RECONOCIMIENTO DE GESTOS EN EL LENGUAJE DE SEÑAS UTILIZANDO DESCRIPTORES BASADOS EN PRIMITIVAS LOCALES DE MOVIMIENTO Y FORMA

Jefferson David Rodríguez Chivatá

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2018

RECONOCIMIENTO DE GESTOS EN EL LENGUAJE DE SEÑAS UTILIZANDO DESCRIPTORES BASADOS EN PRIMITIVAS LOCALES DE MOVIMIENTO Y FORMA

Jefferson David Rodríguez Chivatá

Una tesis presentada en cumplimiento de los requisitos para el grado de Ingeniero de Sistemas e Informática

Director:
Fabio Martínez Carrillo, Ph.d
Profesor escuela de Ingeniería de Sistemas e informática

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2018

AGRADECIMIENTOS

El autor expresa su agradecimiento:

Al grupo de investigación en ingeniería biomédica (GIIB) y al semillero de investigación en análisis de movimiento y visión por computador (MACV), principalmente al profesor Fabio Martínez Carrillo por ser un gran guía, por su paciencia, dedicación, esfuerzo y orientación. También quisiera agradecer por la preocupación de la formación de habilidades integras y profesionales que siempre ha inculcado, sin él no hubiera sido posible la realización de este trabajo.

A todos mis amigos de infancia y universidad y a aquellas personas que fueron parte de mi formación, que de una u otra manera me han permitido construir el ser humano que soy, por su fiel compañía y sincera amistad.

A la escuela de Ingeniería de Sistemas e Informática (EISI) y a la Universidad Industrial de Santander (UIS) por la gran formación que me han brindado y lo cual ha sido fundamental para ser un buen profesional.

Finalmente quiero realizar un agradecimiento especial a mis padres, hermanas y demás familiares por sus enseñanzas, confianza y apoyo, elementos que han tenido una verdadera importancia para seguir cada día.

CONTENIDO

LISTA DE FIGURAS

RESUMEN

Título: Reconocimiento de gestos en el lenguaje de señas utilizando descriptores basados en primitivas locales de movimiento y forma [1]

Autor: Jefferson David Rodríguez Chivatá[2]

Palabras Clave: Análisis de movimiento, reconocimiento de señas, *Shape Difference VLAD*, primitivas locales

DESCRIPCIÓN:

El reconocimiento automático en el lenguaje de señas (SLR) es una tarea fundamental para ayudar en la inclusión de la comunidad sorda en la sociedad, facilitando en la actualidad, muchas tareas de interacción multimedia convencionales. Sin embargo, el reconocimiento de gestos continúa siendo un problema abierto debido a las múltiples variaciones entre persona dadas por su cultura, historia y las interpretaciones particulares según las regiones. Tales variaciones implican grandes desafíos para entender y asociar etiquetas del lenguaje semántico a los gestos espacio-temporales. Además, los escenarios en línea requieren predicciones en cada instante de tiempo necesitando reconocer los gestos mientras se desarrollan. Este trabajo presenta un enfoque novedoso para reconocer gestos predominantes en el lenguaje de señas. Este reconocimiento puede realizarse para secuencias de video completas, así como también para secuencias parciales e incompletas. El método comienza computando parches volumétricos que contienen información cinemática de diferentes primitivas de flujo y de apariencia. A continuación, se aprenden varios intervalos secuenciales para llevar a cabo la tarea de reconocimiento parcial. Para cada nuevo vídeo, se obtiene una representación acumulativa utilizando la estrategia Shape Difference VLAD en diferentes intervalos del vídeo. Cada descriptor SD-VLAD recupera la media y la varianza de la información de movimiento como firma del gesto calculado. El enfoque propuesto fue evaluado en un conjunto de datos públicos con 64 clases diferentes, registrados en 3200 videos. El enfoque propuesto es capaz de reconocer gestos en señas usando sólo 40% de la secuencia con una precisión promedio de 54%. Para secuencias completas, alcanza un promedio del 85 %.

---

[1] Trabajo de Grado

[2] Facultad de Ingenierías Físicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Fabio Martínez Carrillo, Ph.d.

ABSTRACT

Title: Sign Language Gesture Recognition using descriptors based on local primitives of motion and shape[1]

Author: Jefferson David Rodríguez Chivatá[2]

Keywords: Motion analysis, sign recognition, Shape Difference VLAD, local primitives.

DESCRIPTION:

Automatic Sign Language Recognition (SLR) is a fundamental task to assist in the inclusion of the deaf community in society, currently facilitating many conventional multimedia interaction tasks. However, the recognition of gestures remains an open problem due to the multiple variations between people given by their culture, history and particular interpretations according to regions. Such variations imply great challenges in understanding and associating labels of semantic language with spatial-temporal gestures. In addition, on-line scenarios require predictions at every moment of time and need to recognize gestures as they unfold. This paper presents a novel approach to recognizing gestures predominant in sign language. This recognition can be performed for complete video sequences as well as for partial and incomplete sequences. The method begins by computing volumetric patches containing kinematic information from different flow and appearance primitives. Several sequential intervals are then learned to carry out the partial recognition task. For each new video, a cumulative representation is obtained using the Shape Difference VLAD strategy at different intervals in the video. Each SD-VLAD descriptor retrieves the mean and variance of the movement information as signature of the calculated gesture. The proposed approach was evaluated in a public data set with 64 different classes, recorded in 3200 videos. The proposed approach is capable of recognizing sign gestures using only 40 percent of the sequence with an average accuracy of 54%. For complete sequences, it averages 85%.

---

[1] Research Work.

[2] School of Physical-Mechanical Engineering. Department of Systems Engineering and Informatics. Advisor, Fabio Martínez Carrillo, Ph.d.

INTRODUCCIÓN

La comunidad de sordos y personas con alguna limitación auditiva en todo el mundo se estima en más de 466 millones de acuerdo con la Organización Mundial de la Salud (OMS) [1]. La lengua de señas es el principal recurso de comunicación e interacción para estas personas, siendo tan rica y compleja como cualquier lengua hablada. Este lenguaje está compuesto por gestos espacio-temporales coherentes y continuos que resumen los movimientos articulados de las extremidades superiores, las expresiones faciales y las posturas del tronco. A pesar de la importancia de la interpretación automática de las señas, la caracterización de gestos se ve limitada debido a las múltiples variaciones entre personas. También, diferentes factores como la cultura, historia y las interpretaciones particulares según las regiones pueden introducir variaciones externas en los gestos. Tales variaciones implican grandes desafíos para entender y asociar etiquetas del lenguaje semántico a los gestos espacio-temporales. Además, para las interacciones reales, las interpretaciones automáticas actuales exigen aplicaciones en línea para reconocer los gestos mientras se desarrollan. En tal sentido, el desafío es aún mayor porque las estrategias computacionales propuestas deben tener la capacidad de predecir respuestas a información incompleta mientras permanecen robustos a problemas típicos en el procesamiento de video como cambios de iluminación, perspectiva de los sujetos, oclusión de los articuladores durante la conversación, entre muchos otros. Estos problemas relacionados limitan el uso de las metodologías de aprendizaje automático, la usabilidad de las herramientas multimedia y ponen a la comunidad sorda en desventaja para explorar gran parte de la información en las plataformas multimedia.

El reconocimiento y clasificación de señas ha sido abordado en la literatura por múltiples enfoques que incluyen representaciones de forma global que segmentan todos los articuladores del lenguaje pero con limitaciones debido a las oclusiones y dependencias de escenarios controlados [2]. Otras estrategias han desarrollado el análisis de gestos con representaciones locales que incluyen la caracterización de puntos de interés [3, 4] y el análisis de apariencia y primitivas geométricas para capturar la forma de los gestos en videos [5, 6]. Por ejemplo, Zahedi *et. al.* [7] propusieron un reconocimiento computando descriptores de apariencia que junto con gradientes

de primer y segundo orden caracterizaban señas particulares. Sin embargo, este enfoque depende de la apariencia y perspectiva del sujeto en la secuencia de vídeo. Una extensión de este trabajo desarrolló el análisis de información multimodal para recuperar la forma en secuencias RGB-D y también calculando trayectorias con acelerómetros para complementar la descripción de la seña [2]. A pesar de las ventajas del análisis 3D, las secuencias de profundidad se limitan a escenarios controlados y los acelerómetros externos pueden alterar el movimiento natural de los gestos. La caracterización de movimiento ha sido fundamental en el desarrollo de estrategias para reconocer los gestos siendo robustas a variaciones de apariencia y a los cambios de iluminación [4, 8]. Por ejemplo, en las secuencias de video en [3, 8] se caracterizaron los gestos a partir de las relaciones de primer orden de las velocidades de apariencia capturadas del campo de flujo óptico de Lukas-Kanade. Sin embargo, este enfoque es propenso a errores debido a la sensibilidad del flujo a pequeños desplazamientos de la cámara y también a la naturaleza dispersa del enfoque donde se capturan pocos puntos de desplazamiento que dificultan cualquier análisis estadístico. En la misma línea, Jakub Konecn'y *et. al.* [8] integra información de forma local con histogramas de flujo óptico para describir gestos. Este enfoque logró una representación a nivel de frame, pero perdió información local y regional para representar gestos. Jun Wan *et. al.* [3] propusieron un diccionario de palabras dispersas codificadas a partir de puntos SIFT salientes y complementadas con descriptores de flujo capturados alrededor de cada punto. Esta representación logra un rendimiento adecuado en el reconocimiento de gestos, pero sigue siendo limitado para cubrir gran parte de la variabilidad de los gestos. En [4] se logró una descripción de movimiento local en cada frame de un LS en particular computando las trayectorias de movimiento a lo largo del video. Sin embargo, el computo de trayectorias se ve limitado con el seguimiento de los movimientos en intervalos de tiempo cortos, perdiendo continuidad en las descripciones.

