

Two Hands Static Mexican Sign Language Dataset for Object Detection

Víctor Martínez-Sánchez, Ivan Villalon-Turrubiates,
Francisco Cervantes

Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO),
Tlaquepaque, Jalisco
Mexico

{ng683728, villalon, fcervantes}@iteso.mx

Abstract. Currently, there is a trend to use machine learning and deep learning techniques for image classification and object detection into computer-vision systems. Creating and training such models require a considerable amount of images. Gathering hand-gesture data for sign language is not an easy task for researchers because unlike popular belief, sign language is not universal. This is partly due to there are no standard signs for all languages. Hence, these datasets must be created according to a specific language. In this paper we propose a Mexican Sign Language (MSL) dataset that includes 21 static hand-gestures using both hands, each representing a letter in the Spanish alphabet. The dataset consists of 10,000 high resolution colored images in JPG format captured at different environments. The dataset is evaluated in a deep learning model with an overall accuracy of 99.46%.

Keywords: Mexican sign language, dataset, hand-gestures, computer vision.

Conjunto de datos de lengua de señas mexicana estática con dos manos para la detección de objetos

Resumen. Actualmente, existe una tendencia a utilizar técnicas de aprendizaje automático y aprendizaje profundo para la clasificación de imágenes y la detección de objetos en sistemas de visión por computadora. La creación y el entrenamiento de dichos modelos requieren una cantidad considerable de imágenes. La recopilación de datos sobre gestos manuales para la lengua de señas no es una tarea sencilla para los investigadores, ya que, contrariamente a la creencia popular, la lengua de señas no es universal. Esto se debe, en parte, a que no existen señas estandarizadas para todos los idiomas; por consiguiente, estos conjuntos de datos deben crearse en función de un idioma específico. En este artículo, proponemos un conjunto de datos de Lengua de Señas Mexicana (LSM) que incluye 21 gestos manuales estáticos realizados con ambas manos, cada uno de los cuales representa una letra del alfabeto español. El conjunto de datos consta de 10.000 imágenes en

color de alta resolución en formato JPG, capturadas en diversos entornos. Dicho conjunto de datos se evalúa mediante un modelo de aprendizaje profundo, alcanzando una precisión global del 99,46 %.

Palabras clave: Lengua de señas mexicana, conjunto de datos, gestos manuales, visión por computadora.

1 Introduction

According to the World Health Organization (WHO) in 2018 [1], over 6.1% of the world's population, around 466 million people, has disabling hearing loss; 432 million adults (93%) and 34 million children (7%). Approximately one third of people over 65 years of age are affected by disabling hearing loss. It's estimated that by 2050 over 900 million people, one in every ten people, will have disabling hearing loss. Projections were calculated based on the WHO 2008 prevalence estimates.

Most people with disabling hearing loss live in low- and middle-income countries. México is one of the countries included in this groups. According to the 2018 Encuesta Nacional de Dinámica Demográfica (ENADID) results [2], In México 6% (7.1 million) have disabilities where 33% (2.4) million have disabling hearing loss.

Mexico is a developing country. Special education and inclusion in Mexico has been increasing over the last years. It started with an inclusive education movement based on the special education system. In 1971 the educational reform proposed to create the general direction of special education (Dirección General de Educación Especial, DGEE) [3]. This entity was responsible for implementing an educational subsystem to support people with disabilities and outstanding capacities. In 1996, a national project for educational integration in the country was proposed [3].

The results were not encouraging. The study revealed deficiencies in the inclusion methods. Mexico is still so far from having a success inclusion and reasons can be categorized in two main points: teacher training on inclusion techniques and the tools used in the process. If we focus on inclusion tools, we notice that enough solutions do not exist which allow communication with deaf people. Current tools tested in Mexico consist of electronic devices based on instrumented gloves with flexion sensors [4].

However, in the end, they are tools that involve a cost for a society that, in most cases, does not have sufficient resources to implement these solutions. On the other hand, society is moving towards new technologies. The main attributes are: practicality, simplicity and low cost.

So we understand there exists a need to design devices to translate sign language into a regional language. This article is intended to be the basis for training deep learning models in the future. As we know, one of the requirements to train models is to have a reliable dataset based on different people and images in multiple environments. Images shown in this article were not captured in a controlled environment. This improves assertiveness in recognizing gestures by getting closer to reality.

