Reconocimiento de la Lengua de Señas Mexicana mediante redes neuronales

Autores: Kenneth Mejía Pérez, Diana Margarita Córdova Esparza

Facultad de Informática, Campus Juriquilla, Av. De las Ciencias S/N, Querétaro, Qro.
Correspondencia: kmejia09@alumnos.uaq.mx , diana.cordova@uaq.mx

Resumen

En el presente trabajo de investigación se desarrolló un sistema para el reconocimiento automático del abecedario dactilológico de la Lengua de Señas Mexicana (LSM), esto mediante el uso de redes neuronales recurrentes y el uso de una cámara de profundidad. Para la clasificación automática del conjunto de señas se utilizó una red neuronal recurrente (RNN, por sus siglas en inglés). Para evaluar el rendimiento del clasificador se calculó la precisión, la recuperación y la exactitud.

Palabras clave: Lengua de señas, cámara RGB-D, redes neuronales

Abstract

In the present research work, we developed a system for automatically recognizing the Mexican Sign Language (LSM)'s dactylological alphabet using a depth camera and recurrent neural networks (RNN) and its variations, such as LSTM and GRU. We used precision, recall, and accuracy to evaluate the classifier's performance.

Keywords: Sign Language, RGB-D camera, neural networks

Artículo arbitrado	
Recibido:	Aceptado:
15 de febrero de 2023	23 de febrero de 2023

1. Introducción

La lengua de señas es uno de los medios de comunicación más utilizado por las personas con discapacidades auditivas en todo el mundo. La lengua de señas se diferencia de una lengua oral debido a que utiliza el canal de comunicación viso-gestual en lugar del audio-vocal (Tovar, 2001). Por tal motivo, la lengua de señas, como cualquier otro idioma, tiene su propia estructura fonológica. No obstante, el término fonología hace referencia a sonidos verbales (Burquest, 2009). Por esta razón, en la lengua de señas el concepto de fonología se sustituye por el de querología, el cual describe las unidades combinatorias elementales o queremas que constituyen las palabras y signos de las lenguas de señas. La querología analiza las señas mediante los siguientes parámetros en función de rasgos geométricos y de movimiento (Rodríguez-González, 1992):

1. Queirema (configuración): forma que adopta la mano de acuerdo con la configuración de los dedos.
2. Kinema (movimiento): como algunas señas son dinámicas, el kinema se refiere a la forma en la cual se realiza el movimiento (circular, zig zag, lineal, etc.) que produce diferencias semánticas.
3. Toponema (ubicación): espacio donde se hace la seña.
4. Kineprosema (dirección): movimiento de las manos hacia un sentido.
5. Queirotropema (orientación): orientación de la mano con respecto al cuerpo.
6. Prosoponema (rasgos no manuales): se refiere a todos aquellos rasgos que no utilizan las manos, principalmente movimiento corporal y expresiones faciales.

La lengua de señas, también conocida como lengua de signos, clasifica las señas en unimanuales y bimanuales. Las unimanuales pueden ser estáticas o dinámicas; mientras que las bimanuales pueden ser simétricas o asimétricas (Cruz y Ramírez, 2017).

El presente trabajo de investigación se encuentra estructurado en cuatro secciones. En la primera sección se realiza una introducción sobre la lengua de señas y su estructura querológica. En la segunda sección se presenta la metodología utilizada para desarrollar el sistema de reconocimiento del abecedario dactilológico, el cual consiste en las siguientes etapas: adquisición de las imágenes, preprocesado de los datos, almacenamiento de la información, entrenamiento del clasificador y análisis e interpretación de los resultados. En la tercera sección se muestran los resultados obtenidos y, finalmente, en la cuarta sección se describe la discusión y las conclusiones derivadas de esta investigación.

1.1 Lengua de Señas Mexicana

El desafío de las personas con discapacidad auditiva para comunicarse con su entorno limita su desarrollo educativo, profesional y social; como resultado, se ven condicionadas sus oportunidades de inclusión. Ante esta necesidad, las personas sordas han desarrollado su propia forma de comunicación: la Lengua de Señas Mexicana (LSM). A pesar de que ésta les permite comunicarse entre sí, no siempre facilita la relación con el resto de la comunidad, sobre todo, con los oyentes que desconocen esta lengua.

