RECONOCIMIENTO DE ACTIVIDADES EN VIDEO UTILIZANDO UN DESCRIPTOR LOCAL DE COVARIANZA VOLUMÉTRICA

OSCAR MAURICIO MENDOZA CASAS

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2020

RECONOCIMIENTO DE ACTIVIDADES EN VIDEO UTILIZANDO UN DESCRIPTOR LOCAL DE COVARIANZA VOLUMÉTRICA

OSCAR MAURICIO MENDOZA CASAS

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

Director:
Fabio Martínez Carrillo
Ph.D en Ingenieria de Sistemas y Computacion

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2020

AGRADECIMIENTOS

El autor expresa su agradecimiento:

A mi madre, Patricia Casas Fernández, por estar ahí desde el primer día, por el apoyo incondicional y su hermoso carisma que me impulsa a seguir adelante con mis metas y sueños.

A mi difunto padre, que en paz descanse, Gerardo Mendoza Mendoza, el cual, que, aunque no pudo estar presente en mi vida universitaria, sé que estaría orgulloso.

A mi hermano, que es incondicional y único en la vida.

A mi director de proyecto, Fabio el cual me guio desde el comienzo, siendo un excelente profesor, paciente, constante y brindarme un acompañamiento estupendo durante gran parte de mi carrera.

A Kim, por brindarme su apoyo en todo momento que fue necesario y estar a mi lado, impulsándome a seguir adelante.

Al ingeniero Wilson Moreno, con el empezó todo el camino del presente proyecto.

A mi familia, por su apoyo único y especial.

A mis amigos Leonardo, Andrés y Daniel, los cuales han sido desde hace muchos años hermanos indiscutibles en esta vida.

A mis compañeros, amigos y colegas de Ingeniería de sistemas, principalmente Edgar, Henry,

Rubén, Douglas, Santiago, Andrés, Juan Sebastián, Diego, entre muchos otros.

A mis amigos del grupo de investigación $BIVL^2AB$ Jefferson, Guayacán, Gustavo, Isail, Alejandra, Lina y demás miembros, los cuales me brindaron apoyo cuando lo requerí.

Por ultimo y no menos importante, a la escuela de Ingeniería de Sistemas e Informática(EISI)., por brindarme las herramientas y los medios para formarme como profesional.

Índice general

Índice de figuras

Pág

---

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

[2] Mark Sandler y col. "Mobilenetv2: Inverted residuals and linear bottlenecks". En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, págs. 4510-4520.

---

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

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

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

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

Índice de cuadros

Pág

RESUMEN

TÍTULO: RECONOCIMIENTO DE ACTIVIDADES EN VIDEO UTILIZANDO UN DESCRIPTOR LOCAL DE COVARIANZA VOLUMÉTRICA [*]

AUTOR: OSCAR MAURICIO MENDOZA CASAS [**]

PALABRAS CLAVE: COVARIANZA, RECONOCIMIENTO DE ACCIONES, TRAYECTORIAS DENSAS.

DESCRIPCIÓN: La caracterización de acciones involucra reconocer gestos, actividades cotidianas e interacciones entre humanos, objetos o agentes presentes en un video. Sin embargo, esta caracterización es compleja debido a las múltiples variaciones de las acciones, el cambio de iluminación, la superposición de objetos, variaciones de movimiento, entre otros factores. Existen numerosos métodos que han sido propuestos para el reconocimiento y clasificación de acciones, tales como algoritmos densos de aprendizaje mediante redes convolucionales profundas. Estos trabajos son ampliamente analizados en imágenes, sin embargo, su extensión a un análisis volumétrico ha sido poco explorada. Además, estos trabajos requieren extensas bases de datos para aprender representaciones volumétricas y sus procesos de entrenamiento suelen ser redundantes y complejos. En el presente trabajo, se presenta un método computacional, el cual reconoce acciones a partir del modelamiento local de covarianzas, que resumen patrones densos convolucionales de forma local, usando un soporte temporal guiado por trayectorias de movimiento. Para ello, cada cuadro del video es descrito por un conjunto de activaciones de arquitecturas convolucionales pre-entrenadas. Las regiones salientes, que siguen trayectorias de movimiento, son utilizadas como entrada en las matrices de covarianza. Para el cálculo de la covarianza se utilizó una estrategia integral que permite permanecer eficientes en cuanto al costo computacional. Entonces, para cada secuencia se calculan M covarianzas locales, las cuales son representadas por K, K« M centroides, que conforman el descriptor de video. El método propuesto logro en la base de datos publica UT-Interaction, una exactitud de 83.3 %, una sensibilidad de 86.1 %, y una especificidad de 91.3 % para su primer segmento, y para el segundo, una exactitud de 83.3 %, una sensibilidad de 83.3 %, y una especificidad de 96.4 %.

---

[*] Trabajo de investigación

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Fabio Martínez Carrillo, Ph.D.

ABSTRACT

TITLE: VIDEO ACTION RECOGNITION USING A LOCAL VOLUMETRIC COVARIANCE DESCRIPTOR [*]

AUTHOR: OSCAR MAURICIO MENDOZA CASAS [**]

KEYWORDS: COVARIANCE, ACTION RECOGNITION, DENSE TRAJECTORIES.

DESCRIPTION: Video action recognition involves recognizing gestures, everyday activities and interactions between humans, objects or agents present in a video. However, this characterization is complex due to the multiple variations of actions, the change of lighting, the overlapping of objects, variations of movement, among other factors. There are numerous methods that have been proposed for the recognition and classification of actions, such as dense learning algorithms using deep convolutional networks. These works are widely analyzed in images, however, their extension to a volumetric analysis has been little explored. In addition, these works require extensive databases to learn volumetric representations and their training processes are often redundant and complex. In the present work, a computational method is presented, which recognizes actions from local covariance modeling, which summarizes dense convolutional patterns locally, using a time support guided by motion paths. For this purpose, each frame in the video is described by a set of pre-trained convolutional architecture activations. The outgoing regions, which follow motion paths, are used as input to the covariance matrices. For the calculation of covariance, a comprehensive strategy was used to remain computationally cost efficient. Then, for each sequence, M local covariances are calculated, which are represented by K, K « M centroids, which conform the video descriptor. The proposed method achieved in the public database UT-Interaction, an accuracy of 83.3 %, a sensitivity of 86.1 %, and a specificity of 91.3 % for its first segment, and for the second segment, an accuracy of 83.3 %, a sensitivity of 83.3 %, and a specificity of 96.4 %.

---

[*] Research work

[**] Faculty of Physical-Mechanical Engineering. School of Systems and Computer Engineering. Advisor: Fabio Martínez Carrillo

INTRODUCCIÓN

El reconocimiento de acciones es fundamental en aplicaciones como la biomedicina, los deportes, la video-vigilancia, la robótica entre otras áreas [1]. Sin embargo, esta tarea resulta ser compleja debido a la alta variabilidad de movimientos, inestabilidad de la cámara y el fondo, además de cambios de iluminación y subjetividad en la descripción de acciones.

En la literatura encontramos numerosos trabajos orientados al reconocimiento de acciones, entre ellos están los métodos que usan ingeniería de caracteristicas, los basados en representaciones profundas por cuadro, los basados en representaciones volumétricas y aquellos orientados a explotar la representación y relación temporal a lo largo de la secuencia [2][3][4]. Las estrategias basadas en ingeniería de caracteristicas se han dedicado a describir variaciones locales y espacio-temporales que sean significativas en la representación de videos, como los histogramas 3D de gradientes orientados, que operan en secuencias de flujo óptico [5], o los histogramas de fronteras de movimiento [6]. En esta línea, los trabajos más representativos utilizan trayectorias densas de

---

[1] Hà Quang Minh y Vittorio Murino. "Covariances in computer vision and machine learning". En: *Synthesis Lectures on Computer Vision* 7.4 (2017), págs. 1-170.

[2] Suraj Prakash Sahoo y Samit Ari. "On an algorithm for human action recognition". En: *Expert Systems with Applications* 115 (2019), págs. 524-534.

[3] Gül Varol, Ivan Laptev y Cordelia Schmid. "Long-term temporal convolutions for action recognition". En: *IEEE transactions on pattern analysis and machine intelligence* 40.6 (2017), págs. 1510-1517.

[4] Jue Wang y col. "Video representation learning using discriminative pooling". En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, págs. 1149-1158.

[5] Ivan Laptev y col. "Learning realistic human actions from movies". En: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, págs. 1-8.

[6] Heng Wang y col. "Dense trajectories and motion boundary descriptors for action recognition". En: *International journal of computer vision* 103.1 (2013), págs. 60-79.

movimiento para describir patrones temporales a lo largo del video [7]. Por otra parte, la representación compleja e invariante de acciones se destaca mediante el uso de métodos convolucionales [48]. Muchos de estos trabajos usan representaciones a nivel de cuadro para modelar instancias independientes en cada imagen [910]. Un trabajo saliente se basa en el uso de arquitecturas paralelas para el análisis de secuencias RGB y de flujo óptico *two stream learning (Aprendizaje en dos corrientes)*[11], que simultáneamente codifica información de apariencia y de movimiento relacionada con las acciones. Sin embargo, estos enfoques no codifican movimientos descritos en secuencias largas, lo cual puede limitar la descripción de ciertos movimientos complejos. Por otra parte, los enfoques basados en convoluciones 3D han permitido modelar movimientos largos y característicos de las acciones objeto de estudio en un video [3]. Estas arquitecturas sin embargo requieren fuertes esquemas de sub-muestreo que pueden perder información relevante de las acciones. Finalmente, los modelos basados en codificación temporal han sido fundamentados sobre estructuras de aprendizaje de arquitecturas neuronales recurrentes [12]. Normalmente en estos trabajos, se codifican los cuadros independientes y luego se aprende la relación tem-

---

[7] Heng Wang y Cordelia Schmid. "Action recognition with improved trajectories". En: *Proceedings of the IEEE international conference on computer vision*. 2013, págs. 3551-3558.

[8] Amin Ullah y col. "Action recognition in video sequences using deep bi-directional LSTM with CNN features". En: *IEEE Access* 6 (2017), págs. 1155-1166.

[9] Andrej Karpathy y col. "Large-scale video classification with convolutional neural networks". En: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014, págs. 1725-1732.

[10] Limin Wang y col. "Towards good practices for very deep two-stream convnets". En: *arXiv preprint arXiv:1507.02159* (2015).

[11] Karen Simonyan y Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos". En: *Advances in neural information processing systems*. 2014, págs. 568-576.

[12] Jeffrey Donahue y col. "Long-term recurrent convolutional networks for visual recognition and description". En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, págs. 2625-2634.

poral para cuantificar el descriptor de video [13]. Una limitación principal de estos estudios es que usan acercamientos globales a nivel de cuadro, codificando en muchas ocasiones el fondo y descartando puntos locales salientes representativos de las acciones. Además, estos trabajos son computacionalmente complejos, lo cual limita su modelamiento en escenarios restrictivos y con requerimientos de tiempo real.

Como principal contribución se establece un método para el reconocimiento de acciones basado en covarianza espacio temporal con soporte en trayectorias densas de movimiento. Para esto, el presente enfoque cuantifica caracteristicas visuales, como las activaciones de las convoluciones del primer bloque de una red neuronal a lo largo de la secuencia de video. Además, para la representación del movimiento, se calculan las trayectorias densas de movimiento sobre el video. Posteriormente, se codifican localmente matrices de covarianza de las caracteristicas profundas estimadas a lo largo de las trayectorias. El conjunto de covarianzas volumétricas calculadas coexiste en el espacio de Riemann y representan cada acción particular codificada en el video. Luego, las covarianzas son proyectadas al espacio euclidiano para operar con algoritmos de aprendizaje de máquina. El descriptor final de video es obtenido a partir del conjunto de covarianzas volumétricas más representativo mediante un proceso de agrupamiento. Las covarianzas son mapeadas a un clasificador previamente entrenado, para obtener una etiqueta asociada a la acción más probable en el video.

---

[13] Joe Yue-Hei Ng y col. "Beyond short snippets: Deep networks for video classification". En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, págs. 4694-4702.

1. PLANTEAMIENTO Y JUSTIFICACIÓN DEL PROBLEMA

El reconocimiento de acciones presenta variaciones complejas que dependen de cambios de iluminación, representación de los objetos en el video y variaciones de movimiento, que dificultan la tarea de determinar automáticamente las secuencias de video. En aplicaciones como video vigilancia, realidad aumentada, análisis deportivo se ha propuesto diferentes métodos computacionales, los principales se describen a continuación: Los métodos clásicos para llevar a cabo esta tarea resultan en descriptores de alta dimensionalidad, afectando el costo computacional[14][15]. Los descriptores basados en la covarianza pueden ser una alternativa compacta para la representación de acciones mostrando ventajas en cuanto a costo computacional y eficiencia algorítmica[16][1]. En la literatura encontramos métodos basados en covarianza con diferentes enfoques, desde acercamientos globales donde la covarianza describe el total de la acción a lo largo de todo el video; hasta regionales, donde se centra en áreas específicas de la acción plasmada en la secuencia de video. Estas perspectivas podrían sugerir una pérdida de información en ciertas secciones de la imagen, o por el contrario, puede representar el agregar información irrelevante para la acción, como lo es el fondo o interacciones humanas fuera de la acción principal.

