Translation of Mexican sign language into Spanish using machine learning and internet of things devices

Héctor Caballero-Hernández, Vianney Muñoz-Jiménez, Marco Antonio Ramos-Corchado
Department of Computer Science, Faculty of Engineering, Autonomous Mexico State University, Toluca, Mexico

ABSTRACT
This paper describes an application based on internet of things (IoT) and gesture recognition through the MediaPipe library to communicate people with hearing and speech disabilities who use Mexican sign language (MSL) with people who do not have a disability. The system is made up of two modules, on the one hand, there is an Apple watch with the ability to obtain text and voice inputs in Spanish to stand for the translation into sign language through video and images. The system can set up a connection with an application in charge of performing sign recognition by capturing images from a camera, the IoT user can connect so that communication between both modules is carried out bidirectionally through Firebase database. During the development of the experimental tests, the visual recognition module was able to recognize signs belonging to the Spanish alphabet, digits, and greetings, with a 96% of precision, while the IoT device correctly displayed voice translations and text to symbology compared to the MSL.

Keywords: Cloud computing, Internet of things, Language recognition, Language sign, Machine learning

This is an open access article under the CC BY-SA license.

Corresponding Author:
Héctor Caballero-Hernández
Department of Computer Science, Faculty of Engineering, Autonomous Mexico State University
Instituto Literario 100, Centro, 50000 Toluca, Mexico
Email: hcaballeroh045@alumno.uaemex.mx

1. INTRODUCTION
There are currently more than 450 million people worldwide with hearing problems, of which 25% are elderly. In general, this segment of the population has different hearing ranges [1]. Due to the communication problems experienced by deaf-mute people, they are often discriminated against in different sectors of society [2], almost 80% of deaf-mute people live in low and middle-income countries. One of the communication problems of deaf-mute people in Mexico is that [3] they are excluded from work activities, education, or social circles due to the difficulty they have in communicating, this situation is accentuated in rural places, generating elevated levels of stress [4], this problem is common in different Spanish-speaking countries of the American continent [5]. Mexico currently has 2.3 million people with hearing disabilities, of which only 40% of people with disabilities of productive age are economically active. This sector of the population has access to Mexican sign language (MSL) in some television programs [6], allowing them to access certain information content [7], but they have not aid on government websites for interpreting the MSL, as well as commercial tools to translate sign language, increasing the technological gap with the rest of the population.

Nowadays artificial intelligence provides a series of diverse tools to help solve problems such as hearing and speech disabilities, one of these tools is machine learning (ML), which extracts real-time information that generates predictive models and, where proper, acts accordingly [8]. One of the advances in ML is deep learning, which can process copious amounts of data and respond to very precise search
patterns [9]. On the other hand, convolutional neural networks (CNNs), whose logic is based on the way the human cortex works, through the approximation of continuous functions, which allows for a monitoring of a way [10]. ML algorithms require specific validation, some of the validation metrics as the precision measures the quality of the ML model in classification tasks (1), the recall supplies information about how much the ML model can find (2), F1 handles combining the metrics, precision, and lands on a single value, it compares the combined performance of precision and completeness versus different solutions (3), and the accuracy handles recording the percentage of cases that the model has gotten right (4).

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (1)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (2)
\]

\[
F_1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (3)
\]

\[
\text{Accuracy} = \frac{TN + TP}{TP + FP + TN + FN} \quad (4)
\]

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

When computational problems require message analysis, the natural language processing (NLP) tool is used, this branch of AI and linguistics uses automatic language to analyze the structure and meaning of text, which allows it to extract information from people, places, and events that take place in digital spaces where text sources are obtained [11]. The main area of study is the interactions using natural language between the language of people and machines, using lexical analyzers and grammatical resources such as morphology, syntactic structures, among others [12]. Due to the variety of tools developed for the interpretation of sign language, a process has been developed to select those relevant works that focus on computer vision and internet of things (IoT). Figure 1 presents the work selection diagram for subsequent work analysis.

