RECONOCIMIENTO DE ACTIVIDADES EN VIDEO UTILIZANDO UN DESCRIPTOR REGIONAL DE COVARIANZA

Wilson Daniel Moreno Prada

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

RECONOCIMIENTO DE ACTIVIDADES EN VIDEO UTILIZANDO UN DESCRIPTOR REGIONAL DE COVARIANZA

Wilson Daniel Moreno Prada

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

Director:
FABIO MARTÍNEZ CARRILLO
Ph.D en Ingeniería de Sistemas y Computación

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 su amistad y consejos, sin el no hubiera sido posible la realización de este trabajo.

A todos mis amigos de infancia y universidad por su apoyo incondicional y motivación 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 me han permitido entrenarme para ser un buen profesional.

Finalmente quiero realizar un agradecimiento especial a mis padres y hermana por todo por sus enseñanzas y confianza que han tenido en mi por ser siempre la fuente de mi apoyo y motivación.

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

Título: Reconocimiento de actividades en video utilizando un descriptor regional de covarianza [1]

Autor: Wilson Daniel Moreno Prada[2]

Palabras Clave: Covarianza espacio-temporal, reconocimiento de actividades humanas, análisis de movimiento, primitivas de bajo nivel

DESCRIPCIÓN:

El reconocimiento de actividades es una de las áreas predominantes en visión por computador cuyo principal objetivo es la caracterización y cuantificación de patrones de movimiento y de apariencia involucrados en las actividades desarrolladas en video. Estos principios han sido utilizados en una gran variedad de aplicaciones, tales como: la video-vigilancia, el análisis deportivo, los sistemas de interacción persona ordenador, entre muchos otros. A pesar del amplio espectro de propuestas descritas en el estado del arte, existen aún problemas abiertos en cuanto a la descripción de actividades en contextos específicos, la caracterización de patrones frente a cambios de iluminación, la cuantificación de la variabilidad de los objetos de interés, las variaciones de movimiento, entre otros. Por otra parte, los enfoques clásicos son computacionalmente costosos y la precisión en su clasificación depende de la dimensionalidad de los descriptores.
Este trabajo presenta un descriptor de covarianza compacto que permite analizar características espacio-temporales que modelan y caracterizan las actividades. Inicialmente se capturan un conjunto de primitivas de bajo nivel que describen la secuencia de video. El descriptor propuesto es calculado en cada cuadro del video de forma eficiente utilizando una representación de imagen integral. Una vez calculado el descriptor en cada cuadro de la secuencia se obtiene una estimación media de la covarianza utilizando la geometría de Riemann que representa la actividad registrada en el video. Finalmente el descriptor propuesto es mapeado hacia un algoritmo de clasificación para realizar una clasificación automática de las actividades. El enfoque propuesto fue evaluado sobre un conjunto de datos públicos (UT-interacción) con un esquema de validación cruzada (k-fold) obteniendo una precisión promedio de $70,83\%$ para todo el conjunto de datos, con un tamaño del descriptor de 275 valores por secuencia de video.

---

[1] Trabajo de Grado

[2] Facultad de Ingenierías Físicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: FABIO MARTÍNEZ CARRILLO.

ABSTRACT

Title: Frame-level covariance descriptor for action recognition[1]

Author: Wilson Daniel Moreno Prada[2]

Keywords: Spatio-temporal covariance, human activity recognition, motion analysis, low-level primitives.

DESCRIPTION:

Activity recognition is a fundamental task in areas such as video-surveillance, gesture recognition, robotics, multimedia applications among much others. Such task reamins as an open problem because the high variability of dynamic actions, and appareance of actors. Also the considerable illumination changes in real scenarios of capture diifucult the proper recognition of activities. Despite favorable results in recent works for several academic datasets, the proposed methodologies require a huge number of training samples and the output descriptor result in a hig dimensional array that difficult its direct application in real time scenarios. This work proposes a spatio-temporal descriptor that models and characterizes human activities by using a fast regional covariance representation for each frame. At each frame, a set of motion and geometrical map measures are quantified into a pyramidal regional structure to describe the instantaneous action. Such low-level primitive maps are codified into a integral covariance map that allows a fast and compact description of local correlation among features. The set of pyramidal-frame-covariances along the video sequence represent a manifold that coexist in a positive Riemmanian space. Then, a set of Riemmanian means are computed for each regional covariance sequence to represent a very compact action descriptor. The proposed action descriptor uis mapped to a Euclidean space to perform an automatic classification using a Support vector Machine. The porposed approach was tested over a public dataset (UT-interaction) with a k-fold cross-validation scheme obtaining an accuracy of 70.83% with a descriptor size of just 275 features per video sequence.

---

[1] Research Work.

[2] School of Physical-Mechanical Engineering. Department of Systems Engineering and Informatics. Advisor, FABIO MARTÍNEZ CARRILLO.

Introducción

El reconocimiento de acciones es una área fundamental en visión por computador con amplias aplicaciones en video-vigilancia, análisis deportivo, vehículos inteligentes, sistemas HCI (*Human computer interaction*), entre otros [1]. Esta tarea abarca, sin embargo, muchos desafíos en cuanto al modelamiento y complejidad computacional que incluyen el modelado complejo de la variabilidad de la iluminación, la representación de objetos y los cambios de movimiento que dificultan el etiquetado automático del vídeo. Asimismo, los métodos tradicionales son la mayoría de veces computacionalmente costosos debido a la cuantificación exhaustiva de patrones de movimiento y de apariencia, lo cual restringe el desarrollo de aplicaciones en linea y su uso en tecnologías con requerimientos limitados de computación. Además, en estas estrategias se evidencia altos puntajes en la precisión, los cuales dependen de la dimensionalidad de los descriptores.

En el estado del arte se han propuesto múltiples estrategias para reconocer actividades que pueden clasificarse como métodos de reconocimiento global y métodos de reconocimiento local. Los métodos de representación global se han centrado en la caracterización y cuantificación de extensas regiones de interés o incluso en secuencias de vídeo completas. En cuanto a estos métodos, en la literatura se han propuesto métodos basados en la sustracción y seguimiento de siluetas que representan el cuerpo humano durante el video que han permitido abordar problemas de detección de peatones [2–4]. Tambien Bobick y Davis [5] propusieron el cálculo de descriptores de ocurrencia de movimiento a partir de la extracción de siluetas en secuencias de imágenes. Ademas, Wang *et. al* [6] propusieron un nuevo descriptor para el reconocimiento de actividades basado en la extracción de siluetas humanas binarias utilizando la transformada R para representar las características de bajo nivel. Estos descriptores son robustos a oclusiones, a las siluetas disjuntas y agujeros sobre la forma. Souvenir y Babbs [7] ampliaron aún más este trabajo al considerar los contornos de la imagen, mejorando la caracterización de las actividades pero incrementando el costo computacional. Estrategias adicionales han propuesto la caracterización de siluetas a partir de múltiples cámaras, pero que requieren una calibración exhaustiva para los dispositivos de adquisición [8–10]. En términos generales, los métodos globales cuan-

tifican el movimiento postural basado en siluetas a nivel del frame pero son sensibles al ruido, la oclusión parcial y a la variabilidad según el punto de referencia de la cámara. También, estos enfoques son totalmente dependientes de la captura apropiada de las siluetas, lo cual puede involucrar escenarios controlados para su apropiada captura [1].

Por otra parte, en la literatura se han propuesto varios métodos basados en la detección del punto de interés y parches locales que evitan relativamente los cambios a la apariencia, la perspectiva y son robustos a las oclusiones parciales, entre otros problemas de los enfoques globales [1]. Por ejemplo, Laptev y Lindeberg [11] capturan múltiples puntos de interés a diferentes escalas en el dominio espacio-temporal que permite la detección de estructuras locales para la representación de eventos en secuencias de vídeo. Dollár *et al.* [12] propuso un método para codificar cuboides 3D espacio-temporales. Luego se utilizan un conjunto de histogramas de cuboides para representar las acciones en vídeo y se calcula una distancia euclidiana entre los histogramas para clasificar las acciones. Laptev [13] utiliza la caracterización de geometría local a múltiples escalas para calcular cuboides salientes en secuencias de vídeo. Los puntos salientes se asignan a una máquina vectorial de soporte (*SVM*) para clasificar automáticamente las acciones. Gowayyed *et al.* [14] propuso un método para el reconocimiento de acciones utilizando la posición de las articulaciones con respecto a un esqueleto humano. Este método describe trayectorias de articulaciones humanas en 3D basadas en histogramas de desplazamientos orientados (*HOD*). También, Robertson y Reid [15] propusieron combinar las características de trayectoria (es decir, posición y velocidad) como un conjunto de descriptores de movimiento local para el reconocimiento de la acción humana. Liu *et al.* [16] propuso un método basado en el aprendizaje regularizado de tareas múltiples que codifica implícitamente las características visuales locales y la estructura del cuerpo humano como pequeños bloques de información, que se representan como una bolsa piramidal de palabras (*PPBoW*). Este enfoque logra resultados competitivos por la robustez a la apariencia y a la geometría pero sigue dependiendo de la apariencia. En términos generales, los métodos basados en la caracterización local evidencian un alto costo computacional que limita el desarrollo de aplicaciones en línea. Además, en estos enfoques la precisión requiere descriptores de alta dimensiónalidad para lograr una predicción de la acción adecuada.
Otros enfoques han centrado la atención en el desarrollo de descriptores compactos para representar objetos o caracterizar eventos particulares en secuencias de video, los cuales podrían ser extendidos a aplicaciones de reconocimiento y detección de acciones en video. Por ejemplo, Once *et. al.* propuso un descriptor de covarianza local para detectar y clasificar objetos a partir de la información de textura. En dicho enfoque se introdujeron las covarianzas integrales para cálculos regionales rápidos, y también fue evaluada en problemas de seguimiento de objetos de interes, utilizando como primitivas la información de primer y segundo orden de los cuadros del video [17]. Además, Bingpeng et al. [18] implementó un descriptor basado en la covarianza para

el reconocimiento facial mediante el uso de cámaras individuales y múltiples. En este enfoque se combinaron características bio-inspiradas y se codificaron en un descriptor de covarianza para la representación de la persona, reportando la robustez de los cambios de fondo y de iluminación.

