UNIVERSIDAD NACIONAL DE LA PLATA
Facultad de Informática

Reconocimiento de gestos dinámicos y su aplicación al lenguaje de señas

Autor:
Lic. Franco Ronchetti

Directores:
Lic. Laura Lanzarini
Dr. Alejandro Rosete

Tesis presentada para obtener el grado de
Doctor en Ciencias Informáticas

Diciembre 2016

Agradecimientos

Quisiera agradecer especialmente a mis directores, Laura y Alejandro, por todo el trabajo y disposición durante este tiempo.

A Facu Q., gran compañero de trabajo de los últimos tiempos.

A Ayelén, el amor de mi vida que siempre me acompañó.

A mi familia y amigos que siempre me apoyaron durante este transcurso.

A todos los chicos del III-LIDI que se ofrecieron como voluntarios para construir la base de datos. Particularmente gracias a César, Augusto, Germán, Majo, Juli, Ale, Vero, Nico y Santi.

Resumen

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.

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. Además, debido a que cada región presenta gramáticas de lenguaje específicas, se requiere la disposición de una base de datos de la Lengua de Señas Argentina (LSA), inexistente hasta el momento. En base a los motivos mencionados anteriormente, esta Tesis tiene como objetivo general desarrollar un proceso completo de interpretación y traducción de la Lengua de Señas Argentina a través de videos obtenidos con una cámara RGB.

En primer lugar se realizó un estudio del estado del arte en el reconocimiento de gestos. Se investigaron técnicas inteligentes para el procesamiento de imágenes y video así como los diferentes tipos de descriptores existentes en la actualidad. Como trabajo preliminar se desarrolló una estrategia capaz de procesar acciones humanas capturadas con un dispositivo MS Kinect. La estrategia desarrollada implementa una red neuronal SOM probabilística (ProbSOM) con un descriptor específicamente diseñado para retener información temporal. Este trabajo permitió superar los resultados existentes hasta el momento para dos bases de datos reconocidas.

En el campo de la lengua de señas se realizaron dos aportes principales. En primer lugar se desarrolló una base de datos específica para el reconocimiento de señas argentinas. Esto incluyó una base de datos de imágenes con 16 configuraciones de las más utilizadas en el lenguaje, junto con una base de datos de videos de alta resolución con 64 señas distintas, con un total de 3200 videos. Estas bases de datos se grabaron con 10 intérpretes diferentes y varias repeticiones, permitiendo así su uso con técnicas clásicas de aprendizaje automático. Además, en estas bases de datos los intérpretes utilizaron guantes de color, en forma de marcador. Esto se realizó con el fin de facilitar la tarea de segmentar las manos de las imágenes/videos y así poder avanzar con el resto de las etapas de clasificación. De este modo, se da la posibilidad a nuevos investigadores de evaluar otros algoritmos de reconocimiento sin la necesidad de preocuparse por esta etapa de segmentación.

En segundo lugar, se diseñaron e implementaron dos métodos de clasificación de señas, los cuales fueron evaluados satisfactoriamente en las bases de datos antes mencionadas. El primer método está dedicado a la clasificación de configuraciones de manos (gestos estáticos). Aquí se utilizó un agrupamiento probabilístico para clasificar correctamente las 16 configuraciones posibles de la base de datos, logrando un reconocedor simple y potente. El segundo modelo de clasificación permitió la clasificación de señas segmentadas en videos. Este último consta de un sistema probabilístico basado en la información capturada de las dos manos, donde para cada una se evalúan tres componentes principales: la posición, la configuración y el movimiento de las manos. Esta discriminación permitió tener un sistema modular, con diferentes sub-clasificadores capaces de intercambiarse y evaluarse de modo independiente. Para lograr obtener descriptores adecuados para estos subsistemas, es necesario realizar un procesamiento que involucra la correcta segmentación y seguimiento de las manos del intérprete, clasificación de las distintas configuraciones y una correcta representación de la información del movimiento.

Para evaluar los modelos desarrollados se realizaron diversas pruebas sobre las bases de datos desarrolladas. En primer lugar se realizaron pruebas de validación cruzada utilizando un porcentaje de las pruebas como entrenamiento y el resto para testeo. Adicionalmente se realizó también una evaluación de cuán robusto es el sistema al incorporar nuevos intérpretes, desconocidos hasta el momento. De este modo, 9 de los 10 individuos de la base de datos fueron utilizados como datos de entrada del sistema, evaluando con el individuo restante. Todos estos experimentos mostraron excelentes resultados, con una tasa de error menor al 5 %. Por otro lado, para evaluar la eficacia del modelo implementado, se cambiaron algunos de los sub-clasificadores por técnicas más conocidas en la literatura como Modelos de Markov o Redes Neuronales FeedForward, mostrando solidez en las estrategias propuestas en esta Tesis.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1 Motivación

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 [102]). 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 [71].

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 imágen, 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 [29].

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 [104, 126], tales como MS Kinect. Otros simplemente utilizan una imagen 2D a partir de una cámara [29, 153]. 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 [87], [153] y [29] 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 [153]. En este sentido, por ejemplo, el seguimiento de la cabeza es un problema mayormente resuelto [148], 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 [135], y añadiendo la posibilidad de segmentar las manos [30], incluso en presencia de oclusiones mano-mano [169].

La información de la configuración de una mano de un signo está compuesta por una secuencia de poses de esa mano [153]. 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 [126]. 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 [153].

Algunos investigadores se centran en el reconocimiento dactilológico (fingers-

pelling) [126], 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 reconocimeinto 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 ([153, 29]).

1.2 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.

1.3 Alcances y Limitaciones

El reconocimiento automático de gestos dinámicos involucra un gran abanico de temáticas que es necesario abordar. En primer lugar, existen numerosos métodos de captura para sensar los movimientos realizados por una persona. Dentro del presente trabajo se abordará principalmente el procesamiento de video con cámaras RGB tradicional. El procesamiento de señales de cámaras de profundidad se analizará con un caso de estudio puntual en el capítulo 5.

Una vez obtenidas y filtradas las imágenes con las que se captura un gesto, es necesario convertir el dominio espacial de una imagen para obtener la información del objeto que se está queriendo observar. Por ejemplo, en el caso del lenguaje de señas, el principal problema existente en este punto es cómo segmentar la información de cada mano y la cabeza. Aquí existen diversas aproximaciones:

- Con respecto a la detección de la cabeza existen soluciones muy buenas como por ejemplo [148], haciendo uso de los patrones constantes que posee un rostro humano.
- Lo que involucra la detección de manos es un problema mucho más complejo y poco resuelto. Una de las principales soluciones radica en realizar un filtro en base a un modelo de color de piel del intérprete. Estos modelos, si bien existen en numerosos trabajos, suelen ser poco escalables además de tener que lidiar luego con las oclusiones de las manos con la cabeza.
- La estrategia utilizada en esta tesis son los marcadores de color [125]. Los intérpretes utilizan guantes de un color específico que facilitan el filtrado con respecto al resto de los objetos en el video. Luego, se prosigue con el resto del procesamiento que deba hacerse sobra las manos. Estos marcadores de color suelen utilizarse en otros trabajos para ubicar diferentes partes del cuerpo como podría ser la cara, el torso, etc.
- Realizar una detección robusta de las manos basada en la morfología de las mismas es un trabajo complejo, aún no resuelto, y no abordado en este tesis.

Una vez obtenidos los descriptores de los objetos que se consideren necesarios para realizar una correcta clasificación, es necesario definir un método para llevar esto a cabo. Debido a la naturaleza temporal que presentan los datos procesados es necesario aquí abordar estrategias que manipulen correctamente este tipo de información. No obstante, en este trabajo se analizan diversas técnicas tanto para procesamiento temporal como DTW [14] o HMM [128][38], como también técnicas clásicas del aprendizaje automático como Redes Neuronales *FeedForward* [65], algoritmos de clustering, y otras.

Haciendo un análisis de la bibliografía existente, puede notarse que la clasificación de un gesto, como puede ser una seña, se realiza principalmente con gestos segmentados en bases de datos particulares. Esto trae luego el problema de utilizar los métodos propuestos por los autores para realizar una clasificación en tiempo continuo[1] dentro de un video (*video stream*). Existen muy pocos trabajos que detallen este tipo de clasificaciones. En el capítulo 4 se propone un esquema para aplicar los métodos propuestos en esta tesis a videos continuos donde existen diversas señas. No obstante, la aplicación de estas estrategias en un entorno de tiempo real traen aparejados otros problemas que no son contemplados en este trabajo.

1.4 Contribuciones

Las principales contribuciones de esta tesis son las siguientes:

[1] Se llamará *tiempo continuo* en este documento para diferenciar de *tiempo real*, ya que este último término hace referencia también a la velocidad de procesamiento, generalmente con expectativas de funcionar en un entorno real con alta velocidad de respuesta, no siendo esto un objetivo específico de esta Tesis.

- 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.

1.5 Publicaciones

Esta tesis doctoral está avalada por las siguientes publicaciones:

- *Sign Languague Recognition without frame-sequencing constraints: A proof of concept on the Argentinian Sign Language*. Franco Ronchetti, Facundo Quiroga, César Estrebou, Laura Lanzarini, Alejandro Rosete. 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.
- *LSA64: An Argentinian Sign Language Dataset*. Franco Ronchetti, Facundo Quiroga, César Estrebou, Laura Lanzarini, Alejandro Rosete. XXII Congreso Argentino de Ciencias de la Computación. CACIC 2016. San Luis. Argentina. pp794-803. Octubre 2016.
- *Handshape recognition for Argentinian Sign Language using ProbSom*. Franco Ronchetti, Facundo Quiroga, César Estrebou, Laura Lanzarini. Journal

of Computer Science & Technology. ISSN 1666-6038. Editorial ISTEC – RedUNCI. Volumen 16, Número 1, pp01-05. Abril 2016.
- *Distribution of Action Movements (DAM): A Descriptor for Human Action Recognition*. Franco Ronchetti, Facundo Quiroga, Laura Lanzarini, César Estrebou. Frontiers of Computer Science. ISSN 2095-2236. Springer, Higher Education Press. v9. pp956-965. Diciembre 2015.

Se listan a continuación publicaciones previas del tesista relacionadas con la temática:

- *SOM+PSO: A Novel Method to Obtain Classification Rules*. Laura Lanzarini, Augusto Villa Monte, Franco Ronchetti. Journal of Computer Science & Technology. ISSN 1666-6038. Editorial ISTEC – RedUNCI. v15, Número 1, pp. 15-22. Abril de 2015.
- *A Support System for the Diagnosis of Balance Pathologies*. Augusto Villa Monte, Facundo Quiroga, Franco Ronchetti, César Estrebou, Laura Lanzarini, Pedro M. Estelrrich, Claus Estelrrich, Raimundo Giannecchini. VI Workshop Innovación en Sistemas de Software (WISS). XX Congreso Argentino de Ciencias de la Computación. CACIC 2014. La Matanza. Buenos Aires. Octubre 2014.
- *Face recognition based on fuzzy probabilistic SOM*. Laura Lanzarini, Franco Ronchetti, César Estrebou, Luciana Lens, Aurelio Fernández Bariviera. IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint. Edmonton, Canadá. pp. 310-314. IEEE, Junio 2013.

1.6 Organización de la tesis

A continuación se describe la organización del resto del documento:

En el capítulo 2 se analiza la bibliografía existente en la actualidad a cerca de los diferentes temas abordados en esta tesis. Se detallan los métodos para obtener descriptores de imágenes y video, métodos existentes para clasificación de configuraciones de manos y métodos para la clasificación de gestos con cámaras de profundidad y con cámaras RGB. El capítulo concluye con los desafíos aún no resueltos hasta el momento en el estado del arte.

En el capítulo 3 se describen las bases de datos desarrolladas para esta tesis llamadas LSA16 y LSA64. Se describen los modos de construcción de cada base de datos junto con al información relevante en cada caso. Se presentan además de forma gráfica una serie de características de la base de datos principal a efectos de entender su naturaleza y alcances.

En el capítulo 4 se define el modelo propuesto para clasificación de señas. Este modelo incluye diversos sub-módulos que son detallados por separado, incluyendo la genereación de descriptores estáticos para las configuraciones de las manos y descriptores dinámicos como los del movimiento de las mismas. Se presentan también

los alcances y limitaciones que posee de acuerdo a la complejidad del dominio de la lengua de señas.

El capítulo 5 detalla los experimentos llevados a cabo para mostrar la eficacia del modelo planteado. Se analizan estudios preliminares realizados para acciones humanas en bases de datos capturadas con cámaras de profundidad. Luego, se detallan los experimentos realizados en el ámbito del lenguaje de señas separados en dos partes principales: la clasificación de configuraciones de manos y la clasificación de señas segmentadas en videos individuales.

Por último, el capítulo 6 presenta las conclusiones de la tesis junto con las posibles lineas de trabajo futuro.

Capítulo 2

Marco teórico y revisión bibliográfica

En la actualidad, la informática forma parte de la vida cotidiana de las personas de un modo natural, abarcando un número de aplicaciones cada vez mayor. La reducción en los costos de producción de procesadores y sensores, así como el aumento en velocidad de cómputo permiten la existencia de aplicaciones cada vez más complejas, incrementando así la necesidad de interacción humano-computador (*Human- Computer Interaction - HCI*). En este contexto, está comenzando una nueva era donde la dupla mouse-teclado no es la única forma de interacción que existe con una computadora.

Uno de los primeros trabajos aceptados por la comunidad en incorporar gestos en HCI fue [141], donde Sutherland presentó un sistema llamado Sketchpad. Este sistema permitía manipular un marcador tipo bolígrafo para manipular objetos gráficos en un dispositivo tipo tableta. Desde entonces, las investigaciones en software han ido evolucionando al mismo tiempo que la capacidad de adquirir mejores datos con nuevos tipos de sensores. No obstante este crecimiento, el desarrollo de investigación en técnicas avanzadas que permitan una correcta adquisición de datos en estas aplicaciones, no está del todo madura. Los algoritmos y técnicas actuales pueden significar un cuello de botella en la efectiva utilización del flujo de información que es posible obtener con los sensores actuales [129]. Mientras que el reconocimiento del audio (particularmente de la voz) ha tenido notables avances en la actualidad, llegando al punto de existir numerosas aplicaciones comerciales y de fácil acceso, el reconocimiento automático de gestos está recién comenzando su desarrollo [29]. Mientras que los sensores más sofisticados como guantes o trajes electrónicos son sumamente costosos y poco prácticos, los sistemas basados en visión (como el enfoque implementado en esta Tesis) suelen tener problemas aún no resueltos como fondos complejos, problemas de iluminación, entre otros [120].

HCI se ha convertido en un campo de investigación esencial en las últimas décadas [64, 125, 77], y particularmente la interacción mediante gestos es un área

que ha tomado mayor interés actualmente. El reconocimiento de gestos permite la interacción con una computadora para realizar diversas tareas como manipular objetos gráficos, interactuar en un entorno de realidad virtual, navegar por la web, manipular un objeto como un TV u otro dispositivo multimedia, etc. En general, este tipo de interacciones gestuales son relativamente sencillas, con solo un puñado de gestos a reconocer, generalmente bien distintos. Por otro lado, existen aplicaciones más recientes donde el objetivo es reconocer cierto comportamiento que posee una persona de un modo más natural, real, que involucra generalmente un número mayor de gestos y con movimientos mucho más complejos. El ejemplo principal aquí son las lenguas de señas, donde los gestos presentan una complejidad y diversificación superior a cualquier gesto de un sistema de videojuego o de interacción multimedia. Esta disciplina es el campo de aplicación principal que se estudiará en esta Tesis. También es posible encontrar ejemplos de reconocimiento de gestos complejos para monitoreo de pacientes médicos, detección de mentiras, monitoreo de conductores en un automóvil, etc. [102, 120].

Los gestos, y particularmente los gestos hechos con las manos, presentan un modo de comunicación natural y mucho más rico que el que proveen los dispositivos convencionales. Actualmente existen diversas investigaciones en el área, incluyendo algunas revisiones completas como [120], [129], [153], [29] o [102]. Generalmente, el mayor interés está puesto en los gestos hechos con las manos, aunque existen también numerosas investigaciones sobre gestos con el cuerpo completo. No obstante, las aplicaciones reales de estas investigaciones aún son escasas o nulas, siendo por lo general un estudio de laboratorio, con condiciones específicas difíciles de reproducir. Las bases de datos de gestos, y particularmente de lenguas de señas, aún suelen ser pequeñas, y las técnicas de reconocimiento tienen muchos desafíos por delante a superar. Algunos pocos trabajos muy recientes, como por ejemplos el expuesto por Oscar Koller en [87], comienzan a utilizar bases de datos realistas, con cientos de clases y condiciones complejas. En general este tipo de trabajos están recién comenzando, no habiendo documentación precisa sobre cómo proceder ante las diversas aristas a resolver en la temática.

Existe poca información en la bibliografía sobre cómo realizar un proceso de reconocimiento de gestos. Sin embargo, generalmente todos los autores muestran un enfoque similar. La figura 2.1 muestra este esquema de procesamiento, basado en la idea planteada en [18] y [120]. En primer lugar, existe un dispositivo de captura, que podría ser una cámara 2D, de profundidad, o algún otro dispositivo de sensado. Luego, se interpreta la información obtenida con el fin de generar ciertos descriptores que representen correctamente al gesto. Para esto, primero es necesario segmentar las zonas que interesan, como suelen ser las manos, rostro u otras partes del cuerpo. Dependiendo si es un gesto estático o dinámico (ver sección 2.1) habrá una etapa de seguimiento para entender cómo se mueve el cuerpo. Finalmente, una vez obtenidos los descriptores, existe algún método de clasificación encargado de diferenciar los descriptores, identificando de qué gesto se trata. En [153] se especifica un esquema similar, incorporando una etapa de pre-procesamiento de imágenes antes de la

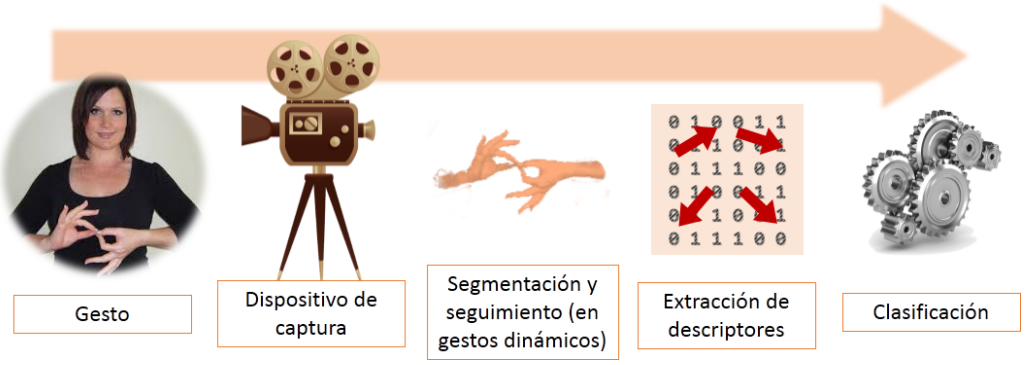


Figura 2.1: Arquitectura típica de un proceso de reconocimiento de gestos. Idea general tomada de [18] y [120].

extracción de descriptores. No obstante, esto podría ser específico del dispositivo de captura utilizado (como se verá más adelante). Por esta razón, aquí se decidió no incorporar esta etapa. Las próximas secciones intentan abarcar completamente estas diferentes etapas de procesamiento en un sistema de reconocimiento de gestos.

El presente capítulo presenta el problema del reconocimiento de gestos dinámicos, con particular interés en la lengua de señas, junto con una revisión de las técnicas y bases de datos existentes en la bibliografía. Todos los temas tratados en este capítulo están interconectados, y es difícil hablar de uno sin hablar de los otros. Los autores, incluso en revisiones específicas, suelen tocar diversos temas al mismo tiempo. No obstante, el capítulo se organiza el siguiente modo. La sección 2.1 presenta las definiciones básicas sobre gestos dinámicos, junto con la presentación de la lengua de señas como área de aplicación concreta. La sección 2.2 presenta una revisión general de los métodos de obtención de descriptores para gestos dinámicos. La sección 2.3 analiza las diferentes estrategias/métodos utilizados por los principales autores en la temática. La sección 2.4 presenta una revisión de las bases de datos existentes para reconocimiento de gestos dinámicos, y especialmente de la lengua de señas. Por último, la sección 2.5 presenta las conclusiones del capítulo.

2.1 Gestos dinámicos

Los gestos son el medio más antiguo de comunicación humana. Junto con la voz, siguen siendo actualmente una de las principales formas de comunicación, ya que son utilizados por las personas de manera inconsciente en la vida cotidiana, incluyendo también situaciones donde únicamente es posible comunicarse a través de este medio. El estudio de los gestos tiene sus orígenes en las ciencias sociales, para comprender y explicar los diferentes modos de comunicación no verbal. Existen diversos estudios sobre la relación de los gestos con el habla, el pensamiento, problemas clínicos, etc. como pueden mencionarse [100, 24]. Desde un punto de vista informático, existen diversos modos de entender un gesto, dependiendo de la aplicación que se quiera

Figura 2.2: Taxonomía de gestos de acuerdo a su característica temporal.

realizar. A continuación se intenta exponer ciertas taxonomías para comprender mejor su naturaleza y cómo afrontar un sistema de reconocimiento.

Existen diversas definiciones en la literatura, dependiendo el enfoque del trabajo que se esté proponiendo. Quizá la más apropiada aquí, sea la de algunos autores, como D'Orazio en [41], que propone un *gesto* como una forma no verbal de comunicación en la cual movimientos particulares del cuerpo codifican un mensaje intencionado. Aquí se diferencia *gesto* de *acción*, entendiendo a esta última como un gesto sin intención. Por ejemplo, los movimientos que hace una persona en su vida cotidiana como correr, saltar, acostarse, levantarse, beber agua, etc. Incluso existen bases de datos con acciones muy específicas en videos reales con escenas de películas, deportes, etc. Esta distinción entre acciones y gestos no resulta del todo importante en lo referente a técnicas de clasificación, pero los sistemas finales suelen tener objetivos diferentes, y el abordaje que se le da al problema no siempre es igual. Generalmente, los estudios realizados para acciones humanas se enfocan a movimientos hechos con todo el cuerpo, mientras que los trabajos enfocados a reconocimiento de gestos tienen particular interés en el rostro y gestos hechos con las manos, como la lengua de señas. Pueden encontrarse diversas revisiones particulares para reconocimiento de acciones como pueden ser [159, 123, 41]. En esta Tesis se prestará principal atención a los trabajos realizados en reconocimiento de *gestos*, entendiendo a estos con la definición antes mencionada. No obstante, se analizan algunos campos de investigación en acciones humanas, incluyendo un trabajo experimental, presentado en el capítulo 5.

Si bien no existe una taxonomía única para clasificar los gestos, existen ciertos criterios básicos donde los autores coinciden para distinguirlos. Estas clasificaciones resultan muy importantes a la hora de desarrollar un sistema informático, para entender qué se debe mirar y cómo se debe afrontar. Algunos autores clasifican los gestos según su interpretación, es decir una taxonomía en base al significado del gesto. Desde el punto de vista del uso comunicacional de los gestos, quizá la clasificación

más importante es la de McNeill [100], donde propone la existencia de cuatro tipos principales de gestos: icónicos, metafóricos, ilustrativos y deícticos. También existen otras clasificaciones donde aparecen términos como *emblemas*, *ilustraciones*, *reguladores*, *pantomimas*, *gestos icónicos*, entre otros (ver [120], [100], o [118]). Por otro lado, desde el punto de vista de las ciencias de la computación, se consideran aspectos tanto enfocados al sensado como al reconocimiento. Una las principales taxonomías existentes aquí radica en su componente temporal. De este modo, como define Pisharady en [120], los gestos pueden clasificarse en base a sus características observables como estáticos, y dinámicos. En la primer categoría se encuentran los gestos que sólo requieren de una imagen instantánea para su formulación. Por ejemplo, es posible encontrar poses de una mano, expresiones faciales (risa, enojo, tristeza, etc.), y otros. Los gestos dinámicos, en cambio, involucran un movimiento con alguna parte del cuerpo, que de no tenerlo, carecen de significado. Este es el ejemplo de un movimiento de saludo con una mano, patear una pelota en un juego deportivo, o casi cualquier gesto de la lengua de señas. La figura 2.2 muestra esta clasificación con alguno ejemplos gráficos. Naturalmente, los gestos dinámicos no pueden representarse totalmente en una imagen bidimensional.

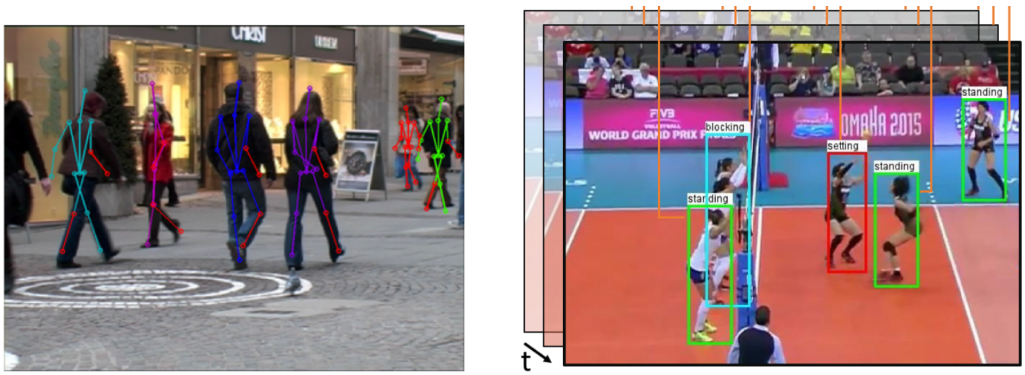
2.1.1 Acciones humanas

Como se mencionó anteriormente, diversas lineas de investigación están enfocadas en el reconcoimeinto de acciones humanas, entendiendo estas como gestos que las personas realizan en su vida cotidiana o entorno específico. Estos trabajos generalmente utilizan videos o bases de datos de entornos reales donde se intenta identificar qué tipo de movimiento está realizando una persona. Algunos enfoques, como en [7], se especializan en reconstruir el esqueleto de la persona, para en una etapa posterior realizar un reconocimiento. Dado un video real de una cámara convencional, identificar personas y reconstruir la información de las diferentes partes del cuerpo, no es un tarea trivial y aún no está resuelto de forma completa. La figura 2.3 muestra dos ejemplos de la bibliografía donde pueden verse aplicaciones de estos trabajos.

Dentro de los trabajos enfocados en reconocimiento de acciones humanas existen dos enfoques principales, dependiendo el tipo de dispositivo de captura utilizado. Los ejemplos mostrados en la figura 2.3 utilizan cámaras convencionales 2D. Como se verá en la sección 2.1.4 existen dispositivos que identifican automáticamente el esqueleto humano mediante cámaras de profundidad. Si bien estos enfoques no pueden realizarse en todos los entornos, reducen la complejidad al eliminar la etapa de estimación de posturas corporales.

2.1.2 Gestos con las manos

La mano es la parte del cuerpo más utilizada para realizar gestos, debido en parte a su naturaleza como medio de comunicación entre las personas [64]. Los movimientos



(a) (b)

Figura 2.3: Ejemplos de la bibliografía sobre trabajos de reconocimiento de acciones con cámaras convencionales 2D. (a) Estimación de poses y seguimiento de personas propuesto por Andriluka en [7]. (b) Reconocimiento de acciones complejas en un partido de Voleyball. Trabajo desarrollado por Ibrahim y Muralidharan en [70].

que realiza son sumamente más ricos que cualquier dispositivo actual de interacción humano-computador como puede ser un *mouse* o un teclado. Si bien, como ya se mencionó anteriormente, existen diversos trabajos de reconocimiento de acciones, los gestos hechos con las manos son el principal interés de los investigadores en la temática. La figura 2.4 muestra un gráfico basado en estudios realizados por Karam en [77], y analizado en [129] y [64]. El gráfico muestra las diferentes partes del cuerpo y otros objetos identificados en la bibliografía como los más utilizados para reconocimiento de gestos. Tomando los trabajos que utilizan solo las manos, llegan a más del 30 %. Considerando además los trabajos que utilizan las manos sumado a otra parte del cuerpo, los trabajos superan el 50 %. En ocasiones, diversos autores utilizan el término *gesto* para referirse solamente a gestos hechos con las manos.

Los gestos hechos con las manos pueden ser estáticos, o dinámicos, dada la clasificación comentada en párrafos anteriores. Generalmente, los gestos estáticos están conformados por poses de las manos con una postura particular de cada dedo. Los gestos dinámicos generalmente constan de ambos componentes: una componente estática en base a la forma que posee la mano, y un movimiento asociado que le da significado al gesto [129]. Esta idea se extiende a continuación.

2.1.3 Lenguas de señas

La lengua de señas es el idioma natural de las personas sordas. Es una lengua de expresión gestual, no verbal, que utiliza principalmente las manos como forma de comunicación. Desde el punto de vista del análisis de gestos, puede considerarse a la lengua de señas como gestos dinámicos hechos con las manos, aunque como se describe a continuación, involucra mucho más que eso.

Si bien el título de esta Tesis utiliza la palabra *lenguaje*, se optó aquí por hablar

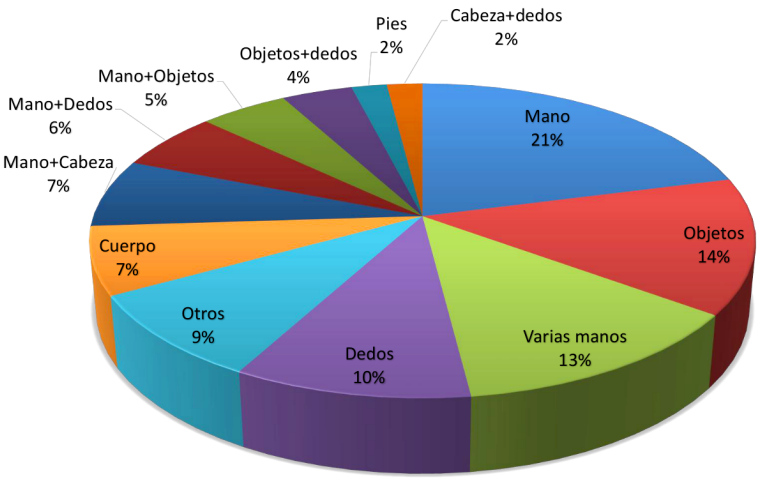


Figura 2.4: Partes del cuerpo u objetos utilizados para reconocimiento de gestos en la bibliografía. Basado en [77].

de *lengua*, ya que recientemente la Confederación Argentina de Sordos (CAS) [1] hace esta distinción. En este contexto la CAS define *lenguaje* como una función del ser humano que nos da la posibilidad de comunicarnos, pensar, reflexionar, etc. La *lengua*, se define entonces como un idioma de cada país, región o grupo humano.

Una particularidad interesante de la lengua de señas, es que al igual que otros idiomas, su alcance es regional. Es decir, cada país, e incluso regiones más pequeñas, tienen su propio idioma o dialecto. De aquí que se hable de "lenguas"de señas, ya que en realidad son muchos idiomas distintos. No obstante, los principios técnicos básicos de todas estas lenguas son los mismos, por lo que en este trabajo se llamará indistintamente *lengua* o *lenguas*. Esta particularidad regional, hace que la implementación de un sistema real que funcione localmente, involucre la existencia de un grupo de datos correctamente creado para este fin. Es decir, en el caso de Argentina, es necesario contar con una base de datos particular con señas del léxico argentino.

La lengua de señas generalmente tiene un espacio cercano al cuerpo donde se realizan los movimientos [37]. En entornos formales, como la traducción en un programa de televisión, o evento, los intérpretes de señas utilizan ropa oscura o neutra con un fondo claro. Esto permite identificar de forma más clara los diferentes gestos realizados por la persona. La figura 2.5 muestra un ejemplo de dos señas de la Lengua de Señas Argentina (LSA).

Las señas, en cualquiera de sus dialectos, constan de cinco componentes principales [37]: movimiento, posición, configuración, orientación y componentes no manuales. No obstante, como se describe más adelante en los ejemplos, en la mayoría de los sistemas informáticos sólo se utilizan los primeros tres. Los movimientos generalmente son cortos y precisos. Suelen ser lineales, con forma circular, o vibrantes, entre otros. La posición es otro elemento importante en la morfología de la seña. Generalmente se considera la posición de las manos en relación con otra parte

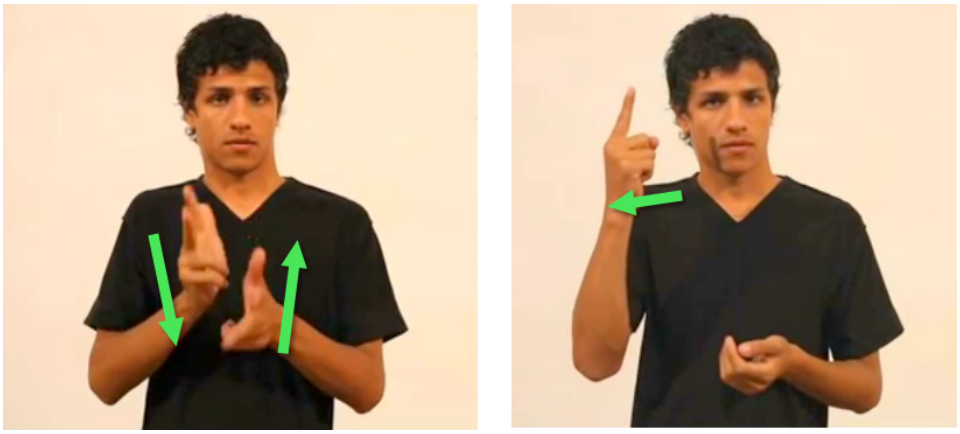


Figura 2.5: Ejemplos de señas realizadas por un intérprete de la Lengua de Señas Argentina (LSA).

del cuerpo como suele ser la cabeza, ojos, pecho, hombros, abdomen, etc. [37][10]. Por otro lado, el tercer componente, la configuración de las manos, es un elemento también clave. Esto es la postura que los dedos de cada mano tienen al realizar el gesto. En un seña, cada mano podría tener configuraciones diferentes, y cada configuración podría cambiar durante la ejecución. Suelen describirse las señas con una configuración inicial, y una configuración final (en cada mano) [2]. Si bien existen infinidades de posturas específicas para los dedos, y sobre todo para las diferentes regiones del mundo, cada lengua tiene su conjunto de configuraciones específicas definidas previamente. La figura 2.6 muestra el sistema de configuraciones completo para la LSA. Como toda lengua viva, esta información va cambiando con los años y en ocasiones los diccionarios y documentación existente debe ir actualizándose con nuevas señas, configuraciones, etc.

Si bien existe un sistema de configuraciones en cada lengua de señas, también existe lo que se conoce como diccionario dactilológico [37]. Es decir, una seña particular para cada letra del alfabeto hablado. Suelen ser unimanuales y generalmente se utiliza para traducir nombres propios y enseñar a las personas sordas el lenguaje verbal articulado de la región. Generalmente, las letras del abecedario suelen ser estáticas, quizá con excepciones de dos o tres letras con pequeños movimientos. Esto hace que los diccionarios dactilológicos se parezcan mucho a los sistemas de configuraciones de un sistema de lengua de señas. Existen diversos trabajos donde sólo se aborda el problema de clasificar un léxico dactilológico semi-estático (por ejemplo [5][76][101][15][161]), y también existen trabajos más robustos, donde se intenta reconocer un léxico dinámicos de señas, que necesariamente deben interpretar las diferentes configuraciones de manos (por ejemplo [144][135][30][145]).

El cuarto componente mencionado anteriormente, la orientación, se refiere generalmente al ángulo que poseen las manos al realizar la seña. Estas orientaciones pueden ser *palma hacia arriba*, *hacia abajo*, *palmas enfrentadas*, *palmas una sobre la otra*, *palmas juntas, hacia afuera o hacia adentro,* etc. [37]. Generalmente

Figura 2.6: Sistema de configuraciones de la Lengua de Señas Argentina (LSA).

esta información resulta redundante en un sistema informático debido al resto de componentes.

Es posible tratar el problema del reconocimiento de lengua de señas como un problema de reconocimiento de gestos con componentes únicamente manuales. De este modo, muchos trabajos se han enfocado en la identificación óptima de descriptores para luego utilizar clasificadores del estado del arte. No obstante, como indica Cooper en [29], la lengua de señas es mucho más que solo un conjunto de gestos. El quinto componente, existente en toda lengua de señas, son los Componentes No Manuales o Rasgos No Manuales (RNM) [1] [87]. Esto se refiere, por ejemplo, a expresiones faciales, orientación del cuerpo, etc. Muy pocos trabajos actualmente tratan este tipo de componentes, sobre todo por la poca disponibilidad de bases de datos que los incluyan. En adición a todos los componentes antes mencionados, la lengua de señas, como todo idioma, posee una gramática particular compleja [10]. Un sistema de reconocimiento robusto de lengua de señas, indefectiblemente necesitaría introducir estos conceptos gramaticales para realizar una traducción apropiada. Actualmente existen muy pocos trabajos que realicen un enfoque gramatical de las lenguas de señas. Quizá debido a que el estado del arte está todavía en un estado previo, intentando solucionar problemas de visión por computador.

Desafíos para construir un sistema de reconocimiento automático

Desde el punto de vista informático, el reconocimiento de lengua de señas es una temática sumamente interesante como campo de aplicación del reconocimiento

de gestos en general, con apenas un par de décadas de investigación [87]. Llevar a cabo un clasificador automático de lengua de señas involucra diversas disciplinas a considerar, todas con un alto grado de complejidad [118]. Dependiendo el objetivo final del sistema a desarrollar, existen diferentes enfoques encontrados en la literatura. El abordaje más simple, pero no menos importante, es el de la clasificación dactilológica, también conocido en el estado del arte como *fingerspelling*. Básicamente se trata de sistemas de clasificación de imágenes (gestos estáticos), donde se intenta discriminar el alfabeto de una lengua hablada representado en lengua de señas. Existen numerosos trabajos que pueden encontrarse en esta área, entre los cuales se pueden mencionar los siguientes. En [76], [126] y [101] se realiza clasificación dactilológica para la lengua de señas de Estados Unidos de América. Bastos en [15] realiza lo mismo para la lengua de señas de Brasil. En [19] se utiliza la lengua de señas colombiana. En [46] la lengua de señas arábica. En [161] puede encontrarse un ejemplo para la lengua de señas china. Por último en [5] se realiza un clasificador dactilológico para la lengua de señas etíope. Como puede apreciarse, existen numerosos esfuerzos aplicados para reconocer las distintas lenguas de cada región. En los siguientes capítulos se presentarán las bases de datos realizadas en esta Tesis para la lengua argentina, junto con diversos trabajos publicados en cuestión.