La LSM representa una de las formas de expresión de la comunidad sorda en México y una herramienta fundamental para la inclusión en los procesos de participación social. La LSM se compone de signos visuales con una estructura lingüística propia con la cual se identifican y expresan las personas sordas en México.

1.2 Aprendizaje automático

El aprendizaje automático (*Machine Learning*) se sitúa dentro de la Inteligencia Artificial (IA) (ver Figura 1) y su objetivo es desarrollar sistemas que "aprenden" por sí mismos, esto en función de los datos de entrada, los cuales pueden provenir de imágenes o videos. Se clasifica en dos tipos de aprendizaje: supervisado y no supervisado.

Generalmente, el aprendizaje automático supervisado se usa para clasificar o hacer predicciones de datos, mientras que el aprendizaje no supervisado se utiliza para encontrar estructuras y patrones dentro de los conjuntos de datos que se están analizando.

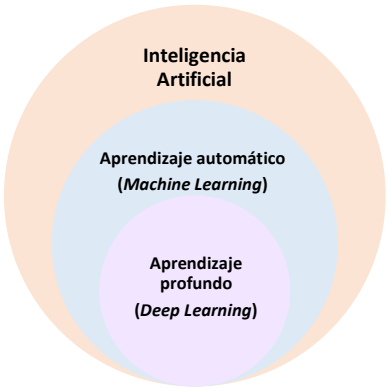

Figura 1. Aprendizaje automático (*machine learning*). Fuente: elaboración propia.

En la Figura 2 se muestra como es el funcionamiento de las técnicas de clasificación que se basan en el aprendizaje automático, el cual consiste en: recibir información proveniente del mundo (entrada), extraer características mediante un descriptor, realizar la clasificación y, finalmente, obtener una salida como respuesta a la técnica de clasificación utilizada.

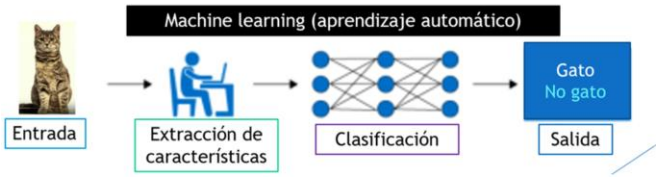

Figura 2. Etapas del aprendizaje automático (*machine learning*). Fuente: elaboración propia.

1.3 Redes Neuronales Artificiales

Las Redes Neuronales Artificiales (ANN, por sus siglas en inglés *Artificial Neural Networks*) se tratan de un modelo computacional inspirado en las redes neuronales biológicas, el cual forma parte de las técnicas del aprendizaje automático y se utiliza en diversas tareas como el reconocimiento de patrones, la detección y clasificación de objetos y la predicción de datos, por mencionar algunas.

En la Figura 3 se muestra la configuración básica de una red neuronal, la cual consta de una capa de entrada (*input layer*), una capa oculta (*hidden layer*) para el procesamiento de la información y una capa de salida (*output layer*).

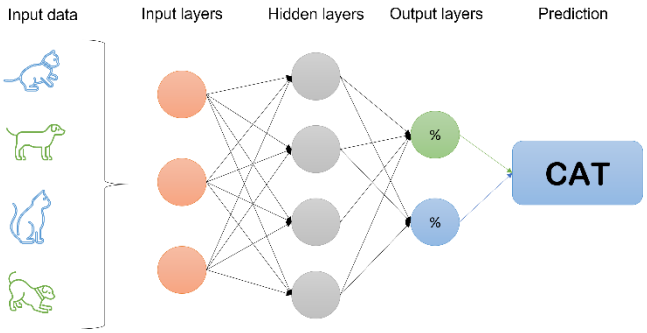

Figura 3. Estructura básica de una red neuronal. Fuente: elaboración propia.

