Reconocimiento aislado de la lengua de señas chilena en tiempo real usando Deep Learning

Real-time isolated recognition of chilean sign language using Deep Learning

Miguel Yánez[1]* https://orcid.org/0009-0009-2965-5629
Ana Aguilera[1] https://orcid.org/0000-0003-0726-1759

Recibido 12 de octubre de 2024, aceptado 06 de noviembre de 2024
Received: October 12, 2024 Accepted: November 06, 2024

RESUMEN

La lengua de señas es parte fundamental de la cultura sorda y tiene sus propias reglas lingüísticas y gramaticales que permiten la perfecta comunicación entre sus hablantes (señantes). En informática se han estudiado algunos mecanismos automáticos para favorecer esta comunicación. Uno de ellos es el Reconocimiento de la Lengua de Señas (SLR) el cual abarca el proceso completo de seguimiento e identificación de las señas realizadas, así como su conversión en palabras y expresiones. En este contexto, este trabajo se centra en desarrollar un conjunto de datos para el Reconocimiento Aislado de Lengua de señas Chilena (LSCh) mediante la segmentación manual de videos de estudiantes sordos, pre-procesándolos con una herramienta de estimación de postura y clasificándolos con modelos de aprendizaje profundo basados en Transformers. Además, explora el impacto del aumento de datos en el rendimiento del modelo y lo compara con resultados del corpus de lengua de señas argentino (LSA64) para evaluar la capacidad de reconocimiento con nuevos señantes. Los resultados muestran la importancia del aumento de datos en conjuntos de datos más pequeños y demuestran que un solo modelo puede aprender de signos realizados tanto por signantes zurdos como diestros. Las métricas de los experimentos muestran una exactitud de 0,75 y 0,94 y una sensibilidad de 0,74 y 0,94 para LSCh y LSA64 respectivamente, utilizando los modelos con mejor rendimiento (Mirror + Angle).

Palabras clave: Aprendizaje profundo, SLR, LSCh.

ABSTRACT

Sign language is a fundamental part of the deaf culture, and it has its own linguistic and grammatical rules that allow perfect communication between its speakers (signers). In computer science, some automatic mechanisms have been studied to favor this communication. One of them is the Sign Language Recognition (SLR) which covers the whole process of tracking and identification of the signs made, and their conversion into words and expressions. In this context, this work focuses on developing a dataset for Isolated Chilean Sign Language Recognition (LSCh) by manually segmenting videos of deaf students, pre-processing them with a pose estimation tool and classifying them with Transformer-based deep learning models. Furthermore, it explores the impact of data augmentation on model performance and compares it with results from the Argentine sign language corpus (LSA64) to assess recognition ability with new signers. The results show the importance of data augmentation on smaller datasets and demonstrate that a single model can learn from signs made by both left-handed and right-handed signers. Metrics from the experiments show an accuracy of 0.75 and 0.94 and recall of 0.74 and 0.94 for LSCh and LSA64 respectively, using the best performing models (Mirror + Angle).

Keywords: *Deep learning, SLR, LSCh.*

[1] Escuela de Ingeniería Informática. Facultad de Ingeniería, Universidad de Valparaíso.
E-mail: miguel.yanez@postgrado.uv.cl; ana.aguilera@uv.cl
* Autor de correspondencia: miguel.yanez@postgrado.uv.cl

INTRODUCCIÓN

Las comunidades sordas se encuentran en un proceso de lucha constante por el derecho al acceso a la educación y todos los ámbitos de la sociedad en sus lenguas de señas [1]. Estas lenguas se pueden considerar como lenguas minoritarias que coexisten junto a la lengua hablada dominante de cada país, lo que crea una barrera lingüística que dificulta la comunicación tanto en la vida cotidiana de las personas sordas como en la atención institucional [2].

Según el Censo de 2012 [3], existen 488.511 personas sordas o con problemas de audición en Chile, de las cuales cerca de un 40% no ha completado la educación básica [4]. Desafortunadamente, no se dispone de información más actualizada en este momento y estas estadísticas no consideran el hecho de que los integrantes de la comunidad sorda no son lo mismo que personas con discapacidades auditivas. Esta confusión se origina principalmente por la falta de reconocimiento de dicha comunidad como una minoría cultural y lingüística [5].

La lengua de señas es parte fundamental de la cultura sorda y tiene sus propias reglas lingüísticas y gramaticales que permiten la perfecta comunicación entre sus hablantes [6]. Además, es importante que no se les niegue a las personas sordas el aprendizaje de la lengua de señas, ni presentar la oralidad como el único medio para el conocimiento y adquisición de la cultura [7]. En Chile, esta lengua se conoce como la Lengua de Señas Chilena (LSCh), la cual es mencionada en la Ley Nº 20.422 [8] de 2010. Sin embargo, sólo a fines del año 2020 se aprobó una importante modificación a este artículo, en la que finalmente se le reconoce como la *"lengua natural, originaria y patrimonio intangible de las personas sordas"*.

En cuanto a la naturaleza de las lenguas de señas, estas son lenguas visuales que están compuestas por gestos manuales, expresiones faciales y movimientos corporales. Estas lenguas no son universales; cada lengua tiene sus propias reglas lingüísticas y estructuras gramaticales y un vocabulario único [9]. Las señas de las lenguas de señas se producen utilizando una cantidad finita de expresiones gestuales, que se pueden agrupar en dos categorías principales: expresiones manuales y no manuales. Las expresiones manuales se refieren al movimiento de las manos y transmiten la mayor parte de la información. Por otro lado, las expresiones no manuales incluyen expresiones faciales y postura corporal que mejoran el significado y contexto de la expresión gestual [10]. Adicionalmente, las señas que requieren movimiento se conocen como señas dinámicas, a diferencia de las señas estáticas que no involucran movimiento alguno para ser identificadas [11].

El reconocimiento aislado de la lengua de señas, conocido como Sign Language Recognition (SLR) en inglés, abarca el proceso completo de seguimiento e identificación de las señas realizadas, así como su conversión en palabras y expresiones [12]. Es un campo de investigación desafiante en la visión artificial. Algunos de los desafíos más importantes en esta área incluyen lidiar con la ocultación de la mano, movimientos rápidos de la mano, cambios en la iluminación y la complejidad del fondo [13]. El SLR se puede categorizar en tres amplias categorías: deletreo de dedos, reconocimiento aislado (ISLR) y reconocimiento continuo (CSLR). Este estudio se centra en el reconocimiento de señas dinámicas aisladas, en donde una sola seña mostrada en una secuencia de movimientos se asocia a su glosa correspondiente [11].

En la última década, los mayores avances en el campo del SLR se han logrado utilizando modelos basados en el aprendizaje profundo. Estos modelos, consisten en algoritmos computacionales de múltiples capas que pueden aprender y representar datos con varios niveles de abstracción, imitando el mecanismo cerebral humano para comprender las complejas estructuras de datos a gran escala [14]. El aprendizaje profundo ha demostrado un gran éxito en el procesamiento de videos e imágenes, especialmente en la estimación de poses y el reconocimiento de gestos [13]. Esto se debe principalmente a la automatización de la creación de características necesarias para identificar las señas realizadas. En lugar de crearlas manualmente desde cero, se trata de aprender automáticamente características a partir de un conjunto de datos [15].

Existen varios conjuntos de datos para el reconocimiento de señas que cubren una gran variedad de modalidades y lenguas. Sin embargo, la disponibilidad de conjuntos de datos especializados en el ISLR es limitada. En el caso de la LSCh,

esta falta de datos es aún más evidente, ya que actualmente el único conjunto de datos disponible es el "Abecedario Lenguaje de Señas Chileno-Español" [16], el cual se enfoca en el deletreo de dedos y no a nivel de palabras. Por esta razón, las principales contribuciones de este trabajo son:

1. La creación de un conjunto de datos para el ISLR de la LSCh, segmentado manualmente a partir de grabaciones de estudiantes sordos de entre 6 a 18 años, en posiciones realistas y desde varias perspectivas.
2. Una evaluación del rendimiento de modelos ISLR entrenados con muestras de señantes diestros y zurdos, así como técnicas de aumento de datos.
3. Una evaluación comparativa de resultados obtenidos al entrenar un modelo basado en Transformers utilizando el conjunto de datos de la LSCh y el corpus LSA64.