Recientemente, estrategias de deep learning han sido propuestas para el reconocimiento de gestos en tiempo real. Por ejemplo, Masood *et. al.* [9] propusieron un modelo convolucional para aprender patrones espaciales y un modelo recurrente para patrones temporales. Este enfoque permite un reconocimiento de múltiples gestos pero no segmenta los principales articuladores aprendiendo elementos que son innecesarios como el fondo de la escena, además los modelos recurrentes tienen limitaciones para adaptarse a cambios bruscos en las señales. De manera similar, en [10] se consideró un modelo convolucional para alcanzar representaciones independientes de la perspectiva pero ignora el análisis dinámico, fundamental para caracterizar las señas. Por otro lado, Liu *et. al.* [11] propusieron una estrategia computacional sobre imágenes RGBD, primeramente realizan segmentación de manos para obtener las respectivas trayectorias, el modelo convolucional aprende los patrones cinemáticos pero se podría mejorar con características cinemáticas de orden superior.

La principal contribución de este trabajo es una representación estadística, a un nivel intermedio, de primitivas cinemáticas y geométricas que logra una descripción local y regional de gestos en el lenguaje de señas. El enfoque propuesto es robusto para describir señas a partir de representaciones incompletas siendo eficiente en aplicaciones en línea. El enfoque propuesto inicialmente calcula un conjunto de parches o volúmenes de movimiento. Estos parches son caracterizados con histogramas de movimiento y apariencia usando primitivas como límites de velocidad, aceleración, características regionales como rotación y divergencia, además de gradientes geométricos espaciales. Estos volúmenes se codifican sobre diferentes diccionarios que representan los principales parches de descripción en los videos. Entonces, el descriptor de gestos es propuesto basado en *Shape Difference VLAD*, el cual recupera los principales patrones descritos por la media y la varianza del movimiento. Esta representación permite recuperar gestos parciales y describir de forma robusta diferentes gestos en un lenguaje de señas concreto. Finalmente, el descriptor de movimiento obtenido se mapea a una máquina vectorial de soporte y se valida en un corpus público LSA64.

Capítulo 1
PLANTEAMIENTO Y JUSTIFICACIÓN DEL PROBLEMA

En el mundo, la comunidad de sordomudos se estima aproximadamente en 70 millones según reporte oficial de la Federación mundial de sordos. Particularmente en Colombia, la comunidad en estado de discapacidad auditiva para el 2013 tenía un aproximado de 131.538 personas sordas y 455.718 personas con limitaciones para oír. A pesar del amplio número de personas con discapacidad auditiva, son muy pocas y limitadas las herramientas de soporte tecnológico a los procesos de interacción con otras comunidades. También, se evidencia poca información con respecto a la automatización de procesos de aprendizaje y el uso de herramientas multimedia.
En términos generales, la lengua de la comunidad de sordomudos está conformado por un conjunto de señas que describen un patrón espacio-temporal de diferentes articuladores, como cabeza, tronco, brazos y manos. Cada seña se definen como un conjunto de gestos que se desarrollan coherentemente y de forma continua a través del tiempo. La caracterización, relación e interpretación automática de gestos representa un desafío para la comunidad académica y científica hoy en día, debido al amplio número de gestos, los diferentes articuladores que componen una seña y la variabilidad inter e intra personas que desarrollan la comunicación. Específicamente, las señas pueden variar entre grupos de personas, y su representación puede tener excepciones espaciales y temporales con significados semánticos similares. En visión por computador, estas variaciones espacio-temporales limitan la representación y modelamiento de las señas. Además, las interpretaciones automáticas actuales exigen aplicaciones en línea para reconocer los gestos mientras se desarrollan. estas representaciones incompletas en video están limitadas a problemas típicos como cambios de iluminación, perspectiva de los sujetos y oclusión de los articuladores durante la conversación.

Capítulo 2
OBJETIVOS

2.1 OBJETIVO GENERAL

Determinar un descriptor basado en la ocurrencia de primitivas locales de movimiento y forma para la caracterización de señas específicas registradas en vídeo.

2.2 OBJETIVOS ESPECÍFICOS

- Representar secuencias de vídeo utilizando un conjunto de volúmenes espacio-temporales que codifican la información local de forma y de movimiento aparente.
- Caracterizar gestos predominantes a partir de un diccionario basado en volúmenes locales representativos de las secuencias de vídeo.
- Calcular un descriptor estadístico de la seña registrada en cada vídeo, basado en la ocurrencia de volúmenes locales del diccionario construido.
- Clasificar un conjunto de señas automáticamente utilizando un algoritmo clásico de aprendizaje de máquina que permita como entrada el descriptor propuesto.
- Validar el método propuesto en cuanto a la predicción de señas en un corpus que contiene un conjunto de señas en un lenguaje de sordomudos particular.

Capítulo 3
MÉTODO PROPUESTO

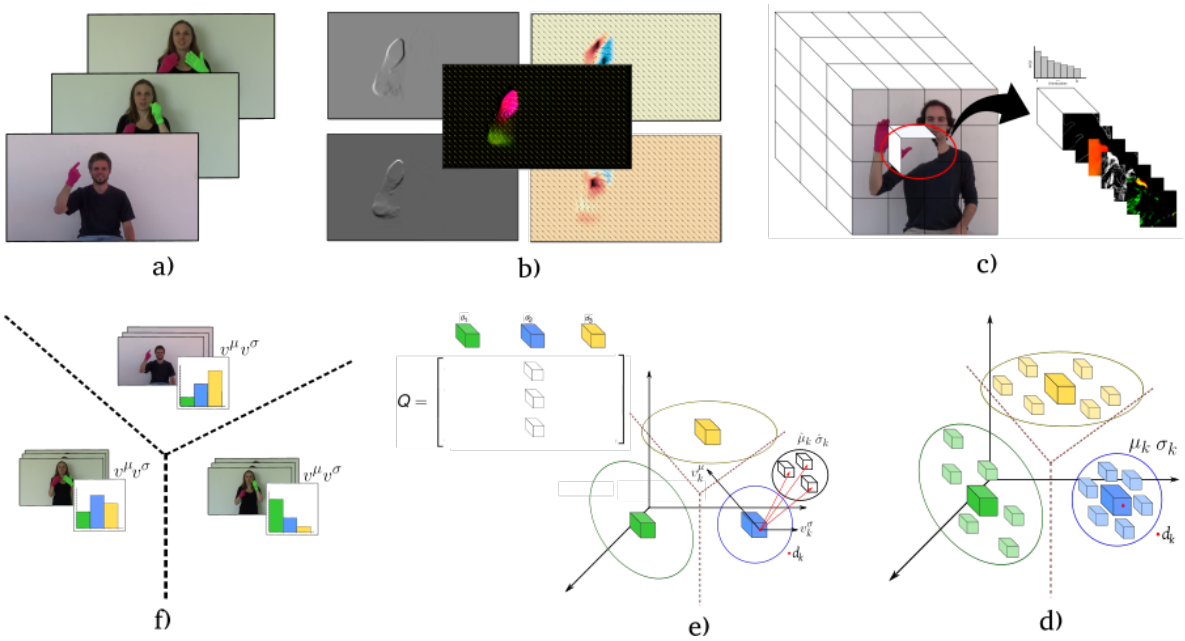
En este trabajo se presenta una estrategia automática para clasificar y reconocer señas utilizando una representación cinemática y geométrica en un nivel de representación intermedio. El enfoque aquí propuesto alcanza una descripción local y regional robusta de señas considerando varias características de movimiento y forma. El enfoque clásico de diccionarios permite codificar parches volumétricos para representar las señas. La representación a nivel intermedio *Shape Difference VLAD* , permite capturar similitudes cinemáticas y geométricas locales entre parches pero también es capaz de recuperar la forma de la distribución de los descriptores asociados [12]. El desarrollo del enfoque propuesto para clasificación se ilustra en la Figura 1.

3.1 ENFOQUE LOCAL PARA LA REPRESENTACION DE SECUENCIAS DE VIDEO

Muchos sistemas de reconocimiento y clasificación usan la perspectiva global para representar la totalidad de la escena procesada, resultando en metodologías eficientes y compactas. Sin embargo, esta perspectiva es sensible a problemas de ruido, oclusión y es dependiente de la calidad de la segmentación para obtener resultados aceptables. Por estas razones, en este trabajo se consideró un enfoque local que permite manejar los problemas globales, considerando más puntos de interés lo cual permite un análisis más detallado. Además, el enfoque propuesto es un insumo para lograr representaciones estadísticas que describan de forma robusta los gestos.

3.1.1 Extracción de características. La caracterización de movimiento y forma en conjunto ha demostrado ser fundamental para el análisis de señas en aplicaciones de reconocimiento gestual (ver figura 2). Una tarea fundamental en este tipo de aplicaciones es la cuantificación de grandes regiones de movimiento desarrolladas por actuadores independientes, tales como brazos, manos, cara o incluso hombros. En este enfoque calculamos un conjunto

Figure 1. Estructura del enfoque propuesto para clasificación de señas. Corpus de señas (a) se codifican a partir de un conjunto de primitivas de movimiento y forma (b). Luego, las primitivas se procesan en cada secuencia generando una descripción local para cada parche volumétrico (c). Se obtiene un diccionario de patrones (d) y finalmente se obtiene una representación global utilizando *Hard assignment* y *Shape Difference VLAD*. El descriptor computado se asigna a una máquina de soporte vectorial previamente entrenada para clasificar las señas (f).



de primitivas cinemáticas y geométricas para describir gestos. El conjunto de características calculadas se describe a continuación:

- **Flujo óptico denso de largo desplazamiento.** Una primera primitiva cinemática considerada fué el campo de flujo óptico denso aparente producido entre cuadros consecutivos. Los enfoques típicos siguen siendo limitados para cuantificar grandes desplazamientos debido a la suposición de movimiento suave en los vecindarios locales. Para evitar estas limitaciones, aquí se implementó un enfoque de flujo óptico robusto capaz de capturar campos de flujo densos pero considerando grandes desplazamientos de gestos [13]. Su cálculo esta sometido a varias restricciones, descritas a continuación:

  - **Color:** Esta es la restricción clásica de $E_{color}(w)$ que consiste en asumir una intensidad de color constante para los píxeles en dos imágenes consecutivas. Esta restricción es comúnmente utilizada por los métodos de flujo óptico variacional y considera que un objeto entre dos cuadros consecutivos únicamente se desplaza pero su información de color permanecerá constante.
  - **Gradiente:** Esta restricción $E_{gradiente}(w)$ indica que el gradiente entre imágenes consecutivas tiene variaciones locales mínimas. Estas variaciones permiten calcular las deformaciones opcionales que se producen en la secuencia de vídeo. En esta restricción, el objeto también se caracteriza utilizando los gradientes de primer nivel y entonces se buscará la mínima diferencia en gradientes entre dos imágenes consecutivas.
  - **Suavidad:** Esta restricción $E_{suavidad}(w)$ cuantifica la diferencia mínima entre los vectores de velocidad dentro de una región. La suposición es que el patrón de velocidad debe ser similar en un determinado vecindario, teniendo en cuenta la dispersión local del campo vectorial.
  - **Regiones no locales:** Esta restricción $E_{desc}(w_1)$ permite buscar grandes desplazamientos locales entre frames consecutivos comparando las regiones coincidentes calculadas a partir de vectores de características. Esta restricción es la que diferencia particularmente el trabajo de Brox, teniendo en cuenta que no solo minimiza flujos en suaves en vecindarios, sino que busca desplazamientos largos en regiones distantes. Estos desplazamientos largos son encontrados a partir de estrategias de emparejamiento de puntos de interes.