Dentro de las redes neuronales existen diversos modelos, como lo son las redes neuronales profundas (DNN, por sus siglas en inglés *Deep Neural Networks*), las redes neuronales convolucionales (CNN, por sus siglas en inglés *Convolutional Neural Networks*), las redes neuronales recurrentes (RNN, *Recurrente Neural Networks*), entre otras.

1.3.1 Red neuronal recurrente

Una RNN permite realizar la clasificación de datos que varían en el tiempo, es decir, es un modelo que tiene la misma premisa que una red neuronal artificial, pero añade recurrencia de la capa de salida de la neurona pasada a la capa de entrada de la siguiente neurona, de esta manera la salida de la neurona pasada influye en el compartimiento de la siguiente red (Grossberg, 2013).

2. Metodología

En la presente investigación se desarrolló el reconocimiento del abecedario dactilológico de la Lengua de Señas Mexicana (LSM) siguiendo la metodología que se muestra en la Figura 4.

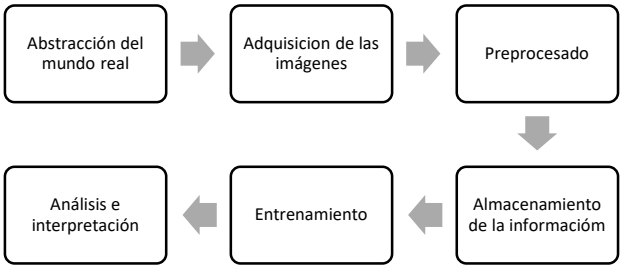

Figura 4. Metodología desarrollada para el sistema de reconocimiento automático. Fuente: elaboración propia

A continuación, de describen cada una de las etapas de la metodología.

Etapa 1. Abstracción del mundo real

Esta etapa es fundamental para poder modelar un sistema de visión por computadora. De manera conceptual, un sistema de visión por computadora o visión artificial se basa en un conjunto de herramientas tecnológicas que permiten captar y registrar imágenes del mundo real, esto con el propósito de poder procesarlas de manera digital a través de una computadora y así obtener información que puede ser relevante para una aplicación en particular. Por tal motivo, es necesario hacer un análisis acerca del tipo de tecnología que se utilizará para realizar la adquisición de estas imágenes, pues debe cumplir con las características pertinentes para la aplicación que se requiere desarrollar.

Etapa 2. Adquisición de las imágenes

Para esta etapa se creó un corpus con el abecedario dactilológico mediante el uso de una cámara de profundidad OAK-D. Este dispositivo consta de tres cámaras: una cámara central para capturar la información RGB y dos cámaras laterales para medir distancias utilizando la disparidad entre las imágenes. Asimismo, para la adquisición de los datos e imágenes se utilizó la librería DepthAI (Documentación de DepthAi, 2021).

Las señas pertenecientes a la LSM que se adquirieron se enlistan en la Tabla 1 junto a una breve descripción querológica.

Tabla 1. Descripción del cuerpo de los datos.

Tipo de seña	Seña	Estática/Dinámica
Abecedario dactilológico	A	Estática
	B	Estática
	C	Estática
	D	Estática
	E	Estática
	F	Estática
	G	Estática
	H	Estática
	I	Estática
	J	Dinámica
	K	Dinámica
	L	Estática
	M	Estática
	N	Estática
	Ñ	Dinámica
	O	Estática
	P	Estática
	Q	Dinámica
	R	Estática
	S	Estática
	T	Estática
	U	Estática
	V	Estática
	W	Estática
	X	Dinámica
	Y	Estática
	Z	Dinámica

Como se puede observar, se adquirieron un total de 27 señas distintas pertenecientes al abecedario dactilológico. De estas señas, seis son estáticas y 21 son dinámicas.

Etapa 3. Preprocesado

En esta etapa se detectaron de forma automática los puntos característicos (*keypoints*) de la cara, cuerpo y manos que se usan al realizar cada una de las señas. Para la detección automática de estos puntos se utilizó la librería MediaPipe (Zhang et al., 2020) (Singh et al., 2021). Cabe mencionar que los puntos de interés seleccionados quedaron distribuidos de la siguiente manera: 20 para la cara, cinco para el cuerpo y 21 para cada mano.