El artículo se estructura de la siguiente manera: en primer lugar, se ofrece una breve descripción del estado del arte en los métodos y conjuntos de datos relacionados con el ISLR. A continuación, se detalla el conjunto de datos de la LSCh y la arquitectura empleada. Seguido de ello, se presentan los resultados obtenidos de los experimentos y se lleva a cabo una discusión sobre los mismos. Finalmente, se exponen las conclusiones y se proponen posibles líneas de trabajo futuro.

TRABAJOS RELACIONADOS

El estado del arte de los estudios relacionados con el ISLR se puede dividir en dos enfoques principales: apariencia y postura. El enfoque basado en la apariencia involucra el procesamiento de los fotogramas de las grabaciones junto con modelos cromáticos como el RGB, modelos de profundidad o una combinación de ambos, conocida como RGB-D [17], [18]. Por otro lado, el enfoque basado en postura se ha destacado últimamente debido a los recientes avances en los modelos de estimación de la postura, el cual consiste en extraer el esqueleto de un individuo formado por las coordenadas de sus articulaciones, generando una representación espacio-temporal de los movimientos realizados por el individuo [11], [19].

Los primeros estudios del SLR implementaron modelos estadísticos de Markov (HMM) [20] y la creación de características manualmente. El éxito de los HMM en otras áreas del reconocimiento automático fue un gran incentivo para la creación de nuevos estudios enfocados en el SLR [21]. Desde entonces, se han aplicado continuamente nuevas técnicas hasta llegar al uso de redes neuronales como las redes convolucionales o Convolutional Neural Networks (CNN) [22], específicamente diseñadas para resolver problemas de visión artificial. Las CNN facilitaron la generación automática de características, marcando un gran avance en el SLR y marcó el comienzo de la creación de modelos híbridos CNN-HMM [23] e incluso la combinación con redes recurrentes o Long Short-Term Memory (LSTM) [24], abordando el desafío de la representación temporal. Eventualmente, surgieron varios estudios que utilizan las redes conocidas como las 3D-CNN [17], que se especializan en resolver problemas espacio-temporales como el reconocimiento de acciones. En años recientes, también se han adoptado nuevas arquitecturas basadas en Transformers. Estas redes neuronales, originalmente propuestas por Vaswani *et al.* en 2017 [25], se han aplicado en investigaciones relacionadas con el SLR [11], [21], consolidándose como el estado del arte actual en este campo.

Las investigaciones en el ISLR se llevan a cabo utilizando conjuntos de datos que contienen videos que muestran únicamente una seña realizada por un único señante. Estos conjuntos de datos desempeñan un papel fundamental y proporcionan ejemplos auténticos de varias lenguas de señas que permiten entrenar y evaluar sus modelos de aprendizaje profundo. La Tabla 1 presenta un resumen de algunos de los conjuntos de datos utilizados en este campo, incluyendo información relevante sobre cada uno de ellos. La información mostrada en la tabla corresponde al año de creación (year), nombre del dataset (dataset), país de origen (country), sigla o nombre corto del dataset (SL), tipo de vídeo (type): grabados con el modelo de color RGB (Rojo, Verde, Azul) o grabados con cámaras de profundidad (D), cantidad de clases o señas (classes), cantidad de muestras (samples), señantes (signers) y tipo de acceso (PA: Acceso Público, CA: Contactar autor, RG: Registro).

El corpus RWTH-BOSTON-50[26] de los Estados Unidos es de acceso público y presenta una variabilidad visual significativa en las señas de la Lengua de Señas Americana (ASL). Por su parte,

Tabla 1. Corpus y conjuntos de datos de la lengua de señas para el ISLR.

Year	Dataset	Country	SL	Type	Classes	Samples	Signers	Access
2005	RWTH-BOSTON-50 [26]	USA	ASL	RGB	50	483	3	PA
2008	Boston ASLLVD [27]	USA	ASL	RGB	2,742	9,794	6	CA
2010	SIGNUM [28]	Germany	DGS	RGB	450	12,150	25	CA
2010	IIITA - ROBITA [29]	India	ISL	RGB	23	–	–	CA
2012	MSR Gesture 3D [30]	USA	ASL	RGB, D	12	336	10	PA
2012	Cooper *et al.* [31]	Greece	GSL	RGB	20	840	6	CA
2012	Cooper *et al.* [31]	Germany	DGS	RGB	40	3,000	15	CA
2013	PSL Kinect 30 [32]	Poland	PSL	RGB, D	30	300	–	PA
2013	PSL ToF 84 [32]	Poland	PSL	RGB, D	84	1,680	1	PA
2015	DEVISIGN-L [33]	China	CSL	RGB, D	2,000	24,000	8	CA
2015	SignsWorld Atlas [34]	Arabia	ArSL	RGB	76	178	4	CA
2016	LSA64 [35]	Argentina	LSA	RGB	64	3,200	10	PA
2016	LSE-Sign [36]	Spain	LSE	RGB	2,400	2,400	2	RG
2020	WLASL2000 [37]	USA	ASL	RGB	2,000	21,083	119	PA
2021	LSFB-ISOL [38]	Belgium	FSL	RGB	395	47,551	85	PA
2023	Balaha *et al.* [39]	Arabia	ArSL	RGB	20	8,467	72	CA

SL: Lengua de señas, ASL: Lengua de señas Americana, DGS: Lengua de señas Alemana, ISL: Lengua de señas India, GSL: Lengua de señas Griega, PSL: Lengua de señas Polaca, CSL: Lengua de señas China, ArsL: Lengua de señas Árabe, LSA: Lengua de señas Argentina, LSE: Lengua de señas Española, FSL: Lengua de señas Francesa.

Boston ASLLVD [27], incluye miles de videos de señas en ASL con anotaciones de inicio y finalización. SIGNUM [28] de Alemania contiene videos tanto de señas aisladas como de oraciones continuas realizadas por diversos señantes de la Lengua de Señas Alemana (DGS). IIITA-ROBITA [29] se empezó a crear desde 2009 en IIIT-Allahabad, conteniendo un total de 23 señas de la Lengua de Señas India (ISL) grabadas en un fondo constante y diversas condiciones de iluminación. MSR Gesture 3D [30] consiste en 12 señas dinámicas de la ASL realizadas por 10 señantes, cada uno realizando las señas 3 veces, presentando amplias variaciones del estilo, velocidad y orientación de la mano. En la investigación de [31] se crearon 2 corpus, el primero consiste en 20 señas de la Lengua de Señas Griega (GSL) realizadas por 6 personas en el mismo entorno con KinectTM, mientras que el segundo corpus está fonéticamente equilibrado de fonemas HamNoSys y consta de 40 señas de la DGS realizadas por 15 personas con diferentes puntos de vista. PSL-Kinect-30 y PSL-ToF-84 [32] se emplearon para el reconocimiento de señas de la Lengua de Señas Polaca (PSL) utilizando sensores de movimiento Kinect y cámaras ToF (Time of Flight) respectivamente, para la detección de profundidad. DEVISIGN-L [33] es el corpus más grande de la Lengua de Señas China (CSL) con 2000 señas y un extenso vocabulario. SignsWorld Atlas [34] es una base de datos de la Lengua de Señas Árabe (ArSL) que contiene varias modalidades: alfabeto, números del 0 al 10, señas estáticas y dinámicas aisladas, oraciones continuas, movimiento de labios y expresiones faciales. LSA64 [35] de Argentina es el único corpus de ISLR en América Latina que ofrece un amplio conjunto de datos con 64 clases de señas y acceso público. LSE-Sign [36] de España es una base de datos en línea gratuita para seleccionar material de estímulo en la LSE, que contiene 2,400 señas y 2,700 no-señas, todos con información detallada. WLASL2000 [37] consiste en videos recopilados de diversos sitios web educativos y tutoriales de YouTube en ASL. LSFB-ISOL [38] de Bélgica es un subconjunto de las 395 señas más recurrentes extraídas de un corpus continuo de la Lengua de Señas Francesa (FSL) introducido en el mismo estudio. Por último, el corpus más reciente es el de [39], que fue creado a través del uso de la cámara de un teléfono móvil en diferentes resoluciones, lugares y fondos. Resultando en 8,467 videos para 20 señas de la ArSL con la colaboración de 72 voluntarios.

