ISSN 1870-4069

Interpretación en tiempo real de lengua de señas mexicana con CNN y HMM

Jairo Enrique Ramírez Sánchez, Arely Anguiano Rodríguez,
Miguel González Mendoza

Tecnológico de Monterrey,
Escuela de Ingeniería y Ciencias,
México

{A01750443, A01752068}@itesm.mx, mgonza@tec.mx

Resumen. La lengua de señas mexicana (LSM) es la primordial forma de comunicación de la comunidad sorda en México. La LSM cuenta con una estructura gramatical diferente al español; además, la expresión facial juega un papel determinante a la hora de complementar el significado basado en el contexto. Esto dificulta que una persona oyente sin conocimientos previos de la lengua comprenda lo que se desea transmitir, presentando una importante barrera de comunicación a las personas sordas. Para lidiar con esto, presentamos la primera arquitectura en considerar los rasgos faciales como indicadores del tiempo gramatical para desarrollar un intérprete en tiempo real de LSM a español escrito. Nuestro modelo usa la librería de código abierto de MediaPipe para extraer marcas de la cara, posición corporal y manos. Se utilizan tres redes neuronales convolucionales 2D para codificar individualmente y extraer patrones, las redes convergen en un perceptrón multicapa para la clasificación. Finalmente, se utiliza un Modelo Oculto de Markov para predecir morfosintácticamente la secuencia de palabras más probable con forme a una base de conocimiento precargada. De los experimentos realizados, se obtuvo una precisión del 94.9 % $\sigma = 0{,}07$ para el reconocimiento de 75 palabras aisladas y 94.1 % $\sigma = 0{,}09$ para la interpretación de 20 frases en LSM en contexto médico. Al ser una aproximación basada en entradas de cámara y al observar que incluso con pocos ejemplos se puede lograr una adecuada generalización, nuestra arquitectura resultaría factible para ser escalada a otras lenguas de señas y brindar posibilidades de una comunicación eficiente a millones de personas con discapacidad auditiva.

Palabras clave: Lengua de señas mexicana, redes neuronales convolucionales, modelos ocultos de Markov, intérprete en tiempo real.

Real-Time Mexican Sign Language Interpretation Using CNN and HMM

Abstract. Mexican Sign Language (MSL) is the primary form of communication for the deaf community in Mexico. MSL has a different

grammatical structure than Spanish; furthermore, facial expression plays a determining role in complementing context-based meaning. This turns it difficult for a hearing person without prior knowledge of the language to understand what is to be transmitted, representing an important communication barrier for deaf people. In order to face this, we present the first architecture to consider facial features as indicators of grammatical tense to develop a real-time interpreter from LSM to written Spanish. Our model uses the open source MediaPipe library to extract marks from the face, body position and hands. Three 2D convolutional neural networks are used to encode individually and extract patterns, the networks converge to a multilayer perceptron for classification. Finally, a Hidden Markov Model is used to morphosyntactically predict the most probable sequence of words based on a preloaded knowledge base. From the experiments were carried out, a precision of 94.9% was obtained with $\sigma = 0.07$ for the recognition of 75 isolated words and 94.1% with $\sigma = 0.09$ for the interpretation of 20 sentences in LSM in a medical context. Being an approach based on camera inputs and observing that even with a few examples an adequate generalization can be achieved, our architecture would be feasible to be scaled to other sign languages and offer possibilities of efficient communication to millions of people with hearing disability.

Keywords: Mexican sign language, convolutional neural networks, hidden Markov models, real time interpreter.

1. Introducción

En México, existen 2.3 millones de personas con discapacidad auditiva (PDA) según datos del Instituto Nacional de Estadística y Geografía para la Información y el Consejo Nacional para la Prevención de la Discriminación 2020. La lengua oficial de la comunidad sorda en México es la Lengua de Señas Mexicana (LSM), destacando que cada país posee un sistema de signos nacional designado para su territorio.

Para lograr una comunicación efectiva y eficiente entre una PDA y una Persona Oyente (PO), no es suficiente con que se aprenda a identificar los movimientos de las manos para designar a cada una de las palabras, ya que los rasgos manuales (RM) representan tan sólo el $20\,\%$ de la comunicación total, mientras que los no manuales (RNM) figuran como el resto [**?**]. Por ello, también es necesario tener conocimiento de su gramática y de los rasgos no manuales, en concreto, los faciales, pues estos son los que indican el tiempo verbal en que la acción se llevó, lleva o llevará a cabo.