La principal contribución de este trabajo es un descriptor de covarianza espacio-temporal que modela y caracteriza las actividades humanas que ocurren en secuencias de video. Para hacerlo, el enfoque propuesto computa una representación de flujo óptico denso a lo largo de la secuencia, permitiendo cuantificar grandes desplazamientos. Luego, se calcula un conjunto de primitivas cinemáticas en cada cuadro para representar el movimiento principal de cada video. Las primitivas cinemáticas en cada frame se codifican en matrices de covarianza regionales que se calculan a partir de una representación gruesa a fina dividiendo iterativamente cada frame. El conjunto de covarianzas calculadas en cada cuadro coexisten en el espacio de Riemann que representan cada acción particular codificada en el video. Luego, el descriptor final se calcula como la media de Riemann de las matrices de covarianza que representan la acción a lo largo del video. El resto del documento está organizado de la siguiente manera: el enfoque propuesto se describe en la sección 3, mientras que la evaluación y el resultado del enfoque propuesto para el estado del arte se informa en la sección 4. Finalmente en la sección 5 se discuten las ventajas y se presentan varias conclusiones del trabajo.

Capítulo 1
PLANTEAMIENTO Y JUSTIFICACIÓN DEL PROBLEMA

La caracterización de acciones humanas involucra reconocer gestos, actividades cotidianas e interacciones entre humanos, entre objetos o entre grupos de agentes presentes en el video. El reconocimiento y clasificación de acciones esta presente en diversas áreas del conocimiento con perspectivas de procesamiento en linea o aprendizaje robusto para detecciones precisas de la acción desarrollada.

Sin embargo, los métodos tradicionales son computacionalmente costosos debido a la cuantificación exhaustiva de patrones de movimiento y de apariencia con dependencia de la dimensionalidad del descriptor para obtener resultados fiables. Además, el reconocimiento de tareas presenta variaciones complejas que dependen de cambios de iluminación, representación de los objetos en el video, variaciones de movimiento, entre otros, que dificultan la tarea de etiquetar automáticamente los vídeos. Por ejemplo, en áreas como la video vigilancia, el reconocimiento se debe realizar de manera instantánea lo cual está limitado por la naturaleza de los descriptores propuestos actualmente en la literatura.

Capítulo 2
OBJETIVOS

2.1 OBJETIVO GENERAL

Proponer e implementar un descriptor espacio-temporal basado en la covarianza de primitivas de bajo nivel para reconocer actividades en video.

2.2 OBJETIVOS ESPECÍFICOS

- Representar las actividades registradas en video utilizando primitivas de bajo nivel en cuanto a movimientos locales y apariencia.
- Caracterizar las primitivas de bajo nivel utilizando un descriptor espacio-temporal de la covarianza.
- Seleccionar e implementar un clasificador que utilice el descriptor propuesto para el reconocimiento de actividades humanas.
- Validar el descriptor propuesto en una base de datos académica en cuanto a la clasificación de actividades obtenidos por el descriptor.

Capítulo 3
MÉTODO PROPUESTO

En este trabajo se introduce un descriptor de covarianza compacto que codifica características espacio-temporales calculadas principalmente a partir de un flujo óptico denso para representar actividades en vídeo. El enfoque propuesto comienza calculando un campo de velocidad densa que permite grandes desplazamientos a lo largo del vídeo. Luego se calcula un conjunto de primitivas cinemáticas a lo largo de la secuencia. A partir de aquí se implementa una covarianza integral eficiente para representar múltiples regiones de un cuadro y codificar las diferentes primitivas cinemáticas. Para lograr una mayor descripción de la acción a nivel de cuadro, en este trabajo se realizó una caracterización de múltiples regiones codificadas en diferentes capas. Inicialmente se toma el cuadro completo como la primera región, luego iterativamente se hace una partición del cuadro para obtener mas regiones que son caracterizadas con matrices de covarianza. Las matrices de covarianza son simétricas, positivas y están definidas en un espacio curvo, definido como el espacio de Riemman. Este hecho implica que el calculo de geodésicas no corresponde directamente a los métodos euclidianos. Por lo tanto, una vez caracterizados los cuadros del video, cada región particular de covarianza se describe utilizando la media de Riemman. El conjunto de medias son concatenadas y forman el descriptor de video. Este descriptor es de baja dimensionalidad y codifica las correlaciones de las diferentes primitivas en las regiones definidas. Finalmente, el descriptor propuesto es mapeado a un espacio euclidiano para ser probado en una maquina de soporte vectorial para obtener una clasificación automática de actividad registrada en el video. La metodología del enfoque propuesto se representa en la Figura 1.

3.1 FLUJO ÓPTICO DE LARGO DESPLAZAMIENTO

Uno de los objetivos principales de este trabajo es la caracterización de las acciones usando únicamente información dinámica registrada en la secuencia. El enfoque propuesto cuantifica mapas de primitivas de movimiento aparente como una representación de bajo nivel en cada

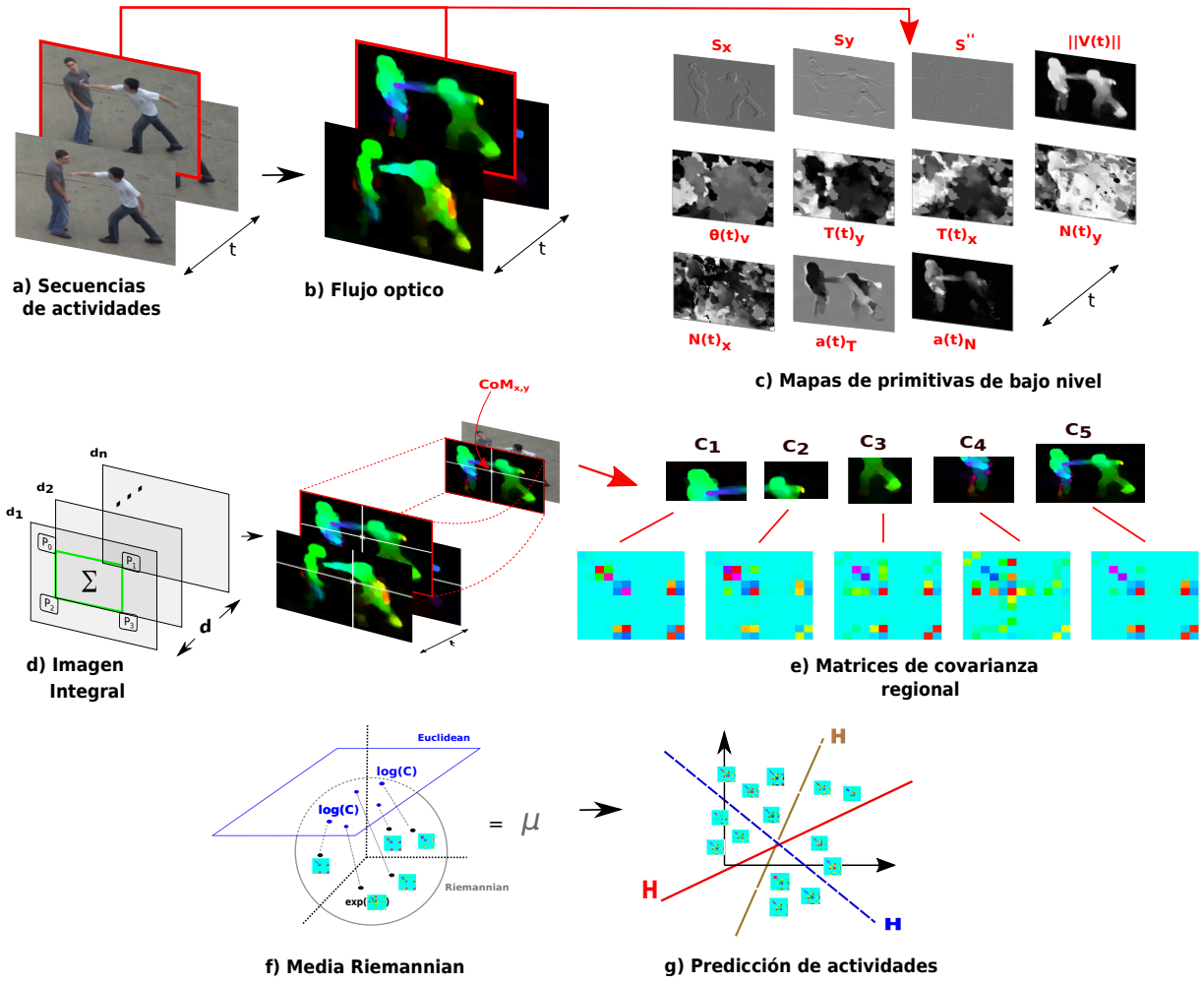

Figure 1. Método propuesto: (a) Selección de vídeos que registran actividades humanas. (b) Cálculo del flujo óptico de largo desplazamiento para toda la secuencia. (c) Cálculo de las primitivas d de movimiento y apariencia de bajo nivel para todo el video. (d) Calculo de las imágenes integrales para cada característica d, luego calculamos la posición del del centro de masa sobre cada cuadro del flujo óptico que nos permite obtener múltiples regiones en cada frame. (e) Calculo rápido de múltiples matrices de covarianza regional utilizando la representación de imagen integral.(f) Luego, se calcula un descriptor de covarianza para cada secuencia estimando el promedio de las matrices de covarianza para cada región sobre la secuencia. (g) Finalmente, el descriptor de covarianza propuesto es validado sobre un algoritmo de clasificación en términos del reconocimiento de actividades humanas.

vídeo. Este enfoque es flexible para admitir cualquier representación de flujo denso que codifique la velocidad aparente en cada frame. En este trabajo se implementó un flujo óptico denso que permite describir grandes desplazamientos a lo largo de la secuencia, considerando varias restricciones [19], 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:

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

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 de-

splazamientos arbitrariamente grandes. Este método es robusto y ha sido ampliamente utilizado en la literatura para diferentes aplicaciones.