Figure 1. Diagram process to select investigation to the state of the art

- Keyword generation: a set of keywords is generated to search for research work, such as: deaf and dumb people, IoT for deaf and dumb people, AI for MSL, gloves for deaf and dumb people.
- Databases chaise: keywords entered search engines such as WoS, Scopus, Latindex, Copernicus are considered to carry out the search for research papers, while those from book chapters and conferences are verified to be edited by Springer, IEEE, and Taylor & Francis.
- Research filtering: the research works are selected according to the technological contribution they propose, in addition to the fact that they must be the 70% of investigation 5 years old from their publication.
- Categories definition: research works selected are divided according to the methods and technologies used to solve the problem.
- Data analysis and proposal: research works are analyzed based on their proposal and the contribution they have regarding the state of the art.

Various investigations have addressed the problem of speech problems through neural network models help to recognize the activities of people with hearing and speech problems, as can be seen in [13]-[15], with you only look once (YOLO) being a recurring solution, although there is a large number of other works that are based on other types of proposals using neural networks with the use of TensorFlow [16], OpenCV [17], among others [18]. Trujillo-Romero and Bautista [19] presented a translator of MSL by recognizing these from the 3D trajectory of the movement of the hands of signers using a Kinect sensor. The authors used a corpus of 53 words, for noise elimination the addition of intermediate points was used and the KNN algorithm was used for the filtering of the general pattern, the words were classified by a multi-layer perceptron ANN. The method was confirmed by K-fold and cross validation and the average accuracy rate.
achieved by this implementation was 93.46%. In other investigations like Ranga et al. [20] a discrete wavelet transformed Gabor filter to extract features was employed. The classifiers like random forest, support vector machine, and K-nearest neighbors were evaluated to classify a dataset conformed by 24 classes of ASL alphabets with an accuracy of 95.8%, 94.3%, and 96.7%. CNN scored 97.01% on the signer-dependent dataset and 76.25% on the signer-independent dataset. Notable works such as Sosa-Jiménez et al. [21] use a Kinect evaluation to create a two-way medical care system with deaf and mute people, with an accuracy of 99%, while in Mejía-Peréz et al. [22] use a 3D camera to calculate the coordinates of the MLS signs to detect the posture of the hands, taking into account 30 signs, in the analysis of results they present a 97% accuracy.

The research of Rivera et al. [23] generated an assistant to deaf-mute people to send instructions and warnings in a manufacturing process, the authors used a CNN as classifiers in the assistant, a synthesized voice and text is employed. A mobile app called García et al. [24] is presented to help deaf children in the educational context and at the same time learning MSL. Jeyasheeli and Indumathi [25] generated a project with NLP to understand sentences, as well as the generated sentence is displayed in the android app. Works as Snehaa et al. [26] present a model feature that uses 3 modules, which are a sensing unit, a manipulation unit, and a voice storage unit, with 92% accuracy. The development of IoT devices is highly widespread for people with hearing impairments [27]-[31], highlighting the work of Varela-Santos et al. [30] with the implementation of a laptop computer with a camera to detect movements made by hands with marked gloves to facilitate the interpretation of MSL, these investigations have the advantage of open architectures that allows the inclusion of several types of hardware such as Arduino for the development of tools for deaf-mute people [32]-[34]. Raspberry is a widely used platform for the use of sign language [35], [36] due to its wide versatility.

