Towards Bidirectional Conversion between Arabic Sign Language and Speech/Text

Souha Ben Hamouda[1,*], Wafa Gabsi[1] and Bechir Zalila[1]

[1]*ReDCAD Laboratory, ENIS, University of Sfax, Tunisia*

Abstract

Sign language is the essential communicating tool for deaf and non-verbal people. Using sign language, deaf and mute people can communicate among themselves but they find it though to face the outside world. The automatic interpretation of sign language and its conversion into text message and voice format remains a challenging task to break the barrier between deaf and the wider majority of non-deaf people. Different approaches and techniques were defined to propose solutions such as android application, smart glove and gesture recognition based on image processing and artificial intelligence. In this paper, we provide an overview of the sign language and other ways used in communication between deaf and non deaf people as well as factors involved in this communication. We then describe our approach, which ensures bidirectional conversion between Arabic sign language and speech/text.

Keywords

Arabic Sign language, Artificial intelligence, Recognition, Image processing, Bidirectional conversion

1. Introduction

Deaf individuals are those with either complete or partial hearing loss, often referred to as hard of hearing or Deaf. To communicate, they rely not only on facial expressions, eyebrows and eye movements, but also a language composed of gestures and signs. Indeed, to express themselves, they use not only facial expressions, eyebrows and eyes, but also a language made up of gestures (or signs) and movements. For this, school, social and professional integration presents a major problem for these people, leading to feelings of isolation, exclusion, and introversion, even within their own families. This comes down to the fact that the majority of people do not know sign language and do not want to learn it. This has led to a conflict between teaching oral language and sign language among educators. In this context, the objective of research in this field is to help these categories of people, using sign language as their main mode of communication, aiming to enhance their participation and integration.

Various researchers have explored technology-driven solutions for automatic bidirectional translation between sign language and spoken/written language [1]. Researchers have tried to develop different solutions translating sign language from Arabic, French, English alphabets

and other written or spoken languages [2, 3]. Various contributions have been made to translate hand and mouth movements, using sensors [4], mobile applications [5] [6] and gloves [7] [8]. Based on image processing and deep learning algorithms, the sign language alphabets were then translated into speech or text. The analysis of these works revealed certain problems such as:

- Most of the work has aimed at recognizing the alphabetical letters of each language while in reality the deaf use words and sentences. Additionally, each of the quoted contributions has its own limitations and costs, rendering them commercially impractical."
- Few works focus on the processing of Arabic sign language.
- Most of the work aims at recognizing gestures and translating them into text/audio but not the other way around.
- Each of the quoted contributions has its own limitations and its own cost which make them commercially unusable.
- Each of the existing works is generally interested in a category of sign language such as static or dynamic gestures. In our work, we address these gaps by proposing a two-way recognition system for Arabic sign language.

As part of our work, to help the deaf and participate in their social and professional integration, we are aiming for the two-way recognition of Arabic sign language.

Our approach offers a two-way translation system from Arabic words to text/audio via mobile development. In a first sense, we designed and developed a 3D hand allowing to translate gestures following a voice recognition of words in Arabic. Our testing involved static as well as dynamic gestures. Additionally, we present a mobile application utilizing image processing for sign recognition and translation to text/audio. Our solution provides an affordable means of facilitating bidirectional communication. The remainder of this paper is structured as follows: Section 2 presents background concepts related to Deep learning and Image processing algorithms. In Section 3, we present an overview of our proposed approach. Then, section 4 gives and discusses results. Finally, Section 5 concludes this paper and gives on going work.

2. Background

Faced with the significant evolution experienced by sign language, innovating technology with a view to improving the properties and performance of communication with the deaf becomes a challenge. First, we begin by presenting the modes of communication of deaf people and the various factors involved. Second, we define and classify different sign languages. Next, we present the importance of body expression and the techniques used for sign language recognition.