El procesamiento y computo de la covarianza cuantifica cuales caracteristicas estan altamente relacionadas y por lo tanto permite identificar las caracteristicas más significativas en la acción, además la covarianza posibilita un correcto seguimiento de localidad a los movimientos sin una perdida significativa de información, debido a sus propiedades de representación de relaciones

---

[14] Gang Yu, Junsong Yuan y Zicheng Liu. "Propagative hough voting for human activity recognition". En: *European Conference on Computer Vision*. Springer. 2012, págs. 693-706.

[15] Khadidja Nour el houda Slimani, Yannick Benezeth y Feriel Souami. "Human interaction recognition based on the co-occurence of visual words". En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2014, págs. 455-460.

[16] Wilson Moreno, Gustavo Garzón y Fabio Martínez. "Frame-Level Covariance Descriptor for Action Recognition". En: *Colombian Conference on Computing*. Springer. 2018, págs. 276-290.

entre variables independientes.

Lo anterior nos conduce a la siguiente pregunta de investigación.

PREGUNTA DE INVESTIGACIÓN

¿Cuál es el aporte de analizar localmente áreas de interés utilizando la covarianza?

2. OBJETIVOS

Objetivo general

Proponer un descriptor de covarianza volumétrica local con soporte en trayectorias de movimiento para el reconocimiento de actividades.

Objetivos específicos

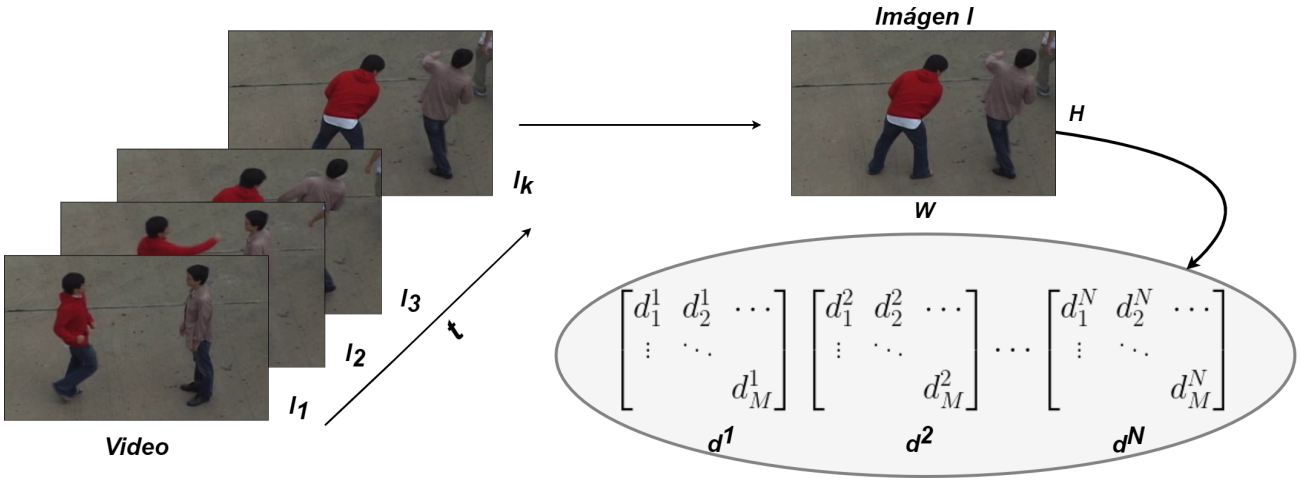
- Seleccionar una base de datos académica con condiciones semicontroladas que permitan desarrollar el proyecto
- Codificar un conjunto de características de movimiento y apariencia para representar las secuencias de video
- Desarrollar un descriptor local de covarianza alrededor de las trayectorias de movimiento.
- Desarrollar una estrategia semisupervisada para agrupar los descriptores volumétricos de covarianza
- Validar el descriptor propuesto en una base de datos académica en cuanto a la clasificación de actividades.

3. MARCO TEÓRICO Y TRABAJOS PREVIOS

3.1. LA COVARIANZA COMO DESCRIPTOR VISUAL

En visión por computador, la covarianza ha sido ampliamente utilizada como un descriptor que correlaciona linealmente diversas fuentes de observación (características o primitivas) y permite el modelamiento de objetos de interés global o regionalmente [17][18]. En el análisis de video, cada uno de los cuadros $I_t(x, y) \in R^{(W \times H)}$ puede ser descrito por un conjunto de características, denotadas por $d^1, d^2, ..., d^N$ (donde todas son de dimensión $W\ X\ H$) Ejemplificado en Figura 1.

Figura 1. Un video es un conjunto de imágenes de dimensión $W \times H$, donde cada imagen será descompuesta en N características $d^{(1)}, d^{(2)}, \cdots, d^{(N)}$ de dimensión $W \times H$. Imágenes tomadas del conjunto de datos *UT-Interaction* [19].



Entonces, dos caracteristicas (d^i, d^j), donde $d^i = [d^i_{l,k}]^{W,H}_{l=1,k=1} = [d^i_1, ..., d^i_M]$, pueden ser relacionadas por su covarianza, como:

---

[17] Bingpeng Ma, Yu Su y Frederic Jurie. "Covariance descriptor based on bio-inspired features for person re-identification and face verification". En: *Image and Vision Computing* 32.6-7 (2014), págs. 379-390.

[18] Mohamed E Hussein y col. "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations". En: *Twenty-Third International Joint Conference on Artificial Intelligence*. 2013.

$$C_{ij} = C_{ij}(d) = \frac{1}{M-1} \sum_{\ell=1}^{M} \left(d_\ell^i - m_i\right)\left(d_\ell^j - m_j\right), \quad 1 \leq i, j \leq N,$$

donde $m_i = \frac{1}{M}\sum_{\ell=1}^{M} d_\ell^i$ es la media de la característica d^i. Por lo tanto, cada cuadro $I_t(x, y)$ se describe por la matriz de covarianza $C_I = [C_{ij}]_{i=1,j=1}^{N,N}$, de dimensión NXN, donde C_I es simétrica y semi-definida positiva($C_{ij} = C_{ji}$). Al ser simétrica esta matriz, solo se necesita su parte triangular superior que se podría ver como un vector de dimensión $\frac{N(N+1)}{2}$. Ahora, si se toma $N << \text{mín}\{W, H\}$ vemos que $\frac{N(N+1)}{2} << W \times H \times N$.

Las caracteristicas resultantes de la matriz de covarianza resultan atractivas para describir propiedades de una imagen o eventualmente de un video con la ventaja de poder compactar y reducir la dimensión de los datos, relacionando de forma eficiente las caracteristicas [1]. Particularmente, para el análisis de video, la secuencia de frames se expresa como $I_1, I_2, \cdots, I_K$, descritas por N características. Así, cada una de las K frames se puede representar por matrices covarianza, como $C_1, C_2, \cdots, C_K$, donde C_i es la matriz de covarianza que le corresponde a la imagen del video I_i para $i \in \{1, 2 \ldots, K\}$.

3.1.1. COVARIANZA INTEGRAL Una limitación de las matrices de covarianza es su descripción global, que en regiones generalmente grandes puede actuar como un filtro pasa bajos. Además, el cálculo, a partir de la expresión original es costosa y redundante para usarla como operador de múltiples regiones (de orden exponencial). Una alternativa eficaz para mitigar este problema, son las covarianzas integrales, que usan representaciones intermedias y precalculadas de las caracteristicas, para la codificación de la covarianza regional [20].

Para obtener la representación integral, inicialmente se puede reescribir la covarianza como la

---

[20] Oncel Tuzel, Fatih Porikli y Peter Meer. "Region covariance: A fast descriptor for detection and classification". En: *European conference on computer vision*. Springer. 2006, págs. 589-600.

expresión:

$$CR(i,j) = \frac{1}{n-1}\left[\sum_{k=1}^{n} z_k(i)z_k(j) - \frac{1}{n}\sum_{k=1}^{n} z_k(i)\sum_{k=1}^{n} z_k(j)\right] \tag{1}$$

Donde:

$$Q = \left[\sum_{k=1}^{n} z_k(i)z_k(j)\right] \tag{2}$$

$$P = \left[\sum_{k=1}^{n} z_k(i)\right] P \Rightarrow \left[\sum_{k=1}^{n} z_k(j)\right] \tag{3}$$

Entonces, cada frame es representado por un conjunto d de caracteristicas con las mismas dimensiones de la imagen. Cada característica d_i se representa como un tensor integral $P \in R^{(W\times H\times d)}$ que representa sumas locales de los pixeles y acumulaciones de estas sumatorias en cada posición, como:

$$P(r_{tl}, i) = \sum_{(x,y)<r_t l} F(x,y,i) \tag{4}$$

Donde, r_{tl} es una posición particular, y por lo tanto la posición $P(r_{tl}, i)$ se acumula la suma de todos los pixeles inferiores a (r_{tl}). Luego, el tensor P es un vector multidimensional d-tamaño que contiene la suma de cada dimensión característica, $P_{x,y} = [P(x,y,1)...P(x,y,d)]^T$(véase en 3) .

Además, la suma del producto de caracteristicas $z_k(i)z_k(j)_{i,j=1...n}$ (véase en 2) se puede expresar con imágenes integrales como un tensor de segundo orden $Q \in R^{W\times H\times d\times d}$.

$$Q(r_{tl}, i) = \sum_{(x,y)<r_{tl}} F(x,y,i)F(x,y,j) \tag{5}$$

con $i, j = 1...d$. El tensor Q es una matriz simetrica de $d \cdot d$ que contiene las representaciones

intermedias de los productos de cualquier par de caracteristicas, expresada como:

$$Q_{x,y} = \begin{pmatrix} Q(x,y,1,1) & ... & Q(x,y,d,1) \\ & . & \\ & . & \\ & . & \\ Q(x,y,d,1) & ... & Q(x,y,d,d) \end{pmatrix} \tag{6}$$

El cálculo de este tensor integral requiere $\frac{d^2+d}{2}$ iteraciones. Una vez obtenida esta representación de las sumas intermedias y locales, codificadas en los tensores P y Q, se puede realizar un cálculo local de la covarianza utilizando simples operaciones aritméticas sobre puntos de interés. Estos puntos de interés, delimitan la región de interés, enmarcada con las esquinas superior izquierda e inferior derecha (ver Figura 2). El cálculo de estas operaciones es $O(d^2)$. Para una representación sencilla, para cada coordenada $(x, y) \Rightarrow r$, donde:

- $r_{tl} \Rightarrow r$ superior izquierdo (top left)
- $r_{tr} \Rightarrow r$ superior derecho (top right)
- $r_{bl} \Rightarrow r$ inferior izquierdo (bottom left)
- $r_{br} \Rightarrow r$ inferior derecho (bottom right)