Outstanding research such as Kumari et al. [37], developed a system based on Raspberry Pi, this proposed integrated wireless GSM card, Bluetooth, camera, and a vibrator. When an image is transferred to the device that helps to know who the person is who comes to a house, in addition, messages can be sent by GSM. Vasanth et al. [38] is performed voice transmission using a microphone, the information is sent to Google's API server, to converts the voice signal into text displayed on an LCD screen, the voice is amplified by a speaker, the voice signal is converted to text to be sent back to a Raspberry Pi. Lee and Lee [39] present an intelligent sign language interpretation system was proposed with a wearable device, which uses flexible, pressure sensors, and a 3-axis inertial motion sensor to recognize ASL alphabet. The information is classified by a support vector machine classifier and displayed in an Android mobile via Bluetooth. The recognition accuracy rate is 98.2%. Hasan [40] developed a deaf-mute electronic assistant system (DMEAS), this proposed use a Thalmic Labs Myo bracelet and a smartphone, the bracelet reads electromyographic signals from the forearm muscles via surface-mounted electrodes (sEMGs) and sends the data to the smartphone by Bluetooth to interpret the hand gestures.

As can be seen from the proposals analyzed, it is important to note the machine vision systems often use advanced motion recognition libraries, the neural networks and CNNs are highly widespread and the IoT devices involve the use of different sensors and artificial intelligence models. Some of the difficulties met in the implementation of sign language translation systems are that most people who do not have a hearing or speech impairment do not know sign language, therefore, it requires contact with a speech-to-text and speech-to-sign language synthesizer. Most of today's applications require a stable internet connection, so a corpus that considers common phrases and words is needed to make a fluent translation. Translation systems usually require gloves or systems that are mounted on the forearms to translate the movements or video-based systems that capture the movements to show the type of sign being represented, the latter being less invasive than the former. Based on the points discussed, a proposal for the translation of signs from MSL is presented in the following section.

2. METHOD

The proposal is based on the implementation of two modules for the facilitation of the interpretation of the MSL. The first module consists of an interpreter of movements with ML and captured by video camera for translation into text and voice on the signs that are made, the second module consists of a voice and text translator implemented in smart watch with image and video display functions that interpret the Spanish language to MSL. The method used for the proposal consists of 4 phases for the translation process of the MSL into Spanish. These phases are dataset formation, ML training, connecting interfaces, and sign recognition, which are presented in Figure 2.

- Dataset conformation: input classes are determined to classify the movements that the sign recognition system will be performing. Entries are short videos less than 3 seconds long.
- ML training: the classes found in the dataset are extracted to extract the body position and determine the body coordinates according to the class being analyzed.
- Interface connection: the connection interfaces are generated between a desktop device and the IoT device for data exchange between the voice and text process and the visual detection system.
- Sign recognition: video captured by the system's camera is compared with the system's training base to display the signal detected through representative images.

According to Figure 3, user 1 stands for a person who has full knowledge of the use of MSL and communicates with user 2 who has no or basic knowledge of MSL, the communication of the modules is through an internet socket for database for the passage of variables between both modules. The smart watch encompasses functions as speech-to-text translation, voice translation based on related images or videos for direct MSL interpreting and message retention for communication with people who are unaware of MSL. The sign detection module is based on capturing the movements of the person who needs to send a message in text or audio. The functions of the module as detection of movements made by the person using the MSL translation of MSL signs into text and speech and display of images standing for the message and intercom with IoT devices for translation of speech and text to contextual images and MSL. The MediaPipe library for the classification of signs uses references to detect different extremities of the body, the contours of the face, arms, thorax, legs, and feet can be visualized for the determination of poses. Figure 4 shows the motion detection diagram, in the first phase the input of the frames that are contained in digital videos is considered, then the detection of the hands is carried out to later generate the extraction of the landmarks, in the next phase a debugging of the data obtained is carried out to move on to the stage of training and data validation. The last phase consists of the data prediction process and the presentation of the information obtained.