2.1. Communication methods

Deaf and hard of hearing people have different profiles. They make choices based on their situation and personal history. These choices are neither definitive nor exclusive. The same person can communicate differently during his life or according to the contexts (family, work,

friends...). People who are deaf and hard of hearing can combine the following elements to communicate:

- Lip reading allows deaf individuals to better perceive speech. A specific learning allows the mastery of this reading.
- Sign language is a separate language that involves learning each specific language and has its own grammar and vocabulary.

Besides sign language, there are several means of communication to get in touch with a deaf person such as: e-mail, video-interpretation services (SVIsual), mediation centers for deaf people, messages of mobile phone text, letters, etc.

Often parents of deaf people also learn how to sign. The native sign language of their parents will be their first language and this before any spoken language. In addition, parents, brothers and sisters of deaf children learn to sign to communicate with them. Many people also learn sign language in their spare time because they have deaf friends.

2.2. Factors involved in communication

During the conversation with a deaf person several factors intervene to establish a communication:

- **Brightness:** The deaf person must be in a strategic position in order to provide him with a general visual overview of the place where he is and the place must be well lit so that he has good visibility.
- **Eye contact:** It is important to have eye contact with the deaf person. We must avoid excessive movement. There is no more eye contact between people. If contact is not possible because of the distance, there are several other ways to get their attention: Knocking hard on the floor or gently on the table so that they feel the vibrations, or turning off the lights moving the arm in the visual field of the deaf person.
- **Speed of speech:** Do not speak too quickly or too slowly. We must vocalize clearly using short and simple sentences for a good understanding of the subject of the conversation.
- **Way of speaking:** You have to speak without impeding the mouth so that the deaf person can read the lips.
- **Facial expression:** Is an element of great help, as well as the components that complement verbal speech, gesture, writing, etc. If the deaf person does not understand, the message sent must be expressions differently.
- **Position in relation to the others:** If several people are going to intervene in the conversation, it is advisable to stand in a circle to facilitate good visibility. To attract the attention of the deaf, you have to touch their arm or shoulder. Never touch a deaf person, either in the back or in the head.

One can always ask for help, guidance and assistance in federations and associations of deaf people. This is the best way to optimize communication according to the different particularities of groups of people or each group of deaf people.

2.3. Sign language

Sign language is a language that incorporates the implementation of hand movements, body orientation and facial expressions for communication, without relying on sound waves.

2.3.1. Principle

Sign language is a communication system used by deaf and hard of hearing people to communicate not only with each other but also with the hearing world. It is a language open to all. It is not necessarily limited to deaf people but can be used by parents, companions, specialized educators, doctors. Apprenticeships are offered to learn sign language to enable deaf people to communicate with each other and with their loved ones. Likewise, they also allow hearing people to communicate with the hearing impaired.

2.3.2. Classification of sign languages

The sign languages have been classified by families according to the language. A classification was established by Henri Wittmann in 1991 [9]. The latter offers the following list of families:

- **French Sign Language Family:** Also known as the Francosign Language Family. It descends from the old French Sign Language which developed in France from the 17th century.
- **Arabic sign language family:** It mainly includes sign languages from the Arabic-speaking Middle East.
- **German sign language family:** This includes German and Polish sign languages.
- **British Sign Language family:** It is historically derived from a prototype variety of British Sign Language (BSL).
- **Japanese sign language family:** It includes Japanese, Taiwanese and Korean sign languages. There are few difficulties in communication between these three languages.
- **Lyons sign language family:** It only includes the Lyons sign language and the sign languages of French-speaking and Flemish Belgium.

Even there are many sign languages, they are all based on some mouvements of the fingers, hands and lips which is called body expression.

2.3.3. Importance of body expression

Different factors and parts of the body are involved in producing sign language:

- **Fingers:** The difference between the gesture of the letter V and the X, in the position of the folded fingers for the X.
- **Hands:** We can consider different positions of the hands such as above, in front and in the shape of a fist to perform static and dynamic gestures.
- **Movements:** which make the hand turn and the fingers move.
- **Location:** The space behind and in front of the speaker is used to express time.

