PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

RECONOCIMIENTO DE ACCIONES EN VIDEOS MEDIANTE TÉCNICAS DE COMPOSICIÓN JERÁRQUICAS

SANTIAGO LARRAIN PÉREZ

Tesis para optar al grado de
Magíster en Ciencias de la Ingeniería

Profesor Supervisor:
ALVARO SOTO, PH.D.

Santiago de Chile, Octubre 2017

© MMXVII, SANTIAGO LARRAIN PÉREZ

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

RECONOCIMIENTO DE ACCIONES EN VIDEOS MEDIANTE TÉCNICAS DE COMPOSICIÓN JERÁRQUICAS

SANTIAGO LARRAIN PÉREZ

Miembros del Comité:
ALVARO SOTO, PH.D.
HANS-ALBERT LÖBEL, PH. D.
BILLY PERALTA, PH. D.
AMADOR GUZMÁN, PH. D.

Tesis para optar al grado de
Magíster en Ciencias de la Ingeniería

Santiago de Chile, Octubre 2017

© MMXVII, SANTIAGO LARRAIN PÉREZ

Dedicado a Esperanza

ÍNDICE

ÍNDICE DE FIGURAS VI

ÍNDICE DE TABLAS VII

ABSTRACT VIII

RESUMEN IX

Capítulo 1. INTRODUCCIÓN 1

1.1. Motivación 1

1.2. Método propuesto 2

Capítulo 2. REVISIÓN BIBLIOGRÁFICA 5

Capítulo 3. ANTECEDENTES 8

3.1. Descriptores locales espacio-temporales 8

3.2. Selección de key frames 9

3.3. Método de segmentación 10

Capítulo 4. MÉTODO PROPUESTO 13

4.1. Primer nivel de jerarquía 14

4.2. Segundo nivel de jerarquía 16

4.3. Vector característico 17

Capítulo 5. RESULTADOS EXPERIMENTALES CUANTITATIVOS 19

5.1. Set de datos 19

5.2. Resultados cuantitativos 21

5.2.1. Resultados generales 21

5.2.2. Análisis de sensibilidad sobre tamaños de diccionarios 22

5.2.3. Análisis de ablación sobre canales de información 23

5.2.4. Configuración óptima del modelo 25

5.2.5. Análisis de sensibilidad sobre regularizador 26
5.2.6. Benchmark y comparación con otros trabajos 27
5.3. Resultados cualitativos 29
5.3.1. Análisis del uso de palabras visuales 29
5.3.2. Análisis de matriz de confusión 31
5.3.3. Visualización de palabras visuales destacadas 34
Capítulo 6. CONCLUSIONES Y TRABAJO FUTURO 39
REFERENCIAS 41

ÍNDICE DE FIGURAS

3.1. Representación de descriptores espacio-temporales 8

3.2. Ejemplo de selección de *key frames* 10

3.3. Ejemplo de segmentación de videos 12

4.1. Arquitectura de la solución 14

4.2. Generación de áreas del segundo nivel 16

5.1. Set de datos *UCF11* 19

5.2. Set de datos *Olympic Sports* 20

5.3. Test de ablación 25

5.4. Resultados de sensibilidad 27

5.5. Uso de palabras visuales por nivel 30

5.6. Pesos asignados por el clasificador a cada palabra 31

5.7. Matrices de confusión 32

5.8. Videos visualmente similares 33

5.9. Videos visualmente diferentes 34

5.10. Visualizaciones de palabras del primer nivel usando *k-means* 36

5.11. Visualizaciones de palabras del segundo nivel usando *k-means* 37

ÍNDICE DE TABLAS

5.1. Descripción de los set de datos 20

5.2. Resultados experimentales cuantitativos 22

5.3. Resultados variando los tamaños de diccionarios k_1 y k_2 23

5.4. Resultados concatenando nivel 1 y 2 23

5.5. Configuración óptima del modelo 26

5.6. Comparación con métodos del estado del arte 28

5.7. Resultados variando los tamaños de diccionarios k_1 y k_2 35

5.8. Resultados concatenando nivel 1 y 2 35

5.9. Comparación con métodos del estado del arte 38

5.10. Resultados experimentales cuantitativos 38

5.11. Configuración óptima del modelo 38

ABSTRACT

Human action recognition in video sequences is a challenging task due to high intra-class variability between instances. This variability is due mainly to camera movements, changes in luminosity and differences in resolution, among other causes. To solve this problem, this thesis proposes a machine learning approach based on hierarchical techniques, using visual dictionaries. The proposal is based on the use of two levels of hierarchy and explores the use of two different methods to learn the dictionaries: *k-means* and *sparse coding*. Experimental results on two benchmark datasets, *UCF11* and *Olympic Sports*, shows that the concatenation of the feature vectors obtained by both hierarchy levels improves the classification results of the algorithm, compared with the use of a unique level. In both datasets, *k-means* algorithm gives better classification results than *sparse coding*, when used to learn the dictionaries. State-of-the-art results are achieved with the *UCF11* data set.

Keywords: computer vision, action recognition, hierarchical representation, sparse coding, BoW, machine learning.

RESUMEN

El reconocimiento de acciones en secuencias de video corresponde a un problema desafiante debido a la alta variabilidad intra-clase que existe entre las instancias. La variabilidad se debe principalmente a movimientos de cámara, cambios de luminosidad y diferencias en resolución. Para atacar este problema, esta tesis propone un algoritmo de aprendizaje de máquina basado en técnicas jerárquicas, utilizando diccionarios visuales. El algoritmo utiliza dos niveles de jerarquía y evalúa dos métodos diferentes para el aprendizaje de diccionarios: *k-means* y *sparse coding*. Los resultados experimentales con dos bases de datos, *UCF11* y *Olympic Sports*, muestran que el uso de un segundo nivel de jerarquía mejora la calidad de la clasificación de las secuencias, al concatenar los vectores descriptivos de ambos niveles. En ambas bases de datos, *k-means* entrega mejores resultados que *sparse coding*, al ser utilizado para el aprendizaje de diccionarios. La tasa de clasificación obtenida por el algoritmo en la base de datos *UFC11* supera las tasas de clasificación de las metodologías del estado del arte.

Palabras Claves: visión por computador, reconocimiento de acciones, representaciones jerárquicas, *sparse coding*, *BoW*, aprendizaje de máquina.

CAPÍTULO 1. INTRODUCCIÓN

1.1. Motivación

Hoy en día la producción de videos va en aumento. A modo de ejemplo podemos considerar los videos generados por cámaras de vigilancia en operación continua, o los miles de videos compartidos diariamente en redes sociales. Es tal la cantidad de datos generada, que usar personas para procesar videos es costoso e ineficiente. Esto ha llamado la atención de la comunidad científica, la que ha generado modelos de inteligencia artificial capaces de procesar, entender y automatizar ciertas tareas asociadas a videos (Poppe 2010).

Una de las tareas asociadas al procesamiento de videos es la detección y clasificación de acciones. Dicha tarea puede ser útil en detección automática de incidentes en temas de vigilancia, como también en la búsqueda y análisis de videos en las redes sociales.

La clasificación de acciones en videos comprende un gran desafío, debido a la alta variabilidad intraclases. Entre las causas de esta variabilidad podemos encontrar: el movimiento del fondo de cada escena, los cambios en la luminosidad entre videos grabados al interior y exterior de un inmueble, la rotación y el tiempo de grabación de cada video, entre otros.

Con el propósito de generar modelos robustos a los desafíos planteados anteriormente, se han desarrollado técnicas análogas a las utilizadas en el procesamiento de imágenes. Entre ellas, en el procesamiento de videos se ha popularizado el uso de descriptores locales, como por ejemplo, los histogramas de gradiente (HOG3D) (Klaser, Marszałek y Cordelia Schmid 2008), histogramas de flujo óptico (HOF) (Laptev y col. 2008), histogramas de movimientos de bordes (MBH) (Dalal, Triggs y Cordelia Schmid 2006b), etc. Al igual que en la detección y clasificación de objetos en imágenes, se han introducido niveles de abstracción o de representaciones intermedias, como el uso de diccionarios visuales (Leung y Malik 2001). Dichas representaciones intermedias, pueden ser generalizadas a más de

una capa, generando modelos con información jerárquica, como en el trabajo de Taralova *et. al.* (Taralova, De la Torre y Hebert 2014), el cual utiliza técnicas de *Bag of Visual Words* (BoVW) mediante diccionarios visuales aprendidos con el algoritmo de *clustering k-means*.

Yang *et. al.* (Jianchao Yang y col. 2009) ha demostrado empíricamente que el uso técnicas de *sparse coding* en la generación de diccionarios visuales mejora las tareas de clasificación en comparación al uso de *k-means*. En particular, uno de los objetivos de esta tesis es estudiar si este resultado se mantiene al utilizar *k-means* y *sparse coding* para el aprendizaje de diccionarios en modelos jerárquicos, en la clasificación de acciones en videos. Para esto, se estudia el comportamiento de estas técnicas utilizando diferentes tamaños de diccionarios, y diferentes valores de regularizador en el caso de *sparse coding*. Se aplicarán dichos algoritmos en el uso de diccionarios con jerarquías de uno y dos niveles.

Esta tesis también busca comprobar si el uso de dos niveles de jerarquía mejora los resultados en el actual estado del arte en tareas de clasificación. Dado que los descriptores del primer nivel corresponden a un conjunto de sub-descriptores, se analiza cuál es la configuración óptima para realizar tareas de clasificación, en cuanto a la mejor combinación de los descriptores.

1.2. Método propuesto

Una de las técnicas usadas en la representación de videos es el uso de diccionarios visuales, la cual ha sido explorada en áreas de clasificación de videos (Peng, L. Wang, Qiao y col. 2014), como también en la clasificación de imágenes (Boureau y col. 2010). Dichas técnicas han logrado mejoras en el estado del arte al incorporar una abstracción intermedia entre los *pixeles* del video y el clasificador final. Estas estructuras intermedias son capaces de detectar diferentes patrones que hacen que un tipo de acción sea diferente a otra.

La base del método propuesto es el uso de diccionarios en forma jerárquica, donde diferentes diccionarios son estructurados generando distintos niveles de abstracción intermedia, lo que permite discriminar entre distintos tipos o categorías de acciones. Esta jerarquía es análoga a la composición de un texto, donde cada uno de ellos es descompuesto en párrafos.

El método descrito en esta tesis, propone clasificar videos basado en diccionarios jerárquicos, descomponiendo el video en áreas que, a su vez, se componen de descriptores visuales de bajo nivel (sección 3.1). Dicha abstracción jerárquica es desarrollada mediante diferentes diccionarios visuales para cada nivel.