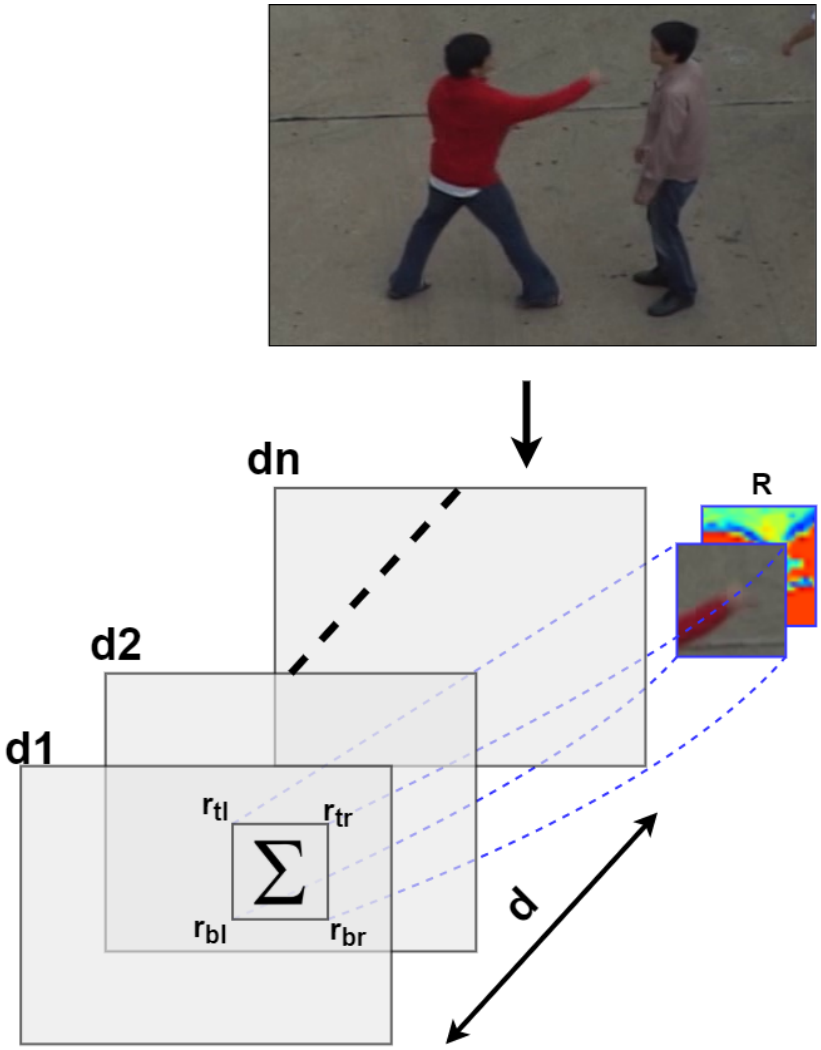
Para obtener una representación regional de la covarianza, a partir de los tensores integrales, se procede simplemente a resolver la siguiente expresión:

$$\begin{aligned} C_R(r_{tl}; r_{br}) = \frac{1}{n-1}[(Q_{r_{br}} + Q_{r_{tl}} - Q_{r_{tr}} - Q_{r_{bl}}) \\ -\frac{1}{n}(P_{r_{br}} + P_{r_{tl}} - P_{r_{tr}} - P_{r_{bl}})(P_{r_{br}} + P_{r_{tl}} - P_{r_{tr}} - P_{r_{bl}})^T] \end{aligned} \tag{7}$$

Donde n es el número de pixeles presentes en la región. Dicha expresión implica cálculos más rápidos para la región especifica en el cuadro completo con pocas operaciones aritméticas, de

esta forma el método integral representa una mejora a nivel computacional sobre el cálculo de la covarianza. Además, esta representación intermedia permite el cálculo de múltiples covarianzas regionales sin un costo computacional adicional.

Figura 2. Esta imagen representa el cálculo local por medio del método integral; cada característica tiene su representación integral.



3.1.2. OPERACIONES CON MATRICES DE COVARIANZA Una vez representado cada una de las imágenes en matrices de covarianza, estas pueden ser operadas para tomar descripciones y métricas generales sobre estas representaciones. Las matrices de covarianza son

simétricas, positivas y están definidas en un espacio curvo, conocido como el espacio de Riemann. Este hecho implica el cálculo de las geodésicas, que no corresponde a medidas en el plano euclidiano directamente y por lo tanto muchos de los métodos tradicionales no son directamente operables con estas matrices. A continuación, se describen las operaciones para proyectar puntos en el espacio-de Riemann a un plano euclidiano tangente.

Espacio de Riemann y espacio Euclidiano Una matriz de covarianza puede ser proyectada a un plano tangente en el espacio euclidiano, utilizando una proyección espectral, como:

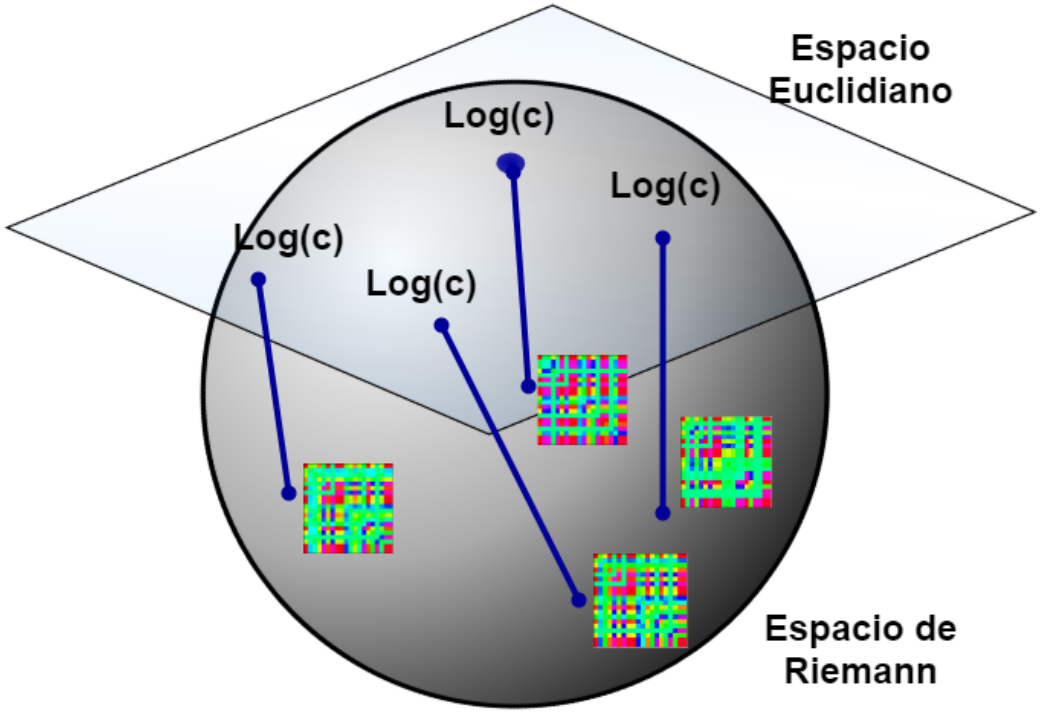
$$log(p) = \Sigma DIAG(log(\lambda_i))\Sigma^T \tag{8}$$

Donde Σ son los auto vectores de la matriz y *lambda* son los autovalores de la misma. Esta proyección es general, y los datos son proyectados con respecto a la matriz identidad. De forma inversa, un punto proyectado en este plano tangente puede ser retornado al espacio de Riemann, usando la operación inversa, definida como:

$$exp(p) = \Sigma DIAG(exp(\lambda_i))\Sigma^T \tag{9}$$

Una ilustración de estas proyecciones es ilustrada en la figura 3. Utilizando estas proyecciones es entonces posibles calcular medidas estadísticas y utilizar algoritmos de aprendizaje de máquina convencionales sobre los puntos de covarianza.

Figura 3. Representación: Proyección las covarianzas, de la variedad al plano euclidiano



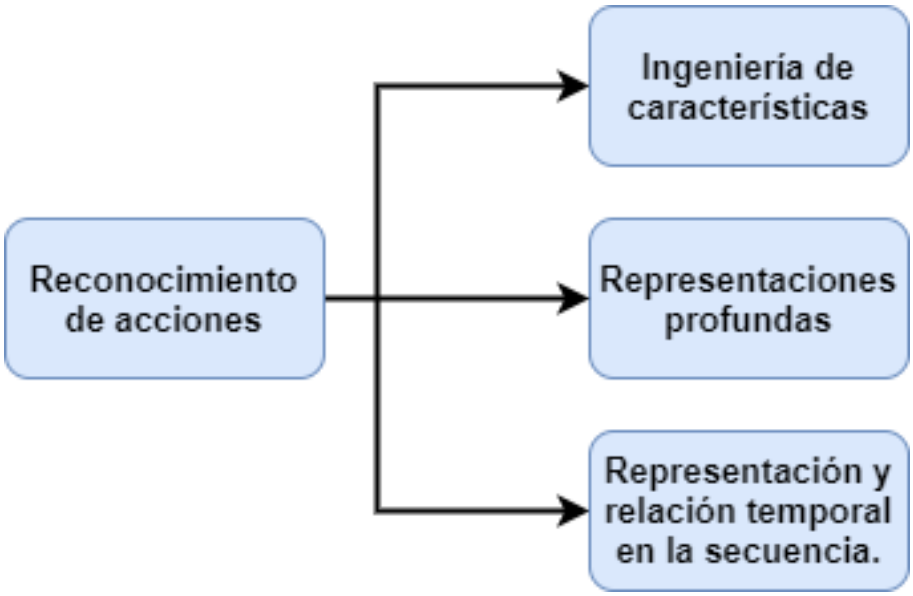
3.2. RECONOCIMIENTO DE ACCIONES EN VIDEO

El reconocimiento de acciones en video consiste en identificar acciones de una serie de observaciones. Este campo ha sido foco de interés de muchos investigadores desde los años ochenta, debido al gran número de aplicaciones para la cual es útil, como la medicina[21] interacción

[21] Akin Avci y col. "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey". En: *23th International conference on architecture of computing systems 2010*. VDE. 2010, págs. 1-10.

hombre-computador[22], vigilancia[23], o incluso sociología[24]. El reconocimiento de acciones es un tema de alta importancia en visión por computador, con múltiples líneas de investigación y focos de trabajo especializados de acuerdo a cada una de las aplicaciones. Teniendo en cuenta el enfoque de este trabajo, a continuación, se exponen los trabajos relacionados en el estado del arte, los cuales son clasificados en cuatro dominios principales, como se ilustra en la Figura 4.

Figura 4. Diagrama vertientes del reconocimiento de acciones humanas.



3.2.1. Ingeniería de caracteristicas La tarea de rastrear y etiquetar las acciones en un video resulta una tarea compleja teniendo en cuenta las múltiples variaciones de iluminancia

---

[22] Siddharth S Rautaray y Anupam Agrawal. "Vision based hand gesture recognition for human computer interaction: a survey". En: *Artificial intelligence review* 43.1 (2015), págs. 1-54.

[23] Sarvesh Vishwakarma y Anupam Agrawal. "A survey on activity recognition and behavior understanding in video surveillance". En: *The Visual Computer* 29.10 (2013), págs. 983-1009.

[24] Claudio Coppola y col. "Social Activity Recognition on Continuous RGB-D Video Sequences". En: *International Journal of Social Robotics* (2019), págs. 1-15.

y apariencia que representan los objetos que desarrollan las acciones [25]. Como línea base para el reconocimiento de acciones, la ingeniería de caracteristicas ha jugado un rol primordial para diseñar descriptores específicos, que acentúan la relevancia en las caracteristicas que mejor representan las acciones. Por ejemplo, los histogramas de flujo óptico han permitido ser invariantes a la apariencia relativa de los video, y enfocarse en las orientaciones principales que definen el movimiento [526]. Normalmente estos métodos resumen una acción según el patrón de velocidad aparente relativa, codificados en un conjunto de bines de orientación. Estos histogramas han sido calculados en diferentes regiones [27], para diferentes intervalos de tiempo [28], y combinados con otros descriptores que complementan la información del video [29]. Una de las limitaciones principales de estos métodos es la dependencia del flujo óptico de base, la limitación solo de describir patrones de velocidad y la variabilidad limitada para representar acciones.

Recientemente se han desarrollado métodos que permiten seguir patrones vectoriales del flujo, los cuales permiten describir trayectorias locales y densas de movimiento[6]. Estas trayectorias agrupan vectores de velocidad en cuadros consecutivos, permitiendo realizar análisis cinemáticos de mayor orden. Por ejemplo, en ocasiones se han utilizado estas trayectorias como base de información, sobre las cuales se calculan histogramas de movimiento y descriptores de apariencia. Estos descriptores pueden ser usados para desarrollar una representación intermedia, como una

---

[25] Debapratim Das Dawn y Soharab Hossain Shaikh. "A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector". En: *The Visual Computer* 32.3 (2016), págs. 289-306.

[26] Li-Jia Li y Li Fei-Fei. "Optimol: automatic online picture collection via incremental model learning". En: *International journal of computer vision* 88.2 (2010), págs. 147-168.

[27] Navneet Dalal y Bill Triggs. "Histograms of oriented gradients for human detection". En: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE. 2005, págs. 886-893.

[28] Moustafa Meshry, Mohamed E Hussein y Marwan Torki. "Linear-time online action detection from 3d skeletal data using bags of gesturelets". En: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2016, págs. 1-9.

[29] Jutta Willamowski y col. "Categorizing nine visual classes using local appearance descriptors". En: *illumination* 17 (2004), pág. 21.

bolsa de palabras, la cual permite describir una secuencia de video en histogramas de ocurrencias [30][31].

