Article

Towards a Recognition System for the Mexican Sign Language: Arm Movement Detection

Gabriela Hilario-Acuapan , Keny Ordaz-Hernández * , Mario Castelán , Ismael Lopez-Juarez *

Posted Date: 14 April 2025

doi: 10.20944/preprints202504.0981.v1

Keywords: mexican sign language; dynamic signs; pattern analysis; pose-based approach; pose estimation; computer vision; machine learning; cnn; yolov8; arm movement

Article

Towards a Recognition System for the Mexican Sign Language: Arm Movement Detection

Gabriela Hilario-Acuapan, Keny Ordaz-Hernández *, Mario Castelán and Ismael Lopez-Juarez *

Robotics and Advanced Manufacturing Department, Centre for Research and Advanced Studies (CINVESTAV), Ramos Arizpe 25900, Mexico

* Correspondence: keny.ordaz@cinvestav.edu.mx (K.O.-H.); ismael.lopez@cinvestav.edu.mx (I.L-J.)

Abstract: This paper describes ongoing work in the creation of a recognition system for the Mexican Sign Language (LSM). We propose a general sign decomposition into three parts: hand configuration (HC), arm movement (AM) and non-hand gestures (NHG). This paper focuses on the AM features and reports the approach created to analyze visual patterns in arm joint movements (wrists, shoulders and elbows). For this research, a proprietary dataset —that do not limit the recognition of arm movements— was developed, with active participation from the deaf community and LSM experts. We conduct analysis on two case studies of three sign subsets. For each sign, the pose was extracted to generate shapes of the joint paths during the arm movements and feeded to a CNN classifier. YOLOv8 was used for pose estimation and visual patterns classification purposes. The proposed approach, based on pose estimation, shows promising results for constructing CNN models to classify a wide range of signs.

Keywords: mexican sign language; dynamic signs; pattern analysis; pose-based approach; pose estimation; computer vision; machine learning; cnn; yolov8; arm movement

1. Introduction

Deafness or hearing loss is the partial or total loss of the ability to hear sounds in one or both ears. The World Health Organization's most recent World Hearing Report [1] estimates that more than 1.5 billion people have some degree of hearing loss. Approximately 430 million people have moderate or greater hearing loss in the better ear, and it is expected to increase to 700 million people by 2050.

According to the Ministry of Health [2], approximately 2.3 million people in Mexico have hearing disabilities. This vulnerable group faces significant levels of discrimination and limited employment opportunities. Additionally, this health condition restricts access to education, healthcare, and legal services, further exacerbating social inequalities and limiting opportunities for integration. One of the primary challenges faced by the deaf community is communication with hearing individuals, as linguistic differences hinder social and workplace interactions. While technology has proven useful in reducing some of these barriers, deaf individuals often rely on the same technological tools as the hearing population, such as email and text messaging applications. However, these tools are not always effective, as not all deaf individuals are proficient in written Spanish.

In the Americas, the most widely studied sign languages are the American Sign Language (ASL) and the Brazilian Sign Language (LIBRAS), which have facilitated research and technological advancements aimed at improving communication with the deaf community. An example of such innovation is SLAIT [3], a startup that emerged from a research project at Aachen University of Applied Sciences in Germany. During this research, an ASL recognition engine was developed using MediaPipe and recurrent neural networks (RNNs). Similarly, [4] announced an innovative project in Brazil that uses computer vision and artificial intelligence to translate from LIBRAS to text and speech in real time. Although this technology is still undergoing internal testing, the developers claim that after four years of work, the system has reached a significant level of maturity. This technology was developed by Lenovo researchers in collaboration with the Center for Advanced Studies and Systems in Recife (CESAR), which has already patented part of this technology [5]. The system is capable of recognizing the positions of arm joints, fingers, and specific points on the face, similar to SLAIT. From this data, it processes facial movements and gestures to identify sentence flow and convert it from sign language into text. CESAR and Lenovo consider that their system has the potential to become a universally applicable tool.

Compared to speech recognition and text translation systems, applications dedicated to sign language (SL) translation remain scarce. This is partly due to the relatively new nature of the field and the inherent complexity of sign language recognition (SLR); which involves visual, spatial, and gestural elements. Recognizing sign language presents a significant challenge, primarily due to limited research and funding. This highlights the importance of promoting research in the development of digital solutions that improve the quality of life of the deaf community (c.f. [6]). However, researchers agree that the key factor for developing successful machine learning models is data (c.f. [7]). In this regard, for SLs as the LSM, existing databases are often inadequate in terms of both size and quality, which hinders the advancement of these technologies. Also, the sensing technology has a fundamental role, in the reliability of the incoming data. This is the main reason that SLR is broadly divided into two branches: contact sensing and contactless sensing.

Sign data acquisition with contact depend on gloves [8], armbands [9], wearable inertial sensors [10,11], or Electromyographic (EMG) Signals [12]. In contrast, *contactless sign data acquisition* is mainly divided into two types, depending on the kind of hardware: simple hardware (color or infrared cameras) vs specialized hardware (depth sensors, optical 3D sensors [13], commercial WiFi devices [14], ultrasonic devices [15]).

This classification is similar to the one presented by ([16], Fig. 1), except that their sign data acquisition approaches are divided into sensor-based approaches and vision-based approaches. We present several examples of sign language research and related work, along with various approaches to sign data acquisition, as detailed in Table 1.

Table 1. Sign Language research and related work.

Ref.	SL	Sign group*	Sign type	Sign features†	Sensor/Tool
Chiradeja et al. (2025) [8]	-	S	Dynamic	HC	Gloves
Rodríguez-Tapia et al. (2019) [10]	ASL	W	Dynamic	HC	Myoelectrical bracelets
Filipowska et al. (2024) [12]	PSL	W	Dynamic	HC	EMG
Umut and Kumdereli (2024) [9]	TSL	W	Dynamic	HC, AM	Myo armbands (IMU + sEMG)
Gu et al. (2024) [11]	ASL	W, S	Dynamic	HC, AM	IMUs
Urrea et al. (2023) [17]	ASL	L, W	Static	HC	Camera/MediaPipe
Al-Saidi et al. (2024) [16]	ArSL	L	Static	HC	Camera/MediaPipe
Niu (2025) [18]	ASL	L	Static	HC	Camera
Hao et al. (2020) [14]	-	W	Dynamic	HC	WiFi
Galván-Ruiz et al. (2023) [13]	LSE	W	Dynamic	HC	Leap motion
Wang et al. (2023) [15]	CSL	W, P	Dynamic	HC	Ultrasonic
Raihan et al. (2024) [19]	BdSL	L, N, W, P	Dynamic	HC	Kinect
Woods and Rana (2023) [20]	ASL	W	Dynamic	AM, NHG	Camera/OpenPose
Eunice et al. (2023) [21]	ASL	W	Dynamic	HC, AM, NHG	Camera/Sign2Pose, YOLOv3
Gao et al. (2024) [22]	ASL, TSL	W	Dynamic	HC, AM, NHG	Camera, Kinect
Kim and Baek (2023) [23]	DGS, KSL	W, S	Dynamic	HC, AM, NHG	Camera/AlphaPose
Current study	LSM	W, P	Dynamic	AM	Camera/YOLOv8

* L: alphabet letters; N: numbers; W: words; P: phrases; S: sentences. †HC: hand configurations; AM: arm movement; NHG: non-hand gestures. SLs names are provided in the Abbreviations section. Top part: sign data acquisition with contact sensing. Bottom part: Contactless sign data acquisition.

In Table 1, we have included information regarding the features of signs that are included in the sign data acquisition, for each reported work. Instead of using the separation employed by [22] (facial, body and hand features), we propose our own decomposition into hand configurations (HC), arm movement (AM) and non-hand gestures (NHG), see Figure 1. This is a fundamental concept of our research, so this decomposition is discussed in more detail in Section 1.1.2. The facial, body and hand features separation is a concept commonly seen in pose estimators —such as MediaPipe [24]– that are also common in SL research, as presented in Table 1. It is also possible to observe, that most SL research is focused on the HC features.

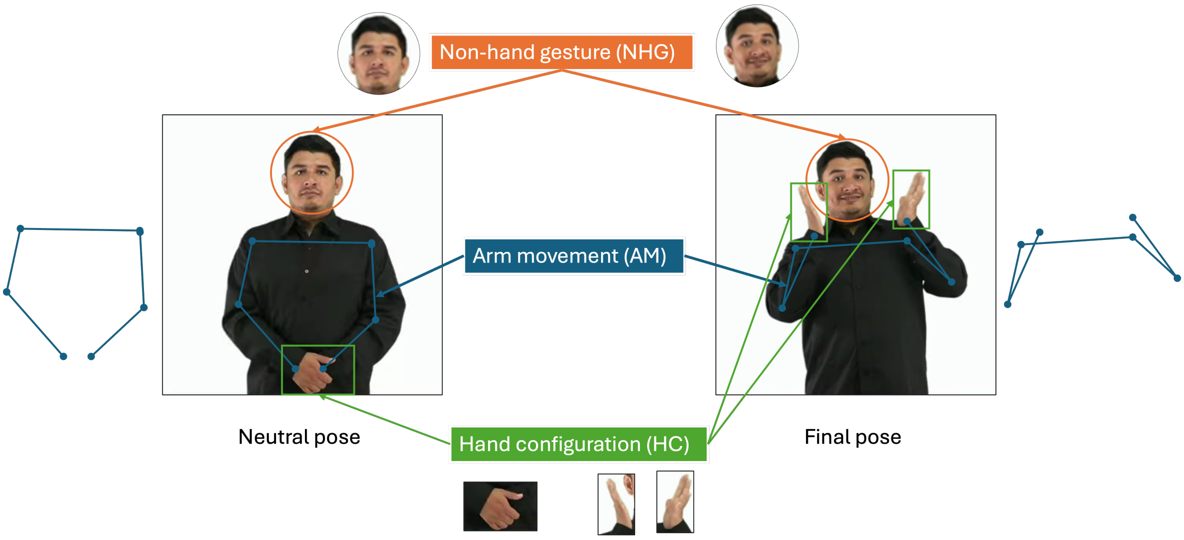

Figure 1. Sign Features: hand configurations (HC), arm movement (AM) and non-hand gestures (NHG). "Surprise!" sign images were taken from screenshots of the corresponding YouTube video of the GDLSM [25], see Appendix A.

We will now present the scientific context of the LSM research. First, we present the known datasets and then studies about LSM recognition and analysis.

The LSM is composed by two parts: dactylology (fingerspelling) and ideograms ([26], p.12). Dactylology is a small subset of the LSM and basically consists of letters of the alphabet, where the most part are static signs. A few signs for numbers are also static. Due to the small, nevertheless important, role of dactylology, we are interested in LSM ideograms datasets. To the best of our knowledge, there are three public available ideogram-focused datasets. Two of them are visual: (i) the MX-ITESO-100 preview [27] with videoclips of 11 signs from 3 signers (out of 100 signs, but not all are currently available), and (ii) the Mexican sign language dataset [28,29] with image sequences of 249 signs from 11 signers. The third dataset, consisting in keypoints, is provided by [30]; this dataset has 3000 samples of 30 signs from 4 signers (8 letters, 20 words and 2 phrases). This was constructed by processing the RGBD data into keypoints by means of MediaPipe [24] tool, but the unprocessed visual data is not provided. A comparison of these datasets, along with LSM glossaries are provided in Tables 2 and 3.

Table 2. LSM Datasets and Glossaries.

