Multimodal Technologies and Interaction

MDPI

Article

Exploring a Novel Mexican Sign Language Lexicon Video Dataset

Víctor Martínez-Sánchez [1], Iván Villalón-Turrubiates [1], Francisco Cervantes-Álvarez [1] and Carlos Hernández-Mejía [2,*]

[1] Instituto Tecnológico de Estudios Superiores de Occidente (ITESO), Guadalajara 45604, Mexico; ng683728@iteso.mx (V. M.-S.); villalon@iteso.mx (I.V.-T.); fcervantes@iteso.mx (F.C.-Á.)
[2] Tecnológico Nacional de México/ITS de Misantla, Misantla 93850, Mexico
* Correspondence: cmahernandez@gmail.com; Tel.: +52-221-573-1858

Abstract: This research explores a novel Mexican Sign Language (MSL) lexicon video dataset containing the dynamic gestures most frequently used in MSL. Each gesture consists of a set of different versions of videos under uncontrolled conditions. The MX-ITESO-100 dataset is composed of a lexicon of 100 gestures and 5000 videos from three participants with different grammatical elements. Additionally, the dataset is evaluated in a two-step neural network model as having an accuracy greater than 99% and thus serves as a benchmark for future training of machine learning models in computer vision systems. Finally, this research provides an inclusive environment within society and organizations, in particular for people with hearing impairments.

Keywords: Mexican sign language; dataset; hand gestures; computer vision

Citation: Martínez-Sánchez, V.; Villalón-Turrubiates, I.; Cervantes-Álvarez, F.; Hernández-Mejía, C. Exploring a Novel Mexican Sign Language Lexicon Video Dataset. *Multimodal Technol. Interact.* **2023**, *7*, 83. https://doi.org/10.3390/mti7080083

Academic Editor: Mu-Chun Su

Received: 6 July 2023
Revised: 12 August 2023
Accepted: 15 August 2023
Published: 19 August 2023

1. Introduction

In Mexico, the incorporation of deaf people in education is lacking, since only 14% of the deaf population in the age group between 3 and 29 years old access education with the support of a hearing aid. Additionally, those who have been incorporated frequently face inappropriate educational strategies which insufficiently use Mexican sign language (MSL) and therefore academical success is difficult and opportunities for insertion in the workplace are few.

Sign language has enabled effective communication with people who have hearing loss. The current challenge is focused on the identification of static or dynamic gestures in real time using systems based on machine learning techniques. For instance, in [1], a novel automatic sign language recognition system has been proposed that incorporates multiple gestures, including hand, body, and face gestures, to overcome the limitations of focusing solely on hand gestures. Using a depth camera and recurrent neural networks, the system achieves high accuracy, with the best model reaching 97% accuracy on clean test data and 90% accuracy on highly noisy data after thorough evaluation and ablation studies.

The classical problem encountered by researchers in the realm of image or image sequence categorization revolves around the pursuit of an appropriate dataset that aligns with the specific requirements of their study [1–3]. The most important factors to consider in this quest encompass language compatibility, format suitability, environmental characteristics, and dimensions, as well as the presence of static or dynamic signs. Presently, the acquisition of a comprehensive dataset catering to the Mexican sign language (MSL) has proven to be a challenging endeavor. Additionally, the format of the dataset plays a pivotal role in determining its compatibility with the research objectives. The dataset should ideally have a standardized format that enables seamless integration with various computational algorithms and models. This ensures that the data can be efficiently processed, analyzed, and classified, thereby facilitating meaningful insights into the categorization of MSL-based images.

The design of expert computer vision systems brings challenges to the acquisition of information. Data collected through videos and images constitute the core of this work. However, in most cases, metadata provide valuable information to assist in the classification of signs. Metadata acquisition can be achieved through thermal and depth sensors. According to the literature, Kinect is one of the most common sensors used by researchers to create datasets. This device was introduced by Microsoft in 2010 and it consists of an RGB camera with an infrared depth sensor used for the recognition of human body gestures. In addition, the device can model a person's skeleton, with an emphasis on joints. Therefore, researchers find Kinect very useful for this purpose. In [4], this device has been used for MSL recognition by collecting color images with a size of 115 × 115 pixels. These images were pre-processed in order to be sent to a classification model. Finally, the system can recognize gestures in real time and display text on the screen related to their meaning.

During the development of a machine learning model, the use of a missing dataset presents an additional challenge for researchers. As a result of this, certain authors prefer to create their own datasets. In [5], researchers have generated a dataset for Mexican signs. The Kinect device was used to carry out this task. Metadata provide complementary information about an individual's skeleton. According to the results, the precision reported for this work is greater than 95 percent and therefore proves that the metadata contribute significantly. In [1], MSL recognition was explored by producing 3000 samples with data from both the hands, body and face of about 30 signs using the Kinect device. However, these data were stored in a personalized format that makes it impossible to reuse the content.

On the other hand, in [6], a video dataset has been established in order to extract a specific sequence of frames. Subsequently, they used image segmentation and feature extraction from three regions of interest to generate geometrical features. This approach eliminates the need to use additional devices such as Kinect sensors. This dataset allows one to carry out machine learning techniques such as [7] Support Vector Machine (SVM), Nearest Neighbor (NN), Bayesian methods, and k-dimensional tree. Therefore, a dataset is essential for the success of methodologies used to recognize sign language. According to [8], there are two categories for these kinds of methodologies: Continuous Sign Language Recognition (CSLR) and Isolated Sign Language Recognition (ISLR).