  Finalmente, la suma de todas las ecuaciones de energía permite encontrar el flujo óptico denso de largo desplazamiento calculado sobre todas las secuencias de vídeo. Por lo

tanto, un modelo completo se define como un problema de optimización único, que se realiza minimizando el método de variación en:

$$E(w) = E_{color}(w) + \gamma E_{gradiente}(w) + \alpha E_{suavidad}(w) + \beta E_{Match}(w, w_1) + E_{desc}(w_1) \tag{3.1}$$

Donde $\{\gamma, \alpha, \beta\}$ representan constantes de regularización con valores entre $[0, 1]$. Esto muestra que el modelo puede manejar deformaciones, discontinuidades del movimiento, oclusión y desplazamientos arbitrariamente grandes. Este método es robusto y ha sido ampliamente utilizado en la literatura para diferentes aplicaciones.

- **Campos de divergencia** Adicionalmente a la descripción del campo de velocidad, en este trabajo también se consideró el patrón físico de divergencia computado sobre el campo de flujo. La característica resulta de derivar los componentes de flujo (u, v) en cada punto x a lo largo de las direcciones espaciales (x, y), descritas como:

$$div(p_t) = \frac{\partial u(p_t)}{\partial x} + \frac{\partial v(p_t)}{\partial y} \tag{3.2}$$

Esta característica captura la expansión local del campo de flujo, y resulta útil para caracterizar independientemente los articuladores del cuerpo a lo largo del desarrollo de la seña.

- **Campos de rotación** Las medidas rotacionales de campo de flujo también fueron consideradas. A partir de cada punto local de campo estimado se mide la rotación alrededor de un eje perpendicular [14, 15]. Estos patrones rotacionales destacan los gestos circulares, comúnmente reportados en la lengua de señas. Además, esta medida estima la rigidez del flujo, útil para distinguir los movimientos articulados. La rotación de campo puede expresarse como:

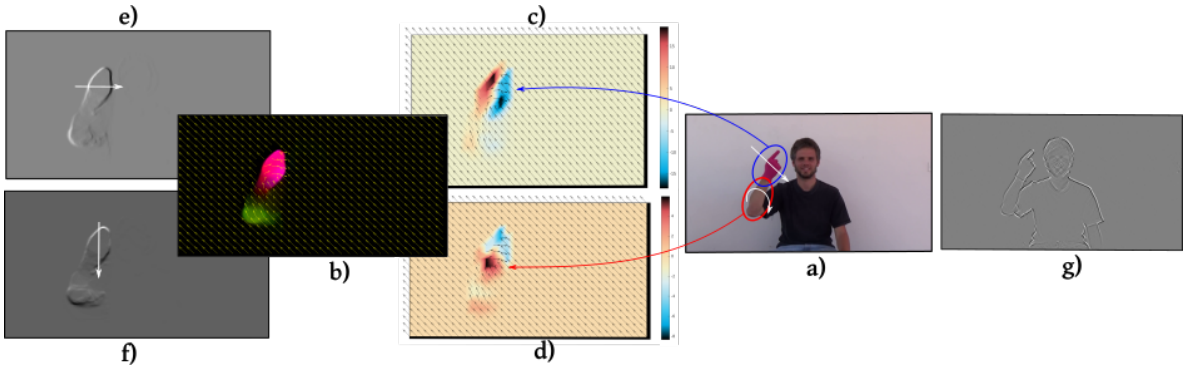
$$curl(p_t) = \frac{\partial v(p_t)}{\partial x} - \frac{\partial u(p_t)}{\partial y} \tag{3.3}$$

- **Límites de movimiento** Derivar espacialmente los componentes de flujo [16] de la siguiente manera: $\frac{\partial u(p_t)}{\partial x}$, $\frac{\partial u(p_t)}{\partial y}$, $\frac{\partial v(p_t)}{\partial x}$, $\frac{\partial v(p_t)}{\partial y}$ permite identificar las zonas donde se produce el mayor cambio en la velocidad del movimiento. Es decir, se captura características de aceleración computadas espacialmente para ser consideradas como información cinemática de las señas, codificando el movimiento relativo entre píxeles. El gradiente de flujo elimina la información de movimiento constante, mientras que permanecen los cambios de

velocidad. Esta primitiva también resalta los principales movimientos del articulador.

- **Gradientes de apariencia** Los gradientes son una característica global ampliamente usada para reconocer objetos, en nuestro problema los gradientes pueden detectar los bordes o contornos de los articuladores presentes en la seña brindando información geométrica que permita describir más adecuadamente las señas. Las derivadas $\frac{\partial p_t}{\partial x}, \frac{\partial p_t}{\partial y}$ de cada pixeles p_t en las direcciones (x, y) son calculadas por la convolución de la imagen o frame del vídeo y un kernel específico. En este trabajo se uso el kernel o operador Sobel de dimensión 3×3.

Figure 2. Primitivas de movimiento y forma. En esta imagen se ilustra las características usadas para describir los gestos en las señas. De la imagen (a) se extrae el flujo óptico (b), divergencia (c), rotacional (d), límites de movimiento (e,f) y gradientes (g).



3.1.2 Modelado de gestos con parches volumétricos espacio-temporales.

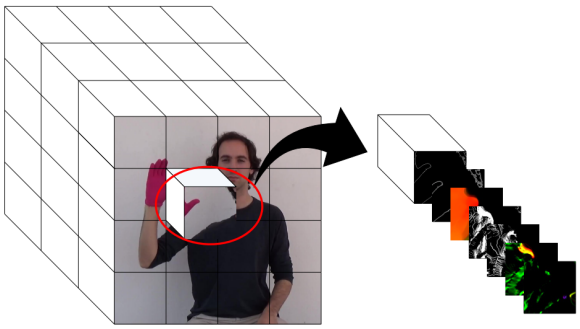
El principal inconveniente de la caracterización de gestos a nivel global es la sensibilidad a la oclusión de los articuladores y a las perturbaciones de la escena mientras se describe la seña. Las características extraídas en este trabajo son calculadas por cada pixel sobre la totalidad de la imagen, pero analizadas en regiones volumetricas de forma local.

El enfoque propuesto se basa en una representación gestual local, a partir de la cual, un conjunto de parches locales pueden representar tanto la descripción temporal como espacial del gesto en la seña (ver figura 3). En este trabajo, un gesto particular se define como un conjunto de n parches espacio-temporales no-sobrelapados $S = \{p_{1...n}^{(c,j)} : j \in [t_1 - t_2]; c \in [x_1, x_2]\}$ limitados en un intervalo temporal j y distribuidos espacialmente en una región c que contienen información parcial de las características extraídas.

3.1.2.1 Parches volumétricos de interés.

Procesar los vídeos desde el enfoque local implica analizar una mayor cantidad de elementos que tratando globalmente el vídeo.

Figure 3. Volúmenes espacio-temporales. En esta figura se aprecia como un gesto es modelado localmente a partir de la división de la escena en volúmenes no-sobrelapados que contienen información parcial de las primitivas.



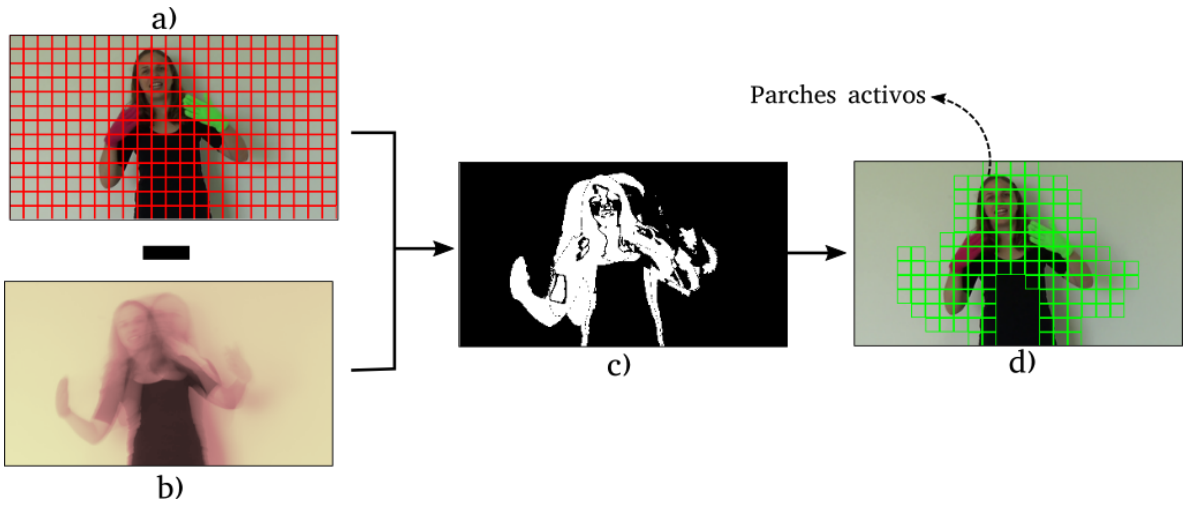
A menudo gran parte de estos elementos contienen información inadecuada y no aporta discriminancia a la descripción final. Por tal motivo, un paso inicial esencial es la segmentación de puntos de interés.