3.2.2. Representaciones Profundas Los algoritmos de aprendizaje profundo son hoy en día el estado del arte para la representación y análisis de imágenes y video. En cuanto al reconocimiento de acciones se ha logrado evolucionar en arquitecturas que permiten el análisis a nivel volumétrico de los videos, el desarrollo de representaciones de movimiento y la codificación de cuadros para su posterior análisis temporal [9][10]. Particularmente las representaciones basadas en caracteristicas profundas representan cada cuadro como un conjunto de activaciones de arquitecturas conocidas y pre-entrenadas en otros dominios. Por ejemplo, en [4] se establece un método basado en la idea de que una acción en un video puede ser descrita a partir de ciertas caracteristicas extraídas en secciones cortas de las secuencias de imágenes y que, además, en dicho video existe al menos una sección que cumpla con dicha función. Una vez representado cada segmento, se propone un agrupamiento discriminativo en el video utilizando hiperplanos no lineales para separar caracteristicas consideradas discriminatorias. Luego, una máquina de soporte vectorial permite la separación e indexación de las acciones aprendidas a partir de estas caracteristicas, del mismo modo se han trabajado enfoques similares al agrupamiento discriminativo ya mencionado, como por ejemplo el agrupamiento por rangos[32], que propone un esquema de extracción de caracteristicas novedoso, el cual permite mantener un factor de organización por medio de una característica aprendida a lo largo del tiempo con funciones de valor vectorial, la cual referencian como *VideoDarwin*, haciendo alusión a una característica que evoluciona, como resultado

---

[30] Gustavo Garzón y Fabio Martínez. "Online Action Recognition from Trajectory Occurrence Binary Patterns (ToBPs)". En: *The International Conference on Advances in Emerging Trends and Technologies*. Springer. 2019, págs. 409-418.

[31] Heng Wang y col. "Action recognition by dense trajectories". En: 2011.

[32] Basura Fernando y col. "Modeling video evolution for action recognition". En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, págs. 5378-5387.

final, *VideoDarwin* aprende a lo largo del video, los cambios e información discriminativa a lo largo del video. Por otro lado, las imágenes dinámicas [33] representan un método que resume la complejidad de una acción en una sola imagen 2D, producto de una función de rangos aplicada a los cuadros del video, la cual, durante el entrenamiento de la red, aprende la información temporal del video, a partir de la integración y el reordenamiento de la intensidad de los pixeles según su rango.

Recientes arquitecturas convolucionales 3D analizan volúmenes completos de videos y permiten un mayor análisis en patrones espacio-temporales. Por ejemplo, la arquitectura LTC (Long-term temporal convolutions)[3] establece un conjunto de capas convolucionales en 3D que operan sobre secuencias completas. Estas operaciones son jerárquicamente operadas para obtener patrones de mayor nivel que encuentra relaciones espacio-temporales, al final son obtenidas caracteristicas volumétricas de movimiento para la descripción de la acción. En [34] se utilizan arquitecturas que aprovechan videos que incluyen profundidad y múltiples vistas para el procesamiento y reconocimiento de acciones sobre secuencias de video simples, con las *Two-stream 3D convNet*[35] se introduce un método el cual funciona independientemente del tamaño de la secuencia, utilizando en su arquitectura una capa de fusión que codifica información espacial y de movimiento con la que es entrenada la red. En Estos enfoques, sin embargo, requieren de etapas de configuración sofisticadas y un tratamiento complejo de los datos, lo cual es limitado para el uso en escenarios cotidianos. También, en *Ji Shuiwang et al.* proponen una red convolucional 3D, aplicando kernels 3D en cuadros de video sobre el eje del tiempo en la búsqueda de la captura de la información espacial y temporal del video, este acercamiento termina demostrando que

---

[33] Hakan Bilen y col. "Dynamic image networks for action recognition". En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, págs. 3034-3042.

[34] Du Tran y col. "Learning spatiotemporal features with 3d convolutional networks". En: *Proceedings of the IEEE international conference on computer vision*. 2015, págs. 4489-4497.

[35] Xuanhan Wang y col. "Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length". En: *IEEE Transactions on Multimedia* 20.3 (2017), págs. 634-644.

puede capturar movimiento e información de flujo óptico, pues los cuadros en el video, están conectados al final, con capas completamente conectadas de la red [36].

3.2.3. Representación y relación temporal del video También desde la perspectiva de aprendizaje profundo se han propuesto arquitecturas que buscan el modelamiento temporal de caracteristicas capturadas en cada uno de los cuadros que representan un video. Estas arquitecturas tienen la ventaja, a diferencia de las cadenas clásicas de Markov, de operar sobre historias más largas y tener funciones de olvido y correlación no lineal. Por ejemplo, con la arquitectura LSTM, en [37] se desglosa el fundamento y funcionamiento de una red neuronal recurrente, la cual mantiene la relación temporal a periodos cortos, de ahí su nombre *Long short term memory*, básicamente en este trabajo se utiliza en la red una celda de memoria que controla el flujo de información dentro y fuera de la red, entonces la información se retiene un tiempo determinado, según esta celda lo indique, en otras palabras, las LSTM permiten a las redes neuronales recordar sus entradas por un periodo determinado. También, *Simoyan et al* en "Two-stream convolutional networks for action recognition in videos"[11],utilizan dos fuentes principales de información: apariencia y flujo óptico. Cada una de estas fuentes de información son representadas por arquitecturas convolucionales que sirven de entrada a una arquitectura recurrente. Esta representación permite aprovechar las dos fuentes de información, lo cual resulta robusto para representar videos no controlados, donde las caracteristicas visuales como el movimiento pueden aportar en la representación de acciones.

---

[36] Shuiwang Ji y col. "3D convolutional neural networks for human action recognition". En: *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012), págs. 221-231.

[37] Ralf C Staudemeyer y Eric Rothstein Morris. "Understanding LSTM–a tutorial into Long Short-Term Memory Recurrent Neural Networks". En: *arXiv preprint arXiv:1909.09586* (2019).

4. PLANTEAMIENTO Y JUSTIFICACIÓN DEL PROBLEMA

El área de visión por computador, más concretamente el reconocimiento de acciones ha sido abordado a lo largo de los años de numerosas formas y se han propuesto diversos métodos para solventar esta tarea de manera efectiva. Los métodos comunes para llevar a cabo esta tarea resultan en descriptores de alta dimensionalidad, lo cual es contraproducente computacionalmente. Los descriptores basados en la covarianza pueden ser una alternativa compacta para la representación de acciones mostrando ventajas en cuanto a costo computacional y eficiencia algorítmica. En la literatura encontramos métodos basados en covarianza con diferentes enfoques, desde acercamientos globales donde la covarianza describe el total de la acción a lo largo de todo el video; hasta regionales, donde se centra en áreas específicas semi-globales de la acción plasmada en la secuencia de video. Estas perspectivas podrían sugerir una pérdida de información en ciertas secciones de la imagen, o, por el contrario, puede representar el agregar información irrelevante para la acción, como lo es el fondo o interacciones humanas fuera de la acción principal.

Mientras que los métodos tradicionales son computacionalmente costosos, el procesamiento y computo de la covarianza representa una ventaja en este aspecto, sin una perdida significativa de información si se realiza un correcto seguimiento de localidad a los movimientos, esto mismo, debido a sus propiedades únicas de representación de relaciones entre variables independientes. Además, el reconocimiento de acciones 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 determinar automáticamente las secuencias de video. Por ejemplo, en áreas como la video vigilancia, realidad aumentada, análisis deportivo y más.

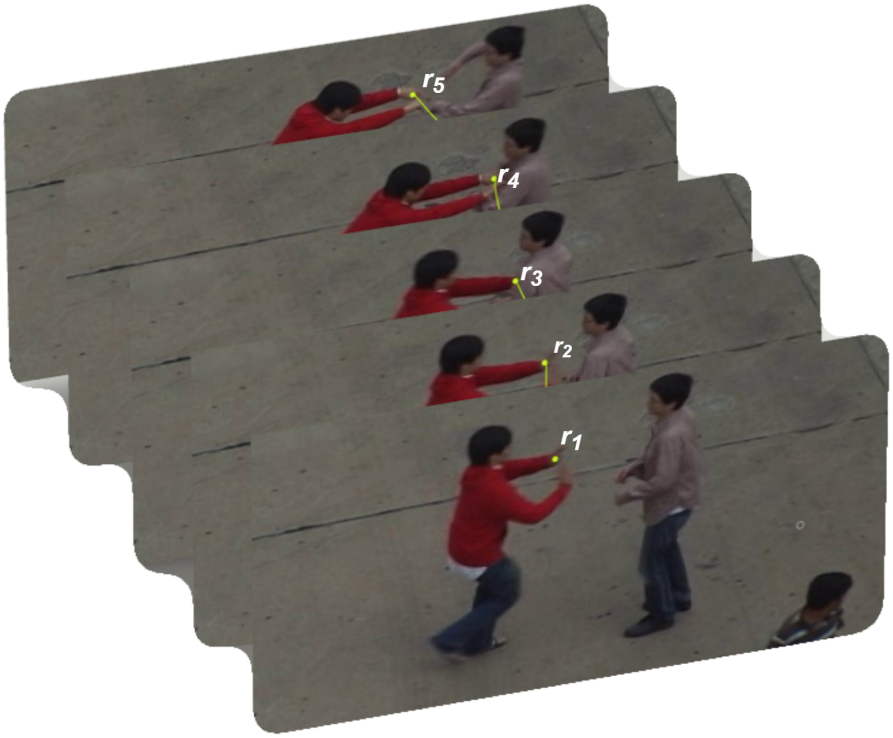
5. MÉTODO PROPUESTO

El presente trabajo establece un método de reconocimiento de acciones que representa un video como un conjunto de covarianzas volumétricas que operan localmente a lo largo del video, siguiendo trayectorias con movimiento de interés. En términos generales, el video es representado por un conjunto de caracteristicas profundas obtenidas de una red convolucional pre-entrenada. Además, para la secuencia particular del video, se calculan trayectorias de movimiento que sirven como soporte y referencia para delinear bloques locales en el espacio de características profundas. La trayectoria se describe a través del tiempo, permitiendo calcular regiones espaciales a lo largo del video, obteniendo así una información volumétrica localizada. Entonces, se propuso una covarianza integral volumétrica que permite representar patrones de video localizados $2D + t$. Una vez es representado el video por un conjunto de N covarianzas volumétricas locales, estas son mapeadas a un algoritmo de agrupamiento para obtener una representación compacta de K covarianzas, siendo $K << N$. El descriptor de video es entonces usado como entrada en un algoritmo de clasificación, como lo puede ser un clasificador Gaussiano, un Random Forest o una máquina de soporte vectorial.

5.1. COVARIANZA VOLUMETRICA INTEGRAL

En este trabajo se extiende el concepto de covarianza integral, para ser operado sobre volúmenes locales de información. Esta representación permite obtener una relación espacio-temporal de los patrones locales relacionados con una acción y una descripción a través de múltiples caracteristicas profundas. Supongamos, que de una secuencia de video se pueden seleccionar puntos relevantes y seguirlos a través del tiempo $\{T_i\}_{i=1...M}$, los cuales describen en su mayoría la acción. Entonces, una trayectoria $T_i = \{(r_1), (r_2) \dots (r_{fr})\}$, q concatena posiciones espaciales (r_i) a través de un conjunto de f cuadros, como en la figura 5.

Figura 5. Representación de una trayectoria densa de movimiento a nivel de cuadro.



Entonces, podemos describir un patrón local de covarianza alrededor de la trayectoria, siendo las posiciones de las trayectorias el centro de cada región en cada cuadro. Entonces el conjunto de cuadros que enmarcan la trayectoria, constituye una región volumétrica de interés, la cual puede ser descrita por una covarianza. Para ello, se precalculan los tensores integrales Q_{fr}, P_{fr} en cada cuadro, y la covarianza volumétrica entonces esta descrita por la suma de tensores integrales que definen el patrón espacio-temporal, como:

$$C_{tr} = \frac{1}{(n \cdot m) - 1}(Q_{fr1} + Q_{fr2} + ... + Q_{frm)} \\ -\frac{1}{n * m}(P_{fr1} + P_{fr2} + ... + P_{frm})(P_{fr1} + P_{fr2} + ... + P_{frm})^T \tag{10}$$

Donde:

- n = Numero de pixeles en el recuadro.

- m = Número de puntos rastreados en la trayectoria.
- fr = frame o cuadro en la trayectoria.