In the case of MSL, there are different words and phrases that can be used to refer to the same objects/concepts, i.e., lexical variations. For instance, the position of objects or persons can be indicated by the dominant hand, followed by the main verb. This means that the non-dominant hand indicates the relative position of an object. Hand motion determines how an object executes the action. For example, as shown in Figure 1, the sign used to represent the CAMINAR (WALK) action, usually in a straight horizontal movement, stands for SALTAR (JUMP) if the hand performs a vertical rocking motion up and down using the palm of the non-dominant hand.

Sometimes objects or words lack a direct sign defined in the MSL. When this happens, the object is spelled out using the alphabet. However, certain signs need no spelling. It is sufficient to quote the first letter of the word to refer to the object. For example, to represent the sign LUNES (MONDAY), the letter L is combined with the gesture of the DIA_DE_LA_SEMANA (WEEKDAY). Alternatively, you can use a suffix to name negative shapes of objects. This involves moving the palm of the open hand downwards.

Pronouns use an index finger to identify the subject of the action. There are two methods of carrying out the action. When the reference is present and visible, the hand moves toward it. If the reference is neither present nor visible, an arbitrary reference address is subsequently chosen to constitute an agreement. Pronouns can be categorized into both manual pronouns and non-manual pronouns [9].

Manual pronouns consist of hand movements, usually with a specific number of fingers, so INDEX-1 would indicate the singular pronoun in the first person. Non-manual pronouns are composed of movements of the eyes, body, and hands; in one sign, the glance

provides additional information about the context of the affected object. For example, INDEX-3 looking to the right followed by the IR gesture indicates the person has gone.

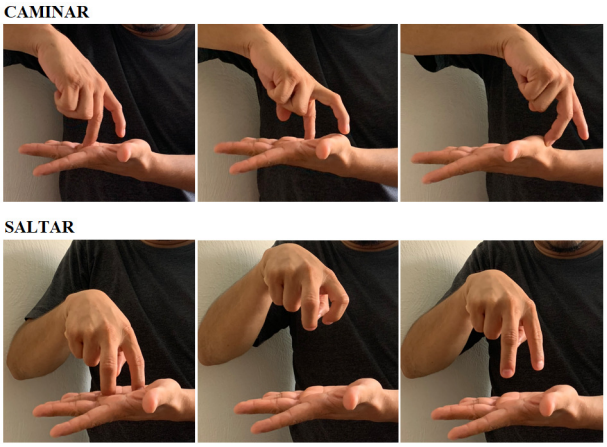

Figure 1. Movement of the hands for representing two different signs.

The fingers of the dominant hand serve to indicate the number of people who have been referred to. Thus, three fingers mean three people. The ownership of an object is represented by the sign of the dominant object, the spelling of the word D-E (OF), and the owned object. For example, for God's home, the gestures of CASA D-E DIOS should be displayed.

In order to better understand MSL, the researches in [10,11] have created support dictionaries. Serafin et al. [11] describes a lexicon of signs commonly used in MSL. The author uses pictures containing information on the configuration, location, motion, representation, and definition of each gesture. In the document, the signs are arranged in seven manual configurations represented by a letter where the shape of the letter is maintained when the sign is generated. The movements of the hands, face, and body complement the meaning of the sign. Moreover, Hawayek et al. [10] offers a bilingual dictionary divided into two sections: MSL-Spanish and the other way around, Spanish-MSL. This dictionary is made up of a glossary of labeled videos and each sign is associated with a word in Spanish. The access to the videos is via an online graphical interface.

According to the information that has been shown so far, it is possible to establish that there is currently no video dataset with lexicon variations in uncontrolled environments for the MSL and therefore this represents a research challenge.

The manuscript is organized as follows: Section 2 briefly describes the relationship that the MX-ITESO-100 dataset has with similar work. In Section 3, the novel Mexican sign language lexicon video dataset is introduced. In Section 4, the experimental procedures are presented. Section 5 presents the experimental results. Finally, in Section 6, some conclusions are drawn.

2. Related Work

Continuous Sign Language Recognition (CSLR) datasets are based on videos that contain a sequence of gestures rather than individual signs for the purpose of simulating real-world scenarios and testing dynamic continuous sign language recognition systems. Datasets from [2] have been produced in German, Chinese, Greek, and English. The German Sign Language datasets Phoenix-2014 and Phoenix-2014-T come directly from German TV and contain 1231 classes of nine participants with a resolution of 210 × 210 pixels. In addition, the RWTH-PHOENIX-Weather dataset is an extension containing 190 recordings directly from the weather channel which were collected between 2009 and 2010. The RWTH-BOSTON American Sign Language [12] dataset (ASL) has been designed as a subset of the BU-ASL corpus at Boston University and consists of 843 annotated classes in XML format that can be used to identify dynamic gestures in the same sequence. These videos were

recorded by a variety of people in controlled environments at a resolution of 210 × 60 pixels and are accessible to the general public. The research in [13] was based on two datasets described above: CSL and RWTH-PHOENIX. In this way, it offers a two-stage system for the recognition of video image sequences. First, it uses convolutional networks to extract functionality and second, it uses attention networks to automatically generate sentences.