El tiempo pasado se expresa a través de un semblante de que la acción ha sido consumada, en este se puede observar una expresión relajada, donde el labio inferior sobresale del superior asintiendo suavemente con la cabeza; por otra parte, en el tiempo presente se proyecta una expresión neutra, alzando ligeramente las cejas y acorde con el mensaje; por último, el tiempo futuro se manifiesta a través de una expresión de duda o pensamiento, donde el cuello se rota ligeramente y los ojos se dirigen hacia cualquiera de las dos esquinas superiores del plano visual. Acerca de la gramática LSM, los verbos ser y estar se ven omitidos, por lo que se expresa únicamente el sustantivo y adjetivo o, en su defecto, sustantivo y lugar.

Igualmente, es relevante que el movimiento de cada seña se realiza con los dedos en posición de la letra con la cual inicia el vocablo asignado a la seña, o bien, destacando características sensoriales de la palabra que se trata. Como se puede percibir, para una persona no inmersa en un ambiente y sin un contacto previo a la comunidad sorda, es de suma dificultad alcanzar un entendimiento total de lo que la PDA busque comunicar.

Esto, aunando al hecho de que en México solo existan 42 intérpretes certificados en LSM [**?**], hace que la comunicación para las PDA esté prácticamente cercada. Existe una vasta investigación en lengua de señas americana (ASL, por sus siglas en inglés), lengua de señas británica (BSL, por sus siglas en inglés), lengua de señas china (CSL, por sus siglas en inglés), entre otras; sin embargo, la incursión en el estudio de LSM es más que limitada.

En nuestra investigación, abordaremos el análisis y reconocimiento de señas en tiempo real haciendo uso de redes neuronales de arquitectura convolucional para la clasificación de señas tomando en cuenta tres aspectos: la expresión facial, movimientos de las manos y posición corporal; la interpretación es mejorada por un Modelo Oculto de Markov para el enriquecimiento con contexto y gramática. El presente artículo está organizado de la siguiente manera: La sección dos aborda la revisión del trabajo previo sobre el reconocimiento en tiempo real de lenguas de señas.

La sección tres contiene la descripción de nuestro modelo propuesto, así como las fases del procesamiento. La sección cuatro explica la forma en cómo se realizó la adquisición de los datos, el número de participantes y las condiciones de prueba. En la sección cinco se presentan los experimentos propuestos y sus respectivos resultados. Finalmente, en la sección seis se abordan las conclusiones y el trabajo futuro.

2. Trabajos relacionados

2.1. Métodos de captación

Los estudios sobre la lengua de señas pueden ser divididos en tres aproximaciones: basadas en entradas de cámara [**?**], sensores externos [**?**,**?**] y en entradas por medio de dispositivos como guantes especializados [**?**,**?**]. En orden, el primer tipo de aproximación ofrece la ventaja de ser de fácil y barata implementación a costa de requerir grandes cantidades de datos para generalizar los distintos tamaños de manos y colores de piel.

En segundo lugar, los basados en sensores externos (comúnmente, Microsoft Kinect sensor) presentan una mejora en la generalización, sin embargo, aumentan los costos y la complejidad de implementación. Finalmente, los guantes suelen ser la cúspide de la generalización llegando también a ser la opción de más difícil implementación, además de ser un considerado como un método intrusivo.

2.2. Técnicas utilizadas

Para lograr la clasificación del movimiento en las lenguas de señas han sido propuestas diversas aproximaciones, algunas de ellas basadas en Modelos Ocultos de Markov (HMM) como [**?**], quienes utilizan estos modelos estadísticos para la predicción del gesto manual más probable de la lengua de señas americana.

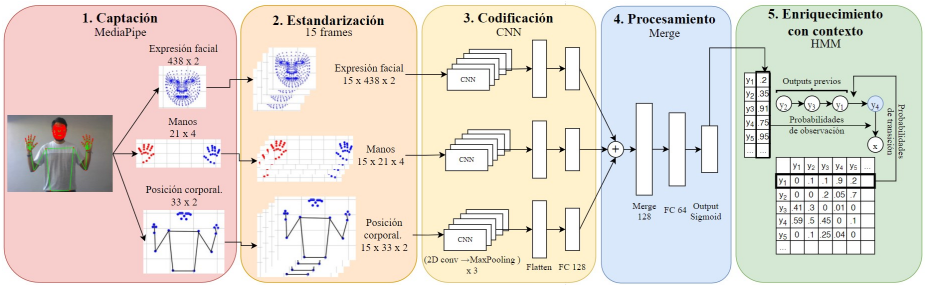


Fig. 1. Etapas de la detección.