Se ha visto que en la mayoría de estos corpus fueron grabados en ambientes controlados, incluyendo aspectos como la iluminación, postura y la velocidad de los movimientos. Este enfoque beneficia la obtención de resultados consistentes dentro del mismo corpus, lo que es importante en la comparación de métodos de preprocesamiento y clasificación. Por otro lado, esto también puede dificultar la implementación de los modelos en ambientes de producción, que carecen de los elementos de los ambientes controlados y son impredecibles. Algunos de los corpus más diversos, como WLASL2000 [37], buscan resolver estas limitaciones y ofrecer una mayor representatividad de cada seña, con una gran cantidad de señantes en distintas configuraciones.

Considerando esta iniciativa, el corpus de la LSCh [40] descrito en las siguientes secciones tiene como objetivo principal presentar una ejecución realista de las señas, en donde aspectos como la postura, velocidad e incluso la misma ejecución de la seña es decisión de los participantes.

METODOLOGÍA

A continuación, se detalla el proceso de preparación de los datos y experimentos realizados en este estudio. La Figura 1 ilustra este proceso de manera visual.

Datos

Utilizamos el corpus LSA64 [35] y un corpus de la LSCh [40], el cual es parte de las Pautas de Evaluación de LSCh en estudiantes sordos, elaboradas en el marco del proyecto *"Centro de asesoría y recursos para establecimientos educacionales con estudiantes sordos/as integrados en el aula"* (2020/2023), desarrollado por la Fundación En Señas del Instituto de la Sordera [41], financiado por la Fundación Olivo [42].

El corpus consta de 48 vídeos, con una duración promedio de 12 minutos. En estos vídeos, estudiantes menores y mayores de 10 años realizan señas siguiendo distintas pautas, algunas de las cuales se repiten en ambas. Estas pautas consisten en una lista de imágenes que representan una idea, por ejemplo: *"Profesor"*, *"Flor"* y "*Noche*". Las imágenes son mostradas a los estudiantes y ellos realizan las señas de la manera que ellos crean correcta.

Para este trabajo, se optó por considerar la pauta de mayores de 10 años, ya que esta categoría incluye un mayor número de señantes, y se seleccionaron los vídeos después de aplicar un filtro que tuvo en cuenta la consistencia y la visibilidad de los gestos realizados. Esto dio como resultado la selección de un total de 16 señantes que cumplían con los criterios establecidos.

Finalmente, las grabaciones de los videos se llevaron a cabo a una velocidad de 30 fotogramas por segundo, con los señantes realizando cada seña en un rango de 1 a 3 repeticiones. Este corpus también presenta la particularidad de contar con señantes zurdos y señantes que adoptan diferentes ángulos de visión y posturas, lo que agrega realismo y complejidad a las señas y mejora su representatividad.

Segmentación

Realizamos la segmentación del corpus de la LSCh de manera manual, siguiendo el formato utilizado en otros corpus del ISLR, como el conocido corpus LSA64 [35]. El resultado de este proceso fue la obtención de un conjunto de datos que consta de un total de 272 vídeos y un vocabulario compuesto por 17 señas distintas, en donde la mayoría de los videos duran menos de 2 segundos. Utilizamos el programa CapCut para recortar los videos, dejando solo las partes donde se realizan las señas. Se ajustó el video para dejar solo al señante en el centro y se recortaron en los fotogramas en donde inicia y termina una seña.

La Figura 2 muestra el desarrollo de la segmentación manual en CapCut. El inicio de una seña se definió

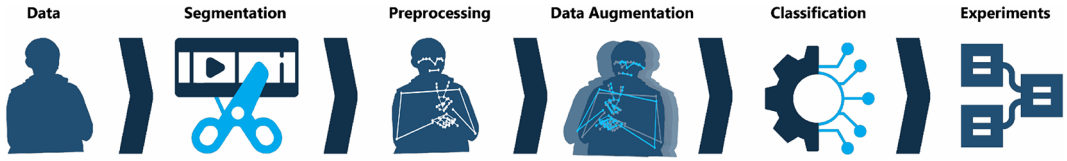


Figura 1. Metodología para la clasificación de señas aisladas utilizando estimación de pose y aprendizaje profundo.

Figura 2. Segmentación de los videos de la LSCh en CapCut.

como el último fotograma en donde el señante levanta las manos para comenzar a realizar una seña, mientras que el término de una seña se definió como el primer fotograma en donde el señante baja las manos después de realizar una seña. De este modo, se descarta el estado transitorio en el que el señante pasa de estar inactivo a estar realizando una seña y viceversa, lo cual es muy importante si se quiere aislar completamente una seña.

Preprocesamiento

En este trabajo se utilizó un enfoque basado en la estimación de la pose de la persona en los videos. Este enfoque fue considerado dado los beneficios que presenta [43]. Dado que los sistemas de ISLR suelen ser computacionalmente pesados y requieren hardware potente para funcionar en tiempo real, los autores proponen un método que es *"2 veces más eficiente computacionalmente y casi 11 veces más rápido que otros métodos en comparación"*.

Algunos trabajos proponen el uso de un estimador de pose para el preprocesamiento de los videos [11], [19]. De este modo, el problema se vería reducido drásticamente a un conjunto de puntos clave que indicarían la posición de las articulaciones de la persona en un momento del vídeo. Adicionalmente, la estimación de pose ayuda también a centrarnos en el problema de la *"postura"* sobre la *"apariencia"* de las personas. De este modo, al capturar la posición de las articulaciones cuando las personas sordas realizan las señas, nos enfocamos en las variables más importantes que diferencian cada seña, como la postura y el movimiento corporal, en lugar de centrarnos en características físicas o estéticas que suelen afectar el proceso de extracción de características.

Debido a su popularidad en los trabajos anteriores, se optó por utilizar la herramienta MediaPipe Holistic [44], que es capaz de estimar la posición de las articulaciones de una persona en un video o incluso en una grabación en tiempo real. Esto la convierte en la herramienta ideal para el reconocimiento de señas ya que uno de los grandes problemas de modelos anteriores es el uso de estos en tiempo real a través de dispositivos portátiles de hardware común. En MediaPipe Holistic, lo primero que se estima es la pose de la persona, formando un esqueleto de 33 puntos clave, como se muestra en la Figura 3. Luego se estiman las manos, cada una de ellas forman un esqueleto de 21 puntos clave. Resultando en 75 puntos clave por fotograma, cada uno de ellos con su respectivo eje x e y. Si tomamos cada número por separado, se puede decir que tenemos un total de 150 valores por imagen. Sin embargo, en este trabajo se descartan algunas coordenadas irrelevantes como las piernas del sujeto, por lo que en total se tienen 118 valores (59 puntos clave). En la Figura 4 se muestran los resultados de este proceso para cada