Isolated Sign Language Recognition (ISLR) datasets are based on videos that contain sequences of single sign frames. It is important to mention that this publication has based its proposal on this category of dataset. The CSL-500 Chinese Sign Language dataset includes videos of 500 signs of 50 people captured in 2016. The ASL MS-ASL and WASL datasets contain videos distributed in 3000 signs of 341 participants from the YouTube platform using uncontrolled environments. The dataset for Turkish Sign Language contains videos captured in 2020 from 226 signs and 43 participants at a resolution of 512 × 1512 pixels. The Arabic Sign Language dataset was created in 2021 and consists of 502 signs from three participants. There are 75,300 video instances in total. Among today's most popular datasets is the ASL Lexicon Video Dataset [3], which belongs to the ISLR category and includes a total of 3800 color video signals within a controlled environment and at different angles. All the CSLR and ISLR datasets that have been discussed above are summarized in Table 1.

Table 1. Summary of CSLR and ISLR datasets.

Dataset	Description	Language	Signs
CSLR	RWTH-PHOENIX	German Sign Language	1231
CSLR	RWTH-BOSTON	American Sign Language	843
ISLR	CSL-500	Chinese Sign Language	500
ISLR	ASL MS-ASL and WASL	American Sign Language	3000
ISLR	AUTSL	Turkish Sign Language	226
ISLR	KArSL	Arabic Sign Language	502
ISLR	ASL Lexicon Video Dataset	American Sign Language	3800

There are authors in the literature who suggest a system to recognize sign language based on static gestures [5,14–16]. In other words, they use images instead of videos to train their models. For example, Fregoso et al. [14] uses a standard ASL dataset called ASL-MNIST that has a total of 34,627 images. It is an alphabet of 21 signs and color images obtained from 18 participants with a resolution of 32 × 32 pixels. This dataset was used to optimize a neural network with Particle Swarm Optimization (PSO), obtaining a better recognition rate than other investigations that used the same dataset. On the other hand, Alejandro et al. [15] experiments with a 21-sign MSL dataset using 10,500 images of both hands with a dimension of 1080 × 720 pixels from 10 participants. Finally, the research carried out in [16] implements an Indian Sign Language recognition system based on a dataset containing static alphabetic signs. There are 25 color images of the hands in a controlled environment and the model evaluation goes through the TensorFlow API for object detection.

As far as MSL semantics are concerned, it is not necessary to use concatenation as it is enough to follow a syntactic convention during sentence formation. Furthermore, people learning MSL generally have no problem using it throughout Mexico because the geographic and generational differences are relatively small. In [17], a comparison of 100 words from different parts of Mexico was carried out in order determine the variations of each one. No less than 50 words have the same meaning and around 90% share a lexical identity. For this reason, MSL has been shown to have the linguistic components to be considered a language or dialect. MSL lexicon resembles ASL. In [18], it was established that MSL is closely related to ASL. However, [19] finds a significant difference when using MSL compared to ASL. A quick lexical comparison of 100 signs in both MSL and ASL shows that 12 out of 100 signs in ASL need an initializer, while 37 out of 100 signs need an initializer in MSL. Signs for initializing use the first letter of the word in the corresponding language before executing the gesture. In [9], MSL spelling and initialization has been discussed.

Finger spelling is not used as much in MSL as it is in ASL. However, it is regularly used to express the names of people and places when one of the participants does not know the sign. MSL lexicon consists of native and non-native vocabulary, where words with initializers are considered native. K. Faurot found that 14% of deaf Americans who know ASL can recognize MSL gestures. Although this is positive, it is far from being an open and universal language. It is important to mention that understanding the Spanish language is not required to use MSL. Despite this, there are important grammatical differences such as verb conjugation due to MSL having no verb tenses. These are agreements that people have come to. In addition, there is a concordant agreement that verbs follow when they relate them to other objects. In the case of a verb with an argument, this indicates what or who it is. For verbs that have two arguments, the first indicates the consistency of the object of the action and the second indicates that it is affected.

3. MX-ITESO-100

The MX-ITESO-100 dataset is a meticulously curated collection of 5000 videos, thoughtfully organized into 100 distinct folders. This dataset stands out for its comprehensive coverage, capturing a wide range of signs and expressions representative of the selected Mexican lexicon. Each folder within the dataset is dedicated to a specific sign, ensuring a coherent and systematic arrangement. Within each folder, one can find a consistent compilation of 50 videos, all pertaining to the same sign. The videos themselves were meticulously generated, involving the collaborative efforts of two individuals with expertise in sign language. Moreover, it is important to mention that the MX-ITESO-100 dataset serves as a valuable resource for various domains and applications. Researchers, linguists, and computer vision experts alike can leverage this extensive collection to advance the development of sign language recognition systems, gesture-based interfaces, and other related technologies. Furthermore, its incorporation of the Mexican lexicon ensures the dataset's relevance and applicability within the cultural context it represents.

The MX-ITESO-100 dataset is consistent with other authors and provides a video dataset for the most representative words. Furthermore, it is oriented to MSL, with the intention of being used in a real-time recognition system. In [20], a MSL dataset has been established in order to identify dynamic gestures using support vector machine (SVM) and it is composed of videos recorded by 22 people in a controlled environment. This research uses the segmentation technique through which geometric features of both hands are extracted. Although the similarities are considerable, our MX-ITESO-100 dataset explores the inclusion of a rich and varied lexicon. In order to do this, it is necessary to generate ISLR content under uncontrolled conditions similar to real-world viewing environments. Lexicon is essential for creating a sign language dataset. MX-ITESO-100 contains the most basic grammatical elements that allow us to create simple but complete ideas, such as nouns, pronouns, verbs, adjectives, adverbs, conjunctions, and commonly used phrases. Due to the fact that there is great variation in the construction of a model, MX-ITESO-100 presents different versions for the same gesture. Currently, the dataset is being oriented towards research in care networks. However, this does not restrict the fact that it can be shared publicly for benchmarking.