Entonces, para una región local específica, centrada en la posición $(r_{fr}) \in T_i$ y definida por sus posiciones extremas, se operan sobre las posiciones específicas en cada Q_{fr_i} y para P_{fr_i}, como:

$$\begin{aligned} Q_{fr_i} &= Q_{r_{br}(i)} + Q_{r_{tl}(i)} - Q_{r_{tr}(i)} - Q_{r_{bl}(i)} \\ P_{fr_i} &= P_{r_{br}(i)} + P_{r_{tl}(i)} - P_{r_{tr}(i)} - P_{r_{bl}(i)} \end{aligned} \tag{11}$$

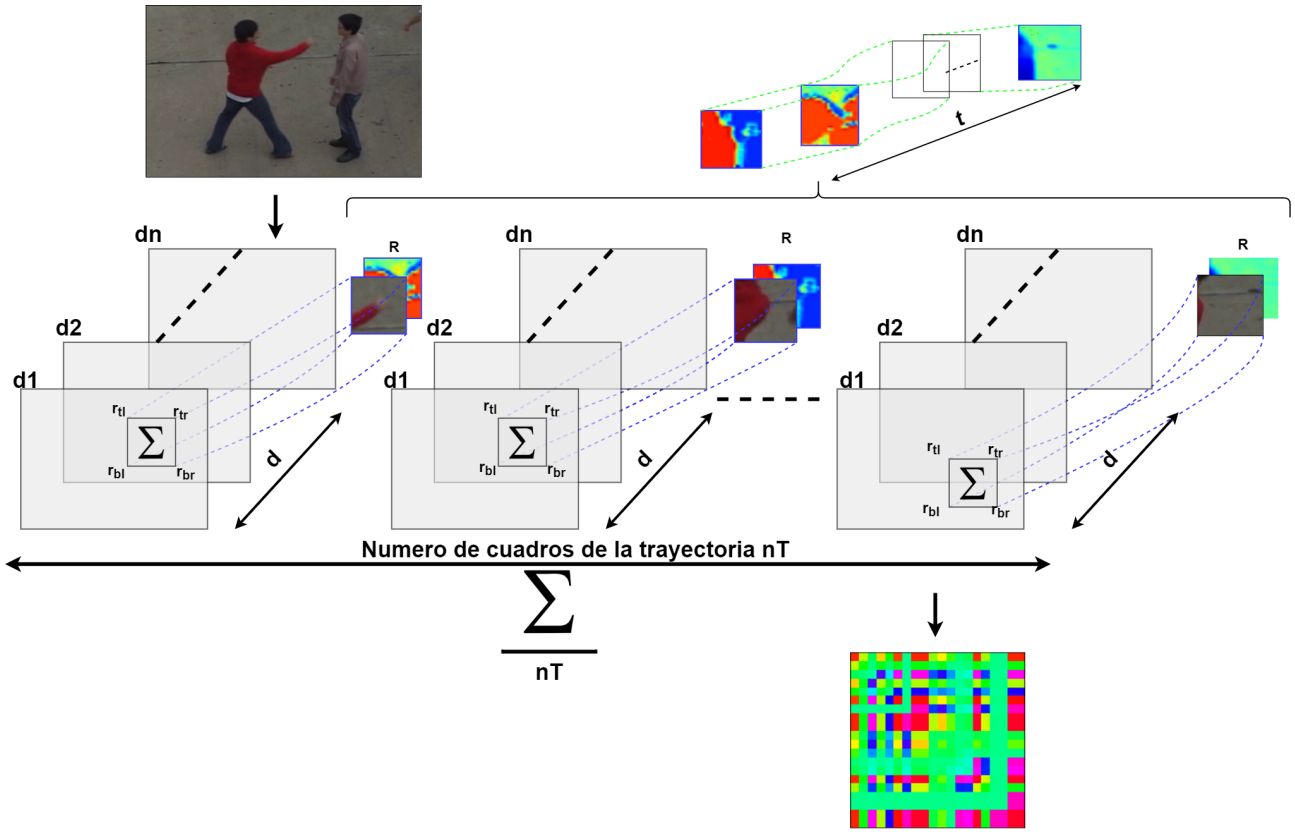
Resultando de forma homóloga a la representación de la imagen integral, pero con la ventaja de responder a la necesidad de capturar patrones en un segmento volumétrico que se desarrolla, alrededor de cada cuadro. De tal forma, para la ecuación 10 obtenemos un análisis volumétrico de la covarianza, con soporte en trayectorias densas de movimiento (como se muestra en la figura 6), en la cual se toma en cuenta la profundidad y el área rastreadas en la secuencia.

Entonces, a partir de la ecuación 10 obtenemos la siguiente formula general, para el cálculo de la covarianza sobre una trayectoria, utilizando los tensores correspondientes Q y P de los frames en los que pasa la trayectoria(fr_i):

$$C_{tr} = \frac{1}{(n \cdot m) - 1} \left[\sum_{i=1}^{m} Q_{fr_i} - \frac{1}{n \cdot m} (\sum_{i=1}^{m} P_{fr_i})(\sum_{i=1}^{m} P_{fr_i})^T \right] \tag{12}$$

Finalmente obtenemos una covarianza única por trayectoria existente en el video, cabe resaltar que cada trayectoria que no cumpla con los requisitos mínimos en cuanto al área (en caso que esta se exceda en las dimensiones del video) es excluida y descartada del cálculo.

Figura 6. Representación del cálculo de la covarianza a nivel volumétrico nT como el número de cuadros en profundidad

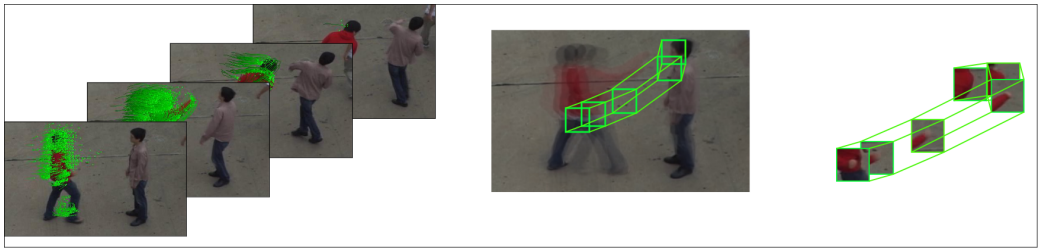


5.2. TRAYECTORIAS DENSAS DE MOVIMIENTO

Uno de los puntos fundamentales en este trabajo es el cálculo de patrones de covarianza volumétrica local. Sin embargo, un enfoque denso sobre todas las secuencias de video puede resultar en un modelamiento excesivo del fondo, perdiendo relevancia el objeto de interés o el patrón más correlacionado con la acción. Por lo tanto, es necesario encontrar puntos de interés que sigan patrones relevantes de movimiento y puedan agrupar localmente regiones de interés altamente correlacionadas con las acciones. En este trabajo se seleccionaron trayectorias densas de movimiento, que resultan del seguimiento de patrones de flujo aparente, que concatena un conjunto de puntos de interés en el tiempo, como $(x_{t_0}, y_{t_0}), (x_{t_1}, y_{t_1}), ..., (x_{t_{nT}}, y_{t_{nT}})$. Para este trabajo se utilizaron las trayectorias densas de movimiento propuestas por *Wang et al.* en [6].

Estas trayectorias rastrean la historia de un punto de interés sobre cierto número de cuadros y entrega información de localidad a lo largo de una acción de interés. Como se ilustra en la figura 7, esta representación local permite calcular y describir volúmenes de interés centrados en la trayectoria.

Figura 7. Representación: Trayectoria densa de movimiento



En el presente trabajo, se implementaron un conjunto de trayectorias densas que soportan la base local de representación de la covarianza volumétrica. Estas trayectorias son primitivas locales de movimiento basadas en el seguimiento de puntos destacados a lo largo de la secuencia de video.

El cálculo de trayectorias comienza calculando el flujo denso de Farneback, que representa el desplazamiento local de pixeles como una expansión polinomial [38]. En este método el vecindario de cada pixel de la imagen se aproxima mediante una expresión polinomial de orden cuadrática, definida como: $f(x) \sim x^T Ax + B^T + c$, donde $x = (x, y)^T$ son pixeles, A son coeficientes desconocidos de representación con un bias b. Luego, el desplazamiento del campo d es obtenido como la diferencia de los coeficientes cuadráticos en cuadros consecutivos, como $d = -\frac{1}{2}A_1^{-1}(b_2 - b_1)$. Este algoritmo presenta buena relación entre velocidad de cómputo y precisión en cuanto a la representación de velocidad aparente.

Las trayectorias entonces resultan de seguir vectores de los campos vectoriales en cuadros consecutivos. Para esto, los puntos de caracteristicas $P_t = (x_t, y_t)$ son muestreados en una gri-

[38] Gunnar Farnebäck. "Two-frame motion estimation based on polynomial expansion". En: *Scandinavian conference on Image analysis*. Springer. 2003, págs. 363-370.

lla y rastreados a través del tiempo en diferentes escalas espaciales. El rastreo de filtrado de media con respecto al campo de flujo óptico denso $w = (u_t, v_t)$, puede ser expresado como: $P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * w)|_{\bar{x}_t, \bar{y}_t}$, donde M es el kernel de filtro de medias, y $(\bar{x}_t, \bar{y}_t)$ es la posición redondeada de (x_t, y_t). Los puntos consecuentes son concatenados para formar una trayectoria $(P_t, P_{t+1}, P_{t+2}, ..., P_{t+n})$. Algunas trayectorias con fuertes desviaciones en su desplazamiento son descartadas y removidas de su análisis.

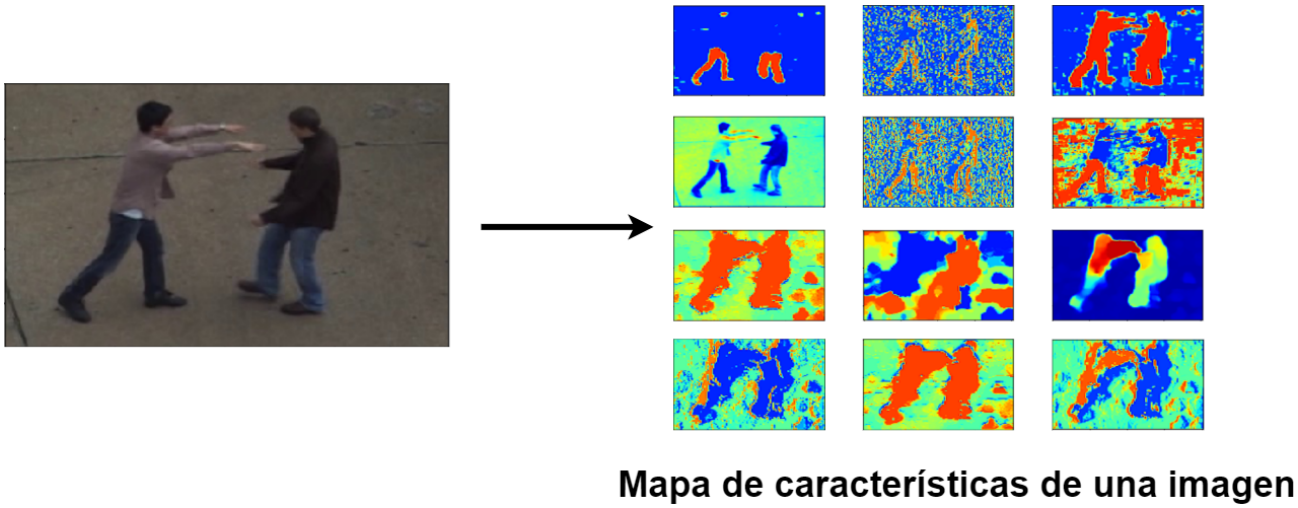
5.3. MAPAS DE CARACTERISTICAS DENSAS

En este trabajo es fundamental la representación de las acciones en cada cuadro. Estas caracteristicas pueden provenir desde la ingeniería de caracteristicas, pero limitándose a descripciones específicas para escenarios particulares. Hoy en día, las caracteristicas profundas, de arquitecturas convolucionales pre-entrenadas resultan una alternativa robusta que permite recopilar representaciones generales de diferentes tipos de acciones. Uno de los objetivos de este trabajo entonces fue lograr la caracterización de las acciones utilizando información extraída del primer bloque de una arquitectura convolucional pre-entenada. Estos productos convolucionales permiten una representación densa de la información, donde cada cuadro se codifica en diferentes bandas de respuesta, con la principal ventaja que los filtros obedecen a respuestas no lineales [39]. Este método de extracción de características es replicable a cualquier arquitectura convolucional, y en el presente se realizan experimentos con la información extraída de las redes: InceptionV3 y MobileNet.

---

[39] Bing Xu y col. "Empirical evaluation of rectified activations in convolutional network". En: *arXiv preprint arXiv:1505.00853* (2015).

Figura 8. Una imagen es representada por K matrices de $W \times H$, donde cada una, es una característica d, esto lo llamamos *Mapa de características*



5.3.1. Arquitectura convolucional InceptionV3 Esta arquitectura convolucional desarrollada por un equipo de Google ha resultado uno de los modelos más exitosos en el desafío más importante de clasificación de imágenes naturales: ImageNet[40]. El conjunto de datos de ImageNet es de aproximadamente un millón de imágenes con más de mil diferentes clases, lo cual resulta interesante en las metodologías de aprendizaje profundo. En el desafío del año 2016 *ImageNet* en uno de sus sub segmentos, este modelo resulto favorecido con una exactitud superior a 78.1 % y una tasa de error de 23.4 % en la tarea de clasificación.