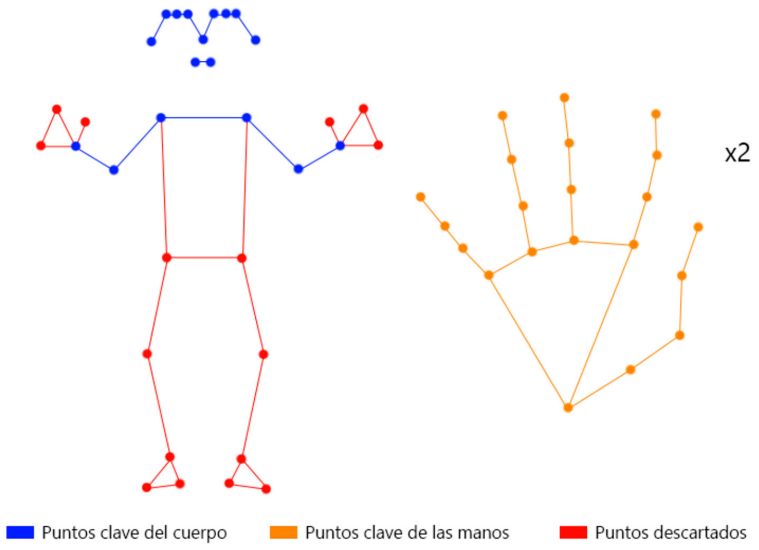


Figura 3. Estimación de pose en MediaPipe Holistics.

Figura 4. Estimación de pose para los señantes del corpus LSCh.

hablante del corpus LSCh, en donde se escogió la seña *"Profesor"* que es realizada por todos los participantes, y se escogió el fotograma con mejor representación para cada una.

Los corpus se reducen a matrices de 3 dimensiones, una dimensión para las muestras, otra para los fotogramas en cada muestra y otra para la cantidad de valores posicionales en cada fotograma (*Muestra* x *Fotogramas* x *Valores*) . Los videos tienen distintas duraciones, esto afecta directamente a la cantidad de fotogramas que contiene cada video, por este motivo se le asigna el valor más grande para cada corpus, siendo 60 fotogramas para el LSCh y 242 para el LSA64. Finalmente, la cantidad de valores se mantiene siempre en 118, resultando en dos matrices iniciales distintas (sin contar el aumento de datos): (272 x 60 x 118) para el LSCh y (3200 x 242 x 118) para el LSA64.

Aumento de datos

El aumento de datos es una práctica común en el aprendizaje profundo, consiste en crear variaciones en los datos a través de transformaciones para generar nuevos datos y reducir el sobreajuste. Es ampliamente implementado en los trabajos de SLR y SLT (Sign Language Transaltion) [43], [45], debido a la poca cantidad de datos disponibles y la gran cantidad de posibles variaciones en la representación de una seña. Por este motivo, aplicamos 2 tipos de transformaciones a los datos. En la Figura 5 se ilustra el resultado de las transformaciones Espejo (*Mirror*) y Ángulo (*Angle*) en comparación a la muestra sin modificar (*Original*).

Espejo

En las lenguas de señas existen señantes zurdos y diestros, ambos ejecutan las señas de la misma forma, pero espejada. Dependiendo de cuál sea la

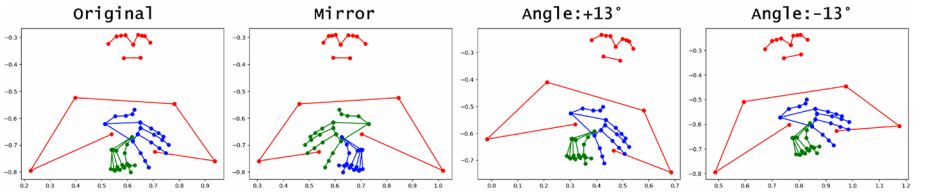


Figura 5. Transformaciones del aumento de datos.

mano dominante, el señante puede realizar la acción principal de una seña con una mano o la otra, y utilizar su mano no dominante para la acción de soporte [46]. Ejemplificando este fenómeno, podemos observar a los señantes (8), (12) y (14) en la Figura 4 realizando señas de forma espejada al resto. Por esta razón, se aplica la transformación de espejo a los corpus con el objetivo de considerar ambos casos. Esta transformación se aplicó directamente a los valores horizontales de los puntos que definen la posición de cada articulación en los datos preprocesados. Para esto, primero se intercambiaron todos los índices de la parte derecha del esqueleto con la izquierda, para prevenir que los modelos aprendan de muestras realmente "espejadas", sino más bien que su formato sea igual al que se obtendría de las muestras de señantes realmente zurdos. Luego se aplicó la transformación definida en la ecuación (1).

$$x = 2x_0 - x \tag{1}$$

En donde x_0 es la posición horizontal de referencia para espejar los valores x. En este trabajo se optó por utilizar la posición en x de la nariz de los señantes como x_0, ya que es el punto más centrado del esqueleto.

Ángulo

Se realizaron otras 2 transformaciones adicionales para considerar situaciones donde el hablante está inclinado, en donde se inclina el esqueleto de los señantes en 13 y –13 grados. Estas transformaciones se han implementado en otros trabajos basados en la pose de los señantes [43], en donde se aplicó esta misma transformación con un efecto positivo. En este trabajo, estas transformaciones se aplicaron sobre los datos preprocesados, por lo que se utilizaron la ecuación (2) y la ecuación (3) para inclinar los puntos del esqueleto.

$$x' = x_0 + \cos(A)\cdot(x - x_0) - \sin(A)\cdot(y - y_0) \tag{2}$$

$$y' = y_0 + \sin(A)\cdot(x - x_0) + \cos(A)\cdot(y - y_0) \tag{3}$$

Donde (x_0, y_0) es el punto correspondiente a la posición de la nariz del esqueleto, (x, y) es el punto que se quiere transformar, (x', y') es el resultado y A es el ángulo de inclinación.

Clasificación

El ISLR es un problema de clasificación supervisada que requiere de modelos de aprendizaje profundo para clasificar las señas y un conocimiento previo de estas, que corresponde al etiquetado de las muestras en los corpus [47]. Se han aplicado varias arquitecturas y métodos para los modelos de clasificación de señas, siendo el uso de Transformers lo más reciente y eficaz.

En general, los modelos *Transformer* adoptan una estructura de codificador-decodificador, diseñada originalmente para la tarea de transducción de secuencias, la cual implica la conversión de una secuencia de entrada en una secuencia de salida. Sin embargo, en el ámbito del ISLR, los enfoques que emplean estos modelos neuronales tienden a utilizar únicamente la parte del codificador, seguida por una red neuronal para la clasificación secuencial. Por este motivo, en la Figura 6 se describe la arquitectura utilizada en este estudio, siguiendo la estructura de modelos de otros estudios similares [11], [48]. La arquitectura de este modelo se basa en funciones de atención multicabezal (*Multi-Head Attention*) ilustrada en la Figura 7, que consiste en varias funciones de atención (*Scaled Dot Product Attention*) ejecutándose en paralelo. Una función de atención recibe como parámetros las matrices Q, K y V obtenidas de las transformaciones lineales entrenables $Q = XW_q$, $K = XW_k$ y $V = XW_v$ dada una secuencia de entrada $X \in \mathbb{R}^{N\times d}$ con una cantidad de N elementos y una cantidad de d características por elemento. Luego, las funciones de atención se calculan con la ecuación (4).