El segundo enfoque, está orientado a la clasificación de señas segmentadas, entendidas como gestos dinámicos. Aquí ya no alcanza con la descripción y clasificación de una imagen, ya que la seña necesita de una variable temporal. Más allá del dispositivo de captura utilizado (como se verá en las próximas secciones), es necesario analizar todos o algunos componentes del dinamismo de la seña como se mencionó anteriormente. Esto incluye también la información de configuración de la mano, por lo que muchos trabajos de este tipo necesitan de trabajos como los mencionados en el párrafo anterior para analizar parte del gesto. Generalmente se suele decir *señas segmentadas* ya que existe un cuerpo de datos con un video por cada seña que se quiere reconocer. La clasificación entonces se realiza por cada muestra. De este modo, se distingue de lo que se conoce como *videos continuos* o *sentencias*, donde varias señas pueden aparecer en una misma muestra. El mayor problema aquí es distinguir cuándo termina una seña y comienza otra, teniendo que lidiar con los movimientos no deseados entre estos huecos temporales. En [153], von Agris explica detalladamente esta diferencia entre clasificación de señas aisladas y continuas. No obstante, en ocasiones estos dos enfoques se mezclan, y los autores no dan demasiados detalles de si clasifican sólo señas segmentadas o también en sentencias continuas. Muchas veces esto tiene que ver con el método de clasificación utilizado. Al igual que para las señas dactilológicas, aquí existen numerosos trabajos que se revisarán con mayor detalle en las próximas secciones. Entre los más recientes es posible mencionar [88], [30].

Por otro lado, de un modo muy similar a lo ocurrido en la voz humana, la articulación de una seña generalmente varía en velocidad y amplitud [153]. Esto se da también con otros tipos de gestos dinámicos. Incluso cuando la misma persona realiza el mismo gesto una y otra vez, podría ocurrir que lo realice de diferente

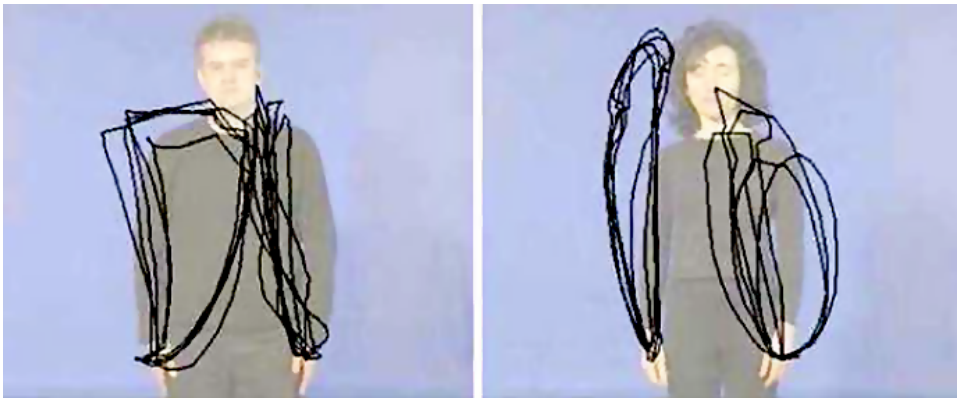


Figura 2.7: Una misma seña ejecutada por dos intérpretes distintos. Ejemplo tomado de [153], donde se ejecuta la palabra *tenis* de la Lengua de Señas Británica cinco veces por cada sujeto. Las lineas negras muestran la trayectoria de las manos.

forma. La figura 2.7 muestra un ejemplo donde dos interpretes ejecutan la misma seña repetidas veces. Es posible notar la trayectoria de las manos al realizar cada una de las repeticiones. En general cada repetición es diferente pero siguiendo una tendencia particular al modo en que ejecuta la seña el sujeto. Dependiendo el sistema que se quiera desarrollar, esto podría no ser un problema tan grande, por ejemplo, si al sistema lo usaran siempre las mismas personas. De todos modos, esto no sería lo habitual. Por estas razones, este es uno de los desafíos más grandes en un sistema de reconocimiento, ya que debe generalizar adecuadamente para adaptarse a esta plasticidad que puede ocurrir en la dinámica de las señas. Generalmente se conoce este problema como *dependencia al sujeto*. Un conjunto de datos correctamente construido debería tener diversos sujetos experimentales para poder realizar este tipo de validación.

Por ejemplo, en [76] se reporta una tasa de acierto (dependiente al sujeto) casi perfecta de 99,9 %, para una base de datos actual y reconocida en gestos estáticos. Sin embargo al reportar cómo se comporta el sistema con un nuevo sujeto, esta precisión cae hasta un 84 %. Esto es algo usual en la mayoría de los trabajos. Incluso diversas publicaciones no reportan resultados experimentales al ingresar un nuevo sujeto al sistema. En la sección 2.3 se analizan algunas de las herramientas más utilizadas para superar estos problemas.

Uno de los enfoques llevados a cabo en los últimos años intenta encontrar *sub-unidades* léxicas en el idioma. Similar a identificar vocablos, o fonemas, en un lenguaje hablado. Cada seña, entonces, está dividida en varias sub-unidades, que representan una pequeña parte del gesto, y que podría estar incluida en otra seña. La idea de este enfoque es reducir la cantidad de datos necesarios para entrenamiento. Del mismo modo, el tamaño del vocabulario podría agrandarse fácilmente por la concatenación de nuevas configuraciones de sub-unidades. No obstante, como explica von Agris en [153], la información de qué sub-unidades conforman cada seña no siempre está disponible, o mejor dicho, casi nunca. Menos aún una descripción

semántica de qué está representando una sub-unidad. Los primeros trabajos existentes son relativamente recientes, en el año 1995 con Waldron y Kim [154], en el año 1999 con Vogler [149] y 2004 con [17]. Más recientemente existen estos enfoques en trabajos como el de Roussos en [135], Pitsikalis en [122], el de Cooper en [30], o Koller en [86].

2.1.4 Tecnologías de captura

Existen diversos dispositivos de captura que se han usado y siguen usando para capturar gestos. Dependiendo el dispositivo, será el formato que tendrán los datos y será más simple o complejo su procesamiento. Berman y Stern realizan una amplia taxonomía en [18] sobre diferentes tipos de sensores existentes. En el trabajo, se presenta una clasificación que separa a los dispositivos en eléctricos, ópticos, acústicos, magnéticos, y mecánicos. Quizá los menos frecuentes en esta clasificación sean los acústicos. Estos sensores utilizan ondas de ultrasonido y el efecto Dopler para calcular la posición de las diferentes partes del cuerpo humano. Generalmente son baratos pero poco precisos. Los dispositivos magnéticos y mecánicos, también suelen ser poco frecuentes, aunque diversos sensores pueden encontrarse en guantes u otros tipos de trajes con acelerómetros, sensores de presión, etc.

Los sensores electricos son muy populares en la actualidad, sobre todo por su utilización en teléfonos inteligentes de uso cotidiano. Estos se encuentran por ejemplo en las pantallas táctiles. Estas ueden existir de muchos tipos como eléctricas, capacitivas, etc. Muchos dispositivos actuales utilizan pequeños gestos manuales para reaccionar con alguna operación específica del sistema operativo o aplicación. Generalmente son gestos pequeños, hechos con los dedos de una o dos manos, con el objetivo de llevar a cabo interacción directa con una aplicación. Por ejemplo, en una cámara de fotos con pantalla táctil es posible enfocar, acercar o alejar la imagen, disparar la foto, etc.

Guantes de datos y Marcadores de color

Los guantes de datos, o sistemas de control basados en guantes, son dispositivos que se colocan en las manos para poder sensar la posición y movimientos de la misma. Fueron especialmente diseñados para reconocimiento de gestos, y en muchos casos se han aplicado particularmente a la lengua de señas [18]. Los guantes generalmente poseen sensores táctiles o de otros tipos, capaces de realizar un mapeo exacto del movimiento realizo por las diferentes articulaciones de las manos [125]. Una gran ventaja de esto es no necesitar una etapa de procesamiento de datos para obtener descriptores, como puede ser el caso de una imagen obtenida de una cámara.

Premaratne define en [125] dos categorías principales de guantes: activos y pasivos. Podrían considerarse también invasivos o no invasivos, siguiendo las categorías de Berman en [18]. Los sistemas activos incluyen a todos los guantes que poseen algún tipo de sensor, para capturar la flexión de los dedos, o acelerómetros para

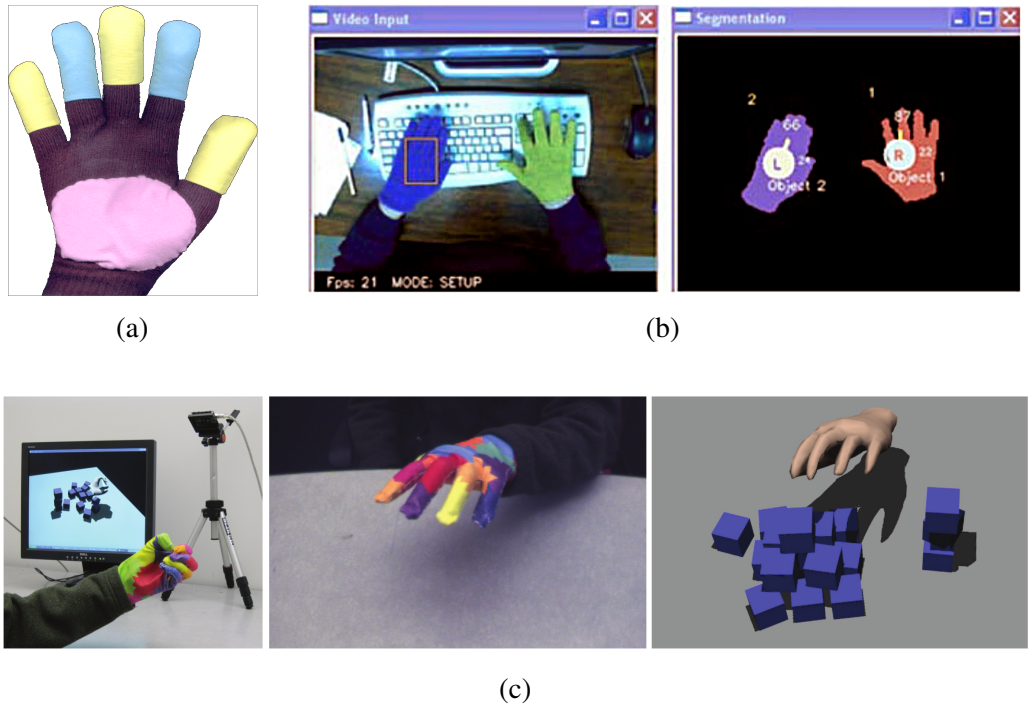
(a) CyberGlove III [3]. (b) AnthroTronix acceleGlove [4].

Figura 2.8: Ejemplos de guantes de datos (*data-gloves*) actuales.

sensar el movimiento. Tradicionalmente esto guantes tenían además varios cables que debían conectarse a una computadora, lo cual resultaba muy molesto. Hoy en día, con las tecnologías inalámbricas (o *wireless*) este problema ya casi no existe. Puede encontrarse una revisión muy detallada de este tipo de dispositivos en [40], donde se analizan también las principales aplicaciones en las cuales se utilizan estos guantes. Particularmente en al área de la lengua de señas hubo un particular interés en la década de 1990, por ejemplo [143]. Luego, estos dispositivos dejaron de ser foco de interés de investigadores para esta temática, en general por su alto costo y baja practicidad a la hora de introducirlo en un entorno real. No obstante, siguen existiendo nuevos trabajos. Por ejemplo en [138] se utiliza un nuevo dispositivo específicamente creado para reconocer la lengua de señas de Malasia. Diversos trabajos utilizan también estos dispositivos como interfaz para control robótico y más recientemente para análisis médico.

Existen numerosos ejemplos de estos dispositivos tanto comerciales como de uso académico. La figura 2.8 muestra dos ejemplos de guantes de datos de los más renombrados actualmente. El CyberGlove [3] es un guante desarrollado por CyberGlove Systems, siendo uno de los más utilizados comercialmente. En la actualidad han llegado a desarrollar el CyberGlove III (el tercero en su tipo). Este dispositivo fue particularmente desarrollado para la industria del cine y animación gráfica, pudiendo reproducir de modo muy preciso los movimientos que hace una persona, para mapearlos en una computadora. EL MIT AnthroTronix acceleGlove [4] es un guante con acelerómetros de costo relativamente bajo, comparado a otros. Puede sensar el movimiento 3D de todos los dedos de la mano.

Por otro lado, los guantes pasivos, o no invasivos, hacen referencia a dispositivos no electrónicos, que solo poseen ciertos marcadores de color para facilitar la identificación en una imagen [109]. Estos dispositivos se utilizan en un sistema de visión por computador, es decir capturados por algún tipo de cámara que genera una imagen 2D o 3D, como se detalla en la siguiente sección. Generalmente son guantes de tela, látex u otro material simple y económico. Los primeros de este tipo



(a) (b)

(c)

Figura 2.9: Ejemplos de guantes de color encontrados en la bibliografía. (a) Guante con tres colores desarrollado por Lamberti en [91]. (b) Guantes de dos colores para segmentar gestos simples [55]. (c) Guante desarrollado por el MIT en [158] para tareas de realidad virtual.

de guantes se remontan a 1980, cuando el MIT Media Laboratory creó el MIT-LED glove [140]. Se basaba en un guante con leds (diodos emisores de luz) para luego poder reconocerlos a través de una computadora. Fue usado principalmente para capturar movimiento y no para procesos de control.

En la actualidad, los sistemas de reconocimiento de gestos basados en visión abarcan la mayoría, en comparación a los que utilizan sistemas invasivos como los guantes de datos. Diversos autores siguen utilizando y desarrollando nuevos guantes como marcadores de color para facilitar la segmentación en una imagen. La figura 2.9 muestra tres ejemplos recientes de la bibliografía. Lamberti y Camastra desarrollaron en [91] un guante simple con tres colores distintos para poder reconocer la palma de la mano junto con los dedos (figura 2.9-a). En el trabajo segmentaban los distintos colores del guante bajo el modelo de color HSI [56], y luego lograron reconocer más de 900 gestos distintos. Genç et al. presenta en [55] un sistema simple que utiliza dos guantes con colores llamativos para segmentar las manos y luego utilizar gestos simples para hacer de interfaz en un sistema de demanda de video (figura 2.9-b). Por último, Wang y Popovic en [158] desarrollaron un guante simple pero un poco más sofisticado, con diferentes marcadores de color, orientado principalmente para ser una interfaz en el manejo de aplicaciones de realidad virtual (figura 2.9-c).

El guante, como se ve en la figura, tiene un patrón muy particular, que los autores usaron para simplificar el problema de la estimación de la postura de la mano. Este tipo de accesorio, pero con un sólo color, es el que se utiliza en el trabajo principal de esta Tesis para facilitar la segmentación de las manos en las bases de datos de la Lengua de Señas Argentina.

A modo de resumen, cabe destacar que si bien ambos tipos de guantes (activos y pasivos) parecen similares a primera vista, el modo de recolectar los datos es muy diferente. Mientras que los guantes de datos suelen capturar información vectorizada y precisa de movimientos como ángulos, aceleración, posiciones 3D, etc., los guantes no invasivos son simples marcadores de color que requieren de un proceso posterior de visión por computador. Claramente, los primeros son más sofisticados, costosos e incómodos de usar que los segundos, aunque también más precisos.

Cámaras RGB y RGB-d

Sin duda, los dispositivos más ampliamente utilizados en la actualidad para captura de gestos son las cámaras RGB y RGB-d [125]. Las cámaras fotográficas digitales o de video actuales utilizan un sensor llamado CCD en forma de grilla cuadriculada, que capturan la luz en tres canales principales de luz: roja verde y azul, formato generalmente conocido como RGB (por sus siglas en ingles, *Red, Green, Blue*) [56]. La imagen resultante, sin importar el formato, generalmente se organiza como una matriz de pixeles, cada uno con sus tres valores RGB. Luego, cualquier intento de identificar un objeto dentro de la imagen resultará en un algoritmo de visión por computador, donde básicamente se trata de darle un significado al orden e intensidad de color de los diferentes pixeles. Los marcadores de color, como los guantes antes mencionados, facilitan este proceso de segmentación.

Una variación de las cámaras convencionales son las cámaras de profundidad, o también llamadas RGB-d en algunos casos. Estas, además de los tres canales de color, presentan un cuarto canal que representa información de profundidad, es decir, la distancia entre el foco y el objeto capturado. Si bien se han utilizado diversas cámaras de profundidad para sistemas de reconocimiento de gestos, tradicionalmente han tenido un alto costo [120]. En los últimos años, el nuevo dispositivo creado por Microsoft, el Kinect, ha bajado notablemente los costos, siendo un gran atractivo para muchos investigadores [167]. Este dispositivo fue creado para la consola de video-juegos *Xbox 360* a fines del año 2010, con el objetivo innovador de poder interactuar sin la necesidad de un dispositivo físico manual con una aplicación. La consola fue muy utilizada pero más allá de su furor dentro del entorno de los juegos, pasó a tener un gran atractivo por la comunidad científica.

El Kinect posee una cámara RGB convencional formato VGA, es decir, con una resolución de 640x480 pixeles (relativamente baja comparada a las cámaras actuales) a 30 frames por segundo. Por otro lado, posee un emisor de rayos infrarrojos, junto con una cámara que los detecta. De este modo, puede computar la distancia existente entre el foco y el objeto donde rebotan esos rayos. En la figura 2.10 puede apreciarse

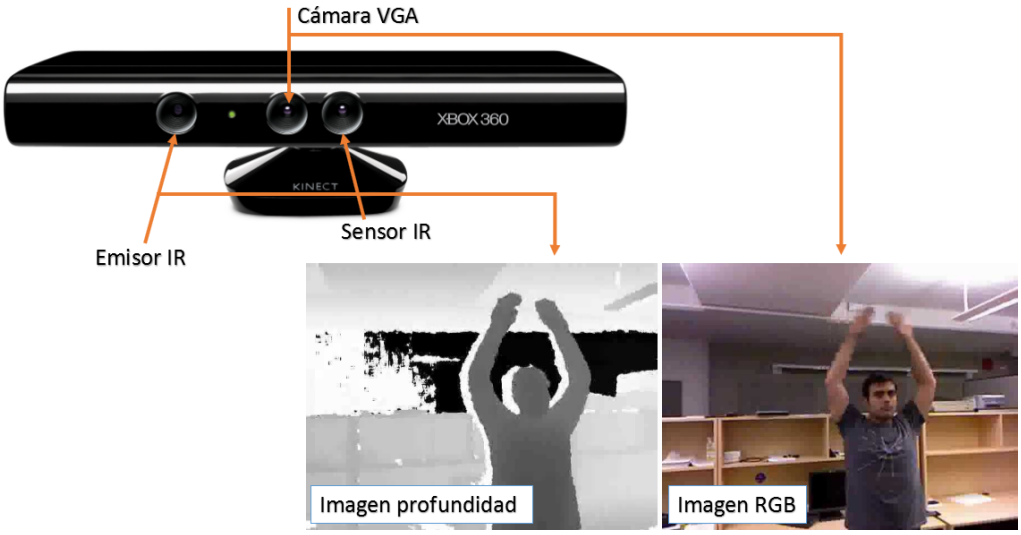


Figura 2.10: Descripción general del MS Kinect y sus componentes, junto con un ejemplo de captura.

el dispositivo junto con sus componentes principales y un ejemplo de sensado. Al utilizar luz infrarroja no resulta invasivo para la persona que lo está utilizando, ya que no es visible por el ser humano. De este modo se obtiene, junto con la imagen RGB, una imagen de profundidad, lo cual resulta de gran ayuda para identificar de forma general los objetos en la escena.

Un gran atractivo para los investigadores, además, es que el Kinect posee una librería de desarrollo que automáticamente estima el esqueleto de una persona, junto con las articulaciones del cuerpo, siempre que el sujeto se pare a una distancia medianamente estipulada, y en entornos controlados. En ocasiones incluso las bases de datos actuales poseen datos erróneos, como esqueletos que se deforman mientras se ejecuta un gesto. No obstante, su simpleza y bajo costo lo hace atractivo para numerosas investigaciones y bases de datos. La figura 2.11 muestra un esquema de las 20 articulaciones que el dispositivo logra reconocer. Esta información vectorizada, permite computar rápidamente la posición del cuerpo sin la necesidad de algoritmos de visión por computador. Las etapas de preprocesamiento de imágnes

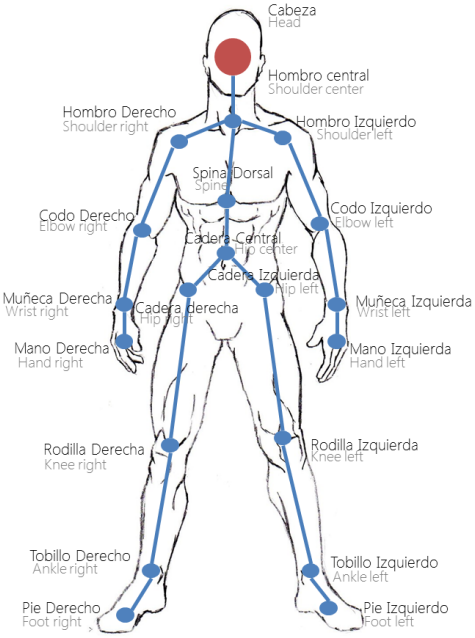


Figura 2.11: Esqueleto reconocido por el MS Kinect junto con las diferentes articulaciones.

(como se verá en secciones siguientes) no son necesarias aquí, siempre y cuando lo único que interese sea el movimiento corporal. Cabe destacar además, que si bien se ha utilizado el Kinect para diversos estudios de acciones humanas, gracias a este potencial de estimar el esqueleto, su utilización para reconocimiento de gestos con las manos es más acotado, ya que la mano posee un área muy pequeña que resulta difícil de identificar en la imagen. También se ha utilizado este dispositivo para seguimiento de objetos y reconstrucción 3D de escenarios [62].

Existen numerosos trabajos desarrollados con este dispositivo, incluyendo diversas revisiones. En [41], D'Orazio et al. proponen un interesante análisis de cómo la información de profundidad en una imagen pueden mejorar los enfoques tradicionales de reconoicmiento de gestos, haciendo una gran revisión bibliográfica. Si bien el enfoque de esta Tesis está puesto principalmente en la clasificación utilizando cámaras RGB convencionales, en las próximas secciones se comentan algunos trabajos que utilizan este dispositivo para reconocer acciones humanas. Luego, en el capítulo 5 se presenta un trabajo experimental de la estrategia desarrollada en esta Tesis sobre bases de datos obtenidas con el Kinect.

2.2 Obtención de descriptores para gestos

Como se mencionó a comienzos de este capítulo, el cómputo de descriptores en un gesto es una etapa clave en el proceso de reconocimiento. Si bien los métodos de clasificación merecen generalmente la principal atención en este tipo de sistemas, no segmentar las partes del cuerpo y generar adecuadamente los descriptores de un gestos podría significar en aumentar exponencialmente la complejidad. En esta sección se analizan algunos de los métodos más utilizados en la bibliografía para estas tareas.

La idea de obtener un descriptor radica en tener una representación vectorizada de la información relevante a la muestra. En el caso de los gestos, la información generalmente está en formato imagen/video. Por esta razón es necesario transformar del dominio de los pixeles a una representación más sencilla de segmentar/clasificar. En ocasiones, la idea de clasificador y descriptor puede estar mezclada. Por ejemplo, es posible utilizar la salida de un clasificador simple, como parte de un descriptor para formar una representación de alto nivel, y así alimentar a un clasificador más robusto.

Generalmente, los descriptores utilizados dependen mucho del resultado al que se quiere llegar, o del tipo de base de datos utilizada. Por ejemplo, no es el mismo problema si se quiere segmentar la mano de una persona en un fondo negro y constante que si se encuentra en un entorno real no controlado, con fondo complejo. Además, la diferencia entre descriptores estáticos y dinámicos es muy sutil, ya que en los gestos dinámicos la información es similar que en los estáticos, pero agregando una nueva variable. No obstante, como el principal interés en esta Tesis es el reconocimiento de lengua de señas en videos RGB, resulta importante hacer un

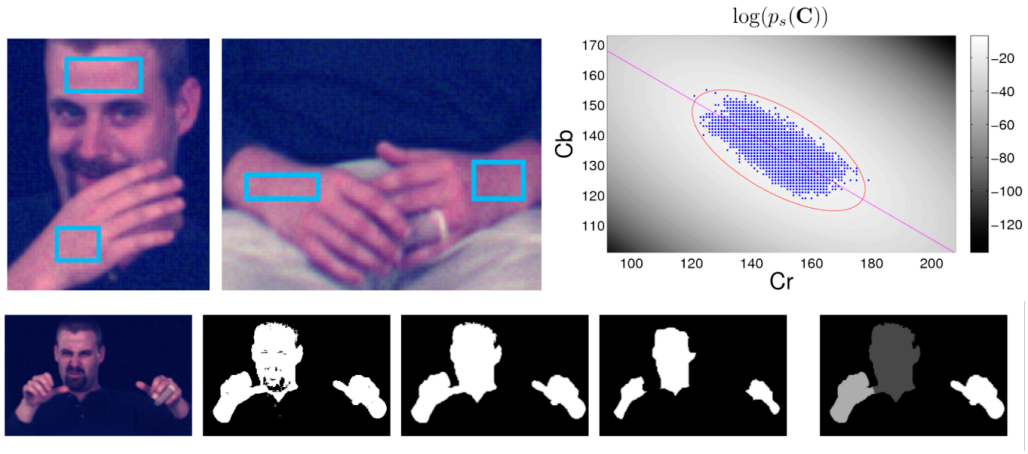
relevamiento de cómo segmentar las manos de un sujeto, los descriptores existentes en imágenes, y particularmente utilizados para configuraciones (o posturas) de manos. Por esto motivos, se comienza analizando las técnicas de segmentación y seguimiento de manos y rostros, para luego revisar los diferentes descriptores existentes para imágenes como así también en gestos dinámicos, tanto para lengua de señas como para acciones humanas.

2.2.1 Segmentación y Seguimiento

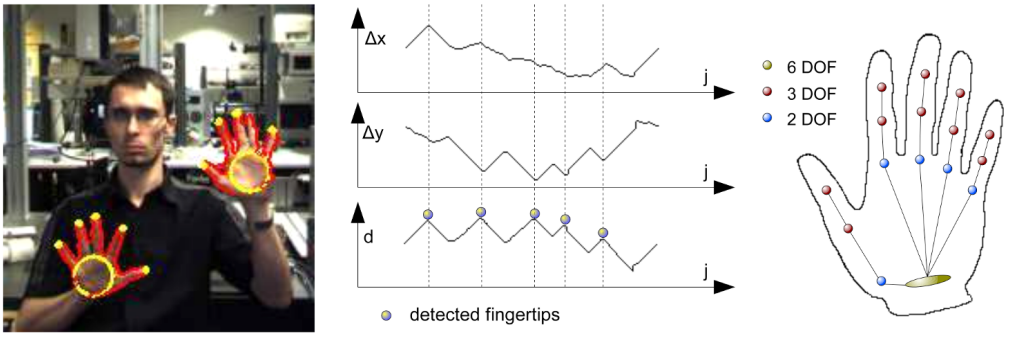
La primer etapa en un sistema de reconocimiento de gestos basado en visión, es la segmentación de las regiones de interés. Particularmente para la lengua de señas esto involucra dos puntos claves: la segmentación de ambas manos y la segmentación del rostro. Esto puede ser un gran desafío, ya que no resulta para nada sencillo construir algoritmos capaces de lidiar con las muchas situaciones diferentes que pueden darse en el mundo real. Por ejemplo, diferencias en iluminación entre una escena y otra, diferentes vestimentas de los sujetos que realizan los gestos, etc. Por otro lado, para gestos dinámicos, es necesario hacer un seguimiento de las manos del sujeto. Esto es una tarea por de más compleja. Las manos son objetos deformables que cambian de posturas al igual que de posición relativamente rápido. Esto involucra muchas veces suciedad en la imagen, dependiendo el dispositivo de captura utilizado. Incluso, existe superposición entre las manos y entre manos y rostro. A continuación se comentan diferentes aproximaciones encontradas en la literatura sobre segmentación y seguimiento (*tracking*) de manos y rostros. Algunos enfoques sólo utilizan información estática, es decir, imágenes RGB sin contexto temporal, mientras que otros hacen uso del dinamismo del video, entendiendo este como una colección de imágenes, para llevar a cabo un seguimiento más robusto.

Separar la información relevante, como las manos, el rostro, etc. de la irrelevante, como cualquier otro objeto en el fondo de la escena, puede resultar totalmente trivial para la mente humana, pero aún es muy complejo para una computadora. Generalmente, las diferentes estrategias desarrolladas en la actualidad, traen aparejadas condiciones particulares en las que funcionan. Quizá la más fuerte en cuanto a gestos y particularmente a lengua de señas sea sobre el fondo y la vestimenta. Si la persona se encuentra ubicada en un ambiente real, la detección de las manos y/o cara puede ser muy complicada debido al fondo complejo. Por ejemplo, en [26] se segmenta la mano en un entorno con fondo real suponiendo gestos no estáticos. Esta es una asunción bastante común pero poco escalable, ya que el algoritmo no funciona correctamente al quedar la mano quieta. Cabe destacar aquí, que la gran mayoría de sistemas de reconocimiento de lengua de señas usan la premisa de tener un ambiente totalmente controlado, con fondo uniforme, y sujetos con vestimenta generalmente oscura. Esto simplifica mucho el problema y es un escenario bastante usual, por ejemplo en interpretación a señas en programas televisivos.

Quizá el modelo más ampliamente utilizado para segmentar las manos y rostros es el que utiliza el color de la piel (ampliamente conocido en el estado del arte



(a)



(b)

(c)

Figura 2.12: Ejemplos encontrados en la literatura para detección y segmentación de manos. (a) Segmentación por color de piel utilizado en [135]. (b) Segmentación utilizando modelo de color de piel sumado a información de profundidad obtenida con multiples cámaras en [8]. (c) Cámara termal analizada en [9].

como *skin detection*) [73]. Este método, se centra en buscar el tono de piel del sujeto/intérprete en la imagen. En [135] Roussos et al. utilizan un modelo estadístico obtenido a partir de segmentos tomados de los propios videos utilizados. Estos modelos, como puede verse en este trabajo, tiene el problema de generar muchas oclusiones al pasar la mano una sobre otra o una sobre el rostro (figura 2.12-a). Generalmente existen muchas aproximaciones para lidiar con esto, pero ninguna completa. Particularmente en el trabajo citado, se utiliza el *know how* del problema, haciendo asunciones del estilo "la cabeza siempre es más grande que las manos".

Las imágenes de profundidad, como las que genera el MS Kinect, obviamente son también una solución. Algo muy usual es utilizar diversas cámaras para generar mapas de profundidad. Por ejemplo en [60] Hadfield y Bowden utilizan un sistema de visión estereoscópica (dos cámaras) para segmentar las manos de una persona, utilizando la asunción de que las manos son los objetos más cercanos a las cámaras. En [8] se utiliza un modelo de piel sumado a un sistema de tres cámaras para obtener información de profundidad. La información adicional de la visión estereoscópica ayuda, pero como bien dicen los autores, involucra un aumento en complejidad computacional con diversos parámetros nuevos a calibrar. Algo parecido ocurre en [83] donde se utilizan cuatro cámaras para mejorar la segmentación de las manos para gestos estáticos de la lengua de señas de India. En [9] se compara incluso la utilización de cámaras termales para segmentación de manos y rostros. Esto, además de tener problemas similares a los modelos de color de piel, se agrega el altísimo costo de este tipo de dispositivos. La figura 2.12 ilustra gráficamente algunos de los ejemplos planteados.

Los filtros por color de piel pueden ser principalmente de dos tipos: pixel a pixel, o basados en regiones [125]. El primero simplemente clasifica cada pixel en base al modelo de color como positivo o negativo individualmente, sin tener en cuenta los pixeles vecinos. El segundo intenta tener en cuenta la disposición de los pixeles para mejorar la detección. Por ejemplo, en [124] se utiliza la noción de superpixel, para agrupar pixeles similares. Cada superpixel es clasificado como piel o no-piel agregando evidencia basada en histogramas Bayesianos, similares a lo propuesto en [73]. En [160] se utiliza un enfoque similar, pero utilizando además información temporal, comparando fotogramas consecutivos para obtener regiones homogéneas.

Si bien las imágenes obtenidas por una cámara suelen estar en formato RGB (como ya se mencionó en secciones anteriores), los filtros por color, ya sea para detectar la piel u otros colores, suelen transformar la imagen a otros modelos, debido a que las distancias en el espacio RGB no son análogas a la percepción humana [47]. Además, el modelo RGB no separa la luminancia de la crominancia, es decir, la información de color, de la información lumínica. Esto hace que los canales verde rojo y azul estén muy relacionados y buscar un color específico sea muy complejo. Generalmente para solucionar este problema se eligen los modelos HSV o YCbCr [56].

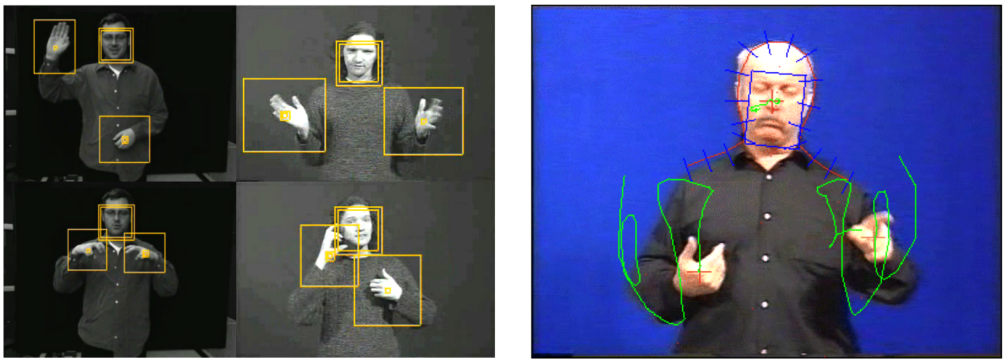
Claramente, los modelos de color de piel funcionan muy bien en entornos contro-

lados, donde el fondo de la escena es conocido, al igual que el tono de color de piel del sujeto a segmentar. Sin embargo, este último punto es algo casi incontrolable en un sistema medianamente robusto. En [47] Elgammal et al. grafican con ejemplos las diferencias del espectro de color para muchas personas, yendo desde gente asiática, hasta gente africana, dejando en evidencia, que el tono de piel de las personas puede tener un rango de valores tan amplio, que resulta imposible de generalizar. Si el rango se amplia tanto, entonces el filtro pierde calidad, obteniendo también valores de color candidatos eventualmente no deseados. Además, es muy común encontrar el tono de piel en diferentes texturas como el cabello, algún tipo de vestimenta, etc., lo que genera una enorme cantidad de falsos positivos [125].

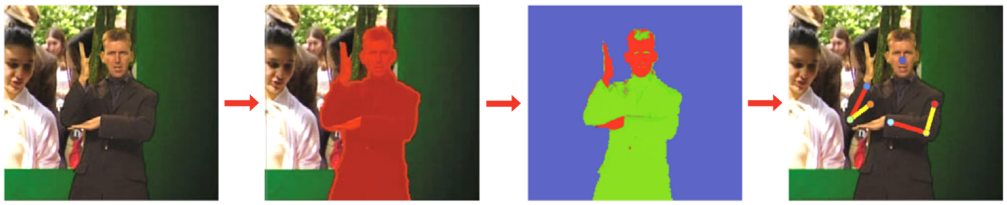
No obstante las dificultades encontradas con los modelos de piel, han sido ampliamente utilizados para realizar seguimiento de manos. En [61] se utiliza un modelo de color de piel para segmentar manos y rostro utilizando un filtro de Kalman [75] para solucionar los problemas de oclusión. En [66] Holden et al. utilizó un modelo de piel junto a un algoritmo de contorno activo llamado *snake*, sumado a información de movimiento para llevar a cabo el seguimiento de manos evitando las oclusiones de manos y rostro y clasificar señas de la lengua de señas australiana. En el trabajo se inicializan los *snakes* como elipses en las posiciones de las manos en el fotograma anterior utilizando un flujo óptico basado en gradiente, moviendo las elipses hacia las nuevas posiciones. Si bien mostró resultados favorables, estos algoritmos son muy susceptibles a fondos complejos y pueden confundirse con ropa de manga corta en los intérpretes [29].

Existen también trabajos más robustos, con segmentación basada en detección morfológica de manos. Por ejemplo, en [74], Kadir et al. proponen un reconocimiento de manos basado en algoritmos de *boosting* [136]. La idea de estos algoritmos es generar un clasificador "robusto" como una combinación lineal de clasificadores más "débiles". De este modo, en el trabajo se utiliza una cascada de clasificadores tanto para detectar las manos como la cara del intérprete. En este caso se utilizaron clasificadores *Viola-Jones* [148] que utilizan descriptores tipo *Wavelet de Haar* [94]. Si bien este trabajo presenta un enfoque mucho más sofisticado de segmentación, requiere miles de imágenes para entrenamiento (tanto positivas como negativas) que no siempre son simples de conseguir. En general, estos trabajos resultan difíciles de reproducir. En [25] se utiliza la idea de co-segmentación propuesta en [134], junto con Random Forest [21], para estimar las articulaciones de los brazos en un entorno real de interpretación de lengua de señas. En [42], Dreuw et al. propone un nuevo método de seguimiento de objetos, aplicado particularmente a la lengua de señas, basado en programación dinámica. Se compara el modelo con estrategias clásicas de alineamiento en reconocimiento de voz. La figura 2.13 resume gráficamente algunos de los trabajos mencionados en el párrafo.

Como ya se mencionó en secciones anteriores, otro modo de segmentar rápidamente las manos del intérprete es a través de guantes de color. Esto posibilita una segmentación casi perfecta, incluso con situaciones de iluminación cambiante,



(a)

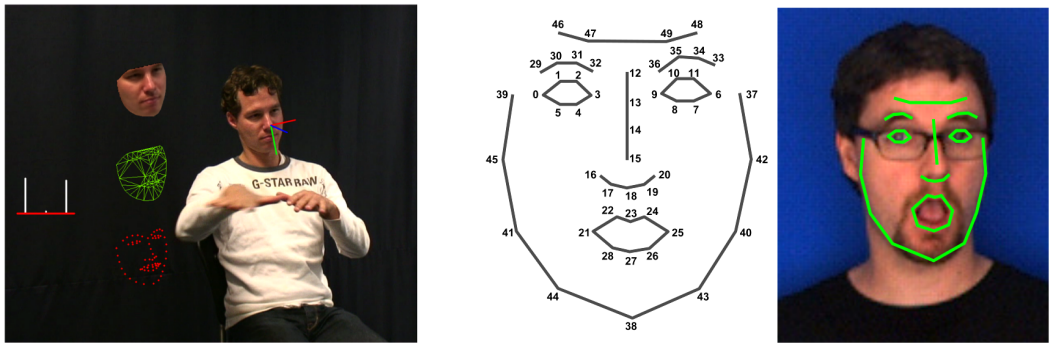


(b)

Figura 2.13: Ejemplos encontrados en la literatura para seguimiento de manos y rostros. (a) Seguimiento basado en cascada de clasificadores tipo Viola-Jones propuesto en [74]. (b) Co-segmentación y Random Forest utilizados en [25] para estimación y seguimiento de articulaciones de brazos en un entorno real de lengua de señas en TV.

fondos complejos, o vestimenta no controlada. En reconocimiento de lengua de señas la segmentación de manos por guantes de color resulta simple, rápida y eficiente.