Los algoritmos de Data Time Wraping (DTW) presentan una mejora de clasificación ya que, al realizar la comparación contra todos los movimientos almacenados, logran generalizar señas de manera independiente a la duración temporal, en [**?**] fue alcanzada una precisión de 98.57 % en para 20 palabras de la lengua de señas mexicana, sin embargo, por la naturaleza de DTW hace que al aumentar el dataset, el tiempo de cómputo crezca de forma polinómica.

Por último, las Redes Neuronales Convolucionales (CNN) han demostrado obtener un correcto funcionamiento en la clasificación en tiempo real de acciones humanas en general [**?**]. La ventaja que presentan las CNN con respecto a otros métodos es la amplia capacidad de generalización, las capas de convolución extraen las características más importantes identificando patrones en las imágenes y las capas Fully Conected realizan la clasificación.

En [**?**] es utilizado el modelo VGG Network para clasificar 26 letras del alfabeto dactilológico americano alcanzando una precisión de 98.56 %. En [**?**] es propuesta una arquitectura que combina una imagen de la persona que realiza la seña con el extracto de la posición corporal y de las manos realizada por un MS Kinect, el cual consigue una precisión 94.2 % para 25 palabras por separado en lengua de señas americana (caso ASL).

Adicionalmente, en [**?**] se presenta una solución para la lengua de señas china (caso CSL) una arquitectura que combina una codificación convolucional con un procesamiento por medio de una red neuronal recurrente con módulo de atención, la cual fue entrenada con más de 25 mil videos creados por 50 intérpretes, mostrando una precisión de 82.7 %.

2.3. Soluciones para LSM en México

En México existe poco trabajo relacionado a la lengua de señas nacional, enfocándose principalmente en el reconocimiento estático de letras del abecedario dactilológico [**?**,**?**,**?**,**?**,**?**,**?**]. En [**?**] se aborda la identificación de 15 palabras aisladas haciendo uso de un MS kinect para identificar puntos corporales; el algoritmo AdaBoost realizaba la clasificación temporal de cada una de ellas. Por su parte, [**?**] presenta una aproximación al reconocimiento de 20 palabras con el algoritmo DTW obteniendo una precisión media de 98.57 %, sin embargo, se torna inviable para escalar el dataset debido al tiempo de computación.

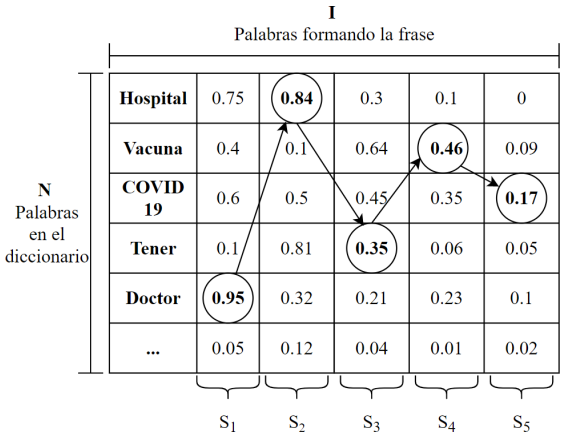


Fig. 2. Ejemplificación del algoritmo de Viterbi aplicado a la determinación de la secuencia de palabras más probable conforme a la gramática de LSM, en este caso: Doctor hospital tener vacuna COVID-19.

Finalmente, sólo [**?**] aborda la interpretación de 5 frases en tiempo real por medio del procesamiento de la posición manual y corporal utilizando momentos geométricos, dichas salidas son integradas por medio de un HMM alcanzando una sensibilidad del 86 %.

En nuestro trabajo, se propone la primera arquitectura que utiliza la expresión facial como determinante del tiempo verbal basado en entradas de cámara con la capacidad de interpretar morfosintácticamente 75 palabras en tiempo real, cuya combinación permite generar más de 50 frases, de las cuales, el rendimiento de 20 frases en el contexto médico es analizado en este estudio.

Para lograr esto, se generó un dataset de 49 palabras (cada uno de los 13 verbos incluidos era conjugado en los tiempos gramaticales dando lugar a tres clases por cada uno) con un promedio de 35 ejemplos para cada una, señadas por seis voluntarios. Adicionalmente, se obtuvo una base de conocimiento con 100 de las frases comunes LSM, la cual aporta los patrones sobre la estructura gramatical propia.

