Real-Time Mexican Sign Language Interpretation Using CNN and HMM

Jairo Enrique Ramírez Sánchez
Tecnológico de Monterrey
Monterrey, Nuevo León, México
A01750443@tec.mx

Arely Anguiano Rodríguez
Tecnológico de Monterrey
Monterrey, Nuevo León, México
A01752068@tec.mx

Miguel González Mendoza
Tecnológico de Monterrey
Atizapán, Estado de México, México
mgonza@tec.mx

1. Introduction

In order to achieve effective and efficient communication between a PHI and a Hearing Person (HP), it is not enough to learn to identify the hand movements to designate each of the words, since manual features (MF) represent only 20% of the total communication, while non-manuals features (NMF) figure as the rest [1]. Therefore, it is also necessary to have knowledge of its grammar and non-manual features, in particular, facial features, since these indicate the verb tense. The **past tense** is expressed through the lower lip protrudes from the upper one nodding softly with the head. On the other hand, in the **present tense**, a neutral expression is projected. Finally, the **future tense** is manifested through an expression of doubt or thought. Regarding MSL grammar, the verbs *'ser'* and *'estar'* are omitted, so only the noun and adjective are expressed or, failing that, noun and place. As can be seen, for a person not immersed in an environment and without previous contact with the deaf community, it is extremely difficult to reach a full understanding.

In our research, we will address the analysis and recognition of signs in real time using convolutional neural networks for sign classification three aspects are taken into account: facial expression, hand movements and body position. Finally, interpretation is enhanced by a Hidden Markov Model for context and grammar enrichment.

2. Proposal

The processing model presented is divided into 5 stages as shown in the figure 1.

The capture uses the open source library *MediaPipe* presented in [4]. This library allows to build perception flows with *OpenCV* visualization tools [5]. Both allow the detection of marks on hands, body and face in real time with a wide generalization capability regardless of skin color, hand size and height of the person. Thus, the capture system blends the benefits of camera inputs (economic feasibility) and external sensors (generalization) in a non-intrusive approach.

For standardization, the coordinates of the marks are extracted selecting 15 frames per second. The coordinates of the facial expression with dimension 234 x 2, the hands 21 x 4 and the body position 33 x 2.

Subsequently, the matrices are independently encoded with a neural network with three layers of 2D convolution each one followed by MaxPooling. This performs the identification of general patterns, allowing similar signs to be encoded in an analogous way. Despite the initial dimensions of the three matrices, each CNN reduces it to a Flatten vector with 128 entries in order to be proper merged in the next stage.

The encodings are integrated by an addition layer. Hands encoding provide meaning, body position spatial location and the face the verbal tense. In this last part of the architecture a multilayer perceptron is used to perform the classification. The last layer of the network uses a sigmoid function as activation, which assigns an observation probability vector $\vec{O}_i$ whose entry O_i^j refers to the word j in the dictionary of length N for each time step i.

Finally, a HMM is used to perform the classification based on the previous context. A transition matrix T between words of the dataset calculated with 100 common sentences in MSL extracted from [6]. With $\vec{O}_i$ and T the selection of the most probable word with a morphosyntactic meaning is performed, the set of possible states for each