Text inputs on the smart watch, as well as video inputs expressed in text after processing, go through an analysis applying NLP, for the purpose of language detection, as well as to obtain the behavior of users in their communication habits. The text inputs are analyzed to decide if it belongs to the Spanish language, to carry out the proper translation process, the message goes through a decomposition process, where the chains are transformed to decide if there are objects in the knowledge base and the message goes through a process to decide whether the content is positive, negative, or neutral. With the processing of the NLP, the elements that make up the input sentence are found, these are named to form a flexible system that allows to evolve with respect to the needs of the user, the NLP shows lexical elements such as pronouns,
auxiliary, verbs, adverbs, numbers, articles, among others, as well as the identification of the language. Figure 5 shows the concept of the evolutionary matrix, which increases its size according to the syntactic categories added, increasing its size in both rows and columns automatically, the first elements of each module are organized according to their corresponding category as assigned by the NLP analyzer, the intersection between each category stands for a relationship between each of them and is related to the times have been associated, to verify which are the most used syntactic categories. In the data flow diagram represented in Figure 6, it is shown that the user when executing the data reception process, the classification of the input is activated, to result in the analysis of the data being activated, to give as a result if it is necessary to translate the signs to text, or speech to text to begin with the analysis, in the analysis the derivations of the words are generated in order to develop the construction of the evolutionary matrix. In case a syntactic category is not included, it is added dynamically, and finally the user sees the output classification.

**Figure 4. Visual recognition module diagram**

**Figure 5. Learning matrix**

<table>
<thead>
<tr>
<th>Category</th>
<th>Alphabet</th>
<th>Digits</th>
<th>Personal pronouns</th>
<th>Verbs</th>
<th>Adverbs</th>
<th>Adjectives</th>
<th>Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphabet</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Digits</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Personal pronouns</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Verbs</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Adverbs</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Adjectives</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Objects</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

**Figure 6. Process of the evolutionary matrix of the system**
3. RESULTS AND DISCUSSION

The proposal has been coded in various languages, the sign recognition module has been developed in Python 3.11 implementing the MediaPipe library, while the software loaded on the smart watch has been made using swift 5, in WatchOS 8.8.1 linked to an iPhone 8 with iOS 16.1.1, database used is Firebase. The training of the sign recognition module classes is trained by 2000 videos in MP4 format with a resolution of 900x600 pixels. These videos were obtained from a group made up of 45 people who cooperated with their generation, while the rest of the videos were obtained from websites. The two modules hold a corpus composed of 800 elements based on the display of images and videos to stand for words, phrases, letters, and digits, the corpus of video sign recognition is limited to 44 classes, which include digits, letters of the Spanish alphabet, greetings, and an object and the visual recognition is performed in an environment with constant lighting.

The software has been tested in two scenarios, in the first scenario the operation of the artificial vision module is verified with different variations of the light source, while the second scenario combines visual recognition with the smart watch to communicate to a deaf-mute person with a person without limitations. Artificial vision module has been configured for a personal computer with a 5 megapixels video camera, and the IoT device is used with another person without hearing or speech limitations. To classify the input signs MSL was used a recognition algorithm evaluation the metrics F1, precision, recall, and accuracy using the random forest [41] (5), ridge [42] (6), logistic regression [43] (7), and SVC [44] (8) and (9) algorithms to obtain the best performance, the equations of the classification algorithms are presented below.

$$\rho \sigma^2 + \frac{1-\rho}{B_\tau} \sigma^2$$  \hspace{1cm} (5)

Where \(\rho\) is pairwise correlation, \(B\) are random variables, and \(\sigma^2\) the variance.

$$\text{minimize} = \|y - Xw^2 + \alpha \sigma^2\|$$  \hspace{1cm} (6)

Where \(y\) is the vector of the values of the response variables, \(X\) is a matrix of predictor variables, \(w\) is a vector of coefficient estimates and \(\alpha\) represents the regularization parameter that controls the strength of the penalty term.

$$p = \frac{1}{1 + e^{-(b_0 + b_1x)}}$$  \hspace{1cm} (7)

Where \(p\) is a probability of 0 to 1, \(b_0\) is a constant that moves the curve left and right and \(b_1\) defines the slope of the curve.

$$\min(w, b, d) = \frac{1}{2} w^T w + C \sum_i d_i$$  \hspace{1cm} (8)

$$y_i(w^T \phi(x_i) + b) > 1 - d_i$$  \hspace{1cm} (9)

Where \(d_i\) forms distances with \(d_i = 0, 1 = 1, \ldots, n\), \(C\) is the regularization of the parameter, \(w^T w\) denotes a normal vector \(\|w\|\), \(\phi(x_i)\) is the input spatial vector, \(b\) is a bias parameter, \(y_i\) is the i-th objective value.

