CÁLCULO DE PRIMITIVAS DENSAS DE MOVIMIENTO EN 3D UTILIZANDO IMÁGENES RGB-D

FABIÁN ALONSO CASTILLO CUADRA

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2018

CÁLCULO DE PRIMITIVAS DENSAS DE MOVIMIENTO EN 3D UTILIZANDO IMÁGENES RGB-D

FABIÁN ALONSO CASTILLO CUADRA

Trabajo de grado para optar por el título de Ingeniero de sistemas

Director:
FABIO MARTÍNEZ CARRILLO, PH.D
Profesor escuela de ingeniería de sistemas e informática (EISI)

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2018

CONTENIDO

INTRODUCCIÓN **10**

1 METODOLOGÍA **13**

1.1 CARACTERIZACIÓN DEL FLUJO DE ESCENA EN SECUENCIAS RGB-D 13

1.2 CÁLCULO DE TRAYECTORIAS LARGAS DE MOVIMIENTO 16

1.3 CARACTERÍSTICAS CINEMÁTICAS DE LAS TRAYECTORIAS 18

1.4 REPRESENTACIÓN PARA EL RECONOCIMIENTO 19

1.5 DATASET PROPUESTO 20

2 EXPERIMENTOS Y RESULTADOS **22**

3 CONCLUSIONES **27**

REFERENCIAS **28**

BIBLIOGRAFIA **31**

LISTA DE FIGURAS

Figura 1 Flujo de escena. 14
Figura 2 Trayectorias 3D+t de movimiento 17
Figura 3 Reconocimiento a partir de trayectorias 3D+t 19
Figura 4 Dataset propuesto 21
Figura 5 Trayectorias 3D para el gesto aro. 22
Figura 6 Resultados cuantitativos 23
Figura 7 Precisión vs centroides 25
Figura 8 Matriz de confusión 26

LISTA DE TABLAS

Tabla 1 Clasificación de reconocimiento utilizando gestos para 4 personas, entre 3 clases y utilizando 8 bins en el descriptor final. Obteniendo una media de 76 %. 23
Tabla 2 Clasificación de reconocimiento de gestos para 4 personas, entre 4 clases y utilizando 8 bins en el descriptor final. Obteniendo una media de 80 %. 24
Tabla 3 Clasificación de reconocimiento de gestos para 4 personas, entre 5 clases y utilizando 8 bins en el descriptor final. Obteniendo una media de 77 %. 24
Tabla 4 Clasificación de reconocimiento de gestos para 3 personas, entre 5 clases y utilizando 8 bins en el descriptor final. Obteniendo una media de 75.33 %. 25

RESUMEN

Título: CÁLCULO DE PRIMITIVAS DENSAS DE MOVIMIENTO EN 3D UTILIZANDO IMÁGENES RGB-D[1]
Autor: FABIÁN ALONSO CASTILLO CUADRA[2]
Palabras Clave: RGBD, flujo de escena, trayectorias.

DESCRIPCIÓN:

Los sensores RGB-D permiten capturar escenas 3D, codificando y haciendo una correlación entre la profundidad y la clásica información de apariencia (RGB). En la literatura se han reportado diversas estrategias para representar la información de la escena, pero estas requieren complejos procesos de calibración y asumen observaciones independientes. En cuanto a la caracterización del movimiento, las típicas estrategias utilizadas están limitadas a capturar el flujo de escena para describir el movimiento local. Sin embargo, estas estrategias sólo capturan información de movimiento entre dos pares consecutivos de imágenes. Limitando el análisis coherente en largos desplazamientos en el tiempo. En este trabajo, se presenta una novedosa estrategia para calcular primitivas densas de movimiento 3D como primitivas cinemáticas fundamentales para representar secuencias de video. El enfoque propuesto empieza calculando flujos densos de movimiento para capturar los campos de velocidad aparente en cada cuadro. Luego, se realiza un muestreo denso sobre una grilla en la que se seleccionan un conjunto de puntos espaciales que son seguidos de acuerdo a la información de velocidad local. Estos puntos seguidos son filtrados utilizando un kernel mediano para remover el ruido del movimiento presente en periodos cortos de tiempo. Cada trayectoria contiene información coherente de movimiento la cual es caracterizada mediante el cálculo de primitivas cinemáticas de movimiento. Donde cada trayectoria representa primitivas cinemáticas que juntas describen acciones complejas realizadas en secuencias de video. Estás características cinemáticas fueron procesadas en una metodología Bag of Words (BoW) para obtener histogramas que describen videos. Finalmente, se validó el método propuesto a través de un nuevo dataset con 5 acciones y 100 videos. El descriptor basado en trayectorias de movimiento 3D+t alcanzó una precisión promedio de 77%.

---

[1] Trabajo de Grado
[2] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Fabio Martínez Carrillo, Ph.D, PhD.

ABSTRACT

Title: 3D+t DENSE MOTION TRAJECTORIES AS KINEMATIC PRIMITIVES TO ANALYZE DEPTH VIDEO SEQUENCES [1]
Author: FABIÁN ALONSO CASTILLO CUADRA[2]
Keywords: RGB-D, scene flows, dense motion trajectories, tracking.

DESCRIPTION:

RGB-D sensors are able to capture 3D scenes by coding and correlating depth and classical optical RGB information. Such novel representation have allowed attack many classical problems in computer vision such as segmentation, scene representations and human interaction, among much others. In the literature have been reported many strategies to represent scene information but requiring complex calibration process and assuming independent observations at each time. Regarding motion characterization, typical strategies are limited to namely analyze global shape changes and capture scene flow fields to describe local motions in RGBD sequences. Nevertheless, such strategies only recover motion information among a couple of frames, limiting the analysis of coherent large displacements along time. This work presents a novel strategy to compute 3D+t dense long motion trajectories as fundamental kinematic primitives to represent video sequences. The herein proposed strategy starts by computing dense flow maps to capture appareance velocity fields for each frame. Then, from a imposed grid is selected a set of spatial points that are tracked according to the local velocity information. The followed points are filtered using a moving median kernel to remove peaks of motion present in relative short periods of time. Every coherent motion trajectory was locally characterized by computing kinematic primitives such as the average of the speed and angle of motion, the curvature and torsion. Then each motion trajectory represent a kinematic word primitives that together can describe complex actions developed along videos. Such kinematic words were processed into a bag-of-kinematic-word framework to obtain a occurrence video descriptor. In a dataset with 5 gestures and 100 videos, the video descriptor based on 3D+t motion trajectories achieved an average accuracy of 77%.

---

[1] Bachelor Thesis
[2] Faculty of Physics-Mechanics Engineering. School of Systems Engineering and Informatics. Advisor: Fabio Martínez Carrillo, Ph.D, PhD.