In the following sections we will delve into the state of the art and compare the related works through a table. Later, we will describe in detail our work in the dataset section, the main challenges and advantages against other datasets. We will detail the

Table 1. Sign language datasets used in recent articles.

Name	Dataset Type	Language	Details
LSA64 [8]	video	Argentinian	64 signs
ASLLVD [9]	video	American	3000 signs
DSGS [12]	video	Swiss German	100 lexical items
ISL-HS [13]	video	Irish	23 hand-shapes
MU HandImages [10]	images	American	36 gestures
BdSL [11]	images	Bangladesh	36 characters
ArSL2018 [14]	images	Arabic	32 signs
THISL [16]	images	Indian	26 gestures
MSL [15]	video	Spanish	249 words

necessary steps followed in evaluating the dataset in a deep learning model. Finally, we will conclude the paper by talking about future work.

2 Related Work

Sign language image recognition begins with one of the early pioneers Jason Isaacs who worked on pattern classification for American Sign Language (ASL) using the wavelet transform [5]. Sign language is a non-verbal language used by deaf people to communicate with others which involves challenges and difficulties for researchers [6]. In addition, people usually use more than one sign to represent the same word. This causes conflicts and little progress in standardization, even within the same region. Despite the government's effort to include deaf people in society, unfortunately very few people can communicate in sign language. Solutions such as the use of interpreters is limited to the high cost that makes difficult to implement it in a country like Mexico.

Commonly deaf people have trouble reading and writing. Therefore, inclusion towards the digital age is not easy for them. An example of the above is multimedia content. It is easy to notice the increase in the number of videos.

However, the content is focused on hearing people. To improve communication between deaf people and hearing people, researchers are needed to develop sign recognition systems. The goal is generating mobile devices accessible to all types of people. In Mexico there is no comprehensive solution. Some of the current work is limited to applications that detect, for example, junction points on the hands [7]. Hence, we are still far from a robust comprehensive sign language recognition system.

A previous step to the development of these applications is the generation of a dataset. Machine learning needs to be powered by thousands of images. In addition to the volume in the number of images, signs involve variability in captures such as lighting conditions, posture, participants, environment, angles, and rotation [8]. Additionally we need to append an extra variable, the language. Table 1 provides an overview of the main latest works.



Fig. 1. Images of MSL static dataset for Spanish alphabet.

These works can be classified into two categories: video and image datasets. Video datasets generally are focused on detecting words rather than static signs. In most cases RGB conventional cameras are used as the capture source [8-12,14]. Other solutions use depth cameras like those of Kinect [12]. In most cases, images datasets contribute as an initial effort for video datasets. Then, the first step is to create a static dataset. Our initial effort is to replicate the work proposed in [16] where they create a two hand indian sign language (THISL) dataset.

3 Dataset

Currently datasets generated by researchers around the worlds are focus on the ASL for image. There are very few public datasets that use the MSL dictionary as a reference. We generated a dataset for object detection which contains 10,000 images based on 21 static gestures of the MSL alphabet. The images have a dimension of 1920x1080 pixels. They are 24-bit deep images stored in JPG format with an approximate weight of 270KB each. Only 21 out of 27 signs (letters) were included as part of the MSL dataset. It is because some letters require hand movement and face pose to generate the sign.



Fig. 2. Letter A from the MSL dataset performed by a girl.

MSL dataset was generated using both hands as depicts in Figure 1. In the first collage image, the fingers are positioned on the palm and only the thumb is open pointing to one side. This represents the letter A in the MSL. Similarly, each of the subsequent images represents a specific sign of the alphabet. These images were taken on multiple locations and different environments.

The image capture process was implemented through cycles. Each cycle consisted of capturing twenty-five images with the right hand and twenty-five images with the left hand from a person, making a single sign, in a previously selected environment. 10 environments in total. Each of them with distinct attributes, highlighting the following: natural versus artificial light, vegetation versus furniture, paintings and home decoration, background with stone wall versus background with a plain wall in a single color and two colors, windows with bars, light exposure tiers, and objects on the wrist. Inclination and rotation was forced in both hands to ensure the sign is recognized independently of the angle of the hand. The idea was to include the natural hand movement. Six people in total participated remotely building this dataset. Images were stored into a cloud-based repository where later the images were downloaded for training.

The twenty one images displayed in Figure 1 are a sample of the full MSL dataset. They are cropped to display only the hand-gesture. However, the original images cover more space than just the one used by the hands, as shown in figure 2.