- **Expressions of the face, the eyes, the movements of the eyebrows or the mouth** are also important to express emotions, questions or feelings.

In addition to facial and body expressions, deaf person's location is also an important factor in understanding the messages exchanged.

A facial expression will even, in some cases, make it possible to tell the difference between two words signed in the same way. For this, the deaf must position themselves well in front of their interlocutor when they express themselves with sign language. Simply turning their head away can cause the person they are communicating with to miss many of the intricacies of the conversation. All of these body expressions are gestures formed to spell out letters, words, numbers and also sentences. In our context, we will focus on hand gestures to produce static and dynamic gestures.

2.3.4. Types of gestures

Hand gestures can be categorized.

- **Static gestures** A static gesture is a particular position of the hand, represented by a single image [10]. Hand gestures express certain information through a form of static movement.
  For this, researchers have been inspired to utilize temporal models. The objective of which is to correct errors in the recognition of static hand gestures [11]. Temporal patterns are also present in gestures.

Figure 1: Arabic alphabets signs [12].

- **Dynamic gestures**
  Dynamic gestures are gestures in motion, represented by a sequence of various images. Dynamic hand gesture means that we have gesture recognition using dynamic hand [10]. Hand gestures express certain information through dynamic movement of the arm, wrist and fingers. A dynamic hand gesture of finger movements can be considered as a temporal sequence of static hand gestures [11].

3. Description of the proposed approach

Our main goal is to propose an arabic sign language recognition approach based on words recognition with lower costs. For that, we propose a dual way communication system based on translation from signs of Arabic words to spoken language based on image processing and vice versa. In this section, we present an overview of our proposed approach in both ways of communication. Then, we give details about each of them.

Our approach is bidirectional allowing translation in both directions between speech and signs. For this, we have broken down our approach into two axes. The first axis aims to translate Arabic words into sign language. To do this, we propose the creation of a database on sign language. To realize the movements of the hands, we propose an arduino application. The second axis aims to translate Arabic sign language into audio or text based on image processing. We briefly describe each of these axes in this section.
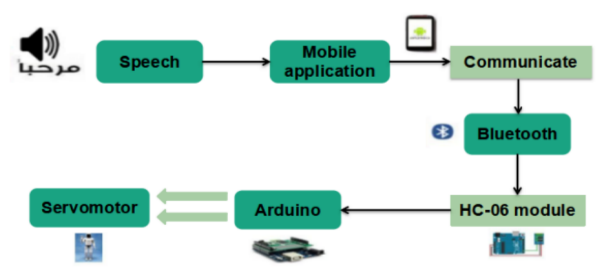
3.1. From audio into sign language

Figure 2: Block Diagram of Speech to sign language Synthesis.

For the transformation of spoken words into gestures, we have proposed a process based on different stages, as shown in Figure 2. From the spoken words, we start with voice recognition allowing the generation of the corresponding text. For voice recognition, we used a Voice Recognition Module V3 microphone. This module accepts spoken word as input and generates the corresponding text as output. For the translation of gestures from text, we used an Arduino board and we designed a 3D hand. The arduino uno card makes it possible to transform the words into order of movements based on servomotors. Servomotors are motors of a particular type, very popular for turning something up to a very precise position. These movements make it possible to move the different fingers of a 3D hand in the positions indicated according to the gestures associated with the spoken words.

This section includes screenshots of some word interfaces processed by the mobile application and the equivalent gesture generated by the 3D hand accompanied by a brief description.