En este trabajo se seleccionaron un conjunto de parches relevantes basados en la información cinemática que contenían. Para dicha selección, se utilizó un esquema de detección de movimiento de primer orden que permite modelar el fondo del video, es decir elementos que permanecen mayormente estáticos a lo largo del video son descritos como fondo. Para esto se calcula el promedio de todos los frames que conforman el vídeo como: $B(\widehat{x}, y) = \frac{1}{t}\sum_{t=1}^{t} f_t(x, y)$, donde para cada frame t, se obtiene la información de interés en la escena por una simple resta, que mantiene aquellas zonas donde se presentó movimiento, entre el frame actual y el fondo modelado, de la siguiente manera: $|f_t(x, y) - B(\widehat{x}, y)| > \tau$ (ver en la Figura 4). Diferentes valores de τ (obtenidos experimentalmente) permiten capturar movimientos pequeños importantes. Para propósitos de reconocimiento en línea, el fondo promedio puede ser calculado a partir de un estimador recursivo de la media.

3.1.2.2 Descripción cinemática y geométrica de los parches.

Cada uno de estos parches volumétricos se describe utilizando la información local de movimiento y forma, codificada como histogramas. Entonces, para cada primitiva cinemática y geométrica se

Figure 4. Sustracción de fondo. La figura ilustra los parches de interés (d) obtenidos del frame actual (a) por la multiplicación de la máscara (c) resultante de la sustracción del fondo modelado (b).



consideró una representación local del histograma, como:

$$h(p) = \sum_{\mathbf{x} \in p} R_b(\mathbf{x}) W(\mathbf{x}), b = \left\{1, 2, \cdots, \frac{2\pi}{\Delta\theta}\right\}$$
$$R_b(x, y) = \begin{cases} 1 \;\; if \;\; (b-1)\Delta\theta \leq \theta(\mathbf{x}) < b\Delta\theta \\ \\ 0 \;\; \text{elsewhere} \end{cases} \tag{3.4}$$

donde $R_b(\mathbf{x})$ es una función de activación que determina el bin particular que codifica la primitiva local, mientras que $W(\mathbf{x})$ corresponde al peso particular que suma en ese bin del histograma. En particular, para los histogramas de flujo óptico (**HOOF**) el bin b corresponde a la orientación, mientras que $W(\mathbf{x})$ es definida por la norma de cada vector, como se propuso en [17]. De manera similar, los límites de movimiento son codificados como histogramas **MBH**, cuantificados para cada componente x, y, como se propuso en [16]. Para la divergencia y curvatura, las primitivas se acumulan estadísticamente definiendo los bines como: $\{\max, \frac{\max}{2}, 0, \frac{\min}{2}, \min\}$. En tal caso, el histograma de curvatura **HCURL** cuantifica el movimiento principal alrededor del eje perpendicular, mientras que el histograma de divergencia **HDIV** resume los principales momentos de divergencia presentes alrededor de cada parche espacio-temporal. Para la divergencia se lleva a cabo un simple conteo de ocurrencias, mientras que para la rotación se pesa la ocurrencia de acuerdo a la velocidad angular. Para cuantificar los gradientes de apariencia calculados, se uti-

lizó un enfoque basado en el histograma orientado a gradientes (**HOG**) propuesto en [18], donde por cada pixel $\mathbf{x}$ se obtiene un gradiente con magnitud $W(\mathbf{x})$ y orientación perteneciente a un bin b. El descriptor final para cada parche se forma como la concatenación de todos los histogramas descritos.

3.2 CARACTERIZACION DE GESTOS BASADA EN DICCIONARIOS

El esquema de caracterización basada en diccionarios ha sido ampliamente utilizado en el framework Bolsa de características, como el paso intermedio para codificar los parches globalmente. En este trabajo los diccionarios de representación fueron calculados usando el algoritmo *k-means*. Este algoritmo calcula los principales K centroides volumétricos que mejor representan la población de parches sobre un conjunto de videos de entrenamiento.
En la representación de diccionarios obtenida, los parches son descritos por el conjunto de histogramas cinemáticos y geométricos. Entonces, el *k-means* como estrategia no supervisada agrupa los parches según sus características, resaltando patrones predominantes para describir gestos. Tales patrones predominantes son la base para generar la representación final. En este trabajo se evaluó el enfoque propuesto en dos diferentes escenarios: la clasificación de secuencias de video completas y el reconocimiento continuo de señas durante la reproducción de un video. Por lo anterior, dos diferentes esquemas de construcción de diccionarios fueron definidos, como se describe a continuación.

3.2.1 Diccionarios para clasificación.

Para la tarea de clasificación (figura 1), se consideró construir un diccionario volumétrico único en la fase de entrenamiento utilizando la totalidad de los descriptores locales calculados. Es decir, se determina un conjunto de K parches representativos $D = [d_1, d_2, ...d_K] \in \mathbf{R^{K \times d}}$ recuperados del conjunto total de N parches locales representados por un descriptor de dimensión $\mathbf{d}$, denotado por $X = [x_1, x_2, ...x_N] \in \mathbf{R^{N \times d}}$. Como se mencionó anteriormente el algoritmo usado es el clásico *k-means*, método de clustering donde se agrupan elementos similares bajo el criterio Euclidiano para la distancia y además se cumple que $K << N$.

3.2.2 Diccionarios para reconocimiento.

En los últimos años ha tomado mayor relevancia las aplicaciones que funcionan en línea. Debido a la alta interacción entre dispositivos tecnológicos digitales y las personas se creó la necesidad de brindar respuestas aproximadas mientras la información se procesa, es decir generar salidas a información incompleta. Para

tal objetivo, se adapta el modelo anterior a uno nuevo donde se puedan realizar varias clasificaciones de las señas a lo largo del vídeo (ver figura 5). Para reconocer temporalmente las señas, un conjunto de diccionarios acumulados $\mathbf{\Lambda} \in \mathbb{R}^{t \times K \times d}$ de información parcial son construidos a partir de diferentes intervalos en las secuencias del vídeo en el entrenamiento. Entonces, $\mathbf{\Lambda} = [D_1, D_2, \ldots D_t]$ tiene t diccionarios temporales generados en un marco progresivo acumulativo cada 20% de los parches activos. Es decir, D_1 es un diccionario construido solo con el primer 20% de la información, D_2 resume la representación del 40% de los parches activos y así sucesivamente. Cada diccionario $D^i = [d_1^i, d_2^i, ...d_k^i], \in \mathbf{R^{K \times d}}$ tiene K centroides representativos de dimensión $\mathbf{d}$. Cada diccionario D^i es construido a partir de N parches $X^i = [x_1^i, x_2^i, ...x_N^i]$ los cuales se van acumulando e incrementando a medida que la seña es desarrollada. Se considera que el conjunto de K parches en cada partición temporal es suficiente para representar gestos particulares a partir de información incompleta. El diccionario progresivamente alcanza una representación más fina debido a la mayor densidad de parches adaptándose adecuadamente a la dinámica presente en las señas.

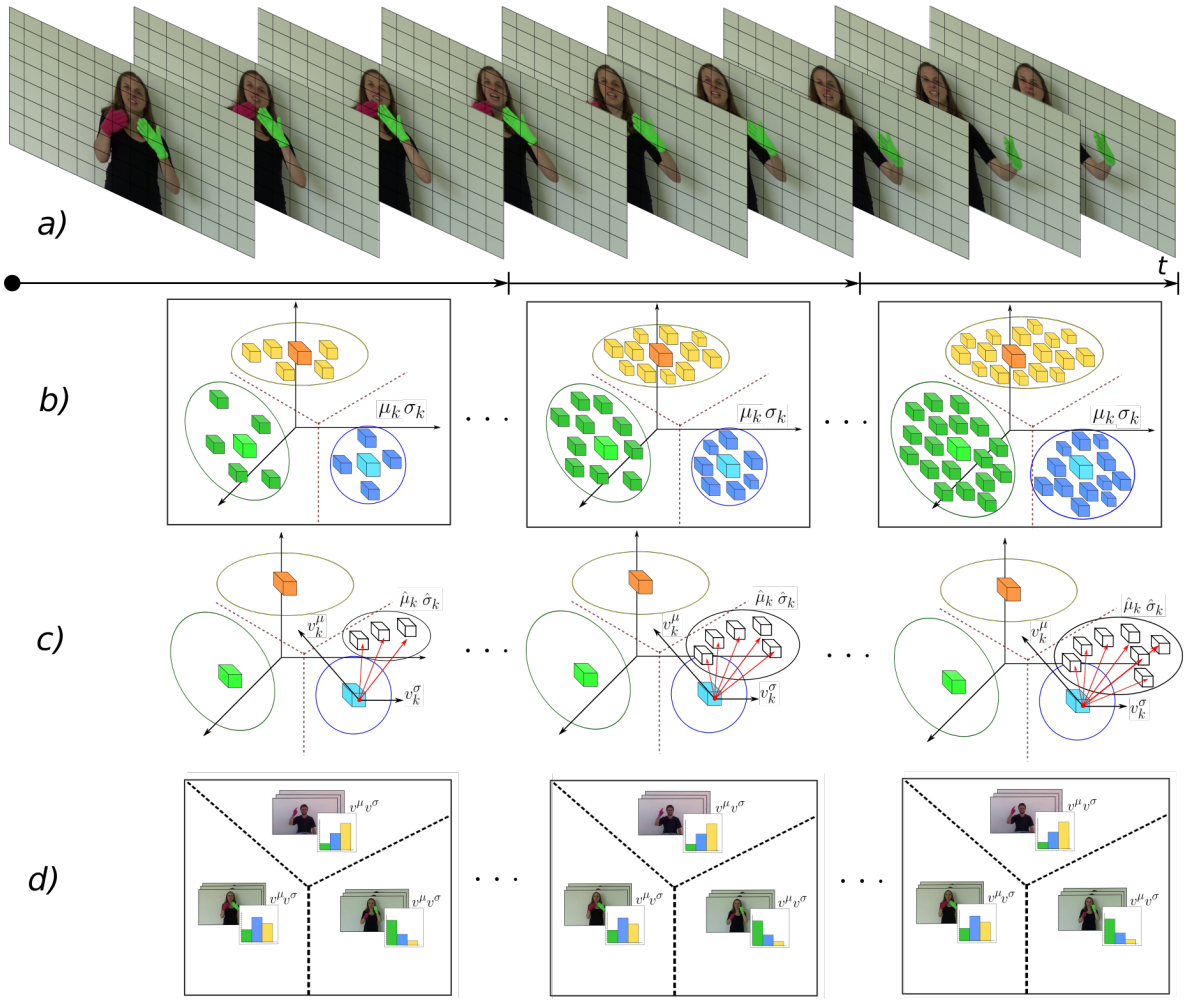
3.3 DESCRIPTORES ESTADISTICOS BASADOS EN LA OCURRENCIA

En este trabajo se exploraron varias estrategias de representación final. Los enfoque más básicos son basados en la ocurrencia, comúnmente llamados métodos de orden cero. Estos esquemas son eficientes en tiempo pero carecen de descripción de forma y localización espacial en sus representaciones. Es por esto que se han propuesto otros algoritmos más robustos conocidos como representaciones de orden superior, los cuales agregan información relevante sobre la distribución de los descriptores locales.