According to the results obtained in the evaluation of the algorithms F1 Figure 7(a), accuracy Figure 7(b), recall Figure 7(c), and precision Figure 7(d) which can be seen in Figure 7, it was obtained that random forest obtained an overall performance of 0.96 points in all the evaluations, this algorithm is the one that has been chosen as a reference for the evaluation of sign detection. The experimental results show different behaviors of the software analyzed, which are broken down below. Visual recognition module was able to obtain the information of the signs that were made at the time of capture by the video camera. Figures 8(a) to (f) show the signs of u, a, 1, 5, e, good morning, and house, respectively. The scenarios shown in Figure 6 were taken in two different lighting scenarios to check the effectiveness of the recognition model, considering when the user presents his face, as well as when the user does not present his face, in this way it is possible ensure the model's ability to recognize the signs made with the users' hands.

In the second scenario, the communication of the sign translation module and the communication through the smart watch are verified. On the one hand, the computer vision module is responsible for interpreting the signals made by the user, while the smart watch is responsible for interpreting for the other user the text or voice to images and videos representative of the MSL. For this process, an Internet connection is needed for communication between both modules. Figure 9(a) shows the greeting "buenas tardes" as well as the response of the smart watch user responding, "buenas tardes", while in sections
Figures 9(b)-(d) of Figure 9 the menu of the smart watch is shown. Smart watch app, the display of the message obtained from the computer vision module and the display of the message represented by an image and a video of the MSL sign.

Figure 7. Evaluation of sign detection algorithms; (a) F1 evaluation, (b) accuracy evaluation, (c) recall evaluation, and (d) precision evaluation

Figure 8. Extract of signs recognized during the experimental process; (a) sign letter “u”, (b) sign letter “a”, (c) sign number “1”, (d) sign number “5”, (e) sign “good morning”, and (f) sign “house”
Figure 9. Sending data from the smartwatch app; (a) communication module connected to smart watch by Firebase, (b) text and voice capture menu of the APP, (c) data display on smart watch “good afternoon”, and (d) image and video display

It should be noted that lighting conditions are important, because a substantial reduction in lighting has an impact on the interpretation of signs, although it is important to keep in mind that the tests were carried out with variations that would allow for similarity in an everyday environment of use within of a home or a place with clear backgrounds to determine how effective and practical the model presented is. Table 1 presents a summary of important works that have addressed the interpretation of the MSL, among the works is that of [19], [21], which highlight the use of Kinect sensors and RGB-3D cameras. To monitor the positions of the signs made by deaf-mute people, in comparison with our proposal the bidirectional communication scheme is contemplated, while in [19] it is not considered. It should be noted that the proposal of [21] presents an accuracy of 99%, in addition to allowing bidirectional communication, but is limited to an environment for medical consultations, on the other hand, tests that were carried out have been in home, while the previous proposals have been developed in a laboratory, which allows us greater certainty that the operation will be more appropriate in everyday environments. Considering the works of [22], [30], it is possible to observe that bidirectional communication has not been contemplated, in addition to the fact that in [22] the proposal for the use of gloves is conditioned to the amount of light that is reflected in the surface of the gloves for correct interpretation of signs. It is important to note that the previously mentioned proposals have emphasized only the recognition of signs but have not included in them an analysis of NLP treatment as has been addressed in this proposal. This point is important to consider, because it allows you to simplify the analysis of sign language information that is processed by the system for subsequent use in tasks such as translation into other languages, as well as detection of emotional aspects.