We ran into difficulties when we generated the MSL dataset. At the beginning, it was not possible to know precisely the moment when an image was captured.

That generated displaced, blurred or poorly formed images as well we faced multiple misunderstandings capturing gestures because people were mistaken for the sign. Also, after some time capturing gestures, we lost the image count.

Hence, additionally we implemented a custom solution to capture images through a solution based on Python programming language. It's a simple algorithm that triggers image captures using configurable parameters. e.g. capture time slot, total required captures and a box to display the current gesture and hand side. Each captured image was stored into a custom folder using the gesture name as reference.

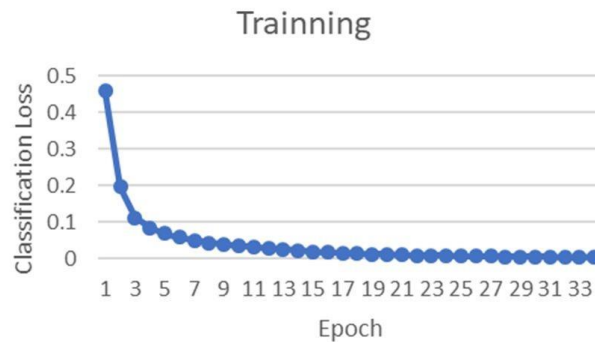


Fig. 3. Classification loss versus 33 epochs implemented with the resnet152 as backbone in the training over the MSL dataset.

The dataset is focused on object detection. Images themselves contain a hand-gesture from the Mexican alphabet. However, it is necessary to know the location of the gesture within the image in the training stage. In our case, each image is linked to an annotation file saved as XML that includes bounding box detailed information. The selected format for annotation files is Pascal VOC. Unlike the COCO annotation format [17], It generates an annotation file per image which makes it easier to grow and maintain the dataset. Additionally, the image label corresponding to represented hand-gesture is also included in the file. The annotations were made using the LabelImg [18] tool which is a graphical solution created in Python and the open-source widget toolkit for creating graphical user interfaces QT [19].

4 Experiments and Results

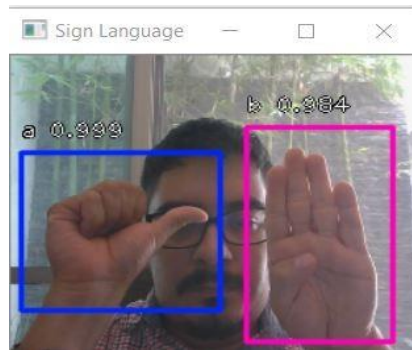
A model was trained using RetinaNet, presented by Facebook AI Research (FAIR) [20]. RetinaNet is one of the most recent deep learning models. It allows object detection in real time. In the same way as other models such as YOLOV3 or Mask RCNN, RetinaNet exhibits a robust ability to identify objects in a single stage. Moreover, RetinaNet shows a significant accuracy rate along with higher detection rate [21].

It was implemented on a virtual machine hosted in the cloud. It contains eight Tesla V100 cards destined to the high-performance computing (HPC) field. Each of these graphic processing units has 5,120 Compute Unified Device Architecture (CUDA) cores capable of developing a Power of 7.5 Floating point operations per second (TFLOP) in double-precision floating-point format (FP64). The full capacity of the virtual machine is 768GB of RAM.

The training consisted on 33 iterations of approximately 43 minutes each. To train the model, the MSL dataset was divided into two parts; the first with 9,000 images for training and the second with 1,000 images used for validation. The training was performed using a deep residual neural called resnet152 [23] that is based on convolutional neural networks (CNN) with 71,511,192 configurable parameters in total. CNN are especially useful for locating patterns in images in order to recognize

Table 2. First epoch results using distinct backbones to train a deep learning model using the MSL dataset.

Backbone	Color/Gray	Epoch Time	Loss	Classification Loss
resnet50	color	25:34	1.6276	0.4773
resnet50	gray	25:33	1.6257	0.4684
resnet152	color	42:39	1.5623	0.4585
vgg16	color	29:39	3.2998	2.1012
densenet201	color	1:37:40	2.1113	0.5945

**Fig. 4.** Custom implemented application to predict hand-gesture using a webcam.

objects, hand-gestures in our case. In the end, the accuracy achieved after 33 epochs was 99.46% as depicts in Figure 3.