Figures 4 and 3 show the voice recognition of the two words إِثنَان and أَربعة respectively with their static sign language gestures that suit them. While, figures 5 and 6 show the voice recognition of the two words عَشرة and عَمَل جَيد respectively with their very similar static sign language gestures. Regarding the dynamic gestures, we give in figure 7 an example of the gesture conversion corresponding to the word السّبت. We notice that there are many similar
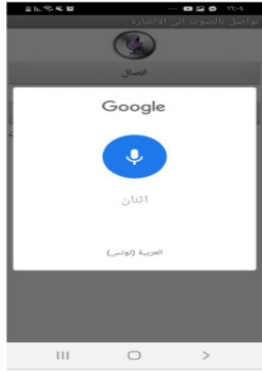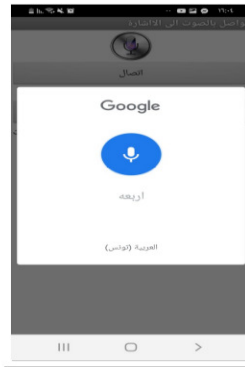
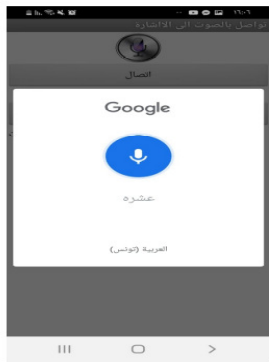**Figure 3:** Sign of the word إِثنَان
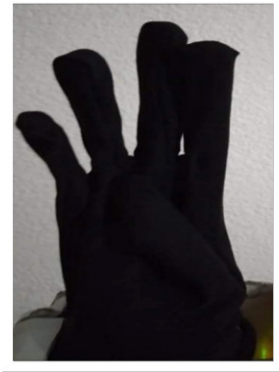

Figure 4: Sign of the word أَربعة

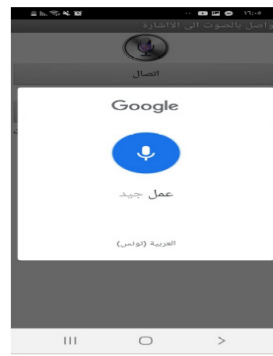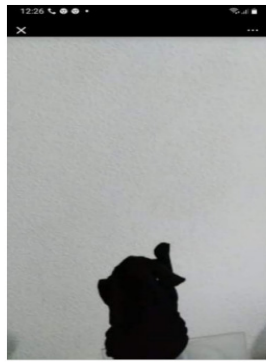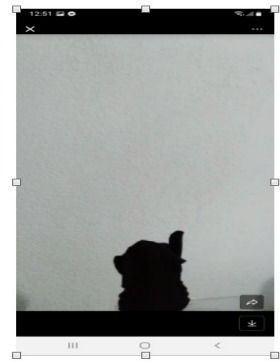

Figure 5: Sign of the word عَشرة

Figure 6: Sign of the word عَمَل جَيد

gestures for different words and meaning. For that, in future work, we aim tio focus on meaning of words and giving conversion of sentences.
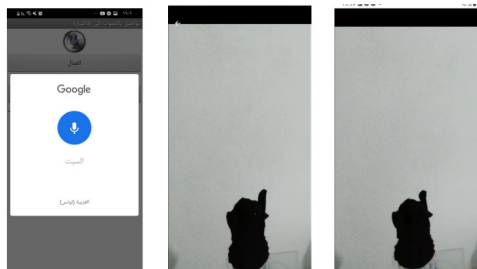

Figure 7: Speech word recognition السّبت and its corresponding gesture

3.2. From Arabic sign language into audio or text

Similarly, in the opposite direction, the proposed approach for translating gestures into words follows different steps as shown in Figure 8. First, we built a database of gestures. The second step is the pre-processing of this database to move on to the processing phase. The third step is the use of the algorithms necessary for the classification. Similarly, the same steps must be applied for the words of each sign. In this phase, it is necessary to concretely define the steps of our application. We mention in the rest of this paper the different tasks carried out to develop this work.
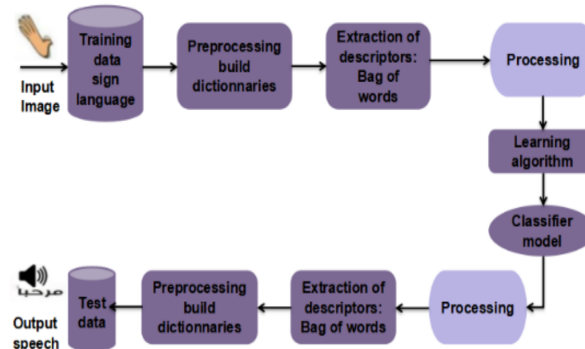

Figure 8: Block Diagram of sign language to Speech Synthesis.