Etapa 4. Almacenamiento de la información

Los datos capturados se almacenaron en tablas de valores separados por comas (csv), las cuales están estructuradas en 20 filas y 201 columnas, en donde

cada archivo representa una única repetición de una seña individual y cada fila representa la información obtenida en un solo fotograma.

Las filas, por otra parte, están estructuradas con la información obtenida del cuerpo (cinco puntos), rostro (20 puntos), mano izquierda (21 puntos) y la mano derecha (21 puntos). Cabe mencionar que cada punto se representa por sus coordenadas (X, Y, Z). Esta información de distancia tiene como unidad de medida el metro y están calculadas respecto al punto central de la imagen capturada. Además, los datos toman valores negativos en los ejes -X, -Y, -Z, lo cual resulta muy útil cuando se quiere conocer la dirección del movimiento realizado.

Etapa 5. Entrenamiento

Debido a que el abecedario dactilológico consta de señas estáticas y dinámicas, se optó por utilizar un modelo neuronal recurrente (descrito en la sección 1.3.1), el cual permite clasificar información que varía con el tiempo. Asimismo, es importante denotar que mediante el sistema de adquisición se captura información tridimensional de cada una de las señas, lo cual representa una ventaja con respecto a los trabajos que se encuentran en el estado del arte, esto debido a que los datos adquiridos son invariantes a escala y cambios lumínicos.

Después de capturar el corpus del abecedario dactilológico, se dividió el conjunto de datos en tres partes: 70 % para datos de entrenamiento, 15 % para validación y 15 % para pruebas. Se realizaron entrenamientos empleando la red neuronal recurrente (la cual se muestra en la Figura 5) y haciendo variaciones en la capa de entrada y la capa oculta (ver Tabla 2) para encontrar el mejor modelo neuronal para las señas adquiridas. Se siguió la siguiente configuración para los cuatro modelos entrenados: 32 y 64; 64 y 128; 128 y 256; 256 y 512.

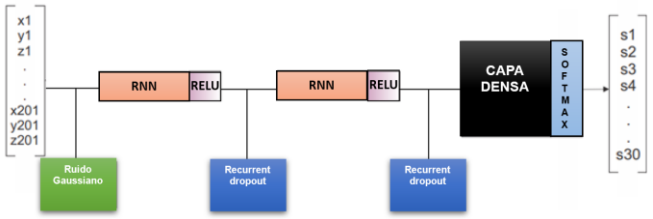

Figura 5. Arquitectura de la red neuronal. Fuente: elaboración propia

Como configuración general del modelo neuronal, se utilizó un diseño de 500 épocas y detención temprana con el fin de evitar un sobreajuste en el entrenamiento de los datos, esto con una paciencia de 50 épocas; también, se empleó la función de pérdida de entropía cruzada categórica para medir la pérdida entre las distribuciones probabilidad y, finalmente, se utilizó el optimizador de Adam para reducir el error en la red mediante las librerías de Keras (Chollet et al; 2018) y Tensorflow (Abadi et al; 2018).

Etapa 6. Análisis e interpretación

En esta etapa se realizó un análisis de los resultados con el propósito de inferir el comportamiento de los parámetros del modelo neuronal en relación con las métricas calculadas para su validación.

Como fue descrito con anterioridad, los modelos resultantes de la etapa de entrenamiento fueron probados con el conjunto de datos de prueba (15 % de los datos). Para la validación del modelo se calculó un promedio de la exactitud obtenida de cada en red entrenada.

La exactitud (*Accuracy*) mide la frecuencia con la que las predicciones coinciden con las etiquetas, es decir, el porcentaje de valores predichos que se corresponden con los valores reales. La exactitud se calcula mediante la ecuación (1).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

Tabla 2: Variaciones en la arquitectura del modelo neuronal utilizado para la clasificación de las señas.

Modelo Neuronal	Capa 1	Capa 2	Exactitud
RNN	32	64	0.7925
	64	128	0.7950
	128	256	0.7432
	256	512	0.5432

Por otra parte, se calcularon los valores de las métricas: precisión (*precision*) y recuerdo (*recall*), para poder realizar una comparativa entre los distintos experimentos.