3. Propuesta

El modelo de procesamiento presentado se divide en 5 etapas como se muestra en la figura **??**. La captación utiliza la librería de código abierto MediaPipe presentada en [**?**]. Dicha librería permite construir flujos de percepción con herramientas de visualización de OpenCV [**?**]. Ambas permiten la detección de marcas en las manos, cuerpo y cara en tiempo real con una amplia capacidad de generalización sin importar el color de piel, tamaño de las manos ni altura de la persona.

Así, el sistema de captación mezcla los beneficios de las entradas por cámara (factibilidad económica) y los sensores externos (generalización) en una propuesta no intrusiva. Para la estandarización se extraen las coordenadas de las marcas para ser colocadas en tres matrices, seleccionando 15 frames en cada segundo.

Tabla 1. Señas utilizadas.

Categoría	Seña	Manos	Ejemplos	Categoría	Seña	Manos	Ejemplos
Básicas	Hola	Una	30	Verbo	Hacer	Ambas	40
	Gracias	Ambas	30		Comer	Una	40
	Día	Ambas	40		Pensar	Una	30
	Horario	Una	40		Creer	Ambas	30
Persona	Niño	Una	35		Sentir	Una	30
	Hombre	Una	35		Ir	Ambas	40
	Mujer	Una	35		Contagiar	Ambas	40
	Intérprete	Ambas	40		Pagar	Ambas	40
	Adulto mayor	Ambas	40		Trabajar	Ambas	40
Pregunta	Qué	Una	30		Gustar	Una	40
	Quién	Ambas	40		Poder	Una	40
	Dónde	Ambas	30		Tener	Una	40
	Cómo estás	Ambas	30		Comprar	Ambas	35
Emergencia	Accidente	Una	35	Estado	Feliz	Una	30
	Alergia	Una	40		Bien	Una	30
	Fiebre	Una	35		Mal	Una	30
	Vacuna	Ambas	35		Enfermo	Ambas	40
	Doctor	Ambas	40		Triste	Una	30
	Emergencia	Ambas	40		Enojado	Ambas	30
	Infección	Una	35		Nervioso	Ambas	40
	Medicina	Una	35	Lugares	Escuela	Una	30
	Operación	Ambas	35		Casa	Ambas	30
	COVID-19	Ambas	40		Hospital	Ambas	35
	-	-	-		Calle	Ambas	30

Las coordenadas de la expresión facial con dimensión de 438×2, las manos de 21×4 y la posición corporal 33×2. Posteriormente, las matrices se codifican de manera independiente con una red neuronal con tres capas de convolución 2D cada una seguida por MaxPooling. Esto permite identificar patrones generales, propiciando que señas parecidas, sean codificadas de forma similar.

Las codificaciones se integran por medio de una capa de adición. La codificación de las manos aporta significado, la posición corporal la ubicación espacial y la cara el tiempo verbal. En esta última parte de la arquitectura se utiliza un perceptrón multicapa para realizar la clasificación.

La última capa de la red utiliza un función sigmoide como activación, misma que asigna un vector probabilidades de observación $\vec{O}_i$ cuya entrada O_i^j se refiere a la palabra j en el diccionario de longitud N para cada paso temporal i. Finalmente, se utiliza un HMM para realizar la clasificación con base en el contexto previo. Es empleada una matriz de transición T entre palabras del dataset calculada con 100 frases comunes en LSM extraídas de [**?**].

Tabla 2. Descripción de los participantes en la generación del DataSet.

Persona	Sexo	Edad	Conocimiento de LSM	Captación
1	Hombre	57	Básico	Manual y facial
2	Mujer	35	Intermedio	Manual y facial
3	Hombre	39	Básico	Manual y facial
4	Hombre	19	Básico	Manual y facial
5	Hombre	20	Básico	Manual
6	Mujer	18	Intermedio	Manual