INTRODUCCIÓN

Usualmente, las técnicas de visión por computador se encuentran basadas exclusivamente en la clásica información de intensidad (RGB). Estas depende en gran medida de la información de apariencia, la cual tiene limitantes en ciertos escenarios tales como: segmentación de objetos de colores similares, detección de escenarios dinámicos, sensibilidad a cambios de luminosidad, incluso una alta variabilidad en el reconocimiento de diferentes perspectivas. Los dispostivos de captura de RGB-D (Kinect) han permitido introducir un nuevo tipo de información para una mejor representación de escenas, la cual contiene la información complementaria de profundidad de objetos de interés. Este nuevo análisis multimodal ha permitido abordar nuevas perspectivas para resolver problemas clásicos de visión por computador como: reconstrucción en 3D, seguimiento humano, interacción hombre-computador, entre otras. Este tipo de análisis ayuda a superar problemas típicos como lo son las variaciones en iluminación y perspectiva. El cálculo de primitivas RGBD resulta de gran utilidad para el entendimiento de escenarios complejos, capturando y caracterizando de manera precisa los distintos objetos de interés en algún problema particular. Sin embargo, los sensores de profundidad son limitados en resolución además de proveer mediciones erróneas causadas por ligeras perturbaciones en el ambiente. Por otra parte, el uso de la información de profundidad es una tarea compleja debido a la correlación entre la profundidad y la apariencia capturadas a diferentes escalas, dependiendo de parámetros intrínsecos de los dispositivos de captura. Además, el cálculo de primitivas de movimiento implica la asociación de temporal de imágenes RGB a lo largo de secuencias, incrementando la complejidad de las aproximaciones computacionales.

En la literatura se han propuesto distintas estrategias que recuperan características de bajo nivel y producen descriptores basados en información RGBD. Por ejemplo, Blum *et. al.* [1] introdujo un algoritmo de clustering para recuperar patrones regionales caracterizados por puntos SIFT. Sin embargo, esta aproximación es dependiente de las

características del cluster y los puntos SIFT no utilizan la información de profundidad. En [2] se analizó la representación de escenas calculando puntos de interés en apariencia y profundidad de manera independiente. Donde interesantemente se reportó que la mejor caracterización para actividades de reconocimiento se obtiene utilizando exclusivamente características de apariencia. Sin embargo, los autores sugieren el uso de puntos de apariencia complementados con la media correspondiente de profundidad en celdas espaciales. En este sentido, en [3] se extendieron los descriptores SIFT y HOG en secuencias RGBD, alcanzando una mejora del 10% tareas de reconocimiento de objetos.

En cuanto al análisis de movimiento, existen varios trabajos seminales como el propuesto en [4], donde se analizan los cambios estimados de siluetas humanas obtenidas a través de mapas de profundidad. Estas aproximaciones eliminan las dependencias de apariencia y resultan eficientes en tiempo de cómputo. Sin embargo, un problema en esa representación global es la dependencia de la perspectiva y de las restricciones w.r.t. Por otro lado, una típica caracterización es el cálculo de la velocidad aparente entre dos cuadros consecutivos. Este campo de movimiento resulta de alguna estrategia de flujo óptico, la cual brinda el desplazamiento de los objetos caracterizados por su apariencia. Utilizando imágenes RGBD, se han propuesto diversas estrategias que estiman movimientos de escena, caracterizando el desplazamiento 3D local [5–7]. Estos flujos de escena aportan importantes primitivas cinemáticas, pero resultan propensos a errores debido la baja resolución de los mapas de profundidad, limitando la correlación con información óptica en la secuencia [8]. En [9] se propuso una aproximación del flujo de escena en la que inicialmente se calcula un clásico flujo óptico denso para luego estimar información de las velocidades 3D utilizando la información de profundidad. Sin embargo, esta aproximación se limita al incluir únicamente restricciones de apariencia, dificultando la correspondencia en profundidad. En [10] se propuso una estrategia simultánea de mapeo y localización (SLAM) para generar una nube de puntos 3D en cada instante. Donde se calculan transformaciones rigidas entre nubes de puntos consecutivas para obtener el grafo correspondiente a la estimación del flujo de escena. Esta aproximación se ha extendido al cálculo de trayectorias para el seguimiento de robots en ambientes controlados. Aunque la representación en grafos resulta computacionalmente compleja y la correspondecia entre nubes de puntos puede fallar en presencia de movimientos inesperados. Por otro lado, Herbest *et. al.* [7] propuso una estrategia para estimar el flujo de escena a partir de operadores no lineales que modelan la apariencia y profundidad. Este enfoque permite obtener flujos densos de escena inclusive en

escenarios uniformes. Una desventaja del fujo de escena es que su caracterización se limita a la descripción entre dos pares de cuadros consecutivos, limitando la descripción cinemática de objetos a primitivas de primer orden.

En este trabajo, se introduce una novedosa estrategia para calcular trayectorias largas 3D+t como primitivas cinemáticas de movimiento en secuencias RGBD. Estas trayectorias de movimiento permiten recuperar información coherente del movimiento local en secuencias utilizadas para el análisis de objetos en escenas. Se empieza calculando el flujo de escena a lo largo de la secuencia, de donde se obtienen campos de movimiento subsequentes. Luego, se realiza un muestreo denso de puntos sobre una grilla de píxeles definida y se siguen de acuerdo a los correspondientes vectores de velocidad. De la concatenación de los puntos muestreados se generan trayectorias 3D de movimiento. Cada una de las trayectorias obtenidas se carecterizó utilizando métricas cinemáticas locales, tales como la media y desviaciones de: velocidades, curvatura y torsión. Estas cinemáticas codifican palabras de movimiento que representan el movimiento de los objetos de interés capturados en videos, en donde a través de una metodología Bag-of-Words (BoW) permiten validar el desempeño de las trayectorias obtenidas. La validación del enfoque propuesto se llevo a cabo en el problema de reconocimiento de gestos capturado en secuencias de imágenes RGBD.

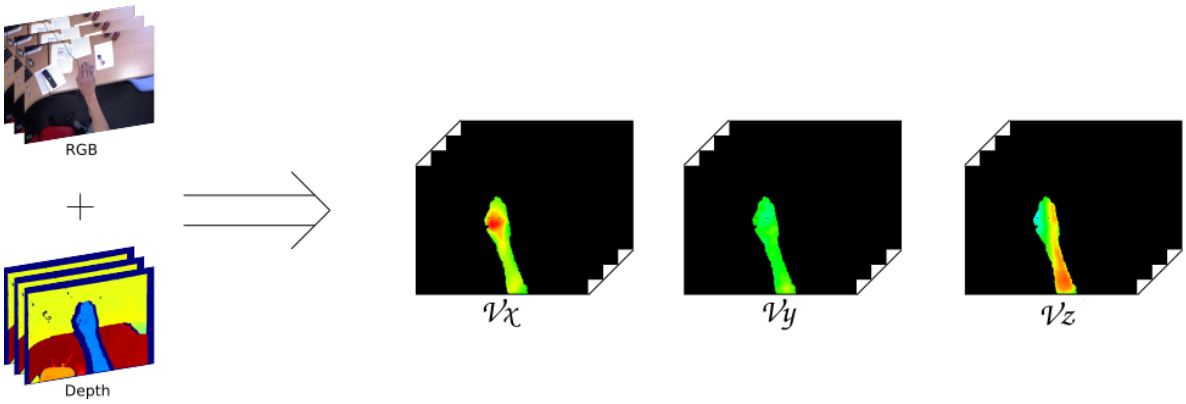
Capítulo 1
METODOLOGÍA

En este trabajo se presentó una estrategia computacional para obtener trajectorias 3D+t que caracterizan localmente secuencias RGBD. El enfoque propuesto empieza estimando los campos de movimiento 3D entre cuadros consecutivos como una caracterización inicial del flujo de escena (subsection 1.1). Estos campos permiten hacer seguimiento a un conjunto de puntos uniformes a lo largo de la secuencia constituyendo trayectorias largas de movimiento (subsection 1.2). Se representaron secuencias de video a partir del cálculo de un conjunto de cinemáticas diferenciales sobre cada trayectoria (subsection 1.3). Finalmente, se utilizó la metodología Bag of Words(BoW) para obtener el descriptor final sobre las trayectorias propuestas y validarlo en el problema de representación de gestos (subsection 1.4). Las siguientes subsecciones describen cada una de las etapas consideras en el método propuesto.