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

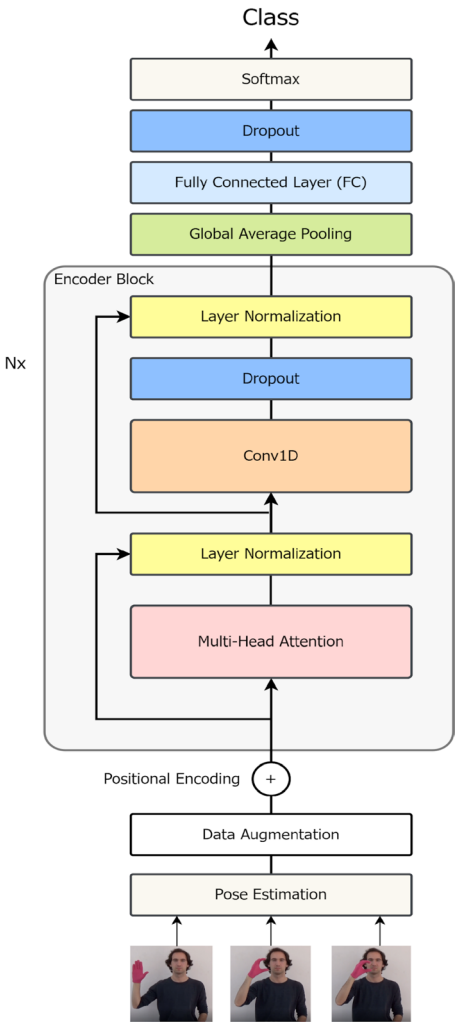


Figura 6. Arquitectura del modelo utilizado, basada en redes *Transformer*.

Fuente. Obtenido de [25].

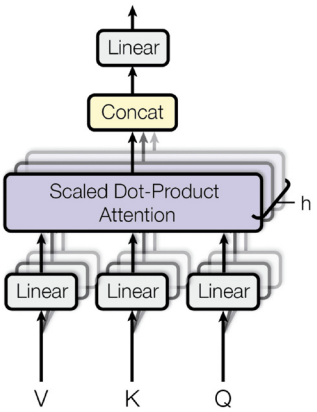


Figura 7. Arquitectura de la atención multicabezal.

El denominador d_k de la ecuación (4) es la dimensionalidad de las características de entrada, y es utilizado para prevenir la saturación debido a grandes incrustaciones en la función *softmax*. Una vez calculadas las funciones de atención, se continúa con la atención multicabezal, representada en la ecuación (5) y ecuación (6).

$$MultiHead(Q,K,V) = Concat(h_1,\ldots,h_n)W^o \quad (5)$$

$$h_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (6)$$

En la atención multicabezal, cada función de atención se realiza en un subconjunto de la entrada *X*, conocida como cabezal *h*. Finalmente, todas las funciones de atención son concatenadas y normalizadas utilizando *Layer Normalization* [49], que mejora la estabilidad y eficiencia del proceso. A diferencia de la normalización por lotes (*Batch Normalization*), la normalización por capas es invariable al cambio y escala de características por caso de entrenamiento.

Finalmente, siguiendo la estructura de modelos de otros estudios similaresv [48], [50], se implementaron capas de redes neuronales convolucionales en una dimensión (*CNN-1D*) que se especializan en analizar y procesar datos en una dimensión. A diferencia de las *CNN-2D* que se utilizan comúnmente para la clasificación de imágenes, las *CNN-1D* son especialmente útiles para trabajar con series de tiempo con propiedades espaciales, como lo son los videos de la lengua de señas. También se aplica la técnica de regularización *Dropout* para prevenir el sobreajuste, descartando unidades en una red neuronal.

Experimentos

En los experimentos, utilizamos el corpus de la LSCh y el corpus LSA64 como referencia. En el primer experimento evaluamos el impacto del aumento de datos en el entrenamiento y rendimiento de los modelos. Se probaron configuraciones sin aumento, con aumentos individuales y con todos los aumentos. En el segundo experimento, evaluamos el rendimiento de los modelos previos en señas realizadas por nuevos señantes, incluyendo señantes no usados en el primer experimento para el entrenamiento.

Experimento 1

El primer experimento se realizó para determinar el efecto que tiene el aumento de datos en el rendimiento

de los modelos en general. Se exploraron cuatro casos: en el primero, no se aplicó ningún tipo de aumento; en el segundo y el tercero, se aplicaron los aumentos por separado, y en el último se aplicaron todos juntos.

Considerando estas configuraciones, se entrenaron 4 modelos distintos por cada conjunto de datos con una división del (70%) de las muestras para el entrenamiento y un (30%) para las pruebas. Adicionalmente, se considera un (15%) de las muestras de entrenamiento para la validación (durante el entrenamiento) y la cantidad de muestras varía según el aumento de datos. En la Figura 8 se muestran las gráficas de la exactitud y pérdida en el entrenamiento de los modelos para los dos corpus (a) LSCh y (b) LSA64. La línea vertical representa la parada temprana del entrenamiento para evitar sobreajuste.

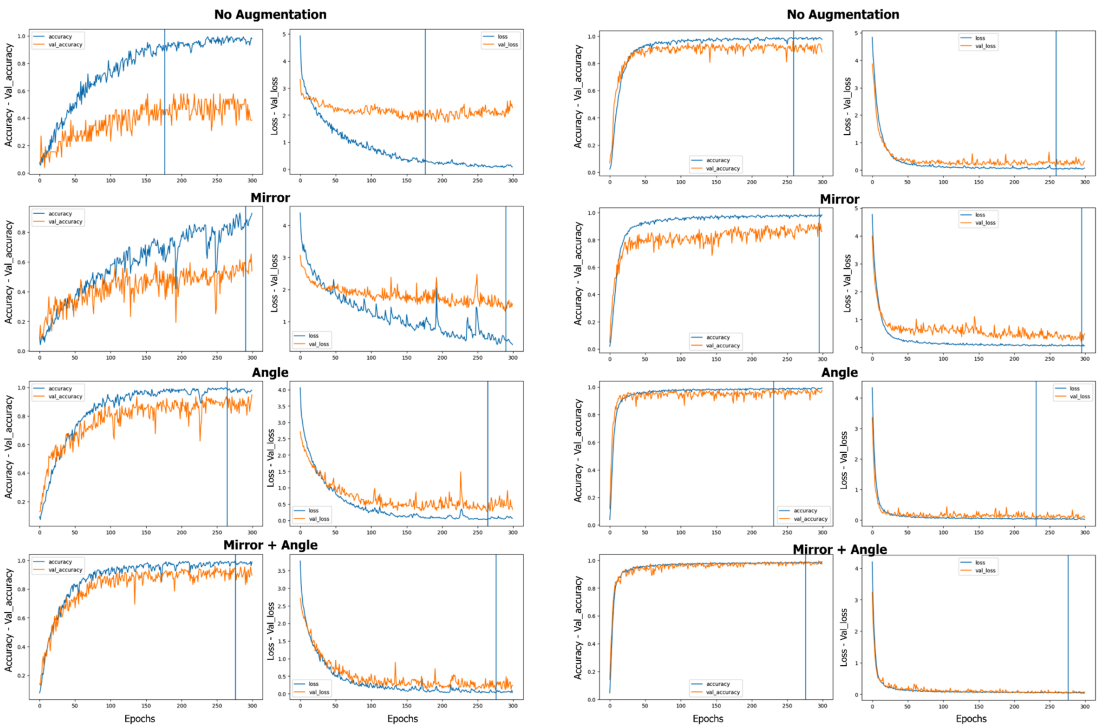
Experimento 2

En el segundo experimento, se realizó una evaluación del rendimiento de los modelos generados en el primer experimento para reconocer las señas realizadas por nuevos señantes. Con el objetivo de determinar el rendimiento del modelo para clasificar señas realizadas por nuevos señantes. Para cada corpus, se utilizó el modelo con mejor exactitud (*Mirror + Angle*), cuyos resultados son comparados con los obtenidos en este experimento.

Los señantes utilizados en este experimento fueron un señante del corpus LSA64 y los señantes (15) y (16) del corpus LSCh, enumerados en la Figura 4, que no fueron utilizados en el primer experimento.