El primer nivel en la jerarquía corresponde a descriptores visuales propuestos por Wang *et. al.* (H. Wang y col. 2011), con los cuales se aprende un diccionario de palabras visuales, las que en este caso corresponden a una serie de descriptores visuales como trayectorias de puntos de interés (*Trajectory*), histogramas de gradiente orientado (*HOG*), histogramas de flujo óptico (*HOF*) e histogramas de movimientos de movimientos de bordes en los ejex x e y (*MBHx* y *MBHy*) . Para el segundo nivel, se utiliza un algoritmo de segmentación (ver sección 3.3), donde cada segmento es representado mediante un histograma de las palabras visuales, aprendidas en el primer nivel. Los descriptores generados para los segmentos son procesados para generar los diccionarios de segundo nivel. Finalmente, cada video es representado como un histograma del uso de las palabras visuales del primer y segundo nivel de forma concatenada, y son clasificados mediante un clasificador del tipo máquina de soporte o *SVM* con kernel lineal. La metodología propuesta se explica en mayor detalle en el capitulo 4.

Los principales aportes del modelo planteado en esta tesis considera: (i) uso de jerarquías composicionales en la detección y clasificación de acciones humanas en videos, dichas jerarquías proponen una forma novedosa de realizar *pooling* en la generación de abstracciones de alto nivel, mediante algoritmos de segmentación visual; (ii) estudio de las estrategias de concatenación de los descriptores visuales de ambos niveles de la jerarquía para la generación del vector característico final; (iii) extensión de la comparación

realizada en el análisis de imágenes, entre el desempeño de *k-means* y *sparse coding* en el aprendizaje de diccionarios visuales.

El documento se estructura de la siguiente forma: el capítulo 2 expone los principales trabajos en el reconocimiento de acciones en videos. El capítulo 3 explica en profundidad los trabajos previos que se utilizaron como base para el desarrollo de esta tesis. El capitulo 4 describe en profundidad el método propuesto para la clasificación de acciones humanas en videos. El capítulos 5 describe los resultados experimentales cuantitativos y cualitativos, obtenidos. Finalmente, el capítulo 6 expone las principales conclusiones del estudio, así como también posibles trabajos futuros en el área.

CAPÍTULO 2. REVISIÓN BIBLIOGRÁFICA

La tarea de reconocer acciones en videos es un problema abierto y bastante estudiado por la comunidad científica (Ziaeefard y Bergevin 2015) y (Aggarwal y Ryoo 2011). Los primeros enfoques que se aplicaron para resolver este problema interpretan las acciones como uniones del cuerpo humano que cambian en el espacio y tiempo (Campbell y Bobick 1995), (Niyogi y Adelson 1994) y (Yacoob y Black 1999). Este tipo de soluciones carecen de la capacidad de generalizar los patrones de las acciones y se consideran *ad-hoc* al intentar modelar el esqueleto humano. El método propuesto en esta tesis genera abstracciones sobre descriptores de puntos de interés, lo cual lo hace generalizable a cualquier tipo de video y no sólo a la clasificación de acciones humanas

Una estructura bastante común para atacar el problema es el uso de *Bag of Visual Words* (BoVW), la que se se compone de las siguientes sub-tareas: (i) extracción de descriptores, (ii) pre-procesamiento de descriptores, (iii) generación de diccionarios visuales, (iv) codificación de descriptores, (v) *pooling* y normalización (Peng, L. Wang, X. Wang y col. 2016). Esta estructura ha mostrado buenos resultados al momento de clasificar acciones.

El proceso de extracción de descriptores de bajo nivel se puede dividir en los procesos de detección de puntos de interés y la descripción de la región detectada. Varias de estas técnicas son extensiones de descriptores aplicados al procesamiento de imágenes. Para las tareas de detección de puntos de interés se encuentran los detectores *Harris 3D* (Laptev 2005), *3D Hessian* (Willems, Tuytelaars y Van Gool 2008), *Cuboids* (Dollar y col. 2005) y *Dense Trajectories* (H. Wang y col. 2011). Para la representación de las regiones alrededor de los puntos de interés se han aplicado las técnicas *HOG3D* (Dalal, Triggs y Cordelia Schmid 2006a), *Local Trinary Pattern* (Yeffet y Wolf 2009), *Motion Boundary Histogram* (MBH) (Laptev 2005), *Histogram of Optical Flow* (HOF) (Chaudhry y col. 2009). El uso de descriptores locales ha mostrado ser una buena base para representaciones intermedias mediante *BoVW*, debido a que éstos son robustos ante perturbaciones locales.

En el flujo típico de técnicas de detección de acciones en video, la detección y caracterización de puntos de interés es luego seguida de una etapa de aprendizaje de diccionarios visuales. Este proceso se puede realizar dividiendo el espacio vectorial, o bien, mediante modelos generativos. La forma más común de dividir el espacio vectorial es mediante el algoritmo de *clustering k-means* (Jun Yang y col. 2007), mientras que *sparse coding* es una técnica generativa (Jianchao Yang y col. 2009).

Otros enfoques para modelar acciones en videos incorporan estructura al problema. Wang y Mori generan un modelo gráfico basado en *Hidden Conditional Random Fields* (HCRF), donde las poses humanas son modeladas utilizando constelaciones condicionales (Y. Wang y Mori 2011). Otro modelo estructural es el de Hu *et. al.*, quienes reconocen acciones humanas mediante estructuras latentes que son aprendidas mediante *Structured-Support Vector Machines* (Hu y col. 2014).

Esta tesis propone un modelo jerárquico basado en diccionarios visuales, los que pueden ser generados mediante *sparse coding* o *k-means*. *Sparse coding* surge del análisis de señales y ha sido utilizado en diversas aplicaciones de procesamiento de imágenes (Huang y Aviyente 2006), logrando grandes avances en el reconocimiento de rostros (Wright y col. 2009), (Ramirez, Sprechmann y Sapiro 2010) y también en clasificación de dígitos y texturas (Mairal y col. 2008).

Se ha mostrado que *sparse coding* también tiene la capacidad de modelar diferentes estructuras para la clasificación de acciones. Gaido *et. al.* representan cada video a través de una secuencia de acciones atómicas llamadas *Actom Sequence Model* (A. Gaidon, Z. Harchaoui y C. Schmid 2011), mediante la técnica de *Bag of Words* (BoW). La desventaja de este modelo es que los átomos del diccionario son rotulados de forma manual. Guha y Ward argumentan que las estructuras generadas mediante *sparse coding* incorporan más información comparado con los métodos de cuantización de vectores (Guha y Ward 2012). Por último Tran *et. al.* representan acciones humanas como un conjunto de partes (Tran, Kakadiaris y S. K. Shah 2011).

Taralova *et. al.* (Taralova, De la Torre y Hebert 2014) incorporan un modelo de dos jerarquías para el reconocimiento de acciones humanas en videos. En cada nivel de jerarquía se genera un diccionario visual mediante el algoritmo de *k-means*, del cual se compone el nivel subyacente. Lo anterior, permite que el modelo genere diferentes vectores característicos, en cada nivel, lo que mejora la precisión en tareas de clasificación.

Considerando lo anterior, esta tesis propone modelar cada video de forma jerárquica mediante diccionarios visuales. Cada elemento de una jerarquía se compone de elementos de la jerarquía anterior, de este modo el modelo aprende a componer videos desde elementos de bajo nivel hacia elementos más sofisticados. La ventaja del modelo propuesto radica en la capacidad de aprender nuevas formas de codificación para cada jerarquía, generando diccionarios compartidos a lo largo de las diferentes categorías de acciones.

El trabajo propuesto en esta tesis se diferencia del trabajo de Taralova *et. al.*, principalmente, en la utilización de *sparse coding* como método de codificación de los descriptores y en la comparación de dicha técnica con *k-means*, realizando un análisis exhaustivo de la mejor forma de combinar los descriptores del primer y segundo nivel para mejorar la clasificación final de cada video.

CAPÍTULO 3. ANTECEDENTES

En este capítulo se describen específicamente las metodologías que se utilizan en la implementación del modelo propuesto en el capítulo 4.

3.1. Descriptores locales espacio-temporales

Los descriptores propuestos por Wang *et. al.* (H. Wang y col. 2011) corresponden al componente de más bajo nivel utilizado en el modelo propuesto. Estos son calculados en base a la detección de puntos de interés, los cuales son seguidos mediante un algoritmo de *tracking* por 15 *frames*. Alrededor de cada trayectoria se define una zona espacio-temporal de dimensionalidad 32 *pixeles* $\times$ 32 *pixeles* $\times$ 15 *frames*. Finalmente, para cada cubo se obtiene y concatenan 5 descriptores diferentes:

(I) *Trajectory* $\in \mathbb{R}^{30}$: incorpora información de movimiento que describe el punto de interés

(II) *HOG* $\in \mathbb{R}^{96}$: histogramas de gradiente orientados

(III) *HOF* $\in \mathbb{R}^{108}$: histogramas de flujo óptico

(IV) *MBHx* $\in \mathbb{R}^{96}$: histogramas de movimiento de bordes en el eje x

(V) *MBHy* $\in \mathbb{R}^{96}$: histogramas de movimiento de bordes en el eje y

La figura 3.1 muestra una representación esquemática de los descriptores.