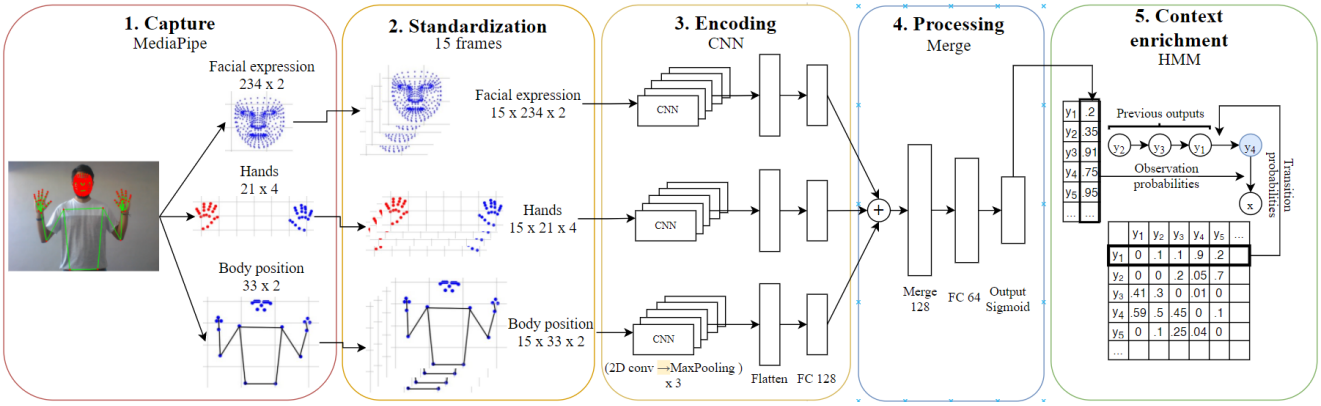

Figure 1. Detection stages.

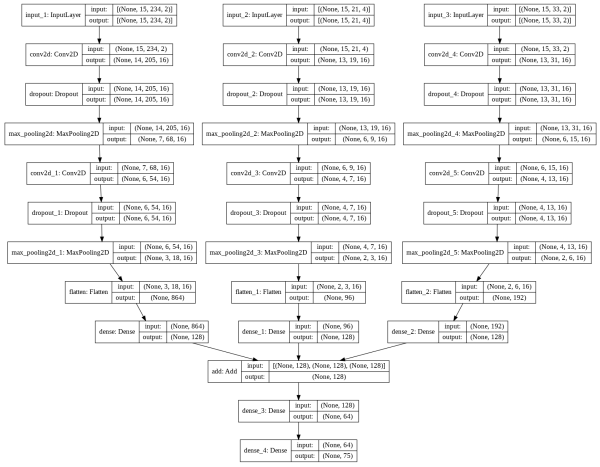

Figure 2. Configuration for CNN and Multilayer perceptron (stages three and four).

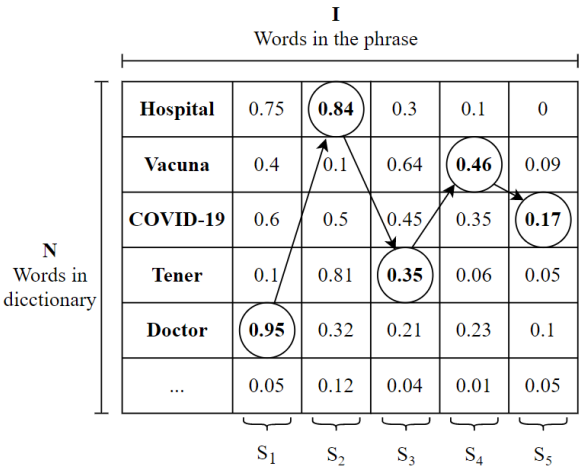

Figure 3. Example of the Viterbi algorithm applied to the determination of the most probable word sequence according to MSL grammar, in this case: *'Doctor hospital tener vacuna COVID-19'*.

time step is represented by $S = (S_1, S_2, ..., S_I)$, with the value S_i being the most probable word. Probability is calculated recursively as shown in the following equation:

$$p(O^j, S^j) = \prod_{i=1}^{I} p(O_i^j|S_i^j)p(S_i^j|S_{i-1}^j) \; \forall j \in 1, ..., N \quad (1)$$

Where:

$$I = \text{Total time steps}$$
$$p(O_i^j|S_i^j) = \text{Observation probability}$$
$$p(S_i^j|S_{i-1}^j) = \text{Transition probability}$$

Finally, a manual implementation of the *Viterbi* [2] algorithm is run to track the highest probability path. An example of this process is shown in figure 3. This implementation reduces the interpretation error by **11.7%** in contrast to isolated word identification.

3. Dataset

The dataset used was generated with the help of six volunteers using *MediaPipe* and *OpenCV* libraries.

The capture for the hand coordinates consists of 15 frames containing the movement of a matrix of 21 points for each hand (15, 21, 4), 33 for the body (15, 33, 2) and 234 points for the face (15, 234, 2). An average of 35 samples was made for one of the 49 signs (13 verbs and 36 words) and 15 for each facial expression for the three verb tenses used (past, present and future). In total, the classification consisted of 13x3 + 36 = 75 classes.