RESULTADOS EXPERIMENTALES

En el corpus LSA64, que consta únicamente de señantes diestros, los resultados demostraron que entrenar un solo modelo para la dualidad de señantes diestros y zurdos no afecta considerablemente el rendimiento. Esto se observa en los casos de *"Sin Aumento"* y *"Espejo"* de la Tabla 2 donde ambos alcanzaron una exactitud (accurary) del 92%. Además, este enfoque equilibró la cantidad de muestras de señantes zurdos y diestros en el dataset LSCh, mejorando la exactitud del modelo del 87% al 95%, como se muestra en la Tabla 3. Por otro lado, el aumento de Ángulo triplicó la cantidad de



(a) Corpus LSCh. (b) Corpus LSA64.

Figura 8. Gráficas del entrenamiento de los modelos.

Tabla 2. Promedio de los resultados para los modelos del primer experimento en el corpus LSA64.

LSA64	Samples	Accuracy	Precision	Recall	F1-score
No Augmentation	2880	0,92	0,94	0,92	0,92
Mirror	5760	0,92	0,93	0,92	0,91
Angle	8640	0,98	0,98	0,98	0,98
Mirror + Angle	17280	0,99	0,99	0,99	0,99

Tabla 3. Promedio de los resultados para los modelos del primer experimento en el corpus LSCh.

LSCh	Samples	Accuracy	Precision	Recall	F1-score
No Augmentation	244	0,51	0,56	0,5	0,49
Mirror	488	0,62	0,61	0,6	0,58
Angle	732	0,87	0,89	0,88	0,87
Mirror + Angle	1464	0,95	0,95	0,95	0,95

muestras (samples) en ambos corpus y aumentó significativamente la exactitud de los modelos, llevando al modelo LSCh del 62% al 95% y al modelo LSA64 del 92% al 99%.

En los resultados presentados en la Tabla 4, se evidencia que el modelo LSCh experimentó una disminución en su exactitud, pasando del 95% al 75%. En contraste, el modelo LSA64 también mostró una disminución en la exactitud, aunque menos significativa, cayendo del 99% al 94%.

Esta ligera disminución en el modelo LSA64 podría atribuirse a la similitud en la postura de los señantes presentes en el corpus LSA64, lo que indica que el modelo ha aprendido de mejor manera esa postura.

También se pudo ver que algunos errores en la clasificación del corpus LSCh ocurrieron debido a señas muy parecidas unas de otras. Por ejemplo, en la Figura 9 se puede ver como la clase 10 (Leer) y

Tabla 4. Resultados del segundo experimento en el corpus LSCh y LSA64 utilizando los modelos con mejor rendimiento (Mirror + Angle).

Models	Accuracy	Precision	*Recall*	F1-score
LSCh	0,75	0,83	0,74	0,73
LSA64	0,94	0,96	0,94	0,94

la clase 8 (Flor) consisten en juntar las manos para realizar un movimiento ascendente. Esta situación también está presente en el corpus LSA64, en donde la clase 6 (Colors) se confunde con la clase 10 (Son), ya que ambas señas son casi idénticas, solo que una se realiza con un dedo y la otra con dos.

DISCUSIÓN DE RESULTADOS

Creamos un pequeño conjunto de datos para el ISLR a partir de la segmentación de videos de estudiantes sordos en el contexto de la LSCh, con

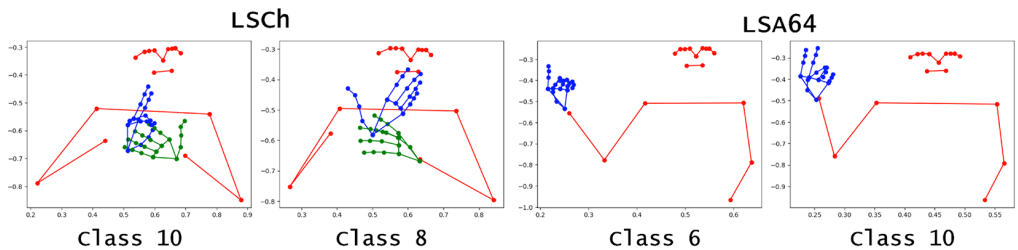


Figura 9. Ejemplos de señas similares en los corpus.

una cantidad inicial de 272 de las cuales 28 fueron utilizadas en el segundo experimento, dejándonos solamente con un total de 244 para el entrenamiento. Sin embargo, la aplicación de técnicas de aumento de datos permitió ampliar la cantidad limitada de muestras en este conjunto de datos a un total de 6 veces su tamaño original, resultando en una escala adecuada para entrenar los modelos, lo cual es muy importante en este trabajo, ya que una de las desventajas de los Transformers es la necesidad de una gran cantidad de datos en el entrenamiento.

Los resultados en el primer experimento muestran una notable capacidad de un modelo para aprender eficazmente tanto las señas de señantes diestros como zurdos simultáneamente, evitando la necesidad de entrenar 2 modelos separados para abordar ambos casos. Este enfoque también contribuyó a mejorar la representación de las señas en algunos casos, como en el modelo LSCh que mejoró su exactitud de su original 51% a un 62%, y no mostró ser un detrimento para el rendimiento en ningún modelo del LSA64, manteniendo su exactitud de 92%.

Sin embargo, es importante reconocer que la cantidad de datos sigue siendo insuficiente para capturar todas las variaciones en la postura y perspectiva de los señantes. La complejidad de la lengua de señas implica una amplia expresividad de las señas que es muy improbable que sea completamente recopilada en un corpus a través de métodos convencionales.

A pesar de este gran desafío, la comparación en el rendimiento de ambos corpus mostró que la dependencia de una gran cantidad de señantes para un mejor rendimiento en la clasificación de señas no es tan importante como lo son las variaciones en la postura y perspectiva, como se pudo ver en las pruebas de nuevos señantes del segundo experimento. Esto sugiere que, en el enfoque de postura, la mejor forma de abordar la escasez de datos es la implementación de más transformaciones del aumento de datos que simulen la participación de nuevos señantes a partir de un conjunto de señantes limitado. Las transformaciones son realizadas en los datos preprocesados, por lo que son fáciles de implementar y replicar en comparación a la modificación directa de los videos. De este modo, no solo se puede mejorar la capacidad de generalización del modelo, sino que también puede mitigar la necesidad de depender exclusivamente de corpus masivos.

Un factor muy importante que se debe considerar es la diferencia de fotogramas por segundo entre el corpus LSCh y LSA64, que son 30 y 60 respectivamente. Lo cual también pudo haber influenciado en los resultados. Considerando lo rápido que se realizan las señas en un escenario realista, es recomendable que se tome en cuenta en la evaluación de los modelos y se sugiere que se use una velocidad de fotogramas de 60 o mayor en futuras grabaciones para crear corpus de señas.

Otro desafío que se identificó en los resultados fueron las señas similares, el cual se presenta en la ocurrencia de señas que son confundidas por otras señas. En las métricas, el "rendimiento" de una seña en individual se puede medir con *Sensibilidad* ya que, en el caso del ISLR, es más importante validar que la seña que se está evaluando se clasifique correctamente en lugar de medir la probabilidad de otras señas de confundirse con esta (*Precisión*). De esta manera, se pudieron identificar individualmente las señas más problemáticas, pero por ahora no se tiene una solución para este desafío en específico.

Las transformaciones de aumento de datos mejoraron drásticamente el rendimiento de los modelos, pero aún existen desafíos en la limitación de los datos y la confusión entre señas similares. La diferencia de velocidad de fotogramas es un factor para tomar en consideración. Es posible entrenar un modelo con una cantidad limitada de datos si se implementan técnicas de aumento de datos, que incluso pueden ser más factibles e igual de beneficiosas que recopilar más muestras con nuevos señantes. Finalmente, se recomienda prestar especial atención a la diversidad y representatividad de las señas, sobre todo al abordar el desafío de la semejanza visual entre señas y que la representatividad no se vea afectada por un tipo de aumento de datos que no sea compatible o genere aún más confusión entre señas.