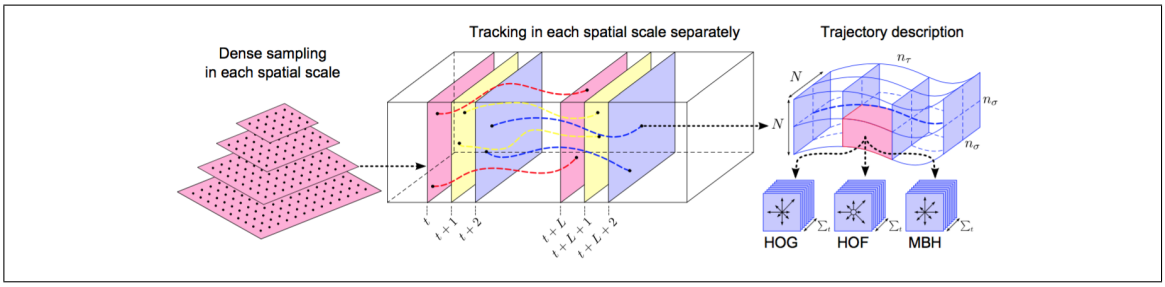


Figura 3.1. Esquema de descriptores espacio-temporales propuestos por Wang *et. al.*

3.2. Selección de key frames

Uno de los problemas relevantes al operar sobre videos, es el gran requerimiento de recursos computacionales para procesar todos sus *frames*. Dada la redundancia de la información que estos contienen, una estrategia típica es operar sólo sobre un subconjunto de *frames* representativos. Con el propósito de resumir los videos sin perder información relevante, se optó por utilizar una adaptación del método propuesto por Elhamifar *et. al.* (Elhamifar, Sapiro y Vidal 2012). En dicha adaptación, se utiliza la solución del problema de optimización descrito en 3.1 para reconstruir los *frames* de cada clase utilizando los *frames* de cada video. De esta forma, se selecciona el mejor conjunto de *frames* que reconstruye el video y también la clase a la cual pertenece. Este método ha mostrado ser eficiente en la selección de *key frames* (Alfaro, Mery y Soto 2016).

$$\begin{aligned} \min_{\mathbf{C}} \quad & ||\mathbf{Y}_c - \mathbf{Y}_v\mathbf{C}||_F^2 \\ \text{s.t.} \quad & ||\mathbf{C}||_{1,2} \leq \tau \\ & \mathbf{1}^T\mathbf{C} = \mathbf{1}^T \end{aligned} \tag{3.1}$$

En 3.1, $\mathbf{Y}_c$ representa todos los *frames* de la clase c, $\mathbf{Y}_v$ representa los *frames* del video v, y $\mathbf{C}$ son los coeficientes de reconstrucción de los *frames* $\mathbf{Y}_v$. Una vez resuelto el problema de optimización, se seleccionan los 3 *frames* que mejor representan al video y a su respectiva clase (Alfaro, Mery y Soto 2016). También se toma en consideración la distribución temporal de los *frames* elegidos. Finalmente, los frames seleccionados se aumentan con información temporal agregando a cada uno de ellos los 15 *frames* contiguos en el tiempo. A modo de ejemplo, la figura 3.2 muestra resultados de *key frames* seleccionado para diferentes videos del set de datos *Olympic Sports*[1], que se describe en mayor profundidad en la sección 5.1.

[1]Niebles, J. C., wei Chen, C. y Fei-fei, L. (2010), disponible en `http://vision.stanford.edu/Datasets/OlympicSports/`

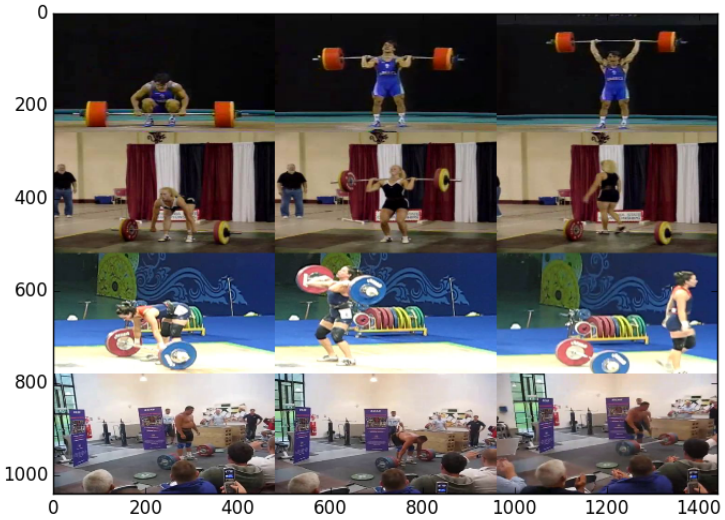


Figura 3.2. Ejemplo de *key frames* elegidos para tres videos de la clase *snatch* (levantamiento de pesas) del set de datos *Olympic Sports*.

3.3. Método de segmentación

El segundo nivel de jerarquía del método propuesto consiste en volúmenes espacio-temporales que son representadas por un conjunto descriptores de *Wang*. Motivados por generar volúmenes entendibles para la visión humana, se utiliza un algoritmo de segmentación visual. Dicho algoritmo permite generar regiones basadas en la intensidad de *pixeles* de cada *frame* y formar áreas consistentes en el tiempo.

El algoritmo utiliza segmentos jerárquicos basados en grafos, según la propuesta de Grundmann *et. al.* (Grundmann y col. 2010). Esta propuesta corresponde a una extensión del método de Felzenszwalb y Huttenlocher (Felzenszwalb y Huttenlocher 2004).

El método de Felzenszwalb y Huttenlocher representa las imágenes como grafos no dirigidos, donde cada *pixel* es un vértice que está unido a sus 8 vecinos. El peso asociado a cada arista corresponde a la similitud entre los *pixeles* vecinos. De manera formal, sea $G = (V, E)$ un grafo no dirigido con vértices $v_i \in V$, aristas $(v_i, v_j) \in E$ y una función

de peso asociado a cada arista de la forma $w(v_i, v_j)$. Este método busca unir en una misma región R todos los *pixeles* que tienen alta similitud tomando en consideración la variación interna de la región. De esta forma la variación interna $Int(R)$ de la región R es definida como el peso máximo de una arista de su árbol de cobertura mínima (*Minimum Spanning Tree (MST)*).

$$Int(R) = \max_{(v_i,v_j)\in MST(R)} w(v_i, v_j) \tag{3.2}$$

Dos regiones (C_1, C_2) son unidas, si el vértice de mayor similaridad, $Dif(C_1, C_2)$, entre ambos sub-grafos es menor a la variación interna de ambas regiones. Formalmente esto se traduce al predicado D definido como:

$$D(C_1, C_2) = \begin{cases} \text{verdadero} & \text{si } Deeplearningforactionandgesturerecognitioninimagesequences : c \\ \text{falso} & \text{e.o.c} \end{cases} \tag{3.3}$$

donde la diferencia interna mínima $MInt(C_1, C_2)$ se define como:

$$MInt(C_1, C_2) = \min\left(Int(C_1) + \frac{k}{|C_1|},\ Int(C_2) + \frac{k}{|C_2|}\right) \tag{3.4}$$

Donde k es un parámetro contante del algoritmo. En el trabajo de Grundmann *et. al.*, este algoritmo de segmentación se extiende para ser utilizado con videos, modelando cada *pixel* del video en un espacio $3D$ con 26 vecinos. Con el fin de generar jerarquías en la segmentación, el algoritmo comienza sobre-segmentando el video, forzando que cada segmento inicial tenga un mínimo de 100 vértices uniendo las aristas de bajo costo. Luego para cada segmento, se calcula un descriptor basado en histogramas de color en el espacio *LAB* con 20 *bins*. Finalmente se vuelve a utilizar el algoritmo de Felzenszwalb y Huttenlocher, utilizando como métrica de similitud la distancia chi cuadrada, χ^2, entre los descriptores. Para cada nivel de jerarquía se itera utilizando el mismo procedimiento.

La figura 3.3 muestra resultados del algoritmo de Grundmann *et. al.* utilizando diferentes niveles de jerarquía para un video de la clase *golf* del set de datos *UCF11*.

Como se aprecia en la figura 3.3, el algoritmo de segmentación comienza sobresegmentando cada *frame* y luego une segmentos similares mediante la distancia χ^2 de los histogramas de colores. En cada iteración, los segmentos se vuelven más estables y consistentes a lo largo del tiempo.

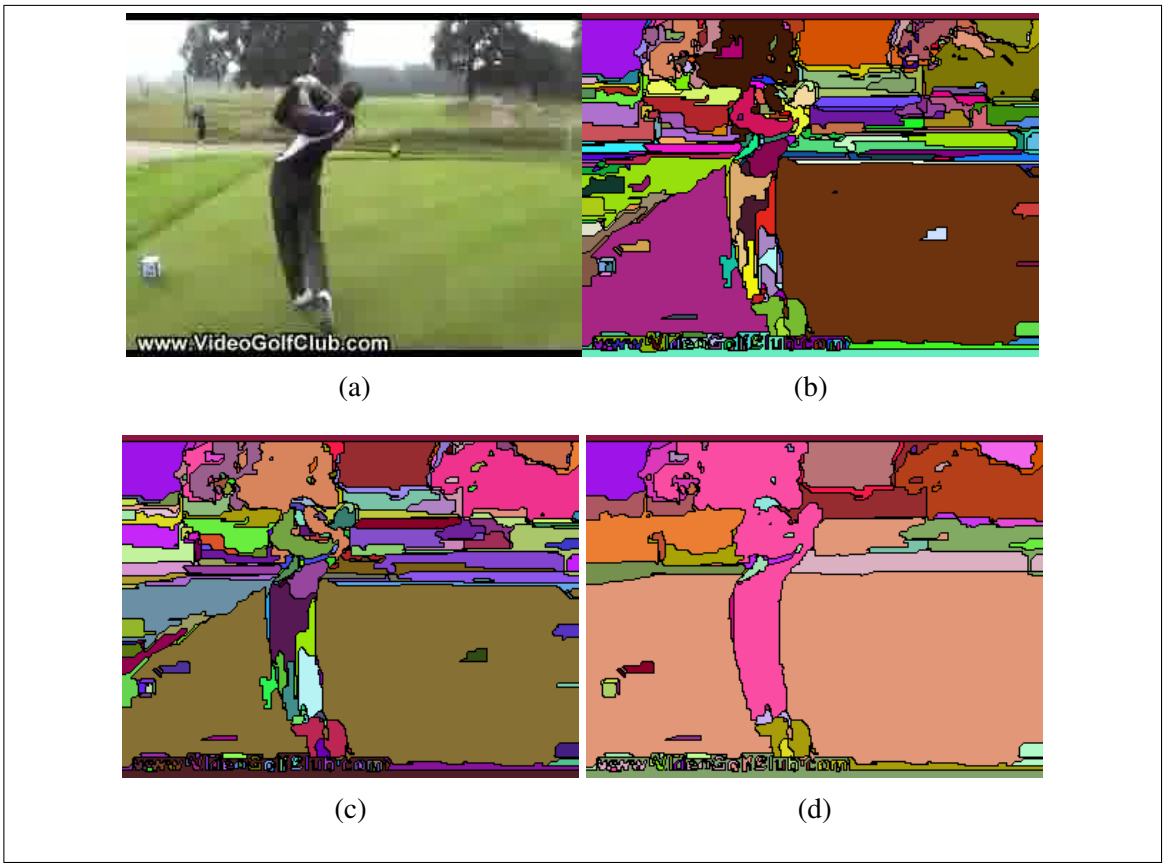


Figura 3.3. Resultado de segmentación utilizando diferentes niveles de jerarquía para un video de la clase *golf* del set de datos *UCF11*. La figura (a) muestra el *frame* original; (b) tiene un nivel de segmentación del 10 %; (c) tiene un nivel de segmentación del 25 %; (d) tiene un nivel de segmentación del 50 %.