Due to the particular spatial meaning, each word has a different temporal duration. To make the dataset uniform,

$n = 15$ frames per second were selected for each sample. Keval H. et al. showed in [3] that the minimum number of frames per second to identify a human action is 8.

3.1. Results Experiment 1: Focus on Isolated Words

Each word was signed 10 times by each of the six volunteers, measuring performance in a binary manner, in other words, if it was predicted correctly it scored 100%, otherwise 0%.

For the case of verbs, the way of measuring performance was modified to obtain a more representative metric. Therefore, we assigned a 100% if the word and the predicted grammatical tense matched with the expected one. 70% if the word matched, but not the tense. 30% if the tense was correct, but not the word. 0% if the prediction was completely wrong. An example is shown in table **??**.

An average accuracy of 94.9% with $\sigma = 0.07$ was obtained for the nine categories. HMM was not used for this experiment. The results are shown in table 1.

Table 1. Results experiment 1.

Category	Average accuracy	σ
Basic	0.987	0.03
People	0.943	0.1
Places	0.937	0.05
Mood	0.99	0.01
Present tense verb	0.91	0.09
Past tense verb	0.94	0.06
Future tense verb	0.89	0.11
Question	0.962	0.07
Emergency	0.981	0.04

3.2. Results Experiment 2: Focus on Sentences

Participants performed sign and facial expression of 20 sentences taking five samples. Accuracy was used as a performance metric. The results are shown in https://drive.google.com/file/d/1z5nQYcMjgQseDbxXU6FO07QxAfmydyGL/view?usp=sharing.

An average accuracy of 94.1% with $\sigma = 0.09$ was obtained for the 20 sentences. In contrast, the interpretation with isolated words (without context enrichment) reached an accuracy of 82.4% with $\sigma = 0.12$.

4. Conclusions

As has been discussed throughout this paper, real-time interpretation of sign languages is a complex process involving several factors such as body position, facial expression, hand movements and the specific context. Our study lays the groundwork for a first approach to the creation of a complete sign language to spoken language interpreter, being the only work in Spanish so far that considers facial expression as an indicator of grammatical tense. Moreover, in terms of feasibility, our model demonstrated that even with relatively few samples (on average, 35 for each sign and a knowledge base of 100 sentences) it is possible to obtain an acceptable performance. Including an accuracy of 94.9% for recognition of 75 isolated words and 94.1% for 20 test sentences in the medical context. Consequently, validating both the wide generalization capacity of the CNN coding architecture and the context-identifying HMM. This positions our work as an economically viable - since it only uses a computer webcam -, easy to implement and fully scalable to other sign languages option. Especially those corresponding to countries with little or no study in the field, improving the inclusion of millions of hearing impaired people.

As future work, we intend to design, implement and include a statistical translation machine that will allow us to move from MSL grammatical structure to Spanish, resulting in a significant contribution to further enhance effective communication tools.

References

[1] Miroslava Cruz. *Gramática de la Lengua de Señas Mexicana.* Centro de Estudios Lingüísticos y Literarios, Colegio de México., 1 edition, 2008. 1

[2] G. David Forney. The Viterbi Algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973. 2

[3] Hina Keval and M Angela Sasse. To catch a thief - You need at least 8 frames per second: The impact of frame rates on user performance in a CCTV detection task. In *MM'08 - Proceedings of the 2008 ACM International Conference on Multimedia, with co-located Symposium and Workshops*, pages 941–944, 2008. 3

[4] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo Ling Chang, Ming Guang Yong, Juhyun Lee, Wan Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A framework for building perception pipelines, 2019. 1

[5] M Naveenkumar and V Ayyasamy. OpenCV for Computer Vision Applications. *Proceedings of National Conference on Big Data and Cloud Computing (NCBDC'15)*, (March 2015):52–56, 2016. 1

[6] Maria Serafîn and Raúl González. *Diccionario de LEngua de Señas Mexicana*, volume 38. México D.F., 2011. 1