CONCLUSIONES

El ISLR juega un papel fundamental en superar las barreras lingüísticas para la comunidad sorda, contribuyendo a la inclusión y accesibilidad. Las lenguas de señas, que poseen una complejidad lingüística única, no son simplemente herramientas auxiliares, sino lenguas naturales que reflejan la cultura y la identidad de la comunidad sorda. Las investigaciones en este campo impulsan el desarrollo

de tecnologías que facilitan la comunicación efectiva entre personas sordas y oyentes, que en un futuro actuarán como un puente entre dos mundos lingüísticos y culturales.

En este estudio se creó un conjunto de datos mediante la segmentación manual de videos de estudiantes sordos realizando señas de la LSCh en diversas posturas y perspectivas. Se utilizó la herramienta *MediaPipe Holistic* para estimar la postura de los señantes en los videos y con estos datos se entrenaron modelos basados en Transformers con aprendizaje profundo para crear modelos de clasificación de señas, aplicando técnicas de aumento de datos. Los resultados fueron comparados con los obtenidos en el corpus LSA64 y resaltan los desafíos en el ISLR para identificar todas las variaciones de postura y perspectiva que se encuentran en un escenario realista, pero también mostró la importancia del aumento de datos para mejorar la capacidad de reconocimiento de señas y su efectividad para entrenar modelos duales que pueden clasificar señas tanto de señantes diestros como de señantes zurdos.

Este enfoque propone una solución escalable al problema de la escasez de datos para el ISLR, donde la prioridad no recae en la recopilación de vídeos de la lengua de señas, sino en la extracción de sus puntos clave y la transformación de estos puntos para abordar nuevas posturas y perspectivas a través del aumento de datos. Sin embargo, aún se requiere de más experimentación para refinar y validar plenamente la eficacia de esta metodología.

En las futuras investigaciones, se plantean diversas tareas. Estas incluyen la expansión del conjunto de datos de la LSCh, la investigación y desarrollo de enfoques de aumento de datos innovadores para potenciar el reconocimiento de la lengua de señas. En cuanto al método de clasificación, se requiere la exploración de otras técnicas que aborden el desafío de señas similares y mitiguen el riesgo de sobreajuste en modelos de reconocimiento. Debe considerarse también especial atención a la variabilidad en las perspectivas y posturas de los señantes, focalizándose en el entrenamiento de modelos robustos capaces de abordar estas diferencias de manera efectiva. Otra línea que seguir, complementaria a la propuesta en este trabajo, es la implementación de las técnicas utilizadas en el corpus no segmentado, abordando el desafío del CSLR.

REFERENCIAS

[1] M. González, A.P. Cuello, J.L. Marín, and C. Villavicencio, "Towards the recognition of Chilean sign language," *The legal recognition of sign languages: Advocacy and outcomes around the world*, M. De Meulder, J.J. Murray and R.L. McKee, Eds., Pennsylvania, USA: Multilingual Matters, 2019, pp. 129-144, doi: 10.21832/9781788924016-010.

[2] M. Vázquez-Enríquez, J.L. Alba-Castro, L. Docío-Fernández and E. Rodríguez-Banga, "Isolated Sign Language Recognition with Multi-Scale Spatial-Temporal Graph Convolutional Networks," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* Nashville, TN, USA, 2021, pp. 3457-3466, doi: 10.1109/CVPRW53098.2021.00385.

[3] Servicio Nacional de la Discapacidad, "Resultados CENSO 2012 en Discapacidad," senadis.gob.cl. [En línea]. Disponible en: www.senadis.gob.cl/sala_prensa/d/noticias/2990/

[4] Servicio Nacional de la Discapacidad, "II Estudio Nacional de la Discapacidad," senadis.gob.cl. [En línea]. Disponible en: www.senadis.gob.cl/pag/355/1197/

[5] M. Lissi *et al.*, "La exclusión de la comunidad sorda del sistema de atención en salud mental. Hacia un modelo que respete sus características culturales y lingüísticas", en *Propuestas para Chile. Concurso de Políticas Públicas 2020*, Santiago, Chile: Pontificia Universidad Católica de Chile, 2021, pp. 17-47. [En línea]. Disponible: https://politicaspublicas.uc.cl/publicacion/propuestas-para-chile-2020/

[6] P. Levin Blanco, "Aprendizaje de la Lengua de Señas Chilena (LSCh) y cultura sorda, en la escuela de sordos 'Santiago Apóstol' ", Tesis de Grado, Facultad de Humanidades, UCINF, Santiago, Chile, 2016. [Online]. Disponible en: https://repositorio.ugm.cl/handle/20.500.12743/1028

[7] M.L. Mora-Olate, "La comunidad sorda: un desafío bilingüe e intercultural crítico para la educación", *Revista Electrónica de Investigación Educativa*, vol. 24, Art. no. e3r, 2022, doi: 10.24320/redie.2022.24.e3r.5737.

[8] Biblioteca del Congreso Nacional, "Biblioteca del Congreso Nacional: Ley

20422. Establece normas sobre igualdad de oportunidades e inclusión social de personas con discapacidad", bcn.cl. [En línea]. Disponible en: www.bcn.cl/leychile/navegar?idLey=20422

[9] N. Sarhan and S. Frintrop, "Unraveling a Decade: A Comprehensive Survey on Isolated Sign Language Recognition," in *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW),* Paris, France, 2023, pp. 3202-3211, doi: 10.1109/ICCVW60793.2023.00345.

[10] N. Adaloglou *et al.*, "A Comprehensive Study on Deep Learning-Based Methods for Sign Language Recognition," in *IEEE Transactions on Multimedia*, vol. 24, pp. 1750-1762, 2022, doi: 10.1109/TMM.2021.3070438.

[11] S. Alyami, H. Luqman, and M. Hammoudeh, "Isolated Arabic Sign Language Recognition Using A Transformer-based Model and Landmark Keypoints," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 1, 2023, doi: 10.1145/3584984.

[12] M.J. Cheok, Z. Omar, and M.H. Jaward, "A review of hand gesture and sign language recognition techniques," *International Journal of Machine Learning and Cybernetics*, vol. 10, pp. 131-153, 2019, doi: 10.1007/s13042-017-0705-5.

[13] R. Rastgoo, K. Kiani, and S. Escalera, "Hand pose aware multimodal isolated sign language recognition," *Multimedia Tools and Applications*, vol. 80, pp. 127-163, 2021, doi: 10.1007/s11042-020-09700-0.

[14] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert Systems with Applications*, vol. 164, Art. no. 113794, 2021, doi: 10.1016/j.eswa.2020.113794.

[15] A. Wadhawan and P. Kumar, "Deep learning-based sign language recognition system for static signs," *Neural Computing and Applications*, vol. 32, pp. 7957-7968, 2020, doi: 10.1007/s00521-019-04691-y.

[16] G. Ailarus y M. Espinoza, "Abecedario Lenguaje de Señas Chileno-Español," kaggle.com. [En línea]. Disponible en: www.kaggle.com/datasets/bpablo27/abecedario-lenguaje-de-seas-chilenoespaol

[17] O.M. Sincan and H.Y. Keles, "Using motion history images with 3D convolutional networks in isolated sign language recognition," *IEEE Access*, vol. 10, pp. 18608-18618, 2022, doi: 10.1109/ACCESS.2022.3151362.

[18] M. De Coster, M. Van Herreweghe, and J. Dambre, "Isolated Sign Recognition from RGB Video using Pose Flow and Self-Attention," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA, 2021, pp. 3436-3445, doi: 10.1109/CVPRW53098.2021.00383.

[19] D. Laines, M. Gonzalez-Mendoza, G. Ochoa-Ruiz, and G. Bejarano, "Isolated Sign Language Recognition based on Tree Structure Skeleton Images," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* Vancouver, BC, Canada, 2023, pp. 276-284, doi: 10.1109/CVPRW59228.2023.00033.