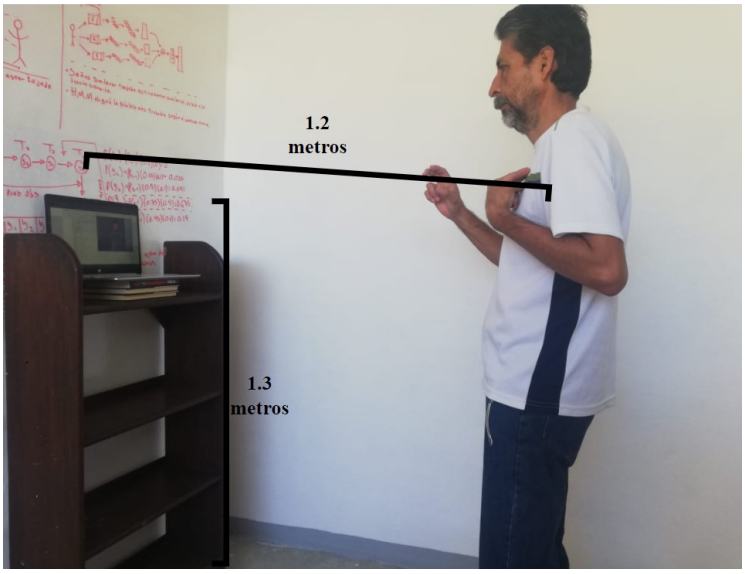


Fig. 3. Ejemplo del montaje para la colección de los datos.

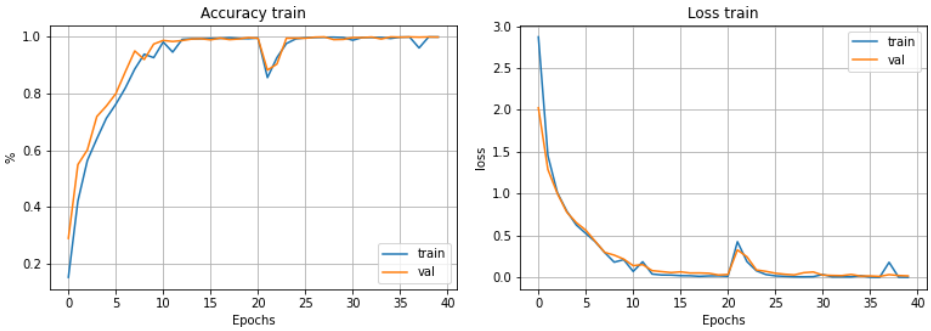
Con $\vec{O_i}$ y T se realiza la selección de la palabra más probable con un sentido morfosintáctico, el conjunto de estados posibles para cada paso temporal se representa por $S = (S_1, S_2, ..., S_I)$, siendo el valor S_1 la palabra más probable. La probabilidad se calcula de forma recurrente como se muestra en la siguiente ecuación:

$$p(O^j, S^j) = \prod_{i=1}^{I} p(O_i^j|S_i^j)p(S_i^j|S_{i-1}^j) \;\; \forall j \in 1, 2, ..., N, \tag{1}$$

donde:

I = total de pasos temporales,
$p(O_i^j|S_i^j)$ = probabilidad de observación,
$p(S_i^j|S_{i-1}^j)$ = probabilidad de transición.

Por último, se ejecuta una implementación manual del algoritmo de Viterbi [**?**] para rastrear el camino de mayor probabilidad. El proceso se ejemplifica en la figura **??**. Dicha implementación reduce el error de interpretación en un 11.7 % en contraste con la identificación de palabras aisladas.



(a) (b)

Fig. 4. Curvas de aprendizaje para set de entrenamiento y validación utilizando dropout (0.3) como regularización. El entrenamiento se consideró suficiente debido a la tendencia asintótica gradual para la reducción del costo.

4. DataSet

4.1. Descripción

El dataset utilizado en el presente trabajo fue generado con ayuda de seis personas utilizando las librerías de MediaPipe y OpenCV. Ver sección **??** para la descripción de los participantes. La captación para las coordenadas de las manos consta de 15 frames que contienen el movimiento de una matriz de 21 puntos para cada mano (15, 21, 4), 33 para el cuerpo (15, 33, 2) y 438 puntos para la cara (15, 438, 2). Se realizó un promedio de 35 ejemplos para una de las 49 señas (13 verbos y 36 palabras) y 15 para cada expresión facial de los tres tiempos verbales utilizados (pasado, presente y futuro). En total, la clasificación constaba de $13 \times 3 + 36 = 75$ clases.

4.2. Participantes

Participaron seis voluntarios con conocimiento de LSM con edades entre 18 y 57 años. Todos los participantes fueron informados del propósito de la investigación y otorgaron su consentimiento. El sexo, nivel de conocimiento, edad y tipo captación se presenta en la tabla **??**.

4.3. Adquisición de datos

Para la obtención de los datos se desarrolló un código en el lenguaje de programación de Python para realizar la identificación de puntos en tiempo real. Cuando el usuario estaba listo, se corría la función y al terminar la captación los resultados eran almacenados. Los participantes se colocaban a 1.2 metros de distancia de la computadora con el montaje que se muestra en la figura **??**.