The lexical structure for establishing the dataset includes one hundred signs of grammatical elements which have been distributed as shown in Table 2.

Each video recording is related to the representation of an idea by the motion of the hands and the body and facial expressions. In addition, the actions are carried out over a period of time, when the final positions of the hands or body or the facial expression vary with respect to the initial signal. Therefore, the dynamic gestures consist of two stages to establish an idea and each stage defines the specific configuration of the movement, i.e., the dynamic gesture is always performed from an initial stage to a final stage, as shown in Figure 2.

Table 2. Distribution of grammatical elements in MX-ITESO-100.

Grammatical Element	Quantity
Verbs	30
Adjectives	29
Nouns	25
Adverbs	6
Pronouns	5
Phrases	4
Conjunctions	1

Figure 2. Initial and final state of a dynamic gesture.

This proposal was based on dynamic actions as opposed to the representations of ideas from unique states such as static gestures. However, we have previously established datasets with static signs that have represented each letter of the alphabet.

It is important to mention that the selection of the lexicon was based on the first 100 words/signs defined by a small inclusive community. This community determined the minimum number of signals required for effective communication in an active environment. The MX-ITESO-100 dataset contains at least fifty different versions of the same dynamic movement for the same sign. Differences include variations in hand positions, dominant hand selection, body positions, and facial expressions, as shown in Figure 3.

Figure 3. Different characteristics of the environment and the participants.

Regarding the format of video recordings in digital files, we have chosen Moving Picture Experts Group (MPEG-4) because it is an international standard for audiovisual coding according to ISO/IEC 14496-12:2001. Additionally, the frame size for each sample in the video is 512 × 512 pixels, the recording duration is five seconds with thirty images per second, and finally, the average size of each video in MPEG-4 format is approximately 4 MB.

As it can be seen in Table 2, the grammatical elements represent different percentages with respect to the general structure of the lexicon. Next, the single words of each gram-

matical section are shown in boxes for the Spanish (top) and English (bottom) versions. Firstly, thirty percent of the lexicon consists of signs that represent the actions of objects. This is the biggest group of grammatical items called verbs. As part of that group, it has the following elements.

ABRIR OPEN	APARTAR SEPARATE	APRENDER LEARN	AYUDAR HELP	BAILAR DANCE
BESAR KISS	BORRAR DELETE	COCINAR COOK	COMER EAT	CONSTRUIR BUILD
CONTAR COUNT	CORRER RUN	DAR GIVE	ENCENDER TURN ON	ENTRAR ENTER
ESCRIBIR WRITE	ESTAR BE	EXPLICAR EXPLAIN	GUSTAR LIKE	HABLAR TALK
HACER DO	JUGAR PLAY	LEER READ	MANEJAR DRIVE	OLER SMELL
PASEAR SIGHTSEEING	PONER PUT	SALTAR JUMP	TOSER COUGH	VER SEE

After verbs, the second largest group of signs in the dataset corresponds to the adjective grammar element, which has the purpose of altering nouns. There are twenty-seven signs in this group.

AZUL BLUE	BLANCO WHITE	BONITO NICE	BUENO GOOD
CAFÉ BROWN	CANSADO TIRED	CHISTOSO FUNNY	CLARO CLEAR
COQUETO FLIRTATIOUS	DURO HARD	EGOÍSTA SELFISH	ENANO DWARF
FÁCIL EASY	FELIZ HAPPY	FEO UGLY	FLACO SKINNY
FRÍO COLD	GORDO FAT	GRANDE BIG	GRIS GRAY
LARGO LONG	MAL EVIL	MORENO BROWN	NEGRO BLACK
OSCURO DARK	SEDIENTO THIRSTY	TRISTE SAD	

The third group of signs is established by nouns and it represents the fourth part of the dataset with twenty-five signs. Nouns designate objects or abstractions independently such as some colors and days of the week. These items are as follows.

AGUA WATER	ANARANJADO ORANGE	AVIÓN PLANE	BANDERA FLAG	BICICLETA BIKE
CALOR HEAT	COLOR COLOR	DEPORTE SPORT	DERECHA RIGHT	DÍA DAY
DOMINGO SUNDAY	JUEVES THURSDAY	LUNES MONDAY	MARTES TUESDAY	MIÉRCOLES WEDNESDAY
MUJER WOMAN	NIÑO CHILD	NOCHE NIGHT	PERRO DOG	PIANO PIANO
QUESADILLA QUESADILLA	SÁBADO SATURDAY	TELEVISIÓN TV	VIERNES FRIDAY	ZAPATO SHOE

In the same way as adjectives, adverbs lead to the verb or adjective and change the idea. Eight adverbs have been incorporated into the lexicon and they are the following.

AQUÍ HERE	AYER YESTERDAY
BIEN GOOD	COMO HOW
HOY TODAY	MAÑANA TOMORROW
MUCHO MUCH	POCO LITTLE

A pronoun is the lexicon structure that supersedes the name. It is important to mention that the selection of pronouns for this lexicon is limited to personal pronouns. Possessive, undefined, relative, and interrogative pronouns are discarded. Personal pronouns are as follows.

EL HE	ELLA SHE	ELLOS THEY	TU YOU	YO I

According to MSL, pronouns are associated with visible and present objects, where the final stage indicates position in relation to the object. For the MX-ITESO-100 dataset, the signs belonging to this grammatical element indicate the general idea of the agreed relationship between the speaker and a non-visible person. This must be taken into account by researchers for the generation of new models.