1.1 CARACTERIZACIÓN DEL FLUJO DE ESCENA EN SECUENCIAS RGB-D

El cálculo de campos de movimiento utilizando estrategias de flujo óptico es un de los problemas más relevantes para la representación de objetos de interés en visión por computador. Estas estrategias son relativamente independientes a la apariencia, resultando robustas a algunos cambios de interés y permiten obtener información adicional de características para complementar tareas de reconocimiento. Las aproximaciones clásicas para el cálculo de flujo óptico incluyen a Lucas-Kanade [11], la cual resuelve el problema lineal de encontrar vectores de movimiento sobre las esquinas de la escena. Por otro lado, Horn-Schunck [12] introduce restricciones globales y un método

Figure 1. **Flujo de escena.** A partir del flujo de escena se obtiene el movimiento 3D entre dos pares de imágenes RGBD consecutivos



variacional para obtener estimaciones densas de movimiento. En estos dos trabajos seminales se han basado cientos de aplicaciones novedosas y estrategias para cuantificar el movimiento [9] [13] [7].

A partir de secuencias de imágenes RGBD, las estrategias de caracterización de flujo de escena permiten recuperar campos de velocidades locales 3D como el conjunto de vectores de desplazamiento tridimensionales entre cuadros consecutivos. En este trabajo, se implementó una estrategia variacional que incluye restricciones locales y globales para preservar discontinuidades del movimiento y obtener estimaciones del flujo de escena [9]. Esta estrategia de flujo 3D modela el movimiento rotacional y traslacional usando restricciones locales y globales que están espacialmente relacionadas mientras que los puntos de escena permiten una mejor estimación de movimientos no rigidos de objetos, como las rotaciones tridimensionales.

Esta aproximación usa un regularizador variacional para obtener un campo de flujo denso, preservando discontinuidades de movimiento. Además, utiliza un conjunto de correspondencias 3D no locales para tratar largos desplazamientos y recuperar movimientos inesperados que representan la firma cinemática de algunos objetos. Particularmente, el modelo variacional de flujo de escena se expresa como el siguiente problema de miminización de energía:

$$E(v) = E_D(\mathbf{v}) + \alpha E_M(\mathbf{v}) + \beta E_R(\mathbf{v}) \tag{1.1}$$

donde $E_D(v)$ es una restricción local en la nube de puntos, $E_M(v)$ modela una restricción

de matching global y $E_R(v)$ actua como regularizador local en el campo de movimiento espacial. La restricción $E_D(v)$ incluye información de apariencia y profundidad de manera global. A partir de la apariencia se modela una típica restricción de brillo (BCA) como: $\rho_I(\mathbf{x};\mathbf{v}) : I_{t+1}(W(\mathbf{x};\mathbf{v})) = I_t(\mathbf{x})$ la cual considera el color de un objeto constante en una secuencia consecutiva de cuadros. Al igual que en los flujos ópticos clásicos, esta restricción permite obtener patrones de movimiento 2D pero en este caso se encuentra alineado con información de profundidad incluida por una restricción de velocidad en profundidad (DVD):

$$\rho_Z(\mathbf{x};v) : Z_{t+1}(W(\mathbf{x};v)) = Z_t(\mathbf{x}) + D^T\mathbf{v}(\mathbf{x}) \tag{1.2}$$

donde $v_z(\mathbf{x})$ denota la componente de movimiento Z en el pixel $\mathbf{x}$ y $D^T = (0,0,1)$. En esta aproximación, se incluye un penalizador *Charbonnier* para lidiar con oclusiones, señales ruidosas e inconsistencias de movimiento. A partir del penalizador *Charbonnier*, las restricciones interactúan como:

$$E_D(v) = \Sigma\Psi\left(|\rho_I(\mathbf{x};\mathbf{v})|^2\right) + \lambda\Psi\left(|\rho_Z(\mathbf{x};\mathbf{v})|^2\right) \tag{1.3}$$

El segundo término incluye una función regularizadora de matching $\alpha E_M(v)$ que recupera largos desplazamientos coherentes. Se considera una suposición no local realizando un matching entre puntos SURF 2D en cuadros consecutivos y se miden los desplazamientos locales alrededor de los puntos. La restricción de velocidad en regiones no locales fuerza a recuperar sólo patrones de movimiento coherentes. En este sentido, primero se obtiene el desplazamiento del flujo de los puntos $\mathbf{x}$ y se realiza un matching de $m(\mathbf{x})$ a $\delta_{3D}(x, m(x))$. Este flujo estimado es comparado con el calculado del análisis local $v(\mathbf{x})$ como una nueva restricción no local:

$$E_M(\mathbf{v}) = \Sigma\rho(\mathbf{x})\Psi\left(|\delta_{3D}(\mathbf{x}, m(\mathbf{x})) - v(\mathbf{x})|^2\right) \tag{1.4}$$

Finalmente, el término regularizador toma en cuenta las discontinuidades del movimiento local y define la diferencia local en el desplazamiento del flujo:

$$E_R(\mathbf{v}) = \sum_{\mathbf{x}} \omega(\mathbf{x})|\nabla v(\mathbf{x})| \tag{1.5}$$

donde $\omega(\mathbf{x})$ pondera la estimación w.r.t a la información de profundidad como: $\omega(x) = exp\left(-\alpha|\nabla Z_1(x)|^\beta\right)$. Aunque el movimiento 3D ha demostrado ser útil para describir es-

cenarios dinámicos, estas primitivas están limitadas entre pares consecutivos de cuadros. En la Figura 1 se ilustra un cálculo de flujo de escena clásico entre dos pares de cuadros. Utilizando información de apariencia y profundidad, la estrategia implementada alcanza una representación del campo 3D de la escena. Mostrando los mapas de color que representan las cantidades de las componenentes (V_x, V_y, V_z).

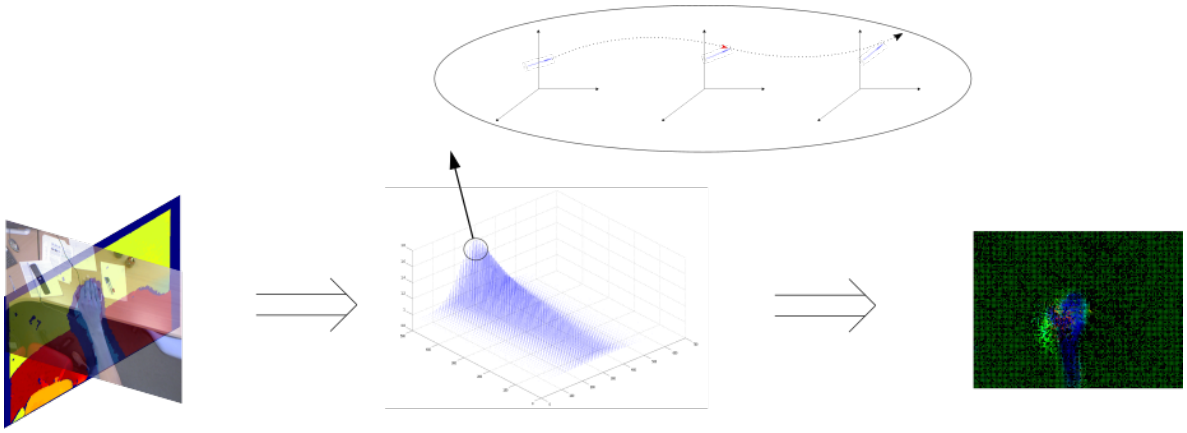
1.2 CÁLCULO DE TRAYECTORIAS LARGAS DE MOVIMIENTO