Tabla 3. Resultados obtenidos para el set de prueba.

Set	F1 Score
Entrenamiento	0.981
Prueba	0.978

Tabla 4. Resultados experimento 1.

Categoría	Precisión media	σ
Básicas	0.987	0.03
Persona	0.943	0.1
Lugar	0.937	0.05
Estado	0.99	0.01
Verbo presente	0.91	0.09
Verbo pasado	0.94	0.06
Verbo futuro	0.89	0.11
Pregunta	0.962	0.07
Emergencia	0.981	0.04

4.4. Estandarización del dataset

Debido al significado espacial particular, cada palabra cuenta con diferente duración temporal en ser señada. Para uniformar el dataset fueron seleccionados n = 15 frames distribuidos a lo largo de los k frames de longitud para cada muestra. Keval H. et al. muestra en [**?**] que el mínimo número de frames por segundo para identificar una acción humana es de 8, así utilizar casi el doble de frames asegura un nivel de detalle aceptable.

5. Experimentos y resultados

5.1. Entrenamiento

Durante el entrenamiento del modelo propuesto se realizó una partición del dataset en 75 % entrenamiento (a su vez dividido en 10 % para validación) y 25 % prueba. El modelo se entrenó por 40 épocas en GPU sustentado por el software de cómputo en la nube de Google Colaboratory [**?**]. La curva de aprendizaje obtenida es mostrada en la figura **??**. Para la medición del rendimiento del modelo se utilizó el F1 Score mostrado en la ecuación **??**. Los resultados se observa en la tabla **??**:

$$\text{Precision} = \frac{T_p}{T_p + F_p}, \tag{2}$$

$$\text{Recall} = \frac{T_p}{T_p + F_n}, \tag{3}$$

$$\text{F1Score} = 2\,\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{4}$$

Tabla 5. Resultados experimento 2.

Frase español	Fase LSM	Precisión media	σ
El niño está en el hospital porque está enfermo	Niño hospital, enfermo	0.95	0.1
El adulto mayor se sintió mal y fue al hospital	Adulto mayor sentir (pasado) mal, hospital ir (pasado)	0.975	0.05
El adulto mayor se sintió bien con la vacuna	Adulto mayor sentir (pasado) bien vacuna	0.975	0.03
El niño se sintió bien y fue a la escuela	Niño sentir (pasado) bien, escuela ir (pasado)	0.925	0.15
El adulto mayor irá al hospital por la vacuna para el COVID-19	Adulto mayor ir (futuro) hospital COVID-19 vacuna	0.962	0.03
El niño enfermo puede contagiar al doctor	Niño enfermo poder (presente) contagiar (presente) doctor	0.962	0.1
El hombre contagió a la mujer en la casa	Hombre contagiar (pasado) mujer casa	0.937	0.12
El doctor tiene medicina para la operación	Doctor tener (presente) operación medicina	0.987	0.02
El niño estará feliz de ir a la escuela	Niño feliz, escuela ir (futuro)	0.977	0.02
El hombre tiene una infección	Hombre tener (presente) infección	0.8	0.14
Yo pagaré la medicina para el COVID-19	Yo pagar (futuro) COVID-19 medicina	0.83	0.11
El hombre fue a su casa porque se sintió mal	Hombre ir (pasado) casa, sentir (pasado) mal	1.0	0.0
El hombre está nervioso porque la mujer tiene COVID-19	Nervioso hombre, mujer tener (presente) COVID-19	0.825	0.12
La intérprete tuvo una emergencia y fue al doctor	Intérprete tener (pasado) emergencia, ir (pasado) doctor	0.95	0.1
El niño se fue a su casa porque se sintió mal en la escuela	Niño ir (pasado) casa, escuela niño sentir (pasado) mal	0.96	0.04
El hospital tendrá vacunas COVID-19	Hospital tener (futuro) COVID-19 vacuna	0.95	0.1
El doctor está feliz porque la mujer se podrá ir a su casa	Doctor feliz mujer poder (futuro) ir casa	1.0	0.0
¿Qué medicina tiene que comprar el hombre?	¿Medicina hombre tener comprar (presente) qué?	0.95	0.1
El niño tiene fiebre por la infección	Niño tener (presente) infección fiebre	0.862	0.11
El doctor irá a su casa porque se siente enfermo	Doctor casa ir (futuro) sentir (presente) enfermo	0.987	0.025

5.2. Resultados experimento 1: Enfoque en palabras aisladas

Cada palabra mostrada en la tabla **??** se señó 10 veces por cada uno de los seis voluntarios, midiendo el rendimiento de manera binaria, es decir, si era predicha correctamente obtenía una calificación de 100 %, caso contrario 0 %.