En lengua de señas, al igual que en otros dominios de reconocimiento de gestos, es muy importante segmentar correctamente el rostro de la persona. Incluso si no se utilizara información facial, generalmente es necesario considerar las posiciones de las manos en relación a la posición del rostro, ya que este se utiliza para marcar diferentes gestos. Un error de calibración de cámara podría ser fatal al querer reconocer una seña, sin tener en consideración la distancia de las manos al rostro, incluso a posiciones específicas del rostro, como la boca, nariz, ojos, etc. Generalmente, la mayoría de los trabajos mencionados en párrafos anterior, no sólo segmentan las manos, sino también el rostro de los sujetos. La detección y segmentación de rostros es una tarea medianamente sencilla desde ya hace unos años [148]. Incluso existen aplicaciones comerciales, por ejemplo en casi todos los dispositivos móviles o cámaras fotográficas actuales, que reconocen en tiempo real el rostro. Esto se debe principalmente al patrón que poseen todos los rostros: dos ojos, nariz, boca, siempre siguiendo la misma disposición para toda persona. No obstante la segmentación, el cómputo de descriptores faciales no son tan triviales, como se verá más adelante.



(a) (b)

Figura 2.14: Extracción de características faciales con *Active Appearance Models*. (a) Ejemplo tomado de [119]. (b) Gráfico de 50 puntos de interés y su aplicación a un intérprete de señas.

Existen numerosos trabajos que intentan capturar información no manual para incorporar a un clasificador de lengua de señas. En [153] puede encontrarse un gran resumen de diferentes técnicas en esta temática.

Quizá, uno de los métodos más utilizados en los últimos años para segmentación se rostros, sean los modelos de apariencia activa (*Active Appearance Models - AAM*) formulados originalmente por Cootes, Edwards y Taylor en [32]. AAM son modelos generativos estadísticos que posibilitan relacionar con gran precisión los rostros de una persona luego de entrenar el modelo con un conjunto de imágenes de entrenamiento. Quizá la desventaja más grande sea que es necesario, como primer etapa, crear un modelo de rostro marcando una serie de puntos de interés de forma manual. No obstante, una vez entrenado el modelo, puede ajustar rápidamente los cambios que el rostro produce para todo tipo de expresiones, estimando la variación en los puntos de interés. Esto permite extraer rápidamente descriptores de alto de nivel sobre la expresión de una persona. Generalmente la búsqueda de los puntos se realiza de forma local, por lo cual en ocasiones primero se realiza un proceso de segmentación de rostros a través de algoritmos como el *Viola-Jones* [148]. Por ejemplo, en [119], Piater et al. utilizan esta técnica para construir un modelo capaz de reconocer rostros en una base de datos de la lengua de señas holandesa. En [151] von Agris et al. analiza la importancia de la utilización de rasgos no manuales en el reconocimiento de lengua de señas, siendo uno de los primeros trabajos en estudiar estos aspectos a fondo.

2.2.2 Configuraciones de manos

Algunos autores, como Cooper en [29] afirman que en los videos donde los intérpretes de lengua de señas están de cuerpo completo, la resolución del video suele no ser suficiente para capturar los detalles de las manos, haciendo imposible identificar con fiabilidad las diferentes configuraciones de las señas. Mientras que

Cooper afirma esto hace sólo cinco años, la evolución en dispositivos de captura es tan rápida, que deja en la actualidad la frase en duda. Los trabajos de investigación en reconocimiento de gestos son todos relativamente recientes, y a esto se le adiciona la evolución de estos dispositivos así como velocidad de cómputo.

Luego de realizar una correcta segmentación de la mano de un sujeto, obtener los descriptores que representen la forma que posee la mano no es una tarea para nada trivial. Claramente, esto depende de la cantidad y variación de los gestos que se quieren reconocer. Cuanto mayor sea la cantidad, más robusto deberán ser los descriptores para lograr una correcta separación de clases. De cualquier forma, como define Premaratne en [125], los descriptores deben ser lo más compactos posibles, y cada *cluster*, es decir cada conjunto de datos que representa una clase, debe estar lo más separado del resto posible. Existe una multiplicidad de técnicas distintas que se han ido usando los últimos años para llevar esta tarea a cabo. Se mencionan a continuación las más utilizadas en la literatura.

Descriptores geométricos de la mano

Quizá los descriptores más triviales de computar, que se utilizan en diversas investigaciones, sean los descriptores geométricos que generalmente se calculan sobre el contorno de la mano. En primer lugar es necesario aplicar algún algoritmo de filtrado de imágenes que permita identificar el contorno. Esto es una tarea relativamente sencilla hoy en día. Existen numerosos filtros llamados "filtros de detección de bordes" que realizan esta tarea eficientemente. Luego, diversas características de la forma pueden calcularse. Las más comunes son: el área, centro de coordenadas, orientación del eje principal, excentricidad, compacidad, etc. Un ejemplo de estos descriptores puede encontrarse en [151]. La figura 2.15 muestra un esquema gráfico de este trabajo.

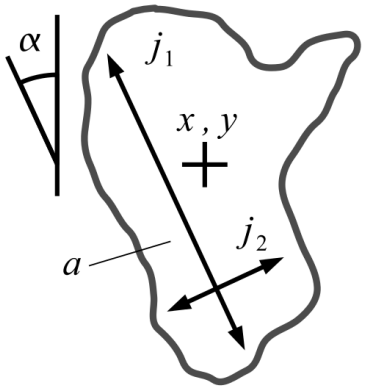


Figura 2.15: Ejemplo tomado de [151] del contorno de una mano segmentada para computar características geométricas.

Descriptores de Fourier

Los descriptores de Fourier han sido aplicados exitosamente a muchas tareas de visión por computador [56], principalmente como representación de formas en trabajos de reconocimiento óptico de caracteres y otros problemas de clasificación de imágenes. Al ser robustos al ruido, además de ser fáciles de derivar y simple de normalizar, han sido utilizados en numerosas aplicaciones. En 1972, Granlund los utilizó para reconocer huellas digitales debido a su simplicidad describiendo contornos [57]. Para el caso de configuraciones de manos, la idea es estimar los

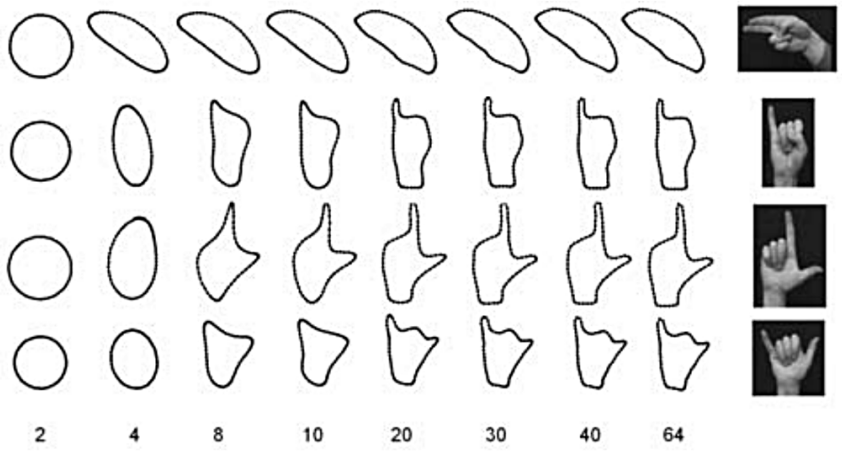


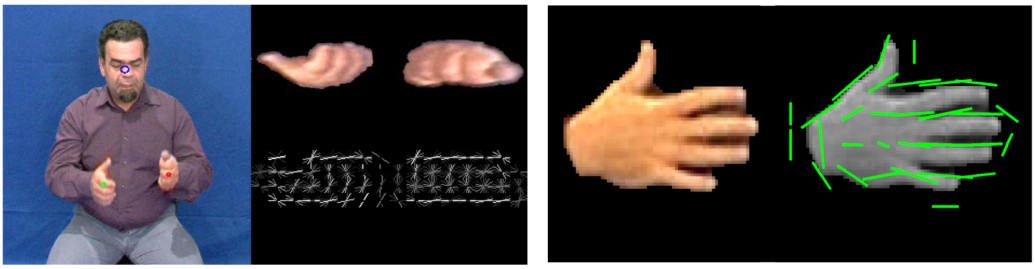
Figura 2.16: Ejemplo de reconstrucción de contornos de manos utilizando descriptores de Fourier. Imagen tomada de [28].

coeficientes de Fourier dado el contorno. Un poco más recientemente, McElroy et al. en [98] utilizaron descriptores de Fourier para reconocer el alfabeto de la lengua de señas de Estados Unidos de América. Al aplicar la transformada de Fourier a una imagen particular del contorno de una mano, se generan una serie de coeficientes, que numéricamente describen la forma que posee el contorno en un dominio frecuencial. La figura 2.16 muestra esto con un ejemplo. Las frecuencias más bajas codifican la curvas más suaves, la idea general del contorno, mientras que las frecuencias más altas codifican los detalles. Una particularidad de estos coeficientes para describir contornos de manos, es que resultan invariantes a la rotación, escala y traslación [81].

SIFT (*Scale-invariant feature transform*)

Un descriptor SIFT es un histograma espacial 3D de los gradientes de una imagen, que caracteriza la apariencia de un punto de interés. Para ello, con el gradiente de cada pixel se calcula un descriptor más elemental formado por la ubicación del pixel y la orientación del gradiente. La técnica fue propuesta por Lowe en 1999 [97] para detectar regiones descriptivas en una imagen y así poder realizar segmentación de objetos, o seguimiento en video. Dado un posible punto de interés, estos descriptores elementales son pesados por la norma del gradiente y acumulados en un histograma 3D que representa el descriptor SIFT de la región alrededor del punto de interés. Al formar el histograma, se le aplica a los descriptores elementales una función de peso gaussiana para darle menos importancia a los gradientes que están más lejos del centro del punto de interés.

Los descriptores SIFT han sido aplicados a varias tareas de visión por computadoras, incluyendo el reconocimiento de configuraciones de manos [168] y reconocimiento de rostros [92].



(a) (b)

Figura 2.17: Extracción de características con Histogramas de Gradientes Orientados (HOG). (a) Ejemplos tomado de [30]. (b) Ejemplo tomado de [22].

Histogramas de gradientes orientados. (*Histogram of Oriented Gradients - HOG*)

En relación con los vectores SIFT, debido a su búsqueda de contexto de formas, los vectores HOGs caracterizan una imagen por la distribución de los gradientes de una serie de regiones. Para esto, este método definido originalmente por Dalal y Triggs en [36], divide la imagen en pequeñas regiones cuadradas llamadas celdas, donde a cada una le calcula las directrices de los gradientes para los pixeles dentro de ella. El resultado es una serie de vectores que indican la magnitud de los gradientes para cada celda. Esto da la posibilidad de describir los diferentes contornos que hay en una imagen. La cantidad de orientaciones que se calculan es un parámetro del algoritmo, siendo lo normal 9 segmentos de orientación, aunque existen trabajos con menos y con más, dependiendo lo que se desee obtener. Para evitar disparidades debido a variaciones en iluminación, los histogramas de las celdas se normalizan en bloques más grandes que contienen varias de ellas.

Buehler et al. [22] utiliza descriptores HOGs para describir cada mano segmentada por separado en un entorno de reconocimiento de lengua de señas. Cada imagen tiene una resolución de $80x80$ pixeles y para calcular los HOGs utilizaron una grilla de $10x10$ celdas con 4 orientaciones. Cooper et al. [30] utiliza estos descriptores con una resolución de $8x8$ celdas de $32x32$ pixeles cada una, con 9 segmentos de orientación. La figura 2.17 grafica estos dos trabajos, con imágenes tomadas de los artículos. Puede verse claramente la diferencia en la cantidad de segmentos de orientación utilizados

Si bien los descriptores HOG pueden encontrarse en numerosos trabajos logrados eficazmente, traen aparejada la desventaja de tener una fuerte dependencia a la traslación y rotación, debido a su naturaleza espacial dentro de la imagen. Esto, generalmente trata de evitarse rotando con anterioridad la imagen, para llevarla a una orientación canónica (como es el ejemplo de la figura 2.17-a). En ocasiones, estos descriptores se utilizan para detectar una objeto conocido, como una persona dentro de una imagen, pero su aplicación para clasificar diferentes configuraciones de manos no es trivial.

Transformada de Radon

La transformada de Radon definida en el espacio $\mathbb{R}^2$ para aplicar a imágenes digitales, se define como una integral de linea sobre la imagen. La idea se basa en integrar diferentes lineas con una ordenada al origen y un ángulo determinado. El resultado es un descriptor con información de frecuencia de puntos para diferentes ángulos. Para reconocimiento de configuraciones, resulta muy útil, sobre todo si se utiliza el contorno de la mano, ya que es posible encontrar rápidamente las zonas donde hay mucha frecuencia de lineas, es decir, donde están los dedos de las manos.

Ha sido muy utilizada en reconstrucción de imágenes, con particular aplicación a reconstrucción de imágenes médicas [108]. Ha sido utilizada también para reconocer objetos y para identificar a personas en base a las características de su mano. Por ejemplo, en [54] se aplica la transformada de Radón al contorno de la mano para realizar un proceso de autenticación biométrica. En [105] se utiliza un enfoque similar, pero se calcula el ángulo óptimo y solo se computa la transformada con ese ángulo, dejando un descriptor en forma de vector, en lugar de ser una matriz. Algo muy similar a estas estrategias se utilizada en esta Tesis para configuraciones de manos en la Lengua de Señas Argentina. Se detallan algunos aspecto sobre este descriptor en el capítulo 4 donde se define el modelo desarrollado, junto con los descriptores utilizados.

Otros descriptores para configuraciones

Claramente, existen diversos trabajos donde se utilizan otros descriptores a los presentados anteriormente. En [115] Ong y Bowden presentan una combinación de detección de manos junto con un clasificador de configuraciones, basado en un clasificador en cascada tipo *boosting*. Los niveles más bajos del árbol de clasificadores permiten separar la imagen de la mano entre varios conjuntos posibles utilizando distancia basada en contexto de formas.

En [161] Yang propone una combinación de cinco descriptores distintos para caracterizar aspectos geométricos y visuales: histogramas de color, momentos Hu, wavlet de Gabor, descriptores de Fourier y descriptores SIFT. El sistema luego es utilizado para reconocer configuraciones de la lengua de señas china.

En [79] Kelly et al. proponen una nueva función de espacio propio (*igenspaces*) aplicada a los momentos HU, luego de segmentar y calcular el contorno de las manos, en un entorno de reconocimiento de lengua de señas irlandesa.

En [135] Roussos et al. proponen un complejo proceso de caracterización de la forma de la mano a través de un novedoso método que los autores proponen como un balance entre “forma” y “apariencia” [33]. El método combina una modificación de los *Active Appearance Models* con un modelado explícito de variación de poses de las manos incorporando transformaciones afines (*Affine transformation*) de las imágenes. El método, luego de realizar una aproximación de la imagen segmentada a los diferentes modelos de configuraciones que posee, realiza una reducción de cardinalidad utilizando Análisis de Componentes Principales (*PCA*).

2.2.3 Descriptores de gestos

Llevar a cabo un proceso completo de clasificación de gestos dinámicos, implica una correcta caracterización de las diferentes partes que componen el gesto. Si los descriptores cumplen con la doble condición de generalizar las clases al mismo tiempo de separar los patrones, el proceso de clasificación resulta mucho más simple. Mientras que algunos autores intentan caracterizar el gesto como un todo, interpretando los fotogramas de forma global, la mayoría de los autores intentan identificar primero las diferentes regiones como rostro y manos para luego describir el comportamiento de los movimientos realizados. Para el caso concreto de la lengua de señas, el análisis de descriptores suele estar enfocado en los 3 componentes principales antes mencionados: posición, configuración de las manos y trayectoria realizada. Algunos autores sólo utilizan la trayectoria. Otros omiten la posición. Aunque en general los mejores resultados se obtienen con una mezcla de los tres componentes. Muchos autores proponen también descriptores complejos que conllevan una cardinalidad muy grande. Algo muy usual en los últimos años es intentar reducir estos valores utilizando Análisis de Componentes Principales (PCA). A continuación se presentan los trabajos más relevantes encontrados sobre caracterización de gestos dinámicos, con particular interés en la lengua de señas. Si bien existen trabajos que utilizan cámaras de profundidad para reconocimiento de lengua de señas, no son los más frecuentes de encontrar ya que no es un dispositivo que suela utilizarse en entornos que se trabaje con esta temática, sumado a que generalmente se caracterizan por sensar el movimiento de todo el cuerpo, cosa que en la lengua de señas carece de sentido. Por esta razón, se dejó la revisión de descriptores para estos dispositivos sólo para el caso de acciones humanas con todo el cuerpo.

Uno de los trabajos más recientes y con diversas publicaciones fue realizado por Helen Cooper, Eng-Jon Ong, Nicolas Pugeault y Richard Bowden, entre otros miembros. En [30], Cooper presenta una revisión de alguno de sus trabajos de años anteriores haciendo un interesante análisis del comportamiento de diversos descriptores para bases de datos de lenguas de señas. La figura 2.18 muestra algunos ejemplos gráficos.

En primer lugar, usa una grilla para discretizar las diferentes posiciones de la mano dentro del video. La grilla se posiciona luego de detectar la cara, dando una distancia relativa a la misma. Para codificar la variación de las configuraciones de manos, utiliza momento Hu [67] junto con tres tipos más de momentos. Estos vectores se calculan por fotograma y luego son concatenados para formar un descriptor global en forma de matriz. Por otro lado, en el mismo trabajo se comparan estos descriptores con la utilización de descriptores HOG para las configuraciones de manos sumado a información de movimiento y posición. Una vez segmentadas las manos, a las diferentes trayectorias se la aplican reglas determinísticas para generar etiquetas de la convención HamNoSys [78] que establece los diferentes movimientos, configuraciones, etc. que puede haber en las lenguas de señas.

En ocasiones, diversas características geométricas de las manos y sus trayectorias

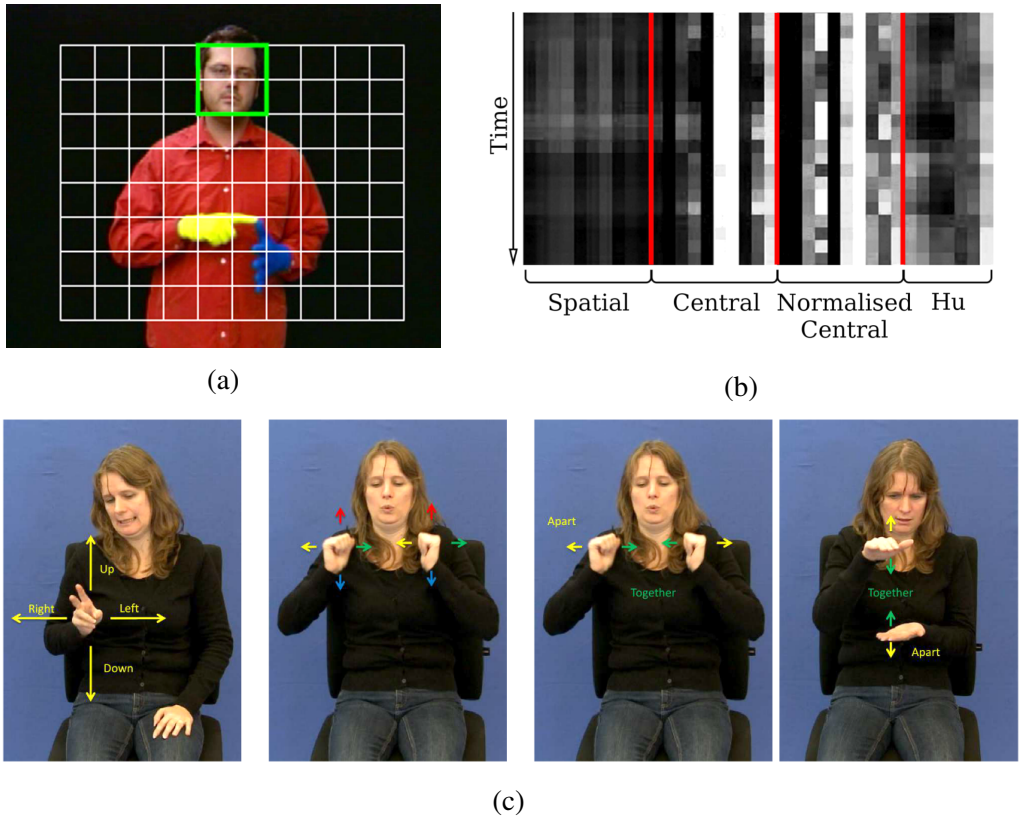
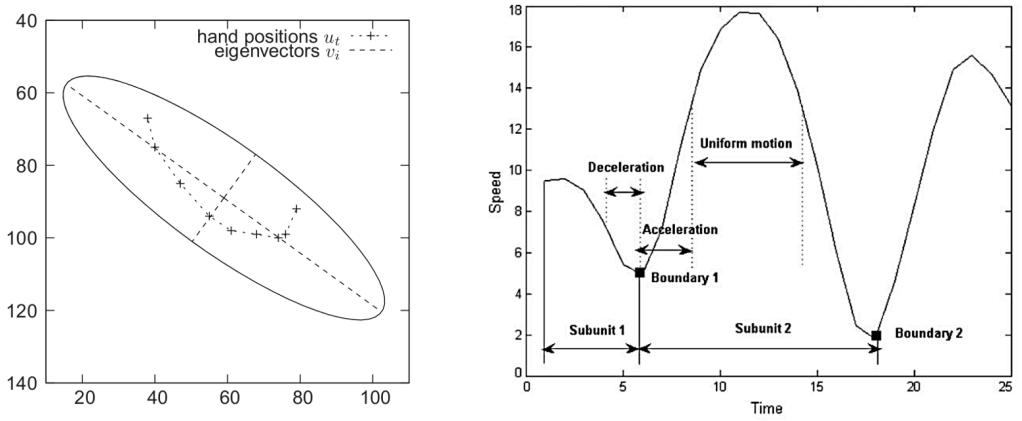


Figura 2.18: Diferentes descriptores para lengua de señas investigados por Cooper et al en [30]. (a) Grilla discreta para identificar posiciones de las manos. (b) momentos HU, espacial, central y central normalizado para describir las formas de manos en el tiempo. (c) Etiquetado discreto de los movimientos de las manos en base a la norma HamNoSys.

son calculadas. Por ejemplo, en [45] Dreuw et al. analiza la interpretación de de diferentes enfoques orientados al reconocimiento de voz aplicados a la lengua de señas. Se utiliza una aproximación mejorada de la definida originalmente por [150]. Aquí se calculan descriptores geométricos de la trayectoria de la mano, como autovalores y autovectores. Luego, combina estos con otros descriptores como las posiciones de las manos, trayectoria y velocidad, agregando contexto temporal a través de una ventana deslizante. En [63] Han et al. utiliza la trayectoria de la mano para segmentar sub-unidades léxicas (fonemas). Busca puntos de discontinuidad en las curvas de velocidad y trayectoria. Luego utiliza Distorsión de Tiempo Dinámico (DTW) como método de interpretación para saber qué bloques no son realmente sub-unidades. La figura 2.19 ilustra estos ejemplos.

Siguiendo con la idea de describir el movimiento de las manos con descriptores de alto nivel, Kadir et al. definen en [74] un descriptor temporal basado en etiquetas determinísticas sobre cuatro grupos de características: posición relativa de las manos



(a) (b)

Figura 2.19: Ejemplos de descriptores de trayectorias de las manos. (a) Autovalores y Autovectores de la trayectoria propuestos en [45]. (b) Puntos de discontinuidad en las curvas de velocidad y trayectoria de la mano para encontrar sub-unidades léxicas en [63].

entre ellas (se define como HA), la posición de las manos en relación a puntos específicos del cuerpo como torso, cabeza, hombres, etc. (se define como TAB), movimiento de las manos como por ejemplo “manos se mueven hacia arriba”, “manos se mueven hacia abajo”, “manos se mueven juntas”, etc. (se define como SIG), y por último la configuración de la mano, previamente clasificada (se define como DEZ). Estos cuatro descriptores “HA-TAB-SIG-DEZ” forma un sólo vector binario, dependiendo si cada ítem existe o no, que se calcula por cada fotograma del video. Lamentablementeno se da demasiado detalle de cuáles son las reglas determinísticas para calcular estos movimientos, o distancias, lo cual hace difícil su reproducción. La figura 2.20 muestra gráficamente este descriptor.

Se mencionó en la sección anterior los descriptores HOG que identifican los diferentes gradientes de orientación en regiones específicas de la imagen. Existen una variedad de estos descriptores llamados HOG3D, definidos en [84], donde se propone calcular estos gradientes no sólo en el espacio bidimensional de una imagen, sino agregando la variable tiempo. Es decir, se interpreta el video como una secuencia de imágenes dando por resultado un descriptor general para un segmento de video determinado. Koller et al. utiliza estos descriptores en [87], acompañados de 7 fotogramas adyacentes como agregado de contexto temporal. Luego se reduce la dimensionalidad utilizando PCA. Forster et al. en [50] utiliza una combinación de cuatro descriptores: el primero es la imagen completa de la mano segmentada. El segundo son los Histogramas de Gradientes Orientados en espacio 3D (HOG3D). Tercero, se calcula la trayectoria de la mano, se normaliza la distancia con respecto a la nariz del intérprete y se calculan lo autovalores y autovectores del movimiento con una ventana temporal deslizante. Cuarto, un perceptrón multicapa (MLP).

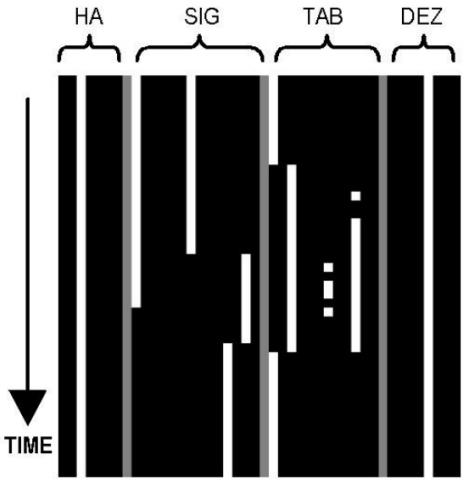


Figura 2.20: Descriptor propuesto por Kadir et al. en [74] para caracterizar una seña.

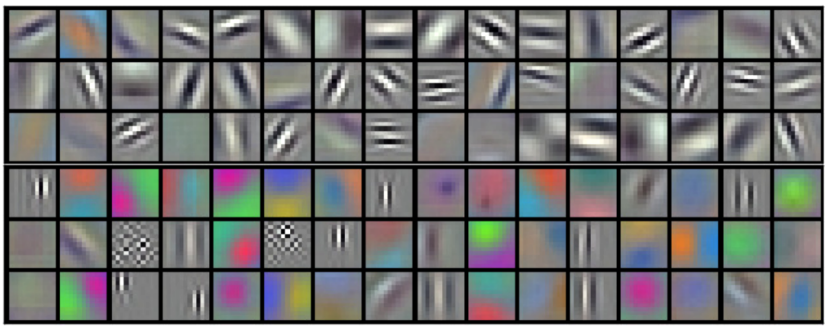


Figura 2.21: Ejemplo de filtros convolucionales generados luego de un proceso de aprendizaje profundo realizado en [89] para clasificar diferentes tipos de imágenes.

Las Redes Neuronales Artificiales son ampliamente conocidas por funcionar como clasificadores lineales o no lineales para una gran variedad de problemas. Existen trabajos donde se utiliza esta herramienta como generador de descriptores, para luego utilizar la salida de la red como entrada para un clasificador general. Por ejemplo, en [59] se alimenta la red con la imagen completa de la mano dominante segmentada. Se reduce la imagen a 32x32 pixeles y se evalúa la utilización tanto del descriptor completo como también una reducción utilizando PCA.

Si bien las redes neuronales han quedado fuera de trabajos de investigación fuerte debido a la creencia de que ya no podían ofrecer mucho más, en los últimos años, se han vuelto a investigar debido a nuevas propuestas que mostraron resultados novedosos. Las nuevas redes neuronales llamadas Redes Neuronales Convolucionales (*Convolutional Neural Networks - CNN*), o de Aprendizaje Profundo (*deep learning*) [93] intentan superar los límites de las tradicional redes de capas ocultas al introducir cientos o miles de neuronas en numerosas capas ocultas, con la intención de que

logren entrenarse y caracterizar problemas complejos. Generalmente estas redes son recurrentes y hasta ahora son utilizadas principalmente para problemas de visión por computador. A diferencia de las estrategias convencional en procesamiento de imágenes, estas nuevas herramientas no utilizan un filtro particular sino que intentan generar diversos filtros nuevos capaces de generar descriptores de alto nivel. Desde este enfoque, un filtro no solo detecta bordes o suaviza una imagen, sino que puede detectar patrones complejos como el bigote de un gato, el faro de un automóvil, o la mano de una persona. La red simplemente se alimenta con las imágenes de entrenamiento, genera una serie de filtros (matrices de convolución) diferentes y luego, las capas más externas finalmente clasifican la imagen. La figura 2.21 muestra un ejemplo de filtros generados bajo un entrenamiento profundo de redes convolucionales en [89]. Generalmente estos algoritmos tienen la contra de necesitar de muchas imágenes de entrenamiento además de un alto costo computacional debido a la cantidad de neuronas que poseen la redes. De hecho, los trabajos más actuales están implementando los algoritmos en GPU para acelerar el cómputo. Existen pocos trabajos de este tipo en la temática de gestos. En [88] Koller et al. utiliza un modelo de CNN con 22 capas ocultas para clasificar diferentes configuraciones de la lengua de señas.

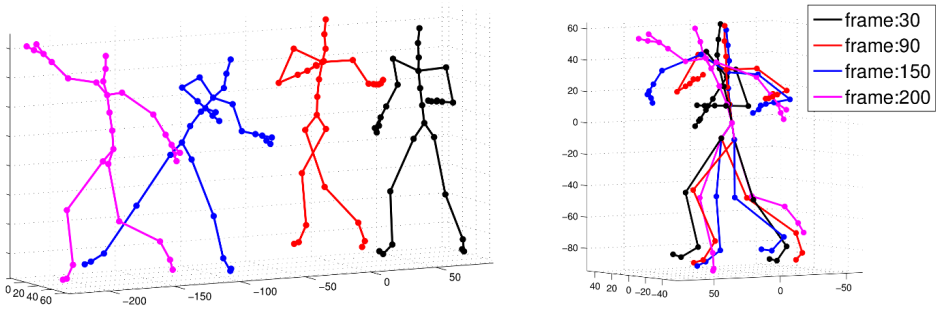
Descriptores para acciones humanas

Para el caso de las acciones humanas, los descriptores encontrados en la literatura suelen ser sutilmente distintos, ya que generalmente se trata de dispositivos que capturan el cuerpo entero, como el MS Kinect, o previamente se aproximan las articulaciones con algoritmos de procesamiento de imágenes (ver [62]). Desde esta perspectiva, generalmente no resulta importante la forma específica de la mano o rasgos faciales, sino que el foco está puesto en entender como se desplazan las diferentes articulaciones corporales en el tiempo.

En este contexto, existen numerosos descriptores desarrollados en los últimos años, principalmente para bases de datos capturadas con el MS Kinect. Algunos tienen en cuenta las articulaciones que el dispositivo calcula, otros el mapa de profundidad, o una combinación de la imagen RGB-d con la imagen RGB convencional. Ye [163] ofrece una revisión interesante y bastante completa sobre estos descriptores.

Diversos autores utilizan sólo la información de profundidad para generar un descriptor apropiado. La figura 2.22-a muestra un ejemplo típico de una acción capturada con una cámara de profundidad. Generalmente, para este tipo de aproximaciones, la extracciones de descriptores apropiados es la pieza clave en el proceso de clasificación. Los mapas o imágenes de profundidad no tienen algunos inconvenientes de que poseen las imágenes convencionales como diferencias en iluminación, pero sin embargo presentan otros, como por ejemplo grandes oclusiones que hacen difícil generar un descriptor global [163]. Esto llevó a los investigadores a desarrollar descriptores semi-locales y robustos a las oclusiones. La mayoría de los métodos que utilizan mapas de profundidad se basan en descriptores de volúmenes espacio-temporales.

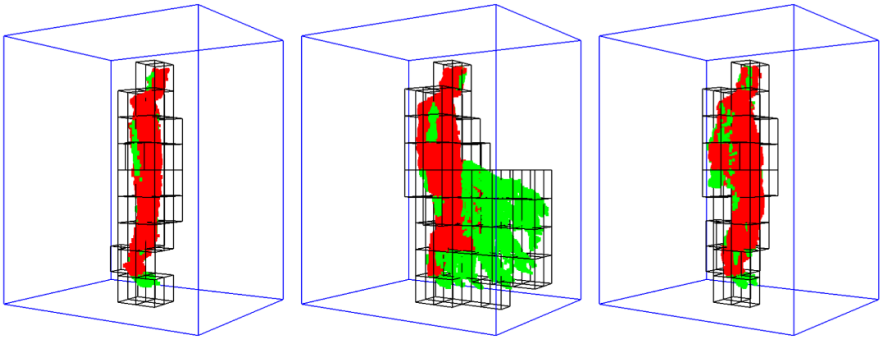
(a)



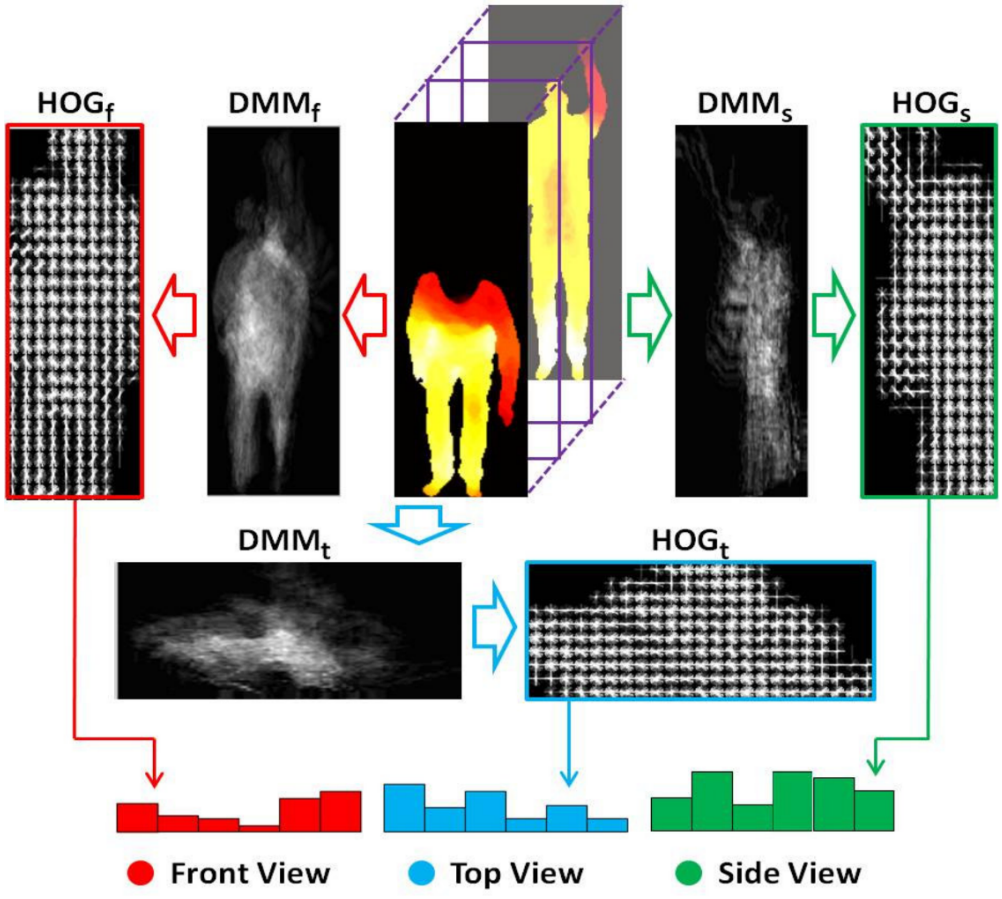
(b)

Figura 2.22: Ejemplos típicos de acciones humanas capturadas con un dispositivo tipo MS Kinect. (a) Ejemplo imagen de profundidad del gesto “saque de tenis”. (b) Ejemplo de esqueleto humano calculado. Los puntos indican las diferentes articulaciones evaluadas. A la derecha, una normalización de orientación propuesta en [27].

Vieira et al. [146] proponen un nuevo descriptor llamado “Space-Time Occupancy Pattens (STOP)”. En el trabajo, la secuencia de profundidad es representada como celdas en cuatro dimensiones para darle contexto temporal a la información 3D. Luego, se utiliza un esquema de saturación para mejorar el rol de las celdas dispersas que típicamente consisten en puntos en las siluetas o partes móviles del cuerpo. la secuencia es dividida en segmentos, donde cada segmentos tiene una cantidad fija de fotogramas. Yang et al. [162] desarrolló un descriptor llamado “Depth Motion Maps” (DMM) que captura la energía del movimiento en las acciones. La imagen de profundidad es proyectada en tres planos cartesianos ortogonales definidos previamente y luego normalizados. Para cada imagen proyectada, un mapa binario es generado calculando la diferencia entre dos fotogramas consecutivos. Luego, los mapas binarios son sumados para obtener el DMM para cada vista. Luego, se aplican descriptores HOG para caracterizar cada vista, y estos son concatenados para formas un descriptor que los autores denominan “DMM-HOG”. La figura 2.23 grafica estos últimos dos ejemplos mencionados.



(a)



(b)

Figura 2.23: Ejemplos sobre descriptores encontrados en la bibliografía basados en imágenes de profundidad. (a) Descriptor STOP propuesto en [146] aplicado a una acción concreta donde una persona ejecuta un golpe con la pierna. (b) Descriptores DMM-HOG desarrollados en [162].

El otro gran grupo de trabajos existentes sobre descriptores para acciones es el basado en el esqueleto humano. La figura 2.22-b muestra un ejemplo típico de cómo dispositivos tipo MS Kinect capturan las diferentes articulaciones del cuerpo humano. Aquí se posee una serie de información sobre la ubicación espacial en el tiempo de estas articulaciones. A diferencia de los trabajos basados en mapas de profundidad, aquí se suele modelar la información temporal de forma explícita.

En [69] Hussein et al. realiza un trabajo basado específicamente en la extracción de descriptores para el caso de acciones humanas en base al esqueleto del cuerpo. Aquí se propone un descriptor basado en matrices de covarianzas de las ubicaciones de lasa articulaciones sobre el tiempo. Para codificar la relación entre los movimientos de las articulaciones y el tiempo, se estructuran múltiples matrices para las sub-secuencias en forma jerárquica.

Otro trabajo muy relevante es el de Olfi et al. en [114] donde se propone un descriptor llamado *Sequence of the Most Informative Joints (SMIJ)* (Secuancia de las articulaciones más relevantes). Criticando otros artículos en base a la utilización de descriptores que no siempre tienen mucho sentido físico con diversas acciones, en este trabajo se intenta identificar qué articulaciones son más relevantes en cada acción particular. Esta selección está basado en características bien observables como la media de las varianzas de los ángulos, la velocidad angular máxima, etc.

