# Reconocimiento de gestos dinámicos y su aplicación al lenguaje de señas

**Franco Ronchetti**
**Directora en la UNLP: Laura Lanzarini**
**Director en la UTH-CUJAE: Alejandro Rosete**
**Fecha de exposición: 23/03/2017**
**Facultad de Informática, Universidad Nacional de La Plata**
**fronchetti@lidi.info.unlp.edu.ar**

## 1. Introducción

El reconocimiento automático de gestos humanos es un problema multidisciplinar complejo y no resuelto aún de forma completa. Desde la aparición de tecnologías de captura de video digital existen intentos de reconocer gestos dinámicos con diferentes fines. La incorporación de nuevas tecnologías como sensores de profundidad o cámaras de alta resolución, así como la mayor capacidad de procesamiento de los dispositivos actuales, permiten el desarrollo de nuevas tecnologías capaces de detectar diferentes movimientos y actuar en tiempo real. A diferencia del reconocimiento de la voz hablada, que lleva más de 40 años de investigación, esta temática es relativamente nueva en el ambiente científico, y evoluciona de forma acelerada a medida que aparecen nuevos dispositivos, así como nuevos algoritmos de visión por computador.

La captura y reconocimiento de gestos dinámicos permite que sean utilizados en diversas áreas de aplicación como por ejemplo monitoreo de pacientes médicos, control en un entorno de videojuego, navegación y manipulación de entornos virtuales, traducción de léxicos de la lengua de señas, entre otras aplicaciones de interés. Particularmente la lengua de señas puede entenderse como un problema particular del reconocimiento de gestos dinámicos, el cual es sumamente apreciado en los últimos tiempos por distintas instituciones, ya que permite una ayuda directa a personas hipoacúsicas.

A grandes rasgos, se pueden distinguir gestos corporales, que se realizan con movimientos de todo el cuerpo, gestos con las manos, como un saludo, gestos con los dedos y las manos, como la lengua de señas y gestos faciales, como los guiños y movimientos de los labios (ver [Mit07]). Otra distinción importante es entre gestos estáticos, comúnmente llamados poses, definidos por una configuración particular del cuerpo en el entorno, y gestos dinámicos compuestos por una serie de movimientos de ciertas partes del cuerpo.

Actualmente, el uso de pantallas táctiles se ha convertido en un estándar para dispositivos móviles en ciertas aplicaciones; el reemplazo de los joysticks tradicionales por interfaces de voz y movimiento en las consolas de juegos se está consolidando. Sin embargo, el retiro del teclado-mouse en las PCs de propósito general por interfaces más naturales basadas en gestos, todavía se encuentra lejos de ser una realidad. En este panorama, las tecnologías más prometedoras para proveer una interfaz hombre-máquina eficiente son el reconocimiento de voz y de gestos en tiempo real [Kar06].

Para poder utilizar un sistema de reconocimiento automático de lengua de señas para traducir los gestos de un intérprete, es necesario afrontar una serie de diversas tareas. En primer lugar, existen diferentes enfoques dependiendo el dispositivo de sensado a utilizar. Si bien existen dispositivos invasivos como guantes de datos, en esta Tesis se analizan sólo dispositivos no invasivos de dos tipos: las cámaras RGB convencionales, y las cámaras de profundidad (con particular interés en los nuevos dispositivos RGB-d). Una vez capturado el gesto se requiere de diversas etapas de pre-procesamiento

para identificar regiones de interés como las manos y rostro del sujeto/intérprete, para luego identificar las diferentes trayectorias del gesto realizado. Además, particularmente para la lengua de señas existe una variabilidad enorme en las diferentes posturas o configuraciones que la mano puede tener, lo cual hace a esta disciplina una problemática particularmente compleja. Para afrontar esto es necesario una correcta generación de descriptores tanto estáticos como dinámicos. Este es uno de los ejes principales investigados en esta Tesis.

Capturar, analizar y responder ante un evento es una tarea sumamente compleja que involucra diferentes áreas de la informática como:

- Procesamiento de imágenes. En todo proceso de reconocimiento en video es necesario contar con un adecuado manejo y filtrado de imágenes. Esto puedo involucrar, entre otras cosas, eliminación de ruido, escalado/rotado de la imagen, filtros frecuenciales para detección de patrones, filtros de color, etc.
- Procesamiento temporal. Ya que se entiende un gesto como una secuencia de movimiento de una o varias partes del cuerpo, es necesario realizar un adecuado procesamiento de la información temporal.
- Sistemas Inteligentes. Para realizar la clasificación de un patrón de video y poder actuar en consecuencia, es necesario la utilización de técnicas inteligentes (*machine learning*).