3.3.1 Codificación de gestos Los diccionarios calculados se utilizan como base para codificar los descriptores locales permitiendo generar una representación global de la señas registrada en video. Existen diferentes estrategias de codificación densa, que se agrupan en los siguientes grupos [19]:

1. **La codificación basada en votación, de orden cero o de ocurrencia** asocia cada parche de un video a la respectiva palabra d_k mas cercano en el diccionario. En este tipo de codificación, el descriptor global de una secuencia es representado, por una función de asignación s, que toma en cuenta el número de ocurrencias de cada centroide d_k. Tipicamente, se realiza un asignamiento rígido de cada parche x del video, al centroide d_k mas cercano, aumentando su occurrencia en 1. Esta relación puede ser expresada como:

Figure 5. Reconocimiento en línea. La figura ilustra el reconocimiento en t instantes de tiempo de las señas (a). Dado por un enfoque parcial y acumulado de diccionarios (b) y representaciones finales (c) se realizan varias clasificaciones (d) a lo largo del video.



$\mathbf{s}(l) = 1.\ if\ \ l = \text{argmin}_k\, ||x - d_k||_2\ \ \textit{s.t.}\ \ ||\mathbf{s}||_0 = 1.$

Este tipo de asignación permite ponderar parches salientes que representan componentes importantes en la señas, mientras descarta centroides comunes entre el conjunto de gestos caracterizados. Sin embargo, en algunos casos este descriptor induce a la pérdida de información relevante. Por otra parte, en el asigmamiento suave, por cada descriptor local x se asignan valores a todas las palabras d_k según su similitud, medida con una distancia Euclidiana. El asignamiento suave es dado por: $\mathbf{s}(k) = ||x - d_k||_2$.

La representación final para la codificación de orden cero, se genera mediante un proceso llamado *pooling*, el cual toma todos los códigos $\mathbf{s}$ y los reduce a un histograma que representa la descripción de la seña. Hay dos métodos comunes para realizar este proceso:

- *Average Pooling*: Donde, el k-ésimo componente del descriptor final es calculado por: $\mathbf{x}_{f_k} = \sum_{n=1}^{N} \frac{\mathbf{s}_n(k)}{N}$.
- *Max Pooling*: usado en el asignamiento suave, donde los componentes k-ésimos del descriptor final es calculado por: $\mathbf{x}_f = \max(\mathbf{s}_1(k)), \ldots, \max(\mathbf{s}_N(k))$.

2. **La codificación basada en super-vectores** logran una representación mas robusta agregando estadísticas de orden superior entre cada parche y el diccionario de representación. Por ejemplo, la representación *VLAD* [20] incluye información de los *clusters* aprendidos alrededor de cada centroide, para describir los gestos como diferencias entre los descriptores locales de cada vídeo y la palabra más cercana d_k. Formalmente estas diferencias son expresadas por cada cluster en un vector característico $v_k^{\mu} = \sum_{j=1}^{n_k}(x_j - d_k)$ con dimensionalidad **d**. Donde n_k indica el número de descriptores locales asociados a la palabra d_k. El descriptor final $\mathbf{x}_f$ es la concatenación de cada vector v_k^{μ} con dimensionalidad $K \times \mathbf{d}$. En clusters con baja variabilidad, los vectores resultantes tienen principalmente valores cero. Este hecho resulta interesante para diferenciar patrones con similitudes dinámicas y espaciales a lo largo de las expresiones. Sin embargo, esta estrategia se limita a capturar la distribución local de los descriptores y es variante a los movimientos con distribuciones simétricas. Por ejemplo, los mismos vectores característicos pueden resultar de diversos gestos.

3.3.2 Shape difference VLAD.

Teniendo en cuenta la complejidad del problema de reconocimiento de señas, debido a la amplia variabilidad en la representación de gestos, en este trabajo se propuso una codificación inspirada en la representación *SD-VLAD* [12]. Esta estrategia permite cuantificar la forma de la distribución de los descriptores locales, hecho fundamental en el reconocimiento de señas. Este descriptor computa y utiliza la desviación estándar a cada cluster para complementar estadísticamente los vectores *VLAD* recobrando los aspectos regionales de

los descriptores que conforman cada cluster. En el esquema de representación, para cada intervalo i del vídeo, definido por el porcentaje de parches, se toma diferente cantidad de descriptores locales a la palabra más cercana K. Para alcanzar la representación de varianza en cada cluster, primero, los vectores característicos *VLAD* se ponderan por la respectiva desviación estándar del cluster K que se este tratando y se normaliza con el número de descriptores asociados a ese mismo cluster, así:

$$v_k^{\mu} = \frac{1}{n_k^i} \sum_{j=1}^{n_k^i} \frac{(x_j^i - d_k^i)}{\sigma_k^i} \tag{3.5}$$

Donde la normalización por n_k^i simula la operación de $pooling$. Para computar el descriptor de varianza, un nuevo cluster $\hat{d_k^i}$ es considerado como proyecciones de los descriptores de los vídeos de prueba que son asignados al patrón d_k^i. Entonces, la varianza de las medias es definida como la diferencia entre el nuevo $\hat{d_k^i}$ estimado y el patrón predominante o palabra d_k^i en un particular intervalo acumulado i, definido como:

$$\begin{aligned} v_{k,\mu}^i = \frac{1}{n_k^i} \sum_{j=1}^{n_k^i} (x_k^i - d_k^i) = \frac{1}{n_k^i} (\sum_{j=1}^{n_k^i} (x_j^i) - n_k^i d_k^i) \\ = \frac{1}{n_k^i} \sum_{j=1}^{n_k^i} (x_j^i) - d_k^i = \hat{d_k^i} - d_k^i \end{aligned} \tag{3.6}$$

Este descriptor estadístico v_k^i es codificado en cada intervalo i con respecto al particular diccionario acumulado D^i. Del mismo análisis, un nuevo vector es agregado al descriptor computando la diferencia entre las desviaciones estándares, como;

$$v_{k,\sigma}^i = \hat{\sigma_k^i} - \sigma_k^i = \left(\frac{1}{n_k^i} \sum_{j=1}^{n_k^i} (x_j^i - d_k^i)^2 \right)^{\frac{1}{2}} - \sigma_k^i \tag{3.7}$$

donde $\hat{\sigma_k^i}$ es la desviación estándar de los nuevos descriptores asociados a la palabra d_k^i y σ_k^i es la desviación estándar del cluster K. Tales diferencias recuperan la forma de la distribución de los descriptores locales.

El descriptor final $\mathbf{x}_f$ SD-VLAD es formado por la concatenación de los vectores característicos v_μ^i y v_σ^i para el intervalo i. La dimensión resultante es $K \times \mathbf{d} \times 2$. Finalmente se aplica la siguiente normalización $f(\mathbf{x}) = sign(\mathbf{x})|\mathbf{x}|^{\frac{1}{2}}$ sobre cada dimensión del descriptor final $\mathbf{x}_f$ sugerida en [21]. Con este método se obtiene una representación tanto para reconocimiento en cada intervalo i del vídeo como para clasificación, considerando solo un único intervalo i.

3.4 MAQUINA DE SOPORTE VECTORIAL

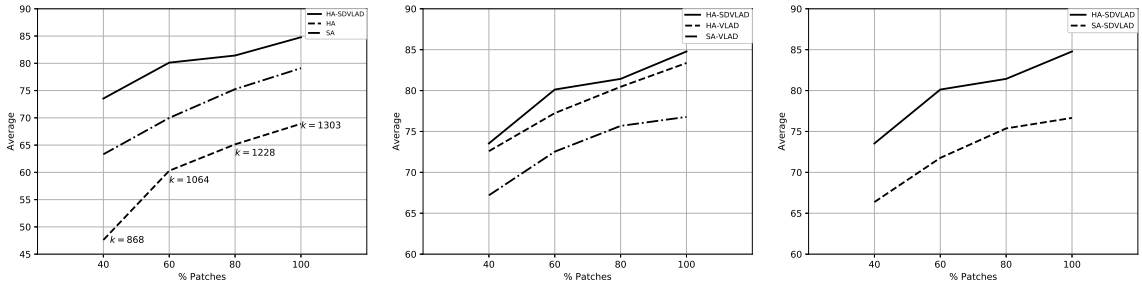
El reconocimiento de cada seña se realiza mediante Máquinas de soporte vectorial (SVM) [22] ya que estas constituyen un equilibrio adecuado entre precisión y bajo costo computacional. El presente enfoque fué implementado usando el esquema *Uno contra uno para clasificación multiclase* con el kernel *Radial Basis Function (RBF)*. Aquí las clases representan las señas particulares descritas por SD-VLAD para lo cual son separadas con hiperplanos óptimos usando la formulación clásica de margen máximo. Para m clases, la estrategia de mayor votación es aplicada para la salida los clasificadores binarios $\frac{m(m-1)}{2}$. Teniendo en cuenta que nuestra representación constituye varias aproximaciones parciales SD-VLAD, para cada intervalo definido se construye un modelo SVM particular. Adicional, se realizó el análisis de sensibilidad para determinar los mejores parámetros (γ, C), utilizando una grilla de búsqueda en un esquema de validación cruzada seleccionando las iteraciones con mayor número de positivos verdaderos.