3.2 MAPAS DE PRIMITIVAS CINEMÁTICAS

El descriptor propuesto primero cuantifica mapas de cinemáticas calculadas a partir del movimiento aparente obtenido por el método de flujo óptico para describir las acciones. También, esta descripción dinámica puede ser complementada por mapas de primitivas geométricas como representaciones de bajo nivel en cada frame.

Las primitivas cinemáticas fueron calculadas a nivel del píxel, lo cual permite una descripción densa a nivel del cuadro que facilitará el calculo de estadísticas y correlaciones regionales. Desde el campo del movimiento denso se obtiene la velocidad $||V(t)||$ y el ángulo $\theta_V(t)$. Estas cinemáticas representan primitivas de primer orden, a partir de las cuales se computan otras representaciones espacio-temporales y de orden superior. Por ejemplo, la derivada de la magnitud de velocidad representa la velocidad de movimiento que corresponde a un valor escalar que relaciona la distancia y el tiempo de seguimiento $S_{\|V(t)\|}$ y la derivada del ángulo de la velocidad representa la dirección de la variación en la velocidad $S_{\theta_v(t)}$ para cada frame.

También se calculó la velocidad unitaria tangencial que se expresa como: $T(t) = \frac{V(t)}{\|V(t)\|}$ y la velocidad unitaria normal que esta definida como: $N(t) = \frac{T^{'}(t)}{\left\|T^{'}(t)\right\|}$. Para cada píxel de un cuadro de flujo óptico existen dos vectores que siempre son ortogonales $N(t)$ y $T(t)$ para cada t que se expanden sobre un plano $\rho(t)$ osculante. Cuando la derivada de la velocidad unitaria tangencial es $T\prime(t) = 0$, la velocidad unitaria normal y el plano osculante no están definidos por lo tanto en nuestra representación no son tenidos en cuenta.

Del mismo modo, cinemáticas de mas alto orden fueron tenidas en cuenta para lograr una mejor caracterización de las acciones en términos dinámicos. En este sentido se calculo la aceleración en términos de velocidades unitarias tangenciales y normales. Esta primitiva existe en el plano osculante siempre y cuando $T(t)$ y $N(t)$ también existan. Expresamos la aceleración de movimiento como:

$$a(t) = a_T(t)T(t) + a_N N(t) \tag{3.2}$$

Donde los coeficientes de aceleración tangencial $a_T(t)$ y normal $a_N(t)$ los expresamos como:

$$a_T(t) = \frac{d}{dt}\left\|V(t)\right\| \tag{3.3}$$

$$a_N(t) = \left\|V(t)\right\| \left\|T^{'}(t)\right\| \tag{3.4}$$

La aceleración tangencial representa la derivada de la rapidez del movimiento mientras que la

aceleración normal representa la derivada de la dirección de la velocidad respecto al tiempo. Ambas cantidades nos permiten obtener la magnitud de la aceleración, como:

$$\|a\|^2 = (a_T)^2 + (a_N)^2 \tag{3.5}$$

Teniendo en cuenta que las primitivas de aceleración son de segundo orden y por lo general implican varios cuadros para su cálculo, en este trabajo se realizó un seguimiento de los vectores de velocidad en donde para un vector particular de velocidad en el tiempo $t+1$ era asociado al vector que había apuntado a esa posición en el tiempo t. Para enriquecer el mapa de primitivas de movimiento, calculamos también la primera derivada del movimiento $||V(t)||$ pero sobre los ejes (x, y) como: $\frac{\partial||V(t)||}{\partial x \partial y}$. Además, los mapas de movimiento cinemático se pueden complementar utilizando mapas de apariencia en cada cuadro. En este trabajo para algunos experimentos se agregaron derivadas de primer y segundo orden sobre cada cuadro. Dichos mapas representan bordes, es decir, la geometría local capturada en cada cuadro.

3.3 CODIFICACIÓN DE COVARIANZA INTEGRAL

La matriz de covarianza constituye un método natural y compacto para combinar múltiples características correlacionadas, que se puede expresar como:

$$C_R(i,j) = \frac{1}{n-1}\left[\underbrace{\sum_{k=1}^{n} z_k(i)z_k(j)}_{Q} - \frac{1}{n}\underbrace{\sum_{k=1}^{n} z_k(i)}_{P}\underbrace{\sum_{k=1}^{n} z_k(j)}_{P}\right] \tag{3.6}$$

,donde $z_{k=1\ldots n}$ es un vector con n muestras para $(i,j) = 1\ldots d$ características y $\frac{1}{n}\sum_{k=1}^{n} z_k(i)$ representa el valor esperado μ. Las matrices de covarianza son simétricas y positivas de dimensionalidad $d \times d$,representadas por solo $\frac{d^2+d}{2}$ valores diferentes, la diagonal principal de la matriz representa la varianza para cada característica. Esta matriz ha sido ampliamente utilizada en diferentes aplicaciones de identificación de objetos, seguimiento y clasificación [18, 20].
Sin embargo, el calculo de la matriz de covarianza requiere un alto costo computacional por las diferentes interacciones entre cada par de características del conjunto d. Por lo tanto para hacer frente a dicha limitación, se implementó una alternativa regional rápida propuesta en [17] para calcular las regiones de covarianza mediante el uso de una representación de imagen integral (como se ilustra en la Figura 2). Estas imágenes integrales son representaciones intermedias y se utilizan generalmente para el cálculo rápido de sumas en una región determinada.

En este trabajo, cada frame F es caracterizado con d características cinemáticas y luego se

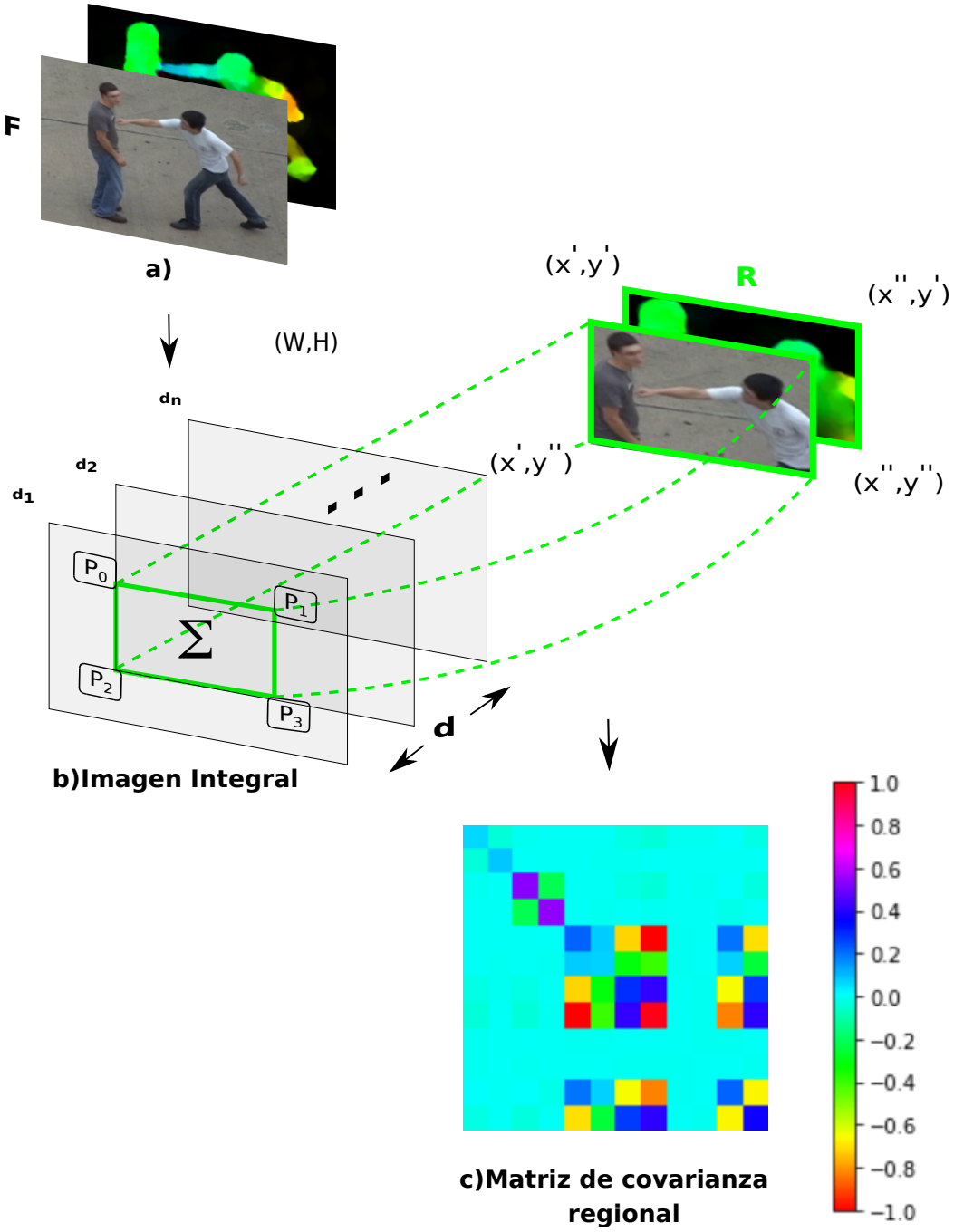

Figure 2. Esta imagen representa el cálculo regional de la covarianza mediante el uso de la representación de imagen integral y un conjunto de mapas de primitivas.