Ref.	Type	Sign group*	Sign Signal	Samples
DIELSEME 1 (2004) [31]	Glossary†	535 W	Visual	1 video per sign
DIELSEME 2 (2009) [32]	Glossary†	285 W	Visual	1 video per sign
GDLSM (2024) [25]	Glossary	27 L, 49 N, 667 W, 4 P	Visual	1 video per sign‡
MX-ITESO-100 (2023) [27]	Dataset	96 W, 4 P	Visual	50 videos per sign
Mexican sign language dataset (2024) [29]	Dataset	243 W, 6 P	Visual	11 image sequences per sign
Mexican Sign Language Recognition (2022) [30]	Dataset	8 L, 20 W, 2 P	Keypoints	100 samples per sign

* L: alphabet letters; N: numbers; W: words; P: phrases. † According to [33], DIELSEME 1 and 2 are actually glossaries and not dictionaries. The three LSM glossaries only have one sample by sign, while datasets have multiple samples by sign. ‡ Their site reports 719 videos, but only 715 were found; also, the 32 videos in the "Estados y capitales" thematic category include two signs per video.

Table 3. LSM Datasets and Glossaries: Sign and signal properties.

Ref.	Sign Features	Signal Properties	File Format	Comments
DIELSEME 1 (2004) [31]	HC, AM*, NHG	320×234 @ 12 fps	SWF videos	
DIELSEME 2 (2009) [32]	HC, AM, NHG	720×405 @ 30 fps	FLV videos	
GDLSM (2024) [25]	HC, AM, NHG	1920×1080 @ 60 fps	videos	Hosted on a streaming platform; c.f. Appendix A
MX-ITESO-100 (2023) [27]	HC, AM, NHG	512×512 @ 30 fps	MP4 videos	Preview only‡
Mexican sign language dataset (2024) [29]	HC, AM*	640×480	JPEG images	Blurred faces
Mexican Sign Language Recognition (2022) [30]	HC, AM, NHG	20×201 array	CSV files	One row per frame, 67 (x, y, z) keypoints

* In those cases, the background and clothing are black, so segmentation of skin (hand and face) are easier, but tracking joints for AM is more difficult. ‡ Only 11 signs (words) are available in the public preview. Also, the 50 samples of every sign are done by a single subject.

Regarding LSM studies, most of the SLR research of the LSM focuses mainly on classifying static letters and numbers using classical machine learning techniques and convolutional neural networks (CNN) [34–41]. Using the classification provided by [16], there are four classes of signs: (i) continuous signs, (ii) isolated signs, (iii) letter signs and (iv) number signs. In the LSM, most of the signs in the three last categories are static signs. But signing in the LSM is generally highly dynamic and continuous, since most signs are ideograms, as mentioned before.

In terms of dynamic sign recognition, early studies focused on classifying letters and numbers with motion. For example, [42] used the CamShift algorithm to track the hand trajectory, generating a bitmap that captures the pixels of the hand path, these bitmaps are then classified using a CNN. Another approach, presented in [43], is to obtain coordinates (x,y) of 22 key points of the hand using Intel RealSense sensor, which are used as training data for a multilayer perceptron (MLP) neural network. Finally, in [44], 3D body cue points obtained with MediaPipe are used to train two recurrent neural network models (RNN): LSTM and GRU.

In more recent research, in addition to letters and numbers, some simple words and phrases have been included. Studies such as that of [45–47], continue to use MLP-type neural networks. While others, such as [30], use more advanced RNN models. In the case of [27], CNNs are used to extract features from the frames of a series of videos, which will be the input data of an LSTM model.

On the other hand, the work of [48] presents a method for the classification of dynamic signs, which involves the extraction of a sequence of frames, which go through a segmentation process using neural networks based on color, resulting on the skin of hands and face. To classify the signs, four classical machine learning algorithms are compared: bayesian classifier, decision trees, SVM and NN.

Although research on LSM recognition has been conducted for several years, progress in this area has been slow and limited compared to other SLs. A common approach is to use computer vision techniques such as CNNs to build automatic sign recognition systems. However, with the recent emergence of pose recognition models such as MediaPipe and YOLOv8, there is a trend in both LSM and other sign languages to use these tools to train more complex models such as RNNs or more sophisticated architectures such as Transformers. A comparison of the studies mentioned here, with additional details, is shown in Table 4.

Table 4. LSM research.

Ref.	Sign group*	Sign type	Sign feature	Sensor/Tool
Solís et al. (2016) [34]	L	Static	HC	Camera
Carmona-Arroyo et al. (2021) [35]	L	Static	HC	Leap Motion, Kinect
Salinas-Medina and Neme-Castillo (2021) [36]	L	Static	HC	Camera
Rios-Figueroa et al. (2022) [37]	L	Static	HC	Kinect
Morfín-Chávez et al. (2023) [38]	L	Static	HC	Camera/MediaPipe
Sánchez-Vicinaiz et al. (2024) [39]	L	Static	HC	Camera/MediaPipe
García-Gil et al. (2024) [40]	L	Static	HC	Camera/MediaPipe
Jimenez et al. (2017) [41]	L, N	Static	HC	Kinect
Martínez-Gutiérrez et al. (2019) [43]	L	Both	HC	RealSense f200
Rodriguez et al. (2023) [44]	L, N	Both	HC	Camera/MediaPipe
Rodriguez et al. (2025) [49]	L, N	Both	HC	Camera/MediaPipe
Martinez-Seis et al. (2019) [42]	L	Both	AM	Camera
Mejía-Peréz et al. (2022) [30]	L, W	Both	HC, AM, NHG	OAK-D/MediaPipe
Sosa-Jiménez et al. (2022) [50]	L, N, W	Both	HC, body but not NHG	Kinect
Sosa-Jiménez et al. (2017) [45]	W, P	Dynamic	HC, AM	Kinect/Pose extraction
Varela-Santos et al. (2021) [51]	W	Dynamic	HC	Gloves
Espejel-Cabrera et al. (2021) [48]	W, P	Dynamic	HC	Camera
García-Bautista et al. (2017) [46]	W	Dynamic	AM	Kinect
Martínez-Guevara and Curiel (2024) [52]	W, P	Dynamic	AM	Camera/OpenPose
Martínez-Guevara et al. (2019) [53]	W	Dynamic	HC, AM	Camera
Trujillo-Romero and García-Bautista (2023) [47]	W, P	Dynamic	HC, AM	Kinect
Martínez-Guevara et al. (2023) [54]	W, P	Dynamic	HC, AM	Camera
Martínez-Sánchez et al. (2023) [27]	W	Dynamic	HC, AM, NHG	Camera
González-Rodríguez et al. (2024) [55]	P	Dynamic	HC, AM, NHG	Camera/MediaPipe
Current study	W, P	Dynamic	AM	Camera/YOLOv8

* L: alphabet letters; N: numbers; W: words; P: phrases.

1.1. Towards a Recognition System for the LSM

We present the sign data acquisition, the hardware selected and the fundamental concepts of our research towards a recognition system for the LSM.

1.1.1. Contactless Sign Data Acquisition with Simple Hardware

Due to the socioeconomic conditions of the main users of the LSM, this research uses contactless simple hardware for the sign data acquisition; i.e. a pure vision-based approach, since color cameras are widely accessible and available in portable devices, that are very common in Mexico. One important remark is —as presented in Table 4— that only one LSM research work [51] uses contact sensing for sign data acquisition.

1.1.2. Sign Features

From a Linguistics perspective, LSM signs present six documented parameters: basic articulatory parameters that simultaneously combine to form signs [31,56–58]. We propose a simplified Kinematics perspective, already shown in Figure 1, that combines four of those parameters into Arm Movement (AM):

1. Hand configuration (HC): the shape adopted by one or both hands. As seen in Table 1 and Table 3, most research focuses on the HC only. Hand segmentation [59] and hand pose detectors are very promising technologies for this feature. The number of HCs required to perform a sign is variable in the LSM, some examples regarding the number of HCs required for a sign are: number "`1`" (1 HC), number "`9`" (2 HCs), number "`15`" (2 hands, 1 HCs), "`grandmother`" (2 hands, 3 HCs). See Appendix A, for samples these signs.
2. Non-hand gestures (NHG): refers to facial expressions (frowning, raising eyebrows), gestures (puffing out cheeks, blowing) and body movements (pitching, nodding). While most signs do not require non-hand gestures, some of the LSM signs do. Some signs that require one or more NHG are: "`How are you?`", "`I'm sorry`", "`Surprise!`" (two NHGs of this sign are shown in Figure 1). See Appendix A, for links to samples of these signs.
3. Arm movement (AM): it can be characterized by tracking the joint movements of wrists, shoulders and elbows. It is enough to obtain the following basic articulatory parameters [31,56–58]:
   (a) Articulation location: the location on the signer's body or space where the signs are executed.
   (b) Hand movement: the type of movement made by the joints from one point to another.
   (c) Direction of movement: the trajectory followed by the hand when making the sign.
   (d) Hand orientation: orientation of the palm of one or both hands, with respect to the signer's body when making the manual configuration.

   This part can be studied from pose-based approaches (c.f. [21,23] with pose estimation using AlphaPose).

Other decompositions have been proposed, in order to simplify sign analysis, such as ([60], Fig. 1) were a LSM sign is decomposed into *fixed postures* and *movements*. We consider that this approach could loose important information, since transitions in hand postures are also important as documented in the Hamburg Notation System (HamNoSys) [61].

The use of pose estimators, in particular the use of MediaPipe, allow having information of face, hands and body features, c.f [22,30]. While, the use of pose estimator is quite frequent in SL research, there are still areas of improvements (c.f. [17], Fig. 8) where they designed a PhBFC to improve mediapipe hand pose estimation) and complementary approaches like bimodal frameworks [22] that show the current limitations of those estimators.

We consider that focusing on a single element to describe the LSM would not be adequate given their meaning and contribution to the sign. But covering everything at the same time is also very complex, as seen in most LSM research. Since most of the LSM work focused on HC, this paper focuses on the AM part and reports the approach created to analyze visual patterns in arm joint movements. Our current work uses YOLOv8 [62,63] for pose estimation. While it is 2D, and MediaPipe is better for 3D; we discuss our decision in Section 2.3.1.

The main contribution of this work is the use of arm movement keypoints, particularly wrist position, as a partial feature for sign language recognition. This is motivated by the observation in [30] that wrist location plays a crucial role in distinguishing similar signs. For instance, the same hand configuration used at different vertical positions (e.g., near the head to indicate headache, or near the stomach to indicate stomachache) conveys different meanings. By isolating and analyzing this spatial feature, we aim to better understand its discriminative power in sign recognition tasks.

An overview of the paper is as follows. Section 2 describes the custom dataset, the experimental design, software and hardware, data processing and methodologies. Section 3 describes the results from the analysis of two case studies. The conclusions and the limitations of our approach are presented in Section 5.

2. Materials and Methods

2.1. Custom Dataset

In this research, a proprietary dataset was developed with the active participation of the deaf community and LSM experts, ensuring no restrictions on recognizing hand configurations, arm movements and facial expressions. The creation of the dataset was reviewed and approved by the Bioethics Committee for Human Research at Cinvestav, and all participants provided written informed consent.

This dataset was divided into three subsets, see Tables 5, 6, and 7 for a list of the signs in each subset.

Table 5. Signs for the first subset.

No.	Semantic field	Sign
1	family	son*
2	greetings	hello*
3	days of the week	Monday*
4	family	godfather*
5	animals	deer*

*These signs are also in the second subset.

Table 6. Signs for the second subset.