Las lenguas de señas pueden entenderse como un caso particular del reconocimiento de gestos dinámicos. Es un problema multidisciplinar sumamente complejo con muchas aristas a mejorar en la actualidad. Si bien recientemente ha habido algunos avances a través del reconocimiento de gestos, todavía hay un largo camino por recorrer antes de poder tener aplicaciones precisas y robustas que permiten traducir e interpretar los signos realizados por un intérprete [Coo11].

La tarea de reconocer una lengua de señas implica un proceso de múltiples pasos, que puede ser simplificado del siguiente modo:

1. El seguimiento de las manos del intérprete
2. La segmentación de las manos y la creación de un modelo de su forma
3. Reconocimiento de las formas de las manos
4. Reconocimiento del signo como una entidad sintáctica
5. Asignación de semántica a una secuencia de signos
6. Traducción de la semántica de los signos a la lengua escrita

Estos pasos se detallan en el capítulo 2. Si bien estas tareas pueden proporcionar información entre ellas, de modo general pueden ser llevadas a cabo independientemente, y de diferentes maneras. Por ejemplo, hay varios enfoques para el seguimiento de movimientos de la mano: algunos utilizan sistemas 3D [Sou14, Pug11], tales como MS Kinect. Otros simplemente utilizan una imagen 2D a partir de una cámara [Coo11, Von08]. La mayoría de los sistemas más antiguos emplean sensores de movimiento tales como guantes especiales, acelerómetros, etc., aunque los enfoques más recientes se centran generalmente en el procesamiento de vídeo. Existen numerosas publicaciones sobre el reconocimiento automático de las lenguas de señas, un campo que comenzó hacia los años 90. Puede verse en [Kol15], [Von08] y [Coo11] algunas revisiones generales del estado del arte en esta temática.

El reconocimiento del lenguaje de señas emplea diferentes tipos de características, generalmente clasificadas como manuales y no manuales. Las características no manuales, como pueden ser la postura, lectura de labios o la cara del intérprete se incluyen a veces para mejorar el proceso de reconocimiento, ya que algunas señales no pueden ser diferenciadas únicamente con información manual [Von08]. En este sentido, por ejemplo, el seguimiento de la cabeza es un problema mayormente resuelto [Vio04], pero su segmentación con respecto a un fondo arbitrario o

en presencia de oclusiones mano-cabeza sigue siendo un problema sin resolver. No obstante, la información manual suele contener la mayor parte de la información en una seña.

Para el seguimiento y la segmentación de las manos, hay mucho interés en la creación de modelos de color de la piel para detectar y realizar un seguimiento de las manos de un intérprete en un video [Rou10], y añadiendo la posibilidad de segmentar las manos [Coo12], incluso en presencia de oclusiones mano-mano [Zie05].

La información de la configuración de una mano de un signo está compuesta por una secuencia de poses de esa mano [Von08]. Luego de la segmentación, la mano debe ser representada en una forma conveniente para el reconocimiento de la configuración. Conseguir una representación adecuada fotograma a fotograma de la mano no es una tarea trivial, existiendo diferentes estrategias que aproximan a una solución óptima. Mientras que la mejor salida posible a partir de este paso sería un modelo completo en 3D de la mano, esto es generalmente difícil de hacer sin múltiples cámaras, sensores o marcadores especiales [Pug11]. En la mayoría de los casos, la configuración de la mano en su lugar se representa como una combinación de características más abstractas basada en propiedades geométricas o morfológicas de su forma o textura [Von08].

Algunos investigadores se centran en el reconocimiento dactilológico (*fingerspelling*) [Pug11], que es esencialmente una tarea de reconocimiento de configuración estática. Mientras que algunos signos de hecho presentan una configuración de la mano estática en una o ambas manos, y no hay movimiento, la mayoría implican muchas formas manuales y sus transiciones, o transformaciones de una sola forma de la mano (rotación y traslación, etc.), y un cierto movimiento de las manos. Para hacer frente a estas señales dinámicas, los sistemas de reconocimiento de gestos (SLR, *Sign Language recognition*) generalmente se basan en Modelos Ocultos de Markov (HMMs), Deformación Dinámica de Tiempo (DTW) o modelos similares, ya sea para reconocer las señales segmentadas o un *stream* continuo ([Von08, Coo11]).