En la literatura, se ha propuesto un nuevo tipo de representación de movimiento para las típicas secuencias RGB. Esta nueva representación consiste en trayectorias de movimiento local que representan primitivas locales las cuales siguen puntos de interés a lo largo de la secuencia. Utilizando características espaciales para describir la información cinemática en más de un par de cuadros consecutivos [14, 15]. Por ejemplo, el KLT-tracker utiliza una extensión piramidal de Lucas-Kanade [16] para seguir vectores de movimiento relevantes, obteniendo trayectorias coherentes con los bordes a lo largo de la secuencia.

Esta representación es sin embargo dispersa y produce una cantidad reducida de trayectorias para representar el movimento de los objetos. También, Sun *et. al.*. Sun [14] propuso, capturar puntos de interés salientes (SIFT) a partir de un matching coherente en la secuencia. Esta aproximación, extiende el desempeño de los puntos SIFT, invariantes a escalas y rotaciones pero representan pocos puntos de la escena. Limitando la representación estadística de los objetos. Por otro lado, Wang *et. al.* propuso una representación a partir de trayectorias densas que siguen campos independientes de movimiento en videos. Esta representación ha demostrado ser exitosa en tareas de representación de acciones en las clásicas secuencias RGB.

Basandose en las trayectorias densas de movimiento, este trabajo presenta una extensión de las trayectorias densas a su cálculo en secuencias RGBD que permiten obtener largas trayectorias (3D+t). Las trayectorias propuestas enriquecen la descripción cinemática de los objetos calculando representaciones de movimiento de orden superior. Además, las trayectorias pueden utilizarse para obtener descriptores locales RGBD. La estrategia propuesta comienza calculando una nube de puntos a partir de la secuencia

Figure 2. **Trayectorias 3D+t de movimiento.** Obtención de trayectorias que describen el movimiento a partir del seguimiento al desplazamiento de puntos 3D.



RGBD para cada instante de tiempo. Esta nube asume una etapa de pre-procesamiento y calibración de las imágenes en apariencia y profundidad. Luego, se realiza un muestreo denso sobre la nube de puntos $P_t = (x_t, y_t, z_t)$ tomando una grilla de puntos espaciales distribuida en cada cuadro.

En este sentido, para cada uno de los puntos $P_t = (x_t.y_t, z_t)$ obtenidos de una grilla densa, se siguen al siguiente cuadro de acuerdo a su respectivo vector de desplazamiento $\mathbf{v} = \{v_x, v_y, v_z\}$, obtenido en el cálculo del flujo de escena. Al igual que en las estrategias basadas en RGB, entre cuadros consecutivos el vector de desplazamiento puede representar movimientos bruscos incoherentes. Por tanto, el desplazamiento de los puntos se filtra con un clásico filtro mediano: $P_{t+1} = (x_{t+1}, y_{t+1}, z_{t+1}) = (x_t, y_t, z_t) + (M * \omega)\,|_{(x_t, y_t, z_t)}$.

Donde el conjunto de puntos se sigue de acuerdo al vector de velocidad asociado para formar trayectorias de movimiento 3D $(P_t, P_{t+1}, P_{t+2}, ...)$. Una representación típica de trayectorias se muestra en la Figura 8, donde el desplzamiento (x, y) se muestra en verde. Para la componente z, cada trayectoria se colorea utilizando un mapa de color donde el verde denota desplazamientos nulos, el azul velocidades negativas (se aleja del sensor) y el rojo velocidad positivas (se acerca al sensor). Este seguimiento 3D introduce características importantes acerca de la geometría de los objetos así como del desplazamiento tridimensional de alguna acción en particular.

Siguiendo el método propuesto por Wang *et. al.*, las trayectorias de movimiento capturadas se dividen en el tiempo, siguiendo un conjunto de puntos 3D en t cuadros. Algunos filtros espaciales se implementaron para remover las trayectorias estáticas y las que presentan largos desplazamientos inespearados. Estos filtros espaciales se implementaron usando como umbral temporal la varianza de las trayectorias. Una ilustración del proceso para calcular trayectorias de movimiento 3D se muestra en al Figura 2.

1.3 CARACTERÍSTICAS CINEMÁTICAS DE LAS TRAYECTORIAS

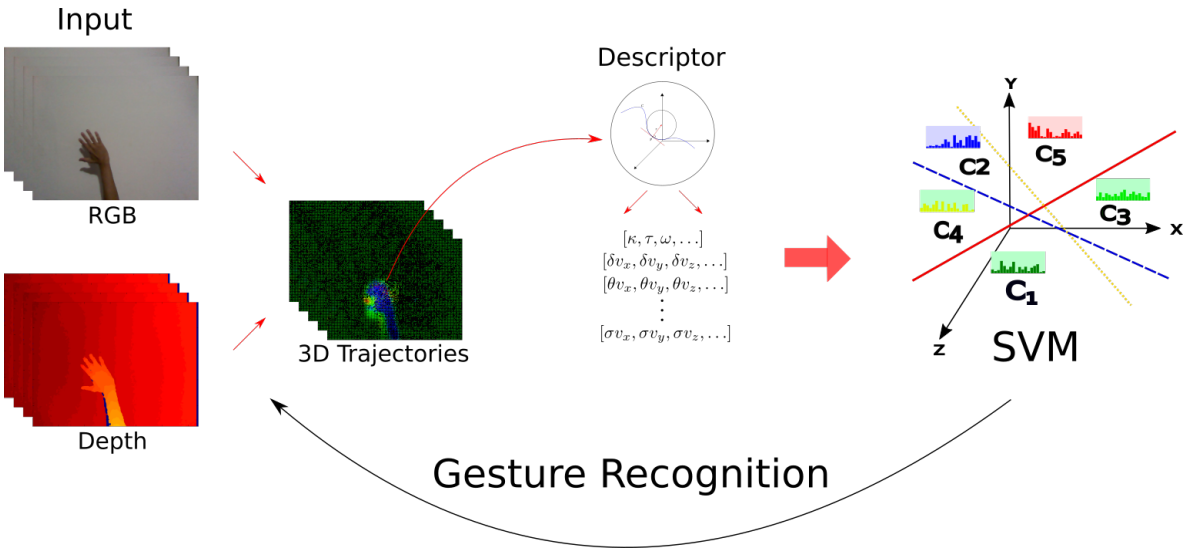
El conjunto de trayectorias (3D+t) extraídas a lo largo del video pueden usarse como palabras cinemáticas independientes para representar acciones en videos. Estas trayectorias contienen abundante información dinámica espacio-temporal y resultan fundamentales para la caracterización del movimiento. A partir de estas, se hace posible el cálculo de un conjunto de medidas cinemáticas diferenciales para obtener las palabras que representan una actividad específica.