No.	Semantic field	Sign	No.	Semantic field	Sign
1	verbs	hug	32	verbs	to arrive
2	adjectives	tall	33	days of the week	Monday*
3	drinks	atole	34	kitchen	tablecloth
4	transport	airplane	35	miscellaneous	sea
5	school	flag	36	fruits	melon
6	transport	bicycle	37	kitchen	table
7	greetings	Good afternoon!	38	verbs	to swim
8	greetings	Good morning!	39	colors	dark
9	cities	capital	40	family	godfather*
10	house†	house	41	animals	bird
11	miscellaneous	sky	42	clothing	pants
12	questions	How?	43	animals	penguin
13	questions	How are you?	44	school	blackboard
14	school	classmate	45	food	pizza
15	house	curtains†	46	room	iron
16	days of the week	day	47	miscellaneous	please
17	house	broom†	48	questions	Why?
18	living room	light bulb	49	time	present
19	animals	rooster	50	professions	president
20	adjectives	fat	51	bathroom	shower
21	adjectives	big	52	living room	living room
22	verbs	to like	53	food	sauce
23	family	daughter	54	cities	Saltillo
24	family	son*	55	clothing	shorts
25	greetings	hello*	56	verbs	to dream
26	time	hour	57	transport	taxi
27	time	today	58	bathroom	towel
28	animals	giraffe	59	animals	deer*
29	verbs	to play	60	house	window†
30	drinks	milk	61	clothing	dress
31	vegetables	lettuce	62	person	widower

*These signs are also in the first training set. †These signs are also in the third subset.

Table 7. Signs for the third subset.

No.	Semantic field	Sign
1	house	garbage
2	house	trash can
3	house	house*
4	house	curtains*
5	house	electricity
6	house	stairs
7	house	broom*
8	house	internet
9	house	garden
10	house	keys
11	house	wall
12	house	floor
13	house	door
14	house	roof
15	house	mop
16	house	window*

*These signs are also in the second subset.

The dataset comprises 74 signs: 73 performed by 17 subjects and one ("iron") performed by 16 subjects. In total we have 1257 color videos (900×720@90 fps). All signs show HCs and AM, and 3 of them have NHGs ("How?", "How are you?", "Why?"). There are four phrases in the dataset: "Good morning!" ("*¡Buenos días!*"), "Good afternoon!" ("*¡Buenas tardes!*"), "How are you?" ("*¿Cómo estás?*"), and "Why?" ("*¿Por qué?*"). The latter is a question word in English; but it is constructed with two words in Spanish, and also in LSM it is a sign composed of two signs with independent meaning. This information is summarized in Table 8.

Table 8. Custom dataset.

Feature	Description
Signs*	70 W, 4 P
Signers	17
Samples	73 signs with 17 samples, 1 sign with 16 samples
Sign features	HC, AM, NHG
Sign signal	Visual
Signal properties	900×720@90 fps
File format	MKV videos
Samples for training	10 samples
Samples for validation	2 samples
Samples for testing	5 samples

* W: words; P: phrases.

2.2. Experimental Design

The goal of these experiments is to classify dynamic LSM signs by detecting and tracking the wrist, elbow and shoulder joints, in order to characterize the AM. For this purpose, since the sign production involves motion and changes in shape in space, we have decided to use a pose-based approach for the sign signals acquisition and CNN for classification.

Two case studies are presented in this experiment. The first only considers shoulders and wrists, as the wrists exhibit the predominant movement while the shoulders serve as base joints with minimal displacement. The second case study includes the elbows, in addition to the shoulders and wrists, as the elbows also experience significant movement.

2.3. *Software and Hardware*

For the study, a pose detector and a CNN classifier framework are needed. For pose estimation, we conducted preliminary experiments to compare the commonly used MediaPipe and the YOLOv8-pose detector. MediaPipe detects 33 keypoints with its *Pose landmarker (Heavy)* model, and it can provide 2D and 3D coordinates. YOLOv8 detects 17 keypoints with its *YOLOv8x-pose-p6* model and provides 2D coordinates. YOLOv8x-pose-p6 keypoints 5–10 are for the shoulder, elbow and wrist joins, and MediaPipe keypoints 11–16 are for the same joints; see Figure 2.

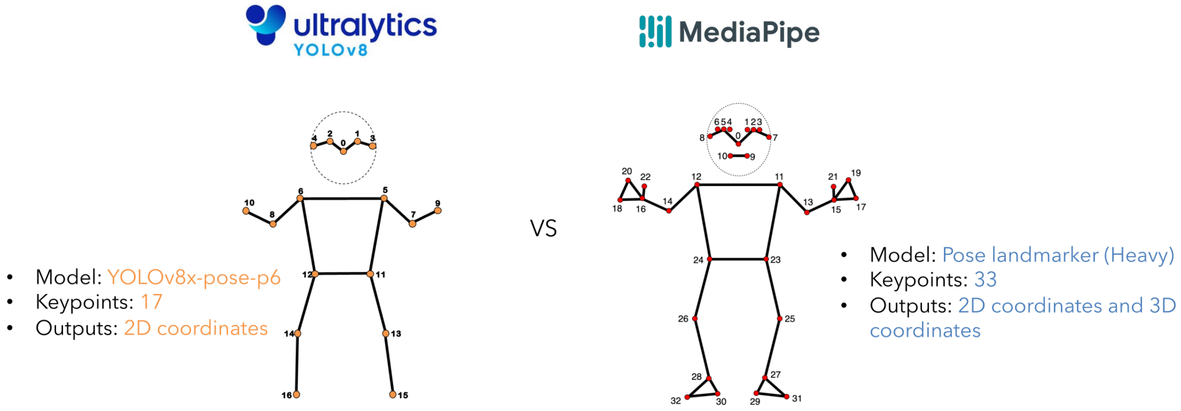

Figure 2. Keypoints in YOLOv8 and MediaPipe.

2.3.1. Comparison Between MediaPipe and YOLOv8 Pose Detection Models

We compared the above mentioned models for pose detection in several signs, and we decided to use YOLOv8 over MediaPipe due to frequent tracking failures of the wrist joint in many of the signs, particularly in occluded conditions of the hands. An example of this issue is shown in Figure 3.

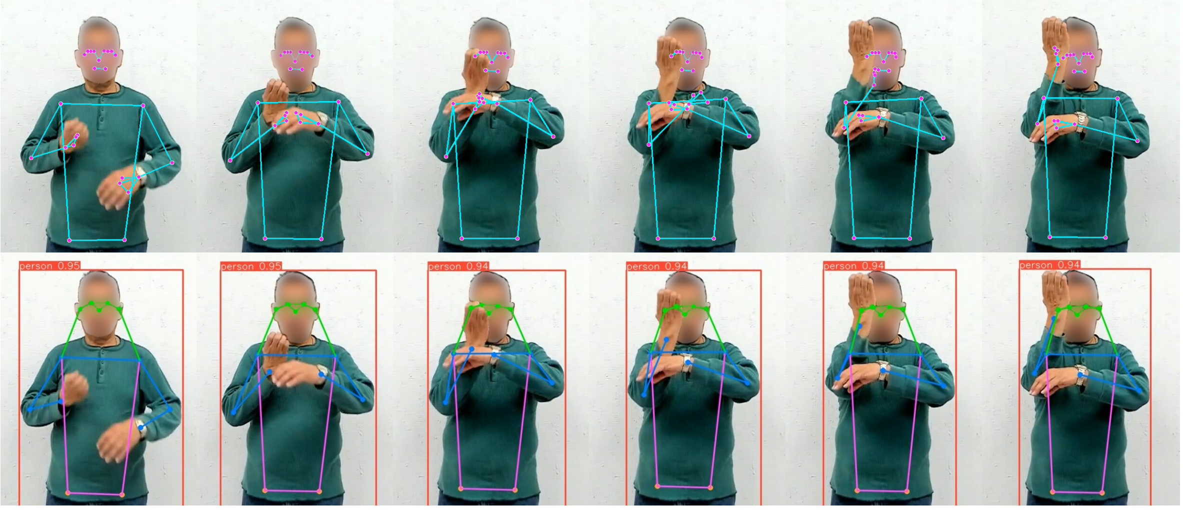

Figure 3. Comparison in wrist joint tracking between YOLOv8 vs MediaPipe. Example with the "`state`" sign. Top row: MediaPipe. Bottom row: YOLOv8 pose detector. Four inner frames: MediaPipe loses track of the wrist joint; while YOLOv8 keeps track of the AM in all frames.

As YOLOv8-pose was selected for pose estimation, we decided to use YOLOv8-cls to analyze visual patterns of the arm joint movements. Using a single technology for multiple tasks offers several advantages: a unified architecture reduces the need for format adaptation between different models, simplifies implementation and streamlines the workflow. Also, reduces the possible problems of training and running multiple models across different frameworks.

A micromamba environment was employed for the installation and implementation of the YOLOv8 pose detection and image classification models used in this work. Table 9 provides a summary of the technical specifications of the hardware and the key required software packages.

doi:10.20944/preprints202504.0981.v1

Table 9. Software and Hardware Specifications.

Software/Hardware	Version/Model
Operating System	Ubuntu 22.04.2
Graphics card	NVIDIA GeForce RTX 2080 Ti
CUDA	12.4
Python	3.11.8
PyTorch	2.2.2
Ultralytics YOLO	8.1.47

2.4. Data Processing

Since YOLOv8 works internally with square images, the scene was cropped to 720×720 pixels (see Figure 4). This adjustment does not affect sign visibility, as all relevant joints remain within the square frame.

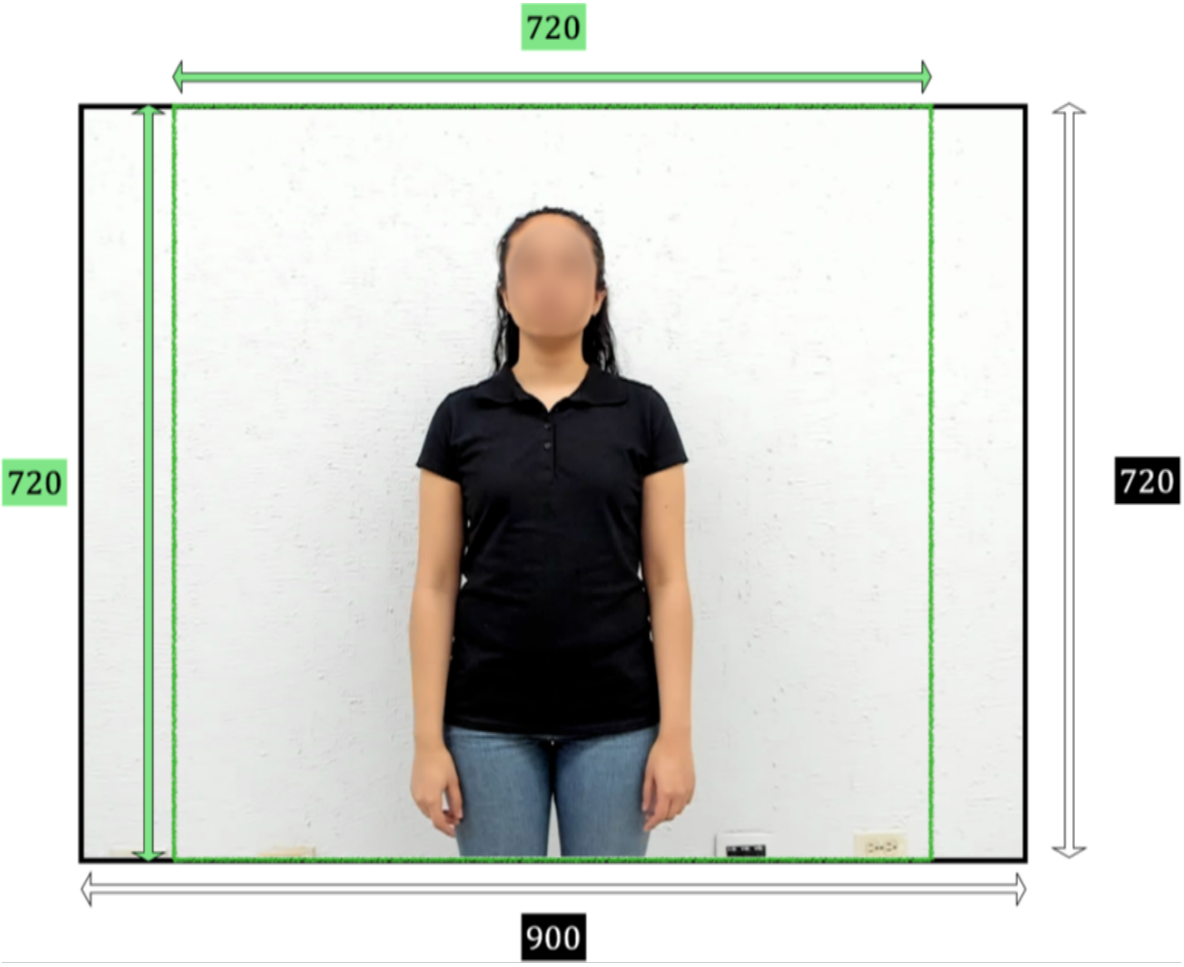

Figure 4. Dimensions of original and cropped frames.