## 2. Objetivos

El objetivo general de esta tesis es desarrollar un modelo de reconocimiento automático de la Lengua de Señas Argentina (LSA). Esto trae aparejados los siguientes objetivos específicos:

- Analizar, describir y comparar las diferentes estrategias existentes en el estado del arte sobre reconocimiento y segmentación de manos.
- Construir una base de datos con fotografías de configuraciones de manos de la LSA utilizando marcadores de color para simplificar la segmentación.
- Realizar un método de clasificación de configuraciones de manos incluyendo la adecuada generación de descriptores.
- Construir una base de datos de la LSA con gestos dinámicos que permitan tanto la implementación de traductores específicos para la región, así como también dar la posibilidad a otros investigadores de poder utilizar el repositorio como herramienta de pruebas para algoritmos de aprendizaje automático.
- Realizar un método de clasificación modular de señas segmentadas que permita reconocer diferentes gestos, así como la posibilidad de intercambiar partes del clasificador para evaluar distintos métodos.

## 3. Contribuciones

Las principales contribuciones de esta tesis son las siguientes:

- Una revisión bibliográfica actualizada sobre diferentes estrategias de clasificación de gestos estáticos y dinámicos, incluyendo descriptores de imágenes, video y algoritmos inteligentes de clasificación, así como una revisión de las bases de datos de gestos existentes en la literatura.
- Dos bases de datos de la Lengua de Señas Argentina inexistentes hasta el momento. LSA16 contiene fotografías de 10 individuos distintos para 16 configuraciones de manos de las más utilizadas en el léxico argentino, con un total de 800 imágenes correctamente etiquetadas para cualquier proceso de aprendizaje automático. LSA64 es una base de datos de señas capturadas con una cámara de video de alta resolución. Contiene 64 señas distintas del LSA interpretadas por 10 sujetos distintos con un total de 3200 videos, correctamente etiquetados y con una versión preprocesada donde se tiene información del seguimiento y segmentación de las manos.
- Un método de clasificación de configuraciones del lenguaje de señas, junto con un conjunto de descriptores que posibilitan reconocer las diferentes formas que puede tener las manos de un intérprete. Este método puede utilizarse tanto para el lenguaje de señas como para cualquier aplicación donde se requiera identificar diferentes posturas de las manos.
- Un método probabilístico para clasificar señas basado en tres componentes principales: la posición, la configuración, y el movimiento de cada mano. Este método, basado en componentes posibilita el análisis de cada módulo por separado, dando la posibilidad de intercambiar sub-clasificadores por otros. El modelo propone un análisis específico de la información, creando descriptores apropiados para cada módulo, junto con métodos de clasificación independientes.

## 4. Base de datos de la Lengua de Señas Argentina (LSA)

En el ámbito de la lengua de señas existen diversas bases de datos dependiendo del problema al cual se dirigen. Aquí se distinguen tres tipos principales: reconocimiento de configuraciones de manos, reconocimiento de una seña, y reconocimiento de una sentencia. Cada tipo de base de datos presenta un desafío mayor que el anterior y permite experimentar con mayores pasos en las etapas de reconocimiento [Von08, Coo11].

Las bases de datos aquí presentadas tienen dos objetivos específicos: por un lado, al estar grabadas con guantes de color en las manos de los intérpretes, permiten la rápida segmentación y seguimiento de las manos, utilizando únicamente un accesorio simple, barato y fácil de conseguir. Por otro lado, el conjunto de datos pretende ser un primer paso en la construcción de una base de datos completa para el léxico argentino, inexistente hasta el momento.

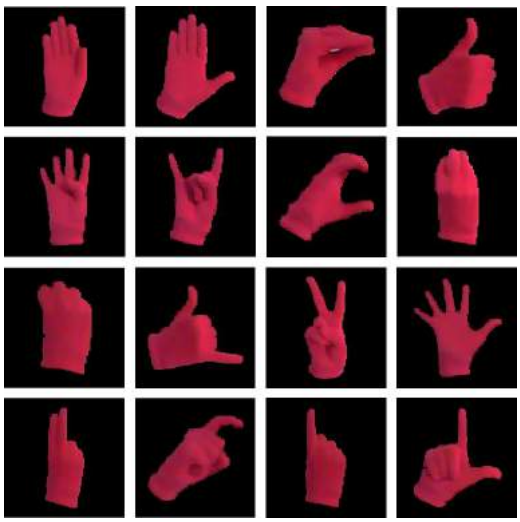

*Figura 1. Las 16 configuraciones de manos en la base de datos LSA16*

Un aspecto esencial de las lenguas de señas es el tipo de configuración que la seña posee. En ocasiones, diversas señas sólo se diferencian por la configuración de la mano, siendo el movimiento que se realiza el mismo que en otras. Esto lleva a la necesidad de tener como parte de un reconocimiento de señas, una etapa de clasificación de configuraciones de manos. Incluso, es normal que una sea comienza con una configuración y termine con otra, o con variaciones de la misma.