At the beginning of the training, we tried out this MSL dataset with distinct backbones. Only one epoch was executed due to time limitation with the server. This is because, training these kind of models like RetinaNet demand a large amount of resources that a local machine does not have. So we decided to use a cloud solution with a dedicated machine containing the necessary resources. Unfortunately, these are high cost solutions. Hence, a single epoch was executed for each distinct backbone, except for the selected resnet152 backbone. In this case, the model was trained using 33 epochs.

During the training we worked with the RetinaNet implementation in TensorFlow 2.0 and Keras as described in [20]. This is a work-in-progress solution found on the GitHub site under the name tf-retinanet [22]. The input parameters used in the training are: the folder where the MSL dataset is located, the folder containing the annotation files, a file with the twenty one classes, and the name of the backbone. Also, within this library there is a Python method that we used to predict hand-gestures on an image. The method returns a tuple with three values which are: boxes, scores, and labels.

For a practical evaluation process of the MSL dataset, a Python application was developed which recognizes gestures through a web camera in real time using the OpenCV library. Regarding the response time on a local machine, for each prediction of a sign, it requires an average time of 200 milliseconds. The application is capable of

recognizing multiple gestures at the same time on the same frame and showing the name of the sign as well as the accuracy, as depicts in Figure 4.

The backbones we tested include: resnet50, resnet152, vgg16 and densenet201. They were selected based on the work in [24]. Additionally, we applied an extra step to test the model using gray-scale images. The one epoch training results are displayed in Table 2.

In such results, we can note the training with gray-scale images versus training with color images does not make much difference. The one epoch elapsed time and loss value are pretty similar in both scenarios. However, we can notice a difference when we use distinct backbones. We got the best loss value in the training using the resnet152 backbone. Although time increased significantly, we see better performance in loss value. Opposite to this result was the vgg16 backbone where the classification loss value was the highest. Therefore, it was ruled out to continue using it in training.

5 Conclusions

The desire to have an MSL dataset to train deep learning models led us to build ours. In this paper we evaluated the two-hand static MSL dataset specialized in object detection using RetinaNet model with distinct backbones. Despite the different positions and angles in the images, the accuracy of the model greater than 99%.

The relevance of generating an MSL dataset lies as part of a contribution to the scientific community of artificial intelligence, so that, they can take this presented dataset as reference, train object detection models and materialize its contribution through applications aimed at the benefit of the Mexican deaf community.

The MSL dataset is planned to be uploaded to a repository that researchers have access to. Likewise, it is intended to make the code of the applications and libraries used in this research available to the community.

It has been demonstrated that it is possible to use object detection models applied to the MSL with positive results. It is possible to predict hand-gestures through low-cost devices such as webcams. It was shown that the prediction of these gestures can be performed in real time on not controlled environments with distinct attributes. Finally, the MSL dataset is able to work with child or adult gestures, different skin color, the right or left hand, and accessories used on them.

6 Future Work

In this preliminary work, the MSL dataset contains only 21 of the 27 static MSL gestures. The reason is because to generate the other 6 MSL gestures (J, K, Ñ, Q, X, Z), hand movement is required. Current research work is planned to continue in future developments. One of the goals is to complement the MSL dataset to recognize dynamic gestures in real time. It is intended to continue working in not controlled environments. An initial idea is to use long short-term memory (LSTM) by adapting an artificial recurrent neural network (RNN) architecture to our model.

We aspire to integrate the entire project into low-cost devices such as mobile phones, tablets, laptops, or even Google Glass. The integration of these devices in Mexican society will benefit the inclusion of people with hearing loss.