Furthermore, in MSL, conjunctions are not frequent because the juxtaposition method suppresses the conjecture and shows nouns. The juxtaposition method in MSL conjunctions involves combining two or more separate signs without any specific transition or linking element. It relies on the spatial arrangement and sequencing of the signs to convey the intended meaning and the relationship between the ideas or concepts being expressed. However, our approach includes a conjunction of the subordinating type through the alphabet letter "Y" in a gesture that requires movement, and therefore it has been classified as a dynamic rather than a static gesture. Finally, the dataset includes dynamic gestures that refer to ideas based on compound words. These expressions are used even by non-deaf people and they are as follows.

GRACIAS THANK YOU	HOLA HELLO
LENGUA DE SEÑAS MEXICANA MEXICAN SIGN LANGUAGE	POR FAVOR PLEASE

With the lexicon of the MX-ITESO-100 dataset, it is possible to recognize sentences using a basic and simple grammatical structure such as the single phrase consisting of

SUJETO + VERBO + PREDICADO
SUBJECT + VERB + PREDICATE

For example, the simple sentences that can be generated or recognized with the lexicon are the following

MAÑANA NOCHE CAMINAR PERRO

This set of signs represents the idea of the sentence: Tomorrow night I will walk with my dog. Additionally, the sentence

HOLA YO APRENDER LENGUA_DE_SEÑAS_MEXICANA FÁCIL

represents the expression: Hello, I am learning MSL very easily.

Finally, a public preview version of the MX-ITESO-100 video dataset for MSL has been uploaded at the following link: https://acortar.link/J0XzZb (accessed on 17 August 2023).

4. Experiments

It is important to emphasize that the principal purpose of this work is not to conduct in-depth experiments on the validation of the MX-ITESO-100 dataset because parallel research is being carried out to insert the data into a machine learning model using neural and attentional networks. This research, in turn, is grounded in ISLR, where the objective is to continuously recognize signs. Furthermore, this investigation aims to extend the capabilities of ISLR by implementing advanced techniques for enhanced sign recognition and interpretation. This endeavor seeks to contribute to the broader field of human–computer interactions and gesture recognition, advancing our understanding of continuous sign language communication.

The experimental evaluation in this work is based on [21]. This research uses three types of recurrent neural network architectures as part of the model, such as a [22] Recurrent Neural Network (RNN) and Long-Short Term Memory (LSTM). In addition, performance measures such as precision, recall, and accuracy are established. According to the results of research, the RNN tends to overfit with fewer units and the highest accuracy score was for the LSTM architecture with 97%.

On the other hand, [8] provides a complete discussion of the methods for carrying out the evaluation of the various Continuous Sign Language Recognition (CSRL)-type datasets. This research examines the following architectures: Two- and Three-Dimensional Convolutional Neural Networks (2D-CNN, 3D-CNN) [23], Bidirectional LSTM Networks (BLSTM) [24], Connectionist Temporal Classification (CTC) [25], a Multi-Cue Spatio-Temporal Network (STMC) [26], and a Sparse ReguLarized Generative Adversarial Network (SRLGAN) [27]. Moreover, this research measures the number of transactions required to transform a string of words using the Word Error Rate (WER). According to the results, the most efficient model is the STMC. However, CNNs are easier to implement with very similar results, and therefore they have been integrated into our model.

The validation process for the MX-ITESO-100 dataset consists of a simple two-step model. The first step implements a CNN model that extracts features from individual frames, e.g., a frame sequence from a video recording. For each frame, there is a sample that is processed independently through a layer distributed in time to later connect to an LSTM network, as shown in Figure 4. Next, in model training, common standard architectures have been exchanged as feature extraction networks, known as core networks. This implementation was carried out using version 2.9.1 of TensorFlow in the Python programming language. For all cases, the initial values of the neural weights were already established, since they come from previous training on the ImageNet dataset. At this point, the architectures that have been assessed are Xception [28], VGG16 [29], VGG19 [29], ResNet50 [29], ResNet101 [30], ResNet152 [30], MobileNet [31], DenseNet121 [32], and EfficientNetV2B0 [33]. In all cases, the dense layer at the end of the network was removed and, in its place, a Flatten-type layer was connected to feed the recurring layers of the next step.