La primera base de datos creada, llamada LSA16, contiene 800 imágenes en donde 10 sujetos realizaron 5 repeticiones de 16 tipos distintos de configuraciones de manos utilizadas en distintas señas del léxico. La figura 1 muestra un ejemplo de cada una de las 16 clases dentro de la base de datos. Las imágenes son la segmentación de las fotografías reales. Cada configuración fue realizada repetidamente en diferentes posiciones y diferentes rotaciones en el plano perpendicular a la cámara, para generar mayor diversidad y realismo en la base de datos. Se utilizó una cámara web Logitech con 640x480 pixeles.

Los sujetos vistieron ropa negra, sobre un fondo blanco con iluminación controlada. Para la simplificar el problema de segmentación de la mano dentro de una imagen, los sujetos utilizaron guantes de tela con colores fluorescentes en sus manos. Esto resuelve parcialmente pero de un modo muy eficaz el reconocimiento de la posición de la mano y carece de los problemas existentes en los modelos de piel. Por otro lado, propone un artefacto simple y económico al momento de realizar pruebas o desarrollar una aplicación real. Los detalles de esta base de datos pueden encontrarse publicados en [Ron16c].

Como se mencionó anteriormente, cada región en el mundo tiene su léxico particular en lengua de señas. Esto hace imposible utilizar una base de datos extranjera si se quiere desarrollar un traductor argentino. Por otro lado, se estableció también que debido a la complejidad que se requiere para segmentar las manos de los intérpretes, las bases de datos actuales resultan difíciles de abordar para evaluar la eficacia de un modelo de clasificación.

En segundo lugar, se creó la base de datos LSA64, primer conjunto de videos específico para la Lengua de Señas Argentina. Este conjunto de datos contiene un total de 3200 videos en formato FullHD con 60FPS de 10 sujetos distintos interpretando 64 señas del LSA. La base de datos está públicamente disponible junto con una versión pre-procesada de la misma, para facilitar a los investigadores algunos pasos de segmentación. Cada intérprete utilizó dos guantes de color diferente con el fin de realizar la tarea de segmentación de forma rápida y eficiente. Esta estrategia puede verse utilizada en trabajos anterior, como por ejemplo en [Wan09], donde se utilizaron guantes no sólo para segmentar la mano sino para facilitar la clasificación de la configuración. La figura 2 muestra dos ejemplos tomados de la base de datos.

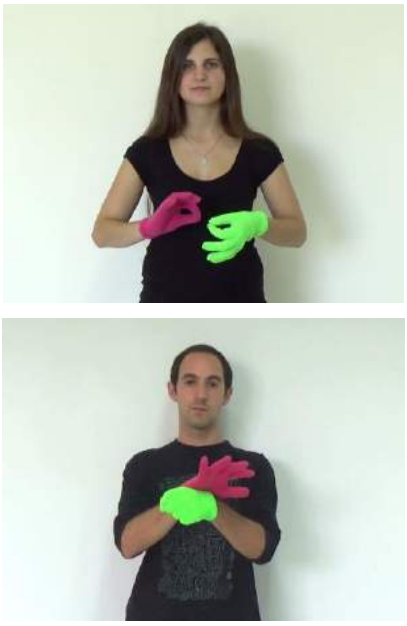

*Figura 2. Dos fotogramas de ejemplos de la base de datos LSA64.*

La base de datos fue construida en dos sets de grabación distintos. En el primero fueron grabadas 23 señas con una sola mano y se utilizó luz natural en un entorno abierto. En el segundo set, se agregaron 41 señas más (22 con dos manos, y 19 con una mano) y se utilizó luz artificial en un entorno semi-cerrado. Estas diferencias en iluminación permiten entrenar un modelo robusto que funcione en diferentes entornos.

Si bien 64 señas no un número particularmente grande para los léxicos reales de lenguas de señas, es un paso inicial para construcción de una base de datos más robusta del léxico argentino, al mismo tiempo que permite un desafío para cualquier sistema de reconocimiento de gestos dinámicos. Las 64 señas elegidas poseen una gran diversidad, presentando superposición tanto de movimientos como en configuración de las manos, siendo necesario analizar todos los aspectos que la componen. Al mismo tiempo los 10 sujetos distintos existentes en la base posibilitan el estudio de un sistema no dependiente al sujeto. Los detalles de esta base de datos pueden encontrarse publicados en [Ron16b].