La precisión indica la proporción de identificaciones positivas que fueron realmente correctas. La precisión se calcula con la ecuación (2).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

El recuerdo (*recall*) representa la proporción de positivos reales identificados correctamente. El recuerdo se calcula por medio de la ecuación (3).

$$\text{Recall} = \frac{\text{TP}}{\text{FP+FN}} \quad (3)$$

3. Resultados

La experimentación se realizó utilizando 27 clases correspondientes al abecedario dactilológico de la LSM, los cuales fueron adquiridos mediante la cámara de profundidad OAK-D. De estos datos, se utilizó el 70 % para el entrenamiento, 15 % para datos de validación y el 15 % para pruebas.

Mediante una red neuronal recurrente, se realizó el entrenamiento variando la capa de entrada y la capa oculta con diferentes cantidades de neuronas. Después de seleccionar los mejores modelos, el experimento se repitió. En esta ocasión se añadió ruido gaussiano de forma aleatoria a los datos entrenados para ayudar a mejorar la robustez de la red neuronal con datos atípicos.

En la Figura 6 se muestra la comparación entre las métricas de precisión y recuerdo de los cuatro modelos entrenados sin ruido. Se puede observar que el mejor modelo corresponde a una red con una capa de entrada de 32 neuronas y una capa oculta formada por 64 neuronas.

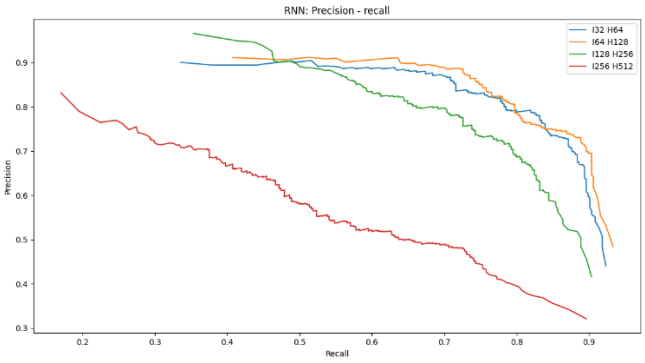

Figura 6. Comparación entre los valores de precisión y recuerdo. Fuente: elaboración propia

En la Tabla 3 se muestra la exactitud de los mejores modelos neuronales entrenados con ruido. Como se puede observar, la exactitud incrementó en el primer caso respecto a los modelos entrenados sin ruido.

Tabla 3: Variaciones en la arquitectura del modelo neuronal con ruido gaussiano, utilizado para la clasificación de las señas.

Modelo Neuronal	Ruido Gaussiano	Capa 1	Capa 2	Exactitud
RNN	0.1	32	64	0.8074
	0.1	64	128	0.7925

Finalmente, se muestra en la Figura 7 la matriz de confusión por clase. En este instrumento se observan en tonalidades más oscuras las señas en las que se obtuvo una mayor exactitud.

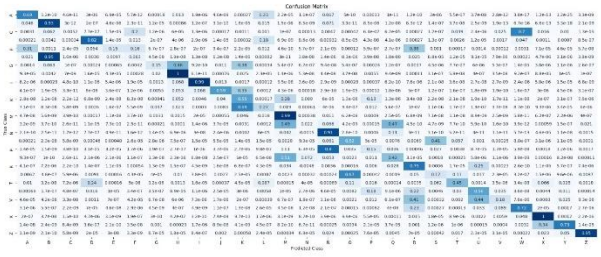

Figura 7. Matriz de confusión para el modelo entrenado sin ruido, con 32 capas de entrada y 64 capas ocultas. Fuente: elaboración propia

Como se puede observar en la Figura 5, las señas más descriptivas son D, I, Ñ, Q y X. Por otra parte, las señas que tienen más problemas para ser reconocidas son las de C y T, las cuales son altamente confundibles con W y S, respectivamente.