Por último, al igual que para los enfoques basados en visión, las redes neuronales profundas están comenzando a investigarse para extraer nuevos descriptores en este tipo de enfoques. En [27] se propone un esquema donde los descriptores son extraídos de cada fotograma en base a tres componentes: las posiciones relativas de cada articulación, las diferencias temporales de las articulaciones, y la trayectoria del movimiento normalizada. Dados estos descriptores, se entrena un perceptrón multicapa híbrido que al mismo tiempo que clasifica reconstruye los datos de entrada, aprendiendo nuevos descriptores. En el trabajo se compara la utilización de *autoencoders* profundos con el clásico PCA, mostrando que los primeros logran capturar mayor información.

2.3 Técnicas de clasificación de gestos

El objetivo principal de un sistema de reconocimiento de gestos, es la interpretación semántica de lo que la mano o distintas articulaciones del cuerpo quieren decir. Luego de generados los descriptores, entonces, el próximo paso en una sistema de de este tipo es la clasificación de los datos. Dependiendo el tipo de gesto que se quiera clasificar será el tipo de clasificador. Diferentes autores han realizado distintas taxonomías en la literatura. Generalmente suele dividirse en estrategias para gestos estáticos, y dinámicos [129, 120]. Generalmente, los gestos estáticos pueden clasificarse con ciertos métodos generales de clasificación lineal o no lineal [129], sin embargo los gestos dinámicos poseen un aspecto temporal que suele necesitar de un clasificador capaz de lidiar con este aspecto.

En [118] se proponen dos esquemas posibles de clasificación de gestos dinámicos: uno donde se realiza una clasificación en una sola etapa, interpretando todo el gesto en un solo descriptor; otro, donde el gesto es descompuesto en diferentes partes clasificadas de forma independiente, para luego ser unificadas en un clasificador a nivel de gesto. Esta último enfoque es el presentado en esta Tesis. Por esta razón, es necesario analizar, como se hizo con los descriptores, los diferentes métodos existentes para configuraciones de manos, trayectoria, etc. En esta sección se resumen las diferentes estrategias existentes para clasificación de gestos estáticos, dinámicos y acciones humanas, con foco puesto en la lengua de señas.

2.3.1 Clasificación de configuraciones de manos

Si bien la clasificación de configuraciones de manos es un desafío muy grande aún en visión por computador, generalmente el foco del proceso está puesto en la correcta generación de descriptores. De este modo, las estrategias utilizadas para el reconocimiento de posturas de manos suelen no ser demasiado complejas, dependiendo de los descriptores obtenidos. Generalmente, los métodos utilizados suelen clasificarse en métodos supervisados, no supervisados, entre otros [120]. También existen taxonomías donde se distinguen en métodos de clasificación lineal y no lineal [125]. A continuación se listan algunos de los métodos más utilizados para reconocimiento de gestos estáticos, particularmente de configuraciones de las lenguas de señas.

Redes Neuronales Artificiales

Las redes neuronales artificial han mostrado ser clasificadores robustos en numerosas aplicaciones [20, 65]. La idea principal de estas redes es la de aproximar una función que relacione los patrones de entrada, con una salida esperada, a través de un algoritmo supervisado. Este algoritmo tradicionalmente se conoce con el nombre de *backpropagation* o propagación hacia atrás por medio del descenso de gradiente en la función de error. La red se organiza en una capa de entrada, donde se alimenta con los descriptores del gesto, una capa de salida con la respuesta de la red, y generalmente existe una capa oculta con diversas neuronas, que permiten un mapeo no lineal de los patrones. Particularmente para posturas de manos han sido muy utilizadas en la década de los 90, donde comenzaron a emerger los estudios de reconocimiento de lengua de señas.

Una de las decisiones a tomar al utilizar una red neuronal es cómo organizar la capa de entrada y la de salida. Un enfoque clásico es tener tantas neuronas de salida como clases esperadas. Entonces, cada neurona de salida responde con un valor binario indicando si el patrón de entrada pertenece o no a la clase. Esto trae aparejados los problemas de poder existir más de una neurona que responda positivamente al mismo tiempo, y por otro lado, la gran cantidad de neuronas necesarias a medida que la cantidad de clases aumenta. En [154] se utiliza una red neuronal para reconocer 36 configuraciones distintas, utilizando una red neuronal con una capa de salida con una

neurona binaria por cada clase. Se analiza la posibilidad de codificar la salida con menos neuronas, pero los autores discuten que esta idea reduce la eficacia del modelo. en [46] se utiliza un enfoque similar para clasificar configuraciones de la lengua de señas arábica. Más recientemente en [15] se propone un clasificador neuronal basado en dos etapas para clasificar configuraciones de la lengua de señas brasilera. La primer etapa recibe patrones basados en descriptores HOGs y discrimina entre 12 diferentes grupos. Luego, al identificar el grupo de pertenencia, una nueva red encargada de reclasificar ese grupo identifica alguna de las 3 o 4 clases que conoce.

Máquinas de Soporte Vectorial

Las máquina de soporte vectorial (*Support Vector Machine - SVM*) son otro de los métodos supervisados de clasificación no lineal más ampliamente utilizados y con resultados exitosos en diversas aplicaciones [65]. La idea principal es la de construir un hiper-plano capaz de separar dos clases donde el margen de separación sea el máximo posible a través de los llamados vectores soporte. Es decir, desde su concepción, las SVM son separadores lineales. No obstante, logran clasificar problemas no lineal extendiendo el espacio del dominio a otro de dimensionalidad mucho más grande. Las SVM fueron originalmente pensadas para clasificación binaria, aunque también son utilizadas adecuadamente para problemas multiclase con algunas adecuaciones. Uno de los cuellos de botella más grande en las SVM es la gran cantidad de vectores soporte necesarios para su correcto funcionamiento. Pueden encontrarse ejemplos donde se utilizan SVM para clasificación de formas de manos en [145], [165], [161], [79] y [121].

Algoritmos de agrupamiento

Los algoritmos de agrupamiento (o *clustering*) son métodos no supervisados, donde una serie de patrones intentan agruparse por característica comunes. Lo más habitual es utilizar alguna medida de distancia aunque también existen enfoques más sofisticados. Como algoritmo no supervizado, resulta interesante su aplicación en situaciones donde no se está seguro de la distribución existente que poseen los datos. En ocasiones se utiliza un agrupamiento como etapa inicial de distirbución, para luego realizar un proceso de clasificación sobre los grupos.

Quizá uno de los más populares algoritmos de agrupamiento existentes es el k-medias (*k-means*). En este método se genera una cantidad inicial estipulada de centros (grupos) y se los inicializa en posiciones aleatorias. Luego, se utiliza la distancia euclídea como medida de similitud, y de forma iterativa se analizan los diferentes patrones moviendo de lugar los centros hasta encontrar una posición estable. Existen diversas estrategias de finalización. Una de las más comunes es cuando ninguna asignación de patrones a los diferentes grupos cambia. Un ejemplo de uso de k-medias en la clasificación de configuraciones de manos puede encontrarse en [135].

Otro algoritmo de agrupamiento ampliamente utilizado son los Mapas Auto-organizados de Kohonen (*Self-Organization Map - SOM*) [65]. Este método de

agrupamiento fue presentado por T. Kohonen en 1982, inspirado en ciertas evidencias de la corteza cerebral sobre la disposición y funcionamiento de las redes neuronales. Es un algoritmo no supervisado competitivo. Existe una cantidad inicial de centros (también llamados neuronas) que compiten entre sí, distribuidos aleatoriamente. Luego, mediante un proceso iterativo cada patrón es asignado a una neurona ganadora, generalmente basado en una función de distancia. Las diferentes neuronas suelen conectarse a través de una topología de grilla. Cuando una neurona resulta ganadora, se acerca al patrón cambiando su ubicación y desplazando también las neuronas que están conectadas con una proximidad establecida. En [58] se utiliza una red SOM para estimación de posturas de manos. En [139] se utiliza una red neuronal llamada "Self-Growing and Self-Organized Neural Gas" (SGONG) que funciona de un modo muy similar a las redes SOM pero comienza con pocos centros y mediante un proceso iterativo se introducen nuevos conectados mediante una malla. La red se usa aquí para ajustar mediante la misma la forma de la mano.

ProbSOM

El ProbSOM es una adaptación estadística de las redes SOM, definido en [49] como un modelo de clasificación para reconocimiento de voz humana, aunque fue utilizado también para aplicaciones de visión por computador [92]. Se basa en una estrategia en dos etapas. En primer lugar se realiza un proceso de agrupamiento de los datos a través de una red SOM. Luego, se analiza la distribución de los patrones de entrenamiento en las distintas neuronas, y se crea una tabla con la cantidad de elementos de cada clase en cada neurona competitiva. Al querer evaluar un nuevo elemento se calcula la distribución de probabilidades para cada clase conocida en base a la distribución de las neuronas para todos los patrones pertenecientes al elemento en cuestión. Cabe destacar, que dado que el método funciona de forma estadística, cuanto más patrones posea un elemento, mayor será eficacia. En el ejemplo original de la voz humana, cada muestra de audio contenía diversos descriptores representados por coeficientes frecuenciales del sonido. En el caso de [92] se utilizaron vectores SIFT para clasificar diferentes rostros de personas. Cada rostro posee una cantidad arbitraria de vectores, pero siempre un número considerable. Un factor importante aquí es la correcta especificación del tamaño de la red. Una red muy chica podría contener pocas neuronas, lo que no sería suficiente para discriminar los patrones. Por otro lado, una red muy grande genera un sobre-entrenamiento lo cual es perjudicial al momento de incorporar elemento nuevos al modelo.

Otros enfoques

Claramente existe una diversidad de otros enfoques encontrados en la literatura, dependiendo de los descriptores computados para la forma de la mano, y de si estos descriptores fueron tomados solo de imágenes o también de secuencias de video. En [30] Cooper et al. utiliza un enfoque tipo Bosque Aleatorio (*multi-class Random Forest*) para clasificar las diferentes configuraciones del léxico junto con diversas variaciones que la forma de la mano puede tener.

En [144] Thangali et al. realiza un interesante trabajo donde se analizan las configuraciones iniciales y finales de las diferentes señas de la base de datos utilizada. No se detalla el tipo de descriptor obtenido. El enfoque del trabajo está en su propuesta de una nueva red bayesiana que llama *Handshapes Bayesian Network (HSBN)* que computa las probabilidades de que un par de configuraciones como incial y final puedan ser de cada seña en el léxico. Incluso proponen un algoritmo de aproximación debido a las variaciones que puede existir en las configuraciones al ejecutarlas en cada seña. Esto resulta sumamente interesante sobre todo considerando un léxico de 1500 señas como se evalúa en el trabajo. El trabajo es casi el único encontrado en la literatura que aborda el dinamismo de las formas de las manos en una seña. Sin embargo el enfoque planteado requiere de numerosas etiquetas marcadas de forma manual por un experto sobre como comienza y termina una seña, y las diferentes variaciones que las configuraciones pueden tener.

2.3.2 Clasificación de gestos dinámicos

Como se vio en secciones anteriores, afrontar un proceso de clasificación de gestos dinámicos implica interpretar correctamente no sólo la información estática de ciertas partes del cuerpo, sino también información dinámica sobre las diferentes trayectorias así como también cambios en expresiones o posturas de manos. Algunos trabajos, como el presentado en esta Tesis, intentan describir la mayor parte de la información temporal en los descriptores del gesto, para luego utilizar un clasificador que no tenga que procesar esta información de forma directa. No obstante muchos trabajos abordan el dinamismo directamente desde un clasificador, intentando tener algún tipo de máquina de estados, o sistema de memoria para describir los diferentes sucesos que ocurren. A continuación se resumen algunos de los métodos más utilizados para clasificación de gestos, particularmente para lengua de señas.

Modelos Ocultos de Markov

Quizá el método más utilizado para clasificación de gestos dinámicos, sean los Modelos Ocultos de Markov (*Hidden Markov Models - HMM*) [128]. Introducidos a mediados de la década de 1960, al igual que para otros sistemas utilizados para reconocimientos de gestos, fueron utilizados por primera vez para reconocimiento del habla.

Los Modelos Ocultos de Markov son modelos gráficos probabilísticos que resultan adecuados para caracterizar los gestos humanos gracias a su naturaleza temporal. Como fue mostrado en problemas de reconocimiento del habla, son flexibles a cambios en velocidad. Como modelo generativo, se puede entender un HMM como una máquina de estados finita que ejecuta una transición por cada instante discreto de tiempo, cambiando de un estado a otro. Al entrar a un nuevo estado, un vector de observación (o símbolos) se genera de acuerdo a una distribución de probabilidad asociada al estado. Las transiciones entre estados son modeladas también de for-

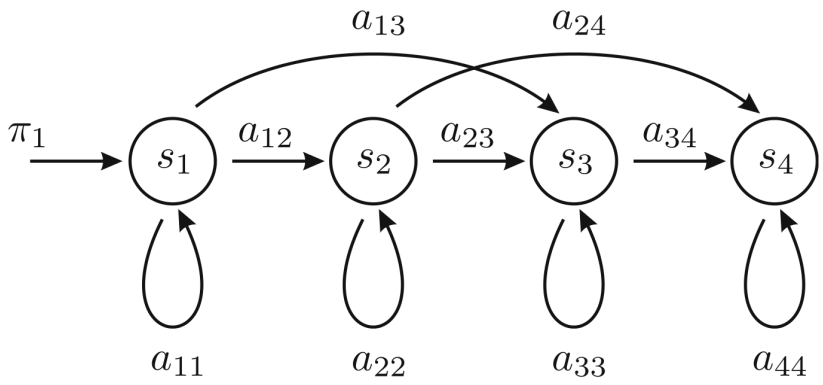


Figura 2.24: Modelo oculto de Markov Bakis.

ma probabilística. Ya que solo la salida del estado es observable, y no el estado propiamente dicho, se dice que la estructura interna es oculta.

Formalmente, un HMM está definido como un conjunto de estados, un conjunto de símbolos de salida, y un conjunto de transiciones entre estados. Cada transición está representada por un estado de salida, un estado de llegada, y una probabilidad de que ocurra el evento. En el contexto de reconocimiento de gestos dinámicos, cada estado podría representar por ejemplo una posición posible de la mano del sujeto, las transiciones podrían representar el cambio de estas posiciones, las observaciones podrían representar una captura específica de la posición mano en el video. Una secuencia de símbolos de salida representaría un gesto. Un HMM puede describirse completamente por los parámetros $k = (A, B, \Pi)$, donde $A = a_{ij}$ representa la matriz de probabilidades de todas las transiciones posibles entre estados; $B = b_j(o_t)$ define la distribución de probabilidades de los símbolos de salida, donde $b_j(o_t)$ representa la probabilidad de generar el vector de observaciones o_t en el instante t, al entrar al estado s_j; $\Pi = \pi_i$, denota la distribución de estados iniciales, donde π_i representa la probabilidad de comenzar en el estado s_i.

Uno de los problemas existentes al utilizar HMM, además de la gran cantidad de datos de entrenamientos requeridos, es el alto costo computacional que poseen si se los entiende como un modelo gráfico probabilístico. En la práctica, existen diversas asunciones que modelan un HMM, y ayudan a reducir este costo [153]. Por ejemplo, la probabilidad de ejecutar una transición hacia un estado sólo depende del estado previo y no de la secuencia entera de estados. Otro ejemplo, llamado "salida independiente", establece que la probabilidad de generar un vector de símbolos de salida sólo depende del estado actual y es estadísticamente independiente de las observaciones previas. Por otro lado, si bien la definición de los HMM permite una gran variedad de modelos, existen ciertos modelos estructurados de cierto modo que han mostrado ser eficaces para diferentes tareas. Uno de los más utilizados en el área de reconocimiento del habla, y de gestos, es el modelo Bakis (figura 2.24). Este modelo puede compensar las variaciones temporales de un gesto con transiciones

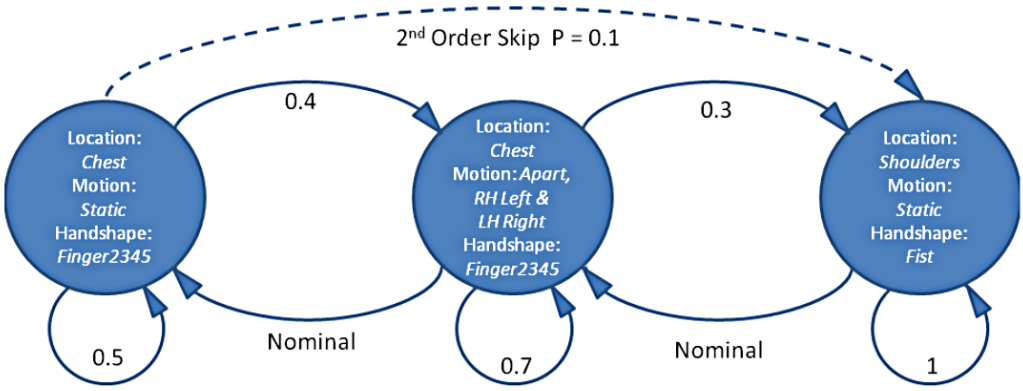


Figura 2.25: Ejemplo de cadena de Markov utilizada en [31] para caracterizar una seña particular de la lengua de señas griega.

que pueden ir al siguiente estado, al mismo, o saltearse un estado entero.

Generalmente en los trabajos se encuentran modelos de Markov tanto con estados ocultos como cadenas de Markov convencionales, dependiendo el modo en que se utilice, o los descriptores calculados. Las cadenas de Markov no poseen estados ocultos, sino que cada estado representa directamente un patrón de salida. En [31] Cooper et al. utilizan cadenas de Markov para modelar las diferentes señas del léxico griego. Puede verse un ejemplo de este trabajo en la figura 2.25. Existen numerosos trabajos existentes que utilizan estos modelos, o variantes de los mismos, para reconocimiento de gestos y lenguas de señas , generalmente bajo un enfoque que utiliza un HMM por cada gesto a reconocer, por ejemplo [59], [11], [99], [127]. Actualmente, muchos autores entrenan los HMM para reconocer sub-unidades (fonemas), como se mencionó en secciones anteriores, en vez de un gesto entero. Pueden encontrarse ejemplos de este enfoque en [122], [16] o [153].

Alineamiento Dinámico del Tiempo (Dynamic Time Warping - DTW)

El Alineamiento Dinámico del Tiempo, o Distorsión Dinámica del Tiempo, más conocido en la literatura como DTW por sus siglas en inglés, es una técnica que tiene como objetivo medir la similitud entre dos señales temporalmente dependientes [103]. La idea aquí es encontrar una medida de similitud al alinear dos señales o series temporales aceptando pequeñas deformaciones de la señal en el dominio tiempo. Generalmente se construye una matriz de distancia y se busca el camino de costo mínimo utilizando programación dinámica. Una particularidad interesante del método para su aplicación en reconocimiento de gestos, es su adaptabilidad a cambios en velocidad de las dos series. Al igual que los HMM fue una técnica muy utilizada para reconocimiento del habla.

En lo referente a gestos dinámicos, esta técnica puede utilizarse por ejemplo para alinear las trayectorias de las manos con un modelo de referencia, entendiendo la

trayectoria como una serie temporal. Sin embargo resulta difícil su utilización con descriptores de imágenes al querer clasificar las diferentes formas que puede tener la mano en un entorno de lengua de señas. Por esta razón, generalmente se utiliza esta estrategia en sistemas con sensores como acelerómetros, u otros tipos de guantes de datos. En [14] Barczewska y Drozd comparan diferentes variantes del DTW como son *Piecewise Dynamic Time Warping* (PDTW) y *Derivative Dynamic Time Warping* (DDTW) para la clasificación de gestos hechos con las manos, capturados con un sensor de movimientos 3D puesto en un dedo. En [68] se puede encontrar otro ejemplo de utilización de esta estrategia con sensores en las manos. En [155] se utiliza una variante de DTW llamada Dynamic Space-Time Warping (DSTW) con el objetivo de analizar el espacio donde se mueve la mano en un entorno de reconocimiento de lengua de señas. En [34] Corrandini propone un enfoque basado en una plantilla que utiliza DTW para alinear y normalizar las señales provenientes del movimiento del brazo, aplicando una transformación temporal específica. En [90] se utiliza esta técnica para clasificar formas de manos provenientes de imágenes alineando el contorno de las mismas, previa segmentación.

Redes neuronales artificiales

Las redes neuronales (definidas previamente) fueron utilizadas mayormente a comienzos de la investigación en reconocimiento de gestos/señas, alrededor de la década de 1990. Actualmente, casi ningún trabajo medianamente importante está usando ya estas técnicas para estas disciplinas. Sin embargo, merece la pena rescatar esta información debido a que siguen siendo un clasificador robusto si se cuenta con un descriptor apropiado.

Algo a tener en cuenta en este caso, es que las redes tipo "Perceptrón Multi-Capa" [65], o también llamadas *feedforward* o *backpropagation* (por su algoritmo de entrenamiento), no suelen manejar correctamente la información temporal. En este caso, sería coherente mapear esta información de algún modo en un descriptor, que luego sea discriminable por una red. Una variante de estas redes son las redes neuronales recurrentes, que poseen conexiones entre neuronas de una misma capa, o de una capa a otra anterior. Esto establece algún tipo de memoria, aunque los algoritmos de aprendizaje supervisado suelen ser más complejos.

Murakami y Taguchi [107] publicaron en 1991 uno de los primeros trabajos existentes en clasificación de lengua de señas. Aquí se utilizó una red neuronal recurrente para clasificar diferentes gestos del léxico japonés, mostrando buenos resultados con un pequeño vocabulario e incluso siendo independiente a los intérpretes.

Waldron y Kim presentaron en [154] una estructura muy particular para reconocimiento de lengua de señas. En primer lugar utilizaron cuatro redes neuronales convencionales tipo *feedforward* para reconocer las diferentes partes de la seña: una para las 36 configuraciones existentes en la base de datos evaluada; otra red para las 10 posiciones estipuladas; otra para las 11 orientaciones; y una última para los 11 tipos de movimientos existentes. Luego, los fonemas reconocidos del comienzo,

medio y fin, fueron utilizados para alimentar una segunda etapa de clasificación, donde se identificaba el gesto. Para esto, se compararon dos enfoques, uno que utiliza una red neuronal *feedforward* y otro con una red auto-organizativa SOM. Con la primera lograron una precisión del 86 % para las 14 señas a reconocer.

Kim et al. propone en [82] una red neuronal "Fuzzy Min Max" para reconocimiento continuo de 25 gestos de la lengua de señas koreana utilizando guantes de datos para sensar los movimientos. El trabajo muestra una tasa de acierto del 85 %, lo que era considerable para el año 1996.

Sistemas de clasificación completos

Como se mencionó a comienzos del capítulo, los sistemas de reconocimiento de gestos, y particularmente de lengua de señas, necesitan considerar diversos aspectos para llevar a cabo una correcta clasificación. Por esta razón, un sistema completo de reconocimiento rara vez sólo posee un método de clasificación basado en las técnicas antes mencionadas. Generalmente, diversos modelos de clasificación abordan los distintos problemas, como reconocimiento de las formas de manos, encontrar el rostro, clasificar los diferentes movimientos, etc. A continuación se mencionan algunos trabajos donde se realizan procesos completos de clasificación de lengua de señas, utilizando sistemas robustos con diferentes etapas o sub-clasificadores.

En [80], Kelly et al. propone un completo trabajo de reconocimiento de lengua de señas irlandesa. El principal interés es realizar un segmentado del flujo de video en tiempo continuo (*spotting*), donde se detecta en qué momento termina una seña y comienza otra, problema sumamente complejo en esta disciplina. Utilizan un sistema débilmente supervisado. Para reconocer las formas de las manos utilizan un algoritmo de agrupamiento sobre los descriptores procesados, para luego clasificar los conjuntos detectados con una Máquina de Soporte Vectorial. Para clasificar la información temporal del movimiento de las manos, utilizan un HMM. Luego combinan los dos clasificadores como un producto de probabilidades.

En [87], Koller et al. utiliza un sistema estadístico utilizando la plataforma RASR de la Universidad RWTH Aachen, que originalmente fue desarrollado para procesamiento de voz humana. El sistema utiliza un modelo visual, basado en diversos HMM que clasifican la información temporal, y un modelo de lenguaje utilizando n-gramas que agrega información gramatical de la lengua de señas a reconocer. Este modelo es aprendido desde un texto fuente.

Cooper et al. propone en [30] una comparación de diferentes modelos. Para reconocimiento de formas de manos utiliza un clasificador multi-modal basado en *Random Forest* con 100 árboles de clasificación, luego de calcular descriptores HOG sobre las manos segmentadas. Cada árbol es entrenado con un subconjunto aleatorio de los datos de entrenamiento. Para la clasificación de los descriptores de movimiento computados, compara la utilización de HMM con *Squential Pattern Boosting*. Se explica que uno de los problemas existentes al utilizar HMM es la dependencia del modelo a una serie exacta de transición sobre los descriptores. Los

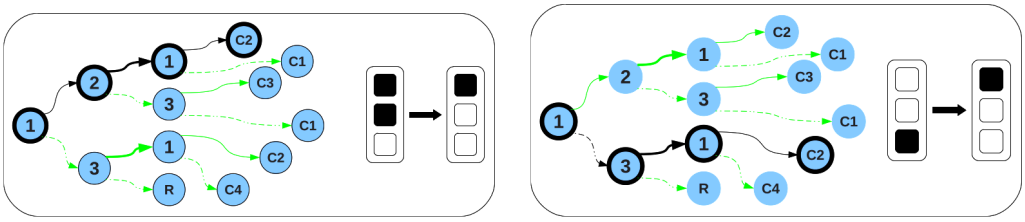


Figura 2.26: Imagen tomada de [116] donde se definen los *Sequential Pattern Trees*. Ejemplo de dos posibles caminos para la misma clase.

Sequential Patterns (SP) buscan los descriptores relevantes, ignorando el resto. En el trabajo se define un método para elegir un posible vector candidato y diversos clasificadores débiles tipo SP son definidos. Luego mediante un proceso de *Boosting* en cascada definido en [115] se encuentra el SP que mejor resultado otorga. En [116] Los mismos autores definen un enfoque similar llamado árboles de patrones secuenciales (*Sequential Pattern Trees*). En estos árboles, cada nodo es selector de descriptor, con excepción de los nodos hojas que representan la clase a reconocer (figura 2.26). Los autores evaluan el modelo propuesto en dos bases de dados de lenguas de señas, una capturada con cámaras RGB y otra bajo el dispositivo MS Kinect.

Por último, con respecto a modelos de reconocimiento para acciones humanas, los métodos generalmente se dividen de acuerdo al descriptor computado [6]. Cuando se calculan descriptores globales, basados en una descripción del espacio-tiempo, suelen usarse modelos discriminativos tipo SVM. En el caso en que se utilice un enfoque secuencial, generalmente se calculan descriptores sobre cada fotograma o diferentes instantes de tiempo que modelen el dinamismo del gesto. En este caso, usualmente se utilizan modelos estadísticos tipo HMM, que modelan el dinamismo de manera explícita. En [163] se ofrece una revisión de diferentes estrategias, donde se analiza SVM y HMM como las dos principales encontradas en la literatura.

2.4 Bases de datos existentes para reconocimiento de gestos

A diferencia de las bases de datos para minería de datos o algoritmos de aprendizaje supervisado en general, en donde diferentes temáticas son representadas con descriptores genéricos, los sistemas de clasificación de gestos dinámicos suelen necesitar una base de datos robusta que contenga las clases que se quieren reconocer. Estas bases de datos varían dependiendo la aplicación a desarrollar. Existen bases de datos sólo de imágenes (para gestos estáticos), de secuencias de videos, videos con gestos segmentados, videos con varios gestos continuos, etc. Cada base de datos suele estar construida para un propósito particular y debe ser abordada considerando el modo en que fue capturada. No obstante las diferencias, toda base de datos necesita tener diversos ejemplos que caractericen cada clase a reconocer, para ser utilizada

en un proceso de aprendizaje automático. Esto generalmente incluye la ejecución de los gestos realizados por diferentes sujetos, para evaluar el comportamiento de los métodos al momento de introducir un nuevo sujeto al sistema. Particularmente para el reconocimiento de lengua de señas, existe el problema de la poca usabilidad de cada base de datos en aplicaciones reales debido a los diferentes léxicos de cada región. Esto implica la necesidad de crear nuevas bases específicas para cada lengua.

En esta sección se presenta una revisión de las bases de datos existentes al momento de realizar esta Tesis para reconocimiento de gestos, separando en tres grandes grupos: bases de datos de formas de manos, de lengua de señas, y de acciones humanas. Claramente, existen otras bases de datos disponibles para reconocimiento de gestos con las manos que no encajan específicamente dentro de estas categorías. Se decidió no hacer referencia a ellas, debido a que no son parte del estudio realizado en este trabajo. No obstante, en [120] se encuentra una revisión de algunas de dichos conjuntos de datos.

2.4.1 Bases de datos de gestos estáticos

Hace unas décadas atrás, había un gran interés por nuevas investigaciones referentes a procesamiento de imágenes. En esos momentos se realizaron numerosas bases de datos para clasificación de imágenes, incluyendo gestos con las manos. Actualmente, el interés es mayor por los gestos dinámicos. Esto hace que la construcción de bases de datos de imágenes sea cada vez menor. Con respecto a sistemas de clasificación de lengua de señas, lo más usual es realizar bases de datos dinámicas, donde los investigadores toman la información estática de los videos. No obstante, existen algunas bases de datos citadas en la bibliografía que merece la pena mencionar.

ASL Finger Spelling Dataset. Quizá una de las más actuales para clasificación de gestos estáticos manuales sea la realizada por Pugeault y Bowden en 2011 [126]. Como su nombre lo indica, la idea aquí fue organizar una serie de imágenes que representan el alfabeto dactilológico de la lengua de señas norteamericana, con excepción de la letra Z que requiere de un comportamiento dinámico. La base de datos fue capturada con un dispositivo MS Kinect en formato video, para luego ser separado en imágenes de cada fotograma. Durante el video cada sujeto mostraba la forma de la mano haciendo leves movimientos para dar variabilidad. De este modo, los autores proponen una base de datos con 24 gestos estáticos, almacenando tanto la imagen RGB como los mapas de profundidad. La base fue grabada por 5 individuos en su primera versión y luego por 9, con un total de más de 60.000 imágenes. La base de datos es públicamente accesible [1]. La figura 2.27 muestra un ejemplo de los diferentes gestos que posee la base con diferentes sujetos. Como se ve en la figura, el dispositivo MS Kinect no posee una gran resolución y esto es notorio en toda la base de datos, en ocasiones generando mucho ruido en las imágenes.

[1] http://empslocal.ex.ac.uk/people/staff/np331/index.php?section=FingerSpellingDataset

Figura 2.27: Ejemplo de imágenes de las base de datos “ASL Finger Spelling Dataset” realizada por Pugeault y Bowden en 2011. Imagen tomada de [126].

MU_HandImages_ASL. Otra base de datos existente para clasificación de gestos estáticos manuales es la creada por Barczak et al. en 2012 [13]. La base de datos consta de 36 clases que representan el alfabeto dactilológico de la lengua de señas norteamericana, incluyendo letras y números. Las imágenes fueron tomadas a cinco voluntarios, con una cámara convencional RGB en un entorno controlado con fondo color verde. El resultado son 2425 imágenes en total para las 36 clases. Las imágenes fueron segmentadas segmentando el color, dando como resultado imágenes que sólo posee la mano con un fondo negro. La base de datos es públicamente accesible [2]. En el trabajo se comenta la posibilidad de extensión a 20 voluntarios para llegar 18.000 imágenes, pero al parecer el proyecto no continuó.

RWTH German Fingerspelling Database. En [43] Dreuw et al. presenta una base de datos para reconocimiento dactilológico de lengua de señas alemana, públicamente disponible[3]. La base de datos fue creada en 2006 en la Universidad RWTH Aachen y presenta 35 gestos que representan el alfabeto alemán, incluyendo las letras, números, y letras especiales particulares de la lengua. La imágenes fueron tomadas en formato video por dos cámaras web con una resolución bastante baja de 320x240 pixeles. 20 diferentes personas ejecutaron los gestos un par de veces cada una, dando un total de 1400 imágenes.

Base de datos de manos sintéticas. En [39], Dilsizian et al. tienen como objetivo realizar una mapeo del espacio 2D de una imagen a una reconstrucción 3D de la mano del sujeto, en un entorno de reconocimiento de lengua de señas. Para esto, en el año 2014 proponen la construcción de una base de datos de configuraciones de manos utilizando un guante de datos *Cyberglove*. Fueron almacenadas 87 configuraciones distintas, incluyendo diferentes orientaciones. No se hace mención sobre la disponibilidad de la base de datos.

Lengua de Señas Flamenca. En [164] Yuan et al. utiliza la base de datos de configuraciones de manos definida en *http://www.cs.bu.edu/groups/ivc/data/hand_shape*.

[2] http://www.massey.ac.nz/∼albarcza/gesture_dataset2012.html
[3] http://www-i6.informatik.rwth-aachen.de/∼dreuw/fingerspelling.php

Las imágenes representan configuraciones de la lengua de señas flamenca y fueron tomadas con una cámara de video dando como resultado alrededor de 1000 imágenes con anotaciones. No se especifica bien cuántas clases son.

Otras. Diversos artículos presentan modelos de clasificación de configuraciones de manos, donde las imágenes procesadas son obtenidas en el contexto del mismo artículo y no suelen darse muchos detalles sobre cómo fueron obtenidos o si el conjunto de datos es públicamente accesible. Ejemplos de esta situación pueden encontrarse en [90] [161] [83].

2.4.2 Bases de datos de Lengua de Señas

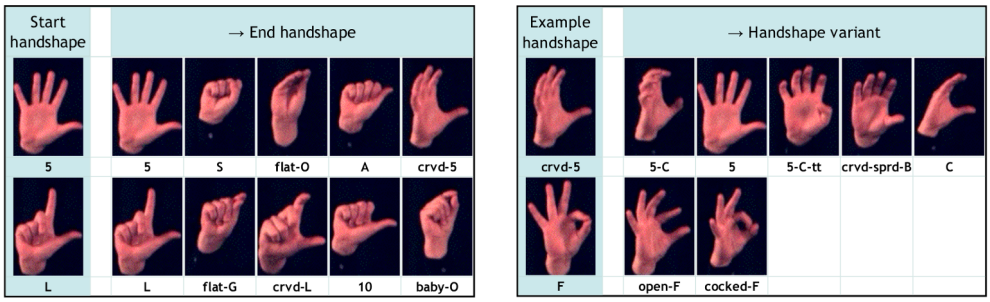
Con respecto a bases de datos para reconocimiento de lenguas de señas, existen numerosos trabajos desarrollados, con enfoques distintos. Por un lado, existen bases de datos orientadas al reconocimiento de patrones, como las que interesan en el marco de esta Tesis. Estas bases de datos suelen estar bien estructuradas, con diversos sujetos experimentales y varias muestras por clase/seña. Por otro lado, existen diversos trabajos donde se realizan bases de datos de lenguas de señas con un enfoque de investigación lingüística [87]. Generalmente aquí se intenta crear un diccionario específico para consulta web, o describir las señas por su estructura, describir variaciones dialécticas, diferencias de pronunciación, etc. Estas bases, en ocasiones no están bien estructuradas para un proceso de aprendizaje automático. A continuación se describen algunas de las bases de datos más relevantes actualmente.

The American Sign Language Lexicon Video Dataset (ASLLVD) . Una de las bases de datos más completas, realizada por la Universidad de Boston [12][111] y disponible públicamente[4]. El trabajo realizado desde el año 2008 contiene más de 3.300 señas organizadas en sentencias. Las sentencias poseen anotaciones lingüísticas como el tiempo exacto donde comienza y termina cada seña, las configuraciones de inicio y fin, sinónimos de la seña, etc. Cada video representa una sentencia de varias señas. Esto hace que cada seña tenga una cantidad variable de sujetos que la realizan (entre 1 y 6) y cantidad variable de repeticiones (promedio de 3). Se utilizaron diversas cámaras sincronizadas y es posible descargar todos los videos para las sentencias que fueron grabadas de ese modo. En [112] publican una interfaz para buscar y descargar los videos a través de la web [5]. La base de datos está enmarcada en un proyecto que tuvo numerosas publicaciones, incluyendo experimentos actuales sobre ella. La figura 2.28 muestra con ejemplos algunos fotogramas del conjunto de datos, y un trabajo realizado sobre las diferentes configuraciones de las manos en cada seña.

Las desventaja más grande que posee este cuerpo de datos es que: la información no está estructurada de un modo uniforme. Como se mencionó, cada sujeto realiza diferentes señas e incluso en diferentes etapas de grabación, dando como resultado

[4] http://www.bu.edu/av/asllrp/dai-asllvd.html
[5] http://secrets.rutgers.edu/dai/queryPages/search/search.php

(a)



(b)

Figura 2.28: The American Sign Language Lexicon Video Dataset. (a) Ejemplo de un fotograma tomado de [12]. (b) Estudio sobre configuraciones iniciales y finales, junto con diferentes variantes que poseen, realizado en [144]

un conjunto de datos complicado para su evaluación. Por otro lado, los videos no poseen ningún tipo de metadatos sobre seguimiento de manos. Sumado al hecho de que los intérpretes utilizan vestimenta no neutral (ropa casual, con diferentes colores), encontrar y segmentar correctamente las manos no es una tarea sencilla. Por último, la base de datos posee pocos sujetos y pocas repeticiones por clase.