## 5. Modelo de Clasificación propuesto

El modelo desarrollado en esta tesis consta de diversas etapas para lograr la clasificación de gestos segmentados. La figura 3 muestra un esquema general del modelo donde pueden observarse los detalles de cada etapa. El esquema de clasificación se basa en un sistema probabilístico que tiene en cuenta la información de ambas manos. Si bien el foco del trabajo está puesto en la clasificación de lengua de señas, es posible adaptar el modelo a cualquier tipo de gesto corporal. De cada mano se utilizan tres componentes esenciales en una seña: la posición, la configuración, y el movimiento de la mano. Para obtener las probabilidades para cada uno de estos componentes se definieron diferentes clasificadores parciales, abordando las características que cada problema presenta. Los detalles del modelo aquí expuesto se encuentran publicados en [Ron16a] y [Ron15].

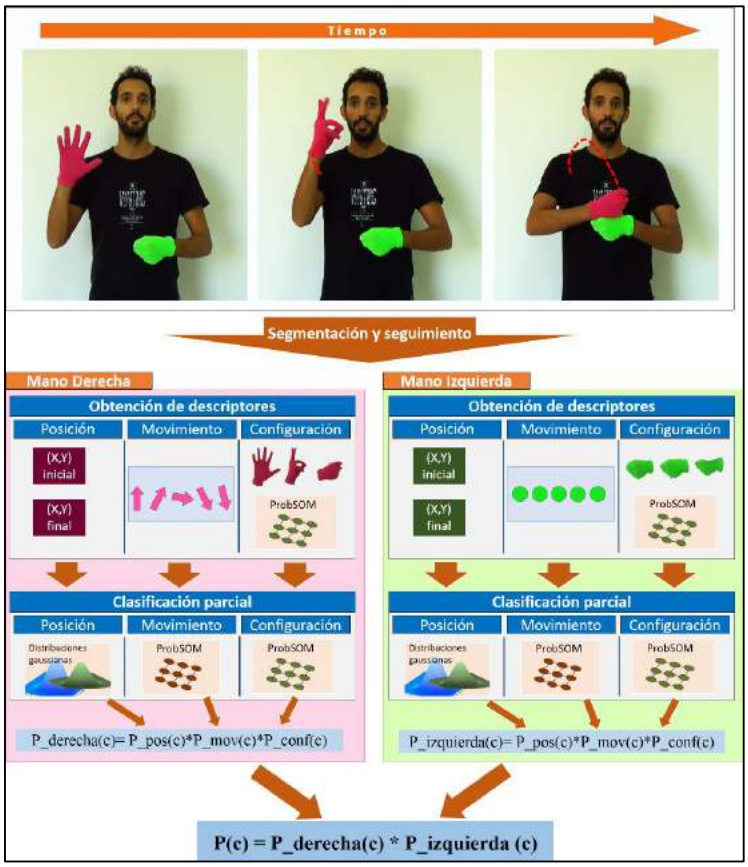


*Figura 3. Descripción general del modelo de clasificación propuesto para señas segmentadas*

El proceso de clasificación de una seña desde el video segmentado puede simplificarse en los siguientes pasos:

1. **Segmentación y Seguimiento**. Inicialmente es necesaria una etapa de segmentación donde cada mano es identificada a partir de un filtro de color. Aquí se obtiene no sólo la posición sino también una máscara con la forma de la mano, lo que facilita luego la clasificación de la configuración. El seguimiento de las manos se realiza fotograma a fotograma almacenando la posición de cada mano relativa a la posición de la cabeza del intérprete.
2. **Generación de descriptores**. En segundo lugar, es necesario generar descriptores apropiados para reconocer los tres componentes principales de la seña. Para describir la posición de cada mano se utilizaron las coordenadas 2D del primer y último fotograma del video. Para describir el movimiento se utilizaron diferencias de percentiles de las posiciones de cada mano. Por último, para describir la configuración de cada mano se utilizó un modelo de clasificación basado en el ProbSOM [Est10].
3. **Clasificación parcial**. En tercer lugar, se realiza un proceso de clasificación parcial. Dado cada conjunto de descriptores, se clasifican de forma independiente, obteniendo una probabilidad parcial de pertenencia a cada clase. Para clasificar la posición de cada mano se utilizó un sistema de distribución de gaussianas, considerando el conjunto de las diferentes posiciones como una distribución normal. Tanto para la clasificación del movimiento como de la configuración de manos se utilizaron redes ProbSOM.
4. **Clasificación de una seña**. Por último, los resultados de los clasificadores parciales son utilizados como entrada para el clasificador probabilístico total. La probabilidad de que una

seña pertenezca a una clase se computa como el producto de probabilidades de cada mano calculada de forma independiente. A su vez, la probabilidad de cada mano se calcula como el producto de probabilidades obtenidas de los clasificadores de posición, configuración y movimiento.

