus, finding a low-cost solution to this challenge is critical. Establishing a model based on computer vision and deep learning using Neural Networks is considered relatively cheaper, as it only requires a computer and a camera to recognize the Arabic Sign Language letters. However, another problem is raised which is the lack of a unified sign language database for Arabic letters. So, we are considering creating a database to gather the largest possible number of images of gestures representing the Arabic Sign Language Letters which will be used to train the network for the recognition process in real time.

In this paper, we will concentrate on Arabic alphabet gestures level, which will assist the deaf and mute people in overcoming communication problems and beginning to learn the Arabic alphabet.

Figure 1. Arabic alphabet signs [4]

RELATED WORK

ArSL is one of the Semitic languages spoken by around 380 million people as their primary official language globally is Arabic, an elegant and interesting language. Arabs have a tenable intellectual and semantic homogeneity [5].

In this is study, the authors focus on capability of Neural Network (NN) to help with ArSL hand gesture recognition [6]. The goal of the study was to demonstrate the usage of several NN models, comprising stationary and dynamic indicators, through the recognition of actual human gestures. They first showed how to use a Feed Forward Neural Network (FFNN) and a Recurrent Neural Network (RNN) in conjunction with their various topologies and totally and moderately repeating systems. The evaluation results showed that the suggested form with the full repeated design does have an implementation with a precision rate of 95% for stationary action recognition, which motivated them to further study their proposed structure.

Proposed the K-nearest neighbor classifier and statistical feature extraction method for the Arabic sign language. The main issue with the method is that it requires users to wear instrumented hand gloves in order to gather information about a specific gesture, which frequently causes the user great distress [7].

METHODS

First of all, the data gathering and preparation must be completed to ensure that the dataset is coordinated and consistent with the model as an input. So, in this section, we will go over how data is collected and preprocessed.

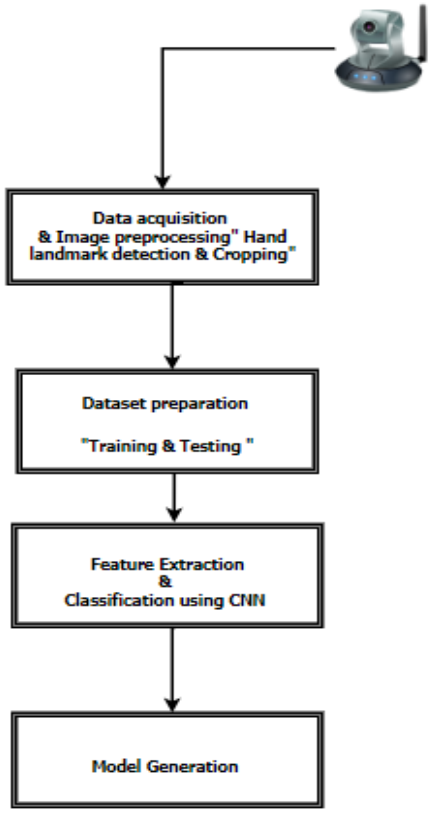

Figure 2. Methodology of proposed system

Dataset Collection and Preprocessing

In order to construct models based on machine learning, gathering data is an essential precondition. It is impossible to perform machine learning without a dataset. Preprocessing the data is the initial step in building a functional deep learning model. It is employed to convert the unprocessed data into a useful and efficient format.

Data Acquisition

A framework of images is the data obtained in vision-based recognition of gestures. Using image-capturing tools like webcams, stereo cameras, and regular video cameras, the input for such a system is gathered.

The proposed sign language detection system is built on the frame captured by a webcam on a laptop or PC. Image processing is done with OpenCV, an open-source collection of computer vision programming tools. To improve accuracy while utilizing a huge dataset, several photos of distinct sign language letters were captured from diverse individuals, perspectives, and lighting circumstances.

Image pre_processing

In this research, we are going to use a data collection program created using OpenCV, Pillow,Cvzone and MediaPipe library packages in Python.

The goal of OpenCV is to create applications using real-time computer vision technologies. Its numerous applications include facial recognition and detection, object identification, classifying human actions, tracking the movements of objects and cameras, 3D object modeling, and more. It has more than 2500 powerful machine learning and computational vision algorithms [8].