Capítulo 4
EVALUACIÓN Y RESULTADOS

Para evaluar el enfoque propuesto hemos utilizado el corpus público **LSA64** [23]. Es un dataset de gestos en el lenguaje de señas argentino, ideal para trabajar en reconocimiento de palabras ya que solo contiene señas aisladas o no continuas. Este dataset contiene un total de 64 señas, estas señas son realizadas por un conjunto de 10 personas no expertas. Cada seña es repetida 5 veces por cada actuador conteniendo un total de 3200 vídeos. La resolución espacial de los vídeos es de 1920×1080 grabados a 60 frames por segundo. Este conjunto de datos involucra movimientos articulados donde se utilizan una o dos manos. Los videos fueron capturados en diferentes escenarios, con algunos cambios de iluminación. Para su evaluación la resolución de los vídeos se redujo a 346×194. Para evaluar el enfoque de reconocimiento en-línea, se consideraron 5 diferentes intervalos en tiempo donde se obtenía información parcial de los gestos, osea por cada 20% de los vídeos se construyó un diccionario y las representaciones SD-VLAD respectivas. Todos los experimentos son computados con parches no sobrelapados de dimensión $15 \times 15 \times 5$ con histogramas de 7 bins para (HOG), (HOOF), MBH por cada dirección y de 5 bins para HDIV y HROT. Un total de 38 bins conforma el descriptor de cada parche. Adicionalmente, Las estrategias de clasificación y reconocimiento fueron validadas usando validación cruzada con 10 iteraciones. En cada iteración, un actor y todas sus correspondientes señas se usaba para probar mientras que los 9 restantes se utilizaron para entrenar el modelo.

El primer experimento consistió en evaluar el comportamiento del método propuesto en términos de clasificación de señas con respecto a diferentes estrategias de codificación. En este caso, para una secuencia completa de video se asigna una única etiqueta que corresponde con la seña más probable. Adicionalmente se dividieron las pruebas en dos grupos. Primero, solo considerando características cinemáticas y después agregando la caracterización de gradientes de apariencia sobre la mejor respuesta obtenida. En la figura 6 se ilustra el rendimiento de las diferentes estrategias de codificación y representación. En la sub-figura 6 a la izquierda se ilustra la propuesta de asignación rígida HA + SD-VLAD con respecto a la asignación rígida (HA)

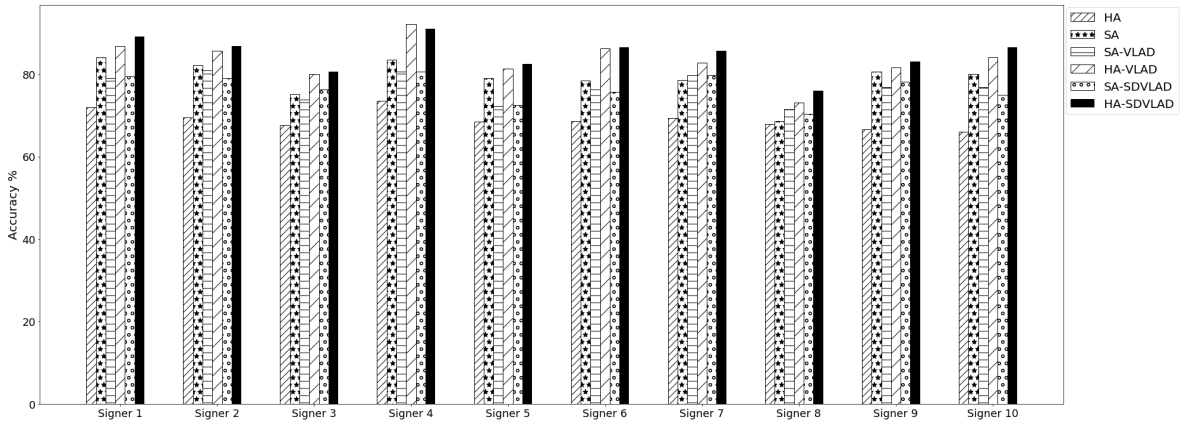
Figure 6. Rendimiento del descriptor usando diferentes estrategias de representación. Esta figura ilustra el rendimiento utilizando los descriptores de orden cero, primer orden y segundo orden sobre las estrategias de codificación rígida y suave tomando subconjuntos de parches de orden aleatorio para analizar la clasificación a información parcial e incompleta.



y asignación suave (SA). El número K de clusters viene dado por la ecuación $K = (\frac{patches}{2})^{\frac{1}{2}}$. En la sub-figura 6 del medio, se muestra el rendimiento de usar descriptores de primer orden utilizando las combinaciones clásicas con la representación VLAD. Finalmente en la parte derecha de la figura 6 se contrasta nuestra contribución contra la asignación suave en el paso de codificación, utilizando un $K = 64$. Como se esperaba en todas los experimentos el mejor rendimiento es alcanzado cuando se realiza una asignación rígida (HA) y SD-VLAD como estrategia de representación. Este resultado es debido a que con esta configuración se maneja de manera adecuada la codificación de los patrones salientes en sentidos locales y regionales manteniendo la independencia entre descriptores y así evitando que los patrones locales se confundan. Es importante notar que el esquema usado es capaz de clasificar adecuadamente las señas con información incompleta, elemento esencial para avanzar a reconocimiento en-línea. Por ejemplo únicamente con el 40% de la información se obtiene un 70% de precisión, con un 60% se obtiene ya resultados elevados de aproximadamente 80% y finalmente como rendimiento máximo se obtuvo un 85% en promedio usando toda la información.

Para profundizar los experimentos realizados, analizamos los resultados por actor en los diferentes esquemas usados (figura 7). En todos lo actores el mejor rendimiento fué alcanzado con HA-SDVLAD. Particularmente, para el actor 8 existen algunas limitaciones de representación debido a variaciones grandes en los descriptores de movimiento ocasionadas por el movimiento del cabello del actor. Estos movimientos indeseados en la representación ocupan un 25% del área capturada, influyendo drásticamente en la dinámica capturada. También se evaluó el comportamiento por características independientes, como se ilustra en la Figura 8. En este experimento se observa que los límites de movimiento (MBH) tienen un mayor impacto en la clasificación. Este hecho puede ser justificado debido a que los gestos tienen aceleraciones espaciales pre-

Figure 7. Análisis individual para cada actor usando las diferentes estrategias de representación en LSA64.



dominantes. Además la dimensión de su descripción está relacionada en términos de los componentes espaciales (x, y). En cuanto a la caracterización global, el hecho de adicionar al descriptor características de forma, incrementó el reconocimiento en un 3%. Sin embargo el mejor resultado en el reconocimiento fué logrado al utilizar únicamente características de movimiento, lo cual comprueba la hipótesis de descripción cinemática de gestos. En el enfoque global, el calcular características de forma en todo el sujeto puede inducir a confundir gestos con posturas similares. La figura 9 muestra la matriz de confusión para el enfoque propuesto, la estrategia logra clasificar de manera perfecta algunas clases, obteniendo rendimiento para varios actores del 90%. Particularmente hay una confusión considerable en la seña "Realizar" y "Comprar" debido a una similitud considerablemente grande entre estas señas.

En lenguaje de señas es bien conocido que los gestos desarrollados por las manos contienen la información más relevante del gesto. Basado en la premisa que las manos son el principal articulador, como se muestra en las imágenes 10, se procedió a realizar una segmentación de las manos y cálculo de la estrategia propuesta únicamente en estas regiones (ver Figura 10). En este experimento se alcanza un rendimiento promedio de 74% para la información de movimiento y 76% para información de movimiento y forma, (ver figura 11). En ese sentido podemos decir que la manos aportan la mayor información pero otros articuladores como los hombros, brazos y cara dan riqueza al descriptor para clasificar adecuadamente las señas. Además, las características de forma tienen un aporte positivo ya que capturan la forma de las manos a diferencia de los resultados anteriores donde tomaba el contorno de otros elementos irrelevantes. Es importante mencionar que con la sustracción de fondo se utilizaron aproximadamente el 22% de los parches y con la segmentación de manos se redujo al 5%, lo que facilita su implementación en escenarios de tiempo real.

Figure 8. Análisis individual para cada actor usando las diferentes características cinemáticas y geométricas en LSA64.

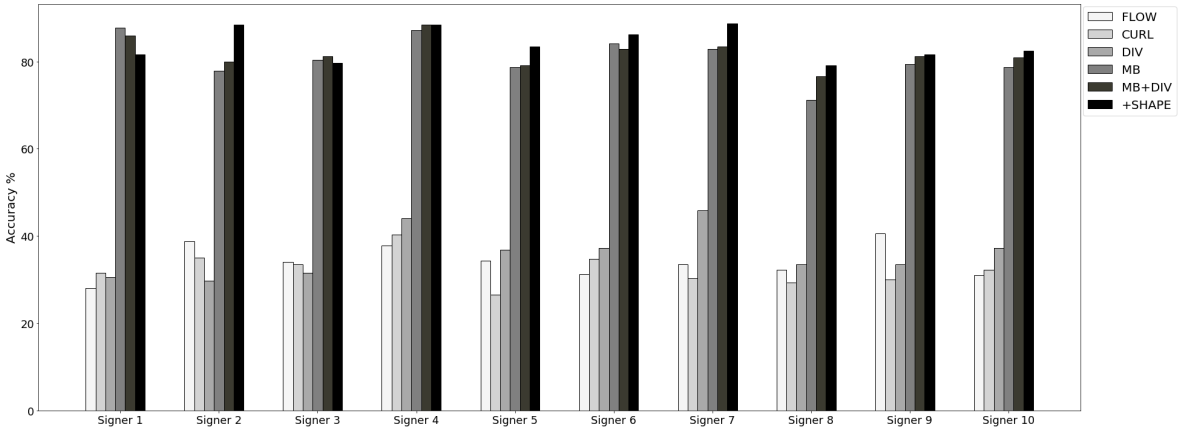


Figure 9. Matriz de confusión obtenida para el dataset LSA64. El enfoque propuesto alcanza una exactitud promedio de 85%.