codifica regionalmente mediante el uso de una representación de covarianza integral. En tal caso, la suma de cada dimensión característica $z(i)_{k=1...n}$, se representa como un tensor de primer orden $P \in \mathbb{R}^{W \times h \times d}$, calculado como:

$$P(x^{'}, y^{'}, i) = \sum_{x<x^{'}, y<y^{'}} F(x, y, i) \tag{3.7}$$

,donde F es un frame $F \in \mathbb{R}^{(W \times H \times d)}$ con $i = 1 \dots d$. Entonces, el tensor P es un vector d-tamaño que contiene la suma de cada dimensión característica, $P_{x,y} = [P(x, y, 1)...P(x, y, d)]^T$. Además, la suma del producto de características $z_k(i)z_k(j)_{i,j=1...n}$ (primera parte de la ecuación 3.6) se puede expresar con imágenes integrales como un tensor de segundo orden $Q \in \mathbb{R}^{W \times H \times d \times d}$

$$Q(x^{'}, y^{'}, i, j) = \sum_{x<x^{'}, y<y^{'}} F(x, y, i)F(x, y, j) \tag{3.8}$$

con $\{i, j\} = 1...d$. Tensor Q es una matriz simétrica $d * d$ que contiene la suma de los productos de cualquier par de características, expresada como:

$$Q_{x,y} = \begin{pmatrix} Q(x,y,1,1) & ... & Q(x,y,1,d) \\ & . & \\ & . & \\ & . & \\ Q(x,y,d,1) & ... & Q(x,y,d,d) \end{pmatrix} \tag{3.9}$$

El cálculo de este tensor integral requiere $\frac{d^2+d}{2}$ interacciones. Entonces, una vez construido las imágenes integrales el calculo de la matriz de covarianza para la región rectangular R (ver Figura 2), delimitada por las esquinas superior izquierda e inferior derecha se puede calcular con un costo computacional de $O(d^2)$. La ecuación 3.6 se puede volver a escribir en términos de tensores integrales como:

$$C_{R(x^{'},y^{'};x^{''},y^{''})} = \frac{1}{n-1}[(Q_{x^{''},y^{''}} + Q_{x^{'},y^{'}} - Q_{x^{''},y^{'}} - Q_{x^{'},y^{''}})$$

$$-\frac{1}{n}(P_{x^{''},y^{''}} + P_{x^{'},y^{'}} - P_{x^{''},y^{'}} - P_{x^{'},y^{''}})$$

$$(P_{x^{''},y^{''}} + P_{x^{'},y^{'}} - P_{x^{''},y^{'}} - P_{x^{'},y^{''}})^T]$$

Donde $n = (x^{''} - x^{'})(y^{''} - y^{'})$. Dicha expresión implica cálculos más rápidos para cualquier covarianza regional en el frame completo con pocas operaciones aritméticas.

Para enriquecer la descripción de cada cuadro, en este trabajo se calculó regionalmente la matriz de covarianza en diferentes regiones. Se obtuvieron un total de cinco matrices de covarianza para representar cada frame F en la secuencia de video (ver la Figura 3). La primera covarianza corresponde al frame completo. Las cuatro restantes corresponden a las subregiones del frame divididas con respecto a la posición del centro de masa CoM dada por el campo de movimiento del frame. Este CoM se calcula como $CoM_{x,y} = \frac{1}{M}\sum_{y=1}^{n}\sum_{x=1}^{n}\|V_{y,x}(t)\|\, r_{y,x}(t)$, donde $\|V(t)\|$ representa la velocidad, $\{x, y\}$ es la posición del cuadro y $M = \sum_{y=1}^{n}\sum_{x=1}^{n}\|V_{y,x}(t)\|$. Este CoM se puede interpretar como la posición espacial con la mayor cantidad de movimiento para cualquier frame dado. Las cuatro subregiones se calculan dividiendo el marco por el CoM. En la figura 3 se representa este cálculo.

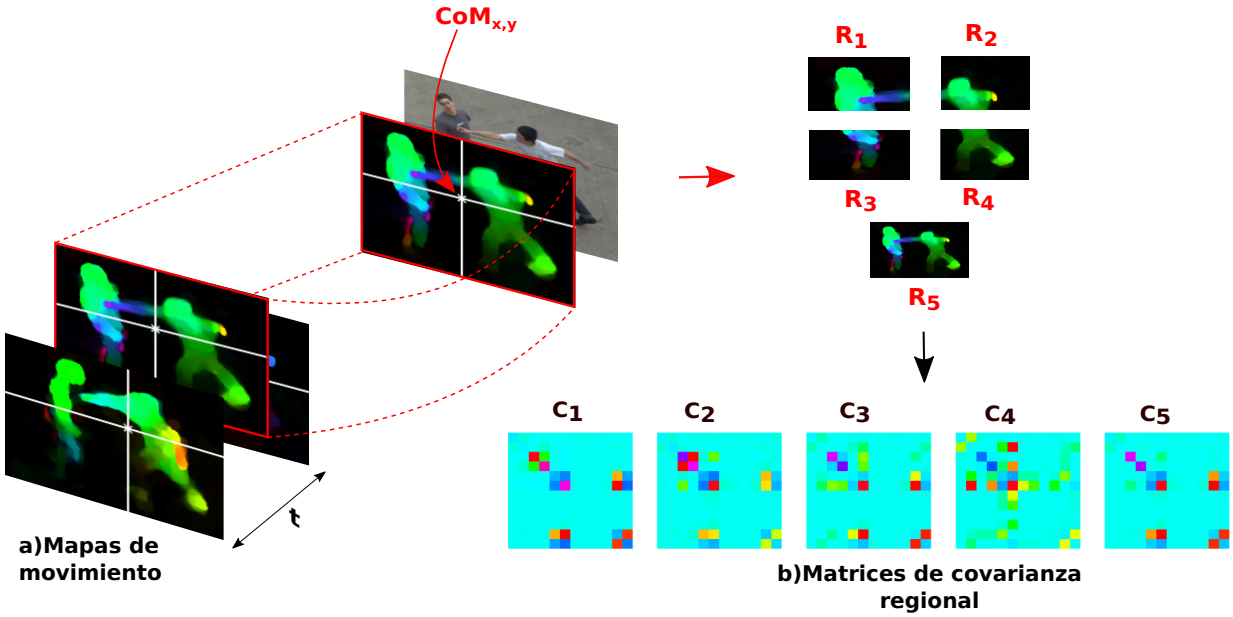

Figure 3. Descriptor de covarianza regional: (a)Calculo de mapas de movimiento y posición del centro de masa CoM. (b)Calculo rápido de múltiples matrices de covarianza regional dado el CoM para cada frame.

3.4 MEDIA DE RIEMANN EN SECUENCIAS DE VIDEO

Las matrices de covarianza calculadas en cada frame $c_1, c_2, c_3, \ldots, c_n$ construyen un descriptor que representa las actividades en video. Debido a las propiedades de las matrices de covarianza, estos descriptores se definen en un espacio de Riemann esférico y no en el espacio euclidiano clásico [21,22], lo que limita el uso de los algoritmos clásicos de aprendizaje maquina y visión por computador. La proyección de una matriz de covarianza c_i en un espacio euclidiano se calcula

como $\log(p) = \Sigma DIAG(\log(\lambda_i))\Sigma^T$, donde $Sigma$ son los vectores propios de la matriz y λ son los valores propios respectivos. De la misma manera, cualquier proyección desde el espacio euclidiano al espacio de Riemann se aproxima como $\exp(p) = \Sigma DIAG(\exp(\lambda_i))\Sigma^T$.

Por lo tanto, el descriptor de video que proponemos como una matriz de covarianza, representa la distancia mínima con respecto al conjunto $c_1, c_2, c_3, \ldots, c_n$, calculado para cada región a lo largo del tiempo. Esta covarianza representativa se calcula luego como la covarianza media intrínseca en el espacio de Riemann, como se muestra en el algoritmo 1 propuesto por [23].