LSM defines that only the upper part of the body is meaningful in signing; so only the upper 13 keypoints (out of 17) are considered, the 4 keypoints for knees and ankles are discarded. If the model fails to detect a joint, it is assigned a null value, which allows to easily discard these missing values in further processing. Below is an example of pose estimation applied to the initial and final poses of "`deer`" sign (Figure 5), as well as the extraction of the 13 keypoints.

doi:10.20944/preprints202504.0981.v1

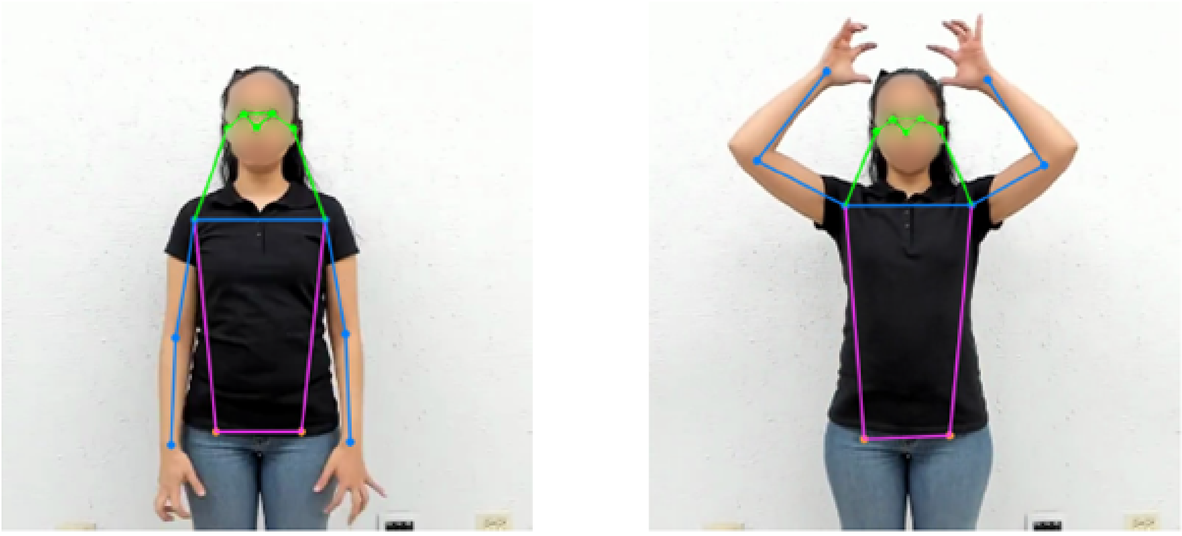

Figure 5. Pose detection in the "`deer`" sign. Left: neutral pose. Right: final pose.

The keypoints are stored in NPY format, a file type used by NumPy for efficiently storing data arrays. These arrays have dimensions of (13, 2, N): keypoints, 2D (x, y) coordinates and the number of frames in each video. From these arrays, the coordinates corresponding to the wrists, shoulders, and elbows are extracted accordingly to each case studies. The position of these coordinates was plotted for each frame, illustrating the movement pattern of each joint, as shown in Figure 6.

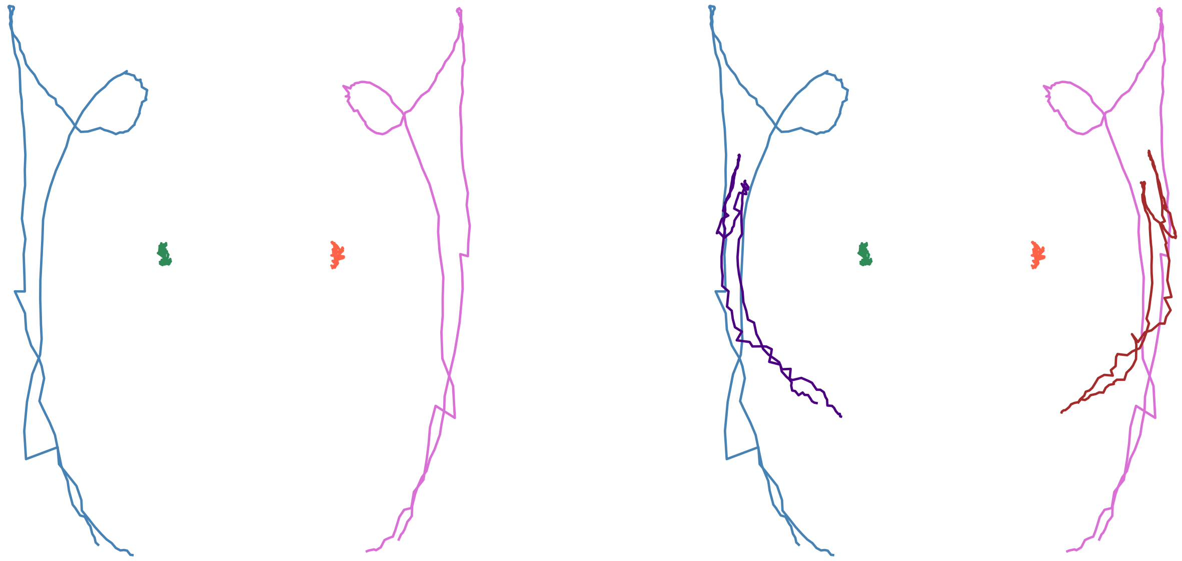

Figure 6. Movement shapes for the "`deer`" sign. Left: only wrists and shoulder. Right: also elbows.

2.5. Neural Network Training

2.5.1. First Subset

This subset consists of a small group of five signs, chosen for their distinguishable shapes based on a qualitative evaluation. The primary objective of this group is to conduct a more controlled evaluation of the neural network, which allows for a clearer analysis of what the network is learning in an environment with fewer variables. Examples of these signs are presented in Figure 7, while the corresponding words are listed in Table 5.

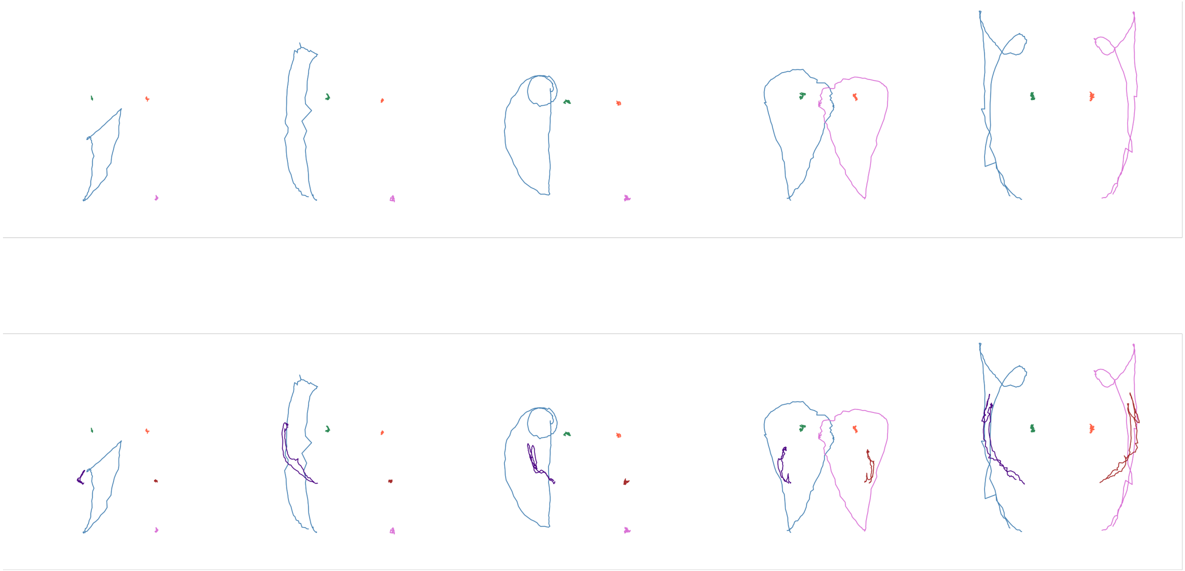

Figure 7. Shapes of the first subset (see words in Table 5). Top: only wrists and shoulders. Bottom: also elbows.

2.5.2. Second Subset

In this group, the signs are similarly distinguishable, but with a larger set consisting of 62 signs. The goal now is to assess whether the neural network's behavior remains consistent with that of the first set, despite the increased number of classes. Some examples of these signs are presented in Figure 8, and the corresponding words are listed in Table 6.

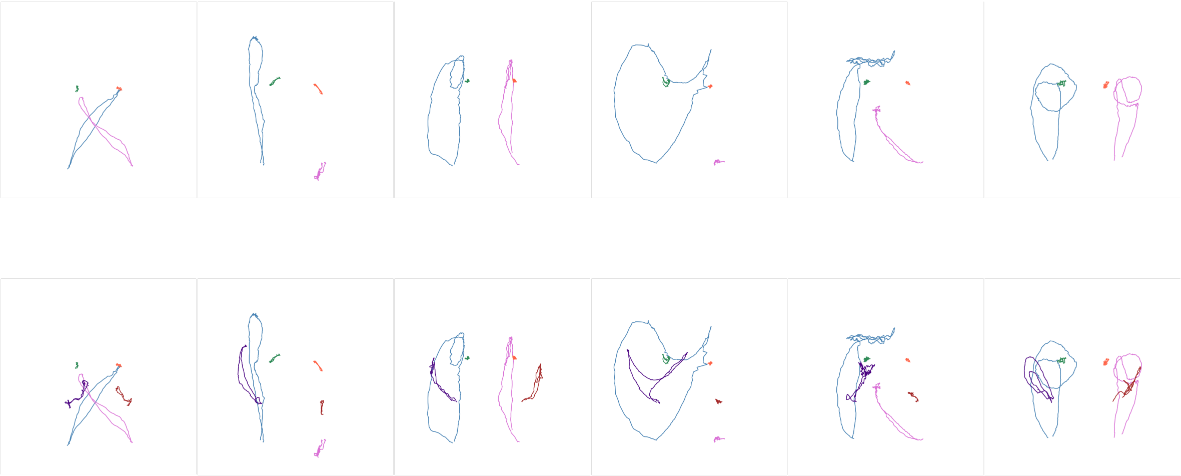

Figure 8. Shapes examples of the second subset (`"hug"`, `"tall"`, `"atole"`, `"airplane"`, `"flag"` and `"bicycle"`). Top: only wrists and shoulders. Bottom: also elbows.