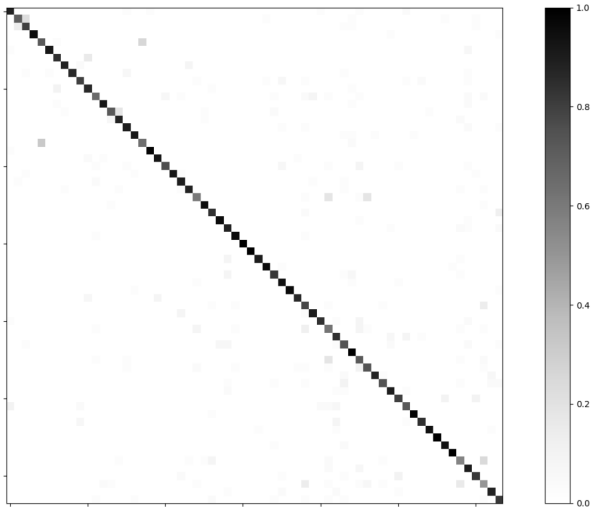
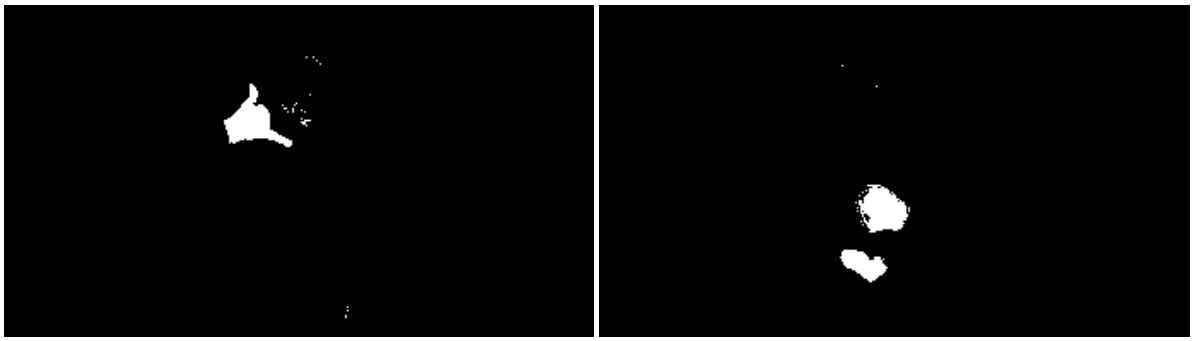


Figure 10. Mascarás usadas en la segmentación de manos. En la izquierda se muestra la máscara para las señas que utilizan una mano y en la derecha un ejemplo cuando se utilizan ambas manos.



Finalmente, en una tercera sección de experimentos se evaluó el rendimiento del enfoque para la tarea de reconocimiento en-línea. Para esto se consideraron dos orientaciones. En el primer enfoque se crearon varios intervalos del vídeo. En donde cada intervalo del 20% acumulado temporalmente del vídeo se generaba un representación SD-VLAD codificada con un único diccionario entrenado con todos los parches de los vídeos de entrenamiento, de hecho es el mismo diccionario utilizado en la tarea de clasificación. En la parte izquierda de la figura 12 se ilustra el rendimiento alcanzado por el enfoque de reconocimiento propuesto, como se esperaba el rendimiento es malo en los intervalos iniciales del vídeo porque hay representaciones sparse que no se tienen en cuenta en la fase de entrenamiento. Sin embargo, para el 60% de el vídeo el enfoque alcanza un rendimiento del 70% en promedio para la exactitud, este resultado resulta ser interesante porque deja en evidencia que con la mitad de la información de la seña se obtiene un reconocimiento aceptable además de encontrar la mayor ganancia en el intervalo de 40 a 60%.
Para la segunda orientación, se consideró una versión completa del enfoque. En tal caso, se entrenaron t diccionarios utilizando información parcial y acumulada de los vídeos. Para las señas se generó un representación SD-VLAD cada 20% como el experimento anterior pero siendo codificada con el respectivo diccionario que se había entrenado en el mismo intervalo de tiempo correspondiente y para el reconocimiento se entrenaba un modelo SVM para cada partición temporal, como se ilustra en la figura 5. Esta orientación alcanza resultados competitivos incluso en los primeros intervalos del vídeo como se muestra en la parte derecha de la figura 12. Por ejemplo, usando solo el 20% y 40% del vídeo la estrategia alcanza en promedio 53.8% y 66.7% respectivamente. Tales intervalos corresponden aproximadamente a 12 frames de las señas grabadas en los vídeos. Finalmente se alcanza el mismo rendimiento máximo pero se mejoran los resultados siendo más estables en etapas tempranas del vídeo con lo que se consigue una aproximación al reconocimiento en línea. En la figura 13 se muestra el rendimiento con una franja

Figure 11. Análisis de rendimiento individual usando segmentación de manos. En la figura se muestra el rendimiento individual, la exactitud más alta se obtuvo con características de movimiento y forma para el actor 4 con un 84%.

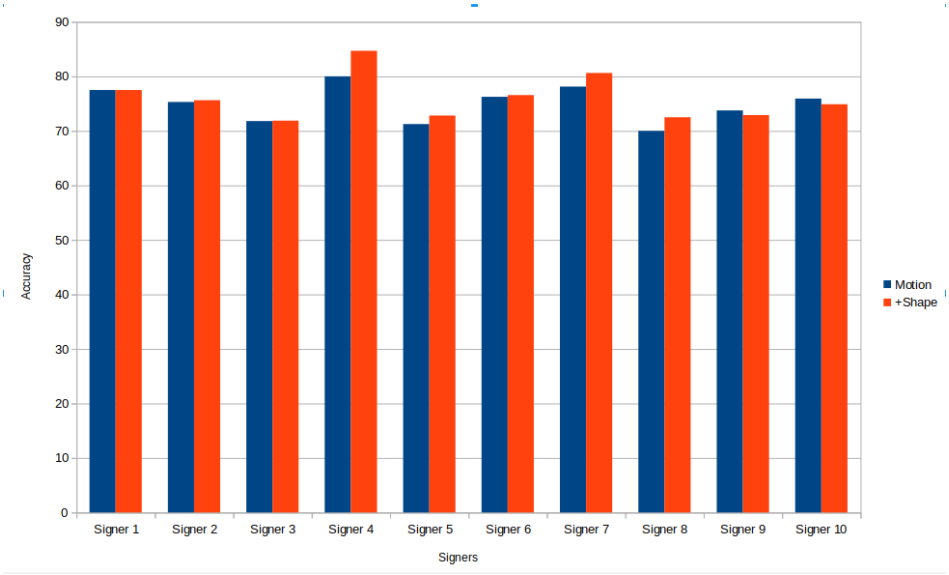
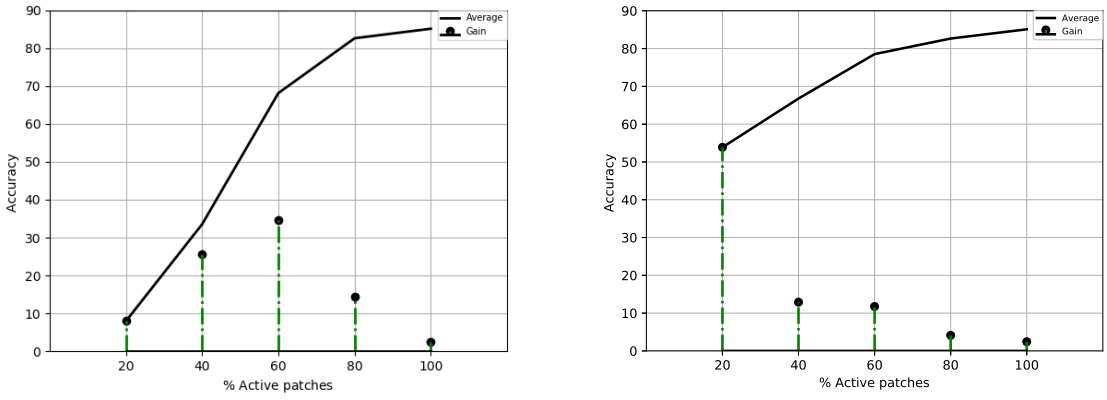
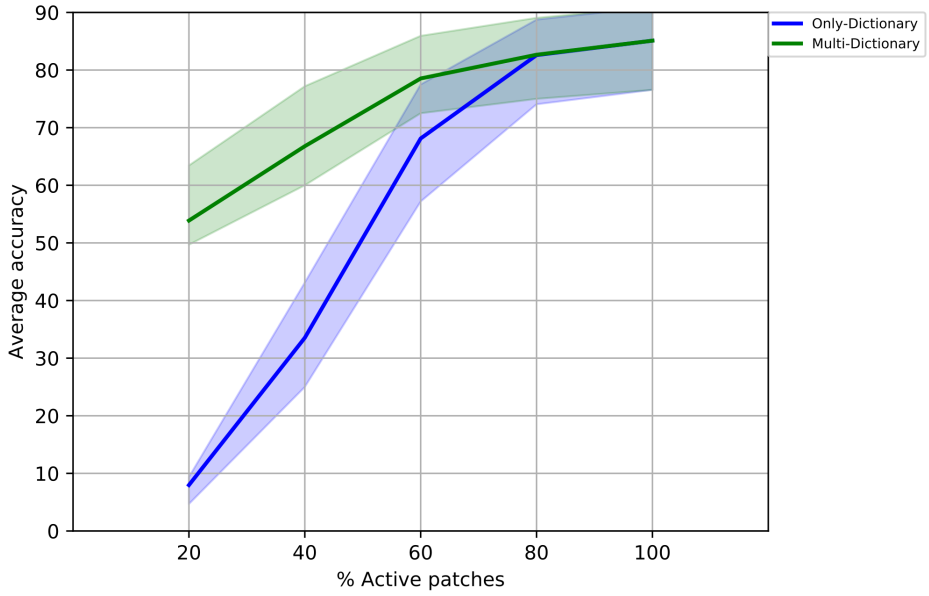


Figure 12. Rendimiento del enfoque propuesto para el reconocimiento temporal. En la izquierda se iluestra el rendimiento usando un solo diccionario. En la derecha se completa el enfoque usando t diferentes diccionarios y en tal caso t modelos SVM para el reconocimiento.



que indica el puntaje mínimo y máximo para cada enfoque.