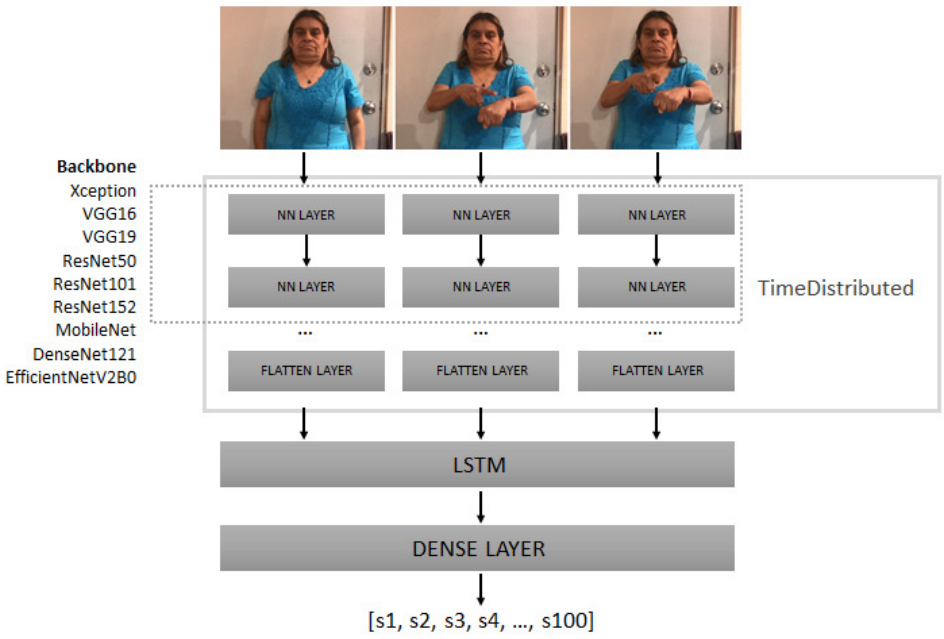

Figure 4. Neural network model for the MX-ITESO-100 dataset validation.

Now, in the second step, a recursive network model for image sequencing is established. An LSTM architecture is chosen due to the results in [21]. This architecture uses a Rectifying

Linear Unit (ReLU) as an activation function. Subsequently, the output of the network is connected to a dense layer that uses the softmax activation function with one hundred neurons. Each of the neurons in the output phase represents a sign from the lexicon of the MX-ITESO-100 dataset.

In this work, all the images were collected from the beginning to the end of the video recording in order to process them through a set of specific Python classes (signs) without discrimination between frames. The sizes of the individual samples were formatted, modified, and converted to a Numpy array. Subsequently, the data structure is incorporated into the model. The training model employed for this dataset adopts the Pareto distribution as its foundation, following a well-established statistical framework. This distribution methodology ensures a systematic division of videos within each sign, allocating approximately 80% of the videos for training purposes and reserving the remaining 20% for validation. It is important to note that both the training and validation sets are derived from a shared pool of 5000 meticulously recorded videos, ensuring consistency and continuity throughout the dataset. Specifically, out of the total 5000 videos, 4000 videos are dedicated to the training set, while the remaining 1000 videos form the validation set. Additionally, it is worth highlighting that people performing sign language are present in both the training and validation sets, enhancing the realism and authenticity of the dataset. Both sets encompass signs from all 100 words contained within the dataset.

For the purpose of carrying out predictions, the most common way to evaluate the performance of a supervised model is through the identification of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FP). TP represents the positive prediction that the model successfully classifies, for example, the model classified the sign ZAPATO within the shoe sign. TN means a negative prediction that the model has successfully classified, e.g., the model did not classify the sign ZAPATO within the shoe sign. FN depicts a negative prediction of an actual value, for example, the model has classified the sign ZAPATO within the piano sign. Finally, FP represents a positive prediction where the actual value is negative, e.g., a sign different from ZAPATO than the model has been classified within the shoe sign.

The parameters for evaluating our model are the same as those used in [21], such as accuracy, recall, and precision. The mathematical expressions for these parameters are shown below.

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{1}$$

The precision metric shown in (1) is the proportion of positive predictions in the total of positive and negative predictions made by the model during sign classification.

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

The recall metric shown in (2) is the proportion of positive predictions that the model successfully classified in the total number of positive and negative predictions. That is, the number of signs that our model qualifies as positive.

$$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{3}$$

The accuracy metric shown in (3) is the proportion of predictions performed correctly by the model regarding the total number of correct and incorrect predictions. That is, how many signs have been successfully classified [34].

It is important to mention that it is not necessary use the F1 classification to evaluate the performance because the intention is to identify the level of precision and recall individually.

5. Results

Measurements of accuracy, precision, and recall have been used in the training process of the model. Training and validation curves are shown in Figure 5.