CAPÍTULO 4. MÉTODO PROPUESTO

Esta tesis plantea un modelo basado en diccionarios visuales composicionales jerárquicos, que se aprenden de forma secuencial. Estos diccionarios son análogos a los utilizados en la clasificación de textos, en el sentido que generan diferentes niveles de abstracción. El nivel más básico equivale al uso palabras en textos, los que en conjunto forman oraciones, que a su vez forman párrafos, y así sucesivamente hasta completar el texto.

El objetivo de los modelos jerárquico es poder generar abstracciones intermedias entre los *pixeles* y el clasificador final, las cuales sirven para capturar información relevante por cada clase de video. También permite que el modelo sea más robusto ante variabilidades intraclases, mediante la transformación de descriptores a espacios vectoriales, los que permite identificar ciertos patrones que facilitan la clasificación de videos.

La figura 4.1 muestra un diagrama general del método propuesto. En la figura, se especifican los 3 pasos principales: (i) generación del primer nivel de jerarquía, (ii) procesamiento del segundo nivel de jerarquía y (iii) generación del vector característico final. En el primer nivel, se obtienen los descriptores de *Wang* para todos los videos de las diferentes clases, donde cada color del diagrama determina un canal de información diferente. En el segundo nivel, un segmentador de videos determina los volúmenes espacio temporales con los cuales se generan los diccionarios del segundo nivel. Finalmente, se genera el vector característico de cada video utilizando técnicas de *BoW* sobre los descriptores del segundo nivel y son clasificados mediante un modelo de máquinas de vectores de soporte (*SVM*).

En lo que sigue, se describe en detalle los principales pasos del método propuesto para el reconocimiento de acciones humanas. Las secciones 4.1 y 4.2 describen el aprendizaje y abstracción del primer y segundo nivel de la jerarquía, respectivamente. La sección 4.3 describe la forma en que se genera el vector característico por cada video o instancia.

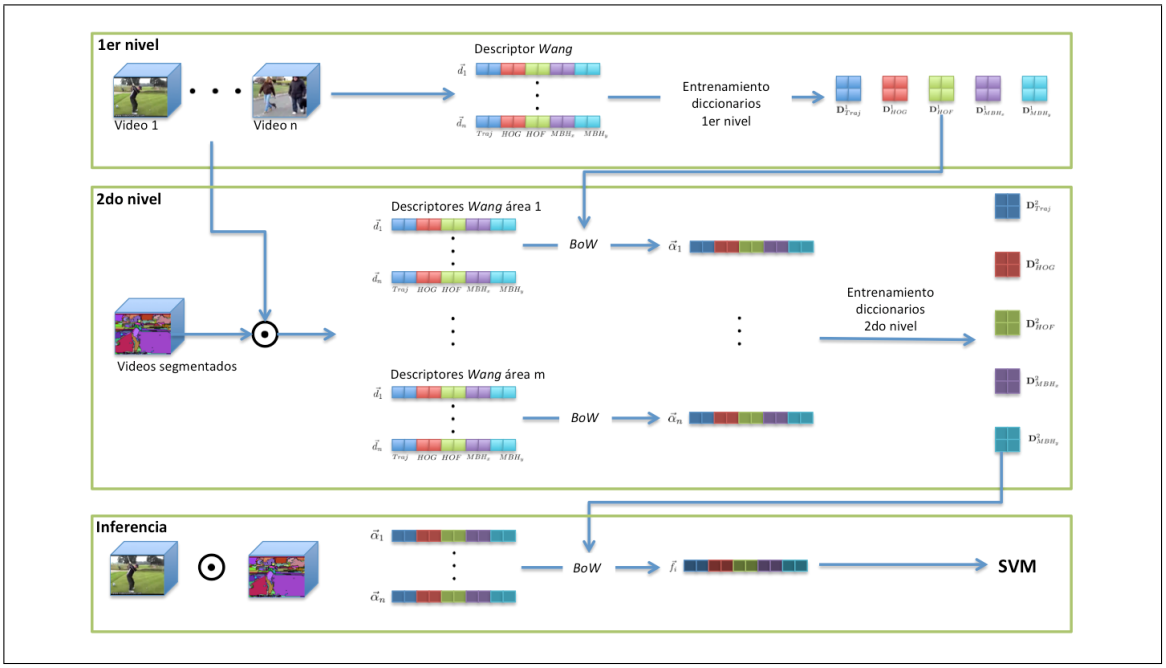


Figura 4.1. Arquitectura del modelo con jerarquía de dos niveles. El primer nivel se compone de los descriptores de *Wang* que son utilizados para generar un diccionario por cada canal de información (*Trajectory*, *HOG*, *HOF*, *MBHx*, *MBHy*). Un segmentador determina los volúmenes espacio temporales del segundo nivel, los que son codificadas utilizando los diccionarios del primer nivel. Nuevamente se genera un diccionario para cada canal de información. El vector característico final es codificado mediante técnicas de *BoW* con los diccionarios del segundo nivel para ser clasificados mediante un modelo *SVM*.

4.1. Primer nivel de jerarquía

El primer nivel de la jerarquía está compuesto por los descriptores espacio-temporales de *Wang*. Con éstos se generan diccionarios visuales, $\mathbf{D}^1$s, que son utilizados para componer los volúmenes espacio-temporales.

Para entrenar los diccionarios de primer nivel se extrajo una muestra de 100.000 descriptores. Dicha muestra se tomó en base a una distribución uniforme sobre una vecindad de 30 *frames* alrededor de los *key frames* de cada video. Los *key frames* de cada video fueron seleccionados mediante el método en (Alfaro, Mery y Soto 2016).

Las dos técnicas que se compararon para la construcción de diccionarios basados en palabras visuales son *sparse coding* y *k-means*.

(I) *Sparse Coding*: para encontrar el grupo de k vectores que mejor reconstruyen los datos iniciales, restringiendo la cantidad de vectores en cada reconstrucción, se debe resolver el problema de optimización en 4.1, donde $\mathbf{D}$ corresponde al diccionario que está siendo optimizado, $\vec{x}_i$ representan los datos de entrenamiento y $\vec{\alpha}_i$ corresponden a los coeficientes de reconstrucción para el dato i sobre el diccionario. El parámetro λ corresponde a un regularizador del tipo *lasso* de modo que cada reconstrucción utilice una cantidad limitada de palabras del diccionario (Lee y col. 2007).

$$\min_{\vec{\alpha}_1,\ldots,\vec{\alpha}_n,\mathbf{D}} \sum_{i=1}^{n} ||\vec{x}_i - \mathbf{D}\vec{\alpha}_i||^2 + \lambda||\vec{\alpha}_i||_1 \tag{4.1}$$

(II) *K-means*: para determinar los k centros de los grupos se debe resolver el problema de optimización en 4.2, done $\mathbf{D}$ es el diccionario producto de la optimización, $\mathbf{D}_i$ corresponde al i-avo $cluster$, $\vec{x}_j$ representan los datos de entrenamiento y $\vec{d}_i$ corresponde a la columna del diccionario y el centro del i-avo $cluster$.

$$\min_{\mathbf{D}} \sum_{i=1}^{k} \sum_{x_j \in \mathbf{D_i}} ||\vec{x}_j - \vec{d}_i||^2 \tag{4.2}$$

Dado que el descriptor de *Wang* se compone de la concatenación de otros descriptores visuales, y cada uno contribuye con diferente tipo de información, se define cada uno de ellos como un canal diferente con su respectivo diccionario visual, es decir, el primer nivel de la jerarquía genera 5 diccionarios diferentes: $\mathbf{D}^1_{TRAY}$, $\mathbf{D}^1_{HOG}$, $\mathbf{D}^1_{HOF}$, $\mathbf{D}^1_{MBHx}$, $\mathbf{D}^1_{MBHy}$.

4.2. Segundo nivel de jerarquía

El segundo nivel introduce los volúmenes espacio-temporales, que corresponden a representaciones vectoriales de los segmentos del video, que se describen en la sección 3.3. Para representar cada segmento, se seleccionan todas las trayectorias de *Wang* obtenidas, que presenten una sobreposición de al menos un 50 % con el volumen espacio-temporal cubierto por el segmento. Se filtran los volúmenes mediante dos criterios: (i) cada región debe tener una estabilidad considerable en el tiempo, lo que implica que debe estar presente en al menos 8 *frames* consecutivos; (ii) para que las regiones sean significativas, cada una debe tener al menos 5 trayectorias. Las trayectorias de *Wang* siguen cada punto de interés a lo largo de 15 *frames*, por lo que al filtrar los volúmenes espacio-temporales menores a 8 *frames* equivale a filtrar todos los volúmenes que no contienen ninguna trayectoria. Por otro lado, al filtrar los volúmenes que contengan menos de 5 trayectorias de *Wang* en los set de datos *UCF11* y *Olympic Sports*, se filtra el 60 % de los segmentos menos representativos, disminuyendo los tiempo de cómputo. La figura 4.2 muestra el proceso de filtro de las trayectorias para un segmento de un video de la clase golf del set de datos *UCF11* [1]

(a) (b) (c)

Figura 4.2. La figura (a) muestra un *frame* con todos sus descriptores. La figura (b) muestra los segmentos obtenidos en dicho *frame*. La figura (c) muestra los descriptores correspondientes al área del torso del golfista.

[1] Jingen Liu, Jiebo Luo y Mubarak Shah (2009), disponible en `http://crcv.ucf.edu/data/UCF_YouTube_Action.php`

Cada volumen es representado mediante técnicas de *BoW* de las trayectorias contenidas en cada una de ellos, utilizando los diccionarios aprendidos en el primer nivel de la jerarquía, para cada canal de información. Dependiendo de la metodología utilizada en el aprendizaje de los diccionarios, la representación se realiza según:

(I) *Sparse coding*: la representación se realiza mediante el uso de *max-pooling* sobre los coeficientes de reconstrucción de las trayectorias dentro del área. Es decir, se reconstruye cada trayectoria con cada diccionario de primer nivel y luego se toma el máximo valor de reconstrucción para cada una de las palabras. Finalmente, el vector es normalizado para que tenga norma ℓ_2 igual a 1.

(II) *K-means*: cada trayectoria es reconstruida sólo por la palabra más cercana en términos de la distancia euclidiana, las que luego son sumadas para generar la representación vectorial. El vector es normalizado para que tenga norma ℓ_2 igual a 1.

Independientemente del método que se utilice, cada área queda asociada a un vector por cada canal de información del mismo tamaño que la cantidad palabras del diccionario de primer nivel.