2.5.3. Third Subset

The fourth set consists of 16 words related to the semantic field of `house`. This group is particularly notable for the high number of variants in its signs. As such, this experiment aims to assess the model's accuracy, as well as its ability to generalize and identify distinctive features within more complex sign language contexts. Examples of the sign forms from this set can be seen in Figure 9, and the corresponding vocabulary is outlined in Table 7.

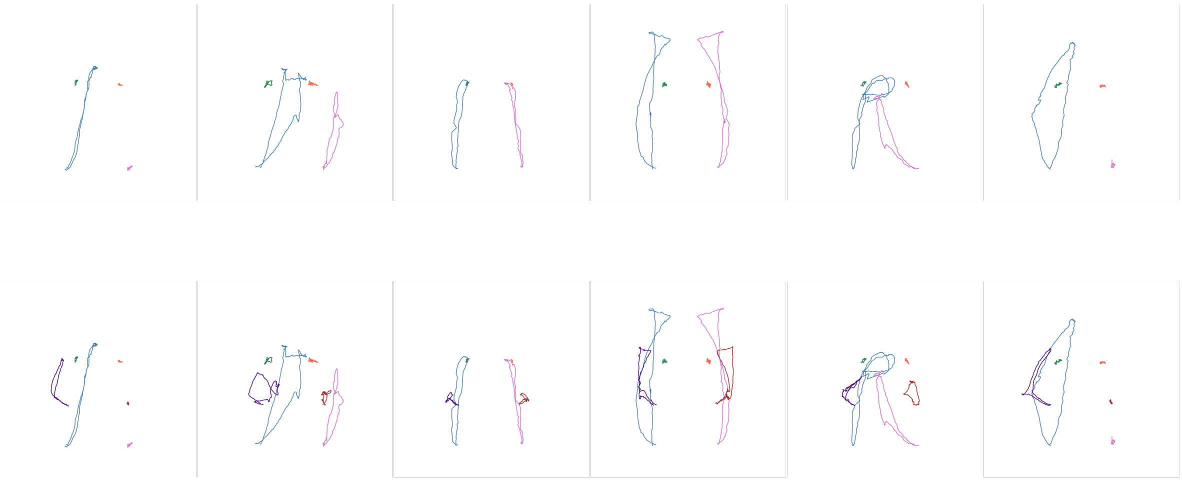

Figure 9. Shapes examples of the third subset ("`garbage`", "`trash can`", "`house`", "`curtains`", "`electricity`" and "`stairs`"). Top: only wrists and shoulders. Bottom: also elbows.

2.6. Training Parameters

The maximum number of examples per sign in all selected sets is 17: 10 examples were used for training, 2 for validation, and 5 for the testing phase. Image classification consists of assigning each image a label within a set of predefined classes. The image classification model used was *yolov8x-cls*. This classifier is the most robust of the classification models and maintains a deep CNN structure. The classifier output is a single class label and a confidence score. Table 10 shows the most relevant hyperparameters for model training and configuration.

Table 10. Training parameters and their descriptions.

Parameter	Value	Description
epochs	50	Number of epochs or training cycles.
batch	16	Number of images processed in each iteration.
imgsz	224	Size of the images input into the model.
patience	100	Number of epochs without improvement before stopping the training.
lr0	0.01	Initial learning rate.
pretrained	True	Indicates that the model uses pre-trained weights (ImageNet).
single_cls	False	If set to True, the model classifies into a single class.
dropout	0.0	Dropout rate. This is a regularization technique used to reduce overfitting in artificial neural networks.

Table 11 details the data augmentation related hyperparameters handled by YOLOv8 (not all parameters are active).

Table 11. Image augmentation parameters and their descriptions.

Parameter	Value	Description
hsv_h	0.015	Hue of the image in the HSV color space.
hsv_s	0.7	Saturation of the image in the HSV color space.
hsv_v	0.4	Brightness of the image in the HSV color space.
degrees	0.0	Random rotation applied to the images.
translate	0.1	Random translation of the images.
scale	0.5	Random scaling factor applied to the images.
shear	0.0	Random shear angle applied to the images.
perspective	0.0	Perspective transformation applied to the images.
flipud	0.0	Probability of flipping the image vertically.
fliplr	0.5	Probability of flipping the image horizontally.
bgr	0.0	BGR to RGB color space correction factor.
mosaic	1.0	Probability of using the mosaic technique to combine images.
mixup	0.0	Probability of mixing two images.
copy_paste	0.0	Technique of copying and pasting objects between images.
auto_augment	randaugment	Type of data augmentation used.
erasing	0.4	Probability of erasing parts of the image to simulate occlusions.
crop_fraction	1.0	Proportion of the image to be cropped. A value of 1.0 indicates no cropping.

2.7. Testing

Once the training stage is completed, the corresponding weights are saved in a custom model, which is then utilized for the subsequent testing phase. During this phase, key performance metrics are collected, such as Top-1 Accuracy and Top-5 Accuracy. Top-1 Accuracy measures how often the model's first prediction is correct, while Top-5 Accuracy evaluates whether the correct class appears among the five most probable predictions. These metrics are crucial for assessing the model's performance in a multi-class classification environment. Additionally, a confusion matrix is generated for each experiment, providing a detailed overview of correct and incorrect predictions for each class. The results, along with their interpretation and analysis, are discussed in the following section.

3. Results

The results are primarily evaluated using Top-1 Accuracy, Top-5 Accuracy, and the confusion matrix, which offer a comprehensive view of the model's performance across each subset. In addition, performance graphs depicting loss and accuracy across training epochs are included, allowing to observe the model's learning curve over time.

3.1. First Set

In the first experiment, five of the most distinguishable classes were selected (see confusion matrices in Figure 10). The results reveal that using only the shoulder and wrist coordinates achieved a Top-1 Accuracy of 0.9599. However, when the elbow coordinates were included, the Top-1 Accuracy decreased to 0.8799, suggesting that the additional information had a negative impact on performance. On the other hand, for Top-5 Accuracy, both configurations achieved a perfect score of 1.0, demonstrating the model's ability to correctly identify the target class within its top five predictions.

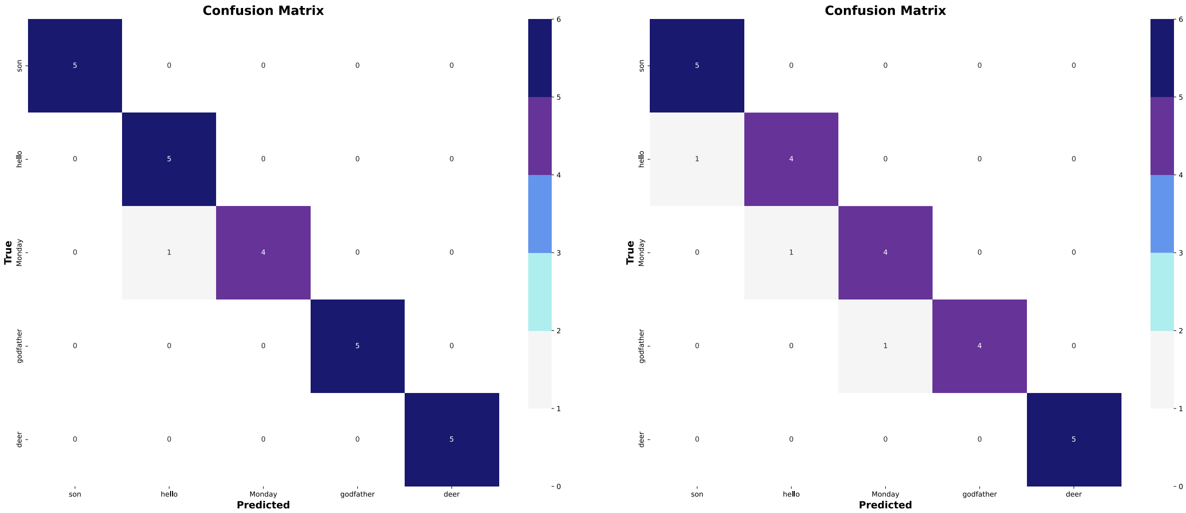

Figure 10. Confusion matrices for the first subset. Left: only wrists and shoulders. Right: also elbows.

Both the "son" and "deer" classes were classified with high accuracy in both case studies. However, slight confusion was observed between the "monday" and "hello" classes in the first case. Additionally, when elbow coordinates were included, the model made errors in three of the five classes, indicating greater difficulty in differentiating between them. The performance graphs show that the accuracy in both models tends to stabilize around the 30th epoch, while the loss continues to decrease. Despite this, the model using only the wrist and shoulder coordinates outperformed the version with elbow coordinates, achieving higher accuracy (see graphs in Figure 11). In summary, the results are highly favorable in the best-case scenario, with a classification rate exceeding 95%. This suggests that the model is capable of effectively distinguishing between a limited number of well-defined classes. However, it is preferable to restrict the analysis to wrist and shoulder data, as including elbow data appears to negatively impact performance.

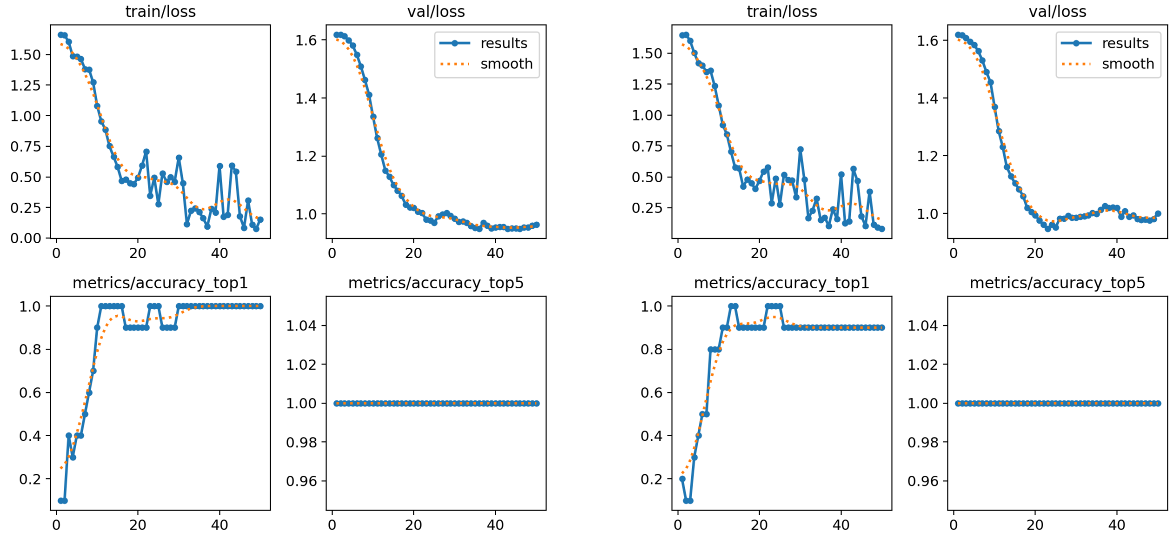

Figure 11. Performance charts for the first subset. Left: only wrists and shoulders. Right: also elbows.