## 6. Trabajos experimentales

Con el fin de validar el modelo desarrollado, se realizaron diversas etapas de experimentación en las bases de datos desarrolladas, como así también en una base de datos de gestos dinámicos capturados con el dispositivo MS Kinect [Ron15]. Todos los resultados aquí mostraron son el promedio de realizar un proceso de validación cruzada aleatoria, con 30 pruebas independientes, 90% de datos para entrenamiento, y 10% para validación.

La tabla 1 muestra los resultados obtenidos de los experimentos realizados en la base de datos LSA16. Se evaluaron los clasificadores ProbSOM, así como también métodos clásicos como Máquinas de Soporte Vectorial (SVM) y Redes Neuronales Feedforward. Dos tipos de descriptores fueron computados sobre las imágenes: vectores SIFT y la Transformada de Radón. Luego, utilizando la mejor configuración obtenida (descriptor Radon y ProbSOM) se llevó a cabo una validación cruzada inter-sujeto, dejando un sujeto para validación y entrenando con el resto. La media de los 10 sujetos con n = 30 repeticiones independientes fue de 87,9%(±4;7%). Como es de esperar, al dejar un sujeto fuera, la tasa de acierto decae, ya que cada persona realiza las configuraciones de forma particular, con tamaños y apariencia de mano propia del individuo. No obstante, el sistema sigue mostrando buenos resultados, dando como posibilidad el reconocimiento correcto de una configuración realizada por un nuevo individuo desconocido por el sistema.

| Método | Precisión |
|---|---|
| ProbSom con Radon | 92,3(±2,05) |
| ProbSom con SIFT | 88,7(±2,50) |
| Random Forest con Radon | 91,0(±1,91) |
| SVM con Radon | 91,2(±1,69) |
| Red Neuronal Feedforward con Radon | 78,8(±3,80) |

*Tabla 1.Precisión del modelo para la base de datos LSA16 de configuraciones.*

Para validar el modelo de clasificación de señas segmentadas, se realizaron una serie de evaluaciones sobre la base de datos LSA64 utilizando el modelo propuesto y comparando los resultados al quitar parte de los sub-clasificadores propuestos. Las diferentes pruebas llevadas a cabo muestran la importancia de cada componente, ya que al quitar alguno la tasa de acierto decae, mostrando que cada componente agrega información no redundante al sistema. Por otro lado, se realizaron evaluaciones de comparación tanto para los descriptores como para el clasificador. Por un lado, la columna Todos-HMM muestra los resultados al reemplazar los subclasificadores de trayectoria y configuración por Modelos Ocultos de Markov (HMM) con Modelos de Mixturas Gausianas (GMM). El resultado aquí fue de casi 96%. Esto muestra que si bien el ProbSOM obtuvo una mejora al reducir el error en un 60% comparado a los Modelos Ocultos de Markov, los descriptores propuestos son una parte esencial en el proceso de clasificación. Por otro lado, la última columna, titulada Todos-BF-SVM, muestra los resultados al cambiar la obtención de descriptores por los *binary features* propuestos por Kadir en [Kad04], clasificador con una Máquina de Soporte Vectorial.

| | Todos | Config | Mov | Pos | Config-Pos | Config-Mov | Pos-Mov | Todos-HMM | Todos-BF-SVM |
|---|---|---|---|---|---|---|---|---|---|
| $\mu$ | 97.44 | 52.97 | 54.03 | 76.05 | 94.91 | 83.59 | 84.84 | 95.92 | 95.08 |
| $\sigma$ | 0.59 | 1.74 | 1.71 | 0.62 | 0.52 | 0.87 | 0.90 | 0.95 | 0.69 |

*Tabla 2. Resultados de los experimentos llevados a cabo sobre la base de datos LSA64*

Otro aspecto importante a considerar en la evaluación de los resultados es si la seña se realizó con una o dos manos. Si bien el modelo se definió para evaluar ambas manos, la base de datos utilizada (LSA64) posee tanto señas con dos manos, como señas con una sola mano (la dominante). En el primer caso, al tener ambas manos, podría conllevar a una ventaja en la clasificación, ya que se está incorporando información al modelo. En este sentido, es importante evaluar cómo se comporta el modelo en ambos casos. Siguiendo esta idea, se dividió la base de datos en dos subconjuntos de datos, uno con las 22 clases de señas con dos manos, y otro con las 42 clases de señas con una mano. Se realizaron experimentos independientes para verificar la tasa de acierto del modelo. Mientras que la media entre ambos tipos de experimentos no difiere significativamente de los resultados de las pruebas generales, cabe rescatar el aumento de la tasa de acierto al tener sólo clases de dos manos, logrando un error de sólo 0,91%. Cabe rescatar que la tasa de acierto al tener señas de sólo una mano llega a casi 96%, mostrando excelentes resultados para las 42 señas con esta característica.