Una vez que se tienen los descriptores por cada área del video, se toma una muestra de 100.000 áreas, entrenándose un diccionario, al igual que en el primer nivel de jerarquía, por cada canal: $\mathbf{D}^2_{TRAJ}$, $\mathbf{D}^2_{HOG}$, $\mathbf{D}^2_{HOF}$, $\mathbf{D}^2_{MBHx}$, $\mathbf{D}^2_{MBHy}$. Con el propósito de ejecutar el algoritmo de forma eficiente, se reduce la dimensión de los vectores de cada volumen por cada canal mediante el método de Análisis de Componentens Principales (PCA).

4.3. Vector característico

Una vez obtenidos los diccionarios del primer y segundo nivel, éstos son utilizados en la generación de un vector característico para cada video, el que será utilizado para su clasificación.

Para obtener el vector característico de cada video, primero se debe generar un vector por cada segmento que cumpla con las condiciones de estabilidad y significancia explicados anteriormente, esto es, todo segmento que esté presente en 8 *frames* consecutivos y que tenga al menos 5 trayectorias en su interior, respectivamente. Luego se calcula un *BoW* utilizando *average pooling*, es decir, se suman todos los coeficientes de reconstrucción sobre los diccionarios de segundo nivel. Finalmente cada vector es normalizado para que tenga norma ℓ_2 igual a 1. Como modelo de clasificación se utiliza *Support Vector Machine* (SVM) con un *kernel* lineal sobre los descriptores finales.

CAPÍTULO 5. RESULTADOS EXPERIMENTALES CUANTITATIVOS

5.1. Set de datos

Para probar el método de forma cuantitativa se utilizan 2 set de datos:

(I) *UCF11*[1]: introducido el año 2009, contiene 11 acciones obtenidas de *YouTube*. Este set se caracteriza por presentar un alto grado de movimiento de cámaras, variaciones de poses y escalas de los objetos, cambios de luminosidad y puntos de vista. Las distintas clases de acciones y videos se pueden ver en la figura 5.1. Al igual que en (Liu, Luo y M. Shah 2009), se generaron 25 grupos aleatorios de videos considerando la misma cantidad de videos por acciones. Luego, se utilizaron 24 grupos para entrenar el modelo y 1 para realizar pruebas de inferencia.

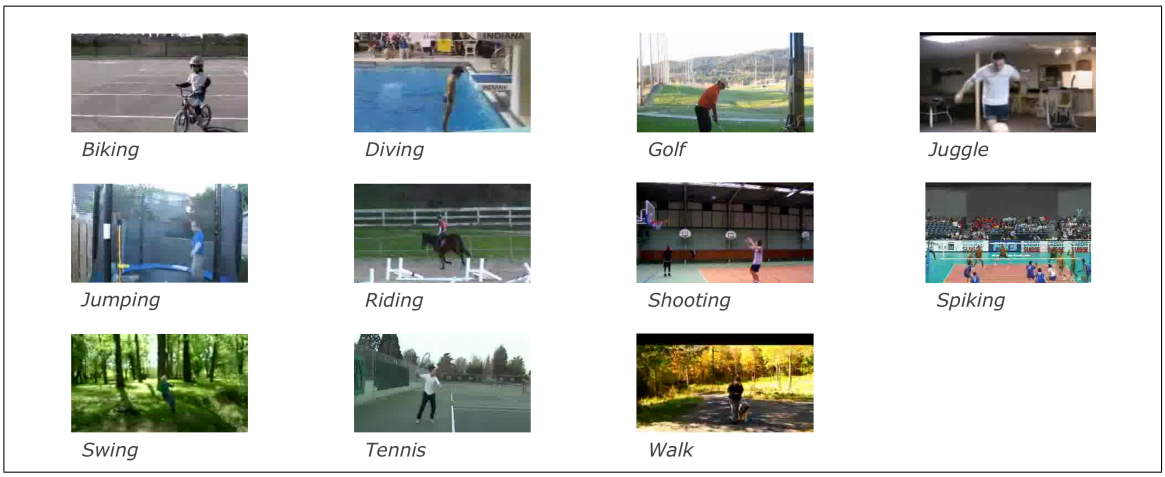


Figura 5.1. Los videos del set de datos *UCF11* se componen de 11 acciones diferentes.

(II) *Olympic Sports*[2]: confeccionado el año 2010, es una colección de videos de personas realizando deportes olímpicos. Los videos fueron obtenidos directamente de la plataforma *YouTube* y fueron rotulados mediante la herramienta *Amazon*

[1] `http://crcv.ucf.edu/data/UCF_YouTube_Action.php`
[2] `http://vision.stanford.edu/Datasets/OlympicSports/`

Mechanical Turk. La figura 5.2 muestra ejemplos de los videos. Al igual que en (Niebles, Chen y Fei-Fei 2010), del total de 743 videos, se utilizaron 649 para entrenar el modelo y 134 para realizar pruebas de inferencia.

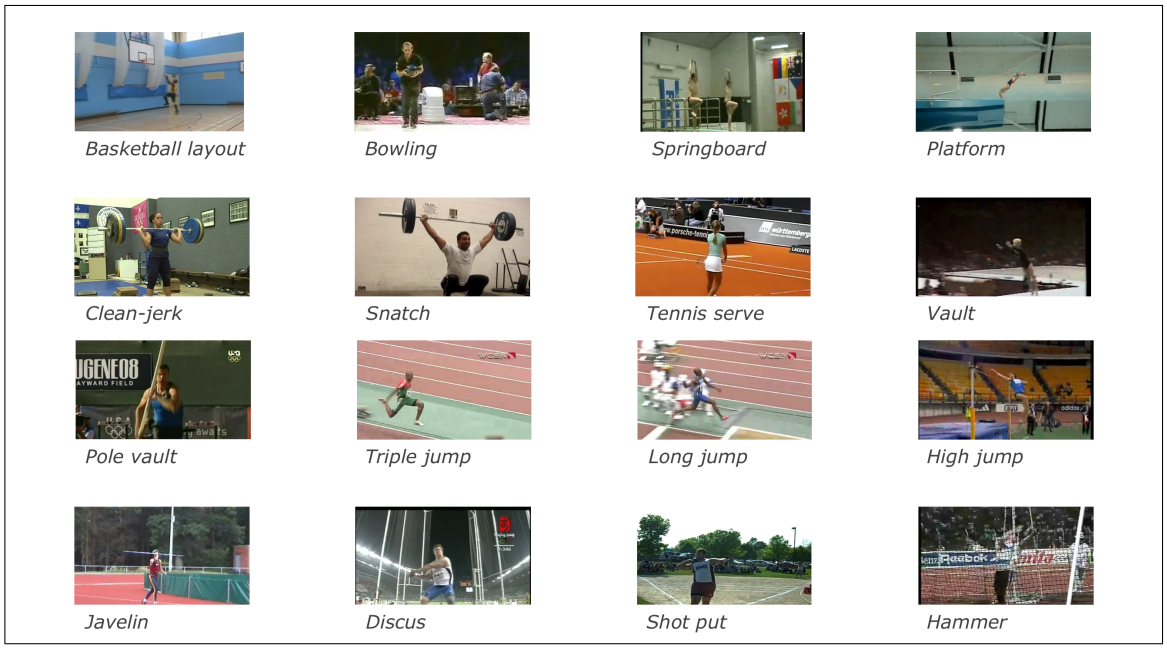


Figura 5.2. Los videos del set de datos *Olympic Sports* se componen de 16 acciones diferentes.

La tabla 5.1 muestra una descripción de los set de datos. En la tabla se observa que los videos pertenecientes al set de datos *UCF11* son de *frames* más pequeños y de menor duración, lo que hace que el tiempo de procesamiento promedio de los videos sea menor. Por otro lado, el set de datos *Olympic Sports* tiene menos videos y mayor cantidad de clases, generando una mayor dificultad debido a la variabilidad de las clases y las pocas instancias de entrenamiento.

Tabla 5.1. Descripción de los set de datos.

Set de datos	$\#(Videos)$	$\#(Clases)$	$\frac{\#(Frames)}{\#(Video)}$	Tamaño Frame (*pixeles*)
UCF11	1,600	11	$191,3$	320×240
Olympic Sports	743	16	$411,9$	688×517

5.2. Resultados cuantitativos

5.2.1. Resultados generales

La tabla 5.10 muestra un resumen de los mejores resultados, utilizando precisión como métrica de evaluación, en la clasificación realizada utilizando cada nivel de jerarquía por separado; utilizando la concatenación de ambos niveles; y el modelo con su configuración óptima, es decir, la configuración de parámetros que entrega mejores resultados (ver tabla 5.11). La misma tabla muestra los mejores parámetros, k_1 y k_2, obtenidos de los diferentes experimentos al utilizar todos los canales de información, excepto en la configuración óptima que utiliza la mejor combinación de los canales. Para el set de datos *UCF11* la combinación óptima de canales de información es *HOG*, *HOF*, *MBHx* y *MBHy* al utilizar *k-means* y *sparse coding*. Al utilizar *k-means* sobre el set de datos *Olympic Sports*, la mejor combinación de canales corresponde a *Trajectory*, *HOG*, *MBHx* y *MBHy*; mientras que al utilizar *sparse coding*, la mejor configuración se obtiene al utilizar *HOG*, *HOF*, *MBHx* y *MBH*.

La métrica de precisión corresponde al porcentaje de instancias clasificadas correctamente. La variante del modelo *Nivel 1* corresponde a una representación *BoW* de los descriptores de *Wang* sobre los diccionarios del primer nivel, concatenando los diferentes canales de información. La variante *Nivel 2* es generada utilizando técnicas de *BoW* de las regiones espacio-temporal sobre los diccionarios del segundo nivel. En la variante *Nivel 1 + Nivel 2* se concatenan los descriptores de ambos niveles. Finalmente, la variante de configuración óptima utiliza los descriptores y niveles descritos en la tabla 5.11.

Para ambos set de datos, el mejor resultado es obtenido utilizando *k-means* como técnica de aprendizaje, concatenando los vectores característicos del primer y segundo nivel (ver fila "Nivel 1 + Nivel 2"de la tabla 5.10). Se observa que el uso del segundo nivel de jerarquía incorpora información que complementa la información del primer nivel. En el set de datos *UCF11*, el uso del segundo nivel de la jerarquía, por sí solo, mostró ser mejor que el primer nivel, al utilizar *sparse coding* para el aprendizaje de los diccionarios.

Tabla 5.2. Mejores resultados en precisión en los set de datos *UCF11* y *Olympic Sports*, para ambos niveles de la jerarquía, como también concatenados.