Algorithm 1 Algoritmo del descenso de gradiente para calcular la media intrínseca a partir del conjunto de covarianzas regionales de cada frame

Salida: $\mu \in C(n)$

for Cada región de secuencias de covarianza j **do**
  $c_1^j, ..., c_N^j \in C(n)$
  $\mu = c_1^j$
  $\tau = 1 \rightarrow$ tamaño paso inicial
  Do
    $X_i = \frac{1}{N}\sum_{k=1}^{N} \log_{\mu_i}(c_k^j)$
    $\mu_{i+1} = \exp_{\mu_i}(\tau X_i)$

    if ($\|X_i\| > \|X_{i-1}\|$) **then**
      $\tau = \tau/2$
      $X_i = X_{i-1}$
    end if

  While ($\|X_i\| > \epsilon$)
end for

Para calcular la media intrínseca, las operaciones se $\log$ y $\exp$ definen respecto al valor calculado μ, que es expresado como: $\exp_\mu(X) = \mu^{\frac{1}{2}} \exp(\mu^{\frac{-1}{2}} X \mu^{\frac{-1}{2}})\mu^{\frac{1}{2}}$ y $\log_\mu(p) = \mu^{\frac{1}{2}} \log(\mu^{\frac{-1}{2}} p \mu^{\frac{-1}{2}})\mu^{\frac{1}{2}}$, respectivamente. Donde se cumple que: $\exp(\log(\mu)) = \mu$ y la expresión representa $\mu^{\frac{1}{2}} = \exp(\frac{1}{2}(\log \mu)$ la matriz inversa.

El criterio de parada iterativo para el cálculo de la estimación a la media es la expresión $\|X_i\|$, que podemos expresar como $\|X_i\| = \sum_{i=1}^{N}(\log(\sigma_i))^2$, donde σ son los valores propios respectivos. El umbral de error los definimos como: $(0.01 < \epsilon < 0.1)$.
En la figura 4 se representa el calculo para la estimación a la media de las matrices de covarianza mediante las operaciones $\log$ y $\exp$.

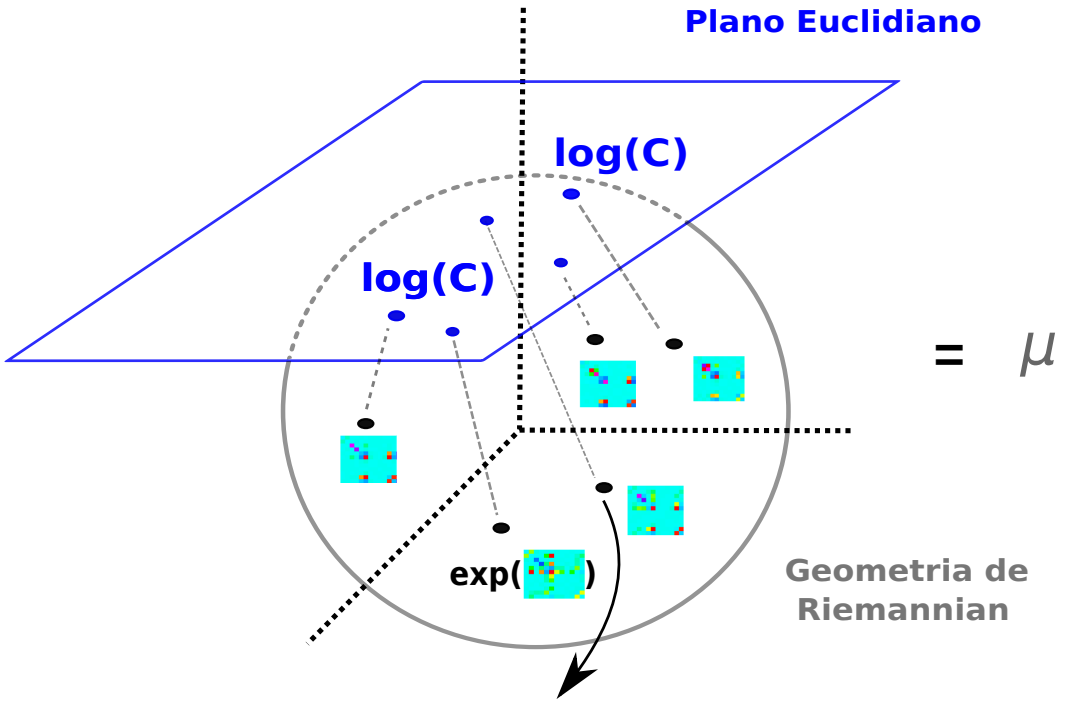

Figure 4. Media de Riemann: En esta figura se muestra el proceso iterativo entre las operaciones $\log$ y $\exp$ para obtener la media de las matrices de covarianza regional.

El descriptor final del video se construye mediante las concatenación del conjunto de matrices de covarianza regionales como $V_d = \{\mu_{c_{R1}}, \mu_{c_{R2}}, \ldots, \mu_{c_{Rn}}\}$. Cada una de las medias $\mu_{c_{Ri}}$ representa la correlación regional del conjunto de primitivas utilizadas para la codificación de la acción. Además, teniendo en cuenta que esta media es la covarianza mínima entre el conjunto de entrada, sus propiedades siguen correspondiendo a una matriz simétrica y positiva. Este hecho implica que únicamente sea necesario usar el triangulo superior/inferior de la matriz para la descripción de la acción.

3.5 MAQUINA DE VECTOR DE SOPORTE

Finalmente, la clasificación de actividades del descriptor de covarianza se logró utilizando una Máquina de Vector de Soporte (*SVM*). La estrategia SVM se ha utilizado ampliamente en problemas de aprendizaje supervisado, clasificación y problemas de regresión [24, 25]. De hecho, este algoritmo es una de las selecciones más comunes para los problemas de reconocimiento de actividades debido a la correlación adecuada entre la precisión y el costo computacional. Dado que el descriptor de actividadesV_d aquí propuesto es construido por el conjunto de ma-

trices de covarianza medias, es necesario proyectar este descriptor al espacio euclidiano como $\log(V_d) = \{\log(\mu_{c_{R1}}), \log(\mu_{c_{R2}}), \ldots, \log(\mu_{c_{Rn}})\}$ mediante el uso de una descomposición espectral $\log(\mu_{c_{Ri}}) = \Sigma \log(\lambda) \Sigma^T$, como se explica en la sección 3.4. De esta forma, las medias de covarianza utilizadas como descriptor permanecerán en el espacio Euclidiano y es posible usar las técnicas clásicas.

El presente enfoque se implementó utilizando un *Clasificación multiclase SVM uno contra uno* con una Función de Base Radial (*RBF*) núcleo (*kernel*) [26]. Aquí, las clases representan las actividades y los hiperplanos óptimos los separan mediante una fórmula clásica de margen máximo. Para las clases k de movimiento, se aplica una estrategia de votación mayoritaria sobre los resultados de los clasificadores binarios $\frac{k(k-1)}{2}$. Se realizó un análisis de sensibilidad del parámetro $(\ gamma a, C)$ con una búsqueda de parámetros utilizando un esquema de validación cruzada y seleccionando los parámetros con el mayor número de verdaderos positivos.

Capítulo 4
EVALUACIÓN Y RESULTADOS

El enfoque propuesto fue evaluado en el conjunto de datos públicos UT-Interaction (*High-level Human Interaction Recognition Challenge*) que exhibe actividades humanas complejas en escenarios reales de vigilancia remota [27]. Este conjunto de datos contiene vídeos de ejecuciones continuas para 6 clases de interacciones humanas, tales como: dar la mano (*shake-hands*), apuntar (*point*), abrazar (*hug*), empujar (*push*), patear (*kick*) y golpear (*punch*). Los vídeos tienen una resolución espacial de 720×480 con 30fps. El conjunto de datos se divide en dos grupos de 60 vídeos. El primer grupo fue capturado en un fondo estático con poco movimiento de la cámara, mientras que el segundo fue capturado en un fondo donde se reportan movimientos de la cámara y también existen movimiento humanos que representan otras acciones en el fondo del video, pero que no hacen parte de la acción principal. En la Figura 5 se muestra un ejemplo de las diferentes actividades del conjunto de datos UT, capturadas en el primer conjunto de datos.

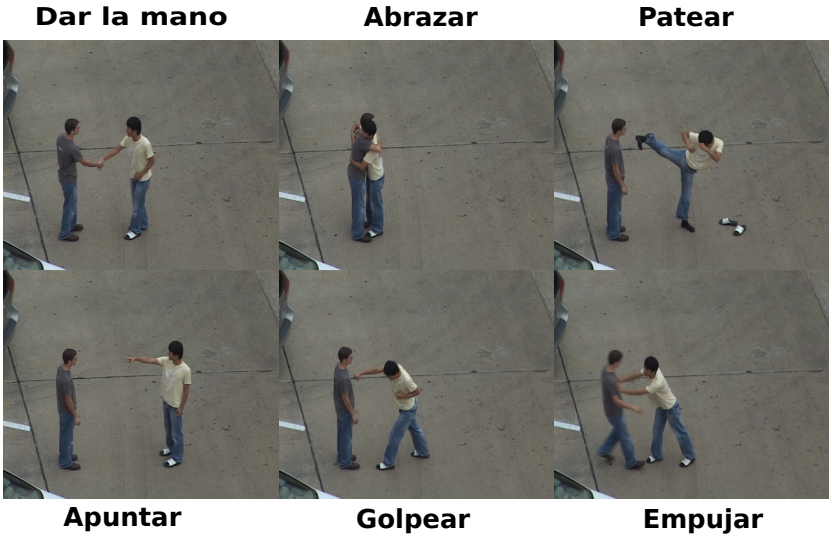

Figure 5. Clases de actividades humanas del conjunto de datos UT-Interaction capturadas para el primer grupo de datos. En total son 6 clases diferentes de actividades para todo UT.