En este trabajo, para cada trayectoria se calculó la media y desviación estandar de las velocidad locales como: $(\mu(||v_x||), \mu(||v_y||), \mu(||v_z||), \sigma(||v_x||), \sigma(||v_y||), \sigma(||v_z||))$. Debido a que las trayectorias siguen el movimiento de puntos a lo largo de la secuencia y representan información local, se hace posible el cálculo y análisis de cinemáticas adicionales. Una medida cinemática complementaria es la curvatura κ, la cual permite describir qué tan rápido se flexiona cada trayectoria durante el video. Además, en este trabajo se considero la torsión como una medida adicional del movimiento 3D de la trayectoria.

El cálculo de estas métricas se implementaron de acuerdo a [17], siguiendo diferencias Euclidianas finitas. En donde, la curvatura es la aproximación al círculo circunscrito en tres puntos consecutivos de la trayectoria P_{i-1}, P_i, P_{i+1}. En esta aproximación, se calculan tres segmentos distintos $a = ||P_{i-1} - P_i||$, $b = ||P_i - P_{i+1}||$ y $c = ||P_{i-1} - P_{i+1}||$. Luego, la curvatura κ se expresa como:

$$\kappa\left(P_i\right) = 4\frac{\triangle abc}{abc} = 4\frac{\sqrt{\hat{s}\left(\hat{s}-a\right)\left(\hat{s}-b\right)\left(\hat{s}-c\right)}}{abc}$$

Figure 3. **Reconocimiento a partir de trayectorias 3D+t.** Primero se lleva a cabo una captura de gestos en imágenes RGBD. Luego, se utiliza un flujo de escena denso y se calculan trayectorias 3D. Después, se calculan descriptores de movimiento en apariencia y profundidad. Finalmente, en se asignan los histogramas concantenados al SVM para predecir el gesto realizado.



, donde $\hat{s} = (a + b + c)/2$ y $\triangle abc$ denotan en triángulo compuesto por los segmentos (a, b, c).

1.4 REPRESENTACIÓN PARA EL RECONOCIMIENTO

El conjunto de trayectorias obtenido, constituye un conjunto de primitivas cinemáticas que representa el movimiento particular de los objetos de interés. Cada una de las trayectorias se codificó utilizando el conjunto de medidas cinemáticas como palabras representativas dentro del modelo Bag-of-kinematic words (BoKW). Este es un modelo de representación de nivel medio, ampliamente usado en diferentes áreas de interés tales como: procesamiento del lenguaje natural, reconocimiento de objetos en imágenes, entre otros. Particularmente, el conjunto de palabras cinemáticas calculado en un conjunto de videos de entrenamiento, se utilizó como entrada a un algoritmo no supervidado k-means para obtener palabras cinemáticas que representen en general el conjunto de acciones en un video [18].

Finalmente, se calculan los histogramas para todo el dataset, *i.e.*, separando los videos

en entrenamiento y pruebas. El conjunto de histogramas de entrenamiento se utilizó para crear un modelo de aprendizaje de máquina, el cual permite realizar una posterior clasificación. Para realizar las tareas de clasificación se implementó una estrategia de máquina de soporte vectorial (SVM) debido a su conocida eficacia y desempeño para obtener resultados [19]. Para clasificar un video se mapean las ocurrencias del histograma a la máquina de soporte previamente entrenada. De tal forma, se obtiene una etiqueta que asigna la clase a la que pertenece dicho gesto. La figura 3, muestra el esquema general usado para el reconocimiento particular de gestos en secuencias de imágenes RGBD a partir de las trayectorias 3D+t de movimiento propuestas.

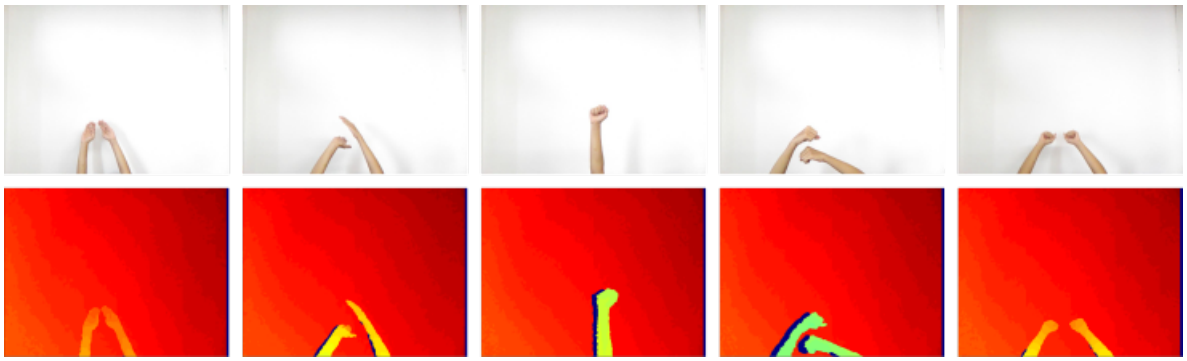
1.5 DATASET PROPUESTO

Aunque en la literatura se han reportado varios datasets RGBD para diferentes propósitos [20–22], estos se suelen capturar exclusivamente para tareas de reconocimiento estático y sólo presentan imágenes independientes que describen objetos. También, se encuentran disponibles algunos datasets que capturan descripciones de gestos, pero con fuertes limitaciones en secuencias de tiempo. En estos casos, se toman algunas imágenes dispersas para representar las poses principales de alguna acción en particular. La mayoría de los datasets RGBD se encuentran limitados a pares consecutivos de imágenes con el fin de probar nuevos algoritmos de flujo de escena.

En este trabajo se introdujo un nuevo dataset RGBD el cual resulta útil para evaluar estrategias que calculan propiedades cinemáticas a lo largo de secuencias. Para la captura del dataset se utilizó la Kinect V1 como dispositivo de captura, de donde se recuperaron las secuencias de imágenes a partir de los frameworks libres *OpenNI* y *libfreenect*. Cada secuencia de video capturada presenta una resolución espacial de 640×480 con una resolución temporal de 30 cuadros por segundo en cada canal: apariencia y profundidad.

El dataset propuesto fue capturado en un ambiente controlado, con fondo uniforme y bajos movimientos relativos de la cámara. Un total de cuatro personas fueron capturadas realizando diferentes acciones utilizando una o ambas manos. Las acciones seleccionadas involucran diferentes gestos con movimientos representativos y desplazamientos en profundidad. Cada gesto se capturó un total de cinco veces para obtener un rango de variación estadística de cada movimiento. El dataset cuenta con un to-

Figure 4. **Dataset propuesto.** En la primera fila se observa el primer frame de cada una de las acciones del dataset propuesto en el canal de intensidad, en la segunda fila su correspondiente mapa de profundidad.



tal de 100 videos. En la figura 4 se ilustra los diferentes gestos capturados en el dataset propuesto. El dataset se encuentra disponible al público y se puede descargar libremente de https://drive.google.com/file/d/1KVhftC1SeKXweaDV8xjO0vm74-19URf5/view?usp=sharing.