| Sujeto | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Media |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mu$ | 94.5 | 93.8 | 87.7 | 93.8 | 91.8 | 92.6 | 89.1 | 90.3 | 88.4 | 94.6 | 91.7 |
| $\sigma$ | 0.66 | 0.83 | 1.05 | 0.79 | 0.65 | 0.41 | 0.91 | 0.70 | 0.85 | 0.66 | 0.75 |

*Tabla 3. Validación independiente al sujeto sobre LSA64.*

Por último, y al igual que para LSA16, se llevaron a cabo una serie de experimentos sobre LSA64 para analizar cómo se comporta el sistema ante la presencia de un sujeto nuevo, con el que no se había entrenado previamente. Para esto se entrenó el sistema con 9 sujetos, dejando el restante para validación, haciendo una evaluación cruzada inter-sujeto con 30 ejecuciones independientes. La tabla 3 muestra los resultados obtenidos para cada uno de los 10 sujetos de la base de datos LSA64, junto con la media de todos los resultados. Como es de esperar, la tasa de acierto decae con respecto a los resultados generales. No obstante, la precisión media conseguida fue de 91,7%(±0,8), mostrando excelentes resultados al introducir un nuevo individuo al sistema.

## 7. Conclusiones y líneas de trabajo futuras

La Tesis aquí resumida cuenta con dos aportes principales: por un lado, un modelo de clasificación para gestos dinámicos específicamente diseñado para la lengua de señas. Por otro lado, una base de datos multimedia de la Lengua de Señas Argentina, inexistente hasta el momento.

La base de datos desarrollada, llamada LSA64, posee 64 señas distintas del LSA. Los intérpretes utilizaron guantes de color para facilitar la segmentación de las manos. Este proceso resulta accesible para cualquiera ya que el guante es una herramienta económica y de fácil acceso. La base de datos propone tanto un diccionario específico para el léxico argentino como una herramienta de base de pruebas para cualquier trabajo de aprendizaje automático. La base de datos consta de 10 sujetos interpretando cada seña 5 repeticiones distintas, dando un total de 3200 videos de alta resolución. Sumado a esto, la base de datos LSA16 contiene 800 imágenes de configuraciones de manos del léxico argentino.

El método propuesto para clasificación de señas propone un esquema modular con subclasificadores parciales capaces de interpretar tres características principales en una seña: la posición, el movimiento y la configuración. Como sub-clasificador de configuración se utilizó también una red tipo PromSOM para clasificar las 16 configuraciones del LSA. Este trabajo fue primero evaluado por separado para general descriptores apropiados que permitieron luego adicionar la información temporal de las diferentes configuraciones que puede tener una seña. Por último, como sub-clasificadores de las posiciones de las señas, se utilizaron distribuciones estadísticas con modelos gaussianos de las posiciones iniciales y finales que cada mano posee en una seña.

Los resultados obtenidos sobre la clasificación de configuraciones mostraron ser relevantes y factibles de utilizar en un entorno real. Además, siendo un clasificador probabilístico, el sistema posee la capacidad de poder ser utilizado como descriptor para el clasificador parcial del modelo propuesto. Los experimentos realizados sobre la base de datos LSA64 fueron extensos y con resultados satisfactorios. Se realizaron diferentes pruebas sobre los clasificadores parciales para observar su comportamiento al igual que diferentes evaluaciones sobre el clasificador completo demostrando su robustez ante diferentes escenarios como por ejemplo la incorporación de un nuevo sujeto al sistema.

Existen diversas líneas de investigación que quedan abiertas luego de la finalización de esta tesis, entre las que se cabe nombrar:

- Focalizarse en la etapa de detección de manos para poder realizar una segmentación sin necesidad de marcadores de color. Esto permitiría realizar pruebas en otras bases de datos existentes donde no existe información de las posiciones de las manos ni tampoco se utilizan estos tipos de marcadores. Una de las estrategias más recientes aplicadas a esta temática son las redes convolucionales, relacionadas con el concepto de aprendizaje profundo (*deep learning*), que recién está emergiendo.
- Si bien se utilizaron algunos descriptores y clasificadores propuestos por otros autores en el estado del arte, generalmente esta tarea resulta sumamente compleja debido a que el modo de obtener los descriptores de una seña está relacionado con el clasificador propuesto.
- Para poder llevar a cabo un traductor más robusto sin duda es necesario aumentar el número de señas en la base de datos. Esto supone un desafío importante, no sólo por el tiempo y puesta en escena de las grabaciones requeridas sino también porque aumentar considerablemente el número de clases en una base de datos implica aumentar el error de clasificación en cualquier proceso de aprendizaje automático.
- Incorporar información no manual, como pueden ser expresiones de la cara, lectura de labios, inclinación del torso, etc. Este es un trabajo que algunos investigadores ya están abordando. El lenguaje de señas no sólo se basa en los movimientos de las manos sino en realizar un diccionario completo involucra también evaluar información no-manual, principalmente del rostro.