Hand tracking is the technique of utilizing computer vision to identify and track Hands' movements in real time. With OpenCV, create a new fully labeled dataset for Alphabetic Arabic sign language (AArSL).

For the purpose of implementing the suggested system, the hand sign images are taken with a webcam to build a dataset from row images for training and testing. The images were taken in various environments (different angles, varying illumination, sharpness and focus, shifting the size and distance of the objects, and taking pictures of various people).

Hand landmarks detection

First, MediaPipe is used for hands and hand key points detection.

Using the MediaPipe Hand Landmarker, one may locate important hand locations and overlay visual effects on the hands by detecting landmarks in an image. In order to produce hand landmarks in image coordinates, hand landmarks in world coordinates, and handedness (left/right hand) of multiple detected hands, it uses a machine learning (ML) model to work on picture data as static data or a continuous stream [9].

To track hands, MediaPipe Hands performs two processes for hands' tracking: palm detection and landmark detection. MediaPipe begins by identifying where the palms are in the input image, so We construct a VideoCapture object and supply the value "0." It is the system's camera IDAt present, the system is linked to a single webcam. Then, by using landmark detection, gives a total of 21 important points for every hand that is found. The Hand landmarks are displayed in Figure 3.

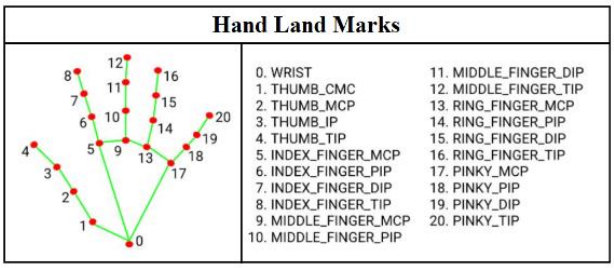

Figure 3. hand landmarks [9]

Crop images

The second part would be to crop the image once we get the hand. In Python, image cropping is done using NumPy array slicing. The slicing process needs to specify the start and end index of the first as well as the second dimension [10].The number of rows or the height of the image is the first dimension. and the number of columns or the width of the image is the second. After that the obtained images as shown in Figure 4, should be adjusted to have the same size using a white image which is generated by NumPy array with size 224*224.

As in figure 5, there is no variance among images of various gestures which is an important requirement of the Neural Network.

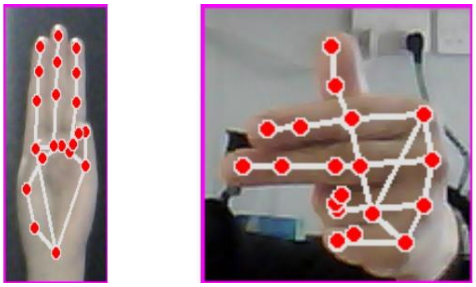

Figure 4. images of different gestures size

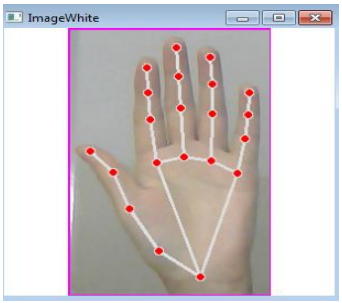

Figure 5. Cropping of hand and plotting landmarks

Feature Extraction

One of the most crucial parts in image processing is to select and extract important features from an image. Note that the Images when captured and stored as a dataset usually take up a whole lot of space as they are a huge amount of data. Feature extraction solves this by reducing the data after the important features are automatically extracted. It also helps

to preserve the precision of the classifier while simplifying its complexity. In the extraction of information from a wide range of data sources that may be represented as structured, multidimensional arrays of measurements have been made, in particular using convolutional neural network (CNN)-based DL approaches [11]. Here, feature extraction of CNN plays an important role in recognition tasks. The pooling layer in the Convolutional Neural Network main idea is to reduce the size of the data set by capturing the important features in which the network should be trained to recognize and extract them from the large input images.

Data Preparation

The dataset we built has 11200 images of Arabic sign language alphabets. For some storage plans, 28 folders are made, and each folder has about 400 images in it that are treated as datasets for the model's training and Testing.