Esta arquitectura Inception involucra la metodología de convoluciones $1 \times 1 \times N$ que permite ser eficientes en el cálculo computacional, disminuir el número de parámetros para ser entrenados y además aprender correlaciones no-lineales entre las activaciones de capas convolucionales. Además, esta arquitectura está formada por bloques de construcción simétricos y asimétricos, con etapas clásicas convolucionales, de *pooling* y módulos Inception que relacionan activaciones. En este trabajo, cada frame entonces fue representado por 32 canales, producto de la .ªctivación

[40] Christian Szegedy y col. "Rethinking the inception architecture for computer vision". En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, págs. 2818-2826.

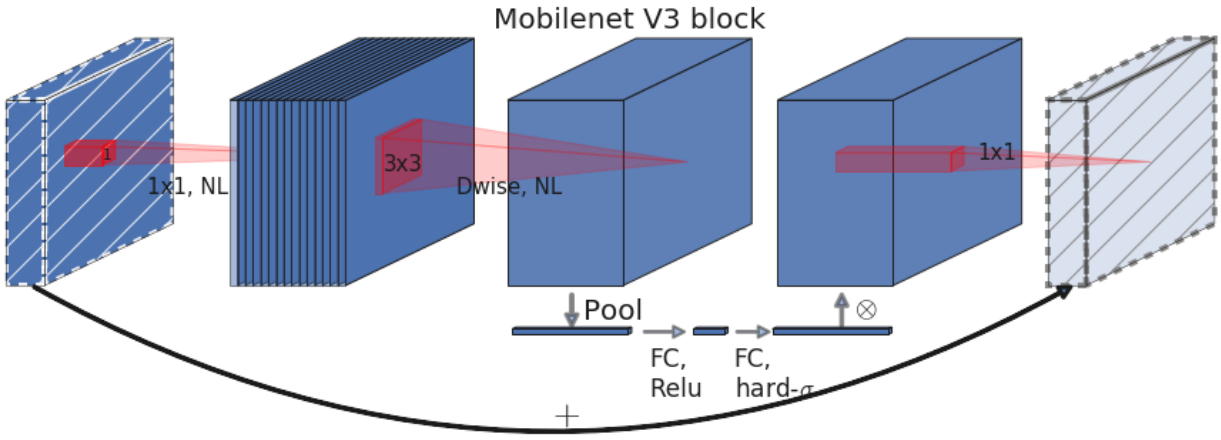
1"del primer bloque convolucional, que resulta después de la función no lineal de tipo RELU". Las covarianzas locales convolucionales son entonces operadas sobre estas activaciones, permitiendo una representación robusta y local de las acciones.

5.3.2. Arquitectura convolucional MobileNet Esta arquitectura convolucional resulta interesante en nuestro trabajo, porque fue diseñada para operar en condiciones de hardware limitados y con operaciones computacionales eficientes. Estas caracteristicas resultan ideales para desarrollar sistemas de reconocimiento de acciones en aplicaciones de tiempo real, que puede ser una aplicación del trabajo desarrollado.

Esta arquitectura fue presentada en el 2018 con el fin de generar un rendimiento sobresaliente en dispositivos móviles, por lo que se infiere que es más compacta, y consta de un número reducido de parámetros, pero suficientes capas profundas [41]. Esta arquitectura está basada en una estructura residual, donde las entradas y las salidas del bloque residual son pequeñas capas cuello de botella (convoluciones 1×1). MobileNet utiliza convoluciones livianas y profundas para filtrar características durante la capa de expansión; Se demostró que es importante remover las linealidades en las capas más estrechas para mantener el poder de representación, esto mejora el rendimiento de la red. Las caracteristicas extraídas de esta red corresponden al primer bloque de la arquitectura con 32 canales, resultado de una función no lineal de tipo RELU".

[41] Mark Sandler y col. "Mobilenetv2: Inverted residuals and linear bottlenecks". En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, págs. 4510-4520.

Figura 9. Visión general de la arquitectura MobilnetV3 que cuenta con un bloque de excitación que reemplaza la función sigmoide clásica con una aproximación lineal por partes, además de la introducción de funciones de activación "hard-Swish"no lineales [42]



6. DISEÑO EXPERIMENTAL

6.1. DATOS: UT-INTERACTION

El método propuesto fue evaluado en el conjunto de datos públicos *UT-Interation(High-level Human Interaction Recognition Challenge)* que exhibe actividades humanas complejas en escenarios simulados de vigilancia remota[43]. Este conjunto de datos contiene secuencias de videos para 6 clases de interacciones humanas, las cuales son: Dar la mano (*Shake Hands*), señalar (*Point*), Abrazar (*Hug*), Empujar (*Push*), Patear (*Kick*) y Golpear con la mano (*Punch*). Los videos cuentan con una resolución de 720x480 pixeles con un ratio de 30 cuadros por segundo. El conjunto de datos se divide en dos grupos balanceados, con 60 videos cada uno. El primer segmento de videos fue capturado en condiciones estáticas, como fondo fijo y poco movimiento de la cámara; por otro lado, el segundo segmento de videos presenta unas condiciones más inestables, con fondo en movimiento, movimientos de cámara, y del mismo modo se reportan movimientos humanos en el fondo de la toma, que no hacen parte de la acción principal. Como ejemplo de las acciones de las diferentes actividades del conjunto de datos UT tenemos la figura 10.

---

[43] Michael S Ryoo y JK Aggarwal. "UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA)". En: *IEEE International Conference on Pattern Recognition Workshops*. Vol. 2. 2010, pág. 4.

Figura 10. Clases de actividades humanas del conjunto de datos UT-Interaction capturadas para el primer grupo de datos. En total son 6 diferentes actividades para todo el dataset.

6.2. MÉTRICAS DE EVALUACIÓN

El método propuesto fue validado siguiendo un conjunto de métricas clásicas. Asumiendo, los verdaderos positivos (VP), los falsos positivos (FP), los verdaderos negativos (VN), y falsos negativos (FN), las métricas se pueden definir, como:

- Exactitud: Es la relación entre las predicciones correctas y el número total de predicciones, definida como: $\frac{VP+VN}{(VP+FP+VN+FP)}$
- Precisión: Evalúa los verdaderos positivos con respecto al total de predicciones obtenidos en la misma clase, definida como: $\frac{VP}{(VP+FP)}$
- Sensibilidad: Es la razón de predicciones correctas sobre el total de clases correcta e incorrectamente clasificadas: $\frac{VP}{(VP+FN)}$
- Especificidad: Es la tasa de verdaderos negativos, definidos como: $\frac{VN}{(VN+FP)}$

- Matriz de confusión: permite representar el desempeño global de la tarea de clasificación, donde cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias de la clase real.

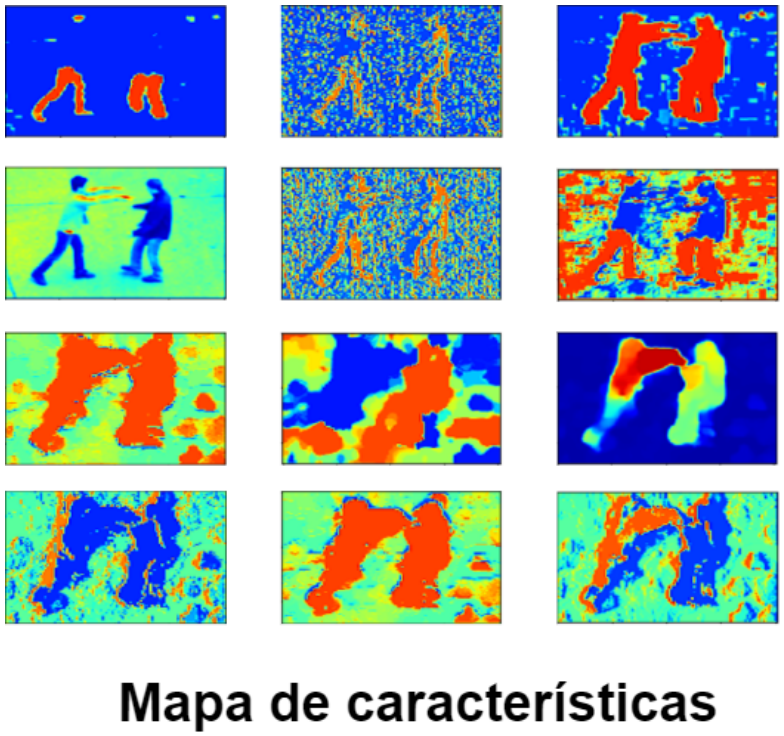
6.3. CONFIGURACIÓN DE PARÁMETROS DEL MÉTODO

Cada video se representa por un conjunto de M covarianzas volumétricas. Cada secuencia puede además tener diferentes números de covarianzas. Entonces se utilizó un algoritmo de clustering para obtener una representación compacta con un conjunto de K covarianzas representativas. En este trabajo se implementó como algoritmo de clustering el *K-Means* y el *Aglomerative Clustering*. En la figura 11 se ilustra la asociación de trayectorias según el agrupamiento de las covarianzas. Como se puede observar, las covarianzas están espacialmente localizadas y además permite agrupar patrones localizados del conjunto de caracteristicas profundas.

Figura 11. Aglomerative clustering con un *K=10* donde se aprecian los agrupamientos en colores diferentes sobre una secuencia.

Para la validación del método propuesto, se utilizó la estrategia de *k-fold cross validation* con $K = 10$. Para el cálculo de caracteristicas se hicieron experimentos independientes con la arquitectura Inception y la MobileNet, con un total de 20 activaciones para el segmento 1 y con un total de 18 activaciones para el segmento 2. En la figura 12 se muestra un ejemplo del cálculo de las caracteristicas generales de las redes neuronales para un conjunto de secuencias del conjunto de datos UT-Interaction.

Figura 12. Ejemplo de mapa de caracteristicas en un segmento de UT, con activaciones aleatorias de la red en una imagen para su visualización.



6.4. CLASIFICACIÓN

Para la clasificación de las acciones se recurrió a un método de clasificación clásico de *Random-Forest*, el cual demostró aprender las caracteristicas únicas de las covarianzas por acción. En la estrategia de Random Forest, un conjunto de algoritmos independientes de árbol de decisión llamados *DT*, se entrenan sobre diferentes partes del espacio de caracteristicas, para reducir la variabilidad de la predicción. Para ello, se implementa una estrategia de agregación bootsrap, que consiste en seleccionar aleatoriamente un conjunto de caracteristicas de covarianza para construir un DT particular. Este método de clasificación tiene la ventaja de la interpretabilidad de los resultados haciendo una propagación de la etiqueta a lo largo de los árboles.

7. RESULTADOS

El método propuesto permite la representación de patrones espacio-temporales locales, los cuales siguen puntos de interés durante la secuencia. En este trabajo se evaluó el método en la tarea de clasificación sobre un conjunto de videos que simulan acciones para el contexto de video-vigilancia. En términos generales el método propuesto logró una exactitud del 83.3 %, siendo un resultado de exactitud sobresaliente y competitivo con respecto al estado del arte. En la tabla 1 se ilustra la matriz de confusión obtenida para el primer conjunto de datos de UT, donde la cámara está relativamente estática y el fondo es constante. Este resultado fue logrado con un conjunto de $K = 300$ covarianzas por video, sobre la arquitectura MobileNet y utilizando un clasificador de RandomForest. Como se puede detallar, el método propuesto alcanza clasificaciones perfectas para cuatro clases y su mayor aporte se reporta en *dar la mano*. Este hecho puede ser atribuido a que esta acción es ampliamente estática, y no se capturan suficientes trayectorias, además que el gesto o postura se confunde fácilmente con abrazar.

Tabla 1. Matriz de confusión obtenida para el conjunto de datos de UT-Interaction número 1 al evaluar el descriptor propuesto. Los resultados están dados en porcentajes

Categoría	DM	AB	PA	AP	GP	EP
Dar la mano	100	0	0	0	0	0
Abrazar	30	50	0	0	0	20
Patear	0	0	66	0	33	0
Apuntar	0	0	0	100	0	0
Golpear	0	0	0	0	100	0
Empujar	0	0	0	0	0	100

En la tabla 2 se muestran las métricas adicionales que permitieron evaluar el primer conjunto de datos sobre el método propuesto. Como se ilustra, las diferentes métricas evidencian un comportamiento estable y sobresaliente para el método propuesto logrando una sensibilidad media de 86,1 % y una especificidad media de 91,3 %.

Tabla 2. Índices de precisión, sensibilidad y especificidad por clase para el segmento 1, los índices estan en %

Métrica	Precisión	Sensibilidad	Especificidad
Dar la mano	76.9	100	93.27
Abrazar	100	50	66.6
Patear	100	66.6	100
Apuntar	100	100	100
Golpear	75.1	100	92.6
Empujar	83.3	100	95.4

En una segunda evaluación, el método propuesto fue evaluado sobre un subconjunto de UT, donde el método es dinámico, con acciones desarrollándose de fondo, pero además la cámara presenta pequeños movimientos. En la tabla 3 se resume los resultados obtenidos por el método propuesto y codificado en una matriz de confusión. En este conjunto de datos, la mayoría de acciones son apropiadamente clasificadas, con errores entre golpear y empujar.