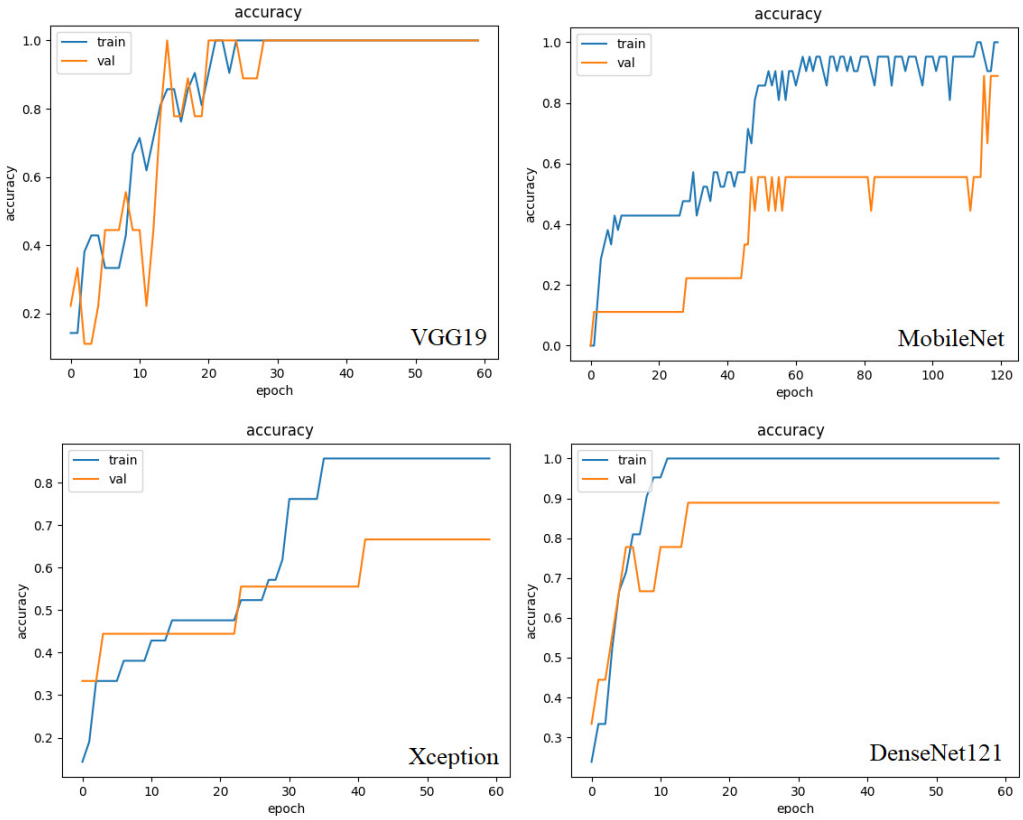

Figure 5. Training and validation curves.

According to the results shown in Figure 5, the best classification performance was achieved using the VGG19 backbone. In this case, the precision measurement value for the training and validation set was greater than 99%. Approximately 30 epochs were required for convergence to this value. Most of the backbone networks converged in 50 epochs. In most cases, simple minor adjustments have been made. However, by default, a batch size of 10, 60 epochs, and an RMSprop optimizer with a training ratio of 5×10^{-5} were selected.

The accuracy of the Mobilenet backbone reached 99% for the training set and above 93% for the validation set. A total of 120 epochs were necessary for convergence in these values. Similarly, the DenseNet121 backbone converged to 99% in the training set. However, the validation set does not exceed 88% accuracy. Unlike MobileNet, the DenseNet121 core network reached stabilization in 18 epochs.

Backbone networks such as Xception and the Residual Neural Network (ResNet) achieved few useful results. For example, the accuracy of the ResNet50 backbone was only 78% in the training set and 35% in the validation set. Despite this, the curve shows a downward trend. The EfficientNetV2B0 backbone shows the worst training behavior. It was unable to converge towards a value greater than 23% despite an increase in the number of epochs and a considerable reduction in time losses.

Finally, the VGG19 backbone was used as a reference for our two-step model. For a total of 5000 videos, the TPS number is 4956, 25 fn and 19 fp. Moreover, the accuracy is 99.12%, the precision is 99.62%, and the recall is 99.5%.

6. Conclusions

Results from the evaluation of the MX-ITESO-100 dataset are similar to the accuracy values reached by other authors. I. Papastratis presented the most significant results using the LSA64 [8] dataset. Through an early comparison, it is possible to establish that the performance values oscillate between 98.09% and 99.9% and therefore the results are similar to our approach. The experimental procedure demonstrated the feasibility of the MX-ITESO-100 dataset for training models based on machine learning. In addition,

dynamic gesture recognition is possible in uncontrolled environments. This provides researchers with an additional resource in order to validate their models in MSL dynamic gesture recognition systems.

The MX-ITESO-100 dataset includes a variety of grammatical items for future research. This is because there is not just one type of grammatical element but the lexicon assists in the enrichment of the semantical process through the generation of complete sentences. The principal contribution of this research is based on the exploration of a novel Mexican sign language lexicon video dataset which represents the first step towards more complete future versions. Future work will be focused on further development of this dataset to include new grammatical elements, additional versions of current signs, more videos recorded in various locations in Mexico, and regionalisms.

Finally, the MX-ITESO-100 dataset offers indirect benefits for inclusivity since it provides a basis for the development of inexpensive devices that facilitate communication with deaf people and therefore forges a more committed and inclusive Mexican society.

Author Contributions: Conceptualization, V.M.-S. and I.V.-T.; Data curation, F.C.-Á.; Methodology, V.M.-S., I.V.-T., F.C.-Á. and C.H.-M.; Formal analysis, V.M.-S., I.V.-T., F.C.-Á. and C.H.-M.; Investigation, V.M.-S., I.V.-T., F.C.-Á. and C.H.-M.; Resources, F.C.-Á. and C.H.-M.; Visualization, V.M.-S., F.C.-Á. and C.H.-M.; Writing-original draft preparation, V.M.-S. and I.V.-T.; Writing-review and editing, V.M.-S., I.V.-T., F.C.-Á. and C.H.-M.. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Instituto Tecnológico de Estudios Superiores de Occidente.