First, we took 50 images from each folder for the model evaluation stage which will be done after completing the training process using Google's Teachable Machine. So, the total training data used is 350 from each class (total of 9800 images), and the remaining 1400 images are used for the evaluation process. By default, the 85% from the dataset passed to Google Teachable Machine (total of 8330 images) are used for training the model and 1470 images (the 15% remains from the passed dataset) will be used by Teachable Machine for testing the model.

Model generation & Classification

The gesture classification stage constitutes the last step of the whole process. The extracted appearance features are compared and classified in order to recognize and identify the presented gesture.

Once the dataset has been meticulously compiled, and all the necessary images for model development have been gathered and prepared, the next crucial step is to develop a machine learning model which can help us predict the hand sign.

Following the preparation stage of image data, the next pivotal phase involves modeling using Google's Teachable Machine to generate a machine learning model highlight three options available on Google's Teachable Machine website for creating machine learning models: image classification, voice classification, and pose classification, all utilizing transfer learning techniques with a pre-trained neural network [12]. In our work, the image classification method was utilized to create the model. We provided various classes with images relevant to sign language and fed them into Teachable Machine to generate the machine learning model, as depicted in Figure 6. This process facilitates easy customization of classes and enables the augmentation of precision in refining the machine learning model before initiating the model generation phase.

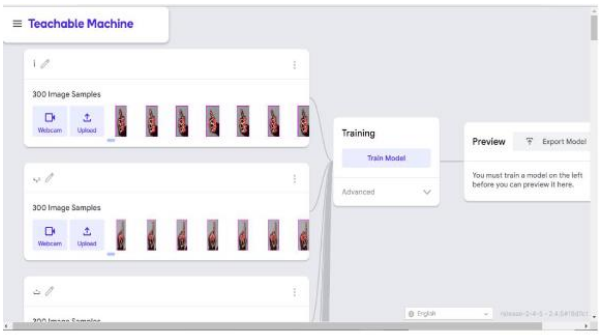

Figure 6. Input data sample and Model Generation

RESULTS

Our proposed approach involves real-time identification of a person's hand within each frame of camera video. Subsequently, we isolate the region of interest from the video frame. Our preprocessing algorithms resize the cropped hand image to dimensions of 224 x 224. Refer to Figure 7 for a sample of the collected data.

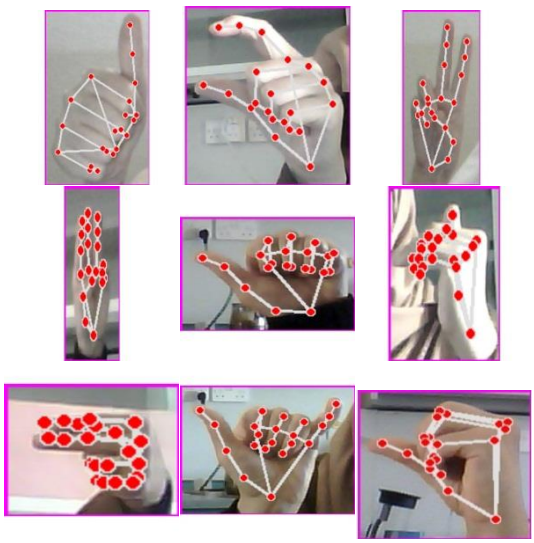

Figure 7. Data sample

After being cropped and preprocessed, the hand image is given into the teachable machine service, which uses the default parameters to train sample data. Table 1 shows some of the hyper parameters used in this study model.

Table 1. Hyper parameters utilized in our project

No	Hyperparameter	Alternative Value
1	Epochs	50
2	Batch Size	16
3	Learning Rate	0.0001

These models employ a technique known as transfer learning. There is a pre-trained neural network, and once you construct your classes, picture them as the neural net's last layer or step. Specifically, the image models are learning from pre-trained mobilenet models. The findings will then be shown on the top right for testing purposes before being exported to a (Ten-sorflow) keras file, which will be utilized to construct the Hand motion detection system.
The model's performance is assessed by comparing its classification of the testing data to the ground truth. The confusion matrix score reflects the model's accuracy in predicting all positive outcomes it generates [13]. Table 2 illustrated confusion matrix.