[20] T. Starner and A. Pentland, "Real-time American Sign Language recognition from video using hidden Markov models," in *Proceedings of International Symposium on Computer Vision - ISCV*, Coral Gables, FL, USA, 1995, pp. 265-270, doi: 10.1109/ISCV.1995.477012.

[21] M. Hrúz, I. Gruber, J. Kanis, M. Boháček, M. Hlaváč, and Z. Krňoul, "One model is not enough: Ensembles for isolated sign language recognition," *Sensors*, vol. 22, no. 13, pp. 5043, 2022, doi: 10.3390/s22135043.

[22] K.M. Lim, A.W.C. Tan, C.P. Lee, and S.C. Tan, "Isolated sign language recognition using convolutional neural network hand modelling and hand energy image," *Multimedia Tools and Applications,* vol. 78, pp. 19917-19944, 2019, doi: 10.1007/s11042-019-7263-7.

[23] A.O. Tur and H. Y. Keles, "Evaluation of hidden Markov models using deep CNN features in isolated sign recognition," *Multimedia Tools and Applications,* vol. 80, pp. 19137-19155, 2021, doi: 10.1007/s11042-021-10593-w.

[24] L. Yin, H. Ying, D.K. Liu, R. li, and M.H. Yang , "Chinese Sign Language Recognition Based on Two-stream CNN and LSTM Network," *International Journal of Advanced Networking and Applications*, vol. 14, no. 6, pp. 5666-5671, 2023, doi: 10.35444/IJANA.2023.14602.

[25] A. Vaswani *et al.*, "Attention is all you need," 2017, arXiv: 1706.03762.

[26] M. Zahedi, D. Keysers, T. Deselaers, and H. Ney, "Combination of Tangent Distance and an Image Distortion Model for Appearance-Based Sign Language Recognition," in *Pattern Recognition. DAGM 2005, en Lecture Notes in Computer Science*, vol. 3663, W.G. Kropatsch, R. Sablatnig, A. Hanbury, Eds., Vienna, Austria, Ago. 2005, pp. 401-408, doi: 10.1007/11550518_50.

[27] V. Athitsos *et al.*, "The American Sign Language Lexicon Video Dataset," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Anchorage, AK, USA, 2008, pp. 1-8, doi: 10.1109/CVPRW.2008.4563181.

[28] U. von Agris and K.F. Kraiss, "SIGNUM Database: Video Corpus for Signer-Independent Continuous Sign Language Recognition," in *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martínez Ruiz, and A. Schembri, Eds., Valletta, Malta, 2010, pp. 243-246. [Online]. Available: https://www.sign-lang.uni-hamburg.de/lrec/pub/10006.pdf

[29] A. Nandy, S. Mondal, J.S. Prasad, P. Chakraborty, and G.C. Nandi, "Recognizing & interpreting Indian sign language gesture for human robot interaction," in *2010 International Conference on Computer and Communication Technology (ICCCT)*, Allahabad, India, 2010, pp. 712-717, doi: 10.1109/ICCCT.2010.5640434.

[30] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *2012 Proceedings of the 20th European signal processing conference (EUSIPCO)*, Bucharest, Romania, August 27-31, 2012, pp. 1975-1979.

[31] H.M. Cooper, E.J. Ong, N. Pugeault, and R. Bowden, "Sign language recognition using sub-units," *Journal of Machine Learning Research*, vol. 13, no. 72, pp. 2205-2231, 2012, doi: 10.1007/978-3-319-57021-1_3.

[32] M. Oszust and M. Wysocki, "Polish sign language words recognition with Kinect," in *2013 6th International Conference on Human System Interactions (HSI)*, Sopot, Poland, 2013, pp. 219-226, doi: 10.1109/HSI.2013.6577826.

[33] X. Chai, H. Wanga, M. Zhoub, G. Wub, H. Lic, and X. Chena, "Devisign: dataset and evaluation for 3D sign language recognition," *Technical report, Beijing, Tech. Rep*, 2015.

[34] S.M. Shohieb, H.K. Elminir, and A.M. Riad, "Signsworld atlas; a benchmark Arabic sign language database," *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 1, pp. 68-76, 2015, doi: 10.1016/j.jksuci.2014.03.011.

[35] F. Ronchetti, F. Quiroga, C. Estrebou, L. Lanzarini, and A. Rosete, "LSA64: A Dataset of Argentinian Sign Language," 2016, arXiv: 2310.17429.

[36] E. Gutierrez-Sigut, B. Costello, C. Baus, and M. Carreiras, "LSE-sign: A lexical database for spanish sign language," *Behavior Research Methods*, vol. 48, pp. 123-137, 2016, doi: 10.3758/s13428-014-0560-1.

[37] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison," in *2020 IEEE Winter Conference on Applications of Computer Vision(WACV)*, Snowmass, CO, USA, 2020, pp. 1448-1458, doi: 10.1109/WACV45572.2020.9093512.

[38] J. Fink, B. Frénay, L. Meurant, and A. Cleve, "LSFB-CONT and LSFB-ISOL: Two new datasets for vision-based sign language recognition," in *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China, 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9534336.

[39] M.M. Balaha *et al.*, "A vision-based deep learning approach for independent-users Arabic sign language interpretation," *Multimedia Tools and Applications*, vol. 82, no. 5, pp. 6807-6826, 2023, doi: 10.1007/s11042-022-13423-9.

[40] F. Otárola *et al.*, "Desarrollo de un instrumento de evaluación de conocimiento léxico para la Lengua de Señas Chilena", *Lengua y Sociedad*, vol. 23, no. 1, pp. 751-778, 2024, doi: 10.15381/lengsoc.v23i1.27540.

[41] Instituto de la Sordera, "Fundación En Señas", institutodelasordera.cl. [En línea]. Disponible en: www.institutodelasordera.cl/instituto/fundacion-en-senas/

[42] P. Matte, "Fundación Olivo", olivo.org. [En línea]. Disponible en: www.olivo.org

[43] M. Boháček, Z. Cao, and M. Hrúz, "Combining Efficient and Precise Sign Language Recognition: Good pose estimation library is all you need," 2022, arXiv: 2210.00893.

[44] I. Grishchenko and V. Bazarevsky, "MediaPipe holistic - simultaneous face, hand and pose prediction, on device," research.google, [Online]. Available: www.research.google/blog/mediapipe-holistic-simultaneous-facehand-and-pose-prediction-on-device/

[45] A. Moryossef, K. Yin, G. Neubig, and Y. Goldberg, "Data augmentation for sign language gloss translation," 2021, arXiv: 2105.07476.

[46] S. Val, "La lingüística de las lenguas de señas: la no-inversión de algunas señas por parte de los señantes zurdos como argumento a favor de una perspectiva rupturista basada en la iconicidad", *Quintú Quimün*, no. 4, Art. no. Q025, 2020. [En línea]: Disponible: https://revele.uncoma.edu.ar/index.php/lingustica/article/view/2920

[47] F. Wei and Y. Chen, "Improving Continuous Sign Language Recognition with Cross-Lingual Signs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 2023, pp. 23612-23621, doi: 10.1109/ICCV51070.2023.02158.

[48] M. De Coster, M. Van Herreweghe, and J. Dambre, "Sign language recognition with transformer networks," in *12th International Conference on Language Resources and Evaluation*, 2020, pp. 6018-6024.

[49] J.L. Ba, J.R. Kiros, and G.E. Hinton, "Layer normalization," 2016, *arXiv: 1607.06450*.

[50] M. Marais, D. Brown, J. Connan, and A. Boby, "An Evaluation of Hand-Based Algorithms for Sign Language Recognition," in *2022 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, Durban, South Africa, 2022, pp. 1-6, doi: 10.1109/icABCD54961.2022.9856310.