3.2. Second Set

In the second experiment, we expanded the number of classes to 62, while ensuring that the shapes remained distinguishable from one another (see confusion matrices in Figure 12). The model using only wrist and shoulder coordinates achieved a Top-1 Accuracy of 0.6375, whereas including elbow information resulted in a slight improvement to 0.6537.

For Top-5 Accuracy, the results were similar, with the first model achieving an accuracy of 0.8640, which improved to 0.8932 when elbow data was included. Performance analysis during training and validation revealed a consistent trend in both models: accuracy steadily increased while loss progressively decreased (see Figures 6.4a and 6.4b), indicating effective learning. The best model achieved an overall accuracy of 65%, which is acceptable but showed variability in class performance. Some classes were classified nearly perfectly, while others exhibited notable precision issues. This suggests that, despite clear visual distinctions between classes, the large number of classes (62) combined with the limited number of examples per class (5) may be hindering the model's ability to generalize effectively. In conclusion, although incorporating elbow information improves classification accuracy, the inconsistent performance underscores the need for more examples per class to optimize the model's results.

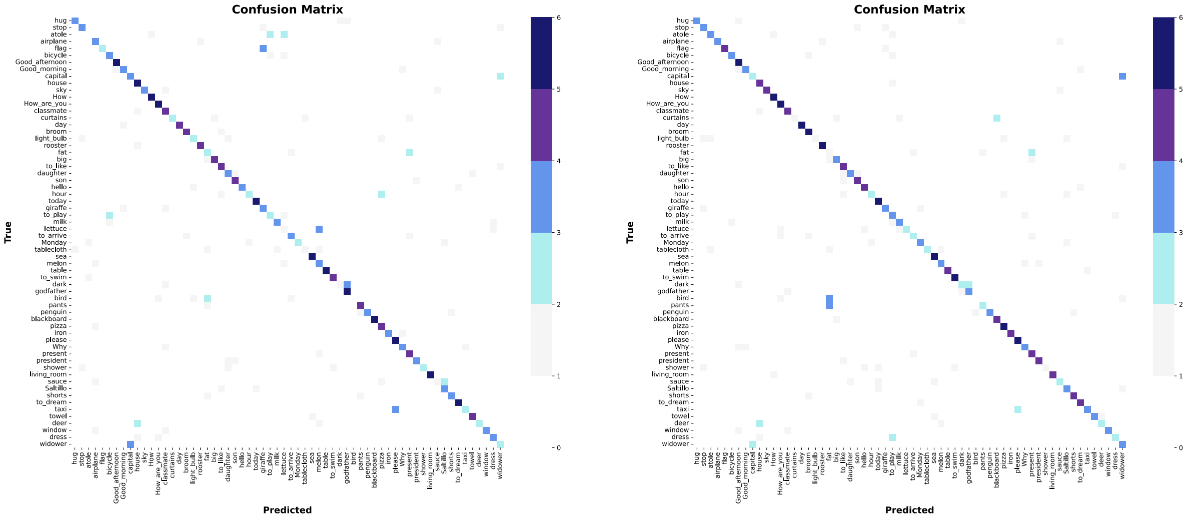

Figure 12. Confusion matrices for the second subset. Left: only wrists and shoulders. Right: also elbows.

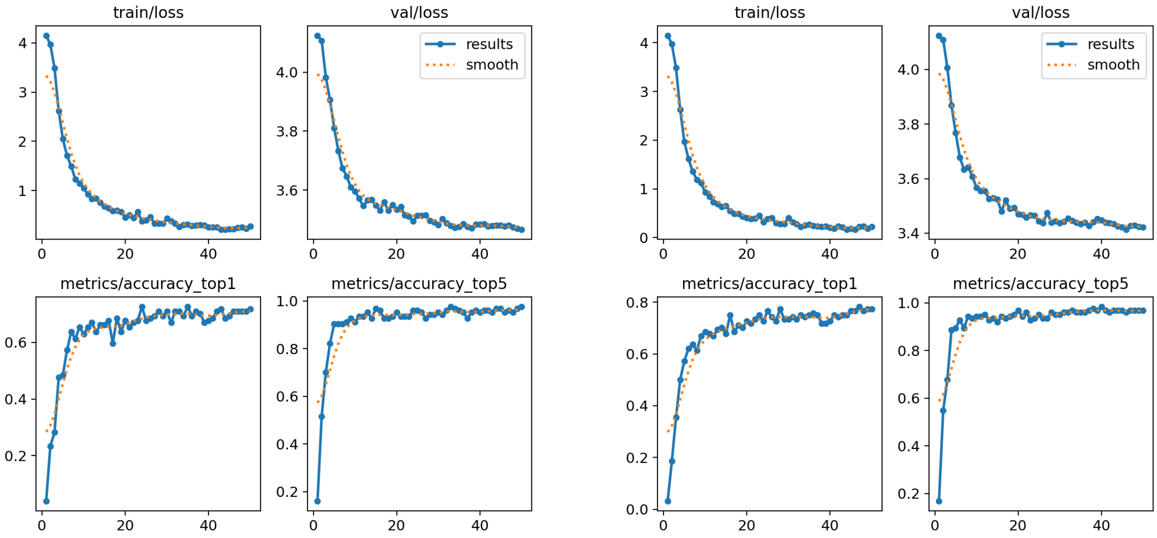

Figure 13. Performance charts for the second subset. Left: only wrists and shoulders. Right: also elbows.

3.3. Third Set

In this experiment, the set is comprised of 16 words in the `home` semantic field. The complexity of this group lies in the fact that some signs have variants. It is interesting to note that in both models, words such as "`internet`", "`keys`", "`mop`" and "`window`" were classified correctly, since they showed less variability. In contrast, words like "`curtains`", "`garden`" and "`wall`" performed poorly, with poor predictions in both models (see confusion matrices in Figure 14).

The model using only wrist and shoulder information achieved a top-1 accuracy of 0.6875, while including the elbow coordinates increased the accuracy to 0.7125. For top-5 accuracy, both models achieved a value of 0.9250.

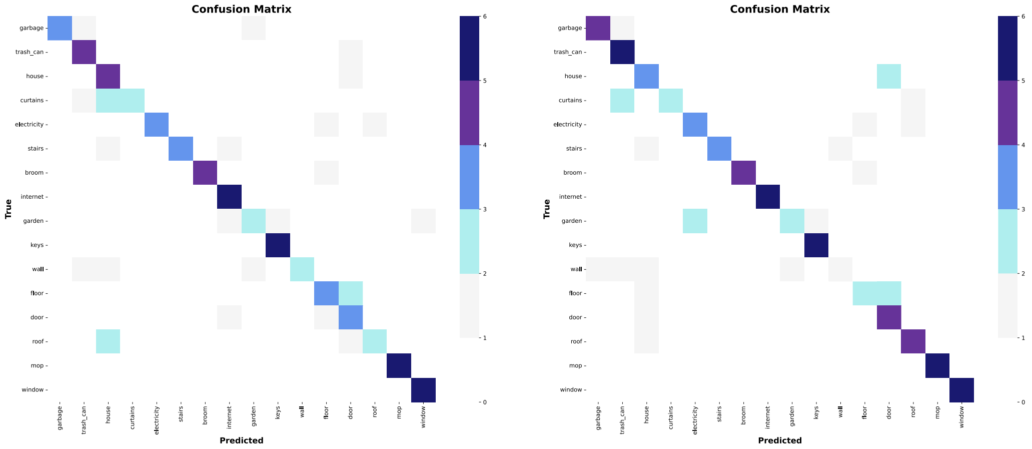

Figure 14. Confusion matrices for the third subset. Left: only wrists and shoulders. Right: also elbows.

The performance in both studies was quite similar (see the graphs in Figure 15), showing fluctuations during training, but with a tendency to stabilize at a constant value towards the later stages. This suggests that the model has managed to learn the main features of the characters, although its generalization capacity is limited by the complexity of the variants within the set. The classification rate reached up to 71% when the elbow information was included, which indicates that this additional information contributes positively to the recognition, although the increase in accuracy is not very significant.

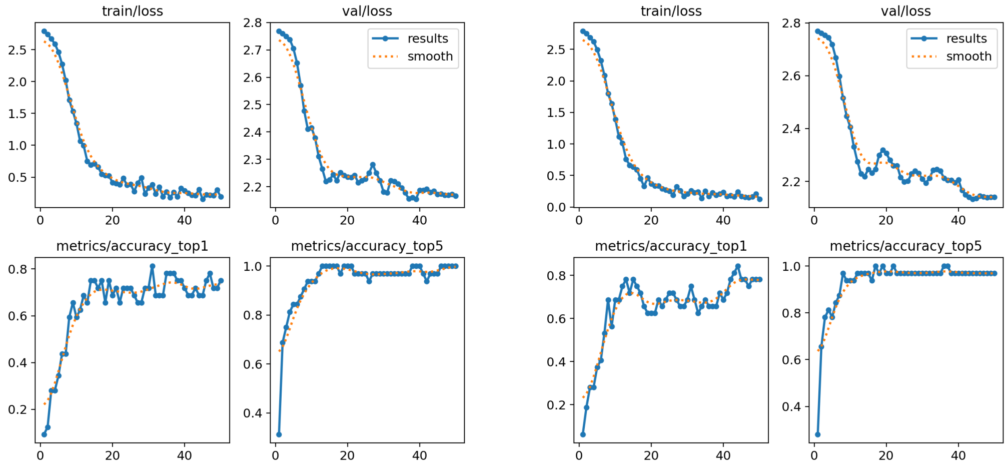

Figure 15. Performance charts for the third subset. Left: only wrists and shoulders. Right: also elbows.

Despite the limitations, the model was able to detect patterns in some cases. However, its ability to generalize across a large number of classes, variants and a limited number of examples is insufficient.

Notwithstanding, the performance graphs reveal a tendency toward stabilization, suggesting that while the model holds potential for certain datasets, it requires additional information —such as finger movements— to enhance its classification accuracy in more complex scenarios.

4. Discussion

Table 12 presents the accuracy values based on the Top-1 Accuracy metric obtained using the YOLOv8 model. The results indicate that including elbow coordinates led to better performance in two out of the three experiments. Although the improvement was modest (ranging from 3% to 4%), it suggests that incorporating additional joint information can contribute to more accurate classifications.

Table 12. Comparative table with the values of Top-1 Accuracy.

Dataset	No. clases	Description	With elbows	Without elbows
1	5	More distinguishable	0.8799	0.9599
2	62	More or less distinguishable	0.6537	0.6375
3	16	group `house`	0.7125	0.6875

The experiments with various datasets allowed us to observe the behavior of the convolutional neural network (CNN) based on the input data. It became evident that the network's performance is heavily influenced by the selection of classes. Using all available classes from the database is not always ideal, as this tends to yield suboptimal results. Therefore, a more focused approach, where only relevant classes are included, is recommended for improving model classification.

Despite certain limitations —such as the small number of examples per class, the presence of variants, and the high similarity between some signs— the neural network was still able to classify a significant number of signs correctly and recognize patterns in the movement data. This demonstrates the potential of the YOLOv8 model for this type of task.

In comparison to other CNNs, YOLOv8 stands out due to its optimized architecture, which allows for the use of pre-trained models on large datasets like ImageNet. This enables the model to achieve high accuracy and efficiency, making it suitable for real-time applications. However, as with any model, performance is largely dependent on the quality and quantity of the input data. In this case, the limited number of examples (17 per class) restricts the network's ability to achieve optimal accuracy.