En la Figura 6 se representa el calculo típico del flujo óptico de largo desplazamiento [19] para un conjunto de secuencias del conjunto de datos de UT-Interaction. Como se puede observar, el flujo denso capturado en cada cuadro de la secuencia logra representar los principales gestos de las actividades registradas en la secuencias de video. El campo vectorial se representa como un mapa de colores que describen la dirección del movimiento, mientras que la intensidad de la imagen representa la magnitud del movimiento. Por ejemplo, para la primera secuencia, los valores en rojo representan desplazamientos principalmente hacia la derecha, mientras en la segunda secuencia los colores en verde representan secuencias a la izquierda. También se puede observar como se hace una apropiada descripción local de los gestos que caracterizan las dos acciones, mientras que el fondo permanece principalmente sin vectores de desplazamiento significativos.

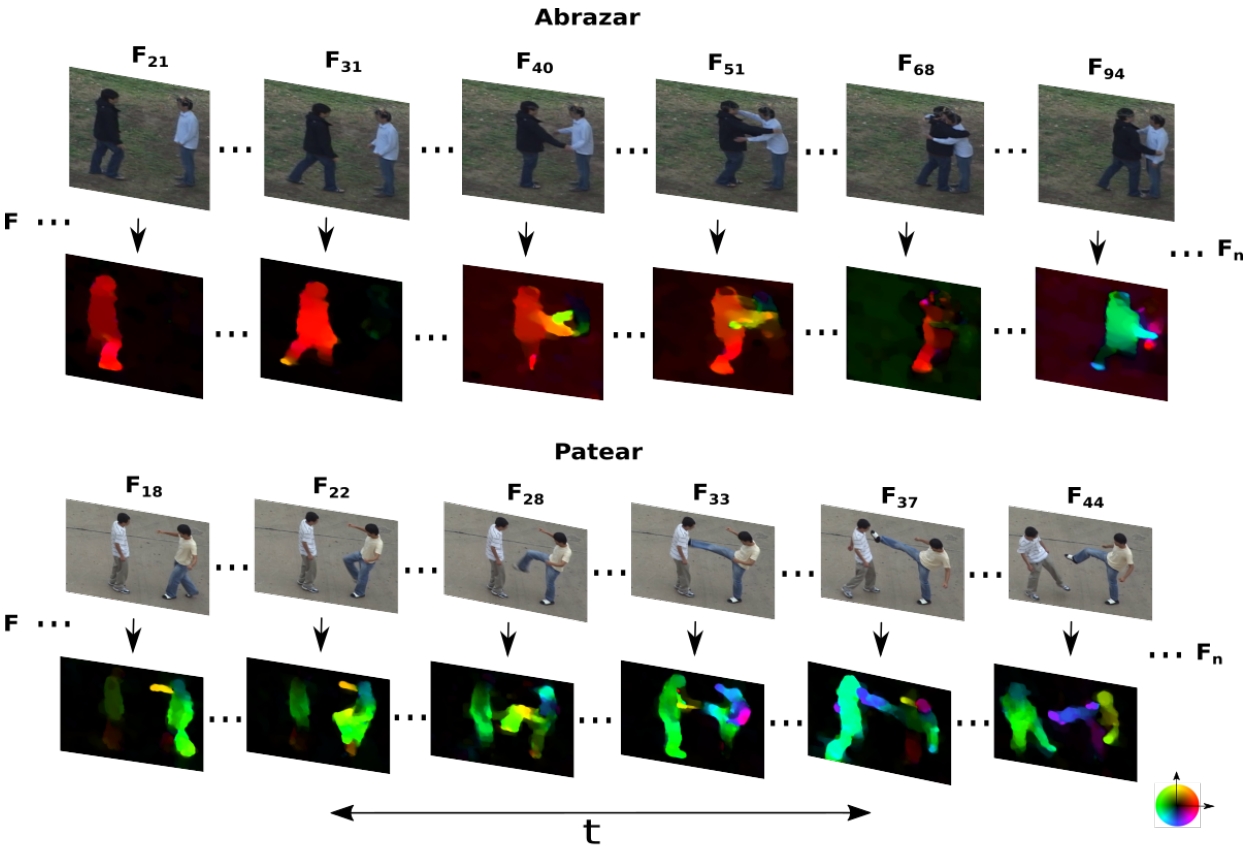

Figure 6. Ejemplo de la representación del calculo de flujo óptico de largo desplazamiento para las actividades abrazar y patear de algunos vídeos del conjunto de datos UT-Interaction

La figura 7 se muestra el cálculo de la matriz de covarianza para cada actividad del conjunto de datos de UT-Interaction. La matriz de covarianza representada en la figura fue calculada de forma global a nivel del cuadro, utilizando las siguientes características de izquierda a derecha: $\|V(t)\|, \theta_V(t), V(t), T(t), N(t), a_T(t)$ y $a_N(t)$. Se observa altas correlaciones de los componentes de velocidad con las aceleraciones calculadas y así como también con las características de rapidez y angulo de la velocidad. Estos patrones se correlacionan de diferente manera para las acciones del conjunto de datos, lo cual puede establecer patrones para realizar una clasificación correcta.

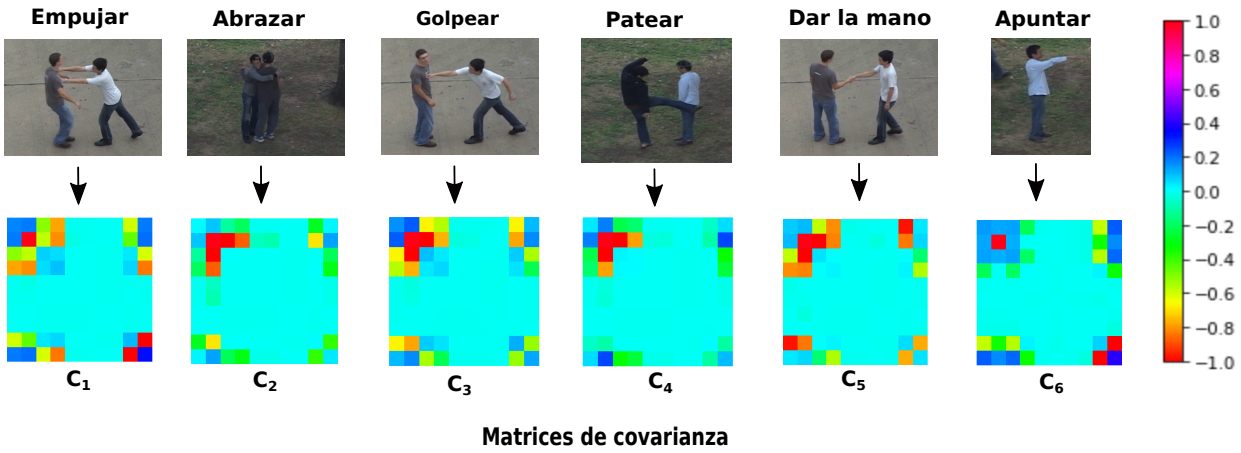

Figure 7. Representación del calculo de las matrices de covarianza de forma global a nivel del cuadro para cada actividad del conjunto de datos UT-Interaction

Para la validación del método propuesto se utilizo la estrategia de *k-fold cross validation*, en la cual se construyen iterativamente dos diferentes conjuntos de datos de entrenamiento y validación. Esta herramienta estadística permite desarrollar diferentes experimentos utilizando varios conjuntos de muestras, los cuales son iterativamente calculados, y permite tener una mejor aproximación de las estadísticas de exactitud. Para la implementación del *k-fold*, nosotros seleccionamos un $K = 10$, que nos permitió obtener la exactitud promedio de cada una de las actividades para diferentes conjuntos de datos.

Las tablas 1 y 2 muestran las matrices de confusión obtenidas al evaluar el método propuesto con el conjunto de datos de UT-Interaction. En promedio, se obtuvo una exactitud de 80.0% y 61.66% para el conjunto de datos uno y dos, respectivamente. En este experimento, el método propuesto fue caracterizado con las siguientes características cinemáticas: $V(t), \|V(t)\|, \theta_V(t), T(t), N(t), a_T(t)$ y $a_N(t)$. Estas primitivas son eficientes computacionalmente y explotan el carácter dinámico de las acciones, siendo robusto a cambios de apariencia. Para el primer grupo de datos UT se logra un 80% de exactitud que es competitivo con respecto al estado del arte, pero con la mayor ventaja de ser un descriptor compacto de 275 valores escalares. Este descriptor resulta ideal para aplicaciones con arquitecturas computacionales de bajo nivel que tienen recursos limitados. Por otra parte, el método propuesto baja su exactitud en el segundo conjunto de datos, debido principalmente a los movimientos de la cámara que afectan el cálculo del flujo óptico. Teniendo en cuenta esta limitación, entonces las primitivas cinemáticas pueden presentar algunas limitaciones para la descripción de las acciones. Por otra parte las acciones registradas como el fondo del video, también pueden afectar la descripción de las acciones utilizando las matrices de covarianza.
En resumen, el enfoque propuesto logra una caracterización dinámica relevante de las diferentes

Categoría	hs	hg	ki	po	pun	pus
Hand Shaking	90	10	0	0	0	0
Hugging	0	100	0	0	0	0
Kicking	0	0	80	0	10	10
Pointing	0	0	0	100	0	0
Punching	20	0	20	0	40	20
Pushing	0	0	20	0	10	70

Table 1. Matriz de confusión obtenida para el conjunto de datos de UT-Interaction numero 1 al evaluar el descriptor propuesto con primitivas de movimiento. Los resultados están en %

Categoría	hs	hg	ki	po	pun	pus
Hand Shaking	60	30	10	0	0	0
Hugging	20	70	0	0	0	10
Kicking	0	0	60	0	40	0
Pointing	0	0	0	100	0	0
Punching	0	0	30	0	40	30
Pushing	10	10	10	0	30	40

Table 2. Matriz de confusión obtenida para el conjunto de datos de UT-Interaction numero 2 al evaluar el descriptor propuesto con primitivas de movimiento. Los resultados están en %

actividades de interacción humana. Sin embargo, tales actividades son a menudo el resultado de combinaciones de patrones de movimiento complejos que pueden ocurrir durante un intervalo de tiempo corto. Del mismo modo, algunas de estas actividades de interacción comparten patrones de movimiento locales que pueden conducir a predicciones erróneas. Por ejemplo, las interacciones como *hand shaking*, *pointing* o *pushing*, comparten movimientos de miembros similares durante cierto intervalo temporal. Por ejemplo, estas tres acciones tienen el levantamiento de los brazos como característica dinámica en común. Por otra parte, en cuanto al conjunto de datos dos de UT-interaction, los escenarios aumentan enormemente la complejidad de la descripción, por ejemplo, consideran un grupo de actores en movimiento con una gran variabilidad en términos de apariencia, interacción y fondo.