Tabla 3. Matriz de confusión obtenida para el conjunto de datos de UT-Interaction número 2 al evaluar el descriptor propuesto

Categoria	DM	AB	PA	AP	GP	EP
Dar la mano	100	0	0	0	0	0
Abrazar	0	100	0	0	0	0
Patear	0	0	100	0	0	0
Apuntar	0	0	0	100	0	0
Golpear	0	0	50	0	50	0
Empujar	0	20	0	0	30	50

Para este segundo conjunto de datos, el método fue validado con métricas adicionales como la precisión, sensibilidad y especificidad. Se logró una sensibilidad media de 83,3 %, y una especificidad media de 96,4 %, estas métricas son resumidas en la tabla 4.

Tabla 4. Índices de precisión, sensibilidad y especificidad por clase para el segmento 2, los índices están dados en porcentajes

Métrica	Precisión	Sensibilidad	Especificidad
Dar la mano	100	100	100
Abrazar	100	100	100
Patear	66.6	100	88.8
Apuntar	100	100	100
Golpear	50	50	90
Empujar	100	50	100

El método propuesto fue además comparado con estrategias del estado del arte que fueron validadas sobre el mismo dataset (ver tabla 5). El método propuesto se muestra competitivo con respecto a los demás enfoques propuestos por el estado del arte, siendo compacto en su descripción, flexible para diferentes aplicaciones, y con una complejidad relativamente baja teniendo en cuenta el conjunto de K covarianzas seleccionadas por video, y la descripción compacta de las covarianzas.

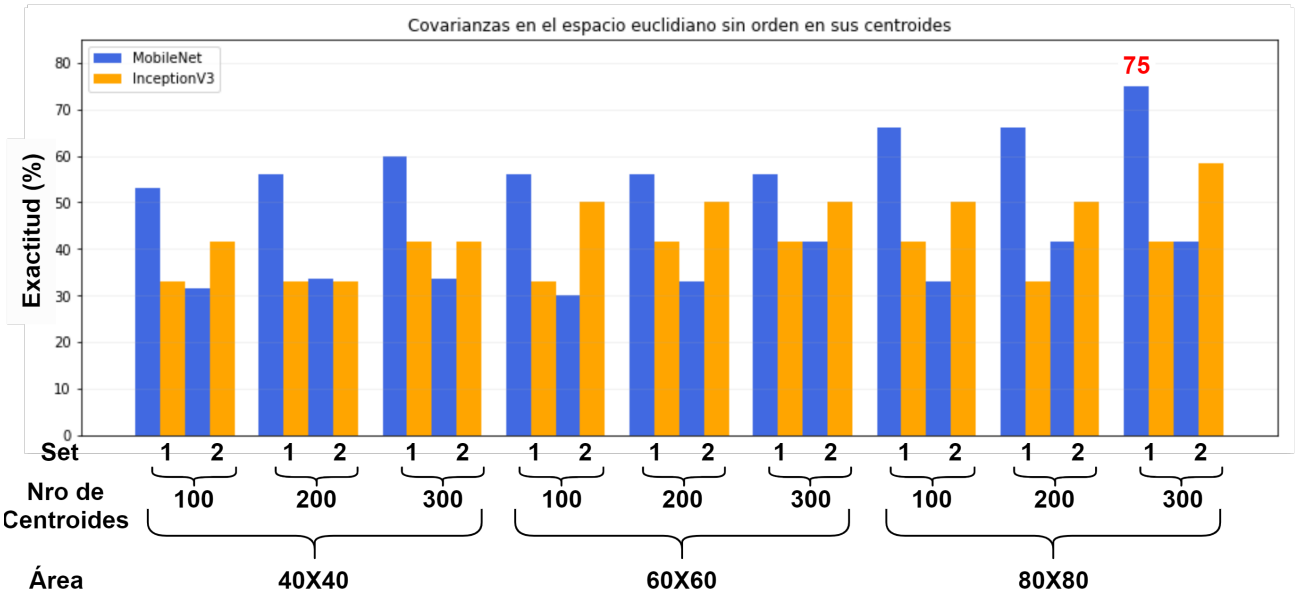
Tabla 5. Precisión promedio para diferentes estrategias informadas en el estado del arte.

Enfoques	Precisión UT Conjunto 1	Precisión UT conjunto 2
Xiaofei [44]	83.3 %	
Enfoque propuesto	**83.3 %**	
Mukherjee [45]	79.17 %	
Ryoo [46]	71.7 %	
Silmani [47]	41 %	
Votacion propagativa [48]	93 %	91 %
Moreno [49]	80 %	61.66 %
Daysy [50]	71 %	51 %
SIFT 3D [51]	63 %	55 %

Cabe destacar que en *Xiaofei et al.* se integran histogramas de ocurrencias de bolsa de palabras(BoW) con histogramas de gradientes orientados(HoG), representando un alto tiempo de cálculo para obtener una representación de acción. Además este enfoque no es escalable para acciones más complejas sobre datasets con mayor variabilidad. A continuación, se presentan experimentos detallados sobre cada uno de los componentes del método propuesto.

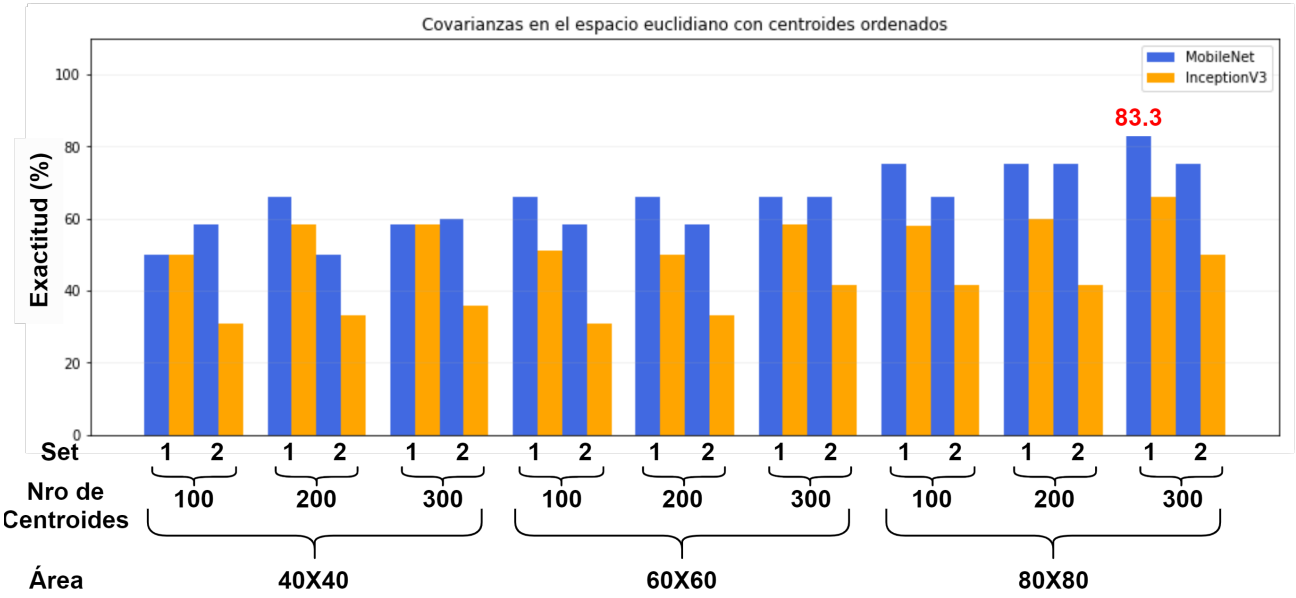
7.0.1. Covarianzas en el espacio Euclidiano Los siguientes resultados fueron obtenidos a partir de la clasificación del descriptor sobre el espacio euclidiano después de hacer la transformación con logaritmo, expuesta en la sección 3.1.2, variando la organización de los centroides extraídos de los algoritmos de agrupamiento en dos configuraciones; por orden de importancia y orden aleatorio. En un primer experimento se concatenaron los centroides sin orden específico, obteniendo los resultados ilustrados en la figura 13. En este caso se logró una exactitud de 75 % para la mejor configuración, usando la arquitectura MobileNet, 300 centroides y regiones de 80×80.

Figura 13. Resultados parciales sobre el dataset *UT-Interaction*[52] con experimentos realizados con diferentes áreas alrededor de las trayectorias y variando el número de centroides sin presentar un orden específico en estos, para su clasificación.



También se realizó un experimento ordenando el descriptor según la importancia de los centroides, determinada por el número de covarianzas en cada grupo. Los resultados son expuestos en la figura 14. Utilizando este orden se lograron incrementar los resultados en un 8.3 %.

Figura 14. Resultados parciales sobre el dataset *UT-Interaction*[53] de experimentos realizados con diferentes áreas alrededor de las trayectorias y variando el número de centroides organizados por importancia en el descriptor, para su clasificación.



7.0.2. Covarianzas en el espacio de Riemann Teniendo en cuenta las regiones relativamente pequeñas para el cálculo de la covarianza, se entendió el espacio de Riemann como un espacio factible para la operación. Por lo tanto, se realizaron experimentos operando directamente con las covarianzas, sin vincular el error numérico asociado a la transformación logarítmica. En este sentido, se realizaron los mismos experimentos. Inicialmente en la figura se ilustran los resultados del descriptor, concatenando las matrices de covarianza sin un orden pre-establecido. En la figura 15 se ilustran los resultados obtenidos bajo esta hipótesis, que resultan más favorables que teniendo en cuenta el plano tangente en el espacio euclidiano. Estos resultados pueden estar justificados en el hecho que el error numérico de proyección logarítmica sobre las M covarianzas del video puede ser predominante.

Además, se realizó el experimento ordenando las matrices volumétricas, según la relevancia de los clúster asociados a cada centroide. La figura 16 muestra los resultados obtenidos con este método, permaneciendo estable en las diferentes configuraciones.

Figura 15. Resultados sobre sobre el dataset *UT-Interaction*[54] de experimentos realizados sobre el espacio de Riemann, con diferentes áreas alrededor de las trayectorias y variando el número de centroides, para su clasificación.

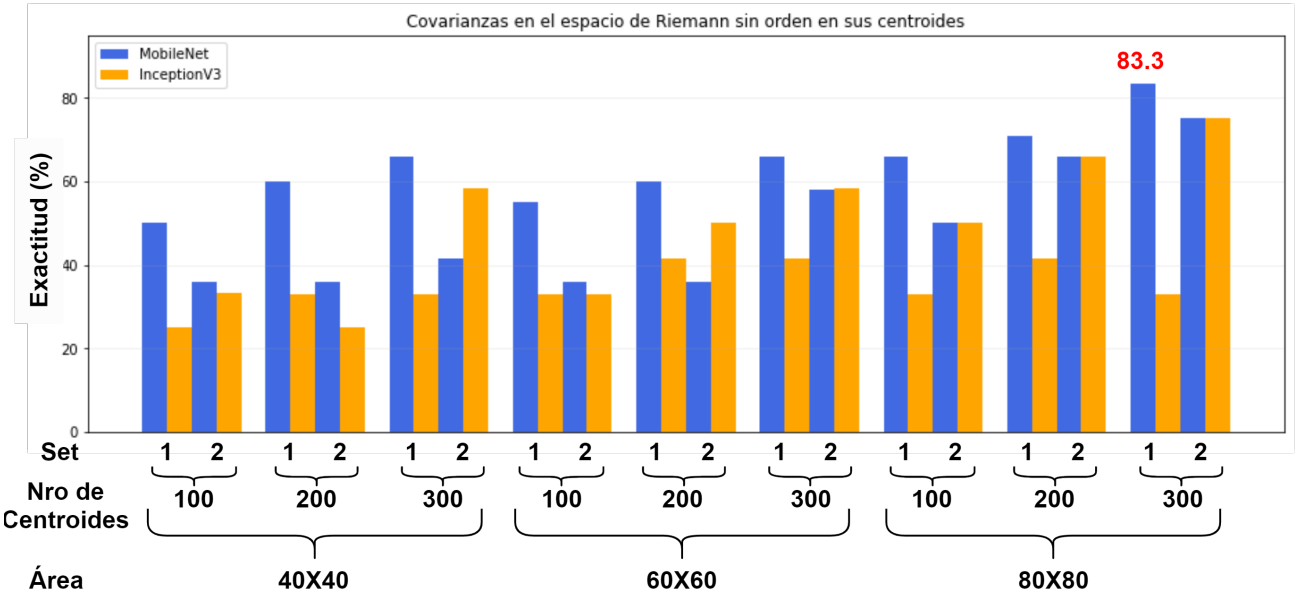
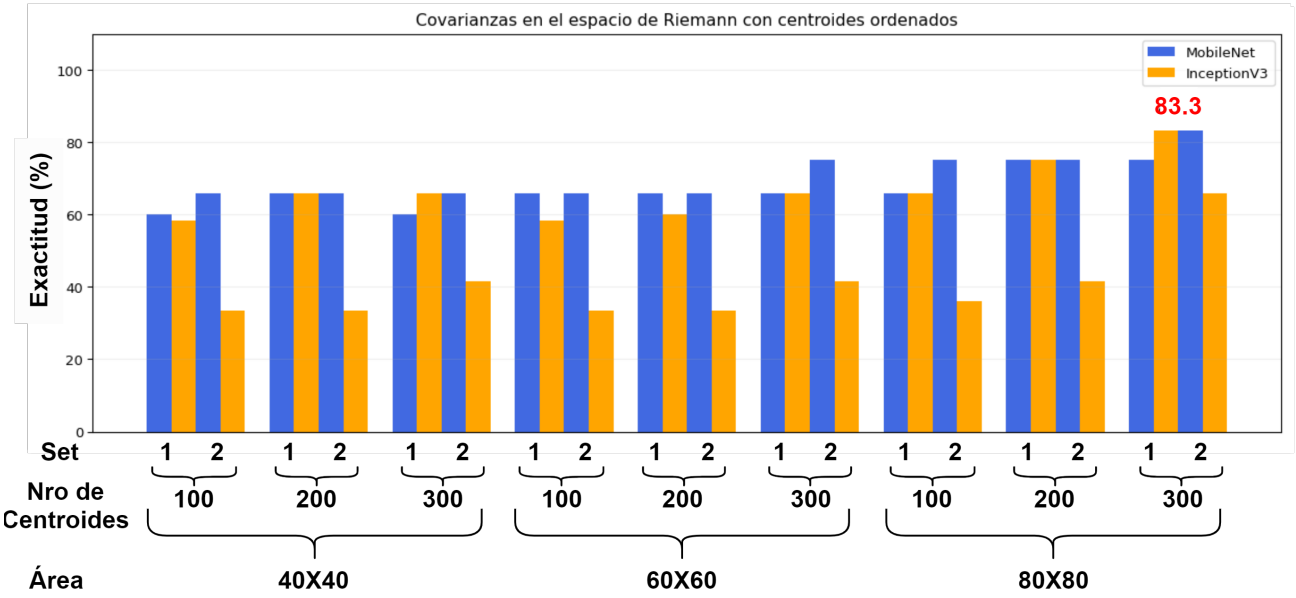