RWTH-Boston-50. Creada originalmente por la Universidad de Boston para investigación lingüística, y luego adaptada para reconocimiento de patrones por la Universidad RWTH Aachen [166], esta base de datos de la lengua de señas norteamericana es públicamente accesible [6]. Está grabada con cuatro cámaras, tres de las cuales son blanco y negro. Dos de estas se utilizan para realizar visión estereoscópica. La cámara color sólo se utiliza para la sección del rostro de los intérpretes. En total son 483 sentencias con 50 señas distintas, interpretadas por 3 sujetos. Los video publicados sólo tiene 195*165 pixeles según se comenta en la documentación. En [44] se presenta una variante de la base de datos titulada “RWTH-Boston-104”donde el objetivo estuvo puesto en el reconocimiento continuo. Contiene 161 sentencias con 104 señas distintas. Si bien estas bases de datos poseen muy baja

[6] http://www-i6.informatik.rwth-aachen.de/aslr/database-rwth-boston-50.php

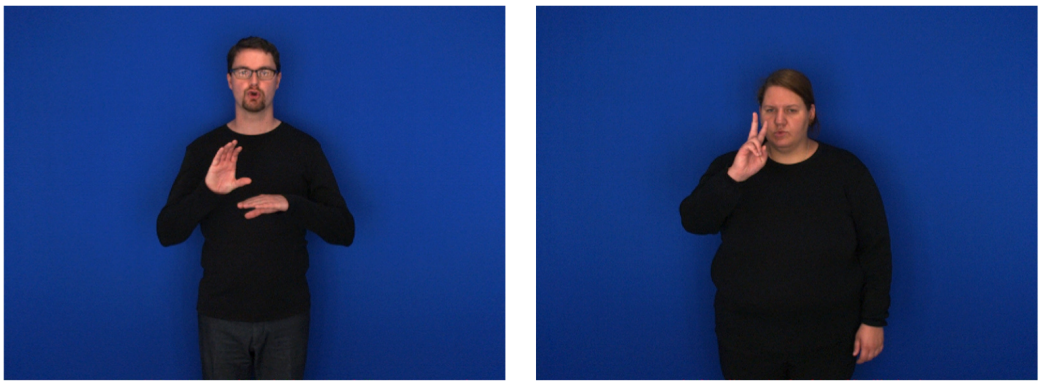


Figura 2.29: Ejemplo de imágenes de la base de datos “SIGNUM” realizada por la Universidad RWTH Aachen [151]

calidad de imagen, los autores adjuntan documentación sobre el seguimiento de las manos y la cabeza

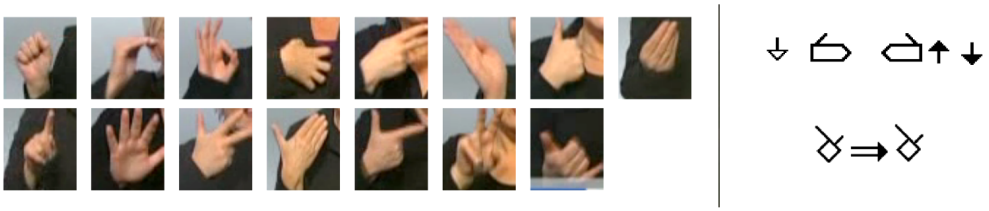
SIGNUM. Esta base de datos, creada por la Universidad RWTH Aachen en el marco de un proyecto gubernamental, fue creada específicamente para reconocimiento de patrones [152][151]. Es una de las bases de datos existentes más grandes de lengua de señas, con 450 señas distintas y 25 intérpretes. La base de datos tiene un volumen de 920Gb, por lo que es necesario contactar al autor para conseguirla[7]. A los intérpretes se les indicó realizar sentencias de actividades de la vida cotidiana como ir al cine, esperar el autobús, etc. En un contexto de evaluación dependiente al sujeto, el intérprete ejecutó 3 veces las 603 sentencias de entrenamiento y las 177 sentencias separadas para evaluación. Para evaluar el comportamiento independiente al sujeto, se llamó a 25 intérpretes que sólo ejecutaron cada sentencia una vez [87]. Los videos fueron tomados por una cámara con una resolución de 780x580 pixeles situada frente a los intérpretes, los cuales vistieron ropa oscura, sobre un fondo azul. La figura 2.29 muestra dos fotogramas ejemplos tomados de la base de datos.

RWTH-PHOENIX-Weather. Al igual que la anterior, desarrollada por la Universidad RWTH Aachen, esta es una de las bases de datos más grandes y realista públicamente disponible[8]. Presentada originalmente en 2012 por Forster et al. en [51], el conjunto de datos fue generado a partir de recortes de videos reales de un canal de televisión donde intérpretes oyentes traducían reportes del estado del clima. Según los autores, es un intento de quitar a los sistemas de reconocimiento del laboratorio y comenzar a utilizar entornos reales. En 2014 se extendió la base de datos [52] incorporando nuevos videos y anotaciones del seguimiento de las manos así como marcadores de diferentes puntos de interés del rostro de los intérpretes. Además, se realizaron anotaciones sobre diferentes posturas y formas de las manos

[7] http://www.bas.uni-muenchen.de/Bas/SIGNUM

[8] https://www-i6.informatik.rwth-aachen.de/web/Software/Databases/Signlanguage/details/rwth-phoenix/index.php

(a)



(b)

Figura 2.30: Base de datos RWTH-PHOENIX-Weather. Imágenes tomadas de [52]. (a) Fotograma de video original. (b) Izquierda, las 15 configuraciones detectadas y anotadas en los videos. Derecha, ejemplo de anotación *SignWriting* para dos señas.

utilizando la notación *SignWriting*[9] [142]. Debido a que la imagen del intérprete es sólo un fragmento del video original, el formato final sólo posee 210x260 pixeles, lo cual es todo un desafío para reconocimiento de patrones. El cuerpo de datos está dividido en un subconjunto de entrenamiento y otro de evaluación, tanto para pruebas dependientes del sujeto como independientes. El tamaño del vocabulario es de 1200 y 9 intérpretes que realizan las señas. No todos los intérpretes realizan las mismas señas, ni la misma cantidad de repeticiones, debido a la naturaleza real de los videos. La figura 2.30 muestra ejemplos de la base de datos junto con anotaciones.

BosphorusSign. Recientemente, en 2016, la Universidad Boğaziçi ha desarrollado un cuerpo de datos de la lengua de señas turca, públicamente accesible[10]. El trabajo publicado en [23] por Camgöz et al. tiene como principal objetivo el uso de una aplicación de ayuda para gente sorda en entornos controlados como un hospital

[9] http://www.signwriting.org/
[10] www.BosphorusSign.com

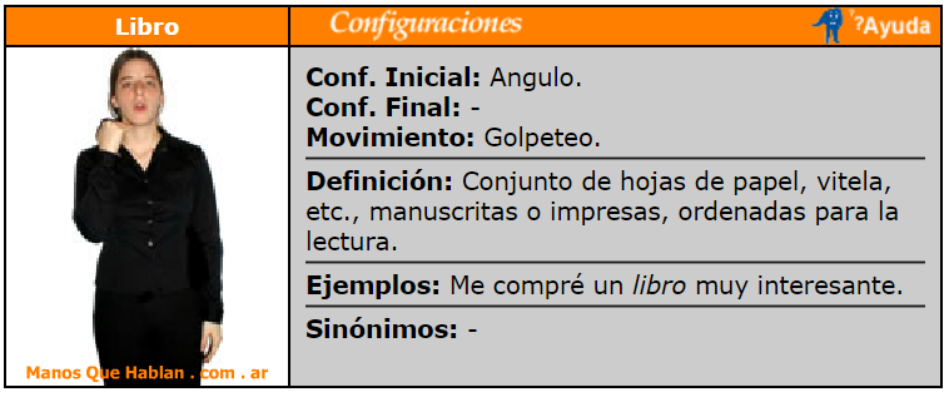


Figura 2.31: Cuerpo de datos "Manos que Hablan". Ejemplo de la palabra "Libro" tomado de [2].

o un banco. Consta de más de 800 señas separadas en tres dominios diferentes: señas utilizadas en una visita a un hospital, señas de carácter financiero utilizadas en un banco, y señas utilizadas en actividades diarias. Las señas fueron capturadas por un dispositivo MS Kinect por 10 intérpretes distintos, y efectuadas 6 veces por cada uno. La base de datos incluye anotaciones de las señas utilizando la notación HamNoSys [78]. Si bien los autores afirman que también puede ser utilizada para reconocimiento de patrones, el principal objetivo de la base de datos es social/lingüístico. El cuerpo de datos es grande, aunque aun no está disponible para descargar.

Otras bases de datos estructuradas. Existen numerosos trabajos donde se utilizan cuerpos de datos descriptos en el mismo trabajo, pero sin posibilidad de descarga de los datos, o sitio web que lo describa. Por ejemplo, en [117] Ong et al. utiliza dos bases de datos de lenguas de señas para evaluar sus métodos. La primera es un conjunto capturado con un dispositivo Kinect de la lengua de señas alemana, que consta de 40 señas interpretadas por 14 sujetos distintos. La segunda, es una base de datos más robusta de 981 señas del léxico griego, capturada con una cámara RGB.

Bases de datos para uso lingüístico. Como se mencionó a principios de la sección, existen algunos trabajos orientados a investigación lingüística que resultan inapropiados para su utilización en sistemas de aprendizaje automático. Por ejemplo, el cuerpo de datos titulado "Corpus NGT" [35] es un conjunto de videos de la lengua de señas holandesa (NGT). Tiene una sorprendente cantidad de 100 intérpretes grabados en diferentes lugares del país, con un total de 72 horas de grabación. En [137] se define el "BSL Corpus", una base de datos de la lengua de señas inglesa[11]. Cuenta con una asombrosa cantidad de 249 sujetos y más de 40.000 items léxicos. Fue creado en el marco de un proyecto gubernamental y desarrollado en la Universidad de Londres con participación de otras universidades de todo el país.

En el contexto de la Lengua de Señas Argentina (LSA), la única base de datos

[11] http://www.bslcorpusproject.org/

Nombre	Lengua	#señas	#sujetos	#MporS	Uso
ASLLVD [12]	Norteamericana	3300	1-6	3,5	Reconocimiento
RWTH-Boston-50 [166]	Norteamericana	50	3	-	Reconocimiento
RWTH-PHOENIX-Weather [52]	Alemana	1200	9	38	Reconocimiento
SIGNUM [152]	Alemana	450	25	74	Reconocimiento
BosphorusSign [23]	Turco	>800	10	6	Ambos
Corpus NGT [35]	Holandesa	-	100	-	Lingüístico
BSL Corpus [137]	Inglesa	-	249	-	Lingüístico
Manos que Hablan [2]	Argentina	-	-	-	Lingüístico
Dicciseñas	Varios	-	-	-	Lingüístico

Tabla 2.1: Comparativa de bases de datos de lenguas de señas. MporS= Muestras por Seña.

digital grande conocida es "Manos que Hablan" [2]. En este caso, fue diseñada particularmente para uso lingüístico/social[12], dando la posibilidad de tener un diccionario on-line de la LSA. El proyecto data del año 2001 y fue creado por un grupo interdisciplinario de personas sin ninguna relación con instituciones, empresas u organizaciones. La sección principal del sitio, posee un diccionario ordenado tanto por orden alfabético, como agrupado por secciones de interés: educación, lugares, pronombres, colores, verbos, etc. Por otro lado, la página web también posee un traductor dactilológico donde es posible escribir una palabra y el sistema la interpreta mostrando imágenes de los diferentes símbolos que representan cada letra. La figura 2.31 muestra un ejemplo tomado del sitio web. El dinamismo de cada seña no fue obtenido mediante una cámara, sino que es representado por varias imágenes, y con una cartilla de información de cómo se realiza el movimiento, la configuración inicial y la final. Claramente, este sitio no resulta útil para un sistema de reconocimiento automático de señas del LSA.

Algo similar a lo anterior ocurre con "Dicciseñas"[13]. Este es un trabajo del Centro de Desarrollo de Tecnologías de Inclusión (CEDETi UC), perteneciente a la Escuela de Psicología de la Pontificia Universidad Católica de Chile (PUC). El trabajo cuenta principalmente con léxico chileno, pero incluye también algunos países latinos como Argentina, México, Uruguay, Costa Rica, y también España. Particularmente para la LSA, tiene diversas palabras agrupadas en diferentes secciones. En este caso, las señas fueron capturadas en videos por intérpretes jóvenes. Al igual que el cuerpo de datos mencionado en el párrafo anterior, no resulta apropiado para un sistema de aprendizaje automático. La tabla 2.1 resume todas las bases de datos para lenguas de señas mencionadas anteriormente.

2.4.3 Bases de datos de acciones humanas

A continuación se nombran las principales bases de datos existentes en la literatura para reconocimiento de acciones humanas. Todos los cuerpos de datos han sido

[12] http://manosquehablan.com.ar/
[13] http://diccisenas.cedeti.cl/

capturados utilizando dispositivos MS Kinect o enfoques similares.

MSRC-12. Esta base de datos desarrollada por Fothergill et al. [53] es un desarrollo conjunto de la Universidad de Cambrige junto con el equipo de investigación de Microsoft y es públicamente accesible[14]. El cuerpo de datos consta de 12 gestos distintos categorizados en dos grupos: gestos icónicos y gestos metafóricos. Los icónicos corresponden a acciones concretas del mundo real, mientras que los metafóricos representan conceptos abstractos. La base de datos fue capturada con 30 sujetos diferentes con un total de 594 secuencias de videos donde cada sujeto ejecuta diversas repeticiones la misma acción. Un aporte interesante realizado en este trabajo radica en el estudio de cómo influyen las instrucciones dadas a los sujetos para realizar los diferentes gestos. Se instruyó a los sujetos con tres tipos de materiales: un texto descriptivo indicando el tipo de movimiento a realizar, un conjunto ordenado de imágenes representativas del gesto, y un video de una persona interpretando el gesto. Así, se analizó el efecto que tuvo el modo de instrucción utilizado en cada caso.

MSR Action3D. Nuevamente un trabajo del equipo de Microsft, en este caso junto con la Universidad de Wollongong, en Australia realizaron esta base de datos de acciones humanas públicamente disponible [15]. El trabajo fue publicado por Li et al. en [95]. Consta de 20 acciones diferentes que podrían encontrarse en un videojuego como patear una pelota, ejecutar un golpe, realizar un saque de tenis, etc. Las acciones fueron ejecutadas por 10 sujetos, dando un total de 567 secuencias distintas. En el artículo se dividen los gestos en tres grupos superpuestos, donde los primeros dos presentan acciones similares, y el tercero intenta ser un grupo que posee todas acciones complejas. La base de datos es muy citada, aunque contiene diversos errores de seguimiento en algunas secuencias que hasta los propios autores indican que se deben quitar.

HDM05. Con casi 10 años de antigüedad, esta base de datos es una de las más citadas creadas con una configuración tipo MoCap (*Motion Capture*). Estos enfoques se basan en poner un traje a la persona con diferentes marcadores en forma de luz visible. Luego diferentes cámaras son puestas alrededor del sujeto y calibradas para obtener información de los marcadores a cada instante de tiempo. El resultado es similar a las articulaciones del esqueleto que otorga el MS Kinect, pero el primer caso con mayor precisión. La base HDM05 fue desarrollada en la Universidad de Bonn (Alemania). Presentada en [106] por Müller et al., provee varias horas de grabación que consisten en diversas secuencias con un total de 70 gestos distintos organizados en categorías realizadas por 5 sujetos, que interpretaron las acciones múltiples veces.

MHAD (Berkeley Multimodal Human Action Database). Esta base de datos fue desarrollada en el año 2013 por la Universidad de California. Fue publicada por

[14] http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/
[15] http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc/

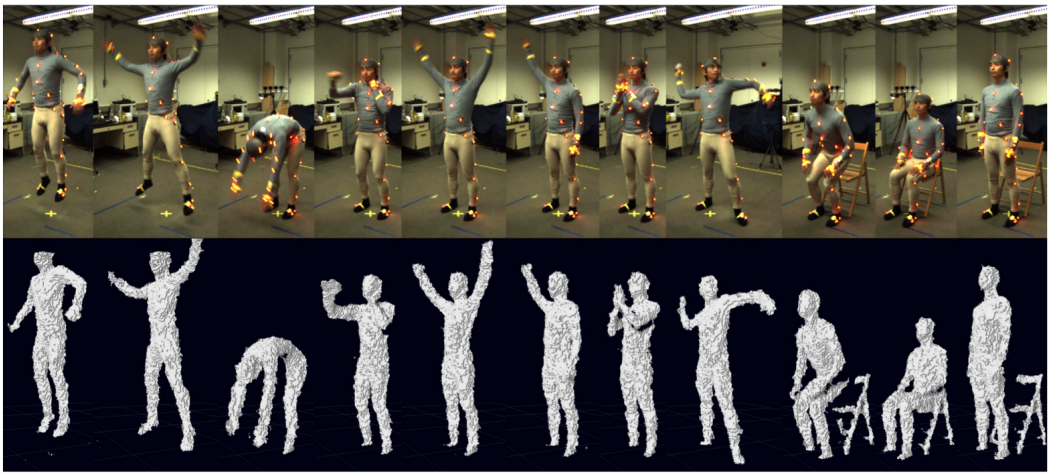


Figura 2.32: Ejemplos de los 11 gestos que posee la base de datos MHAD. Arriba, imágenes tomadas por una cámara RGB. Abajo, mapas de profundidad tomado por un dispositivo Kinect. Imagen tomada de [113].

Ofli et al. en [113] y actualmente es públicamente accesible[16]. Consta de sólo 11 gestos diferentes, pero posee una sofisticada puesta en escena. Los gestos fueron capturados con diversas cámaras RGB estéreo desde diferentes ángulos, varios dispositivos Kinect, un traje tipo *MoCap* y seis acelerómetros puestos en el mismo traje. La figura 2.32 muestra las 11 clases existentes donde se puede apreciar el traje que utilizaron los sujetos junto con mapas de profundidad obtenidos por los Kinects. Los gestos fueron ejecutados por 12 individuos diferentes que hicieron cada gesto 5 veces, dando un total de 80 minutos de grabación. El conjunto de datos es muy interesante por la cantidad de información capturada simultáneamente, aunque resulta poco reproducible en la práctica debido al alto costo y puesta en escena de todos los dispositivos utilizados.

2.5 Conclusiones del capítulo

El reconocimiento de gestos y particularmente de la Lengua de señas implica el estudio de diversos escenarios de dominios diferentes. Al observar las distintas estrategias existentes, se encuentra que no siempre resulta trivial utilizar estas estrategias en forma directa para el reconocimiento de gestos. Sino que por lo general es necesario una adaptación particular debido a la naturaleza multidisciplinar del dominio.

En general, es necesario definir muy bien qué tipo de dispositivo de captura se va a

[16] http://tele-immersion.citris-uc.org/berkeley_mhad

utilizar, y bajo qué condiciones (controladas, no controladas, tipo de iluminación, tipo de fondo, nuevos sujetos, etc). Esta elección del dispositivo de captura y condiciones de trabajo, traerá luego diversos enfoques a considerar para obtener los descriptores, y luego realizar un proceso de clasificación. Del mismo modo, es necesario definir cuántas clases habrá en un conjunto de datos, cuántos sujetos van a ejecutar los gestos, y cuántas repeticiones por gesto. Estos puntos son de vital importancia al momento de realizar un proceso de clasificación.

Para poder clasificar una seña correctamente es necesario realizar diversos procesos como son el reconocimiento de la posición de la cabeza y manos del intérprete, hacer un seguimiento de las manos en el espacio (2d o 3d), reconocer la configuración que posee cada mano, y clasificar utilizando la información temporal de las articulaciones. Todas estas líneas de investigación pueden ser analizadas en forma paralela pero son necesarias todas para la correcta clasificación. Gracias a los trabajos existentes sobre reconocimiento de rostros, ubicar la posición de la cabeza resulta generalmente simple. Sin embargo, al querer reconocer una mano en movimiento en video los trabajos existentes resultan escasos o poco prácticos. Los diferentes algoritmos que utilizan filtros de color de piel son simples de implementar pero presentan dos problemas: por un lado no escalan a diferentes tipos de color de piel, haciendo que sea impracticable para todos los escenarios. Por otro lado resulta difícil solucionar las múltiples oclusiones que las manos presentan con la cara del intérprete. Si se utiliza una tecnología de captura de video con sensores de profundidad, la posición de la mano es simple de computar. Sin embargo estas tecnologías no son aun totalmente precisas, además de ser más costosas para su aplicación en un entorno real. Sumado a esto, las bases de datos existentes no suelen tener información del movimiento de la mano, y ninguna posee algún tipo de marcador de color para facilitar la segmentación.

Clasificar las diferentes configuraciones que una mano puede tener, sigue siendo un problema no resuelto completamente. El principal desafío aquí es la extracción de características de la mano para su posterior clasificación. Existen numerosos trabajos al respecto que utilizan diferentes descriptores, cada uno con sus restricciones de implementación.

Con respecto al reconocimiento de acciones humanas, existen diversas bases de datos. Aquí el problema se convierte en un grupo de puntos que representan las articulaciones del cuerpo y se mueven en el espacio 3D. El enfoque en cómo capturar los descriptores generalmente cambia un poco con respecto a las señas, pero el desafío es similar: lograr clasificar el dinamismo de los movimientos realizados.

Siendo la lengua de señas un área de aplicación relativamente nueva y debido a su natural complejidad, existen pocas bases de datos disponibles públicamente. En general, los autores que realizan grandes investigaciones utilizan bases de datos propias, que conocen bien, y en pocos casos se publica información preprocesada sobre los datos. En general, muchos trabajos se realizan además para un uso lingüístico, sin importar de qué forma están estructurados los datos, o qué meta-información

existe de ellos. Otro problema encontrado en muchos trabajos es la poca cantidad de muestras por clase, lo que dificulta la generalización en un proceso de aprendizaje automático. Del mismo modo, muchos autores utilizan pocos sujetos, lo que imposibilita evaluar la independencia al sujeto de un sistema. Las nuevas bases de datos más grandes no suelen estar bien estructuradas o anotadas. Además, ya que cada región del mundo presenta su propio léxico, la utilización de una base de datos local es necesaria si se quiere llevar el sistema a una utilidad práctica real. Desde este punto de vista, no existen bases de datos argentinas para poder utilizar como validación de un sistema de clasificación.

Si bien existen algunas bases de datos públicas con información del seguimiento de las manos, la segmentación completa de esta sigue siendo un problema debido a diferentes oclusiones que existe entre la mano y el resto del cuerpo. La solución de utilizar un marcador de color, como puede ser un guante, resulta simple de segmentar y fácil de implementar en un entorno real.

Cabe mencionar además los diferentes desafíos mencionados en la sección 2.1, entre los cuales se encuentra la dependencia al sujeto. Como se mencionó, pocos trabajos desarrollan un correcto análisis de la precisión del sistema al incorporar un nuevo individuo.

Capítulo 3

Una base de datos de la Lengua de Señas Argentina

Realizar un proceso completo de recrecimiento automático de lengua de señas es un trabajo extenso que requiere de diversas etapas a abordar. Para evaluar un modelo de aprendizaje supervisado es necesario contar con una base de datos apropiada que permita la correcta extracción de la información relevante que se quiere clasificar. La base de datos aquí descripta se encuentra publicada en [130].

El capítulo se organiza del siguiente modo, en la sección 3.1 se comenta en forma resumida la motivación para la construcción de una base de datos de la Lengua de Señas Argentina. En la sección 3.2 se presenta la base de datos de configuraciones del LSA con 800 fotografías de 16 configuraciones diferentes. En la sección 3.3 se presenta la bases de datos LSA64 con 3200 videos de 64 señas distintas del LSA. Por último en la sección 3.4 se presentan las conclusiones del capítulo.

3.1 Motivación

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 [153, 29].

Como se vio en el capítulo 2, las bases de datos de configuraciones están basadas generalmente en fotografías de un grupo particular de formas de manos que se quiere reconocer. Estos grupos suelen ser el alfabeto de un lenguaje particular, por lo cual, en ocasiones estas bases de datos se conocen como dactilológicas. En algunos casos la base de datos está compuesta por videos de una mano siendo rotada en diversos

ejes, con el fin de identificar una configuración particular desde muchos puntos de vista. Un ejemplo de base de datos de este tipo es la desarrollada por Pugeault y Bowden en [126]. Aquí se usó una cámara tipo Kinect para capturar las diferentes configuraciones del alfabeto norteamericano en imágenes de profundidad. En [13] se presenta una base de datos análoga pero con imágenes RGB y diferentes posturas de las manos.

De modo general, la mayoría de las bases de datos existentes confían en un sistema de filtro por color de piel para segmentar las manos y cara del intérprete. Estos filtros generalmente no son robustos ante variaciones como la ropa del intérprete, u oclusiones comunes de los gestos entre mano-mano o mano-cara. Normalmente estos filtros por color requieren además extracción de características morfológicas como un paso consecuente para identificar la posición y forma de las manos. Una aproximación a este problema son las bases de datos que utilizan cámaras de profundidad (imágenes RGB-D) para obtener detalles de las posiciones. Esto genera una limitación en la utilidad del modelo, además de lo poco precisos que son aún este tipo de dispositivos.

Adicionalmente a todos los inconvenientes mencionados anteriormente, cabe resaltar que los léxicos son únicos por países, e incluso por regiones dentro de un país. Esta característica provoca que incluso pudiendo realizar una correcta clasificación sobre una base de datos existente, sólo podría utilizarse el sistema bajo ese dialecto. Actualmente, la única información digitalizada de Lengua de Señas Argentina son pequeños diccionarios online como por ejemplo *Manos que Hablan* [2]. Estos diccionarios no poseen una estructura fija o estandarizada como para realizar un proceso de aprendizaje automático. Esto imposibilita la creación de un modelo real que permita traducir videos del LSA.

La base de datos aquí presentada tiene dos objetivos específicos: por un lado, al estar grabada con guantes de color en las manos de los intérpretes, permite 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 general para el léxico argentino, inexistente hasta el momento.

3.2 Construcción de una base de datos de configuraciones

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 base de datos presentada en esta sección tiene dos objetivos principales: por un lado introducir una base de datos de imágenes, utilizando marcadores de color

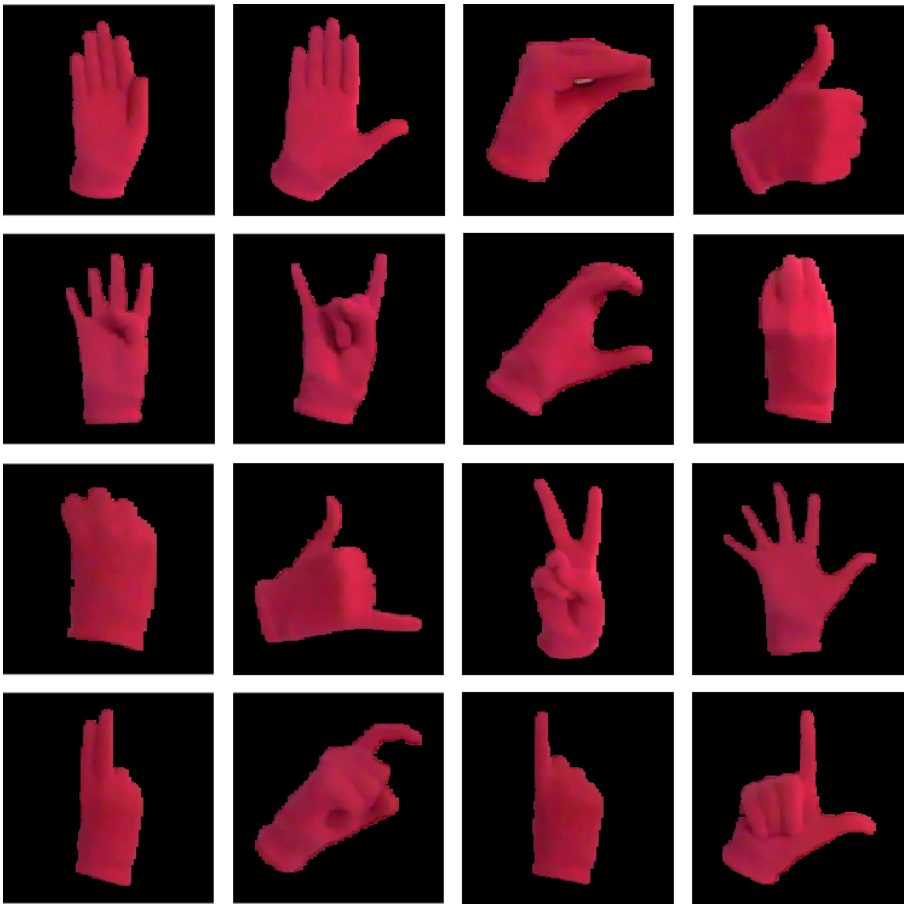


Figura 3.1: Ejemplos de cada clase de la base de datos de configuraciones LSA16

que permite evaluar cualquier clasificador de imágenes en un sistema de aprendizaje automático. Por otro lado, presenta 16 de las configuraciones del LSA más usadas en el léxico (ver [2]), convirtiéndose en un diccionario único a nivel regional. Los resultados del desarrollo de esta base de datos pueden encontrarse publicados en [133].

La base de datos, 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 3.1 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, como se observa en la figura 3.2. Para la simplificar el problema de segmentación de la mano dentro de una imagen, los sujetos utilizaron guantes de tela con colores

Figura 3.2: Imágenes no segmentadas de la base de datos de configuraciones LSA16

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.

La base de datos LSA16 de configuración está públicamente disponible, bajo licencia *Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International*, sólo para uso académico, en *https://midusi.github.io/lsa16/index.html*. En el sitio se encuentra información detallada de las diferentes configuraciones y sujetos de prueba. Es posible descargar la base de datos completa, con las 800 imágenes en formato *PNG* con 640x480 píxeles. También es posible descargar una versión preprocesada que consta de las imágenes de las manos segmentadas con fondo negro. Esto permite realizar experimentos evitando el proceso de segmentado por color.

3.3 Construcción de una base de datos de señas

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.

Es esta sección se detalla la creación de la base de datos “LSA64”, primer conjunto de datos específico para la Lengua de Señas Argentina. En esta sección se

detallan la características de la base de datos, junto con una serie de estadísticas que ejemplifican el alcance y objetivos de la misma.

Como explica S. Mitra en [102], al igual que para escribir o hablar, un gesto realizado por un individuo puede resultar diferente a como lo hace otro. Incluso en cada intento de realizar el mismo gesto en diferentes instancias. Por lo general, los gestos no están perfectamente definidos y esto hace necesario contar con un gran número de muestras en una nueva base de datos.

LSA64 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 [158], donde se utilizaron guantes no sólo para segmentar la mano sino para facilitar la clasificación de la configuración.

Si bien la base de datos no posee el tamaño de algunas bases mencionadas en el capítulo 2 como ASLLVD, RWTH-PHOENIX-Weather o SIGNUM, tiene más ejemplos y sujetos que muchas otras, además de la facilidad para llevar a cabo el proceso de segmentación de manos, como ya se mencionó. La figura 3.3 muestra algunos ejemplos de la base de datos, donde se pueden visualizar los 10 sujetos distintos que interpretaron las señas. Las señas fueron seleccionadas en base a las más utilizadas incluyendo diversos sustantivos y verbos, así como también para ser suficientemente variables en movimientos y configuraciones. La table 3.1 muestra una lista detallada de todas las señas de la base de datos junto con información relevante. En total, hay 22 señas con dos manos, y 42 con una mano.

3.3.1 Grabación y disponibilidad

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. La figura 3.3 deja ver las diferencias en la temperatura de color del balance de blancos. Estas diferencias en iluminación permiten entrenar un modelo robusto que funcione en diferentes entornos.

En ambos sets de grabación los intérpretes usaron ropa oscura sobre un fondo blanco y ejecutaron las señas sentados y parados. Como ya se mencionó, se utilizaron guantes de color para facilitar la segmentación de las manos. Para la mano derecha se utilizó un guante de color rosa fuerte, y para la mano izquierda un guante de color verde intenso. Cada seña fue ejecutada 5 veces distintas por cada sujeto, incluyendo algunas variaciones mínimas para aumentar la diversidad y el realismo en la base de datos. La cámara utilizada fue una Sony HDR-CX240. El trípode estuvo siempre a

Figura 3.3: Ejemplos de señas diferentes de la base de datos LSA64, junto con los 10 sujetos que las realizaron.

ID	Nombre	Manos	ID	Nombre	Manos
1	Opaco	Una	33	Hambre	Una
2	Rojo	Una	34	Mapa	Dos
3	Verde	Una	35	Moneda	Dos
4	Amarillo	Una	36	Musica	Dos
5	Brillante	Una	37	Nave Espacial	Una
6	Celeste	Una	38	Ninguno	Una
7	Colores	Una	39	Nombre	Una
8	Rosa	Una	40	Paciencia	Una
9	Mujer	Una	41	Perfume	Una
10	Enemigo	Una	42	Sordo	Una
11	Hijo	Una	43	Trampa	Dos
12	Hombre	Una	44	Arroz	Dos
13	Lejos	Una	45	Asado	Dos
14	Cajón	Una	46	Caramelo	Una
15	Nacer	Una	47	Chicle	Una
16	Aprender	Una	48	Fideos	Dos
17	Llamar	Una	49	Yogurt	Dos
18	Espumadera	Una	50	Aceptar	Dos
19	Amargo	Una	51	Agradecer	Dos
20	Dulce	Una	52	Apagar	Una
21	Leche	Una	53	Aparecer	Dos
22	Agua	Una	54	Aterrizar	Dos
23	Comida	Una	55	Atrapar	Dos
24	Argentina	Una	56	Ayudar	Dos
25	Uruguay	Una	57	Bailar	Dos
26	Pais	Una	58	Bañarse	Dos
27	Donde	Una	59	Comprar	Una
28	Apellido	Una	60	Copiar	Dos
29	Burla	Dos	61	Correr	Dos
30	Cumpleaños	Una	62	Darse cuenta	Una
31	Desayuno	Dos	63	Dar	Dos
32	Foto	Dos	64	Encontrar	Una

Tabla 3.1: Información básica de cada seña en la base de datos LSA64. La columna Mano indica si la seña utiliza una o dos manos.

2,5 metros de distancia de la pared y a 1,5 metros de altura. El modo de grabación fue establecido para modelar el espacio alrededor del cuerpo que normalmente se utiliza para comunicarse con la lengua de señas.

La base de datos es públicamente accesible bajo licencia *Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International* sólo para uso académico. Se encuentra almacenada en *https://midusi.github.io/lsa64/index.html*. Está disponible la versión original, que consta de 3200 videos en formato Full-HD, con un total de casi 2Gb de información. También es posible descargar una versión pre-procesada, que consta de información del seguimiento de las manos de los intérpretes, junto con videos de las manos segmentadas y con fondo negro. Esta versión tiene un peso de paroximadamente 800Mb. Los archivos fueron computados con el entorno Matlab.

3.3.2 Características principales de la base de datos

En esta sección se muestran algunas características, estadísticas e información de las señas en LSA64 para mejorar la comprensión de la naturaleza y desafío que presenta la base de datos. Las figuras muestran la superposición existente entre las señas en términos de posición movimiento y configuración, lo que conlleva una evaluación no trivial de nuevos modelos de reconocimiento. Toda la información fue computada desde la versión pre-procesada de la base de datos, como se detalla en el capítulo 4.

En las figuras 3.4 y 3.5 se puede observar las diferentes configuraciones de las manos derecha e izquierda respectivamente para cada clase en LSA64. Las imágenes fueron tomadas del primer fotograma de un video particular para cada clase. Claramente puede observarse que existen diversas clases con superposición de configuraciones, aunque su proyección 2D puede ser diferente dependiendo de la rotación de la mano. Por ejemplo, las clases 6, 7, 10, 12, 16, 34, 36, 37, 41, 42, 46 y 62 comienzan todas con la misma configuración en la mano derecha, que se basa en la mano cerrada con el dedo índice levantado. Esto genera una gran ambigüedad que el modelo debe sortear luego, con otras características de la seña.

En algunas ocasiones, la configuración cambia durante el desarrollo de una seña, dando resultados complejos en su interpretación. Cabe notar que si bien la mayoría de las configuraciones utilizadas existen como clases en la base de datos LSA16 descripta en la sección 3.2, la complejidad que presentan las señas involucran nuevas rotaciones y transformaciones que no estaban contempladas en las fotografías. Esto supone un nuevo desafío para un clasificador de señas.

La figura 3.6 muestra una distribución de las posiciones iniciales y finales de cada mano al interpretar una seña. Cada elipse representa la media junto con su co-varianza de cada una de las 64 señas en la base de datos. Aquí no se quiso hacer distinción entre cada clase, ya que sólo es de caracter ilustrativo, para entender la distribución que existen del espacio 2D de las señas. Mientras que algunas pocas

Figura 3.4: Imágenes de las manos segmentadas para cada clase en la base de datos LSA64. Cada imagen muestra la configuración inicial de la mano derecha para cada seña.

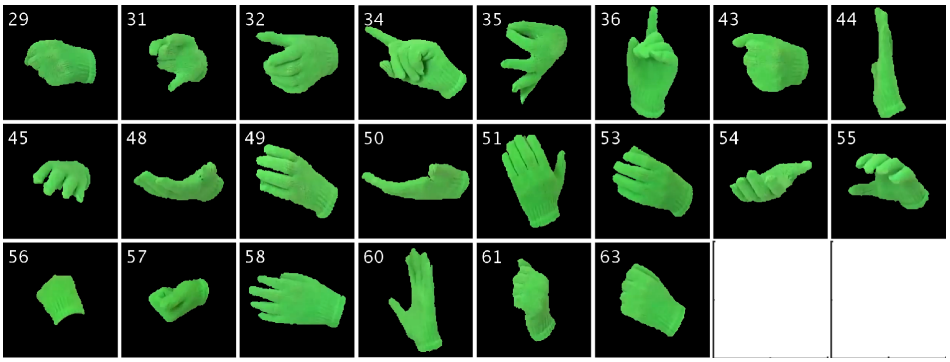
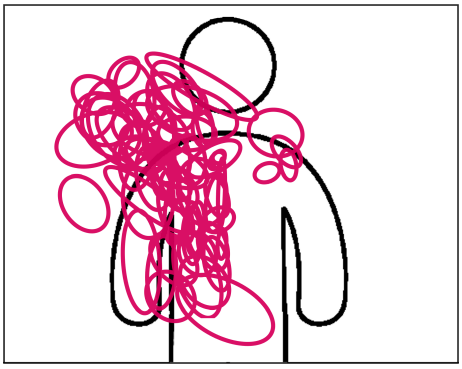
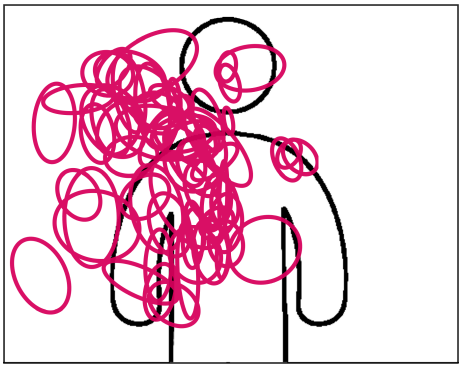


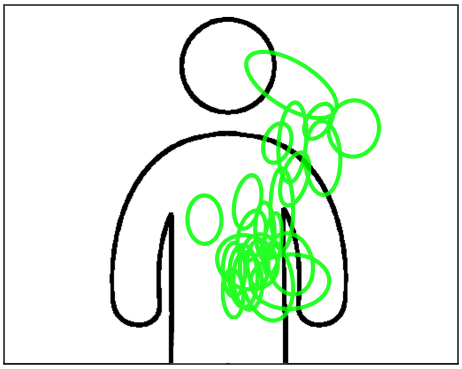
Figura 3.5: Imágenes de las manos segmentadas para cada clase en la base de datos LSA64. Cada imagen muestra la configuración inicial de la mano izquierda para cada seña.



(a) Posiciones finales de todas las señas para la mano derecha.



(b) Posiciones finales de todas las señas para la mano derecha.



(c) Posiciones finales de todas las señas para la mano izquierda.

(d) Posiciones finales de todas las señas para la mano izquierda.

Figura 3.6: Medias de las posiciones iniciales y finales de cada mano. Las elipses representan la media y la covarianza de la distribución 2D, suponiendo una distribución gaussiana.

señas podrían ser distinguidas por sus posiciones iniciales o finales, en general existe una superposición significativa. La mayor parte de las señas utilizan zonas cercanas al hombro, la cabeza y el abdomen. La mano izquierda se utiliza en una menor cantidad de señas, y por lo general acompaña a la mano derecha que se considera dominante.