Teniendo en cuenta la flexibilidad de la matriz de covarianza para agregar características que permitan representar las acciones, en un segundo experimento, se agregaron a nuestro descriptor características de apariencia, tales como: $S_{\|V(t)\|}, S_{\theta_v(t)}$ y $S^{''V(t)}(t)$. Estas características básicamente son calculadas como los gradientes espaciales de cada frame. En promedio, ob-

tuvimos una precisión de 75.0% y 55.0% para los conjuntos de datos 1 y 2 respectivamente. El agregar primitivas de apariencia a nuestro descriptor propuesto no muestra ninguna mejora en la clasificación y reconocimiento de actividades humanas. Este resultado demuestra que para el descriptor propuesto y las acciones registradas en las secuencias de video, las cinemáticas son las principales primitivas que permiten explicar las actividades. Por el contrario, las características de gradiente son similares entre las acciones lo cual limitan al descriptor propuesto para desarrollar una apropiada clasificación. Las tablas 3 y 4 muestran los resultados de las matrices de confusión para nuestro descriptor de primitivas de apariencia y movimiento para cada conjunto de datos UT-Interaction.

Categoría	hs	hg	ki	po	pun	pus
Hand Shaking	90	10	0	0	0	0
Hugging	10	80	0	0	0	10
Kicking	0	10	80	0	0	10
Pointing	0	0	0	100	0	0
Punching	10	10	0	0	40	40
Pushing	0	20	0	0	20	60

Table 3. Matriz de confusión para el conjunto de datos de UT-Interaction numero 1 para el descriptor de covarianza propuesto con primitivas de movimiento y apariencia. Los resultados están en %

Categoría	hs	hg	ki	po	pun	pus
Hand Shaking	30	50	10	0	0	10
Hugging	10	70	0	0	0	20
Kicking	0	0	60	0	20	20
Pointing	0	0	0	100	0	0
Punching	0	0	10	10	10	70
Pushing	0	10	0	0	30	60

Table 4. Matriz de confusión para el conjunto de datos de UT-Interaction numero 2 para el descriptor de covarianza propuesto con primitivas de movimiento y apariencia. Los resultados están en %

La Tabla 5 informa la comparación del descriptor de movimiento propuesto con otras estrategias del estado del arte. Algunos de estos enfoques logran altos índices de precisión en problemas relacionados con el reconocimiento de acciones, pero exigen un procesamiento completo del

video para calcular las características que describen la secuencia. Por ejemplo, el método de votación propagativa [28] basado en la transformada Hough calcula coincidencias del descriptor mediante el uso de árboles de proyección aleatorios, una estrategia precisa que resulta computacionalmente costosa y prohibitiva en muchas aplicaciones donde los requerimientos de hardware son limitados. Como se muestra en la Tabla 6, el enfoque propuesto logra ser competitivo con respecto a otros métodos propuestos por el estado del arte, siendo compacto en su descripción, siendo un descriptor fácil de implementar, flexible para diferentes aplicaciones y eficiente en términos computacionales.

Enfoques	Precisión UT- conjunto 1	Precisión UT- conjunto 2
Votación propagativa [28]	93	91
Enfoque propuesto	80.0	61.66
Daysy [29]	71	51
SIFT 3D [30]	63	55
Slimani 2014 [31]	41	
Ryoo 2011 [32]	71.7	
Mukherjee [33]	79.17	
Xiaofei [34]	83.33	

Table 5. Precisión promedio para diferentes estrategias informadas en el estado del arte. Aunque la votación de propagación logra mejores resultados en términos de precisión, la coincidencia de las características que usan árboles de proyección aleatorios es computacionalmente costosa. En Xiaofei *et. Alabama.* este trabajo integra el histograma de ocurrencias de BoW con HoG, representando nuevamente un alto tiempo de cálculo para obtener una representación de acción. Por el contrario, nuestro enfoque propuesto produce un descriptor compacto que tiene en cuenta diferentes profundidades de intervalo de tiempo usando la misma fuente de primitivas, es decir, un flujo óptico denso. Además, la naturaleza recursiva del enfoque propuesto hace que este estimador se actualice constantemente para que las secuencias parciales puedan predecirse.

Finalmente, teniendo en cuenta el bajo costo computacional y el diseño compacto del descriptor propuesto en este trabajo, se realizó un experimento inicial para evaluar su comportamiento en aplicaciones en linea. Este experimento consistió en estimar las actividades utilizando únicamente segmentos parciales del video. Inicialmente se calculan las covarianzas medias que representan el video utilizando únicamente el 10 % de los cuadros del video, luego se incremento el numero de cuadros y se midió la exactitud obtenida por el descriptor. Para este experimento seleccionamos aleatoriamente 12 vídeos por cada conjunto de UT-Interaction, por cada conjunto tomamos dos vídeos de cada clase. Sobre cada video calculamos el descriptor de covarianza

propuesto cambiando el numero de frames que seleccionamos por video, empezando con el 10% de total de frames, incrementando de a 10% hasta llegar finalmente a seleccionar el 100% de los frames. Por cada interacción calculamos la precisión del descriptor en un clasificador para los conjuntos 1 y 2 respectivamente.

En la Tabla 6 se muestra la estimación parcial obtenida utilizando diferentes porcentajes del video. Como se puede observar en la tabla, utilizando el 60 % de las secuencias del video, 4 de las 6 actividades del conjunto UT se logran clasificar correctamente. En este caso, las acciones que comparten patrones dinámicos similares pueden presentar limitaciones, como se reportan en las secuencias completas. También cabe resaltar que con el 90 % de las secuencias del video se logran los mejores resultados para el reconocimiento de las secuencias en las diferentes acciones.

	Conjunto %			
Frames %	**Test 1**		**Test 2**	
	1	***2***	***1***	***2***
10	0	16.66	0	16.66
20	16.66	16.66	16.66	0
30	16.66	33.33	33.33	0
40	33.33	33.33	33.33	0
50	33.33	33.33	50	16.66
60	66.66	50	66.66	50
70	50	66.66	66.66	50
80	66.66	66.66	83.33	50
90	66.66	66.66	83.33	66.66
100	83.33	66.66	50	50

Table 6. Precisión parcial del descriptor propuesto para cada porcentaje de frames seleccionados dada una secuencia para cada conjunto de datos de UT-Interaction. Se seleccionaron aleatoriamente 12 vídeos por cada conjunto de UT, por cada grupo tomamos dos vídeos de cada clase.

En este experimento inicial, como datos de entrenamiento se utilizaron descriptores de las acciones completas. Para trabajos futuros se requiere hacer un entrenamiento especial con secuencias parciales que permitan caracterizar mejor las acciones.

Capítulo 5
CONCLUSIONES Y PERSPECTIVAS

El descriptor propuesto es compacto y logra describir las acciones utilizando únicamente 275 valores escalares, en el caso de utilizar 5 regiones. Para el uso de una única región el descriptor propuesto puede describir una secuencia completa con únicamente 55 valores, lo cual puede resultar de interés para arquitecturas de bajo costo computacional o en aplicaciones que se requiere el reconocimiento y clasificación en linea. En cuanto a la exactitud del método propuesto, se logra alcanzar hasta un 80% en un conjunto de datos estable y hasta un 61.66 % en un conjunto de datos que involucra escenarios dinámicos y con movimientos de la cámara.

Futuros trabajos involucrarán la validación con otros conjuntos de datos públicos para el reconocimiento de acciones que permitan evaluar el desempeño en otras aplicaciones del reconocimiento de acciones. También se propone evaluar otras metodologías para la selección de regiones salientes que puedan ser representadas por la matriz de covarianza. También se pretende modificar la versión del descriptor actual para lograr un reconocimiento en linea utilizando la dinámica histórica de los cuadros precedentes.

REFERENCIAS

[1] Poppe, R. A survey on vision-based human action recognition. *Image and vision computing 28*, 6 (2010), 976–990.

[2] Enzweiler, M., and Gavrila, D. M. Monocular pedestrian detection: Survey and experiments. *IEEE transactions on pattern analysis and machine intelligence 31*, 12 (2009), 2179–2195.

[3] Gandhi, T., and Trivedi, M. M. Pedestrian protection systems: Issues, survey, and challenges. *IEEE Transactions on intelligent Transportation systems 8*, 3 (2007), 413–430.

[4] Geronimo, D., Lopez, A. M., Sappa, A. D., and Graf, T. Survey of pedestrian detection for advanced driver assistance systems. *IEEE transactions on pattern analysis and machine intelligence 32*, 7 (2010), 1239–1258.

[5] Bobick, A. F., and Davis, J. W. The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence 23*, 3 (2001), 257–267.