Table 2. Assessment using a confusion matrix.

Predicated Class	**Actual Class**	
	Positive (P)	Negative (N)
Positive (P)	True Positive (TP)	False Positive (FP)
Negative (N)	False Negative (FN)	True Negative (TN)

In the multi-class confusion matrix depicted in Figure 8, the highest values for each class level are located along the diagonal. This suggests that the model in Teachable Machine accurately predicted most of the correct values in the dataset.

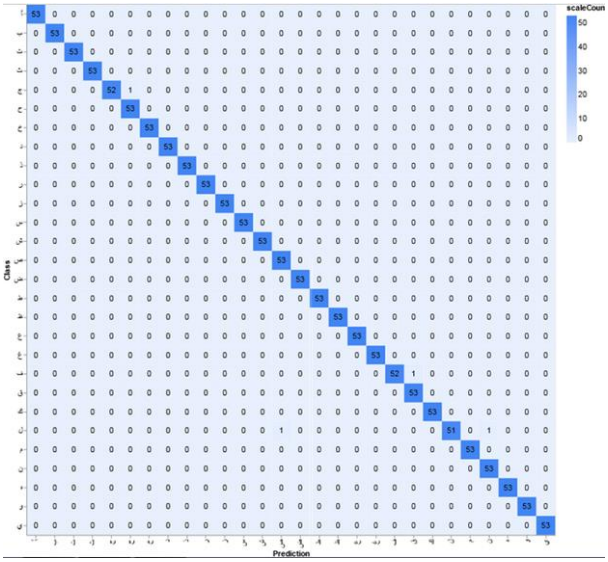

Figure 8. Confusion matrix of the trained model in Teachable Machine

TeachableMachine trains and tests all data within the class prior to exporting it. In this study, researchers acquired the source data in the form of a (Keras_model.h5) file show in Figure 9, which was prepared by Teachable Machine services as source code and then connected to Tensor flow machine learning using the Convolutional Neural Network (CNN) technique. Subsequently, the file is executed on the researchers' computer at the local host.

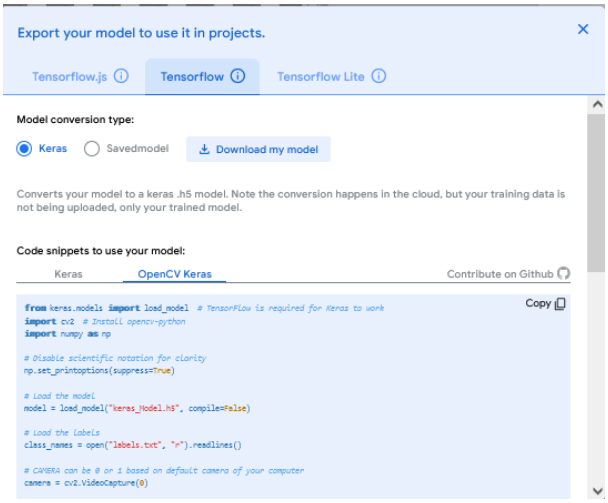

Figure 9. Export model and download (keras_model.h5) file on teachable machine.

Test model result

The proposed approaches were tested on 1400 images from our test dataset, which contains 28 classes, each class containing 50 images. Figures 10 include a confusion matrix for 28 classes to test the efficacy of the proposed teachable machine model.

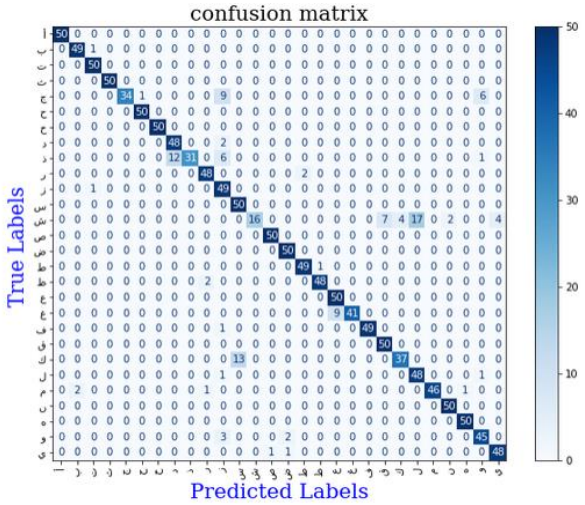

Figure 10. Confusion matrix for the test teachable machine model on our test dataset.