Modelo	*UCF11*		*Olympic Sports*	
	K-means	*Sparse Coding*	*K-means*	*Sparse Coding*
Precisión				
Nivel 1	89,69 %	83,33 %	73,11 %	71,10 %
Nivel 2	87,27 %	90,30 %	67,98 %	64,60 %
Nivel 1 + Nivel 2	90,90 %	89,69 %	74,51 %	75,57 %
Conf. Óptima	**95,15 %**	**95,15 %**	**76,68 %**	75,57 %
Parámetros				
Nivel 1	$k_1 = 500$	$k_1 = 500$	$k_1 = 500$	$k_1 = 500$
Nivel 2	$k_1 = 500$	$k_1 = 250$	$k_1 = 250$	$k_1 = 500$
	$k_2 = 500$	$k_2 = 500$	$k_2 = 500$	$k_2 = 500$
Nivel 1 + Nivel 2	$k_1 = 500$	$k_1 = 250$	$k_1 = 500$	$k_1 = 500$
	$k_2 = 250$	$k_2 = 500$	$k_2 = 500$	$k_2 = 500$
Conf. Óptima	$k_1 = 500$	$k_1 = 500$	$k_1 = 500$	$k_1 = 500$
	$k_2 = 250$	$k_2 = -$	$k_2 = -$	$k_2 = 500$

5.2.2. Análisis de sensibilidad sobre tamaños de diccionarios

La tabla 5.7 muestra la precisión del modelo utilizando ambos niveles de la jerarquía para diferentes valores de k_1 y k_2, siendo éstos los tamaños de los diccionarios de primer y segundo nivel, respectivamente. Se observa que para ambos niveles de jerarquía, siempre es mejor el uso de diccionarios con más palabras visuales, dado que así el modelo puede representar de mejor manera cada instancia, teniendo un espacio de búsqueda mayor.

En el segundo nivel de jerarquía, para diccionarios pequeños ($k_2 = 250$), el uso de *sparse coding* resulta ser mas efectivo que *k-means* A diferencia de *k-means*, *sparse coding* puede reconstruir mejor los datos con menos cantidad de palabras. Al considerar el segundo nivel utilizando *k-means*, se observa que al aumentar la cantidad de palabras visuales se obtienen mejores resultados respecto al uso de *sparse coding*.

Tabla 5.3. Precisión en set de datos *UCF11* y *Olympic Sports* variando los tamaños k_1 y k_2 de los diccionarios.

		K-means			*Sparse coding*		
		Nivel 1	Nivel 2		Nivel 1	Nivel 2	
		-	$k_2 = 250$	$k_2 = 500$	-	$k_2 = 250$	$k_2 = 500$
UCF11	$k_1 = 250$	77,22 %	74,69 %	74,84 %	76,88 %	88,48 %	**90,30 %**
	$k_1 = 500$	**89,69 %**	73,78 %	87,27 %	83,33 %	82,42 %	82,42 %
Olympic Sports	$k_1 = 250$	72,42 %	55,55 %	67,98 %	67,30 %	58,74 %	60,67 %
	$k_1 = 500$	**73,11 %**	55,43 %	64,14 %	**71,10 %**	62,13 %	64,60 %

La tabla 5.8 muestra la precisión en la clasificación al concatenar los descriptores del primer y segundo nivel. La diferencia promedio entre los resultados de la tabla 5.7 implementando sólo el segundo nivel de jerarquía, y la tabla 5.8 al utilizar *k-means*, son de $12,61\,\%$ y $12,14\,\%$ sobre los set de datos *UCF11* y *Olympic Sports*, respectivamente. Mientras que la diferencia promedio al utilizar *sparse coding* es de $1,97\,\%$ y $8,66\,\%$ en los mismos set de datos. Esto sugiere que el primer y segundo nivel no se complementan al utilizar *sparse coding*.

Tabla 5.4. Precisión en set de datos *UCF11* y *Olympic Sports* concatenando descriptores del nivel 1 y 2 para diferentes valores de k_1 y k_2.

		K-means		*Sparse coding*	
		$k_2 = 250$	$k_2 = 500$	$k_2 = 250$	$k_2 = 500$
UCF11	$k_1 = 250$	90,45 %	89,69 %	89,69 %	89,69 %
	$k_1 = 500$	90,90 %	90,00 %	83,78 %	88,18 %
Olympic Sports	$k_1 = 250$	72,17 %	72,49 %	68,27 %	71,08 %
	$k_1 = 500$	72,52 %	74,51 %	65,87 %	75,57 %

5.2.3. Análisis de ablación sobre canales de información

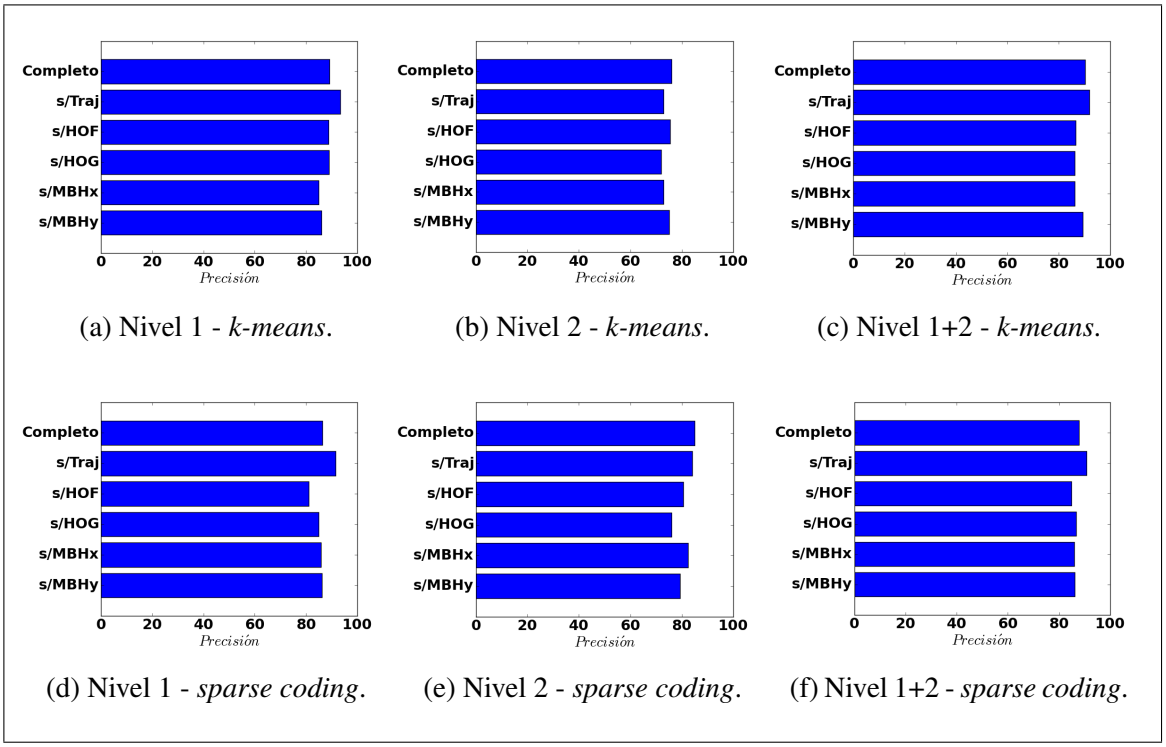
La figura 5.3 muestra los resultados de un test de ablación realizado sobre el set de datos *UCF11*, utilizando los descriptores del primer y segundo nivel de manera independiente y de manera concatenada. El análisis se realizó sobre el promedio de los resultados

experimentales al variar los tamaños de los diccionarios k_1 y k_2, es decir, se realizaron 4 experimentos asignando valores de 250 y 500 para cada tamaño de los diccionarios, manteniendo el valor $\lambda = 0,3$ en el caso de *sparse coding*. Al analizar los promedios del test de ablación se descartan particularidades de la configuración y proporciona información general del modelo.

El test de ablación consiste en extraer partes del modelo de forma independiente y comparar la métrica de precisión en cada caso con el modelo completo. Particularmente, se utilizó este análisis sobre los diferentes canales de información de los descriptores utilizados, es decir, se probó el modelo utilizando todos los descriptores, y luego sacando los descriptores *Trajectory*, *HOF*, *HOG*, *MBHx* y *MBHy*, de manera independiente en todos los niveles de la jerarquía. Al concatenar los descriptores del primer y segundo nivel, los canales son extraído de ambos niveles.

Para ambos métodos de aprendizaje se observa que en el primer nivel los descriptores de Trayectoria proporcionan información contradictoria respecto al resto de los canales. Esto implica que al ser excluido del modelo la precisión aumenta. Este comportamiento no se observa en el segundo nivel, donde las trayectorias generan abstracciones que ayudan a la clasificación de acciones.

También se observa que la concatenación de los descriptores del primer y segundo nivel incorporan robustez al modelo dado que las variaciones de las figuras (c) y (f) son menores que el resto. Un ejemplo claro de este fenómeno se observa al analizar el aporte del canal *HOF* sobre el primer nivel al utilizar *sparse coding*, el cual al ser removido del modelo la precisión disminuye considerablemente. Dicha disminución es menor al concatenar los descriptores del primer y segundo nivel.



(a) Nivel 1 - *k-means*. (b) Nivel 2 - *k-means*. (c) Nivel 1+2 - *k-means*.

(d) Nivel 1 - *sparse coding*. (e) Nivel 2 - *sparse coding*. (f) Nivel 1+2 - *sparse coding*.

Figura 5.3. Resultados de test de ablación en el set de datos *UCF11* sobre diferentes niveles de jerarquía y diferentes métodos. Los gráficos muestran el comportamiento de la métrica de precisión para cada caso.

5.2.4. Configuración óptima del modelo

La tabla 5.11 muestra la mejor configuración de descriptores y tamaños de diccionario para todas las variantes del método y para cada set de datos.

En base a la tabla, se observa que los descriptores *Trajectory* son los que menos se utilizan en las configuraciones óptima, por lo que no son buenos para describir acciones humanas. Dado que los descriptores de *Wang* se calculan de forma densa, existen muchas palabras visuales en los diccionarios de trayectoria que pertenecen al fondo de cada acción que son detectados y que incorporan información contradictoria al resto de los canales, los cuales se centran en las acciones humanas. Este fenómeno es explicado en la sección 5.3.3.

Al incorporar el segundo nivel, se observa que *sparse coding* requiere de tamaños de diccionarios más pequeños y por consiguiente el modelo requiere de menos parámetros.

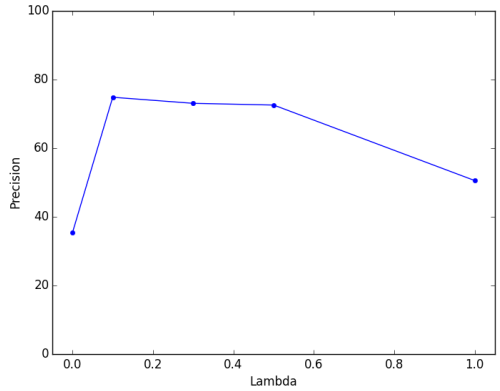
Esto se debe porque *sparse coding* tiene mayor flexibilidad al momento de la reconstrucción de cada dato y puede reconstruir más vectores con menos palabras. Esto conlleva a un compromiso entre cantidad de parámetros y resultados al considerar la misma configuración entre ambos métodos.