Capítulo 2
EXPERIMENTOS Y RESULTADOS

La estrategia propuesta, representa un conjunto de trayectorias largas de movimiento como un conjunto de primitivas cinemáticas para representar gestos capturados en secuencias de imágenes RGBD. Una primera evaluación cualitativa de las trayectorias obtenidas se ilustra en la Figura 8, donde se obtienen trayectorias para un gesto particular realizado con una mano. En este gesto, se realizó un movimiento rotacional de la mano en el eje vertical. Como se observa en la figura, la mayoría de la información importante resulta coherente con el movimiento realizado por la mano. El color de cada trayectoria se representa utilizando un código de colores. El azul representa un desplazamiento en profundidad hacía el sensor, el rojo el desplazamiento alejandose del sensor y el verde la ausencia de movimiento en profundidad. En la segunda fila, se observa las componentes de las trayectorias se pintan para cada uno de los tres ejes, mostrando información cinemática de relevancia en todas las componentes.
En términos cuantitativos, se realizó una evaluación para medir la capacidad de representación de las trayectorias 3D+t propuestas. En donde, cada trayectoria capturada de la secuencia se caracterizó a través de un conjunto de cinemáticas: velocidades, desviaciones, curvaturas. Luego, se utilizó una bolsa palabras cinemáticas (BoW) para

Figura 5. **Trayectorias 3D.** Muestra las trayectorias obtenidas para el gesto aro. El color representa el desplazamiento en la profundidad codificado en la barra de colores.

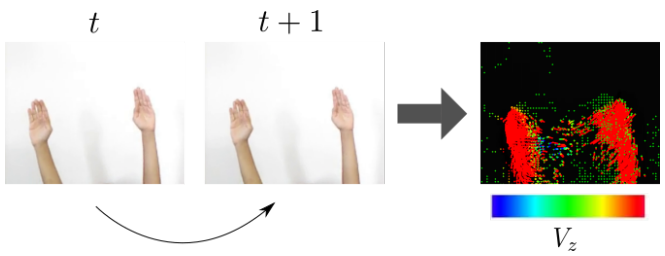


Figura 6. **Vistas trayectorias 3D.** Primera fila: Muestra las trayectorias obtenidas a través del método propuesto. Segudnda fila: Las trayectorias 3D permiten operar múltiples vistas y transformaciones, ofreciendo ventajas sustanciales para la caracterización del movimiento

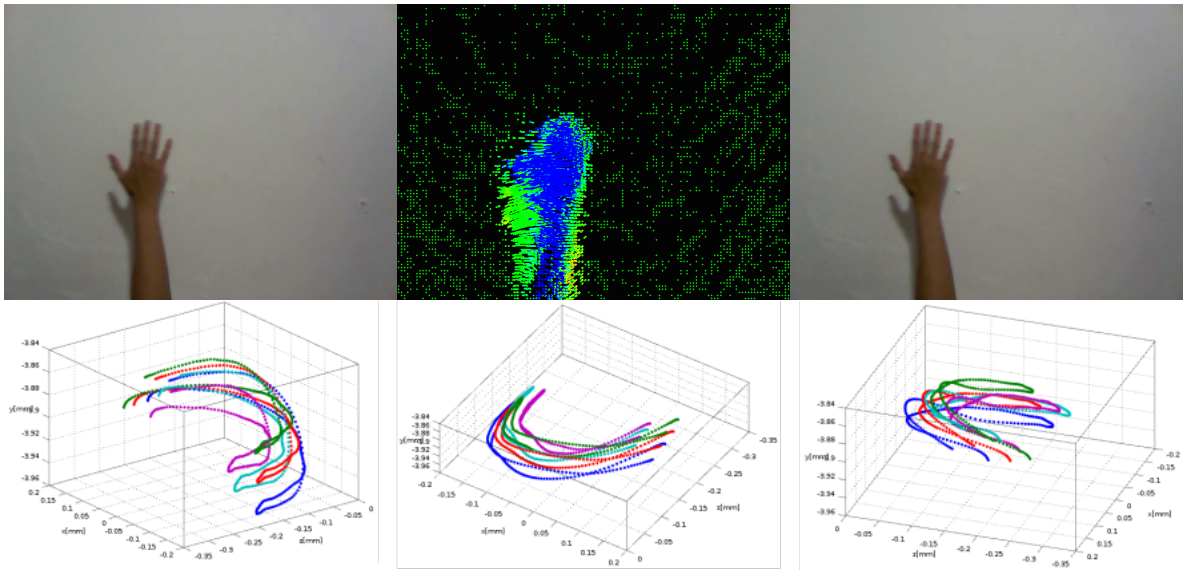


Tabla 1. Clasificación de reconocimiento utilizando gestos para 4 personas, entre 3 clases y utilizando 8 bins en el descriptor final. Obteniendo una media de 76 %.

Subjects	**Accuracy(%)**
P_1	80
P_2	64
P_3	88
P_4	72

representar la descripción de cada gesto del dataset. En este trabajo, se realizó una evaluación utilizando una estrategia de validación cruzada k-fold con $k = 20$. En un primer experimento, se evaluó la capacidad de representación de la estrategia propuesta para los tres gestos más desiguales: aro, triángulo y flexionar. En la Tabla 1 se muestra los resultados obtenidos para cada una de las personas incluida en los experimentos. De manera general, el método propuesto alcanza una precisión promedio de 77% en un total de 100 videos. Para la persona P_2 existen ciertas limitaciones debido al ruido presente en la secuencia, el cual afecta el cálculo apropiado del flujo de escena y resulta en una representación reducida de trayectorias.

En ese mismo sentido, en un segundo expermiento se incluyó un gesto adicional a la evaluación. En la Tabla 2, se reportan los resultados obtenidos para este experimento.

Tabla 2. Clasificación de reconocimiento de gestos para 4 personas, entre 4 clases y utilizando 8 bins en el descriptor final. Obteniendo una media de 80 %.

Subjects	**Accuracy(%)**
P_1	92
P_2	72
P_3	68
P_4	88

Tabla 3. Clasificación de reconocimiento de gestos para 4 personas, entre 5 clases y utilizando 8 bins en el descriptor final. Obteniendo una media de 77 %.

Subjects	**Accuracy(%)**
P_1	80
P_2	76
P_3	64
P_4	88

Interesantemente, el enfoque propuesto alcanza mejores resultados de reconocimiento con un gesto adicional (rotación), con un promedio de 80% en un total de 100 videos. Estos resultados muestran un desempeño estable de la aproximación propuesta en la caracterización de diferentes gestos capturados en secuencias RGBD.

En la Tabla 3, se reporta un tercer experimento en el cual se evaluó el dataset completo incluyendo todos los gestos. En este caso, la precisión promedio del método propuesto es de 77%. Lo cual resulta favorable en términos de la capacidad de representación de gestos. Como se puede observar en la Tabla 3, la persona P_3 es la que obtiene el peor resultado. Esto es debido a que realizó los gestos a diferentes velocidades w.r.t que el resto de personas, además de algunos ruidos en la escena que limitan la representación de las trayectorias.