To evaluate the models, we employed accuracy, Recall, Precision and F1–score, computed as follows:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + false\ Negative} \quad 1$$

$$Precision = \frac{(True\ Positive)}{(True\ Positive + False\ Positive)} \quad 2$$

$$Recall = \frac{(True\ Positive)}{(True\ Positive + False\ Negative)} \quad 3$$

$$F1 - score = \frac{2 * Precision * Recall)}{(Precision + Recall)} \quad 4$$

Figure 11 illustrates the performance of the model on the test dataset.

	precision	recall	f1-score	support
أ alf	1.00	1.00	1.00	50
ب baa	0.96	0.98	0.97	50
ت taa	0.96	1.00	0.98	50
ث tha	1.00	1.00	1.00	50
ج jem	1.00	0.68	0.81	50
ح haa	0.98	1.00	0.99	50
خ kha	1.00	1.00	1.00	50
د dal	0.80	0.96	0.87	50
ذ thl	1.00	0.62	0.77	50
ر raa	0.94	0.96	0.95	50
ز zaa	0.69	0.98	0.81	50
س sen	0.79	1.00	0.88	50
ش shn	1.00	0.32	0.48	50
ص sad	0.98	1.00	0.99	50
ض thd	0.94	1.00	0.97	50
ط tah	0.96	0.98	0.97	50
ظ thh	0.98	0.96	0.97	50
ع aen	0.85	1.00	0.92	50
غ ghn	1.00	0.82	0.90	50
ف faa	1.00	0.98	0.99	50
ق gaf	0.88	1.00	0.93	50
ك kaf	0.90	0.74	0.81	50
ل lam	0.74	0.96	0.83	50
م mem	1.00	0.92	0.96	50
ن non	0.96	1.00	0.98	50
ه haa	0.98	1.00	0.99	50
و wow	0.85	0.90	0.87	50
ي yaa	0.92	0.96	0.94	50
accuracy			0.92	1400
macro avg	0.93	0.92	0.91	1400
weighted avg	0.93	0.92	0.91	1400

Figure 11. Performance of the model

Test in real time

In Figure 12, we observe samples of Arabic alphabets and their real-time classification. We process the video in real-time to predict hand gestures in each frame and classify them into one of the alphabet classes.

Figure 12. Samples of Real-time Arabic sign alphabet recognition.

Sign language recognition is a pivotal component for fostering inclusivity among the hearing-impaired community. This paper presents a real-time **ArSL** recognition system utilizing Convolutional Neural Networks (CNNs), with a focus on simplicity and accessibility through the use of Teachable Machine. The methodology involves data collection and preparation using OpenCV, a computer vision library, ensuring robustness in handling diverse sign gestures. The dataset, encompassing a variety of ArSL signs, was processed to facilitate effective model training on the Teachable Machine platform. This work contributes to advancing communication accessibility for the Arabic-speaking deaf community.

CONCLUSION

Our study focuses on a way to improve communication between individuals with speech difficulties and the general community. The main goal of this research was to create a model that would assist individuals with speech disabilities in communicating more effectively through ArSL while reducing the disadvantages of sign language, such as its lack of spread, the variety of its forms, and the difficulty of understanding it. We developed a model based on the database we established for Arabic Sign Language letters because there is no agreed-upon and authorized database that can be utilized during the training phase. To develop data leads according to the specified requirements, images were taken in variety condition of light and background They were captured in different lighting and background conditions with different individuals.

The database was utilized to train and test the model, and the proportion of real-time predictions was extremely high and acceptable. Our approach may also be used to accurately identify hand gestures for human-computer interaction. During testing, the accuracy on localhost achieved 92%, while in real time it was 93.8%.