[6] Wang, Y., Huang, K., and Tan, T. Human activity recognition based on r transform. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (2007), IEEE, pp. 1–8.

[7] Souvenir, R., and Babbs, J. Learning the viewpoint manifold for action recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (2008), IEEE, pp. 1–7.

[8] Huang, F., and Xu, G. Viewpoint insensitive action recognition using envelop shape. *Computer Vision–ACCV 2007* (2007), 477–486.

[9] Cherla, S., Kulkarni, K., Kale, A., and Ramasubramanian, V. Towards fast, view-invariant human action recognition. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on* (2008), IEEE, pp. 1–8.

[10] Weinland, D., Ronfard, R., and Boyer, E. Free viewpoint action recognition using motion history volumes. *Computer vision and image understanding 104*, 2 (2006), 249–257.

[11] Laptev, I., and Lindeberg, T. Space-time interest points. In *9th International Conference on Computer Vision, Nice, France* (2003), IEEE conference proceedings, pp. 432–439.

[12] Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on* (2005), IEEE, pp. 65–72.

[13] Laptev, I., Caputo, B., Schüldt, C., and Lindeberg, T. Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding 108*, 3 (2007), 207–229.

[14] Gowayyed, M. A., Torki, M., Hussein, M. E., and El-Saban, M. Histogram of oriented displacements (hod): Describing trajectories of human joints for action recognition. In *IJCAI* (2013).

[15] Robertson, N., and Reid, I. A general method for human activity recognition in video. *Computer Vision and Image Understanding 104*, 2 (2006), 232–248.

[16] Liu, A.-A., Xu, N., Su, Y.-T., Lin, H., Hao, T., and Yang, Z.-X. Single/multi-view human action recognition via regularized multi-task learning. *Neurocomputing 151* (2015), 544–553.

[17] Tuzel, O., Porikli, F., and Meer, P. Region covariance: A fast descriptor for detection and classification. *Computer Vision–ECCV 2006* (2006), 589–600.

[18] Ma, B., Su, Y., and Jurie, F. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing 32*, 6 (2014), 379–390.

[19] Brox, T., and Malik, J. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE transactions on pattern analysis and machine intelligence 33*, 3 (2011), 500–513.

[20] Hussein, M. E., Torki, M., Gowayyed, M. A., and El-Saban, M. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Twenty-Third International Joint Conference on Artificial Intelligence* (2013).

[21] Pennec, X., Fillard, P., and Ayache, N. A riemannian framework for tensor computing. *International Journal of Computer Vision 66*, 1 (2006), 41–66.

[22] Pennec, X. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision 25*, 1 (2006), 127–154.

[23] Fletcher, P. T., and Joshi, S. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing 87*, 2 (2007), 250–262.

[24] Suykens, J. A., and Vandewalle, J. Least squares support vector machine classifiers. *Neural processing letters 9*, 3 (1999), 293–300.

[25] Shao, Y.-H., Deng, N.-Y., and Yang, Z.-M. Least squares recursive projection twin support vector machine for classification. *Pattern Recognition 45*, 6 (2012), 2299–2307.

[26] 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.

[27] Ryoo, M. S., and Aggarwal, J. K. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010.

[28] Yu, G., Yuan, J., and Liu, Z. Propagative hough voting for human activity recognition. In *Computer Vision-ECCV 2012* (2012), A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., vol. 7574 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 693–706.

[29] Cao, X., Zhang, H., Deng, C., Liu, Q., and Liu, H. Action recognition using 3d daisy descriptor. *Mach. Vision Appl. 25*, 1 (Jan. 2014), 159–171.

[30] Scovanner, P., Ali, S., and Shah, M. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia* (2007), ACM, pp. 357–360.

[31] Nour el houda Slimani, K., Benezeth, Y., and Souami, F. Human interaction recognition based on the co-occurence of visual words. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2014), pp. 455–460.

[32] Ryoo, M. S. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *2011 International Conference on Computer Vision* (2011), IEEE, pp. 1036–1043.

[33] Mukherjee, S., Biswas, S. K., and Mukherjee, D. P. Recognizing interaction between human performers using 'key pose doublet'. In *Proceedings of the 19th ACM international conference on Multimedia* (2011), ACM, pp. 1329–1332.

[34] Ji, X., Wang, C., Zuo, X., and Wang, Y. Multiple feature voting based human interaction recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition 9*, 1 (2016), 323–334.

BIBLIOGRAFIA

BOBICK, Aaron F.. ; DAVIS, James W.. . The recognition of human movement using temporal templates. IEEE Transactions on pattern analysis and machine intelligence 23.3 (2001): 257-267.

BROX, Thomas; MALIK, Jitendra. Large displacement optical flow: descriptor matching in variational motion estimation. IEEE transactions on pattern analysis and machine intelligence 33.3 (2011): 500-513

CAO, Xiaochun, *et al.* Action recognition using 3D DAISY descriptor. Machine vision and applications 25.1 (2014): 159-171.

CHANG, Chih-Chung; LIN, Chih-Jen. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2.3 (2011): 27.

CHERLA, Srikanth, *et al.* Towards fast, view-invariant human action recognition. Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on. IEEE, 2008.

DOLLÁR, Piotr, *et al.* Behavior recognition via sparse spatio-temporal features. Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on. IEEE, 2005.

ENZWEILER, Markus; GAVRILA, Dariu M. Monocular pedestrian detection: Survey and experiments. IEEE transactions on pattern analysis and machine intelligence 31.12 (2009): 2179-2195.

FLETCHER, P. Thomas; JOSHI, Sarang. Riemannian geometry for the statistical analysis of diffusion tensor data. Signal Processing 87.2 (2007): 250-262.

GANDHI, Tarak; TRIVEDI, Mohan Manubhai. Pedestrian protection systems: Issues, survey, and challenges. IEEE Transactions on intelligent Transportation systems 8.3 (2007): 413-430.

GERONIMO, David, *et al.* Survey of pedestrian detection for advanced driver assistance systems. IEEE transactions on pattern analysis and machine intelligence 32.7 (2010): 1239-1258.

GOWAYYED, Mohammad Abdelaziz, *et al.* Histogram of Oriented Displacements (HOD): Describing Trajectories of Human Joints for Action Recognition. IJCAI. 2013.

XU, Guangyou, *et al.* Viewpoint insensitive action recognition using envelop shape. Asian Conference on Computer Vision. Springer, Berlin, Heidelberg, 2007.

HUSSEIN, Mohamed E., *et al.* Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations. IJCAI. Vol. 13. 2013.

JI, Xiaofei, *et al.* Multiple feature voting based human interaction recognition. Int. J. Signal Process. Image Process. Pattern Recognit 9.1 (2016): 323-334.

LAPTEV, Ivan, *et al.* Local velocity-adapted motion events for spatio-temporal recognition. Computer vision and image understanding 108.3 (2007): 207-229.

LAPTEV, Ivan. International journal of computer vision 64.2-3 (2005): 107-123.

LIU, An-An, *et al.* Single/multi-view human action recognition via regularized multi-task learning. Neurocomputing 151 (2015): 544-553.

MA, Bingpeng; SU, Yu; JURIE, Frédéric. Covariance descriptor based on bio-inspired features for person re-identification and face verification. Image and Vision Computing 32.6-7 (2014): 379-390.

MUKHERJEE, Snehasis; BISWAS, Sujoy Kumar; MUKHERJEE, Dipti Prasad. Recognizing interaction between human performers using'key pose doublet'. Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011.

SLIMANI, K. Nour El Houda; BENEZETH, Yannick; SOUAMI, Feriel. Human interaction recognition based on the co-occurence of visual words. IEEE CVPR CMSI workshop. 2014.

PENNEC, Xavier. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. Journal of Mathematical Imaging and Vision 25.1 (2006): 127.

PENNEC, Xavier; FILLARD, Pierre; AYACHE, Nicholas. A Riemannian framework for tensor computing. International Journal of computer vision 66.1 (2006): 41-66.

POPPE, Ronald. A survey on vision-based human action recognition. Image and vision computing 28.6 (2010): 976-990.

ROBERTSON, Neil; REID, Ian A general method for human activity recognition in video. Computer Vision and Image Understanding 104.2-3 (2006): 232-248.

RYOO, Michael S. Human activity prediction: Early recognition of ongoing activities from streaming videos. Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011.

RYOO, Michael S.; AGGARWAL, J. K. UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA). IEEE International Conference on Pattern Recognition Workshops. Vol. 2. 2010.

SCOVANNER, Paul; ALI, Saad; SHAH, Mubarak. A 3-dimensional sift descriptor and its application to action recognition. Proceedings of the 15th ACM international conference on Multimedia. ACM, 2007.

SHAO, Yuan-Hai; DENG, Nai-Yang; YANG, Zhi-Min. Least squares recursive projection twin support vector machine for classification. Pattern Recognition 45.6 (2012): 2299-2307.

SOUVENIR, Richard; BABBS, Justin. Learning the viewpoint manifold for action recognition. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008.

TUZEL, Oncel; PORIKLI, Fatih; MEER, Peter. Region covariance: A fast descriptor for detection and classification. European conference on computer vision. Springer, Berlin, Heidelberg, 2006.

WANG, Ying; HUANG, Kaiqi; TAN, Tieniu. Human activity recognition based on r transform. Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.

WEINLAND, Daniel; RONFARD, Remi; BOYER, Free viewpoint action recognition using motion history volumes. Computer vision and image understanding 104.2-3 (2006): 249-257.

YU, Gang; YUAN, Junsong; LIU, Zicheng. Propagative hough voting for human activity recognition. European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012.