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

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://acortar.link/J0XzZb (accessed on 17 August 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mejía-Pérez, 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. *Appl. Sci.* **2022**, *12*, 5523. [CrossRef]
2. Dreuw, P.; Neidle, C.; Athitsos, V.; Sclaroff, S.; Ney, H. Benchmark Databases for Video-Based Automatic Sign Language Recognition. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 26 May–1 June 2008; pp. 1115–1120.
3. Athitsos, V.; Neidle, C.; Sclaroff, S.; Nash, J.; Stefan, A.; Yuan, Q.; Thangali, A. The American Sign Language Lexicon Video Dataset. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
4. Sosa-Jiménez, C.O.; Ríos-Figueroa, H.V.; Rechy-Ramírez, E.J.; Marín-Hernández, 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), Ixtapa, Mexico, 8–10 November 2017; pp. 1–6.
5. Carmona-Arroyo, G.; Ríos-Figueroa, H.V.; Avendaño-Garrido, M.L. Mexican Sign-Language Static-Alphabet Recognition Using 3D Affine Invariants. In *Machine Vision Inspection Systems, Volume 2: Machine Learning-Based Approaches*; Wiley: Hoboken, NJ, USA, 2021; pp. 171–192.
6. Espejel-Cabrera, J.; Cervantes, J.; García-Lamont, F.; Ruiz Castilla, J.S.; Jalili, L.D. Mexican sign language segmentation using color based neuronal networks to detect the individual skin color. *Expert Syst. Appl.* **2021**, *183*, 115295. [CrossRef]
7. Ray, S. An Analysis of Computational Complexity and Accuracy of Two Supervised Machine Learning Algorithms—K-Nearest Neighbor and Support Vector Machine. In *Data Management, Analytics and Innovation*; Sharma, N., Chakrabarti, A., Balas, V.E., Martinovic, J., Eds.; Springer: Singapore, 2021; pp. 335–347.
8. Papastratis, I.; Chatzikonstantinou, C.; Konstantinidis, D.; Dimitropoulos, K.; Daras, P. Artificial Intelligence Technologies for Sign Language. *Sensors* **2021**, *21*, 5843. [CrossRef] [PubMed]
9. Plumlee, M. Pronouns in Mexican Sign Language. *Work. Pap. Summer Inst. Linguist.* **1995**, *39*, 81–92. [CrossRef]

10. Hawayek, A.; Del Gratta, R.; Cappelli, G. A Bilingual Dictionary Mexican Sign Language-Spanish/Spanish-Mexican Sign Language. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 17–23 May 2010; pp. 3055–3060.
11. Serafin de Fleischmann, M.E.; Gonzalez Perez, R. *Manos con voz. Diccionario de Lengua de Señas Mexicana*; CONAPRED: Ciudad de México, Mexico, 2011.
12. Forster, J.; Schmidt, C.; Koller, O.; Bellgardt, M.; Ney, H. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; pp. 1911–1916.
13. Huang, J.; Zhou, W.; Zhang, Q.; Li, H.; Li, W. Video-based Sign Language Recognition without Temporal Segmentation. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32, pp. 2257–2264.
14. Fregoso, J.; Gonzalez, C.I.; Martinez, G.E. Optimization of Convolutional Neural Networks Architectures Using PSO for Sign Language Recognition. *Axioms* **2021**, *10*, 139. [CrossRef]
15. Alejandro, S.M.; Antonio, N.C.J. 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), Morelia, Mexico, 9–11 August 2021; pp. 1–7.
16. Srivastava, S.; Gangwar, A.; Mishra, R.; Singh, S. Sign Language Recognition System Using TensorFlow Object Detection API. In *Advanced Network Technologies and Intelligent Computing*; Communications in Computer and Information Science; Springer International Publishing: Cham, Switzerland, 2022; pp. 634–646.
17. Bickford, J.A. Lexical Variation in Mexican Sign Language. *Sign Lang. Stud.* **1991**, *72*, 241–276. [CrossRef]
18. Quinto-Pozos, D. Sign language contact and interference: ASL and LSM. *Lang. Soc.* **2008**, *37*, 161–189. [CrossRef]
19. Faurot, K.; Dellinger, D.; Eatough, A.; Parkhurst, S. The identity of Mexican sign as a language. *J. Lang. Surv. Rep.* **2000**, *1*. Available online: https://www.sil.org/resources/archives/9069 (accessed on 18 August 2023)
20. Cervantes, J.; García-Lamont, F.; Rodríguez-Mazahua, L.; Rendon, A.Y.; Chau, A.L. Recognition of Mexican Sign Language from Frames in Video Sequences. In *Intelligent Computing Theories and Application*; Springer: Cham, Switzerland, 2016; pp. 353–362.
21. Solís, F.; Martínez, D.; Espinoza, O. Automatic Mexican Sign Language Recognition Using Normalized Moments and Artificial Neural Networks. *Engineering* **2016**, *8*, 733–740. [CrossRef]
22. Staudemeyer, R.C.; Morris, E.R. Understanding LSTM - a tutorial into Long Short-Term Memory Recurrent Neural Networks. *arXiv* **2019**, arXiv:1909.09586.
23. Kurmanji, M.; Ghaderi, F. A Comparison of 2D and 3D Convolutional Neural Networks for Hand Gesture Recognition from RGB-D Data. In Proceedings of the 2019 27th Iranian Conference on Electrical Engineering (ICEE), Yazd, Iran, 30 April–2 May 2019; pp. 2022–2027.
24. Syed, F.; Sipio, R.D.; Sinervo, P. Bidirectional Long Short-Term Memory (BLSTM) neural networks for reconstruction of top-quark pair decay kinematics. *arXiv* **2019**, arXiv:1909.01144.
25. Li, H.; Wang, W. A Novel Re-weighting Method for Connectionist Temporal Classification. *arXiv* **2019**, arXiv:1904.10619.
26. Zhou, H.; gang Zhou, W.; Zhou, Y.; Li, H. Spatial-Temporal Multi-Cue Network for Continuous Sign Language Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
27. Shah, A.A.; Venkateswara, H. Sparsity Regularization For Cold-Start Recommendation. *arXiv* **2022**, arXiv:2201.10711.
28. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2016; pp. 1800–1807.
29. Mascarenhas, S.; Agarwal, M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In Proceedings of the 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), Bengaluru, India, 19–21 November 2021; Volume 1, pp. 96–99.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
31. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
32. Huang, G.; Liu, Z.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2016; pp. 2261–2269.
33. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946.
34. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013.