Para el caso de los verbos, se modificó la forma de medir el rendimiento en aras de obtener una calificación ponderada con base en la complejidad que representaba; así, asignamos un 100 % si la seña y el tiempo gramatical predicho coincidían con el esperado; 70 % si coincidía la palabra, pero no el tiempo. 30 % si el tiempo, pero no la palabra. 0 % si la predicción era totalmente diferente. Se obtuvo una precisión media para las nueve categorías de **94.9 %** con $\sigma = 0{,}07$.

5.3. Resultados experimento 2: Enfoque en frases

Los participantes realizaron la seña y expresión facial de 20 frases tomando cinco muestras. Se utilizó la precisión como métrica de desempeño. Los resultados se muestran en la tabla **??**. Se obtuvo una precisión media para las 20 frases de 94.1 % con $\sigma = 0{,}09$. En contraste, la interpretación con palabras aisladas, es decir, sin enriquecimiento con contexto, alcanzó un **82.4 %** con $\sigma = 0{,}12$.

6. Conclusiones

Como ha sido discutido a lo largo de este trabajo, la interpretación en tiempo real de las lenguas de señas es un proceso complejo en el que intervienen diversos factores como la posición corporal, la expresión facial, los movimientos de las manos y el contexto específico en el que se aborda lo anterior.

Nuestro estudio sienta las bases de un primer acercamiento a la creación de un intérprete completo de lenguas de señas a idiomas hablados, al ser el único trabajo en español hasta el momento que considera la expresión facial como indicador del tiempo gramatical.

En suma, en cuestiones de factibilidad, nuestro modelo demostró que aun con relativamente pocos ejemplos (en promedio, 35 por cada seña y una base de conocimiento de 100 frases) es posible obtener una precisión del **94.9 %** para reconocimiento de 75 palabras aisladas y **94.1 %** para 20 frases de prueba en el contexto médico; mismas que validan tanto la amplia capacidad de generalización de la arquitectura codificadora CNN como el HMM identificador de contexto.

Esto posiciona a nuestro trabajo como una opción económicamente viable - debido a que sólo utiliza una computadora -, de fácil implementación y totalmente escalable a otras lenguas de señas, especialmente las correspondientes a países de bajo o nulo estudio en el campo, mejorando la inclusión de millones de personas con discapacidad auditiva.

Como trabajo futuro, se pretende diseñar, implementar e incluir una máquina de traducción estadística que permita pasar de la estructura gramatical de LSM a español, la cual será una aportación significativa con miras a seguir potenciando las herramientas de comunicación efectiva.

Referencias

1. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Sequential deep learning for human action recognition. In: Lecture Notes in Computer Science, vol. 7065, pp. 29–39 (2011) doi: 10.1007/978-3-642-25446-8_4
2. Ben-Jmaa, A., Mahdi, W., Ben-Jemaa, Y., Ben-Hamadou, A.: A new approach for hand gestures recognition based on depth map captured by RGB-D camera. Computación y Sistemas, vol. 20, no. 4, pp. 709–721 (2016) doi: 10.13053/cys-20-4-2390
3. Bisong, E.: Google colaboratory. Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners, pp. 59–64 (2019)
4. Carmona-Arroyo, G., Rios-Figueroa, H. V., Avendaño-Garrido, M. L.: Mexican sign-language static-alphabet recognition using 3D affine invariants. In: Machine Vision Inspection Systems, vol. 2 (2021)
5. Cruz, M.: Gramática de la lengua de señas mexicana. Centro de Estudios Lingüísticos y Literarios, Colegio de México (2008)
6. Dong, C., Leu, M. C., Yin, Z.: American sign language alphabet recognition using Microsoft Kinect. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 44–52 (2015) doi: 10.1109/cvprw.2015.7301347
7. Fels, S. S., Hinton, G. E.: Glove-TalkII-a neural-network interface which maps gestures to parallel formant speech synthesizer controls. IEEE Transactions on Neural Networks, vol. 8, no. 5, pp. 977–984 (1998) doi: 10.1109/72.655042
8. Forney, G. D.: The viterbi algorithm. In: Proceedings of the IEEE, vol. 61, pp. 268–278 (1973) doi: 10.1109/proc.1973.9030
9. Galicia, R., Carranza, O., Jimenez, E. D., Rivera, G. E.: Mexican sign language recognition using movement sensor. In: Proceedings of the IEEE 24th International Symposium on Industrial Electronics, vol. 2015, pp. 573–578 (2015) doi: 10.1109/isie.2015.7281531