4. Discusión y conclusiones

En este trabajo de investigación se implementó un sistema para el reconocimiento de la lengua de señas utilizando una cámara RGB-D. Se recolectó un conjunto de datos de 27 señas pertenecientes al abecedario dactilológico de la Lengua de Señas Mexicana. Para realizar el entrenamiento, validación y prueba del modelo neuronal, para cada seña se extrajeron puntos característicos de las manos, el cuerpo y los rasgos faciales y se realizó la transformación de estos puntos en coordenadas 3D para entrenar el clasificador basado en redes neuronales recurrentes (RNN). Con el clasificador entrenado con datos ruidosos se obtuvo un 80.74 % de exactitud para el mejor de los casos. Estos resultados son congruentes con el análisis querológico de las señas procesadas, debido a que es altamente conocido que las señas pertenecientes al abecedario dactilológico comparten una gran cantidad de rasgos similares.

Con respecto a otros trabajos desarrollados en la literatura que hacen uso de cámaras RGB-D y redes neuronales, se encuentra la propuesta de Galicia et al. (2015) en la cual se propone un sistema para reconocer cinco vocales (A, E, I, O, U) y dos consonantes (B, L) de la LSM mediante el uso de un sensor Kinect y una red neuronal; el resultado fue una exactitud del 76.19 %. Los autores Martínez-Gutiérrez et al. (2019) desarrollaron un software que permite capturar 22 puntos de la mano en coordenadas 3D por medio de la cámara Intel-RealSense f200 y el uso de una red neuronal perceptrón multicapa, del cual obtuvieron una tasa de reconocimiento del 80.11 %.

Cabe mencionar que una de las principales ventajas que ofrecen los sistemas de reconocimiento basados en cámaras de profundidad (RGB-D) es la adquisición de las señas sin colocar dispositivos externos en el usuario, lo cual permite una comunicación más natural.

Referencias

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Tensorflow.org. https://doi.org/10.48550/arXiv.1603.04467

Burquest, D. A. (2009). Análisis fonológico: un enfoque funcional. *SIL International*, *9254*(17). Recuperado el 17 de marzo del 2023, de: https://www.sil.org/resources/archives/9254.

Chollet, F. (2021). Keras: The python deep learning library. Version (2.7.0). Keras. Recuperado el 17 de marzo del 2023, de: https://keras.io/.

Luxonis. (2020). DepthAI's Documentation. versión (2.6.0.0), Recuperado el 17 de marzo del 2023 de: https://docs.luxonis.com/en/latest/

Mercader Flores, C. A., Escobar Dellamary, L., Ramírez Barba, M. del R. G., Pool Westgaard, M., & Cruz Aldrete, M. (2017). C. E. Escobedo Delgado (Ed.), Diccionario de Lengua de Señas Mexicana Ciudad de México (pp. 1–505). INDEPENDI. Recuperado el 17 de marzo del 2023 de: https://pdh.cdmx.gob.mx/storage/app/media/banner/Dic_LSM%202.pdf

Galicia, R., Carranza, O., Jimenez, E. D., & Rivera, G. E. (2015). Mexican sign language recognition using movement sensor. *2015 IEEE 24th International Symposium on Industrial Electronics* (ISIE). DOI: https://doi.org/10.1109/isie.2015.7281531

Grossberg, S. (n.d.). Recurrent neural networks. Scholarpedia. Recuperado de: http://www.scholarpedia.org/article/Recurrent_neural_networks

Martínez-Gutiérrez, M. E., Rojano-Cáceres, J. R., Benítez-Guerrero, E., & Sánchez-Barrera, H. E. (2019). Data Acquisition Software for sign language recognition. Research in Computing Science, *148*(3), pp. 205–211. https://doi.org/10.13053/rcs-148-3-17

González Rodríguez, M. A. (1992). Lenguaje de Signos. Confederación Nacional de Sordos de España.

Singh, A. K., Kumbhare, V. A., & Arthi, K. (2022). Real-time human pose detection and recognition using MediaPipe. Advances in Intelligent Systems and Computing, 145–154. https://doi.org/10.1007/978-981-16-7088-6_12

Tovar, L. A. (2011). La Importancia Del Estudio De Las Lenguas De Señas. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., & Grundmann, M. (2020). *MediaPipe Hands: On-Device Real-Time Hand Tracking*.