The last step of our approach was the realization of our 3D hand. The hand consists of 5 straws representing the fingers. Each of the fingers can be in different positions by varying the gesture. For this, for each word, it is necessary to give the position of the corresponding fingers based on the servos responsible for moving the fingers. As a proof of concept, we have programmed the necessary movements for a few words. The main advantage of our approach is the price of the material necessary for the construction of the 3D hand. Indeed, the hand does not cost too much. It turns around 100$. This makes our solution much less expensive in terms of materials compared to others (gloves or electric hands).

We will detail in what follows, the different steps to translate gestures into text. Our focus is on sign detection, which detects numbers from zero to nine.

- **Creating the dataset for gesture detection** In order to validate our approach, we have created a database of gestures corresponding to Arabic numerals for learning and another database which is the test database.
  Each base contains ten folders containing images captured using a camera consisting of gestures corresponding to the numbers from zero 0 to nine 9.
  The training database contains 2,060 images. For each number, we recorded 206 gesture images by varying the background and the lightings. On the other hand, the test base is composed of 10 gestures.
  On both bases, we perform the necessary experiments. These two databases represent an important source for the development of image recognition applications. They also

provide a resource for future researchers in order to test and evaluate the applications developed for this purpose.

- **Gesture detection**
Now to detect a hand we get the live camera feed using OpenCV (OpenCV is an open source computer vision and machine learning software library.) and create a ROI (region of interest) which is nothing but the part of the frame we want to detect the hand for gestures that will be saved in a directory. Here the gestures directory contains the two folders train and test containing captured images. The blue box provides the live camera feed from webcam.

To differentiate the background, we calculate the accumulated weighted average for the background and then subtract it from images containing an object in front of the background that can be distinguished as the foreground. This is done by calculating the accumulate weight for some images (here for 60 images) we calculate the accumulate avg for the background. Once we have accumulated the average for the background, we subtract it from each frame we read after 60 frames to find any object that is covering the background.

When the edges are detected (or the hand is present in the ROI), we start recording ROI images in the train and the test set respectively for the digit for which we detect it.

- **CNN Training**
On the created dataset, we train a CNN (Convolutional Neural Network). First, we load the data using keras' ImageDataGenerator through which we can use the flow from directory function to load the train and test set data, and each of the names of the digital folders will be the class name for the loaded images.

CNN consists of multiple layers like the input layer, Convolutional layer, Pooling layer, and fully connected layers as it is shown in the figure9. The Convolutional layer applies filters to the input image to extract features, the Pooling layer downsamples the image to reduce computation, and the fully connected layer makes the final prediction. The network learns the optimal filters through backpropagation and gradient descent. The plotImages function is for plotting images of the dataset loaded.
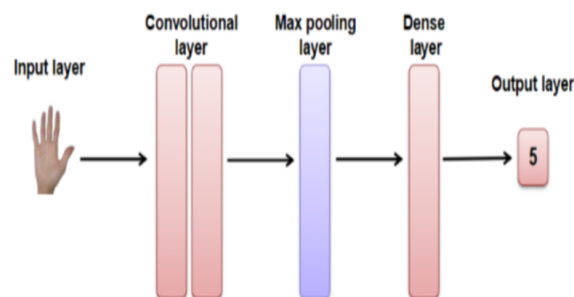

Figure 9: Simple CNN architecture.