Figure 13. Comparación entre los dos métodos propuestos para la tarea de reconocimiento. Se ilustra el rendimiento promedio para las dos orientaciones propuestas donde se evidencia los resultados mínimos y máximos para cada enfoque.



Capítulo 5
CONCLUSIONES Y PERSPECTIVAS

En este trabajo fué presentado una novedosa estrategia para clasificación y reconocimiento de señas a partir de la representación estadística de segundo orden. En esta estrategia, se cuantificaron parches temporales con características de movimiento y forma. Luego, esta representación local fue codificada con un esquema de diccionarios únicos y parciales. El descriptor final fue obtenido basado en una codificación de forma llamada *Shape Difference Vector of Locally Aggregated descriptors*, que permite recuperar la forma de las distribuciones, en los centroides más relevantes. El enfoque es robusto a oclusiones y recupera regiones salientes de los gestos codificados. Como rendimiento promedio máximo para el dataset LSA64 con más de 3000 videos se alcanzó 85% de exactitud. Mientras que en el reconocimiento temporal para el enfoque multidiccionario con múltiples representaciones SD-VLAD y modelos SVM una exactitud de 80% con el 60% de información. El enfoque propuesto es apropiado para ser usado en aplicaciones en-línea requiriendo pocos frames para calcular dichas representaciones. Trabajos futuros incluyen una evaluación recursiva a nivel de frame adecuando más el modelo al enfoque en-línea, segmentaciones temporales y el uso de otros datasets.

CONTRIBUCIONES

- J.Rodríguez, F.Martínez, "A kinematic gesture representation based on Shape Difference VLAD for Sign language recognition". International Conference on Computer Vision and Graphics, ICCVG, 2018. **Aceptado**.
- J.Rodríguez, F.Martínez, "Towards on-line sign language recognition using cumulative SD-VLAD descriptors". Congreso Colombiano de Computación, 13CCC, 2018. **Aceptado**.
- J.Rodríguez, F.Martínez, "Reconocimiento de gestos en el lenguaje de señas utilizando información espacial de primer y segundo orden representada mediante la estrategia BOW". II Congreso internacional de investigación (UDI), 2017.

REFERENCIAS

[1] centre, W. M. Deafness and hearing loss, Mar. 2018.

[2] Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H., and Presti, P. American sign language recognition with the kinect. In *Proceedings of the 13th international conference on multimodal interfaces* (2011), ACM, pp. 279–286.

[3] Wan, J., Ruan, Q., Li, W., and Deng, S. One-shot learning gesture recognition from rgb-d data using bag of features. *The Journal of Machine Learning Research 14*, 1 (2013), 2549–2582.

[4] Martínez, F., Manzanera, A., Gouiffès, M., and Braffort, A. A gaussian mixture representation of gesture kinematics for on-line sign language video annotation. In *International Symposium on Visual Computing* (2015), Springer, pp. 293–303.

[5] Paulraj, M., Yaacob, S., Desa, H., Hema, C., Ridzuan, W. M., and Ab Majid, W. Extraction of head and hand gesture features for recognition of sign language. In *Electronic Design, 2008. ICED 2008. International Conference on* (2008), IEEE, pp. 1–6.

[6] Tofighi, G., Monadjemi, S. A., and Ghasem-Aghaee, N. Rapid hand posture recognition using adaptive histogram template of skin and hand edge contour. In *Machine Vision and Image Processing (MVIP), 2010 6th Iranian* (2010), IEEE, pp. 1–5.

[7] Zahedi, M., Keysers, D., and Ney, H. Appearance-based recognition of words in american sign language. *Pattern recognition and image analysis* (2005), 373–384.

[8] Konecnỳ, J., and Hagara, M. One-shot-learning gesture recognition using hog-hof. *Journal of Machine Learning Research 15* (2014), 2513–2532.

[9] Masood, S., Srivastava, A., and Ahmad, M. Real-time sign language gesture (word) recognition from video sequences using cnn and rnn. In *Intelligent Engineering Informatics*. Springer, 2018, pp. 623–632.

[10] Wu, J., Tian, Z., Sun, L., Estevez, L., and Jafari, R. Real-time american sign language recognition using wrist-worn motion and surface emg sensors. In *Wearable and Implantable Body Sensor Networks (BSN), 2015 IEEE 12th International Conference on* (2015), IEEE, pp. 1–6.

[11] Liu, Z., Huang, F., Tang, G. W. L., Sze, F. Y. B., Qin, J., Wang, X., and Xu, Q. Real-time sign language recognition with guided deep convolutional neural networks. In *Proceedings of the 2016 Symposium on Spatial User Interaction* (2016), ACM, pp. 187–187.

[12] Duta, I. C., Uijlings, J. R., Ionescu, B., Aizawa, K., Hauptmann, A. G., and Sebe, N. Efficient human action recognition using histograms of motion gradients and vlad with descriptor shape information. *Multimedia Tools and Applications*, 1–28.

[13] Brox, T., Bregler, C., and Malik, J. Large displacement optical flow. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (2009), IEEE, pp. 41–48.

[14] Jain, M., Jegou, H., and Bouthemy, P. Better exploiting motion for better action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2013), pp. 2555–2562.

[15] Ali, S., and Shah, M. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence 32*, 2 (2010), 288–303.

[16] Dalal, N., Triggs, B., and Schmid, C. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision* (2006), Springer, pp. 428–441.

[17] Chaudhry, R., Ravichandran, A., Hager, G., and Vidal, R. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (2009), IEEE, pp. 1932–1939.

[18] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (2005), vol. 1, IEEE, pp. 886–893.

[19] Peng, X., Wang, L., Wang, X., and Qiao, Y. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding 150* (2016), 109–125.

[20] Jégou, H., Douze, M., Schmid, C., and Pérez, P. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (2010), IEEE, pp. 3304–3311.

[21] Perronnin, F., Sánchez, J., and Mensink, T. Improving the fisher kernel for large-scale image classification. *Computer Vision–ECCV 2010* (2010), 143–156.

[22] Chang, C.-C., and Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST) 2*, 3 (2011), 27.

[23] Ronchetti, F., Quiroga, F., Estrebou, C. A., Lanzarini, L. C., and Rosete, A. Lsa64: An argentinian sign language dataset. In *XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016).* (2016).

BIBLIOGRAFIA

ALI, Saad; SHAH, Mubarak. Human action recognition in videos using kinematic features and multiple instance learning. IEEE transactions on pattern analysis and machine intelligence, 2010, vol. 32, no 2, p. 288-303.

BROX, Thomas; BREGLER, Christoph; MALIK, Jitendra. Large displacement optical flow. En Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009. p. 41-48.

CHANG, Chih-Chung; LIN, Chih-Jen. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2011, vol. 2, no 3, p. 27.

CHAUDHRY, Rizwan, et al. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. En computer vision and pattern recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009. p. 1932-1939.

DALAL, Navneet; TRIGGS, Bill; SCHMID, Cordelia. Human detection using oriented histograms of flow and appearance. En European conference on computer vision. Springer, Berlin, Heidelberg, 2006. p. 428-441.

DALAL, Navneet; TRIGGS, Bill. Histograms of oriented gradients for human detection. En Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005. p. 886-893.

DUTA, Ionut C., et al. Efficient human action recognition using histograms of motion gradients and VLAD with descriptor shape information. Multimedia Tools and Applications, 2017, vol. 76, no 21, p. 22445-22472.

JAIN, Mihir; JEGOU, Herve; BOUTHEMY, Patrick. Better exploiting motion for better action recognition. En Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013. p. 2555-2562.

JÉGOU, Hervé, et al. Aggregating local descriptors into a compact image representation. En Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010. p. 3304-3311.

KONECNÝ, Jakub; HAGARA, Michal. One-shot-learning gesture recognition using hog-hof. Journal of Machine Learning Research, 2014, vol. 15, p. 2513-2532.

LIU, Zhengzhe, et al. Real-time Sign Language Recognition with Guided Deep Convolutional Neural Networks. En Proceedings of the 2016 Symposium on Spatial User Interaction. ACM, 2016. p. 187-187.

MARTÍNEZ, Fabio, et al. A Gaussian mixture representation of gesture kinematics for on-line Sign Language video annotation. En International Symposium on Visual Computing. Springer, Cham, 2015. p. 293-303.

MASOOD, Sarfaraz; SRIVASTAVA, Adhyan; AHMAD, Musheer. Real-Time Sign Language Gesture (Word) Recognition from Video Sequences Using CNN and RNN. Intelligent Engineering Informatics, 2018, p. 623-632.

PAULRAJ, M. P., et al. Extraction of head and hand gesture features for recognition of sign language. En Electronic Design, 2008. ICED 2008. International Conference on. IEEE, 2008. p. 1-6.

PENG, Xiaojiang, et al. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding, 2016, vol. 150, p. 109-125.

PERRONNIN, Florent; SÁNCHEZ, Jorge; MENSINK, Thomas. Improving the fisher kernel for large-scale image classification. En European conference on computer vision. Springer, Berlin, Heidelberg, 2010. p. 143-156.

RONCHETTI, Franco, et al. Lsa64: An argentinian sign language dataset. En XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016). 2016.

TOFIGHI, Ghassem; MONADJEMI, S. Amirhassan; GHASEM-AGHAEE, Nasser. Rapid hand posture recognition using adaptive histogram template of skin and hand edge contour. Machine Vision and Image Processing (MVIP), 2010 6th Iranian, 2010, p. 1-5.

WAN, Jun, et al. One-shot learning gesture recognition from RGB-D data using bag of features. The Journal of Machine Learning Research, 2013, vol. 14, no 1, p. 2549-2582.

WU, Jian, et al. Real-time American sign language recognition using wrist-worn motion and surface EMG sensors. En Wearable and Implantable Body Sensor Networks (BSN), 2015 IEEE 12th International Conference on. IEEE, 2015. p. 1-6.

ZAFRULLA, Zahoor, et al. American sign language recognition with the kinect. En Proceedings of the 13th international conference on multimodal interfaces. ACM, 2011. p. 279-286.

ZAHEDI, Morteza; KEYSERS, Daniel; NEY, Hermann. Appearance-based recognition of words in american sign language. En Iberian Conference on Pattern Recognition and Image Analysis. Springer, Berlin, Heidelberg, 2005. p. 511-519.