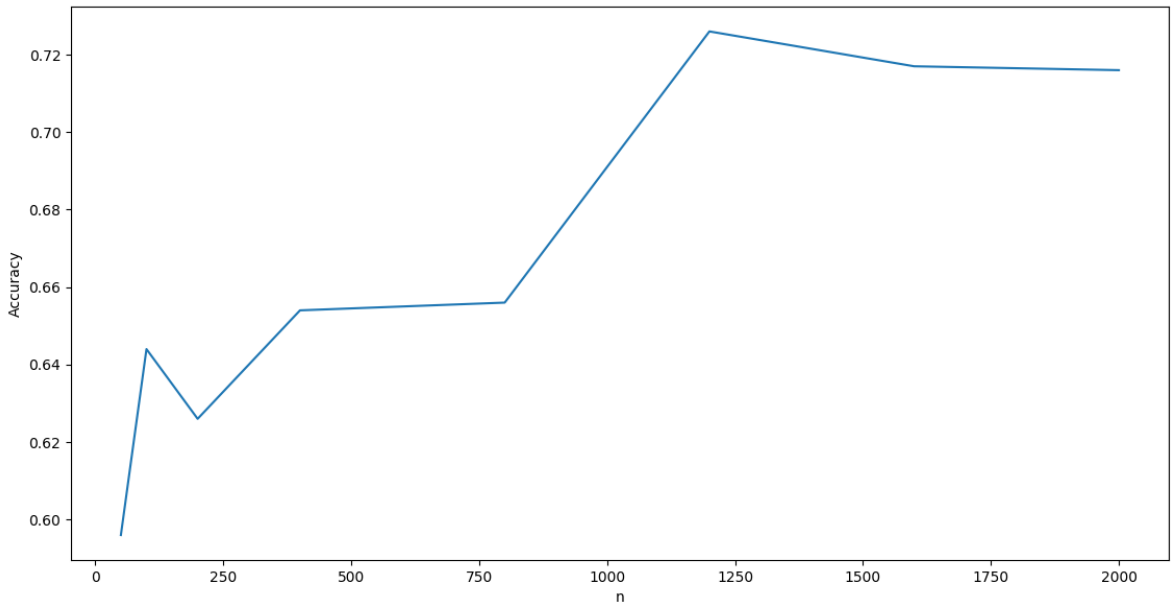
También, en la Tabla 4 se reporta un experimento adicional. En el que se utilizaron todos los gestos usando todas las posibles combinaciones de tres personas $\binom{4}{3} = 4$. En este experimento, se confirmó que la persona P_3 reporta variaciones w.r.t respecto a las demás. Sin embargo, el método propuesto resulta robusto a la captura de estas variaciones y trata de predecir correctamente el gesto realizado.

En la Figura 7, se observa un experimento realizando variaciones en el número de centroides utilizados para construir el diccionario de la bolsa de palabras cinemáticas. Experimentalmente se puede observar que el diccionario con 1000 palabras obtiene el mejor desempeño en el enfoque propuesto, alcanzando una precisión del 0.72%. Sin

Tabla 4. Clasificación de reconocimiento de gestos para 3 personas, entre 5 clases y utilizando 8 bins en el descriptor final. Obteniendo una media de 75.33 %.

Subjects	**Accuracy(%)**
P1,P2,P3	73.33
P1,P2,P4	84
P1,P3,P4	76
P2,P3,P4	68

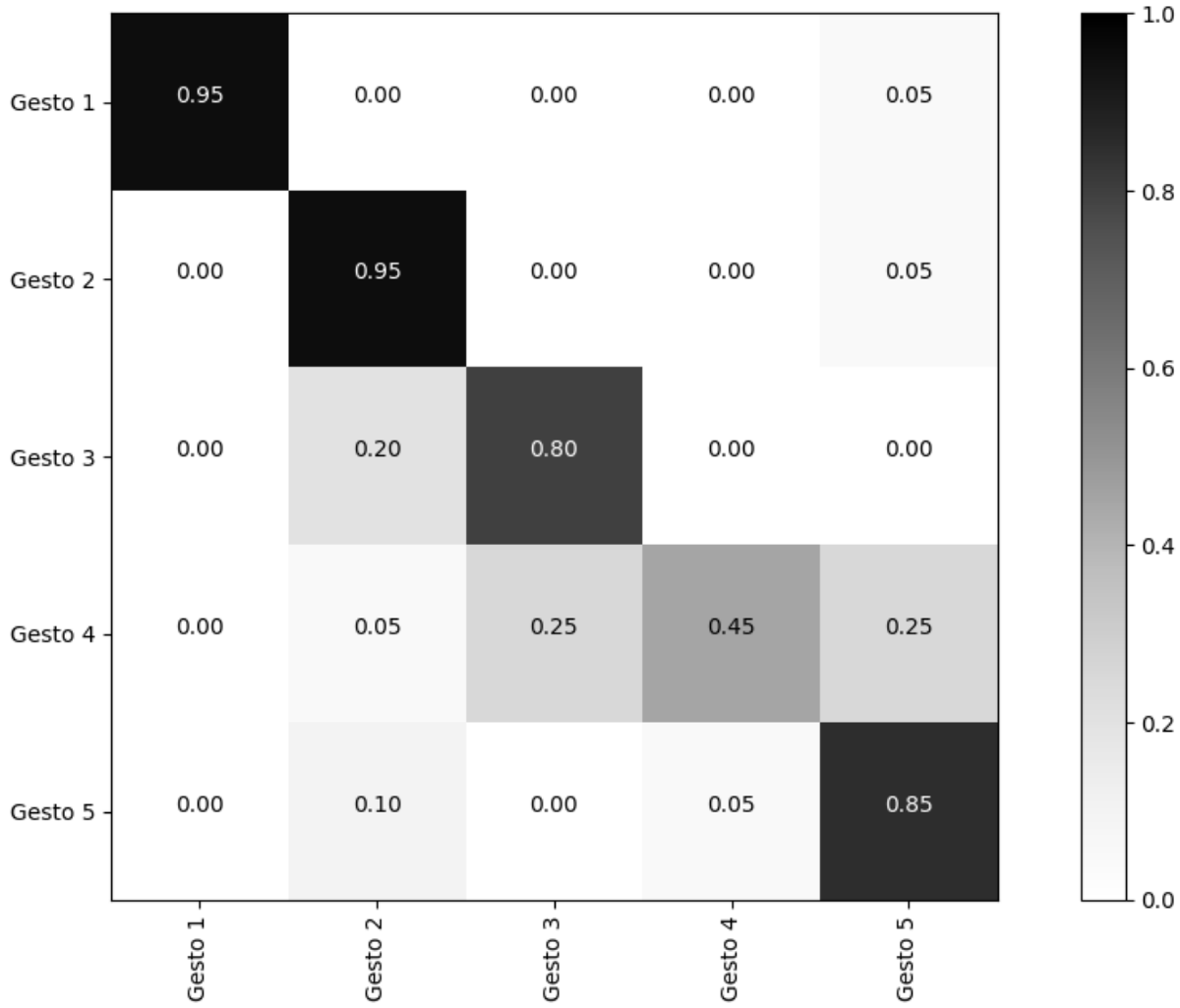
Figura 7. **Centroides.** Precisión del método propuesto variando el número de centroides utilizado por el algoritmo de clasificación.



embargo, utilizando solamente 250 palabras la estrategia propuesta alcanza una descripción competitiva, resultando eficiente para aplicaciones en tiempo real.

Finalmente, se calculó la matriz de confusión para el total de 5 gestos. Donde se puede observar que la mayoría de los gestos alcanzan un reconocimiento por encima del 80%. Particularmente, el gesto flexionar reporta algunas falencias en la clasificación con los gestos triángulo y frotar. En general, las trayectorias 3D+t de movimiento han demostrado ser robustas para la representación cinemática de gestos.