La figura 3.7 muestra un ejemplo de la trayectoria de las manos para cada seña de la base de datos. Para esta figura, los ejemplos fueron tomados del sujeto $n^o 2$. Al igual que para otras estadísticas, puede verse que existe mucha superposición entre las distintas clases. Por ejemplo, las clases 1, 5, 7, 13 y 19 presentan la misma trayectoria en señas de una mano. Las clases 31, 32 y 61 son señas de dos manos que poseen igual trayectoria. Claramente, esta ambigüedad debe eliminarse con el

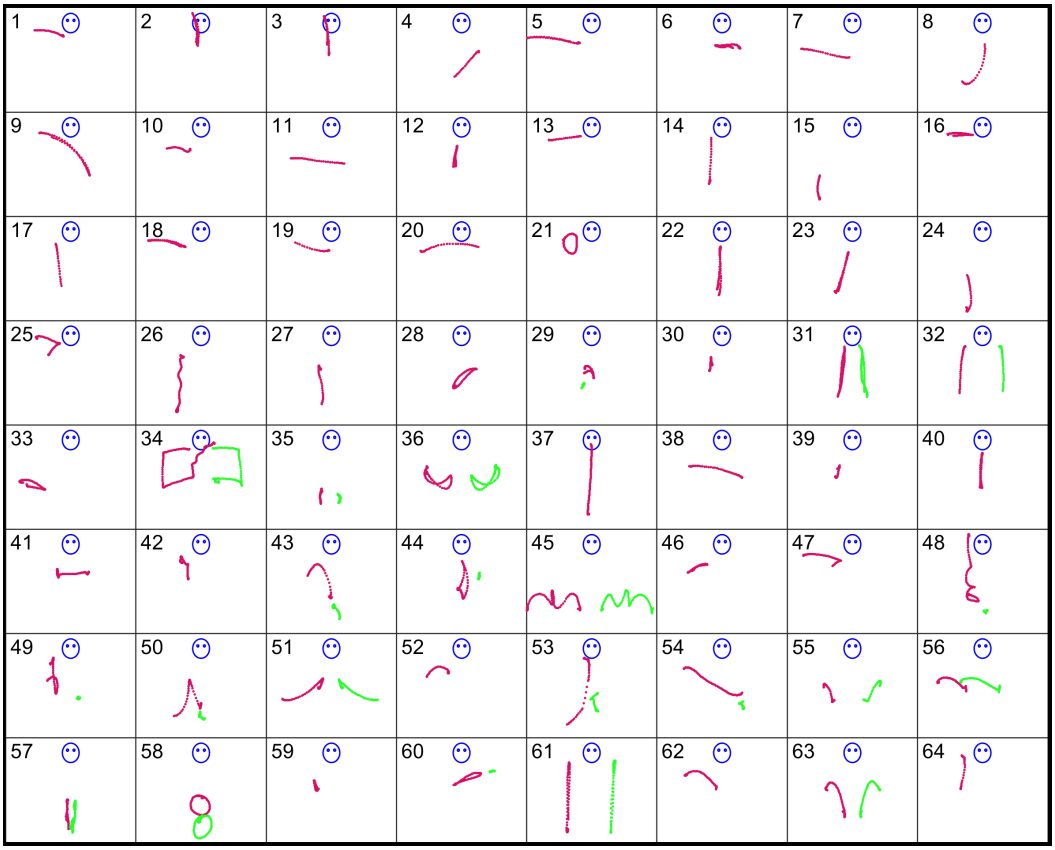


Figura 3.7: Trayectoria de las manos para cada seña de la base de datos. La mano derecha es representada en color rosa fuerte, la mano izquierda en verde claro, y la cabeza como un círculo azul en el medio.

resto de características que componen la seña, principalmente la configuración de las manos.

Por último, la figura 3.8 muestra la cantidad de movimiento de cada mano para cada clase, medida como la máxima distancia entre dos puntos de toda la trayectoria. El gráfico deja ver la natural diferencia que existe en las distintas clases respecto a su amplitud de movimiento. Mientras que algunas clases como la 29 poseen una apertura de movimiento muy pequeña, clases como la 37 se desplazan hasta 4 veces más. Esto podría ser un aspecto más a considerar en el momento de un reconocimiento. Muchos sistemas de clasificación abordan el problema del seguimiento con la premisa de que las manos siempre se mueven. Este gráfico, deja ver que existen señas con muy poco movimiento, las que resultan difíciles de diferenciar entre una seña auténtica o un pequeño movimiento involuntario del brazo del intérprete.

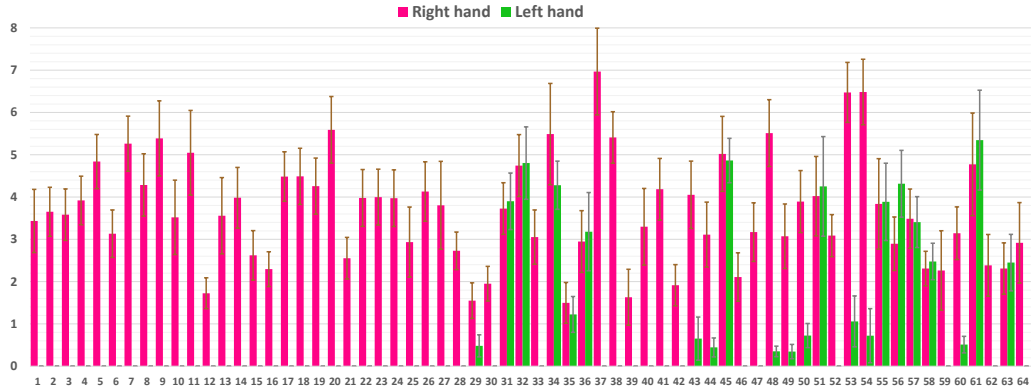


Figura 3.8: Cantidad de movimiento de cada mano para cada clase en la base de datos LSA64. Las barras rosas muestran la cantidad de movimiento de la mano derecha y las barras verdes de la mano izquierda, en las señas que utilizan dos manos.

3.4 Conclusiones del capítulo

En este capítulo se mostró el desarrollo de una base de datos de la Lengua de Señas Argentina (LSA). El objetivo del desarrollo se basó en dos puntos principales: por un lado, no existe una base de datos de léxico argentino que permita realizar un proceso de aprendizaje automático. Como se vio en el capítulo 2, los únicos conjuntos de datos del léxico son pequeñas páginas web con diccionarios visuales. Por otro lado, el uso de guantes de color facilita completamente la segmentación de las manos del intérprete, al mismo tiempo que no involucra un anexo invasivo ni costoso.

La base de datos desarrollada está dividida en tres partes principales: por un lado una sub-base de imágenes para poder clasificar 16 configuraciones diferentes típicas en la LSA. En segundo lugar, un repositorio público de 64 señas segmentadas interpretadas por 10 sujetos diferentes teniendo un total de 3200 videos. Junto con el conjunto de videos, se provee información pre-procesada sobre el seguimiento de las manos en relación a la posición de la cabeza, así como videos de las manos segmentadas. Esto facilita la utilización del cuerpo de datos a otros investigadores o trabajos futuros.

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.

Capítulo 4

Definición del modelo de clasificación

En este capítulo se presenta el modelo de clasificación propuesto en esta tesis para la traducción automática de señas. Debido a la natural complejidad del dominio, el modelo está compuesto por diversos módulos que abordan los diferentes problemas como la segmentación de las manos, seguimiento, cálculo de descriptores y etapas de clasificación.

El capítulo se organiza del siguiente modo: en la sección 4.1 se define el esquema general del modelo propuesto; en la sección 4.2 se describe el proceso de clasificación de configuraciones de manos, lo que luego es utilizado como parte de los descriptores de una seña, desarrollado en la sección 4.3; la sección 4.4 termina de desarrollar los detalles del modelo de clasificación junto con sus clasificadores parciales; la sección 4.5 detalla algunos casos puntuales particulares del lenguaje de señas que es necesario tener en cuenta; por último, la sección 4.6 presenta las conclusiones del capítulo.

4.1 Descripción general del modelo

El modelo desarrollado en esta tesis consta de diversas etapas para lograr la clasificación de gestos segmentados. La figura 4.1 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 gestos 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.

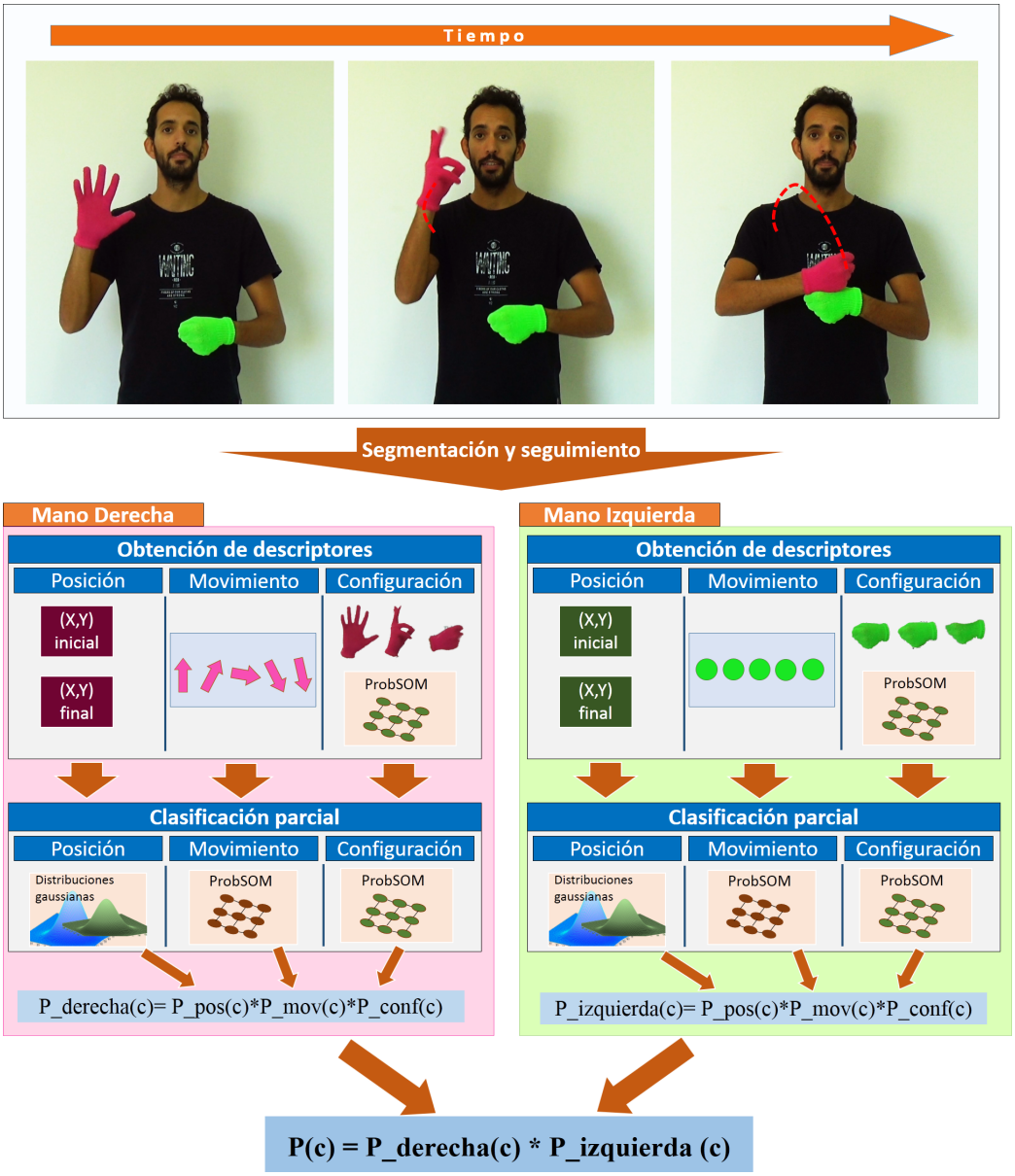


Figura 4.1: Descripción general del modelo de clasificación propuesto para señas segmentadas.

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 facilitará luego la clasificación de la configuración.
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 [49]. Los detalles de los procesos de obtención de descriptores pueden verse en la sección 4.3.
3. **Clasificación parcial.** En tercer lugar se realiza un proceso de clasificación parcial. Dado cada conjunto de descriptores se clasifican de forma independiente. Para clasificar la posición de cada mano se utilizó un sistema de distribución de gausianas, considerando el conjunto de las diferentes posiciones como una distribución normal. Tanto para la clasificación del movimiento como de la configuración se utilizaron redes ProbSOM. Los detalles de estos procesos pueden encontrarse en la sección 4.4
4. **Clasificación de una seña.** Por último, los resultados de los clasificadores parciales son utilizados como entrada para el clasificador probabilístico.

A lo largo del capítulo se hace foco en cada una de estas etapas, describiendo de un modo detallado los diferentes componentes que se aprecian en la figura 4.1. Así mismo se discuten diversos aspectos a tener en cuenta como restricciones y limitaciones sobre las manos, movimientos, etc. También se detallan diferentes técnicas utilizadas previamente como descriptores o clasificadores del estado del arte.

4.2 Clasificación de configuraciones de manos

Como se mencionó anteriormente la configuración de una seña es una característica principal de su composición. Diferentes señas poseen la misma trayectoria pero con distintas configuraciones. En esta sección se aborda el problema de la clasificación de configuraciones de manos, pasando por las etapas de segmentación, generación de descriptores y modelos de clasificación. Los resultados de esta sección pueden encontrarse publicados en [133].

4.2.1 Preprocesamiento de la imagen

Antes de poder calcular los descriptores de una mano, para luego clasificarlos, es necesario realizar una serie de pasos que permitan ajustar y normalizar la fotografía

(a) Imagen original RGB.

(b) Canal H del modelo HSV.

(c) Canal S del modelo HSV.

(d) Canal V del modelo HSV.

Figura 4.2: Ejemplo de conversión de modelo de color RGB a HSV.

de la mano. El primer paso para lograr una correcta clasificación de las configuraciones, es la segmentación de la mano. Como se mencionó anteriormente, este proceso puede llevarse a cabo desde diferentes enfoques. Ya que el trabajo presentado en esta tesis se enfoca en las etapas de generación de descriptores y clasificación, la segmentación fue realizada mediante un proceso simple de filtrado por color. Este filtrado depende claramente de las imágenes utilizadas para evaluar. Como se verá en el capítulo 5 el conjunto de datos utilizados para evaluación en este caso fue la base de datos LSA16 de configuraciones de LSA, descripta en el capítulo 3 de esta tesis.

Para llevar a cabo el proceso de filtrado se transformó el espacio de color RGB de la imagen a un espacio HSV (*Hue, Saturation, Value*). Este modelo permite identificar de un modo más robusto y sencillo el color a filtrar. La figura 4.2 muestra un ejemplo de la base de datos utilizada donde puede verse la imagen original y su conversión al modelo HSV, mostrando sus tres canales de manera separada. Al

(a) Imagen original segmentada. (b) Imagen rotada. (c) Máscara de segmentación. (d) Contorno final.

Figura 4.3: Pre-procesamiento de la imagen de una mano.

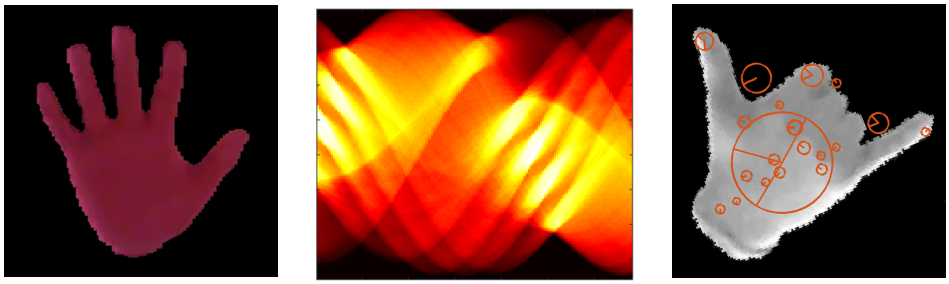
utilizar marcadores de color como son los guantes, es posible identificar la mano mediante la intersección de los tres canales. Por lo general los guantes poseen un tono de color particular, representado por el canal H, una saturación alta, representada por el canal S, y una luminosidad relativamente alta, dependiendo claro de la iluminación de la escena, que está representada por el canal V. Este proceso permite segmentar las manos del intérprete de un modo simple generando una máscara de color.

Como segundo paso en el pre-procesamiento de las imágenes es necesario realizar una serie de procedimientos con dos fines principales: obtener un contorno, y orientar la mano para que todas las imágenes de las mismas configuraciones estén en posiciones similares antes de computar los descriptores. Para esto, en cada imagen se calculan los ejes principales de los pixeles de la mano, y con ellos su inclinación ϕ. Luego, la imagen es rotada $-\phi^{\text{o}}$ para llevarla a una orientación canónica. Debido a que esta orientación es insensible a rotaciones de 180^{o} de la mano, puede ocurrir que la imagen quede orientada hacia arriba o hacia abajo. Para corregir esto, se calcula la cantidad de cruces de cada línea horizontal posible en la imagen, y se estima la posición de los dedos en base si la moda de la cantidad de cruces se encuentra en la parte superior o inferior de la imagen.

Posteriormente, la imagen es remuestreada sin afectar su relación de aspecto a un tamaño de 128×128 pixeles y se re-posiciona de modo que la misma quede centrada. El contorno de la mano se obtiene aplicando un filtro de bordes a la máscara de segmentación de la mano, la cual contiene una sola componente conexa. La figura 4.3 muestra un ejemplo del proceso completo de pre-procesamiento.

4.2.2 Descriptores

Una vez procesada la imagen, es necesario realizar una conversión adecuada que permita obtener una representación vectorial que describa la configuración que se quiere clasificar. Para esto, en este trabajo se analizaron diferentes descriptores del estado del arte. A continuación se describen los dos descriptores principales evaluados, uno basado en la transformada de Radon, y el otro en los *Scale-Invariant Feature Transform* (SIFT). Se realizaron además pruebas con descriptores utilizados



(a) Imagen original segmentada.

(b) Transformada de Radon.

(c) Representación de los vectores SIFT.

Figura 4.4: Ejemplos de descriptores de una mano segmentada.

tradicionalmente para clasificar imágenes, como descriptores de Fourier, Banco de filtros de Gabor, Local Binary Patterns o Histogramas de Gradientes Orientados (HOGs). Los resultados fueron inferiores a los presentados en casi todos los casos. Por esta razón se decidió no desarrollar esta información, y sólo presentar los dos descriptores que mejores resultados otorgaron.

Transformada de Radon

La transformada de Radon ha sido utilizada en el pasado para reconocer objetos y también para identificar a personas en base a las características de su mano. Por ejemplo, en [54] se utilizó esta transformada como descriptor de la mano para realizar autenticación biométrica.

La transformada de Radon de una imagen 2D $f : \mathbb{R}^2 \to \mathbb{R}$ se define como una integral de línea sobre la imagen. La línea L a través de la cual se integra está dada por un par (b, θ), donde b es la distancia al origen de la línea y θ el ángulo con el eje horizontal de la imagen. Matemáticamente queda representada por la siguiente fórmula:

$$R_{(b,\theta)} = \iint_{L_{(b,\theta)}} f(\vec{x})|d\vec{x}| = \int_{-\infty}^{\infty} f(x(t), y(t))dt = \int_{-\infty}^{\infty} f(tsin\theta + bcos\theta, -tcos\theta + bsin\theta)dt \tag{4.1}$$

Aplicando la versión discreta de la misma a la imagen segmentada para todas las combinaciones de valores enteros de (b, θ) posibles ($1, \ldots, 180$ para θ, un valor K dependiente del tamaño de la imagen para b), obtenemos un descriptor $R \in \mathbb{R}^{180 \times K}$. Luego, para reducir la dimensionalidad r se remuestrea a un tamaño fijo $r \in \mathbb{R}^{32 \times 32}$. Esta reducción simplifica la complejidad del dominio sin generar pérdida de información.

Este descriptor se utiliza como global considerándolo un vector en $r' \in \mathbb{R}^{32^2}$, o como 32 descriptores locales tomando cada fila r_i, $i = 1, \ldots, 32$, $r_i \in \mathbb{R}^{32}$ como un descriptor local. Cada r_i entonces contiene una aproximación suave a los $R_{(b,\theta)}$ para todo b, y donde θ corresponde aproximadamente a la media de un subconjunto de ángulos contiguos. La figura 4.4 muestra un ejemplo de aplicación de la transformada de Radón a una imagen segmentada y preprocesada de la base de datos.

En particular, como el clasificador que se utilizó, el ProbSom, tiene como entrada un conjunto de cardinalidad arbitraria de vectores, se utilizaron los vectores r_i como descriptores de entrada. Generalmente, cuanto más vectores de entrada existan para un ejemplo particular, el ProbSOM trabajará mejor, ya que trabaja bajo un modelo estadístico (ver [49]). Por otro lado, para el resto de clasificadores evaluados, se utilizó el vector completo r' concatenando todas las filas para formar un sólo vector característica.

SIFT

Un descriptor SIFT es un histograma espacial 3D de los gradientes de una imagen, que caracteriza la apariencia de un punto de interés. Para ello, con el gradiente de cada pixel se calcula un descriptor más elemental formado por la ubicación del pixel y la orientación del gradiente. Dado un posible punto de interés, estos descriptores elementales son pesados por la norma del gradiente y acumulados en un histograma 3D que representa el descriptor SIFT de la región alrededor del punto de interés. Al formar el histograma, se le aplica a los descriptores elementales una función de peso gaussiana para darle menos importancia a los gradientes que están más lejos del centro punto de interés.

Los descriptores SIFT han sido aplicados a varias tareas de visión por computadoras, incluyendo el reconocimiento de configuraciones de manos [168] y reconocimiento de rostros [92].

4.2.3 Modelo de clasificación

Como se definió en el capítulo 2, ProbSOM [49] es una adaptación probabilística de los mapas auto-organizados de Kohonen(SOM)[85]. Estos mapas son redes competitivas no supervisadas que configuran sus neuronas para representar la distribución de los datos de entrada procesados durante la fase de entrenamiento. Como resultado de esta fase de aprendizaje se obtiene una red donde cada neurona aprende a representar un área del espacio de entrada y agrupa los vectores de datos por su similitud o cercanía.

El proceso de entrenamiento del ProbSOM se realiza de la misma manera que en el algoritmo SOM convencional. ProbSOM agrega una etapa adicional luego del entrenamiento para pesar la proporción de representación de cada neurona. Para ello se repasan todos los patrones de entrada y se agrega a cada una de las neuronas ganadoras información acerca de la clase que representa y en que proporción.

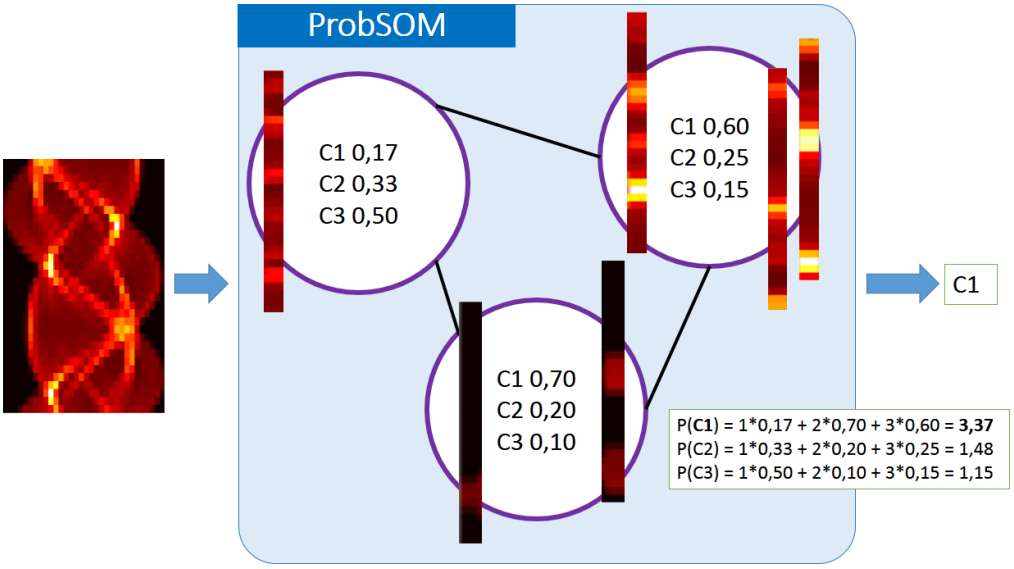


Figura 4.5: Esquema ejemplo de la etapa de clasificación del ProbSOM para configuraciones de manos, utilizando la Transformada de Radon como descriptor de entrada. En el ejemplo sólo existen 3 clases.

El proceso de reconocimiento también es similar a las redes SOM. El mecanismo de respuesta que decide la identificación de una clase consiste en un sistema probabilista. Como cada vector no permite por si solo la identificación de una clase, un conjunto de vectores es requerido. Cuando un conjunto de vectores de características son introducidos en la red, se obtiene un conjunto de neuronas ganadoras donde cada una representa a varias clases con una proporción determinada. La clase identificada será aquella cuya suma de proporciones sea máxima. ProbSOM ha demostrado ser un algoritmo robusto para resolver problemas de clasificación [92, 49, 147] donde las clases se representan por un conjunto de vectores de características.

La figura 4.5 muestra un ejemplo de clasificación del método ProbSOM. Los vectores de entrada son representados por la transformada de Radon de una mano segmentada. En una etapa inicial, debe realizarse el proceso de *clustering* (agrupamiento), donde se genera un esquema de entrenamiento inicial y cada neurona queda asociada con un porcentaje de representación de cada clase. Luego, para clasificar un nuevo patrón, los diferentes vectores quedan ubicados en las neuronas (*centroides*) de la red y se calcula la probabilidad de pertenecer a cada clase, en base a la distribución inicial.

4.3 Definición de descriptores de una seña

Como se mencionó en la sección 4.1 el modelo aquí desarrollado presenta diversas etapas para llevar a cabo la clasificación de señas hechas con las manos. En esta sección se detallan los descriptores que permiten representar vectorialmente

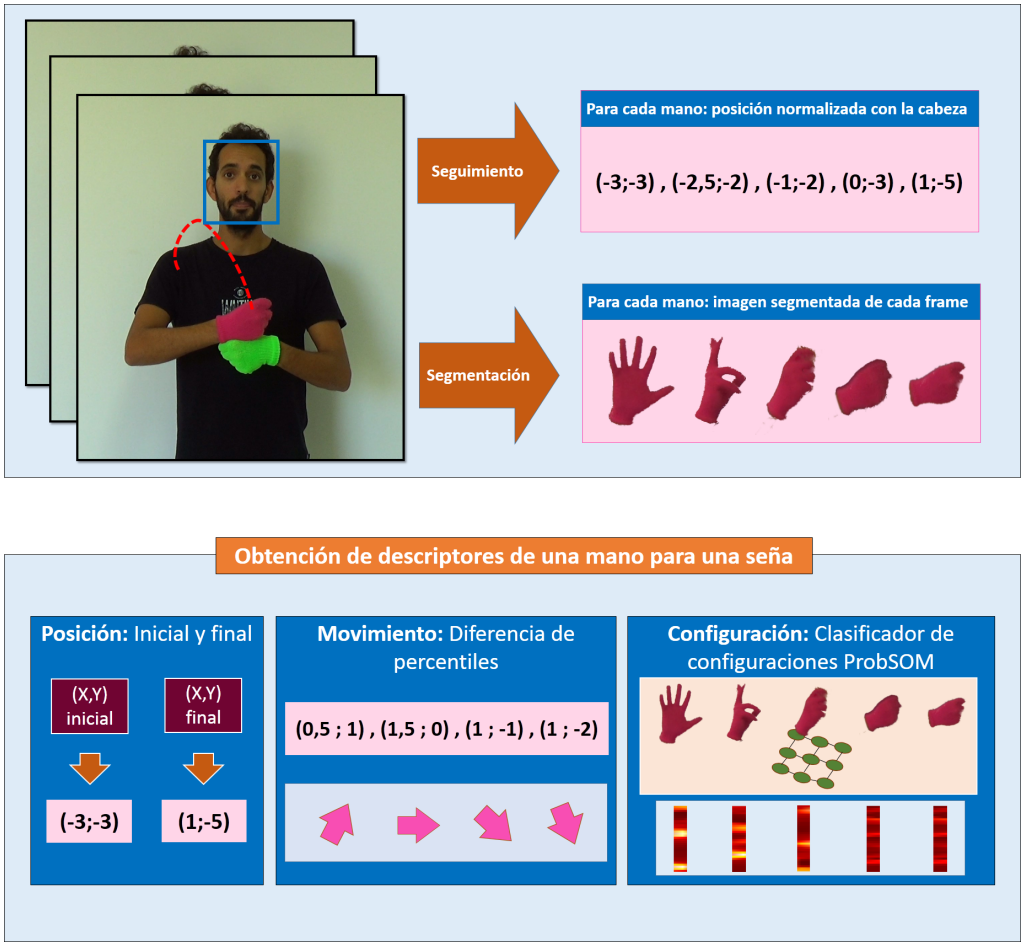


Figura 4.6: Generación de descriptores para una seña. Ejemplo para una sola mano en un video con 5 fotogramas.

la información de una seña. Como se vio en la figura 4.1, los descriptores que se obtienen pueden separarse en 3 partes principales: descriptores de configuración, descriptores de movimiento, y descriptores de posición. Los 3 conjuntos de descriptores se computan para ambas manos.

Para obtener esta serie de descriptores es necesario primero haber realizado un proceso de segmentación y seguimiento tanto de ambas manos, como de la cabeza del intérprete. En primer lugar se detecta la cabeza del intérprete por medio del algoritmo Viola-Jones para detección de rostros [148]. Este método está ampliamente difundido para reconocimiento de objetos en general. Por las características morfológicas y estáticas que presentan los rostros de las personas (dos ojos, una nariz, una boca, etc.) es posible aplicar este algoritmo eficazmente para encontrar los rostros de los intérpretes en los videos.

En segundo lugar, se realiza una segmentación y seguimiento de ambas manos del modo en que se describió en la sección anterior. Aquí, las posiciones de las manos

son desplazadas para establecer como eje de coordenadas la posición de la cabeza. De este modo se evitan pequeños desplazamiento en las posiciones que un intérprete pueda tener frente a la cámara. Como resultado de este proceso se obtienen dos vectores de posiciones 2D, uno para cada mano, con tantos puntos como fotogramas tenga el video. Por último, estos valores son normalizados por una constante en base a la altura del video medida en pixeles. Esto tiene como fin la independencia del formato de video. La figura 4.6 muestra un ejemplo con la etapa completa de obtención de descriptores. La figura muestra como ejemplo sólo la obtención de descriptores de la mano derecha. A continuación se detallan los tres descriptores propuestos.

Descriptores de posición

Los descriptores de posición son quizá los más sencillos de computar. En base a las posiciones obtenidas de las manos, el descriptor de posición son dos puntos en $\mathbb{R}^2$ representando la posición inicial y final de la trayectoria de cada mano. Es decir la posición de la mano en el primer y último fotograma del video.

Descriptores de movimiento

La trayectoria es quizá una de las partes fundamentales de una seña. La correcta clasificación del tipo de movimiento realizado por cada mano permite establecer un punto de partida crucial en la identificación de la seña. Aquí, básicamente es necesario entender los tipos de movimientos realizados por los intérpretes y que limitaciones existen en ellos. Generalmente, estos movimientos son simples y precisos. Por ejemplo, una mano sube, una mano se mueve a un costado, una mano realiza un círculo, etc. Estos movimientos en ocasiones se realizan con ambas manos al unísono, en otros casos con una sola mano, o en muchos casos la mano dominante se mueve y la otra queda quieta.

Para generar una serie de vectores representativos del movimiento de cada mano, se utilizaron las diferencias entre las posiciones de cada fotograma de video. En el ejemplo de la figura 4.6 se puede apreciar la diferencia entre las posiciones de los únicos cinco fotogramas existentes. No obstante, en un caso real los videos suelen tener muchas más muestras. Para llevar esto a cabo se utilizó la diferencia de ciertos percentiles elegidos experimentalmente. Al aumentar el número de percentiles se adquiere mayor información en cuanto a cambios de dirección de una seña, pero se pierde precisión en la magnitud de desplazamiento. En general, 9 percentiles resultó ser una cantidad apropiada tanto para captar todos los cambios de dirección de una seña, como así también para tener buena información relevante al desplazamiento. Esto se detalla en el capítulo 5.

Descriptores de configuración

Para computar los descriptores de configuración, se utilizó el trabajo descripto en la sección 4.2. El método utilizado responde con un vector de tamaño 16 para cada imagen dada como argumento. Este vector salida representa la probabilidad de pertenencia de la imagen a alguna de las 16 clases conocidas por el modelo. Dada una seña, puede ocurrir que contenga diferentes configuración durante su ejecución. Por este motivo, utilizar sólo una imagen del video no sería suficientemente representativo. Entonces, para generar los descriptores de configuración se utilizó la salida del ProbSOM para las imágenes segmentadas de los mismos percentiles utilizados para los descriptores de movimiento.

4.4 Clasificación de señas segmentadas

Como se mencionó en secciones anterior, el modelo planteado en esta tesis es un sistema probabilístico basado en las diferentes características que componen una seña. El modelo combina la salida de dos subclasificadores, uno para cada mano. A su vez, el clasificador de cada mano combina las salidas de tres subclasificadores que modelan la información referente a la posición, el movimiento y las configuraciones. Para alimentar estos tres subclasificadores se utilizan los descriptores mencionados en la sección anterior.

Ya que el modelo propuesto clasifica cada mano de forma independiente, para calcular la pertenencia de una seña x a una clase se define la probabilidad dada una clase c como:

$$P(x|c) = P(x^i|c)P(x^d|c) \tag{4.2}$$

donde x^i y x^d es la información referente a la mano izquierda y derecha respectivamente. Dividir en dos las probabilidades depende de la suposición de que la sincronización entre las manos no es importante para reconocer una seña, o al menos que esa información no es crucial para el reconocimiento.

Como ya se mencionó, la probabilidad para una mano depende de los tres subclasificadores que se enfocan en diferentes características de la seña. Ya que se usaron los mismos tipos de subclasificadores para cada mano, en los siguientes párrafos se usará la notación m para hacer referencia indistintamente a i o d (izquierda, o derecha). Entonces, la probabilidad para una mano puede escribirse del siguiente modo:

$$P(x^m|c) = P(x^m_{pos}|c)P(x^m_{mov}|c)P(x^m_{con}|c) \tag{4.3}$$

donde x^m_{pos} es la información referente a la posición de la mano, $x^m_{mov}|c$ es la información referente al movimiento, y x^m_{con} es la información referente a la configuración. La información de cada etapa está dada por los descriptores propuestos en la sección 4.3.

Nuevamente, la ecuación 4.3 asume independencia entre x^m_{pos}, $x^m_{mov}|c$ y x^m_{con}, lo que significa, por ejemplo, que la posición de la mano no restringe de ningún modo los tipos posibles de movimientos o configuraciones. Si bien en algunos casos esto no sería correcto (por ejemplo, no es posible mover la mano hacia abajo si está tocando el piso), ya que generalmente los intérpretes están restringidos a un cuadrado imaginario donde realizan las señas, esta suposición resulta cierta en la mayoría de los casos.

Entonces, para clasificar una seña x de un video segmentado, se elige la clase c con la máxima probabilidad $P(x|c)$. Es posible reescribir la ecuación 4.2 como:

$$P(x|c) = P(x^i|c)P(x^d|c) = P(x^i_{pos}|c)P(x^i_{mov}|c)P(x^i_{con}|c)P(x^d_{pos}|c)P(x^d_{mov}|c)P(x^d_{con}|c) \tag{4.4}$$

En los párrafos siguientes se describe cómo se calcula la probabilidad para cada subclasificador. Es decir, se define cada subclasificador detalladamente. La tabla 4.1 resume algunos símbolos utilizados en la notación de los párrafos anteriores y siguientes.

4.4.1 Subclasificador de Posición

Para clasificar la seña en base a la posición se utilizaron la primera y última posición, como ya se describió en la sección 4.3. Se realizó una prueba 2D Kolmogorov-Smirnov con la primera y última posición para todos los videos en la BD utilizada y se encontró que con un nivel de confianza del 95 % existe evidencia para rechazar la hipótesis de normalidad en un 30 % de las clases. No obstante al ajustar las posiciones con un Modelo de Mixturas Gaussianas se encontró que para el 79 % de las clases un solo componente provee el mejor BIC (*Bayesian Information Criteria*), y el desempeño en evaluación resultó peor suponiendo dos componentes, posiblemente debido a un sobreajuste. Por consiguiente, se eligió modelar las posiciones (inicial y final de forma separada) como una simple distribución normal 2D.

Con los datos de entrenamiento utilizados, se calculó la media $\mu_{pi,c}$ y $\mu_{pf,c}$ y covarianzas $\Sigma_{pi,c}$ y $\Sigma_{pf,c}$ para las posiciones iniciales (pi) y finales (pf), para cada clase c. Luego, la probabilidad para una nueva seña con los descriptores de posición definidos por x_{pi} y x_{pf}, para una clase c es obtenida por la siguiente ecuación:

$$P(x^h_p|c) = g_{pi,c}(x^m_{pi})g_{pf,c}(x^m_{pf}) \tag{4.5}$$

c	Clase	x	Una seña	m	Mano m (genérico)	$(\cdot)^m$	Información de la mano m
$(\cdot)^i$	Información de la mano izquierda	$(\cdot)^d$	Información de la mano derecha	x^m_{pos}	Posición	x^m_{con}	Configuración
x^m_{mov}	Movimiento	x^m_{tr}	Trayectoria	x^m_{cm}	Cantidad de movimiento	x^m_a	Ausencia de la mano m en testeo
a^m_c	La mano m no es usada en la clase c	$a^h_{c,m}$	Ausencia de movimiento en entrenamiento				

Tabla 4.1: Referencias de notación. La variable x referencia siempre a la información de una seña. Subíndices indican tipo de información. Superíndices indican la mano. La variable a referencia un parámetro del modelo.

donde $g_{pi,c}$ es una función de densidad de probabilidad gausiana 2D con media $\mu_{pi,c}$ y covarianza $\Sigma_{pi,c}$. La función $g_{pf,c}$ es definida análogamente para la última posición

4.4.2 Subclasificador de Movimiento

Se consideraron dos factores para definir la probabilidad basada en movimiento: la trayectoria realizada por la mano (x_{tr}) y la cantidad de movimiento realizado (x_{cm}). Por lo tanto, la probabilidad de movimiento para una clase c se define como:

$$P(x^m_m|c) = P(x^m_{tr}|c)P(x^m_{cm}|c) \tag{4.6}$$

Cantidad de movimiento

Para representar la cantidad de movimiento de una mano m para una seña c se computó la distancia máxima entre todos los pares de posiciones que la mano recorre. En etapa de entrenamiento, se calcula la media $\mu^m_{cm,c}$ junto con su desviación $\sigma^m_{cm,c}$, para cada clase. En etapa de testeo, se penalizan las clases para las cuales la cantidad de movimiento difiere demasiado de $\mu^m_{cm,c}$. Se define entonces, la probabilidad de cantidad de movimiento como:

$$P(x^m_{cm}|c) = g_{cm,c}(x^m_{cm}) \tag{4.7}$$

donde $g_{cm,c}$ es una función de distribución Gaussiana 1D con media $\mu^m_{cm,c}$ y desviación estandar $\sigma^m_{cm,c}$.

Trayectoria

En la sección 4.3 se mencionaron los diferentes descriptores utilizados para representar una seña. Los descriptores de movimiento utilizados son gradientes que representan la dirección de movimiento de cada mano en pequeños fragmentos de video. Para calcular la probabilidad de la trayectoria de una mano $P(x^m_{tr}|c)$, se utilizó el clasificador ProbSOM, similar al modo en que se usó para clasificar las configuraciones (ver sección 4.2). Un descriptor de trayectoria se basa en una serie de vectores 2D, donde cada uno representa un gradiente que indica la dirección y magnitud de la mano en el espacio.