REFERENCES

1. Rautaray SS, Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. 2012Nov6;43:1--54.
2. Z CMJO, Jaward MH. A review of hand gesture and sign language recognition techniques. 2019 Aug 4;10:131--153.
3. Tharwat G, Ahmed AM, Bouallegue B. Arabic sign language recognition system for alphabets using machine learning techniques. 2021 Aug 4;2021:1--17.
4. Aly S, Osman B, Aly W, Saber M. 12th International Computer Engineering Conference (ICENCO); 2016 Aug 4;:pp. 99--104.
5. Mohammed M. A study on Arabic sign language recognition for differently abled using advanced machine learning classifiers. 2021 Aug 4;12:4101--4115.
6. Manar M, Raed AZ. Recognition of Arabic Sign Language (ArSL) using recurrent neural networks. 2008 Aug 4;:478--481.
7. Noor T, Khaled A, null S. Glove-based continuous Arabic sign language recognition in user-dependent mode. 2015 Aug 4;45:526--533.

8. Vision based hand gesture recognition for human computer interaction: a survey. Artificial intelligence review. 2012.
9. Hand landmarks detection guide | MediaPipe [Internet]. 2023 [cited 2024 Feb 6].. Available from: https://developers.google.com/mediapipe/solutions/vision/hand_landmarker.
10. Cropping an Image using OpenCV [Internet]. 2023 [cited 2023 May 30].. Available from: https://learnopencv.com/cropping-an-image-using-opencv/
11. Maxwell AE, Odom WE, Shobe CM, Doctor DH, Bester MS, Ore T. Exploring the Influence of Input Feature Space on CNN-Based Geomorphic Feature Extraction From Digital Terrain Data. 2023 May 4;10:e2023EA002845.
12. Agustian D, Pertama PPGP, Crisnapati PN, Novayanti PD. Implementation of Machine Learning Using Google's Teachable Machine Based on Android. 2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS). 2021.
13. Soni KB, Chopade, Vaghela R. Credit card fraud detection using machine learning approach. 2021 Aug 4;4:71--76.

التعرف على لغة الإشارة العربية في الوقت الآني باستخدام التعلم العميق(CNN)

نورا الحبيب الشريف، ريما عبدالجواد ابوبكر، اسماء مصطفى عبدالجليل

قسم الحاسوب، كلية العلوم، جامعة عمر المختار، البيضاء، ليبيا

المستخلص

في المجتمع العربي، يواجه الصم والبكم صعوبات في التواصل، إذ لم تنتشر لغة الإشارة العربية (ArSL) على نطاق واسع بين أفراد المجتمع. ولحل هذه المشكلة يمكن استخدام التقنيات الحديثة لترجمة إشارات اليد من لغة الإشارة العربية إلى اللغة العربية المكتوبة، مما سيساعد على سد هذه الفجوة وكسر الحواجز التي تعيق التواصل بينهم وبين باقي افراد المجتمع العربي .وتستخدم الأبحاث الحديثة الكاميرات لتسجيل ملامح يد المستخدم والتعرف على إشاراته وترجمتها الى لغة مكتوبة. في هذا البحث، تم تطوير برنامج لتحديد إشارات لغة الإشارة العربية في الوقت الفعلي وترجمتها فوراً إلى الأحرف العربية المكتوبة، باستخدام تقنيات الشبكات العصبية التلافيفية (CNNs) والنماذج القابلة للتعلم كـ "Teachable Machine"من جوجل .وتضمنت منهجية البحث جمع البيانات وإعدادها، حيث تم إنشاء قاعدة بيانات لأبجدية لغة الإشارة العربية، بـ28 فئة تمثل الحروف العربية. وتم التقاط 400 صورة لكل حرف. ثم تم استخدام 87.5% من البيانات لتدريب نموذج"Teachable Machine" ، بينما استخدم 12.5% المتبقية لاختبار النموذج المطور محلياً. وقد بلغت دقة النموذج 92% في الاختبارات المحلية، كما كانت دقة التعرف على الإشارات في الوقت الاني 93.8% في المتوسط.

الكلمات المفتاحية: التعرف على لغة الاشارة العربية , الشبكات العصبية التلافيفية(CNNs) , التعرف على ايماءات اليد, التعلم العميق, الرؤية الحاسوبية.