Now, we design the CNN as follows (or based on trial and error, other hyperparameters

can be used).


```
Please report this to the TensorFlow team. When filing the bug, set the verbosity to 10 (on Linux, `export AUTOGRAPH
  VERBOSITY=10`) and attach the full output.
Cause: 'arguments' object has no attribute 'posonlyargs'
To silence this warning, decorate the function with @tf.autograph.experimental.do_not_convert
207/207 [==============================] - 28s 126ms/step - loss: 1.8231 - accuracy: 0.4190 - val_loss: 0.6655 - val_
  accuracy: 0.8000 - lr: 0.0010
Epoch 2/10
207/207 [==============================] - 18s 88ms/step - loss: 0.7758 - accuracy: 0.7386 - val_loss: 0.3684 - val_
accuracy: 0.8000 - lr: 0.0010
Epoch 3/10
207/207 [==============================] - 20s 96ms/step - loss: 0.4841 - accuracy: 0.8380 - val_loss: 0.1885 - val_
accuracy: 0.9000 - lr: 0.0010
Epoch 4/10
207/207 [==============================] - 19s 92ms/step - loss: 0.3409 - accuracy: 0.8899 - val_loss: 0.0678 - val_
accuracy: 1.0000 - lr: 0.0010
Epoch 5/10
207/207 [==============================] - 19s 92ms/step - loss: 0.2345 - accuracy: 0.9263 - val_loss: 0.0482 - val_
accuracy: 1.0000 - lr: 0.0010
Epoch 6/10
207/207 [==============================] - 21s 99ms/step - loss: 0.1673 - accuracy: 0.9549 - val_loss: 0.0861 - val_
accuracy: 1.0000 - lr: 0.0010
Epoch 7/10
207/207 [==============================] - 22s 105ms/step - loss: 0.0652 - accuracy: 0.9888 - val_loss: 0.0048 - val_
  accuracy: 1.0000 - lr: 5.0000e-04
Epoch 9/10
207/207 [==============================] - 20s 99ms/step - loss: 0.0537 - accuracy: 0.9922 - val_loss: 0.0051 - val_
accuracy: 1.0000 - lr: 5.0000e-04
Epoch 10/10
207/207 [==============================] - 19s 90ms/step - loss: 0.0461 - accuracy: 0.9927 - val_loss: 0.0142 - val_
accuracy: 1.0000 - lr: 5.0000e-04
```

Figure 10: Training precision

In training, we utilize the Reduce Learning Rate (LR) on plateau and early stopping (Stop training when a monitored metric has stopped improving) are used, and both depend on the loss of the validation data set.

After each period, the accuracy and loss are calculated using the validation dataset and if the validation loss does not decrease, the LR of the model is reduced using the Reduce LR to prevent the model to exceed the loss minimums and we also use the earlystopping algorithm so that if the validation accuracy continues to decrease for certain periods, learning is stopped.

The example contains the callbacks used, it also contains the two optimization algorithms used which are SGD (stochastic gradient descent, which means the weights are updated at each training instance) and Adam (combination of Adagrad and RMSProp).

We found that for the SGD model seemed to give higher accuracies. During training, it's evident that we found 100% training accuracy.

- **Predict the gesture**

  To predict the gesture, we create a bounding box for detecting the ROI and calculate the accumulated average as we did in creating the dataset. This is done for identifying any foreground object. Then, we find the maximum contour and if contour is detected that means a hand is detected so the threshold of the ROI is treated as a test image. We load the previously saved model using keras.models.load model and feed the threshold image of the ROI consisting of the hand as an input to the model for prediction. Getting the imports for the gesture model. After that, we load the model that we had created earlier and set some of the variables that we need, i.e, initializing the background variable, and setting the dimensions of the ROI. Function to calculate the background accumulated weighted average (like we did while creating the dataset...)for Detecting the hand now on the live cam feed.