Una particularidad que presenta el clasificador ProbSOM es que cada patrón entra para ser elegido por una neurona, pero no hay relación entre los diferentes patrones al momento de elegir la clase ganadora. Por esta razón, es posible decir que los patrones son desordenados, dando el mismo resultado ingresar los patrones, por ejemplo, *AAB*, que *ABA*. Esta particularidad conlleva a la situación que la temporalidad que existe en los patrones se elimina parcialmente. Si bien cada gradiente posee cierta información temporal debido al modo en que fueron calculados, los diferentes gradientes son desordenados por el clasificador. Esto podría resultar perjudicial en algunos dominios. Particularmente se encontró que en el lenguaje de señas esta característica no afectó el desempeño del clasificador.

4.4.3 Subclasificador de Configuraciones

Para obtener la probabilidad $P(x^m_{con}|c)$ de configuración para una seña particular se utilizó, al igual que para el movimiento, el clasificador ProbSOM. Como se mencionó en la sección 4.3, los descriptores utilizados para la configuración representan la probabilidad para un fotograma particular de que sea algunas de las 16 clases que el modelo de configuraciones previamente entrenado conoce. Estos diferentes vectores utilizados como descriptores nuevamente son desordenados por el clasificador ProbSOM como se mencionó en la sección anterior.

4.5 Limitaciones y casos especiales

Si bien el modelo presentado aplica a todo el dominio del lenguaje de señas (o a cualquier dominio de gestos realizados con las manos), existen algunos casos especiales a tener en cuenta, donde no es posible aplicar el sistema de probabilidades antes mencionado, debido a que resulta inpracticable su cómputo. Estos casos pueden resumirse principalmente en dos: puede ocurrir que una de las manos no exista en el video; y por otro lado, puede ocurrir que una mano no realice movimientos.

4.5.1 Señas con una sola mano

En el caso en que una mano no exista en un video, la ecuación 4.2 deja de tener sentido, ya que la probabilidad $P(x^m|c)$ para la mano inexistente sería imposible de calcular. Esto es algo habitual ya que muchas clases de la base de datos utilizan una única mano, estando oculta la segunda. Podría ocurrir también que una seña realizada con una sola mano tenga la otra dentro del área visible por la cámara, pero en una posición y configuración aleatoria. En este caso, sería deseable no considerar esta información, ya que podría interpretarse por el clasificador como un intento genuino de realizar otra seña. Para evitar estas situaciones, se puede reescribir la ecuación 4.2 como:

$$P(x|c) = P(x^i|c)^{a_c^i} P(x^d|c)^{a_c^d} \tag{4.8}$$

donde a_c^i y a_c^d son parámetros con valores 1 o 0. Establecer $a_c^m = 0$ permite al modelo ignorar la información para la mano m respecto a la clase c, lo que sería equivalente a establecer la probabilidad para esa mano en 1.

Es posible calcular el parámetro a_c^m de varias formas. La más simple es establecer estos parámetros de acuerdo a anotaciones de la base de datos en base a cuándo se utilizó una o dos manos para cada clase. Aunque estas anotaciones no siempre están disponibles. Otra opción podría ser selecionar los descriptores analizando la contribución de cada mano en la clasificación de la seña, y establecer $a_c^m = 0$ siempre que la mano no haya tenido una contribución real para el reconocimiento. Para esta tesis se utilizó una estrategia intermedia: se consideró que una seña fue realizada con una mano si en más del 50 % de los fotogramas del video la mano no fue detectada.

En etapa de entrenamiento es sencillo calcular qué clases utilizan una mano y qué clases utilizan ambas manos. Luego, en etapa de testeo si una mano es usada en una seña particular c (es decir $a_c^m = 1$), pero la mano no se detecta en el 50 % de los fotogramas entonces la probabilidad para esa clase es fijada a 0.

Es posible computar x_a^m como una variable booleana con valores 1 cuando la mano m está presente en etapa de testeo. Luego, es posible definir:

$$P(x_a^m|c) = \begin{cases} 0 & \text{if } a_c^m = 1 \text{ and } x_a^m = 0 \\ 1 & \text{otherwise} \end{cases} \tag{4.9}$$

Es posible entonces agregar dos factores (uno para cada mano) a la ecuación 4.8 para penalizar esta situación:

$$P(x|c) = P(x^i|c)^{a_c^i} P(x_a^i|c) P(x^d|c)^{a_c^d} P(x_a^d|c) \tag{4.10}$$

4.5.2 Señas sin movimiento en alguna mano

Las señas donde una mano no realiza movimiento presentan un problema, sobre todo para al subclasificador de movimiento. Retomando el ejemplo de la figura 4.1, puede observarse que la mano izquierda del intérprete no realiza movimiento. Si una clase c realiza poco o ningún movimiento en una mano, entonces, la probabilidad de trayectoria $P(x^m_{tr}|c)$ no resulta representativa, y probablemente penalice al resto de las clases de manera casi aleatoria. Para evitar esta situación, se utilizó la media de la distribución de cantidad de movimiento por clase($\mu^m_{cm,c}$) calculada en etapa de entrenamiento, y se estableció el siguiente parámetro:

$$a^m_{c,mov} = \begin{cases} 1 & \text{if } \mu^m_{am,c} > 5cm \\ 0 & \text{if } \mu^m_{am,c} \leq 5cm \end{cases} \tag{4.11}$$

donde se estableció el umbral de $5cm$ de manera experimental. Este parámetro puede ser utilizado como exponente para $P(x^m_{tr}|c)$ para neutralizar la situación de las clases que no realizan o realizan muy poco movimiento. Se redefine la ecuación 4.6 como:

$$P(x^m_{mov}|c) = P(x^m_{tr}|c)^{a^m_{c,mov}} P(x^m_{cm}|c) \tag{4.12}$$

De este modo, si una seña no posee movimiento en una mano, se establece $a^m_{c,mov} = 0$ para esa clase y el modelo podrá ignorar la información de trayectoria.

4.6 Conclusiones del capítulo

Se definió un modelo probabilístico para la clasificación de gestos hechos con las manos que permite interpretar tres componentes principales: la posición de las manos, el movimiento que realizan, y sus configuraciones. El análisis del dominio del problema permitió incorporar la información relevante sobre los tres componentes de los diferentes gestos de un modo modular. Las diferentes partes del clasificador permiten desacoplar funcionalidad como podría ser el problema de tener una o dos manos en un gesto, o la necesidad o no de contemplar las configuraciones que las manos poseen. A su vez, sería posible agregar nuevos sub-clasificadores como por ejemplo información no manual sobre la expresión facial, de carácter importante en dominios como la lengua de señas.

Capítulo 5

Trabajos experimentales

En este capítulo se exponen las diferentes etapas de experimentación realizadas, las pruebas y los resultados obtenidos. En primera instancia, en la sección 5.1 se muestran algunos campos de aplicación donde se utilizó parte del clasificador propuesto, haciendo uso de las propiedades temporales de los descriptores. Dicha sección presenta un estudio específico en bases de datos de gestos realizados con todo el cuerpo, grabados con cámaras de profundidad (RGB-d). En la sección 5.2 se desarrollan los resultados principales de la tesis. Se divide la presentación esta sección en los resultados de clasificación de configuraciones (gestos estáticos) y clasificación de señas como una entidad léxica. Finalizando el capítulo se encuentran sus conclusiones.

5.1 Caso de estudio preliminar. Clasificación de acciones

Como trabajo preliminar al reconocimiento de lengua de señas, se llevo a cabo un desarrollo para clasificar gestos humanos hechos con todo el cuerpo. Se utilizaron dos bases de datos reconocidas en la literatura: *MSR Action3D* y *MSRC12 Dataset*. Estas bases de datos fueron obtenidas a través del dispositivo *MS Kinect* (ver capítulo 2). Este dispositivo captura los movimientos de una persona con sensores de profundidad, obteniendo información aceptablemente precisa sobre las articulaciones de todo el cuerpo. Los resultados de esta sección se encuentra publicados en [132].

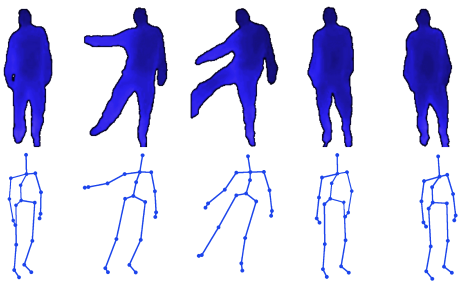


Figura 5.1: Un ejemplo de captura del dispositivo *MSKinect*. Arriba, imagen de profundidad de varios frames. Abajo, esqueleto con la información de las articulaciones.

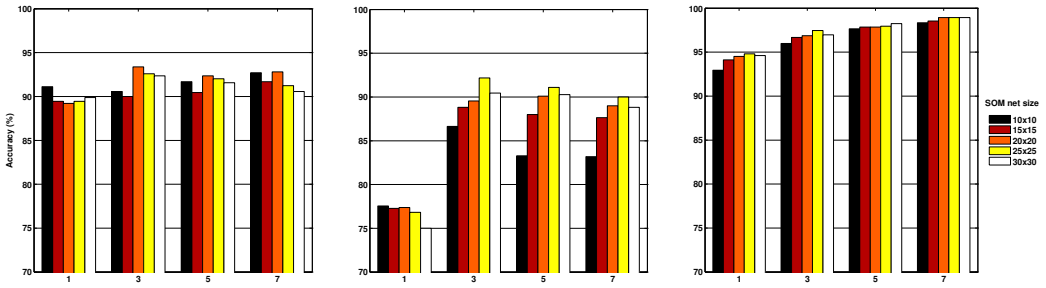
5.1.1 Adaptación del clasificador

Para trabajar en este dominio, el clasificador propuesto en esta tesis tuvo algunas modificaciones leves. En primer lugar, hay que destacar que a diferencia de la lengua de señas, aquí solo interesa la información del movimiento de las articulaciones. Por esta razón toda información sobre configuración de manos y posición resulta irrelevante en estas bases de datos. Por otro lado, ya que existe información sobre 20 articulaciones del cuerpo, ya no es posible modelar sólo la información de una mano. Aquí se concatenó la información de las 20 articulaciones en un sólo vector característica.

Por otro lado, debido a la alta dependencia temporal de los frames en estos gestos, se utilizó un sistema de ventanas para los descriptores de movimiento, de modo de tener fragmentos de información temporal por cada vector característica. De este modo se definió W como el tamaño de la ventana temporal a utilizar. Para cada base de datos, se probó con diversos valores de W para evaluar la importancia de incluir información temporal en el descriptor. La cantidad de percentiles utilizados fue ligeramente diferente que para la Lengua de Señas. Para la base *MSR Action3D* se utilizaron 25, mientras que para *MSRC12* fueron 16.

5.1.2 Resultados en *MSR Action3D Dataset*

La base de datos MSR Action3D ([95]) consta de 20 tipos de acciones diferentes ejecutadas por 10 sujetos distintos, con un total de 567 secuencias. Se excluyeron 10 secuencias que los autores originales quitaron por estar corruptas.



(a) AS1 (b) AS2 (c) AS3

Figura 5.2: Resultados obtenidos para MSR-Action3D, con diferentes valores de tamaño de ventana W, y diferentes tamaños de redes SOM. Cada resultados es el promedio de 30 ejecuciones independientes

Se utilizó una configuración típica en esta base de datos, que divide las acciones en tres grupos de datos llamados AS1, AS2 y AS3. Cada uno está compuesto con acciones de 8 clases, con alguna superposición de clases entre los conjuntos. Se ejecutaron las pruebas en cada conjunto por separado, siguiendo el mismo protocolo que los autores originales donde utilizan una validación cruzada por sujeto de $50/50$

Ventana	1	3	5	7
AS1	89.4 ($\pm$0.9)	92.6 ($\pm$1.4)	92.0 ($\pm$1.5)	91.2 ($\pm$1.5)
AS2	76.8 ($\pm$1.9)	92.2 ($\pm$1.6)	91.1 ($\pm$1.3)	90.7 ($\pm$1.0)
AS3	94.8 ($\pm$0.5)	97.4 ($\pm$0.7)	97.9 ($\pm$1.0)	98.9 ($\pm$0.7)
Media	**87.02**	**94.07**	**93.68**	**93.61**

Tabla 5.1: Precisión promedio del modelo en cada subconjunto de la base de datos MSR-Action 3D, para diferentes tamaños de ventanas W y una red SOM de 25×25 neuronas. La desviación estándar de las 30 ejecuciones independientes aparece entre paréntesis.

([95]). Se realizaron 30 ejecuciones independientes para cada configuración de parámetros cambiando los tamaños de ventanas y los tamaños de las redes SOM del clasificador.

La figura 5.2 muestra el comportamiento del método propuesto al variar los parámetros establecidos. Cada gráfico representa los resultados de un subconjunto (AS1, AS2 y AS3). Cabe notar la relevancia que tuvo los diferentes tamaños de ventanas para la correcta clasificación, siendo menos significativa la cantidad de neuronas utilizadas en la red SOM. La figura 5.3 resume los resultados mostrando el promedio de los 3 subconjuntos, para todas las configuraciones evaluadas. Por último, la tabla 5.1 muestra los resultados obtenidos en los 3 subconjuntos junto con la media, utilizando una red SOM de $25x25$ neuronas (siendo esta la mejor configuración obtenida). El mejor valor obtenido para la media de AS1, AS2 y AS3 fue con $W = 3$, aunque cabe destacar que $W = 5$ y $W = 7$ sobrepasan al segundo mejor método en la literatura (ver tabla 5.2).

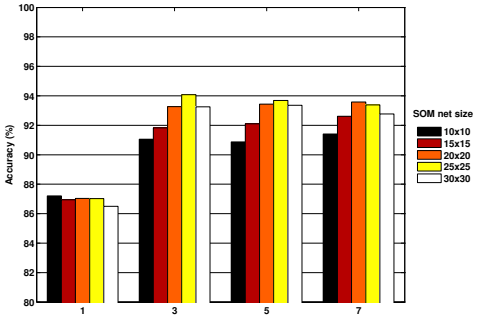


Figura 5.3: Accuracy promedio (*eje y*) en MSR Action 3D para diferentes tamaños de ventana W (*eje x*) y diferentes tamaños de redes SOM (colores). El promedio fue calculado sobre los resultados de los 3 subconjuntos de datos (AS1, AS2 y AS3).

La tabla 5.2 muestra una comparativa de resultados de este promedio de los diferentes autores del estado del arte en esta base de datos junto con los resultados obtenidos con el método propuesto en esta tesis. Con el descriptor propuesto se obtuvo una precisión de 94,07 %, reduciendo el error a $1/3$ del mejor resultado hasta el momento.

Un dato curioso sobre la figura 5.2 es que la precisión oscila fuertemente dependiendo el subconjunto del que se trate. Mientras que para AS3 se obtuvieron

Método	Accuracy (%)
Ellis (logistic regression) [48]	65.70
Li (Action Graph) [96]	74.70
Wang (Random Occupancy Pattern) [156]	86.50
Wang (Actionlets Ensemble) [157]	88.20
Hussein (Cov3DJ) [69]	90.53
Descriptor propuesto	**94.07**

Tabla 5.2: Comparación de resultados en la base de datos MSR-Action 3D.

mediciones cercanas al 99 %, en AS2 están por debajo de 90 %. Esto tiene que ver sin duda con la naturaleza de los gestos que hay en cada subconjunto. La figura 5.4 muestra la matriz de confusión para una ejecución individual en AS2. Aquí es claramente visible qué acciones retornan una peor eficacia en el método. Por ejemplo, acciones como *high arm wave* y *hand catch* son ejecutadas con el mismo grupo de articulaciones y son muy similares en movimiento. Otras acciones como *forward kick* o *two hand wave* son naturalmente más simples de distinguir ya que involucran diferentes articulaciones. Estos patrones de dificultad fueron también descriptos en trabajos previos como [157] o [156].

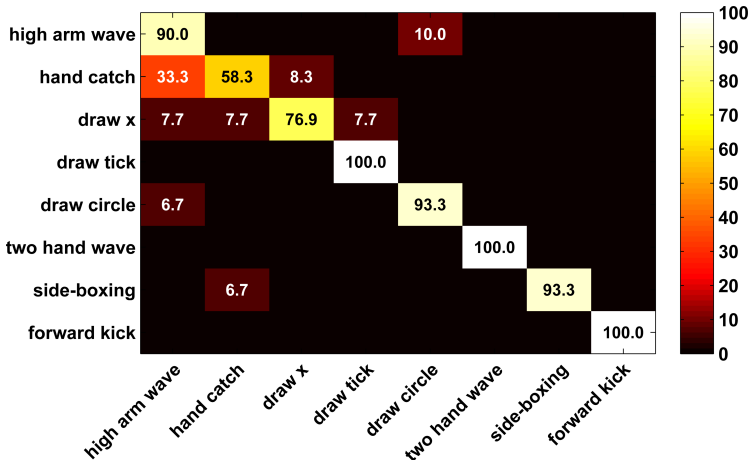
La figura 5.4 muestra también una matriz de probabilidades, resultado del clasificador ProbSOM (ver capítulo 4 para detalles). La matriz muestra la probabilidad promedio de que cada acción de la clase c sea clasificada como cualquier otra clase. La figura deja ver cuan robusto es el clasificador para las diferentes clases, dejando una amplia diferencia entre los posibles candidatos, para algunas clases como *forward kick* o *two hand wave*.

5.1.3 Resultados en *MSRC12 Dataset*

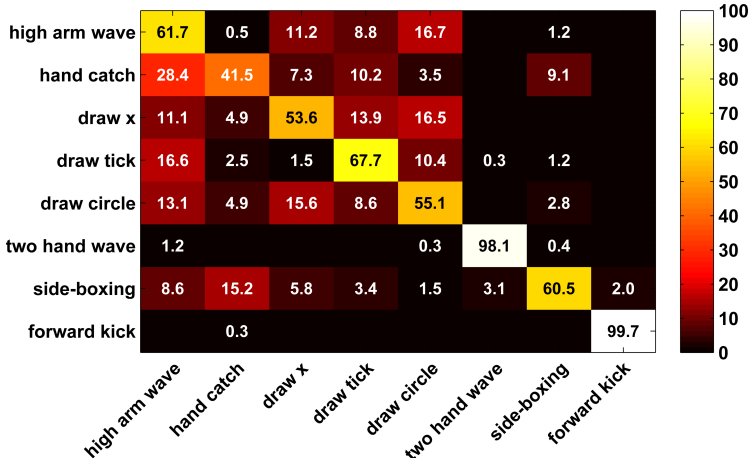
La base de datos MSRC12 fue creada por Fothergill et al. y descripta en [53]. Consiste en 594 secuencias de 12 acciones diferentes ejecutadas por 30 individuos. Cada secuencia posee diversas instancias de la misma clase. Por esta razón, es necesario primero hacer una segmentación adecuada. Siguiendo la especificación de [69] se llevó a cabo la segmentación para extraer cada acción por separado, obteniendo un total de 6244 acciones.

Al igual que con *MSR-Action 3D* se ejecutó una validación cruzada por sujeto con distribución $50/50$ para entrenamiento y testeo, y 30 ejecuciones independientes para que grupo de pruebas. La configuración de parámetros utilizada fue un tamaño de ventana $W = 5$ y una red SOM de 30×30 neuronas. También se llevó a cabo una validación cruzada dejando un sujeto fuera, siguiendo la especificación en [69].

La tabla 5.3 muestra los resultados obtenidos para MSRC-12 con el método propuesto, comparado con los métodos más recientes en el estado del arte, utilizando un test por validación cruzada. Los resultados son el promedio de 30 ejecuciones



(a) Matriz de confusión para todas las clases.



(b) Matriz de probabilidades promedio de asignar una acción de la clase c (filas) a cualquier otra clase (columnas)

Figura 5.4: Resultados de una ejecución independiente en AS2.

independientes. Si bien la precisión logró ser similar al método de Hussein[69], cabe notar la reducción en tamaño del clasificador propuesto, lo que permitiría la posibilidad de una evaluación en tiempo real.

La tabla 5.4 muestra los resultados obtenidos utilizando una validación cruzada dejando un sujeto fuera. Los resultados son el promedio de 30 ejecuciones independientes. Este método de validación permite verificar la eficacia del método en un escenario clásico, donde un nuevo sujeto trata de utilizar el sistema entrenado por otras personas. Con esta configuración, el método propuesto obtuvo una precisión de 93 %, resultados comparables a Negin et al., Hussein et al. y Jiang et al. Cabe notar que este último autor utilizó una validación no dejando un sujeto entero fuera, sino sólo un registro, lo que llevó a obtener una mejor eficacia que el resto.

En ambos tipos de pruebas, la varianza para la precisión promedio del método

Método	Precisión. (%)
Ellis (Logistic reg.)[48]	88.7
Hussein (Cov3DJ) [69]	91.7
Método propuesto	**91.7**

Tabla 5.3: Comparación de resultados del estado del arte para la base de datos *MSRC-12* con validación cruzada.

Método	Tipo de prueba	Precisión (%)
Hussein et al. (Cov3DJ) [69]	deja un sujeto fuera	93.6
Negin et al. (Decision Forest)[110]	deja un sujeto fuera	93.2
Jiang et al. (Hierarchical Model) [72]	deja un registro fuera	94.6
Método propuesto	deja un sujeto fuera	93.1

Tabla 5.4: Comparación de resultados para la base de datos *MSRC-12* utilizando validación cruzada dejando un sujeto fuera.

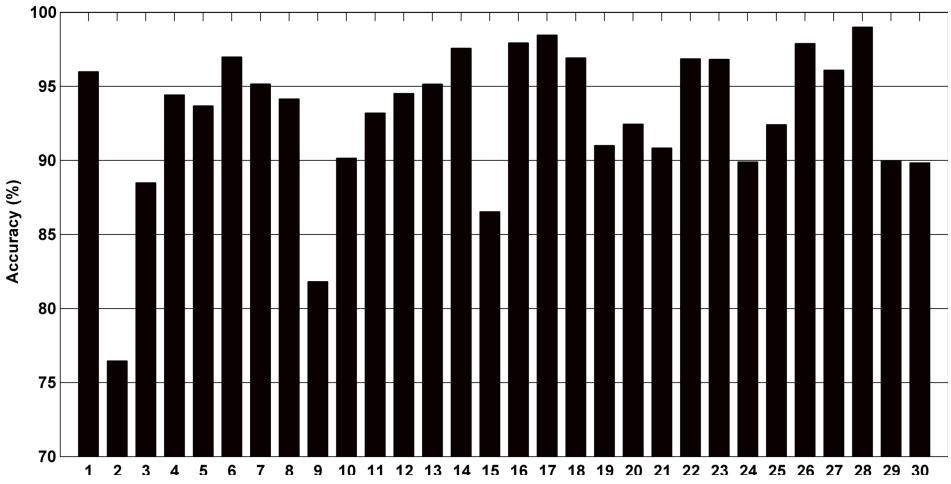


Figura 5.5: Precisión (*Accuracy*) por sujeto para la base de datos *MSRC-12* utilizando validación cruzada dejando un sujeto fuera.

propuesto fue relativamente alta ($\sigma^2 = 5$). No fue posible realizar un test de hipótesis estándar para las diferencias en las medidas con el resto de los autores debido a que dichos datos no se encontraron disponibles.

La figura 5.5 muestra la precisión promedio obtenido para cada sujeto, entrenando con el resto y dejándolo para validación. El gráfico muestra que existe una enorme variación en la eficacia del método dependiendo del sujeto con el que se validó. Esto posiblemente se deba a que los distintos sujetos en esta base de datos fueron instruidos de diversas maneras. A algunos se les dio un video para que imiten el gesto a realizar. A otros se les mostró un texto o una imagen. Por ejemplo, el sujeto 28 tiene una precisión casi perfecta, cercano a 100 %, mientras que en el otro extremo, el sujeto 2 tiene una precisión de tan solo 76 %. Estos resultados permiten mostrar el por qué de la varianza relativamente alta que se encontró en el párrafo anterior.

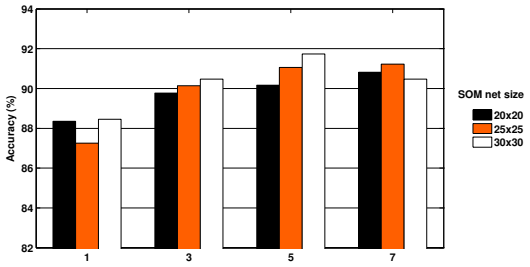
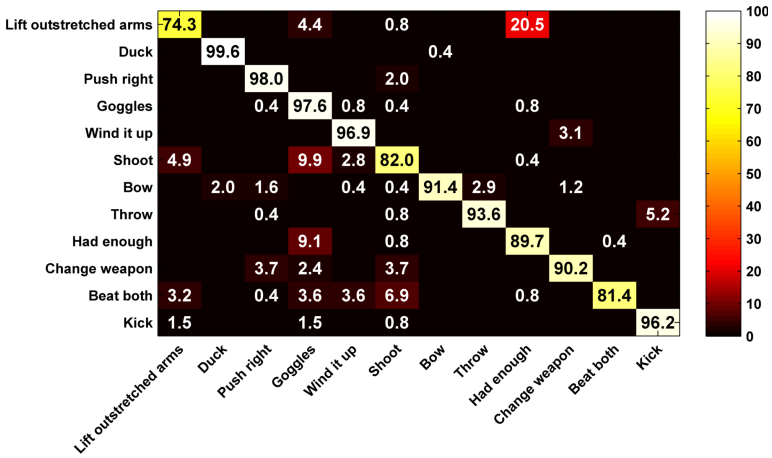


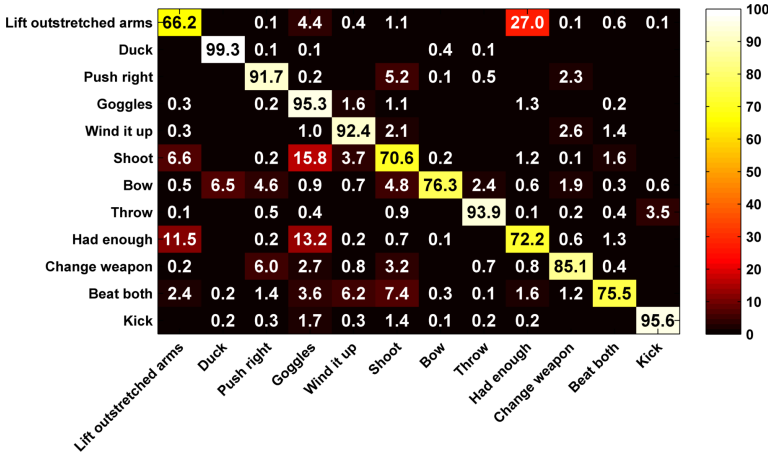
Figura 5.6: Precisión (*Accuracy*) promedio para la base de datos *MSRC-12* con diferentes tamaños de ventanas W y diferentes tamaños de redes SOM.

La figura 5.6 muestra el comportamiento del método utilizando diferentes tamaños W de ventanas y diferentes tamaños de redes SOM. Al igual que lo ocurrido para *MSR-Action 3D*, el gráfico muestra la clara ventaja de utilizar el sistema de ventanas, en vez de sólo el vector de direcciones. Por otro lado, la variación del tamaño del cluster no parece ser significativa. Ya que el éxito del método ProbSOM se determina en última instancia por lo bien que se realizó la estimación de la distribución de las ventanas temporales para cada clase, la cantidad de clusters (neuronas) en realidad depende en gran medida de la cantidad de muestras en el conjunto de datos. Por esta razón fue necesario una red más grande que para *MSR-Action 3D*.

Al igual que para *MSR-Action 3D*, la figura 5.7 muestra la matriz de confusión y matriz de probabilidades luego de una ejecución particular en *MSRC-12*. Algunas acciones como *Duck* o *Push Right* fueron clasificadas casi perfectamente, alcanzando también una probabilidad muy alta, lo que presenta un ejemplo de la robustez del modelo propuesto. Otras clases como *Lift outstretched arms* o *Shoot* presentan más dificultades, presuntamente debido a que usan las mismas articulaciones y movimientos similares. No obstante, la matriz de confusión en 5.7a muestra que los errores de clasificación involucran mayormente dos o tres clases al mismo tiempo. Una segunda capa de clasificación, quizá podría resolver este problema, entrenándose específicamente para diferenciar entre dichas clases.



(a) Matriz de confusión para todas las clases.



(b) Matriz de probabilidades promedio de asignar una acción de la clase c (filas) a cualquier otra clase (columnas)

Figura 5.7: Resultados de una ejecución para la base de datos *MSRC-12*.

5.2 Clasificación de señas en la Lengua de Señas Argentina

A continuación se describen los experimentos llevados a cabo sobre las bases de datos LSA16 de configuraciones de manos y LSA64 de señas de la Lengua de Señas Argentina.

5.2.1 Clasificación de configuraciones de manos

En la sección 4.2 se describió el método definido para clasificar configuraciones de manos, basado en el PromSOM, utilizando como descriptor la transformada de Radon y vectores SIFT. En esta sección se presentan los resultados de la experimentación llevada a cabo con la base de datos LSA16 de configuraciones (ver capítulo 3). Los resultados de esta sección pueden verse publicados en [133].

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

Tabla 5.5: Precisión del modelo para la base de datos LSA16 de configuraciones utilizando validación cruzada aleatoria, con 30 pruebas independientes. 90 % de imágenes para entrenamiento, y 10 % para validación.

Se realizaron diversas evaluaciones tanto con el método propuesto, como con algunos métodos del estado del arte. En el caso del ProbSom, se realizaron pruebas con los descriptores SIFT y Radon. Además, para el descriptor basado en Radon, se realizaron pruebas con Máquinas de Soporte Vectorial (SVM), Random Forest, y Feedforward Neural Networks. [1]. En los casos en que los métodos a comparar se comportan con distinta performance dependiendo de sus parámetros internos, se reportan la mejor configuración.

La precisión del modelo (*accuracy*) se calculó como el porcentaje de ejemplos reconocidos correctamente sobre el total de cada clase. La tabla 5.5 muestra los resultados obtenidos bajo validación cruzada aleatoria estratificada con $n = 30$ repeticiones independientes, utilizando 90 % de las imágenes para entrenar y 10 % para validar. Los resultados muestran una performance comparable del ProbSOM frente a otras técnicas de clasificación. Por otro lado, los descriptores de Radón mostraron ser mucho más representativos que los vectores SIFT. Esto puede deberse a que generalmente los descriptores SIFT buscan puntos con información particular, para luego realizar *template matching* (comparación de plantillas) de imágenes, o describir una situación particular. En las imágenes de LSA16 existen diversos puntos muy similares (como las puntas de los dedos) que resultan comunes a muchas clases, lo que dificulta la utilización de SIFT como se había utilizado en [92] para reconocer rostros, utilizando un modelo de clasificación similar. La figura 5.8 muestra un ejemplo de clasificación con el modelo ProbSOM entrenado. En la ilustración puede versa la probabilidad de cada una de las 16 clases de la base de datos.

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 %($\pm$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

[1] Se realizaron además pruebas con descriptores de Fourier, Banco de filtros de Gabor, Local Binary Patterns (no descriptos en la tesis) con resultados inferiores en casi todos los casos a los presentados.

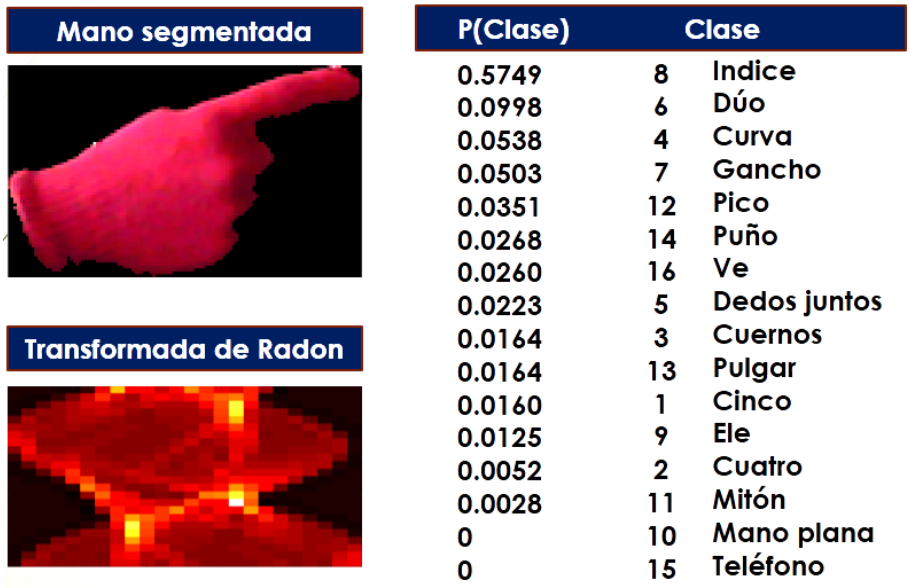


Figura 5.8: Ejemplo de clasificación de una mano segmentada con el modelo Prom-SOM entrenado.

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. La figura 5.9 muestra los resultados obtenidos para cada individuo de la base de datos.

En esta sección se mostró que los descriptores utilizados junto con el modelo de clasificación mostraron ser robustos en la clasificación de las configuraciones de manos en LSA16, incluso con una validación inter-sujeto, dando la posibilidad de incorporar un nuevo individuo desconocido por el sistema. Por otro lado, cabe destacar que la tasa de acierto es similar en todas las clases de la base de datos.

Ya que el ProbSOM funciona de modo probabilístico realizando un ranking de posibles clases candidatas, resulta interesante observar qué ocurre con las imágenes clasificadas erróneamente por el sistema. Si se observa el orden generado por el modelo y la tasa de acierto se obtiene considerando como clasificación correcta tanto a la primer o a la segunda opción, la tasa de acierto general sube de 92,25 % a 96,6 %. Esto demuestra que el modelo, en casi todos los ejemplos de validación, la confusión es entre sólo dos clases. Esto resulta muy interesante si el modelo funciona como un diccionario, ya que podría utilizarse la probabilidad del modelo para mostrar una o dos posibilidades. Del mismo modo, podría volverse a aplicar un clasificador más específico para solucionar la ambigüedad en las situaciones que lo requieran.

5.2.2 Clasificación de señas segmentadas

En esta sección se presentan los experimentos llevass a cabo y los resultados obtenidos con la base de datos LSA64 de señas segmentadas (ver capítulo 3). Los des-

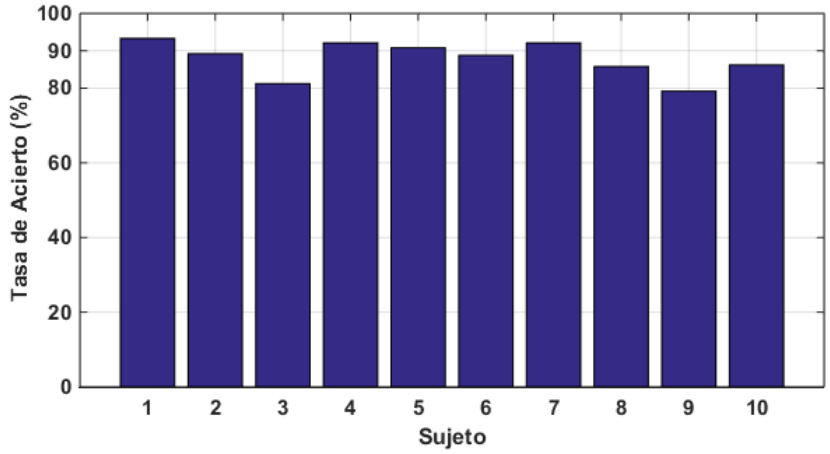


Figura 5.9: Validación cruzada inter-sujeto para LSA16.

	Todos	Config	Mov	Pos	Config-Pos	Config-Mov	Pos-Mov	Todos-HMM	Todos-BF-SVM
μ	97.44	52.97	54.03	76.05	94.91	83.59	84.84	95.92	95.08
σ	0.59	1.74	1.71	0.62	0.52	0.87	0.90	0.95	0.69

Tabla 5.6: Resultados de los experimentos llevados a cabo con los diferentes descriptores propuestos, utilizando 80 % para entrenamiento y 20 % para validación. En cada columna se indican los diferentes descriptores utilizados (y su correspondiente subclasificador). Las últimas dos columnas representan los resultados de utilizar Modelos Ocultos de Markov con los descriptores propuestos, y los descriptores binarios (ver [74]) utilizando Máquinas de Soporte Vectorial.

criptores y métodos utiilizados son los descriptos en el capítulo 4. Los resultados de esta sección pueden verse publicados en [131]. Para realizar todos los experimentos se utilizó una configuración $80/20$ para entrenamiento/validación respectivamente, con $n = 30$ ejecuciones independientes.

La tabla 5.6 muestra los resultados de la experimentación llevada a cabo para los diferentes descriptores propuestos junto con sus respectivos subclasificadores. Los tres componentes principales definidos previamente son: la configuración de la mano en el tiempo, el movimiento realizado, y las posiciones de las manos. La tabla muestra el comportamiento de clasificador al utilizar diferentes configuraciones de estos descriptores/clasificadores. Las diferentes pruebas llevadas a cabo muentran 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. Cabe apreciar la relativamente alta tasa de acierto al utilizar sólo la componente de posición. Esto puede deberse a la distribución particular de la base de datos utilizada, como se vio en el capítulo 3. Por otro lado, la suma de las componentes de posición junto con

la de configuración garantizan una tasa de acierto de casi 95 %, lo que demuestra la importancia de las distintas configuraciones utilizadas en la lengua de señas. Por último, al utilizar las tres componentes del clasificador, se obtuvo una tasa de acierto del 97,4 %.

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) [20]. En este caso se entrenó un modelo para cada clase y cada descriptor, utilizando un algoritmo *EM* [20]. Cada modelo es un modelo izquierda-derecha con 4 estados en todos los casos. 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 escencial 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 [74]. Estos descriptores analizan la información de un modo análogo al propuesto en esta tesis definiendo diferentes tipos de movimientos y posiciones donde las manos pueden estar, al igual que diferentes configuraciones. Luego, se genera un descriptor binario temporal, indicando para cada frame las diferentes situaciones que ocurren con las manos. Para definir si las manos coinciden con un determinado patrón se establecen reglas determinísticas del estilo "si la mano izquierda se mueve hacia arriba y la mano derecha hacia abajo, entonces clase X". Las diferentes matrices generadas por estos descriptores son de tamaño $numdescriptor \times cantframes$. Estas fueron reformateadas para tener un tamaño de $frames = 32$, de modo que todos los descriptores posean el mismo tamaño para cada seña. La matriz resultante como descriptor en este caso fue utilizada como entrada para un clasificador basado en Máquinas de Soporte Vectorial (SVM), con una configuración de uno-contra-todos con un kernel lineal. En este caso se obtuvo una tasa de acierto de 95 %. La figura 5.10 muestra un ejemplo de este tipo de descriptor.