10. García-Bautista, G., Trujillo-Romero, F., Caballero-Morales, S. O.: Mexican sign language recognition using kinect and data time warping algorithm. In: Proceedings of the International Conference on Electronics, Communications and Computers, pp. 1–5 (2017) doi: 10.1109/conielecomp.2017.7891832
11. Grobel, K., Assan, M.: Isolated sign language recognition using hidden markov models. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, IEEE, vol. 1, pp. 162–167 (1997) doi: 10.1109/icsmc.1997.625742
12. Huang, J., Zhou, W., Li, H., Li, W.: Sign language recognition using 3D convolutional neural networks. In: Proceedings of the IEEE International Conference on Multimedia and Expo, IEEE Computer Society, vol. 2015, pp. 1–6 (2015) doi: 10.1109/icme.2015.7177428
13. Huang, J., Zhou, W., Zhang, Q., Li, H., Li, W.: Video-based sign language recognition without temporal segmentation. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, pp. 2257–2264 (2018) doi: 10.48550/ARXIV.1801.10111
14. Kadhim, R. A., Khamees, M.: A real-time american sign language recognition system using convolutional neural network for real datasets. TEM Journal, vol. 9, no. 3, pp. 937–943 (2020) doi: 10.18421/tem93-14
15. Keval, H., Sasse, M. A.: To catch a thief - you need at least 8 frames per second: The impact of frame rates on user performance in a CCTV detection task. In: Proceedings of the 2008 ACM International Conference on Multimedia, with co-located Symposium and Workshops, pp. 941–944 (2008) doi: 10.1145/1459359.1459527
16. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for building perception pipelines (2019) doi: 10.48550/ARXIV.1906.08172
17. Luis-Pérez, F. E., Trujillo-Romero, F., Martínez-Velazco, W.: Control of a service robot using the mexican sign language. Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 7095, pp. 419–430 (2011) doi: 10.1007/978-3-642-25330-0_37
18. Naveenkumar, M., Ayyasamy, V.: OpenCV for computer vision applications. In: Proceedings of the National Conference on Big Data and Cloud Computing, pp. 52–56 (2016)
19. Ordóñez, E.: Asociación de intérpretes en lengua de señas del Distrito Federal: Número de intérpretes de lengua de señas en méxico (2015)
20. Priego-Pérez, F. P.: Reconocimiento de imágenes del lenguaje de señas mexicano. Master's thesis, Centro de Investigación en Computación, Instituto Politécnico Nacional (2012)
21. Rashed, J., Al-Behadili, H.: New method for hand gesture recognition using wavelet neural network (2017)
22. Serafín, M., González, R.: Diccionario de lenguaje mexicano de señas. Revista de Investigación, vol. 38, no. 83, pp. 240 (2011)
23. Solís, F., Martínez, D., Espinoza, O.: Automatic mexican sign language recognition using normalized moments and artificial neural networks. Engineering, vol. 8, no. 10, pp. 733–740 (2016) doi: 10.4236/eng.2016.810066
24. Solís, F., Toxqui, C., Martínez, D.: Mexican sign language recognition using Jacobi-Fourier moments. Engineering, vol. 7, no. 10, pp. 700–705 (2015) doi: 10.4236/eng.2015.710061
25. Sosa-Jimenez, C. O., Rios-Figueroa, H. V., Rechy-Ramirez, E. J., Marin-Hernandez, A., Gonzalez-Cosio, A. L. S.: Real-time mexican sign language recognition. In: Proceedings of the IEEE International Autumn Meeting on Power, Electronics and Computing, pp. 1–6 (2017) doi: 10.1109/ropec.2017.8261606
26. Tolba, A. S., Abu-Rezq, A. N.: Arabic glove-talk (AGT): A communication aid for vocally impaired. Pattern Analysis and Applications, vol. 1, no. 4, pp. 218–230 (1998) doi: 10.1007/bf01234769
27. Álvarez Torres, N.: Kinect V2 como alternativa para desarrollar un traductor de ideogramas de lengua de señas mexicana (2016)