The test of our model has good results. It shows a good recognition for most numbers. Figures 11 and 12 show snapshots of the recognition of the two numbers 1 and 4 respectively. While figure 13, shows that the system mistranslated the digit gesture **8**. It translated to the number **9**. This error may be due to the similarity between the different forms of gestures in Arabic sign language. Like it is the same error in figure 14 for the recognition of number 9.
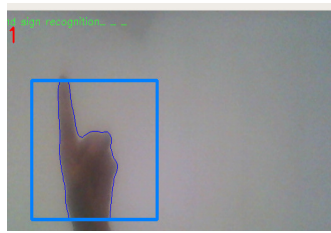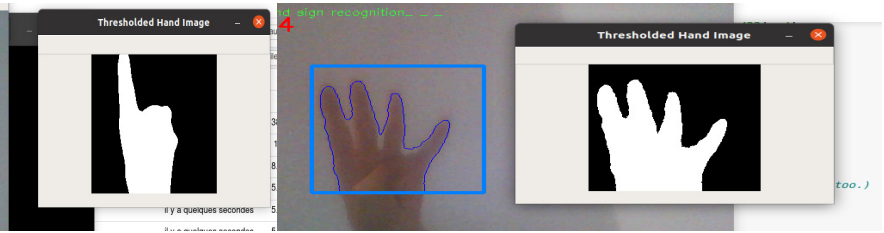
Figure 11: Recognition of number 1

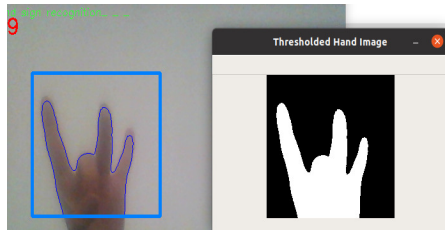**Figure 12:** Recognition of number 4

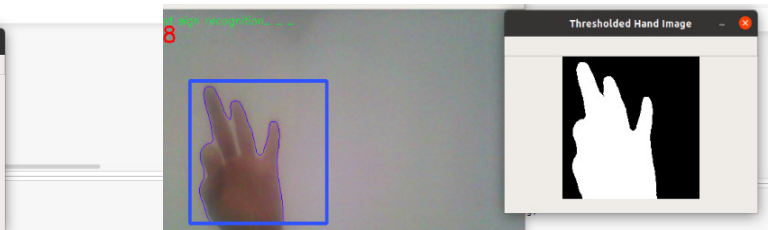
Figure 13: Bad recogntion of number 8

Figure 14: Bad recognition of number 9

4. Evaluation

In this section, we give and discuss results of both proposed approaches to convert speech to gestures and conversely.

4.1. From audio to 3D gesture

To validate our approach, we tested the performance of this approach compared to other approaches.

	Average static gestures	Average dynamic gestures	Average
Number of test images	30 images	30 images	
Rate of correct detection	95.82%	88.33%	93.32%
Rate of false detection	4.14%	11.66%	6.64%

Table 1
Rates of good and bad conversion of both static and dynamic signs

Preliminary tests of our 3D hand showed satisfactory overall accuracy of the correct translation of Arabic words translated into sign language.

Table 1 gives the rates of good detections and false detections. The results are generally satisfactory with an average of approximately 93.32% of good detections for 6.64% of false

detections. The results are generally satisfactory with an average of static gestures around 95.82% of good detections for 4.14% of false detections.

The results are globally satisfactory with on average dynamic gestures about 88.33% of good detections for 11.66% of false detections. Our application developed for sense 1 (from speech to gesture) saves time as it analyzes words and not letters. Moreover, the use of 3D hands helps ordinary people to communicate with deaf people without learning their own language. From the usage point of view, our application is simple to use and solves the problem of cost and availability of material.

4.2. From gesture to Arabic text

We have successfully developed the approach for sign language digit detection. This is an interesting machine learning python project to gain expertise. This can be further extended to detect Arabic words. This approach shows good gesture detection performance in terms of good classification rate and computation time as well.

The results are globally satisfactory with an average of approximately 94.36% of good detections for 5.64% of false detections.

5. Conclusion and perspectives

Sign language is an important area of research that is attracting increasing attention from research communities that aims to make life easier for deaf people. Deaf individuals have limits in terms of communication. For this, researchers have developed translation applications capable of translating sign language into written language and vice versa. Each of the existing solutions admits disadvantages considering certain criteria such as the language treated, the obstacles, the treatment of letters and not of words, the high cost and the unavailability of materials. In addition, there are few works aimed at bidirectional communication and are generally interested in the translation of gestures into speech/text and not the reverse.