References

1. Deafness and Hearing Loss: <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss> (2018)
2. National Institute of Statistics and Geography (INEGI): Mexico National Survey of Demographic Dynamics. <https://en.www.inegi.org.mx/programas/enadid/2018/> (2018)
3. Romero, S., García, I., Chris, F., Lomeli, K.A.: Preparing Teachers for Inclusion in Mexico. How Effective is this Process?. *Journal of Education for Teaching*, 39(5), pp. 509–522 (2013). Doi: 10.1080/02607476.2013.836340.
4. García, F., Tinoco, L., González, H.G., García, M.J., Moreno, J.: Modelado e implementación de un sistema de enseñanza de lenguaje de señas mexicano. *Ciencia ergo-sum*, 14(2), pp. 185–190 (2007)
5. Isaacs, J., Foo, S.: Hand Pose Estimation for American Sign Language Recognition. In: Thirty-Sixth Southeastern Symposium on System Theory, pp. 132–136 (2004). Doi: 10.1109/SSST.2004.1295634.
6. Von, A.U., Zieren, J., Canzler, U., Bauer, B., Kraiss K.F.: Recent Developments in Visual Sign Language Recognition. *Univ Access Inf Soc*, 6, pp. 323–362 (2008). Doi: 10.1007/s10209-007-0104-x.
7. Martínez, M.E., Rojano, J.R., Benitez, E., Sánchez, H.E.: Data Acquisition Software for Sign Language Recognition. *Research in Computing Science*, pp. 205–2011 (2019). Doi: 10.13053/rcs-148-3-17.
8. Ronchetti, F., Quiroga, F., Estrebow, C., Lanzarini, L., Rosete, A.: LSA64: An Argentinian Sign Language Dataset. Instituto de Investigación en Informática LIDI (2016). Doi: 10.48550/arXiv.2310.17429.
9. Neidle, C., Thangali, A., Sclaroff, S.: Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus. *IEEE Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis (CVPR4HB)*, pp. 1–8 (2008)
10. Barczak, A.L., Reyes, N.H., Abastillas, M., Piccio, A., Susnjak, T.: A New 2D Static Hand Gesture Colour Image Dataset for ASL Gestures. *Res. Lett. Inf. Math. Sci.*, 15, pp. 12–20 (2011)
11. Islam, M., Sultana, S.S., Jessan, N., Rabby, A.S., Hossain, S.: Ishara-Lipi: The First Complete Multipurpose Open Access Dataset of Isolated Characters for Bangla Sign Language. *International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–4 (2018). Doi: 10.1109/ICBSLP.2018.8554466.
12. Ebling, S., Camgoz, N., Braem, P., Tissi, K., Sidler, S., Stoll, S., Hadfield, S., Haug, T., Bowden, R., Tornay, S., Razavi, M., Magimai, M.: SMILE Swiss German Sign Language Dataset (2018)
13. Oliveira, M., Chatbri, H., Ferstl, Y., Farouk, M., Little, S., O'Connor, N., Sutherland, A.: A Dataset for Irish Sign Language Recognition. *Proceedings of the Irish Machine Vision and Image Processing Conference (IMVIP), International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8 (2017)
14. Ghazanfar, L., Nazeeruddin M., Jaafar, R., Rawan, A.: ArASL: Arabic Alphabets Sign Language Dataset. *Data in Brief*, 23 (2019). Doi: 10.1016/j.dib.2019.103777.

15. Cervantes, J., García, F., Rodríguez, L., Rendón, A., Lopez, A.: Recognition of Mexican Sign Language from Frames in Video Sequences. Huang and K.-H. Jo: ICIC Part II, LNCS 9772, pp. 353–362 (2016)
16. Teja L.S., Jain L., Sharmay A.: Two Hand Indian Sign Language Dataset for Bechmarking Classification Models of Machine Learning (2019). Doi: 10.1109/ICICT46931.2019.8977713.
17. COCO and Pascal VOC Data Format for Object Detection: <https://towardsdatascience.com/coco-data-format-for-object-detection-a4c5eaf518c5> (2017)
18. Tzutalin: Label Img. Git Code. <https://github.com/tzutalin/labelImg> (2015)
19. QT Open-Source: Widget Toolkit for Creating Graphical User Interfaces as Well Ascross-Platform Applications. <https://www.qt.io/> (2018)
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection. arXiv preprint (2017). Doi: 10.48550/arXiv.1708.02002.
21. Pradipta, A., Imon A.: Performance Comparison of Different CNN Models for Indian Road Dataset. ICGSP 3rd International Conference on Graphics and Signal Processing (2019). Doi: 10.1145/3338472.3338480.
22. TF RetinaNet: GitHub. <https://github.com/fizyr/tf-retinanet> (2020)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). Doi: 10.1109/CVPR.2016.90.
24. Bressen, K., Adams, L., Erxleben, C., Hamm, B., Niehues, S., Vahldiek, J.: Comparing Different Deep Learning Architectures for Classification of Chest Radiographs. arXiv. (2020). Doi: 10.48550/arXiv.2002.08991.