5. Conclusions

This paper presents the ongoing work towards the creation of a recognition system for the LSM. A sign features decomposition is proposed into HC, AM and NHG. Contactless simple hardware was selected for sign signal acquisition. A custom proprietary dataset of 74 signs (70 words, 4 phrases) was constructed for this research. In contrast to most of the LSM research, this paper reports the analysis focused on the AM part of signs, instead of HC focused or a holistic approach (HC + AM + NHG).

The analysis were conducted through a series of classification experiments using YOLOv8, aimed at identifying visual patterns in the movement of key joints: wrists, shoulders, and elbows. A pose detection model was used to extract joint movements, followed by an image classification model (both integrated into YOLOv8) to classify the shapes generated by these movements.

The results, discussed in the previous section, highlight both the potential and the limitations of our approach. The experiments demonstrated that it is possible to classify a considerable number of signs, indicating that this dataset and strategy could serve as a useful tool for training a convolutional neural network (CNN), such as YOLOv8. However, the analysis also reveals that the current structure of the dataset, characterized by a limited number of examples, variations between classes, and high similarity among some signs, presents challenges that must be addressed through alternative approaches.

These experiments are the first stage of a larger project. For now, we are focusing on the analysis of arm movement (shoulders, elbows, and wrists) because it is a less studied feature and information can be extracted from it using a relatively simple methodology.

The comparison between the two case studies was intended to assess whether the inclusion of a greater number of keypoints improves the performance of the model. This seems to indicate that this assumption is correct. The next immediate step is to optimize these results, either by using a different convolutional neural network (CNN) or by exploring different architectures, such as recurrent neural networks (RNN), but keeping the focus on the use of keypoints; i.e. using pose-based approaches.

Later, the goal will be to integrate other essential components of sign language, such as manual configuration and non-hand gesture, to develop a more complete system. Ultimately, this will allow progress towards automatic sign language recognition.

Author Contributions: Conceptualization, G.H.-A., K.O.-H. and M.C.; methodology, G.H.-A., K.O.-H. and M.C.; software, G.H.-A. and K.O.-H.; validation, G.H.-A.; formal analysis, G.H.-A., K.O.-H. and M.C.; investigation, G.H.-A., K.O.-H. and M.C.; resources, K.O.-H. and M.C.; data curation, G.H.-A.; writing—original draft preparation, G.H.-A., K.O.-H. and I.L-J.; writing—review and editing, G.H.-A., K.O.-H., M.C. and I.L-J.; visualization, G.H.-A.; supervision, K.O.-H., and M.C.; project administration, K.O.-H.; funding acquisition, G.H.-A. and I.L-J.. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by CONAHCYT through scholarship grant number 828990.

Institutional Review Board Statement: Ethical review and approval were conducted by the Ethics Committee of Cinvestav (protocol code: 105/2023; date of approval: December 7th, 2023)

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study and due to technical and time limitations. Requests to access the datasets should be directed to keny.ordaz@cinvestav.edu.mx.

Acknowledgments: We thank Felipe Hernández Rodríguez, Ph.D., for providing an space at his institution for the dataset acquisition. We thank Hilda Xóchitl Cabrera Hernández, Ph.D.; Daniela Fernanda Espinoza Ibarra, M.D.; and María Guadalupe Luna Arguello for their help in contacting participants.

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AM	Arm Movement
ArSL	Arabic Sign Language
ASL	American Sign Language
BdSL	Bangladeshi Sign Language
CESAR	Recife Center for Advanced Studies and Systems
CSL	Chinese Sign Language
CNN	Convolutional Neural Network
DGS	German Sign Language (*Deutsche Gebärdensprache*)
EMG	Electromyography
FLV	Flash Video
fps	Frames per second, frame rate
HC	Hand Configuration
IMU	Inertial measurement unit
JPEG	Joint Photographic Experts Group, ISO/IEC 10918
LIBRAS	Brazilian Sign Language (*Língua Brasileira de Sinais*)

LSM	Mexican Sign Language (*Lengua de Señas Mexicana*)
LSTM	Long Short-Term Memory
MKV	Matroska Video
MLP	Multilayer Perceptron
MP4	MPEG-4 Part 14, ISO/IEC 14496-14:2003
NHG	Non-Hand Gesture
NN	Neural Network
NPY	NumPy standard binary file format
PJM	Polish Sign Language
RGBD	Red, Green, Blue and Depth
RNN	Recurrent Neural Network
sEMG	Surface EMG
SL	Sign Language
SLR	Sign Language Recognition
SVM	Support Vector Machine
SWF	Small Web Format
TSL	Turkish Sign Language
YOLO	You Only Look Once

Appendix A. Digital Glossary of LSM

The GDLSM [25] has 747 signs grouped in 19 thematic categories. We provide direct links to some of the signs included in this digital glossary, that were mentioned in Section 1.1.2.

- Numbers (*Números*):
  - 1 one: https://youtu.be/zcd4GfYz-fA
  - 9 nine: https://youtu.be/MgnvumQM-cQ
  - 15 fifteen (first variant): https://youtu.be/yZ3X38cFWUM
  - 15 fifteen (second variant): https://youtu.be/64jBCZXv6rY
- Family (*Familia*):
  - Grandmother (*abuela*) : https://youtu.be/lckOvtr0lZU
- Everyday expressions (*Expresiones cotidianas*):
  - How are you? (*¿Cómo estás?*) https://youtu.be/x7zFMacTe04
  - I'm sorry (*Disculpa*) https://youtu.be/bWwIisAtYCI
  - Surprise! (*¡Sorpresa!*) https://youtu.be/Q0OqTBjoIjU, this sign was used in Figure 1.

References

1. World Health Organization. World Reporting on hearing. https://www.who.int/publications/i/item/9789240020481, 2021. Accessed on 2025-03-31.
2. Secretaría de Salud. 530. Con discapacidad auditiva, 2.3 millones de personas: Instituto Nacional de Rehabilitación. https://www.gob.mx/salud/prensa/530-con-discapacidad-auditiva-2-3-millones-de-personas-instituto-nacional-de-rehabilitacion, 2021. Accessed on 2025-03-31.
3. SLAIT. SLAIT — Real-time Sign Language Translator with AI. https://slait.ai, 2024. Accessed on 2024-03-31.
4. Lenovo. Lenovo's AI-powered sign language translation solution empowers signers in Brazil. https://news.lenovo.com/ai-powered-sign-language-translation-solution-hearing-brazil/, 2023. Accessed on 2025-03-31.
5. Rocha, J.V.; Lensk, J.; Ferreira, M.D.C. Techniques for determining sign language gesture partially shown in image (s), 2023. United States Patent 11,587,362 B2.
6. Mane, V.; Puniwala, S.N.; Rane, V.N.; Gurav, P. Advancements in Sign Language Recognition: Empowering Communication for Individuals with Speech Impairments. *Grenze International Journal of Engineering &amp; Technology (GIJET)* **2024**, *10*, 4978–4984.
7. Krishnan, S.R.; Varghese, C.M.; Jayaraj, A.; Nair, A.S.; Joshy, D.; Sulbi, I.N. Advancements in Sign Language Recognition: Dataset Influence on Model Accuracy. In Proceedings of the 2024 International Conference

on IoT Based Control Networks and Intelligent Systems (ICICNIS), Dec 2024, pp. 1563–1568. https://doi.org/10.1109/ICICNIS64247.2024.10823232.

8. Chiradeja, P.; Liang, Y.; Jettanasen, C. Sign Language Sentence Recognition Using Hybrid Graph Embedding and Adaptive Convolutional Networks. *Applied Sciences* **2025**, *15*. https://doi.org/10.3390/app15062957.
9. Umut, I.; Kumdereli, U.C. Novel Wearable System to Recognize Sign Language in Real Time. *Sensors* **2024**, *24*. https://doi.org/10.3390/s24144613.
10. Rodríguez-Tapia, B.; Ochoa-Zezzatti, A.; Marrufo, A.I.S.; Arballo, N.C.; Carlos, P.A. Sign Language Recognition Based on EMG Signals through a Hibrid Intelligent System. *Res. Comput. Sci.* **2019**, *148*, 253–262.
11. Gu, Y.; Oku, H.; Todoh, M. American Sign Language Recognition and Translation Using Perception Neuron Wearable Inertial Motion Capture System. *Sensors* **2024**, *24*. https://doi.org/10.3390/s24020453.
12. Filipowska, A.; Filipowski, W.; Mieszczanin, J.; Bryzik, K.; Henkel, M.; Skwarek, E.; Raif, P.; Sieciński, S.; Doniec, R.; Mika, B.; et al. Pattern Recognition in the Processing of Electromyographic Signals for Selected Expressions of Polish Sign Language. *Sensors* **2024**, *24*. https://doi.org/10.3390/s24206710.
13. Galván-Ruiz, J.; Travieso-González, C.M.; Pinan-Roescher, A.; Alonso-Hernández, J.B. Robust Identification System for Spanish Sign Language Based on Three-Dimensional Frame Information. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23010481.
14. Hao, Z.; Duan, Y.; Dang, X.; Liu, Y.; Zhang, D. Wi-SL: Contactless Fine-Grained Gesture Recognition Uses Channel State Information. *Sensors* **2020**, *20*. https://doi.org/10.3390/s20144025.
15. Wang, Y.; Hao, Z.; Dang, X.; Zhang, Z.; Li, M. UltrasonicGS: A Highly Robust Gesture and Sign Language Recognition Method Based on Ultrasonic Signals. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23041790.
16. Al-Saidi, M.; Ballagi, A.; Hassen, O.A.; Saad, S.M. Type-2 Neutrosophic Markov Chain Model for Subject-Independent Sign Language Recognition: A New Uncertainty–Aware Soft Sensor Paradigm. *Sensors* **2024**, *24*. https://doi.org/10.3390/s24237828.
17. Urrea, C.; Kern, J.; Navarrete, R. Bioinspired Photoreceptors with Neural Network for Recognition and Classification of Sign Language Gesture. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23249646.
18. Niu, P. Convolutional neural network for gesture recognition human-computer interaction system design. *PLOS ONE* **2025**, *20*, 1–22. https://doi.org/10.1371/journal.pone.0311941.
19. Raihan, M.J.; Labib, M.I.; Jim, A.A.J.; Tiang, J.J.; Biswas, U.; Nahid, A.A. Bengali-Sign: A Machine Learning-Based Bengali Sign Language Interpretation for Deaf and Non-Verbal People. *Sensors* **2024**, *24*. https://doi.org/10.3390/s24165351.
20. Woods, L.T.; Rana, Z.A. Modelling Sign Language with Encoder-Only Transformers and Human Pose Estimation Keypoint Data. *Mathematics* **2023**, *11*. https://doi.org/10.3390/math11092129.
21. Eunice, J.; J, A.; Sei, Y.; Hemanth, D.J. Sign2Pose: A Pose-Based Approach for Gloss Prediction Using a Transformer Model. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23052853.
22. Gao, Q.; Hu, J.; Mai, H.; Ju, Z. Holistic-Based Cross-Attention Modal Fusion Network for Video Sign Language Recognition. *IEEE Transactions on Computational Social Systems* **2024**, pp. 1–12. https://doi.org/10.1109/TCSS.2024.3435693.
23. Kim, Y.; Baek, H. Preprocessing for Keypoint-Based Sign Language Translation without Glosses. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23063231.
24. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.; Lee, J.; et al. MediaPipe: A Framework for Perceiving and Processing Reality. In Proceedings of the Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019, 2019.
25. CDMX, I. Glosario Digital de Lengua de Señas Mexicana. https://lsm.indiscapacidad.cdmx.gob.mx, 2024. Accessed on 2025-03-31.
26. Serafín De Fleischmann, M.; González Pérez, R. *Manos con voz: diccionario de lengua de señas mexicana*; Consejo Nacional para Prevenir la Discriminación, 2011.
27. Martínez-Sánchez, V.; Villalón-Turrubiates, I.; Cervantes-Álvarez, F.; Hernández-Mejía, C. Exploring a Novel Mexican Sign Language Lexicon Video Dataset. *Multimodal Technologies and Interaction* **2023**, *7*. https://doi.org/10.3390/mti7080083.
28. Espejel-Cabrera, J.; Dominguez, L.; Cervantes, J.; Cervantes, J. Mexican sign language dataset. https://data.mendeley.com/datasets/6rj76z6y3n/1, 2023. Accessed on 2025-03-31, https://doi.org/10.17632/6rj76z6y3n.1.
29. Espejel, J.; Jalili, L.D.; Cervantes, J.; Canales, J.C. Sign language images dataset from Mexican sign language. *Data in Brief* **2024**, *55*, 110566. https://doi.org/10.1016/j.dib.2024.110566.