Table 1. Comparative table of results

<table>
<thead>
<tr>
<th>Authors</th>
<th>Interfaces</th>
<th>Methods</th>
<th>Signs</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sosa-Jiménez et al. [21]</td>
<td>Sensor MS Kinect</td>
<td>Hidden Markov models</td>
<td>82</td>
<td>99</td>
</tr>
<tr>
<td>Mejía-Peréz et al. [22]</td>
<td>RGB depth camera (OAK-D)</td>
<td>Recurrent neural networks, GRU, and LSTM</td>
<td>30</td>
<td>97</td>
</tr>
<tr>
<td>Varela-Santos et al. [30]</td>
<td>RGB digital camera and gloves</td>
<td>Neural Networks</td>
<td>20</td>
<td>88</td>
</tr>
<tr>
<td>Proposal</td>
<td>Digital camera and smartwatch</td>
<td>ML through random forest</td>
<td>42</td>
<td>96</td>
</tr>
</tbody>
</table>

Some of the most important limitations that were found during the development of the proposal is the time it takes to collect data to train the ML model, due to the non-existent standardization of a database that serves as a basis for the development of models’ recognition of the MSL, this problem is also observed in the research previously mentioned in Table 1. One of the main advantages of the proposal is that it can be tested in environments with few restrictions in terms of development space, as well as the use of conventional cameras, in addition to preventing two-way communication. On the other hand, it is possible to make an association with several IoT devices, which can exchange information with the online database platform in real time. The implementation of NLP has made it possible to identify the category of input words, both in the MSL sign interpretation module and in the smart watch, generating a structure that allows relating the most used words during the conversation, as well as validating that the language in which the smart watch information is transmitted is Spanish. Other important contribution of this proposal is a flexible system that allows people with hearing and speech disabilities to communicate with people who do not have any disability with non-invasive devices, with the ability to increase their knowledge base without modifying its structure managed to develop the first database designed to train MSL-oriented computer vision problems.
In future iterations, this implementation will detect other input languages on the smart watch to be able to perform translations to Spanish and subsequently to MSL. Another important line that has been considered in this research is to be able to develop software modules that allow deaf-mute people to communicate with other people without disabilities through Web platforms, in addition to increasing the amount of training data to increase the precision of the detection and size of the corpus, allowing it to increase its functionality for real and non-experimental conversations. Considering the results found in this research, it is important to mention that another of the lines that this research will take is to expand the database built, as well as seek its standardization to improve the quality of future research related to the LSM.

4. CONCLUSION

Classification algorithms for posture detection, such as random forest, are efficient when applied in sign language classification tasks. In this research, a performance of 96% was obtained in the classification process of the signs expressed for the MSL; on the other hand, the introduction of IoT devices such as smart watches facilitate the communication of users with speech and hearing disabilities with people without disabilities due to their flexibility to integrate into mixed environments. The presented application can perform voice-to-text and sign language translation, as well as translate signals to text and audio. Information processing using NLP has made it possible to generate a classification of the input text, as well as the validation of the Spanish language to avoid errors in the input data of the smart watch. One of the limitations with the greatest impact on this project was obtaining a database available for training the artificial vision model, because the information is usually available on social networks, and is not developed to be used in computer systems which reflects, on the one hand, the precarious attention of the MSL, as well as the lack of government programs to develop platforms that allow its teaching and dissemination to the target population.

ACKNOWLEDGEMENTS

We thank CONACYT for the financial support granted to carry out this research project with the series 11200/320/2022.

REFERENCES


Translation of Mexican sign language into Spanish using machine ... (Héctor Caballero-Hernández)
Translation of Mexican sign language into Spanish using machine translation ... (Héctor Caballero-Hernández)