## 8. Referencias


Coo11 Helen Cooper, Brian Holt, and Richard Bowden. Sign language recognition. In Thomas B. Moeslund, Adrian Hilton, Volker Krüger, and Leonid Sigal, editors, Visual Analysis of Humans: Looking at People, chapter 27, pages 539 – 562. Springer, 2011.

Coo12 Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. Sign language recognition using sub-units. Journal of Machine Learning Research, 13:2205–2231, Jul 2012.

Est10 Cesar Estrebou, Laura Lanzarini, and Waldo Hasperue. Voice recognition based on probabilistic som. In Latinamerican Informatics Conference. CLEI 2010. Paraguay. October 2010.

Kad04 T. Kadir, R. Bowden, Ej Ong, and a. Zisserman. Minimal Training, Large Lexicon, Unconstrained Sign Language Recognition. British Machine Vision Conference, pages 96.1–96.10, 2004.

Kar06 Maria Karam. PhD Thesis: A framework for research and design of gesturebased human-computer interactions. PhD thesis, University of Southampton, October 2006.

Kol15 Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding, 141:108–125, December 2015.

Mit07 S. Mitra and T. Acharya. Gesture recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 37(3):311– 324, 2007.

Pug11 N. Pugeault and R. Bowden. Spelling it out: Real-time ASL fingerspelling recognition. In 1st IEEE Workshop on Consumers Depth Cameras for Computer Vision, in conjunction with ICCV'2011, 2011.

Ron15 Ronchetti, Facundo Quiroga, Laura Lanzarini, César Estrebou. Distribution of Action Movements (DAM): A Descriptor for Human Action Recognition. Franco Frontiers of Computer Science. ISSN 2095-2236. Springer, Higher Education Press. v9. pp956-965. Diciembre 2015.

Ron16a Franco Ronchetti, Facundo Quiroga, César Estrebou, Laura Lanzarini, Alejandro Rosete. Sign Languague Recognition without frame-sequencing constraints: A proof of concept on the Argentinian Sign Language. Advances in Artificial Intelligence - IBERAMIA 2016: 15th Ibero-American Conference on AI, San José, Costa Rica, November 23-25, 2016, Proceedings. pp338-349. Springer International Publishing. 2016.

Ron16b Franco Ronchetti, Facundo Quiroga, César Estrebou, Laura Lanzarini, Alejandro Rosete. LSA64: An Argentinian Sign Language Dataset. XXII Congreso Argentino de Ciencias de la Computación. CACIC 2016. San Luis. Argentina. pp794-803. Octubre 2016.

Ron16c Franco Ronchetti, Facundo Quiroga, César Estrebou, Laura Lanzarini. Handshape recognition for Argentinian Sign Language using ProbSom. Journal of Computer Science & Technology. ISSN 1666-6038. Editorial ISTEC – RedUNCI. 16:1, pp01-05. April 2016.

Rou10 Anastasios Roussos, Stavros Theodorakis, Vassilis Pitsikalis, and Petros Maragos. Hand tracking and affine shape-appearance handshape sub-units in continuous sign language recognition. In Trends and Topics in Computer Vision - ECCV 2010 Workshops, Heraklion, Crete, Greece, September 10-11, 2010, Revised Selected Papers, Part I, pages 258–272, 2010.

Sou14 Gabriel de Souza Pereira Moreira, Gustavo Ravanhani Matuck, Osamu Saotome, and Adilson Marques da Cunha. Recognizing the brazilian signs language alphabet with neural networks over visual 3d data sensor. In Advances in Artificial Intelligence–IBERAMIA 2014, pages 637–648. Springer, 2014.

Vio04 Paul Viola and Michael J Jones. Robust real-time face detection. International journal of computer vision, 57(2):137–154, 2004.

Von08 Ulrich von Agris, Jörg Zieren, Ulrich Canzler, Britta Bauer, and Karl-Friedrich Kraiss. Recent developments in visual sign language recognition. Universal Access in the Information Society, 6(4):323–362, 2008.

Wan09 Robert Y. Wang and Jovan Popovic. Real-time hand-tracking with a color glove. ACM Transactions on Graphics, 28(3), 2009.

Zie05 Jörg Zieren and Karl-Friedrich Kraiss. Robust person-independent visual sign language recognition. In Pattern recognition and image analysis, pages 520–528. Springer, 2005.