In our work, we proposed a low-cost and easy-to-use two-way communication system based on Arabic word recognition. We were able to develop and implement the first axis of our approach, translating words into gestures based on voice recognition and using an Arduino board. We tested and validated it, achieving a 93% recognition rate. Regarding the second axis, we also developed our own approach allowing the recognition of gestures and their translation into text. We have tested it currently using numbers. We have proven the effectiveness of our approach with a highly satisfactory recognition rate.

For these reasons, we aim in the short term to improve the development of the two axes through larger databases of words and gestures for increased credibility.

In the medium term, we aim to integrate facial recognition and lip reading to improve our approach and and distinguish words that are similar.

In the long term, we aim to improve our approach to consider sentences while processing sequences of words within a reasonable time frame.

References

[1] B. H. Souha, G. Wafa, Arabic sign language recognition: Towards a dual way communication system between deaf and non-deaf people, in: 2021 IEEE/ACIS 22nd International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2021, pp. 37–42. doi:10.1109/SNPD51163.2021.9705002.

[2] A. Er-Rady, R. Faizi, R. O. H. Thami, H. Housni, Automatic sign language recognition: A survey, in: 2017 International Conference on Advanced Technologies for Signal and Image Processing, ATSIP'17, 2017, pp. 1–7. doi:10.1109/ATSIP.2017.8075561.

[3] R. Rastgoo, K. Kiani, S. Escalera, Sign language recognition: A deep survey, Expert Systems with Applications 164 (2021) 113794. doi:https://doi.org/10.1016/j.eswa.2020.11379.

[4] P. Kumar, H. Gauba, P. Pratim Roy, D. Prosad Dogra, A multimodal framework for sensor based sign language recognition, Neurocomputing 259 (2017) 21–38. doi:https://doi.org/10.1016/j.neucom.2016.08.132, multimodal Media Data Understanding and Analytics.

[5] Setiawardhana, R. Y. Hakkun, A. Baharuddin, Sign language learning based on android for deaf and speech impaired people, in: 2015 International Electronics Symposium (IES), 2015, pp. 114–117. doi:10.1109/ELECSYM.2015.738082.

[6] S. Ghanem, C. Conly, V. Athitsos, A survey on sign language recognition using smartphones, in: Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments, PETRA '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 171–176. doi:10.1145/3056540.3056549.

[7] M. Mohandes, J. Liu, M. Deriche, A survey of image-based arabic sign language recognition, in: 2014 IEEE 11th International Multi-Conference on Systems, Signals Devices (SSD'14), 2014, pp. 1–4. doi:10.1109/SSD.2014.6808906.

[8] S. Sarker, M. M. Hoque, An intelligent system for conversion of bangla sign language into speech, in: 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), 2018, pp. 513–518. doi:10.1109/ICISET.2018.8745608.

[9] H. Wittmann, Classification linguistique des langues signÉes non vocalement, Revue québécoise de linguistique théorique et appliquée 10 (1991) 215–288.

[10] J. Mahmood, Z. Tao, H. Md, A real-time computer vision-based static and dynamic hand gesture recognition system, Int. J. Image Graph. 14 (2014) 881–898. doi:10.1142/S0219467814500065.

[11] K. Hu, L. Yin, T. Wang, Temporal interframe pattern analysis for static and dynamic hand gesture recognition, in: 2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019, IEEE, Singapore, 2019, pp. 3422–3426. doi:10.1109/ICIP.2019.8803472.

[12] M. MUSTAFA, A study on arabic sign language recognition for differently abled using advanced machine learning classifiers, JOURNAL OF AMBIENT INTELLIGENCE AND HUMANIZED COMPUTING 2 (2021) 211–226. doi:HTTPS://DOI.ORG/10.1016/J.ESWA.2020.113794.