30. Mejía-Peréz, K.; Córdova-Esparza, D.M.; Terven, J.; Herrera-Navarro, A.M.; García-Ramírez, T.; Ramírez-Pedraza, A. Automatic Recognition of Mexican Sign Language Using a Depth Camera and Recurrent Neural Networks. *Applied Sciences* **2022**, *12*. https://doi.org/10.3390/app12115523.
31. Calvo-Hernández, M.T. Diccionario Español-Lengua de Señas Mexicana (DIELSEME). http://campusdee.ddns.net/dielseme.aspx, 2004. Accessed on 2025-03-31.
32. Álvarez Hidalgo, A.; Acosta-Arellano, A.; Moctezuma-Contreras, C.; Sanabria-Ramos, E. Diccionario Lengua de Señas Mexicana (DIELSEME2). http://campusdee.ddns.net/dielseme.aspx, 2009. Accessed on 2025-03-31.
33. Cruz-Aldrete, M. Hacia la construcción de un diccionario de Lengua de Señas Mexicana. *Revista de Investigación* **2014**, *38*, 57–80.
34. Solís, F.; Martínez, D.; Espinoza, O. Automatic Mexican Sign Language Recognition Using Normalized Moments and Artificial Neural Networks. *Engineering* **2016**, *8*, 733–740. https://doi.org/10.4236/eng.2016.810066.
35. Carmona-Arroyo, G.; Rios-Figueroa, H.V.; Avendaño-Garrido, M.L. Mexican Sign-Language Static-Alphabet Recognition Using 3D Affine Invariants. *Machine Vision Inspection Systems, Volume 2: Machine Learning-Based Approaches* **2021**, pp. 171–192. https://doi.org/10.1002/9781119786122.ch9.
36. Salinas-Medina, A.; Neme-Castillo, J.A. A real-time deep learning system for the translation of mexican signal language into text. In Proceedings of the 2021 Mexican International Conference on Computer Science (ENC), 2021, pp. 1–7. https://doi.org/10.1109/ENC53357.2021.9534825.
37. Rios-Figueroa, H.V.; Sánchez-García, A.J.; Sosa-Jiménez, C.O.; Solís-González-Cosío, A.L. Use of Spherical and Cartesian Features for Learning and Recognition of the Static Mexican Sign Language Alphabet. *Mathematics* **2022**, *10*. https://doi.org/10.3390/math10162904.
38. Morfín-Chávez, R.F.; Gortarez-Pelayo, J.J.; Lopez-Nava, I.H. Fingerspelling Recognition in Mexican Sign Language (LSM) Using Machine Learning. In Proceedings of the Advances in Computational Intelligence; Calvo, H.; Martínez-Villaseñor, L.; Ponce, H., Eds., Cham, 2023; pp. 110–120. https://doi.org/10.1007/978-3-031-47765-2_9.
39. Sánchez-Vicinaiz, T.J.; Camacho-Pérez, E.; Castillo-Atoche, A.A.; Cruz-Fernandez, M.; García-Martínez, J.R.; Rodríguez-Reséndiz, J. MediaPipe Frame and Convolutional Neural Networks-Based Fingerspelling Detection in Mexican Sign Language. *Technologies* **2024**, *12*. https://doi.org/10.3390/technologies12080124.
40. García-Gil, G.; López-Armas, G.d.C.; Sánchez-Escobar, J.J.; Salazar-Torres, B.A.; Rodríguez-Vázquez, A.N. Real-Time Machine Learning for Accurate Mexican Sign Language Identification: A Distal Phalanges Approach. *Technologies* **2024**, *12*. https://doi.org/10.3390/technologies12090152.
41. Jimenez, J.; Martin, A.; Uc, V.; Espinosa, A. Mexican Sign Language Alphanumerical Gestures Recognition using 3D Haar-like Features. *IEEE Latin America Transactions* **2017**, *15*, 2000–2005. https://doi.org/10.1109/TLA.2017.8071247.
42. Martinez-Seis, B.; Pichardo-Lagunas, O.; Rodriguez-Aguilar, E.; Saucedo-Diaz, E.R. Identification of Static and Dynamic Signs of the Mexican Sign Language Alphabet for Smartphones using Deep Learning and Image Processing. *Res. Comput. Sci.* **2019**, *148*, 199–211.
43. Martínez-Gutiérrez, M.E.; Rojano-Cáceres, J.R.; Benítez-Guerrero, E.; Sánchez-Barrera, H.E. Data Acquisition Software for Sign Language Recognition. *Res. Comput. Sci.* **2019**, *148*, 205–211.
44. Rodriguez, M.; Oubram, O.; Ali, B.; Lakouari, N. Mexican Sign Language's Dactylology and Ten First Numbers - Extracted Features and Models. https://data.mendeley.com/datasets/hmsc33hmkz/1, 2023. Accessed on 2025-03-31, https://doi.org/10.17632/hmsc33hmkz.1.
45. Sosa-Jiménez, C.O.; Ríos-Figueroa, H.V.; Rechy-Ramírez, E.J.; Marin-Hernandez, A.; González-Cosío, A.L.S. Real-time Mexican Sign Language recognition. In Proceedings of the 2017 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), 2017, pp. 1–6. https://doi.org/10.1109/ROPEC.2017.8261606.
46. García-Bautista, G.; Trujillo-Romero, F.; Caballero-Morales, S.O. Mexican sign language recognition using kinect and data time warping algorithm. In Proceedings of the 2017 International Conference on Electronics, Communications and Computers (CONIELECOMP), 2017, pp. 1–5. https://doi.org/10.1109/CONIELECOMP.2017.7891832.
47. Trujillo-Romero, F.; García-Bautista, G. Mexican Sign Language Corpus: Towards an Automatic Translator. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*. https://doi.org/10.1145/3591471.

48. Espejel-Cabrera, J.; Cervantes, J.; García-Lamont, F.; Ruiz-Castilla, J.S.; Jalili, Dominguez, L. Mexican sign language segmentation using color based neuronal networks to detect the individual skin color. *Expert Systems with Applications* **2021**, *183*, 115295. https://doi.org/10.1016/j.eswa.2021.115295.
49. Rodriguez, M.; Oubram, O.; Bassam, A.; Lakouari, N.; Tariq, R. Mexican Sign Language Recognition: Dataset Creation and Performance Evaluation Using MediaPipe and Machine Learning Techniques. *Electronics* **2025**, *14*. https://doi.org/10.3390/electronics14071423.
50. Sosa-Jiménez, C.O.; Ríos-Figueroa, H.V.; Solís-González-Cosío, A.L. A Prototype for Mexican Sign Language Recognition and Synthesis in Support of a Primary Care Physician. *IEEE Access* **2022**, *10*, 127620–127635. https://doi.org/10.1109/ACCESS.2022.3226696.
51. Varela-Santos, H.; Morales-Jiménez, A.; Córdova-Esparza, D.M.; Terven, J.; Mirelez-Delgado, F.D.; Orenday-Delgado, A. Assistive device for the translation from mexican sign language to verbal language. *Computación y Sistemas* **2021**, *25*, 451–464. https://doi.org/10.13053/cys-25-3-3459.
52. Martínez-Guevara, N.; Curiel, A. Quantitative Analysis of Hand Locations in both Sign Language and Non-linguistic Gesture Videos. In Proceedings of the Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources, 2024, pp. 225–234.
53. Martínez-Guevara, N.; Rojano-Cáceres, J.R.; Curiel, A. Detection of Phonetic Units of the Mexican Sign Language. In Proceedings of the 2019 International Conference on Inclusive Technologies and Education (CONTIE), 2019, pp. 168–1685. https://doi.org/10.1109/CONTIE49246.2019.00040.
54. Martínez-Guevara, N.; Rojano-Cáceres, J.R.; Curiel, A. Unsupervised extraction of phonetic units in sign language videos for natural language processing. *Universal Access in the Information Society* **2023**, *22*, 1143–1151. https://doi.org/10.1007/s10209-022-00888-6.
55. González-Rodríguez, J.R.; Córdova-Esparza, D.M.; Terven, J.; Romero-González, J.A. Towards a Bidirectional Mexican Sign Language–Spanish Translation System: A Deep Learning Approach. *Technologies* **2024**, *12*. https://doi.org/10.3390/technologies12010007.
56. López-García, L.A.; Rodríguez-Cervantes, R.M.; Zamora-Martínez, M.G.; Esteban-Sosa, S.S. *Mis manos que hablan, lengua de señas para sordos*; Editorial Trillas, México, 2008.
57. Cruz-Aldrete, M. *Gramática de la Lengua de Señas Mexicana*; El colegio de México, 2008.
58. Escobedo-Delgado, C.E., Ed. *Diccionario de Lengua de Señas Mexicana de la Ciudad de México*; INDEPEDI, 2017.
59. Sánchez-Brizuela, G.; Cisnal, A.; de la Fuente-López, E.; Fraile, J.C.; Pérez-Turiel, J. Lightweight real-time hand segmentation leveraging MediaPipe landmark detection. *Virtual Reality* **2023**, *27*, 3125–3132. https://doi.org/10.1007/s10055-023-00858-0.
60. Martínez-Guevara, N.; Rojano-Cáceres, J.R.; Curiel, A. Unsupervised extraction of phonetic units in sign language videos for natural language processing. *Universal Access in the Information Society* **2023**, *22*, 1143–1151. https://doi.org/10.1007/s10209-022-00888-6.
61. Hanke, T. HamNoSys — Representing Sign Language Data in Language Resources and Language Processing Contexts. In Proceedings of the LREC 2004, Workshop proceedings: Representation and processing of sign languages; Streiter, O.; Vettori, C., Eds. European Language Resources Association (ELRA), European Language Resources Association (ELRA), 2004, pp. 1–6.
62. Rasheed, A.F.; Zarkoosh, M. Optimized YOLOv8 for multi-scale object detection. *Journal of Real-Time Image Processing* **2024**, *22*, 6. https://doi.org/10.1007/s11554-024-01582-x.
63. Wang, H.; Liu, C.; Cai, Y.; Chen, L.; Li, Y. YOLOv8-QSD: An Improved Small Object Detection Algorithm for Autonomous Vehicles Based on YOLOv8. *IEEE Transactions on Instrumentation and Measurement* **2024**, *73*, 1–16. https://doi.org/10.1109/TIM.2024.3379090.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.