Figura 8. **Matriz de Confusión.** Matriz de Confusión para los mejores resultados obtenidos en los expermientos realizados con diferente número de gestos.



Capítulo 3
CONCLUSIONES

En este trabajo se presentó una nueva estrategia para caracterizar primitivas largas de movimiento en secuencias RGBD. Estas primitivas representan trayectorias 3D+t que se obtienen a partir del seguimiento de un conjunto de puntos a lo largo de la secuencia, utilizando información del flujo de escena. Estas trayectorias de movimiento se caracterizarón utilizando cinemáticas diferenciales y se incluyeron en una representación de bolsa de palabras. Además, se propuso un nuevo dataset útil para evaluar la información del movimiento de objetos particulares de interés en secuencias RGBD. En tareas de clasificación, las trayectorias demostraron ser robustas en la representación de distintos gestos capturados en RGBD a través de un descriptor compacto. El dataset cuenta con un total de cinco gestos diferentes, desarrollados por cuatro personas diferentes para un total de 100 videos. El método propuesto alcanza una precisión promedio del 77%, utilizando histogramas de solo 8 bins. En trabajos futuros se considerará el cálculo de trayectorias de movimiento que sigan exclusivamente puntos de interés en el mapa de profundidad, así como la evaluación en datasets más amplios y escenarios más complejos.

REFERENCIAS

[1] Manuel Blum, Jost Tobias Springenberg, Jan Wülfing, and Martin Riedmiller. A learned feature descriptor for object recognition in rgb-d data. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1298–1303. IEEE, 2012.

[2] Yang Zhao, Zicheng Liu, Lu Yang, and Hong Cheng. Combing rgb and depth map features for human activity recognition. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–4. IEEE, 2012.

[3] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Depth kernel descriptors for object recognition. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 821–826. IEEE, 2011.

[4] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1297–1304. Ieee, 2011.

[5] Sundar Vedula, Simon Baker, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 722–729. IEEE, 1999.

[6] Frédéric Huguet and Frédéric Devernay. A variational method for scene flow estimation from stereo sequences. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–7. IEEE, 2007.

[7] Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d flow: Dense 3-d motion estimation using color and depth. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 2276–2282. IEEE, 2013.

[8] Kourosh Khoshelham. Accuracy analysis of kinect depth data. In *ISPRS workshop laser scanning*, volume 38, page W12, 2011.

[9] Julian Quiroga, Frédéric Devernay, and James Crowley. Local/global scene flow estimation. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 3850–3854. IEEE, 2013.

[10] Felix Endres, Jürgen Hess, Jürgen Sturm, Daniel Cremers, and Wolfram Burgard. 3-d mapping with an rgb-d camera. *IEEE Transactions on Robotics*, 30(1):177–187, 2014.

[11] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[12] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[13] Julian Quiroga, Thomas Brox, Frédéric Devernay, and James Crowley. Dense semi-rigid scene flow estimation from rgbd images. In *European Conference on Computer Vision*, pages 567–582. Springer, 2014.

[14] Shih-Wei Sun, Yu-Chiang Frank Wang, Fay Huang, and Hong-Yuan Mark Liao. Moving foreground object detection via robust sift trajectories. *Journal of Visual Communication and Image Representation*, 24(3):232–243, 2013.

[15] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[16] Jianbo Shi et al. Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 593–600. IEEE, 1994.

[17] Shandong Wu and You Fu Li. Flexible signature descriptions for adaptive motion trajectory representation, perception and recognition. *Pattern Recognition*, 42(1):194–214, 2009.

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[19] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[20] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. A category-level 3d object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*, pages 141–165. Springer, 2013.

[21] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Rgb-d object recognition: Features, algorithms, and a large scale benchmark. In *Consumer Depth Cameras for Computer Vision*, pages 167–192. Springer, 2013.

[22] Bingbing Ni, Gang Wang, and Pierre Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer, 2013.

BIBLIOGRAFIA

BLUM, Manuel, *et al.* A learned feature descriptor for object recognition in rgb-d data. En Robotics and Automation (ICRA), 2012 IEEE International Conference on. IEEE, 2012. p. 1298-1303.

BO, Liefeng; REN, Xiaofeng; FOX, Dieter. Depth kernel descriptors for object recognition. En Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ

ENDRES, Felix, et al. *et al.* 3-D mapping with an RGB-D camera. IEEE Transactions on Robotics, 2014, vol. 30, no 1, p. 177-187.

HENRY, Peter, *et al.* RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. The International Journal of Robotics Research, 2012, vol. 31, no 5, p. 647-663.

HORN, Berthold KP; SCHUNCK, Brian G. Determining optical flow. Artificial intelligence, 1981, vol. 17, no 1-3, p. 185-203.

HUGUET, Frédéric; DEVERNAY, Frédéric. A variational method for scene flow estimation from stereo sequences. En Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. IEEE, 2007. p. 1-7.

JANOCH, Allison, *et al.* A category-level 3d object dataset: Putting the kinect to work. En Consumer Depth Cameras for Computer Vision. Springer, London, 2013. p. 141-165.

KHOSHELHAM, Kourosh. Accuracy analysis of kinect depth data. En ISPRS workshop laser scanning. 2011. p. W12.

LAI, Kevin, *et al.* RGB-D object recognition: Features, algorithms, and a large scale benchmark. En Consumer Depth Cameras for Computer Vision. Springer, London, 2013. p. 167-192.

LUCAS, Bruce D, *et al.* An iterative image registration technique with an application to stereo vision. 1981.

MIKOLOV, Tomas, *et al.* Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

NI, Bingbing; WANG, Gang; MOULIN, Pierre. Rgbd-hudaact: A color-depth video database for human daily activity recognition. En Consumer Depth Cameras for Computer Vision. Springer, London, 2013. p. 193-208.

QUIROGA, Julian; DEVERNAY, Frédéric; CROWLEY, James. Local/global scene flow estimation. En Image Processing (ICIP), 2013 20th IEEE International Conference on. IEEE, 2013. p. 3850-3854.

QUIROGA, Julian, *et al.* Dense semi-rigid scene flow estimation from rgbd images. En European Conference on Computer Vision. Springer, Cham, 2014. p. 567-582.

SHI, Jianbo, *et al.* Good features to track. En Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on. IEEE, 1994. p. 593-600.

SHOTTON, Jamie, *et al.* Real-time human pose recognition in parts from single depth images. En Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. Ieee, 2011. p. 1297-1304.

SUN, Shih-Wei, *et al.* Moving foreground object detection via robust SIFT trajectories. Journal of Visual Communication and Image Representation, 2013, vol. 24, no 3, p. 232-243.

SUYKENS, Johan AK; VANDEWALLE, Joos. Least squares support vector machine classifiers. Neural processing letters, 1999, vol. 9, no 3, p. 293-300.

VEDULA, Sundar, *et al.* Three-dimensional scene flow. En Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. IEEE, 1999. p. 722-729.

WANG, Heng, *et al.* Action recognition by dense trajectories. En Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011. p. 3169-3176.

WU, Shandong; LI, You Fu. Flexible signature descriptions for adaptive motion trajectory representation, perception and recognition. Pattern Recognition, 2009, vol. 42, no 1, p. 194-214.

ZHAO, Yang, *et al.* Combing rgb and depth map features for human activity recognition. En Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific. IEEE, 2012. p. 1-4.