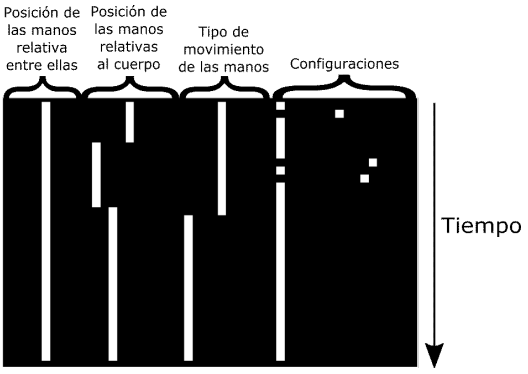


Figura 5.10: Ejemplo de descriptor propuesto por Kadir y Bowden, para una seña ([74]).

La figura 5.11 muestra la matriz de confusión de una ejecución independientes para la base de datos LSA64. No obstante la alta tasa de acierto obtenida, existen algunos patrones visibles en la matriz. Por ejemplo, el modelo confunde las clases 24 y 26, lo que podría resultar evidente al observar que poseen similar movimiento y

Sub-conjunto	1 Mano	2 Manos	Media
μ	95.93	99.09	97.01
σ	1.31	0.77	-

Tabla 5.7: Precisión del modelo para los sub-conjuntos de la base de datos de señas con una mano y dos manos para la base de datos LSA64, junto con el promedio de los resultados.

posiciones, diferenciándose tan sólo en la configuración. La figura 5.12 muestra la misma matriz de confusión pero luego de entrenar el modelo sin el subclasificador de configuraciones. Aquí se hace más evidente aún la confusión entre las clase 24 y 26, ya que sin la información de configuración resulta imposible diferenciarlas. Algo similar ocurre entre las clases 2 y 3 con un trayecto y ubicación idénticos. Esta matriz muestra con un ejemplo la importancia del rol que cumple cada uno de los tres subclasificadores propuestos.

Comparación de resultados para señas con una y dos manos

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 incoprporando 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. La tabla 5.7 muestra los resultados obtenidos luego de 30 ejecuciones independientes. 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.

Pruebas independientes del sujeto (*cross-subject validation*)

Al igual que los resultados presentados en secciones previas, se describe aquí una serie de experimentos realizados 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 5.8 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

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

Tabla 5.8: Validación independiente al sujeto en la base de datos LSA64. Cada columna muestra la precisión media y desviación estándar cuando se valida con cada sujeto, entrenando el sistema con los otros nueve. La columna final muestra la media de todos los experimentos.

acierto decae con respecto a los resultados generales. No obstante la precisión media conseguida fue de $91{,}7\,\%(\pm 0{,}8)$, mostrando excelentes resultados al introducir un nuevo individuo al sistema.

5.3 Conclusiones del capítulo

Durante el desarrollo del capítulo se mostraron los resultados de la experimentación llevada a cabo para validar la eficacia del modelo propuesto. Se realizó un estudio preliminar basado en reconocimiento de acciones humanas con sensores de profundidad. En este caso, hubo que hacer una adaptación del modelo propuesto debido a la diferencia en el dominio de aplicación. Los resultados mostraron ser satisfactorios, comparables a otros del estado del arte y con mejoras en eficiencia.

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.

También se realizaron evaluaciones intercambiando los clasificadores parciales de movimiento y configuración para comparar la eficacia del clasificador ProbSOM por técnicas convencionales como Modelos Ocultos de Markov. Estos resultados mostraron que ambas estrategias se comportan de modo similar. Sin embargo, al ser el ProbSOM un clasificador de tipo bolsa de palabras (*bag-of-words*), los patrones de entrada son barajados, lo que demuestra la no necesidad de tener un orden determinado sobre estos datos, como sí necesitan los Modelos de Markov.

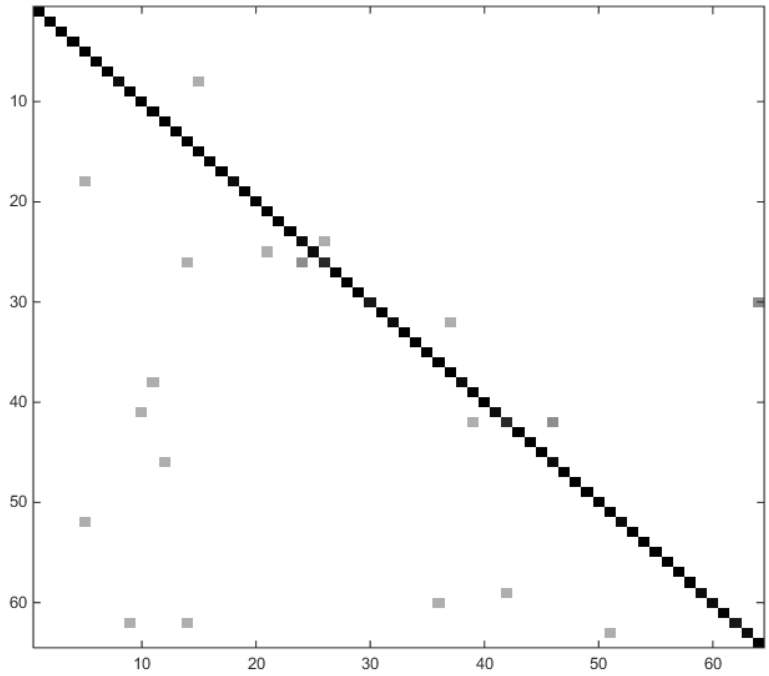


Figura 5.11: Matriz de confusión de una ejecución independientes con la base de datos LSA64.

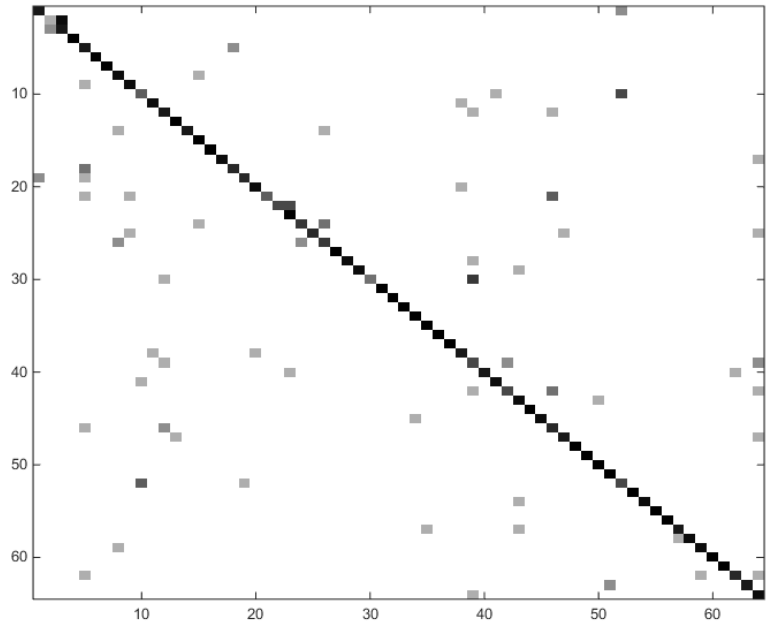


Figura 5.12: Matriz de confusión de una ejecución independientes con la base de datos LSA64 sin el subclasificador de configuracion.

Capítulo 6

Conclusiones y trabajos futuros

6.1 Conclusiones

El reconocimiento automático de gestos dinámicos y particularmente de las lenguas de señas se presentan como un problema multidisciplinar sumamente complejo. Lejos se está aún de resolverlo de forma completa. No obstante ha habido numerosos avances en lo que respecta a resoluciones parciales o de ciertas etapas.

Los diferentes métodos de captura existentes abren un abanico de posibilidades, cada uno con pros y contras. Mientras que las cámaras de profundidad prometen ser una tecnología evolucionada para sensar la posición de distintas partes del cuerpo humano en un video, siguen siendo poco eficientes, mostrando ciertas falencias a la hora de tener información precisa. Las cámaras tradicionales RGB son una herramienta más económica y accesible para el público en general. Además, existen numerosos trabajos realizados bajo estos paradigmas. Esto hace a esta tecnología una de las más estudiadas aún hoy para el reconocimiento de señas.

No obstante los diferentes avances existentes en tópicos como el reconocimiento de rostros, seguimiento de manos, o incluso reconocimiento de un gesto completo segmentado, la integración de todos los elementos propuestos para llevar a cabo un reconocedor automático robusto no es una tarea sencilla. Llevar a cabo un reconocimiento de lengua de señas desde una cámara de video supone una cadena de procesamiento sumamente compleja en la cual es necesario incluir:

1. Reconocimiento y ubicación de partes del cuerpo como manos, rostro, etc.
2. Segmentación de las partes ubicadas con el fin de obtener información relevante.
3. Trasformar la información capturada con el fin de obtener descriptores adecuados que puedan ser utilizados luego como entrada en un clasificador.
4. Clasificar los patrones obtenidos para distinguir una seña.

5. Incorporar gramática del lenguaje con el fin de mejorar la traducción.

Cada tarea mencionada es un desafío en sí mismo y pueden ser abordadas de diferentes modos. Diversos artículos presentan partes de estas tareas resueltas en forma parcial o total. Algunos trabajos proponen métodos completos de reconocimiento donde cada etapa es resuelta con ciertos métodos desarrollados previamente. Generalmente, los trabajos más recientes se focalizan en el modo de interpretar y clasificar la información temporal propia de un gesto/seña, obviamente etapas iniciales supuestas como ya resueltas. Por ejemplo, la detección de un rostro es una tarea resuelta de un modo bastante efectivo hoy en día, siendo posible saltear este paso con sólo algunos detalles. Sin embargo la detección y segmentación de las manos de un sujeto en un video sigue siendo una tarea sumamente compleja y resuelta sólo de modo parcial. Algunos trabajos proponen filtros de color para detectar la piel como indicador de dónde están ubicadas las manos del intérprete. Esto trae aparejado el problema adicional de distinguir luego el rostro de las manos, y las manos entre sí.

El modo de capturar la información y generar descriptores apropiados está relacionado también con la naturaleza de la base de datos. En la actualidad no son demasiadas las bases de datos existentes de lengua de señas para procesos de aprendizaje automático. Las más completas poseen miles de gestos. Sin embargo, ninguna de estas grandes bases de datos posee información precisa del movimiento de las manos. Esto conlleva la dificultad de la etapa de reconocimeinto y segmentación de las manos, que como se dijo anteriormente, no es un proceso trivial. Por otro lado, debido a la naturaleza plurilingüe del lenguaje de señas, para realizar un traductor regional es necesario contar con datos representativos de su léxico.

La base de datos desarrollada en esta tesis, 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 intepretando cada señas 5 repeticiones distintas, dando un total de 3200 videos de alta resolución. Sumado a esto, la base de datos LSA16 continee 800 imágenes de configuraciones de manos del léxico argentino.

En cuanto a los métodos de clasificación existen diferentes aproximaciones, dependiendo la información utilizada para describir un gesto o seña. Cuando se trata de gestos estáticos como detectar una sonrisa o la pose de una mano generalmente se cuenta con información estática pudiendo utilizarse muchos de los algoritmos clásicos de aprendizaje automático como redes neuronales artificiales, algoritmos de clustering, lógica difusa, etc. En cambio, al tratarse de gestos dinámicos como lo que implica el lenguaje de señas o reconocimiento de acciones humanas en general, la información obtenida tiene un dominio naturalmente temporal. Aquí existen diferentes aproximaciones para interpretar esta información. La mayor parte de los trabajos utilizan Redes Ocultas de Markov (*HMM*) o Deformaciones Dinámicas de

Tiempo (*DTW*) debido a su origen propio del procesamiento temporal que poseen. Otras estrategias, como la presentada en esta tesis, se basan en la correcta transformación de los datos para general descriptores que por sí solos posean la información temporal inherente al problema.

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. Se mostraron estudios preliminares donde se utilizaron los mismos descriptores de movimiento para clasificar gestos humanos capturados con una cámara de profundidad. En dicho trabajo se utilizó una red SOM probabilística (*ProbSOM*) para clasificar los distintos gestos de dos bases de datos existentes. Luego, este mismo esquema de descriptores y clasificador se utilizaron como módulo de clasificación parcial para identificar el movimiento de una seña.

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.

Todos los resultados fueron evaluados satisfactoriamente en la base de datos LSA64 desarrollada en esta tesis. Diversos módulos fueron intercambiados por otros para mostrar solidez en los resultados, incluyendo evaluaciones con trabajos de otros autores.

6.2 Trabajos futuros

Existen diversas lineas de investigación que quedan abiertas luego de la finalización de esta tesis:

- 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. En ocasiones, partes de la solución

planteada en un trabajo resultan irreproducibles en bases de datos diferentes a las utilizadas. Por ejemplo, muchos trabajos suponen un esquema de grabación de los intérpretes con un fondo oscuro y sin otros contrastes más allá de la piel del intérprete. Sin embargo existen grandes bases de datos con grabaciones que incluyen vestimentas de colores fuertes en los intérpretes, o un fondo con ambiente real. Por esta razón al utilizar un trabajo existente en la bibliografía es necesario ser cuidadoso en todo el proceso propuesto por los autores. Sumado a esta complejidad está el hecho de que pocos trabajos disponen de código disponible para su evaluación. Uno de los desafíos sin resolver en esta tesis está tanto en evaluar el método propuesto en otras bases de datos existentes como en aumentar el número de métodos del estado del arte para evaluar la base de datos LSA64.

- 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 gramática propia del lenguaje para no tener únicamente un traductor léxico sino un completo traductor del lenguaje de señas.
- 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.
- Por último, implementar el modelo propuesto en un entorno de aplicación real, con el fin de realizar una transferencia concreta para facilitar la traducción de la lengua de señas a personas hipoacúsicas. Esto involucra un trabajo de investigación en sistemas de tiempo real no contemplado en esta tesis. Este punto sería de sumo interés para instituciones como la universidad, debido a la incorporación reciente de personas con estas características y el interés de la institución por mejorar la accesibilidad.

Bibliografía

[1] Confederación argentina de de sordos. *http://cas.org.ar/*, consultado el 10/11/2016.

[2] Manos que hablan. *http://manosquehablan.com.ar/*, consultado el 15/11/2016.

[3] Cyberglove iii. *http://www.cyberglovesystems.com/cyberglove-iii/*, consultado el 22/11/2016.

[4] Mit anthrotronix acceleglove. *https://www.technologyreview.com/s/414021/open-source-data-glove/*, consultado el 22/11/2016.

[5] Y. F. Admasu and K. Raimond. Ethiopian sign language recognition using artificial neural network. In *2010 10th International Conference on Intelligent Systems Design and Applications*, pages 995–1000, Nov 2010.

[6] J.K. Aggarwal and M.S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16:1–16:43, April 2011.

[7] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3d pose estimation and tracking by detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010.

[8] J. Appenrodt, S. Handrich, A. Al-Hamadi, and B. Michaelis. Multi stereo camera data fusion for fingertip detection in gesture recognition systems. In *2010 International Conference of Soft Computing and Pattern Recognition*, pages 35–40, 2010.

[9] Jörg Appenrodt, Ayoub Al-Hamadi, Mahmoud Elmezain, and Bernd Michaelis. Data gathering for gesture recognition systems based on mono color-, stereo color- and thermal cameras. In *Proceedings of the 1st International Conference on Future Generation Information Technology*, FGIT '09, pages 78–86, Berlin, Heidelberg, 2009. Springer-Verlag.

[10] Mark Aronoff, Irit Meir, and Wendy Sandler. The paradox of sign language morphology. *Language*, 81(2), 2005.

[11] Marcell Assan and Kirsti Grobel. Video-based sign language recognition using hidden markov models. In *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 97–109, London, UK, UK, 1998. Springer-Verlag.

[12] Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. *The American Sign Language Lexicon Video Dataset*. 2008.

[13] A. L. C. Barczak, N. H. Reyes, M. Abastillas, A. Piccio, and T. Susnjak. A new 2d static hand gesture colour image dataset for asl gestures. *Research Letters in the Information and Mathematical Sciences*, 15:12–20, 2011.

[14] K. Barczewska and A. Drozd. Comparison of methods for hand gesture recognition based on dynamic time warping algorithm. In *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*, pages 207–210, Sept 2013.

[15] I. L. O. Bastos, M. F. Angelo, and A. C. Loula. Recognition of static gestures applied to brazilian sign language (libras). In *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 305–312, Aug 2015.

[16] B. Bauer and K. F. Kraiss. Video-based sign recognition using self-organizing subunits. In *Object recognition supported by user interaction for service robots*, volume 2, pages 434–437 vol.2, 2002.

[17] Britta Bauer. *Erkennung kontinuierlicher Gebärdensprache mit Untereinheiten-Modellen*. PhD thesis, RWTH Aachen University, 2004.

[18] Sigal Berman and Helman Stern. Sensors for gesture recognition systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(3):277–290, May 2012.

[19] Daniel Betancur Betancur, Mateo Vélez Gómez, and Alejandro Peña Palacio. Traducción automáica del lenguaje dactilológico de sordos y sordomudos mediante sistemas adaptativos. *Revista Ingeniería Biomédica*, 7:18–30, 06 2013.

[20] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[21] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[22] Patrick Buehler, Andrew Zisserman, and Mark Everingham. Learning sign language by watching tv (using weakly aligned subtitles). In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[23] Necati Cihan Camgöz, Ahmet Alp Kındıroğlu, Serpil Karabüklü, Meltem Kelepir, Ayşe Sumru Özsoy, and Lale Akarun. Bosphorussign: A turkish sign language recognition corpus in health and finance domains. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).

[24] Nina C. Capone and Karla K. Mcgregor. Gesture development: A review for clinical and research practices. *Journal of Speech, Language & Hearing Research*, 47(1):173–186, 2004.

[25] James Charles, Tomas Pfister, Mark Everingham, and Andrew Zisserman. Automatic and efficient human pose estimation for sign language videos. 2013.

[26] Feng-Sheng Chen, Chih-Ming Fu, and Chung-Lin Huang. Hand gesture recognition using a real-time tracking method and hidden markov models. *Image and Vision Computing*, 21(8):745 – 758, 2003.

[27] Kyunghyun Cho and Xi Chen. Classifying and visualizing motion capture sequences using deep neural networks. *CoRR*, abs/1306.3874, 2013.

[28] S. Conseil, S. Bourennane, and L. Martin. Comparison of fourier descriptors and hu moments for hand posture recognition. In *Proceedings of the European Signal Processing Conference (EUSIPCO), Poznan, Poland*, pages 1960–1964, 2007.

[29] 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.

[30] 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.

[31] Helen Cooper, Nicolas Pugeault, and Richard Bowden. Reading the signs: A video based sign dictionary. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 914–919, Nov 2011.

[32] T. F. Cootes, G. J. Edwards, and C. J. Taylor. *Active appearance models*, pages 484–498. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[33] T.F. Cootes and C.J. Taylor. *Statistical Models of Appearance for Computer Vision*. Tech. Rep., University of Manchester. 2000.

[34] Andrea Corradini. Dynamic time warping for off-line recognition of a small gesture vocabulary. In *In Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in RealTime Systems (RATFG-RTS'01*, page 82. IEEE Computer Society, 2001.

[35] O. A. Crasborn and I Zwitserlood. The corpus ngt: An online corpus for professionals and laymen. In *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages. ELDA, Paris*, pages 44–49, 2008.

[36] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.

[37] Ministerio de Cultura y Educación de la Nación. *Diccionario Lengua de Señas Argentina-español*. Dirección de equipo técnico, Graciela Alisedo, 1997.

[38] Thomas G. Dietterich. Machine learning for sequential data: A review. In *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, pages 15–30, London, UK, UK, 2002. Springer-Verlag.

[39] Mark Dilsizian, Polina Yanovich, Shu Wang, Carol Neidle, and Dimitris N Metaxas. A new framework for sign language recognition based on 3d handshape identification and linguistic modeling. In *LREC*, pages 1924–1929, 2014.

[40] L. Dipietro, A. M. Sabatini, and P. Dario. A survey of glove-based systems and their applications. *Trans. Sys. Man Cyber Part C*, 38(4):461–482, 2008.

[41] T. D'Orazio, R. Marani, V. Renò, and G. Cicirelli. Recent trends in gesture recognition: how depth data has improved classical approaches. *Image and Vision Computing*, 52:56 – 72, 2016.

[42] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. Tracking using dynamic programming for appearance-based sign language recognition. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 293–298, 2006.

[43] Philippe Dreuw, Thomas Deselaers, Daniel Keysers, and Hermann Ney. Modeling image variability in appearance-based gesture recognition. In *ECCV Workshop on Statistical Methods in Multi-Image and Video Processing*, pages 7–18, Graz, Austria, May 2006.

[44] Philippe Dreuw, Jens Forster, Thomas Deselaers, and Hermann Ney. Efficient approximations to model-based joint tracking and recognition of continuous sign language. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–6, Amsterdam, The Netherlands, September 2008.

[45] Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi, and Hermann Ney. Speech recognition techniques for a sign language recognition system. In *Interspeech*, pages 2513–2516, Antwerp, Belgium, 2007. ISCA best student paper award Interspeech 2007.

[46] N. El-Bendary, H. M. Zawbaa, M. S. Daoud, A. E. Hassanien, and K. Nakamatsu. Arslat: Arabic sign language alphabets translator. In *2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM)*, pages 590–595, Oct 2010.

[47] Ahmed Elgammal, Crystal Muang, and Dunxu Hu. Skin detection – a short tutorial, 2009.

[48] Chris Ellis, Syed Zain Masood, Marshall F. Tappen, Joseph J. Laviola, Jr., and Rahul Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *Int. J. Comput. Vision*, 101(3):420–436, February 2013.

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

[50] Jens Forster, Oscar Koller, Christian Oberdörfer, Yannick Gweth, and Hermann Ney. Improving continuous sign language recognition: Speech recognition techniques and system design. In *Workshop on Speech and Language Processing for Assistive Technologies*, pages 41–46, Grenoble, France, 2013. Satellite Workshop of INTERSPEECH 2013.

[51] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater, and Hermann Ney. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *Language Resources and Evaluation*, pages 3785–3789, Istanbul, Turkey, May 2012.

[52] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the sign language recognition and translation corpus rwth-

phoenix-weather. In *Language Resources and Evaluation*, pages 1911–1916, Reykjavik, Island, May 2014.

[53] Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. Instructing people for training gestural interactive systems. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 1737–1746. ACM, 2012.

[54] A. Gangopadhyay, O. Chatterjee, and A. Chatterjee. Hand shape based biometric authentication system using radon transform and collaborative representation based classification. In *Image Information Processing (ICIIP), 2013 IEEE Second International Conference on*, pages 635–639, Dec 2013.

[55] Serkan Genç, Muhammet Baştan, Uğur Güdükbay, Volkan Atalay, and Özgür Ulusoy. Handvr: a hand-gesture-based interface to a video retrieval system. *Signal, Image and Video Processing*, 9(7):1717–1726, 2015.

[56] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.

[57] G. H. Granlund. Fourier preprocessing for hand print character recognition. *IEEE Transactions on Computers*, C-21(2):195–201, 1972.

[58] Haiying Guan, R. S. Feris, and M. Turk. The isometric self-organizing map for 3d hand pose estimation. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 263–268, 2006.

[59] Y. L. Gweth, C. Plahl, and H. Ney. Enhanced continuous sign language recognition using pca and neural network features. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 55–60, 2012.

[60] Simon Hadfield and Richard Bowden. *Generalised Pose Estimation Using Depth*, pages 312–325. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[61] J. Han, G. Awad, and A. Sutherland. Automatic skin segmentation and tracking in sign language recognition. *IET Computer Vision*, 3(1):24–35, March 2009.

[62] Jungong Han, Ling Shao, Dong Xu, and J. Shotton. Enhanced Computer Vision With Microsoft Kinect Sensor: A Review. *Cybernetics, IEEE Transactions on*, 43(5):1318–1334, 2013.

[63] Junwei Han, George Awad, and Alistair Sutherland. Modelling and segmenting subunits for sign language recognition based on hand motion analysis. *Pattern Recogn. Lett.*, 30(6):623–633, April 2009.

[64] Haitham Hasan and Sameem Abdul-Kareem. Human–computer interaction using vision-based hand gesture recognition systems: a survey. *Neural Computing and Applications*, 25(2):251–261, 2014.

[65] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998.

[66] Eun-Jung Holden, Gareth Lee, and Robyn Owens. Australian sign language recognition. *Machine Vision and Applications*, 16(5):312–320, 2005.

[67] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187, 1962.

[68] Shah Muhammed Abid Hussain and A. B. M. Harun ur Rashid. User independent hand gesture recognition by accelerated dtw. In *2012 International Conference on Informatics, Electronics Vision (ICIEV)*, pages 1033–1037, May 2012.

[69] Mohamed E Hussein, Marwan Torki, Mohammad A Gowayyed, and Motaz El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2466–2472. AAAI Press, 2013.

[70] Moustafa Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[71] Jhilmil Jain, Arnold Lund, and Dennis Wixon. The future of natural user interfaces. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 211–214, New York, NY, USA, 2011. ACM.

[72] Xinbo Jiang, Fan Zhong, Qunsheng Peng, and Xueying Qin. Robust action recognition based on a hierarchical model. *2013 International Conference on Cyberworlds*, pages 191–198, 2013.

[73] Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002.

[74] 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.

[75] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[76] Byeongkeun Kang, Subarna Tripathi, and Truong Q. Nguyen. Real-time sign language fingerspelling recognition using convolutional neural networks from depth map. *CoRR*, abs/1509.03001, 2015.

[77] Maria Karam. *PhD Thesis: A framework for research and design of gesture-based human-computer interactions*. PhD thesis, University of Southampton, October 2006.

[78] Khushdeep Kaur and Parteek Kumar. Hamnosys to sigml conversion system for sign language automation. *Procedia Computer Science*, 89:794–803, 2016. Twelfth International Conference on Communication Networks, {ICCN} 2016, August 19– 21, 2016, Bangalore, India Twelfth International Conference on Data Mining and Warehousing, {ICDMW} 2016, August 19-21, 2016, Bangalore, India Twelfth International Conference on Image and Signal Processing, {ICISP} 2016, August 19-21, 2016, Bangalore, India.

[79] Daniel Kelly, John McDonald, and Charles Markham. A person independent system for recognition of hand postures used in sign language. *Pattern Recognition Letters*, 31(11):1359–1368, 2010.

[80] Daniel Kelly, John McDonald, and Charles Markham. Weakly supervised training of a sign language recognition system using multiple instance learning density matrices. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 41(2):526–541, april 2011.

[81] Laura Keyes and Adam Winstanley. *Shape Description for Automatically Structuring Graphical Data*, pages 256–264. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[82] Jong-Sung Kim, Won Jang, and Zeungnam Bien. A dynamic gesture recognition system for the korean sign language (ksl). *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 26(2):354–359, 1996.

[83] P. V. V. Kishore, M. V. D. Prasad, C. R. Prasad, and R. Rahul. 4-camera model for sign language recognition using elliptical fourier descriptors and ann. In *2015 International Conference on Signal Processing and Communication Engineering Systems*, pages 34–38, Jan 2015.

[84] Alexander Klaser, Marcin Marszalek, and Cordelia Schmid. A Spatio-Temporal Descriptor Based on 3D-Gradients. In Mark Everingham, Chris Needham, and Roberto Fraile, editors, *BMVC 2008 - 19th British Machine Vision Conference*, pages 275:1–10, Leeds, United Kingdom, September 2008. British Machine Vision Association.

[85] Teuvo Kohonen. Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[86] O. Koller, H. Ney, and R. Bowden. May the force be with you: Force-aligned signwriting for automatic subunit annotation of corpora. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6, 2013.

[87] 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.

[88] Oscar Koller, Hermann Ney, and Richard Bowden. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3793–3802, Las Vegas, NV, USA, June 2016.

[89] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[90] A. Kuzmanic and V. Zanchi. Hand shape classification using dtw and lcss as similarity measures for vision-based gesture recognition system. In *EUROCON 2007 - The International Conference on Computer as a Tool*, pages 264–269, Sept 2007.

[91] Luigi Lamberti and Francesco Camastra. *Image Analysis and Processing – ICIAP 2011: 16th International Conference, Ravenna, Italy, September 14-16, 2011, Proceedings, Part I*, chapter Real-Time Hand Gesture Recognition Using a Color Glove, pages 365–373. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[92] Laura Lanzarini, Franco Ronchetti, Cesar Estrebou, Luciana Lens, and Aurelio Fernandez Bariviera. Face recognition based on fuzzy probabilistic SOM. In *IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint*, pages 310–314. IEEE, 2013.

[93] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[94] Stan Z. Li and ZhenQiu Zhang. Floatboost learning and statistical face detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1112–1123, September 2004.

[95] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 9–14, June 2010.

[96] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 9–14. IEEE, 2010.

[97] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.

[98] T. McElroy, E. Wilson, and G. Anspach. Fourier descriptors and neural networks far shape classification. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 3435–3438, 1995.

[99] R. M. McGuire, J. Hernandez-Rebollar, T. Starner, V. Henderson, H. Brashear, and D. S. Ross. Towards a one-way american sign language translator. In *Proceedings of sixth IEEE International Conference on Automatic Face and Gesture Recognition.*, pages 620–625, 2004.

[100] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992.

[101] P. Mekala, Y. Gao, J. Fan, and A. Davari. Real-time sign language recognition based on neural network architecture. In *2011 IEEE 43rd Southeastern Symposium on System Theory*, pages 195–199, March 2011.

[102] 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.

[103] Meinard Müller. *Dynamic Time Warping*, pages 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[104] 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.

[105] A. Mostayed, M. E. Kabir, S. Z. Khan, and M. M. G. Mazumder. Biometric authentication from low resolution hand images using radon transform. In *2009 12th International Conference on Computers and Information Technology*, pages 587–592, Dec 2009.

[106] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, June 2007.

[107] Kouichi Murakami and Hitomi Taguchi. Gesture recognition using recurrent neural networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '91, pages 237–242, New York, NY, USA, 1991. ACM.

[108] Frank Natterer and Frank Wubbeling. *Mathematical Methods in Image Reconstruction*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.

[109] V. Nayakwadi and N. B. Pokale. Natural hand gestures recognition system for intelligent hci: A survey. *International Journal of Computer Applications Technology and Research*, (1):10–19, 2013.

[110] Farhood Negin, Fırat Özdemir, Ceyhun Burak Akgül, Kamer Ali Yüksel, and Aytül Erçil. A decision forest based feature selection framework for action recognition from rgb-depth cameras. In *Image Analysis and Recognition*, pages 648–657. Springer, 2013.

[111] Carol Neidle, Ashwin Thangali, and Stan Sclaroff. Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012, Istanbul, Turkey*, 2012.

[112] Carol Neidle and Christian Vogler. A new web interface to facilitate access to corpora: development of the asllrp data access interface. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012, Istanbul, Turkey*, 2012.

[113] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 53–60, 2013.

[114] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *J. Vis. Comun. Image Represent.*, 25(1):24–38, January 2014.

[115] Eng-Jon Ong and Richard Bowden. A boosted classifier tree for hand shape detection. In *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, FGR' 04, pages 889–894, Washington, DC, USA, 2004. IEEE Computer Society.

[116] Eng-Jon Ong, Helen Cooper, Nicolas Pugeault, and Richard Bowden. Reading the signs: A video based sign dictionary. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2200–2207, 2012.

[117] Eng-Jon Ong, Oscar Koller, Nicolas Pugeault, and Richard Bowden. Sign spotting using hierarchical sequential patterns with temporal intervals. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[118] Sylvie C. W. Ong and Surendra Ranganath. Automatic sign language analysis: a survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):873–891, Jun 2005.

[119] Justus Piater, Thomas Hoyoux, and Wei Du. Video analysis for continuous sign language recognition. In *In: 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, 2010.

[120] Pramod Kumar Pisharady and Martin Saerbeck. Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding*, 141:152 – 165, 2015.

[121] Pramod Kumar Pisharady, Prahlad Vadakkepat, and Ai Poh Loh. Attention based detection and recognition of hand postures against complex backgrounds. *International Journal of Computer Vision*, 101(3):403–419, 2013.

[122] Vassilis Pitsikalis, S. Theodorakis, C. Vogler, and P. Maragos. Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 1–6, June 2011.

[123] Ronald Poppe. A survey on vision-based human action recognition. *Image Vision Computing*, 28(6):976–990, jun 2010.

[124] R.P.K. Poudel, H. Nait-Charif, J. J. Zhang, and D. Liu. Region-based skin color detection. In *VISAPP 2012 - Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 301–306, 2012.

[125] Prashan Premaratne. *Human Computer Interaction Using Hand Gestures*. Springer Publishing Company, Incorporated, 2014.

[126] 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.

[127] Timo Pylvänäinen. Accelerometer based gesture recognition using continuous hmms. In *Proceedings of the Second Iberian Conference on Pattern Recognition and Image Analysis - Volume Part I*, IbPRIA'05, pages 639–646, Berlin, Heidelberg, 2005. Springer-Verlag.

[128] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[129] Siddharth S. Rautaray and Anupam Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54, 2015.

[130] Franco Ronchetti, Facundo Quiroga, Cesar Estrebou, Laura Lanzarini, and Alejandro Rosete. Lsa64: An argentinian sign language dataset. *XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016)*, pages 794–803, 2016.

[131] Franco Ronchetti, Facundo Quiroga, César Estrebou, Laura Lanzarini, and Alejandro Rosete. *Sign Languague Recognition Without Frame-Sequencing Constraints: A Proof of Concept on the Argentinian Sign Language*, pages 338–349. Springer International Publishing, 2016.

[132] Franco Ronchetti, Facundo Quiroga, Laura Lanzarini, and Cesar Estrebou. Distribution of action movements (dam): a descriptor for human action recognition. *Frontiers of Computer Science*, 9(6):956–965, 2015.

[133] Franco Ronchetti, Facundo Quiroga, Laura Lanzarini, and Cesar Estrebou. Handshape recognition for argentinian sign language using probsom. *Journal of Computer Science and Technology*, 16(1):1–5, 2016.

[134] Carsten Rother, Tom Minka, Andrew Blake, and Vladimir Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1* , CVPR '06, pages 993–1000, Washington, DC, USA, 2006. IEEE Computer Society.

[135] 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.

[136] Robert E. Schapire. *The Boosting Approach to Machine Learning: An Overview*, pages 149–171. Springer New York, New York, NY, 2003.

[137] Adam Schembri, Jordan Fenlon, Ramas Rentelis, and Kearsy Cormier. *British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2014 (Second Edition)*. London: University College London, 2014.

[138] Ahmad Zaki Shukor, Muhammad Fahmi Miskon, Muhammad Herman Jamaluddin, Fariz bin Ali Ibrahim, Mohd Fareed Asyraf, and Mohd Bazli bin

Bahar. A new data glove approach for malaysian sign language detection. *Procedia Computer Science. IEEE International Symposium on Robotics and Intelligent Sensors (IEEE IRIS2015).*, 76:60–67, 2015.

[139] E. Stergiopoulou and N. Papamarkos. Hand gesture recognition using a neural network shape fitting technique. *Engineering Applications of Artificial Intelligence*, 22(8):1141–1158, 2009.

[140] D. J. Sturman and D. Zeltzer. A survey of glove-based input. *IEEE Computer Graphics and Applications*, 14(1):30–39, 1994.

[141] I. E. Sutherland. Sketchpad, a man-machine graphical communication system. *Managing Requirements Knowledge, International Workshop on*, 00:329, 1899.

[142] V. Sutton, F.A. Paul, I. Candelaria, and J. Gunderson. *SignWriting Basics*. signWritting Press, 2009.

[143] Tomoichi Takahashi and Fumio Kishino. Hand gesture coding based on experiments using a hand gesture interface device. *SIGCHI Bull.*, 23(2):67–74, 1991.

[144] Ashwin Thangali, Joan P Nash, Stan Sclaroff, and Carol Neidle. Exploiting phonological constraints for handshape inference in asl video. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 521–528. IEEE, 2011.

[145] Paulo Trigueiros, Fernando Ribeiro, and Luís Paulo Reis. *Hand Gesture Recognition System Based in Computer Vision and Machine Learning*, pages 355–377. Springer International Publishing, Cham, 2015.

[146] Antonio W. Vieira, Erickson R. Nascimento, Gabriel L. Oliveira, Zicheng Liu, and Mario F. M. Campos. *STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences*, pages 252–259. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[147] A. Villamonte, F. Quiroga, F. Ronchetti, C. Estrebou, L. Lanzarini, P. Estelrrich, C. Estelrrich, and R. Giannechini. A support system for the diagnosis of balance pathologies. In *Congreso Argentino de Ciencias de la Computación. CACIC 2014. Argentina. October 2014.*, 2014.

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

[149] Christian Vogler and Dimitris Metaxas. *Toward Scalability in ASL Recognition: Breaking Down Signs into Phonemes*, pages 211–224. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[150] Christian Vogler and Dimitris Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *Comput. Vis. Image Underst.*, 81(3):358–384, 2001.

[151] U. von Agris, M. Knorr, and K. F. Kraiss. The significance of facial features for automatic sign language recognition. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–6, Sept 2008.

[152] U. von Agris and K.-F. Kraiss. Towards a video corpus for signer-independent continuous sign language recognition. In Sales Dias and Jota, editors, *GW 2007 The 7th International Workshop on Gesture in Human-Computer Interaction and Simulation*, pages 10–11, Lisbon, Portugal, May 23-25 2007. Poster.

[153] 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.

[154] M. B. Waldron and Soowon Kim. Isolated asl sign recognition system for deaf persons. *IEEE Transactions on Rehabilitation Engineering*, 3(3):261–271, 1995.

[155] Haijing Wang, Alexandra Stefan, and Vassilis Athitsos. *A Similarity Measure for Vision-Based Sign Recognition*, pages 607–616. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[156] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu. Robust 3d action recognition with random occupancy patterns. In *Computer Vision–ECCV 2012*, pages 872–885. Springer, 2012.

[157] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, pages 1290–1297. IEEE, 2012.

[158] Robert Y. Wang and Jovan Popović. Real-time hand-tracking with a color glove. *ACM Transactions on Graphics*, 28(3), 2009.

[159] Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224 – 241, 2011.

[160] Ming-Hsuan Yang, N. Ahuja, and M. Tabb. Extraction of 2d motion trajectories and its application to hand gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1061–1074, Aug 2002.

[161] Q. Yang. Chinese sign language recognition based on video sequence appearance modeling. In *2010 5th IEEE Conference on Industrial Electronics and Applications*, pages 1537–1542, 2010.

[162] Xiaodong Yang, Chenyang Zhang, and YingLi Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 1057–1060, New York, NY, USA, 2012. ACM.

[163] Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, and Juergen Gall. *A Survey on Human Motion Analysis from Depth Data*, pages 149–187. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[164] Q. Yuan, A. Thangali, V. Ablavsky, and S. Sclaroff. Multiplicative kernels: Object detection, segmentation and pose estimation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[165] Y. Yuan and K. Barner. An active shape model based tactile hand shape recognition with support vector machines. In *2006 40th Annual Conference on Information Sciences and Systems*, pages 1611–1616, March 2006.

[166] Morteza Zahedi, Daniel Keysers, Thomas Deselaers, and Hermann Ney. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Deutsche Arbeitsgemeinschaft für Mustererkennung Symposium*, volume 3663 of *Lecture Notes in Computer Science*, pages 401–408, Vienna, Austria, August 2005.

[167] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 19:4–10, 2012.

[168] Xiaolong Zhu and Kenneth KY Wong. Single-frame hand gesture recognition using color and depth kernel descriptors. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2989–2992. IEEE, 2012.

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