Tabla 5.5. Configuración óptima del modelo para cada set de dato. Cada configuración corresponde a la elección de los mejores parámetros para el modelo, tomando en cuenta la cantidad de palabras visuales k_1 y k_2, como también la mejor combinación de canales de información.

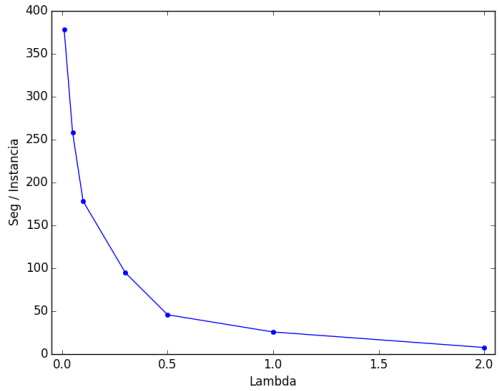
	Método	Descriptores	Tamaño diccionarios	Precisión
UCF11	Nivel 1 [*k-means*]	HOG, HOF, MBHx, MBHy	$k_1 = 500$	94,28 %
	Nivel 2 [*k-means*]	Traj, HOG, HOF, MBHx, MBHy	$k_1 = 500; k_2 = 500$	87,27 %
	Nivel 1 + 2 [*k-means*]	HOG, HOF, MBHx, MBHy	$k_1 = 500; k_2 = 250$	**95,15 %**
UCF11	Nivel 1 [*sparse coding*]	HOG, HOF, MBHx, MBHy	$k_1 = 500$	**95,15 %**
	Nivel 2 [*sparse coding*]	Traj, HOG, HOF, MBHx, MBHy	$k_1 = 250; k_2 = 500$	90,30 %
	Nivel 1 + 2 [*sparse coding*]	HOG, HOF, MBHx, MBHy	$k_1 = 500; k_2 = 250$	92,75 %
Olympic Sports	Nivel 1 [*k-means*]	Traj, HOG, MBHx, MBHy	$k_1 = 500$	**76,68 %**
	Nivel 2 [*k-means*]	Traj, HOG, HOF, MBHx, MBHy	$k_1 = 500; k_2 = 500$	64,14 %
	Nivel 1 + 2 [*k-means*]	HOG, HOF, MBHx, MBHy	$k_1 = 500; k_2 = 500$	75,14 %
Olympic Sports	Nivel 1 [*sparse coding*]	HOG, HOF, MBHx, MBHy	$k_1 = 500$	74,78 %
	Nivel 2 [*sparse coding*]	HOG, HOF, MBHx, MBHy	$k_1 = 250; k_2 = 500$	70,05 %
	Nivel 1 + 2 [*sparse coding*]	HOG, HOF, MBHx, MBHy	$k_1 = 250; k_2 = 250$	**75,57 %**

5.2.5. Análisis de sensibilidad sobre regularizador

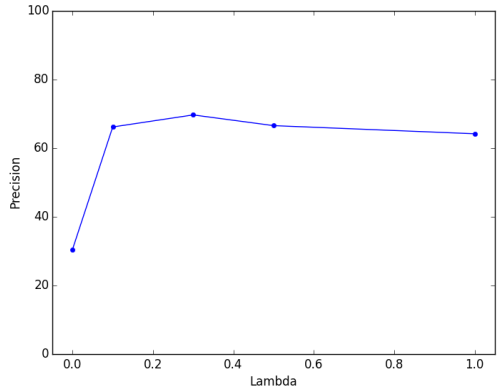
La figura 5.4 muestra un análisis de sensibilidad sobre el parámetro λ en el aprendizaje de los diccionarios de primer nivel utilizando *sparse coding*. Se observa que el tiempo de procesamiento de una instancia decrece con el valor de λ, esto se debe principalmente a que el parámetro controla el nivel de *sparsity* del modelo (grado de densidad de los vectores de reconstrucción). Por consiguiente a menor valor de λ cada instancia se reconstruirá con una mayor cantidad de palabras visuales del diccionario, aumentando así el espacio de búsqueda y ,por consiguiente, el tiempo de procesamiento, mientras que a mayor valor, cada instancia se reconstruirá con menos palabras. El valor óptimo utilizado en los experimentos para ambos set de datos es $\lambda = 0,3$, dado que muestra estabilidad en la precisión y un tiempo de procesamiento manejable por cada instancia.



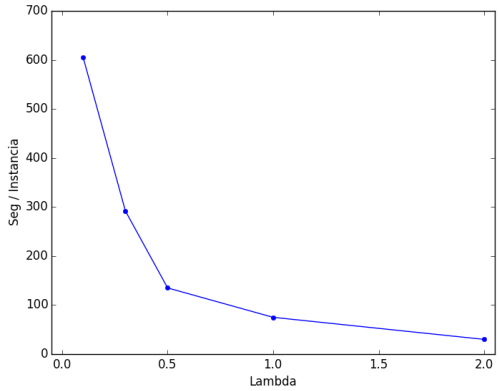
(a) Precisión en el set de datos *UCF11*.



(b) Tiempo de cómputo promedio por video en el set de datos *UCF11*.



(c) Precisión en el set de datos *Olympic Sports*.



(d) Tiempo de cómputo promedio por video en el set de datos *Olympic Sports*.

Figura 5.4. Resultados de análisis de sensibilidad sobre parámetro λ para los modelos de *sparse coding*, en el aprendizaje del 1er nivel, en los set de datos *UCF11* y *Olympic Sports*.

5.2.6. Benchmark y comparación con otros trabajos

En la tabla 5.9 se compara la propuesta de esta tesis con otros algoritmos del estado del arte que comparten elementos con el método propuesto en este trabajo. Para el set de datos *UCF11*, el método presentado, al concatenar los descriptores del primer y segundo nivel utilizando *k-means*, obtiene mejores resultados con una diferencia de $9,75$ puntos,

respecto al trabajo de Wang *et. al.* Sin embargo para el set de datos *Olympic Sports*, el método de esta tesis no mejora ninguno de los resultados base.

Los videos del set de datos *UCF11* se caracterizan por ser videos cortos teniendo 190 *frames* por video, en comparación con el set de datos *Olympic Sports* con 410 *frames* por video. El modelo planteado en esta tesis tienen mejor desempeño en videos cortos, donde la incorporación del tiempo es menos relevante. A medida que los videos tienen más *frames* se vuelve más importante la forma en que el modelo captura información temporal. El modelo de Gaido *et. al.*, hace un buen trabajo incorporando información temporal de forma jerárquica utilizando un árbol de *clusters*, mientras que el modelo planteado divide el video en 3 partes de igual largo, calculando un vector característico para cada parte, para luego ser concatenados formando el vector característico final.

Tabla 5.6. Comparación del trabajo propuesto con métodos del estado del arte.

Método	*UCF11*	*Olympic Sports*
Wang *et. al.* (H. Wang y col. 2011)	85,4 %	77,2 %
Taralova *et. al.* (Taralova, De la Torre y Hebert 2014)	83,8 %	-
Gaidon *et. al.* (Adrien Gaidon, Zaid Harchaoui y Cordelia Schmid 2012)	-	**82,7 %**
Utilizando todos los canales de información		
Nivel 1 [*k-means*]	89,69 %	73,11 %
Nivel 1 [*sparse coding*]	83,33 %	71,10 %
Nivel 2 [*k-means*]	87,27 %	64,14 %
Nivel 2 [*sparse coding*]	90,30 %	64,60 %
Nivel 1 + 2 [*k-means*]	90,90 %	74,51 %
Nivel 1 + 2 [*sparse coding*]	89,69 %	71,08 %
Utilizando parámetros óptimos de la tabla 5.11		
Nivel 1 [*k-means*]	94,28 %	76,68 %
Nivel 1 [*sparse coding*]	95,15 %	74,78 %
Nivel 2 [*k-means*]	87,27 %	64,14 %
Nivel 2 [*sparse coding*]	90,30 %	70,05 %
Nivel 1 + 2 [*k-means*]	**95,15 %**	75,14 %
Nivel 1 + 2 [*sparse coding*]	92,75 %	75,57 %

5.3. Resultados cualitativos

5.3.1. Análisis del uso de palabras visuales

La figura 5.5, a modo de ejemplo, muestra el uso de las palabras de los diccionarios de primer y segundo nivel, de manera independiente, de la clase *biking* del set de datos *UCF11* al utilizar el modelo jerárquico de dos niveles concatenados. Los histogramas representan el uso para diccionarios generados con *k-means* y *sparse coding* por cada canal de información.

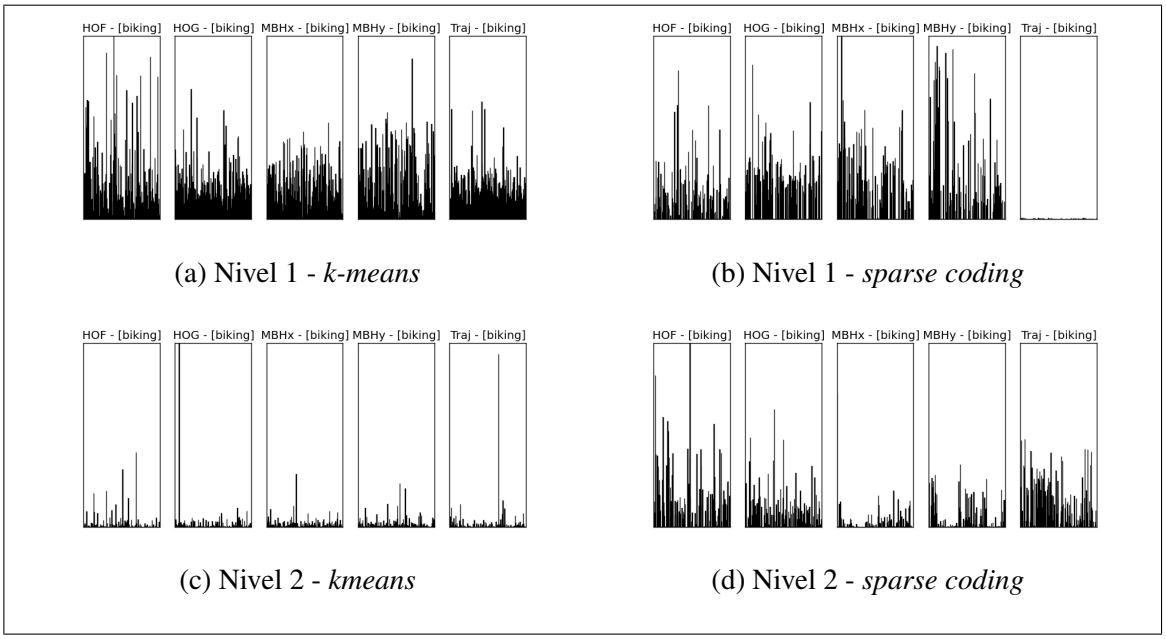
Al considerar el uso de las palabras generadas con *k-means*, se observa que en el segundo nivel el uso de cada una disminuye considerablemente. Esto se debe a la menor cantidad de áreas de segundo nivel en cada video en comparación a las trayectorias de *Wang*. Además, se observa que en el segundo nivel la mayoría de las palabras de los canales *HOG* y *Trajectory* se representan con una sola palabra del diccionario.