Figura 16. Resultados sobre el dataset *UT-Interaction*[55] de experimentos realizados sobre el espacio de Riemann, con diferentes áreas alrededor de las trayectorias y variando el número de centroides organizados por importancia en el descriptor, para su clasificación.



8. CONCLUSIONES Y PERSPECTIVAS

Este trabajo propuso una alternativa para describir localmente puntos de interés espacio-temporales utilizando covarianzas volumétricas integrales. Esta representación es flexible y sigue puntos relacionados con las acciones. Para una descripción robusta, cada cuadro fue representado por el conjunto de activaciones de una arquitectura pre-entrenada. Entonces, cada secuencia de video es representada por un conjunto de covarianzas, que de forma robusta representan diferentes regiones altamente correlacionadas con la acción de interés. Este conjunto denso de covarianzas permite calcular K covarianzas representativas, que constituyen el descriptor de video. Este descriptor es entonces mapeado a un clasificador, previamente entrenado, que permite etiquetar automáticamente las secuencias de video.

El método propuesto logró resultados sobresalientes en la tarea de clasificación sobre el dataset UT-Interaction. Las caracteristicas profundas resultaron salientes para representar regiones, y las matrices de covarianza fueron efectivas para describir patrones espacio-temporales. Como trabajo futuro se esperan hacer representaciones locales que incluyan patrones de movimiento, así como la construcción de diccionarios sobre el conjunto de patrones volumétricos. También se espera hacer una validación exhaustiva sobre otros conjuntos de datos.

BIBLIOGRAFÍA

Avci, Akin y col. "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey". En: *23th International conference on architecture of computing systems 2010*. VDE. 2010, págs. 1-10 (vid. pág. 26).

Bilen, Hakan y col. "Dynamic image networks for action recognition". En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, págs. 3034-3042 (vid. pág. 30).

Cao, Xiaochun y col. "Action recognition using 3D DAISY descriptor". En: *Machine vision and applications* 25.1 (2014), págs. 159-171 (vid. pág. 48).

Coppola, Claudio y col. "Social Activity Recognition on Continuous RGB-D Video Sequences". En: *International Journal of Social Robotics* (2019), págs. 1-15 (vid. pág. 27).

Dalal, Navneet y Bill Triggs. "Histograms of oriented gradients for human detection". En: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE. 2005, págs. 886-893 (vid. pág. 28).

Dawn, Debapratim Das y Soharab Hossain Shaikh. "A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector". En: *The Visual Computer* 32.3 (2016), págs. 289-306 (vid. pág. 28).

Donahue, Jeffrey y col. "Long-term recurrent convolutional networks for visual recognition and description". En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, págs. 2625-2634 (vid. pág. 15).

Farnebäck, Gunnar. "Two-frame motion estimation based on polynomial expansion". En: *Scandinavian conference on Image analysis*. Springer. 2003, págs. 363-370 (vid. pág. 37).

Fernando, Basura y col. "Modeling video evolution for action recognition". En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, págs. 5378-5387 (vid. pág. 29).

Garzón, Gustavo y Fabio Martínez. "Online Action Recognition from Trajectory Occurrence Binary Patterns (ToBPs)". En: *The International Conference on Advances in Emerging Trends and Technologies*. Springer. 2019, págs. 409-418 (vid. pág. 29).

Hussein, Mohamed E y col. "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations". En: *Twenty-Third International Joint Conference on Artificial Intelligence*. 2013 (vid. pág. 20).

Ji, Shuiwang y col. "3D convolutional neural networks for human action recognition". En: *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012), págs. 221-231 (vid. pág. 31).

Ji, Xiaofei y col. "Multiple feature voting based human interaction recognition". En: *International Journal of Signal Processing, Image Processing and Pattern Recognition* 9.1 (2016), págs. 323-334 (vid. pág. 48).

Karpathy, Andrej y col. "Large-scale video classification with convolutional neural networks". En: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014, págs. 1725-1732 (vid. págs. 15, 29).

Laptev, Ivan y col. "Learning realistic human actions from movies". En: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, págs. 1-8 (vid. págs. 14, 28).

Li, Li-Jia y Li Fei-Fei. "Optimol: automatic online picture collection via incremental model learning". En: *International journal of computer vision* 88.2 (2010), págs. 147-168 (vid. pág. 28).

Ma, Bingpeng, Yu Su y Frederic Jurie. "Covariance descriptor based on bio-inspired features for person re-identification and face verification". En: *Image and Vision Computing* 32.6-7 (2014), págs. 379-390 (vid. pág. 20).

Meshry, Moustafa, Mohamed E Hussein y Marwan Torki. "Linear-time online action detection from 3d skeletal data using bags of gesturelets". En: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2016, págs. 1-9 (vid. pág. 28).

Minh, Hà Quang y Vittorio Murino. "Covariances in computer vision and machine learning". En: *Synthesis Lectures on Computer Vision* 7.4 (2017), págs. 1-170 (vid. págs. 14, 17, 21).

Moreno, Wilson, Gustavo Garzón y Fabio Martínez. "Frame-Level Covariance Descriptor for Action Recognition". En: *Colombian Conference on Computing*. Springer. 2018, págs. 276-290 (vid. págs. 17, 48).

Mukherjee, Snehasis, Sujoy Kumar Biswas y Dipti Prasad Mukherjee. "Recognizing interaction between human performers using'key pose doublet'". En: *Proceedings of the 19th ACM international conference on Multimedia*. 2011, págs. 1329-1332 (vid. pág. 48).

Rautaray, Siddharth S y Anupam Agrawal. "Vision based hand gesture recognition for human computer interaction: a survey". En: *Artificial intelligence review* 43.1 (2015), págs. 1-54 (vid. pág. 27).

Ryoo, M. S. y J. K. Aggarwal. *UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)*. http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html. 2010 (vid. págs. 20, 49-51).

Ryoo, Michael S. “Human activity prediction: Early recognition of ongoing activities from streaming videos”. En: *2011 International Conference on Computer Vision*. IEEE. 2011, págs. 1036-1043 (vid. pág. 48).

Ryoo, Michael S y JK Aggarwal. “UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA)”. En: *IEEE International Conference on Pattern Recognition Workshops*. Vol. 2. 2010, pág. 4 (vid. pág. 42).

Sahoo, Suraj Prakash y Samit Ari. “On an algorithm for human action recognition”. En: *Expert Systems with Applications* 115 (2019), págs. 524-534 (vid. pág. 14).

Sandler, Mark y col. “Mobilenetv2: Inverted residuals and linear bottlenecks”. En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, págs. 4510-4520 (vid. págs. 40, 41).

Scovanner, Paul, Saad Ali y Mubarak Shah. “A 3-dimensional sift descriptor and its application to action recognition”. En: *Proceedings of the 15th ACM international conference on Multimedia*. 2007, págs. 357-360 (vid. pág. 48).

Simonyan, Karen y Andrew Zisserman. “Two-stream convolutional networks for action recognition in videos”. En: *Advances in neural information processing systems*. 2014, págs. 568-576 (vid. págs. 15, 31).

Slimani, Khadidja Nour el houda, Yannick Benezeth y Feriel Souami. “Human interaction recognition based on the co-occurence of visual words”. En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2014, págs. 455-460 (vid. págs. 17, 48).

Staudemeyer, Ralf C y Eric Rothstein Morris. "Understanding LSTM–a tutorial into Long Short-Term Memory Recurrent Neural Networks". En: *arXiv preprint arXiv:1909.09586* (2019) (vid. pág. 31).

Szegedy, Christian y col. "Rethinking the inception architecture for computer vision". En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, págs. 2818-2826 (vid. pág. 39).

Tran, Du y col. "Learning spatiotemporal features with 3d convolutional networks". En: *Proceedings of the IEEE international conference on computer vision*. 2015, págs. 4489-4497 (vid. pág. 30).

Tuzel, Oncel, Fatih Porikli y Peter Meer. "Region covariance: A fast descriptor for detection and classification". En: *European conference on computer vision*. Springer. 2006, págs. 589-600 (vid. pág. 21).

Ullah, Amin y col. "Action recognition in video sequences using deep bi-directional LSTM with CNN features". En: *IEEE Access* 6 (2017), págs. 1155-1166 (vid. pág. 15).

Varol, Gül, Ivan Laptev y Cordelia Schmid. "Long-term temporal convolutions for action recognition". En: *IEEE transactions on pattern analysis and machine intelligence* 40.6 (2017), págs. 1510-1517 (vid. págs. 14, 15, 30).

Vishwakarma, Sarvesh y Anupam Agrawal. "A survey on activity recognition and behavior understanding in video surveillance". En: *The Visual Computer* 29.10 (2013), págs. 983-1009 (vid. pág. 27).

Wang, Heng y Cordelia Schmid. "Action recognition with improved trajectories". En: *Proceedings of the IEEE international conference on computer vision*. 2013, págs. 3551-3558 (vid. pág. 15).

Wang, Heng y col. “Action recognition by dense trajectories”. En: 2011 (vid. pág. 29).

Wang, Heng y col. “Dense trajectories and motion boundary descriptors for action recognition”. En: *International journal of computer vision* 103.1 (2013), págs. 60-79 (vid. págs. 14, 28, 36).

Wang, Jue y col. “Video representation learning using discriminative pooling”. En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, págs. 1149-1158 (vid. págs. 14, 15, 29).

Wang, Limin y col. “Towards good practices for very deep two-stream convnets”. En: *arXiv preprint arXiv:1507.02159* (2015) (vid. págs. 15, 29).

Wang, Xuanhan y col. “Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length”. En: *IEEE Transactions on Multimedia* 20.3 (2017), págs. 634-644 (vid. pág. 30).

Willamowski, Jutta y col. “Categorizing nine visual classes using local appearance descriptors”. En: *illumination* 17 (2004), pág. 21 (vid. pág. 28).

Xu, Bing y col. “Empirical evaluation of rectified activations in convolutional network”. En: *arXiv preprint arXiv:1505.00853* (2015) (vid. pág. 38).

Yu, Gang, Junsong Yuan y Zicheng Liu. “Propagative hough voting for human activity recognition”. En: *European Conference on Computer Vision*. Springer. 2012, págs. 693-706 (vid. págs. 17, 48).

Yue-Hei Ng, Joe y col. “Beyond short snippets: Deep networks for video classification”. En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, págs. 4694-4702 (vid. pág. 16).