Analizando el uso de las palabras generadas mediante *sparse coding*, llama la atención que los descriptores del primer nivel no se representan bien con las palabras del canal *Trajectory*, fenómeno que no se repite en el segundo nivel. Bajo este escenario, *k-means* tiene la obligación de reconstruir cada vector con el *cluster* más cercano, mientras que *sparse coding* puede no reconstruir un vector con un diccionario si sus palabras no son representativas, entregando mayor flexibilidad en la reconstrucción. Esto implicaría que este método es más robusto que al utilizar *k-means*.

Globalmente, el primer nivel muestra un comportamiento uniforme salvo con algunas palabras específicas que se interpretan como palabras visuales más especializadas para la representación de la clase. Mientras que el segundo nivel muestra uso excesivo de pocas palabras, especialmente en la representación con *k-means*. Esto implicaría que el segundo nivel representa patrones más especializados por cada clase, agregando información complementaria a la generada en el primer nivel de la jerarquía.

La figura 5.6 muestra los pesos asignados por el clasificador a cada palabra visual sobre los diccionarios del primer y segundo nivel utilizando *k-means* y *sparse coding*. Para la clase *biking* del set de datos *UCF11*, al utilizar *sparse coding* se observa que el clasificar asigna bajos pesos a los canales menos usados. Esto muestra que *sparse coding* es capaz de filtrar los datos menos relevantes para realizar inferencia.

Las figuras 5.5 y 5.6 muestran cómo *sparse coding* es capaz de filtrar el ruido generado por los descriptores de trayectoria para las representaciones de primer nivel, así como también su capacidad de generar patrones más especializados en niveles de abstracción más altos.



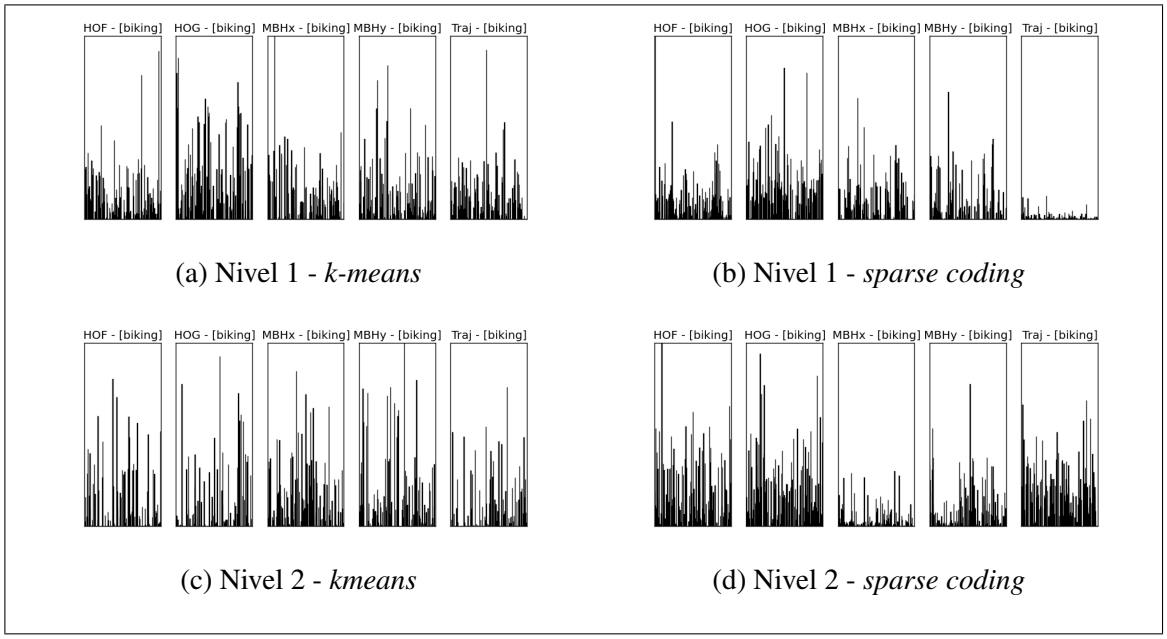
(a) Nivel 1 - *k-means*

(b) Nivel 1 - *sparse coding*

(c) Nivel 2 - *kmeans*

(d) Nivel 2 - *sparse coding*

Figura 5.5. Histograma de uso de palabras sobre el primer y segundo nivel de jerarquía utilizando *k-means* y *sparse coding* sobre la clase *biking* del set de datos *UCF11*.



(a) Nivel 1 - *k-means*

(b) Nivel 1 - *sparse coding*

(c) Nivel 2 - *kmeans*

(d) Nivel 2 - *sparse coding*

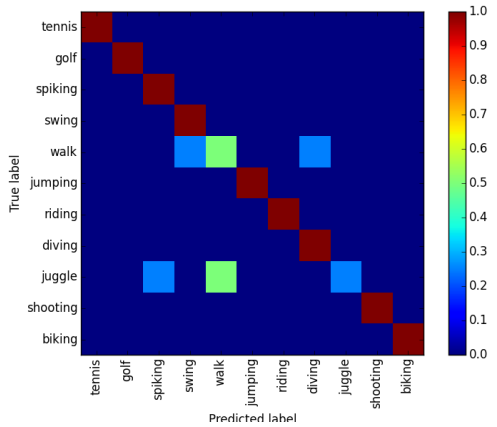
Figura 5.6. Pesos asignados por el clasificador a cada palabra visual utilizando *k-means* y *sparse coding* sobre la clase *biking* del set de datos *UCF11*.

5.3.2. Análisis de matriz de confusión

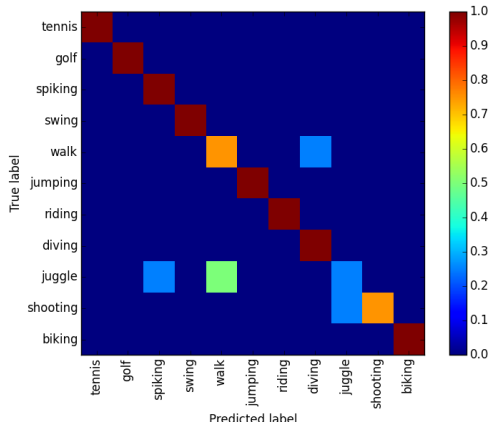
La figura 5.7 muestra la matriz de confusión generada por el modelo en los set de datos *UCF11* y *Olympic Sports* con diccionarios generados con *sparse coding* y *k-means*. En todas las instancias se utilizaron 500 palabras para cada diccionario, concatenando los descriptores del primer y segundo nivel, dado que esta corresponde a la mejor configuración del modelo.

Las matrices del set de datos *UCF11* se muestran muy estables y similares, debido a la alta precisión de ambas variantes. La clase que presenta mayor grado de confusión es *juggle* la cual se confunde con la clase *walk*. Se atribuye esta confusión a que en en las instancias de ambas clases se observan personas de pie con bajo movimiento.

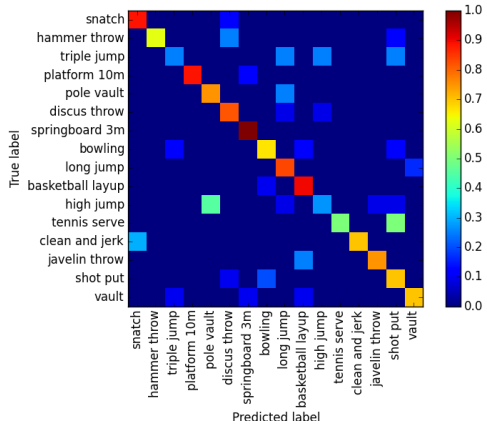
Por otro lado, las matrices asociadas al set de datos *Olympic Sports* muestran un alto grado de complejidad y confusión. Ambos algoritmos de generación de diccionarios muestran dificultad en clasificar acciones cuando estas poseen características similares a



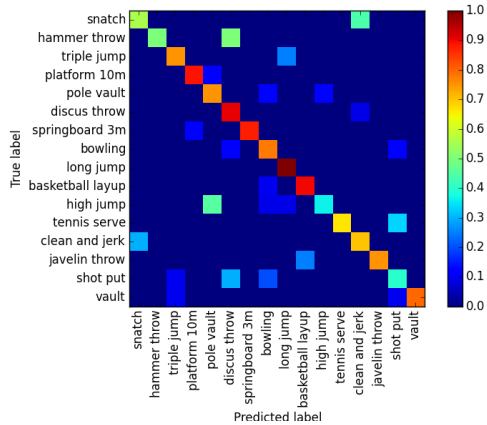
(a) Matriz de confusión sobre el set de datos *UCF11* utilizando *k-means*.



(b) Matriz de confusión sobre el set de datos *UCF11* utilizando *sparse coding*.



(c) Matriz de confusión sobre el set de datos *Olympic Sports* utilizando *k-means*.



(d) Matriz de confusión sobre el set de datos *Olympic Sports* utilizando *sparse coding*.

Figura 5.7. Matrices de confusión utilizando la concatenación de los descriptores del 1er y 2do nivel para los set de datos *UCF11* y *Olympic Sports*.

otras. Las clases donde el modelo tiene menos aciertos son *hight jump* que es confundida mayormente por las clases *pole vault* y *long jump*, en el caso de *sparse coding*. Al utilizar *k-means* la mayor confusión de produce entre las clases *triple jump* y *hight jump*. Éstas clases corresponen a personas corriendo en una pista atlética para realizar un tipo de salto, pudiendo una persona distinguir entre una clase y otra sólo en la parte final de la acción. La figura 5.8 muestra este fenómeno para las clases *high jump*, *long jump* y *triple jump*.

(a) *Frames* de la clase *high jump*.

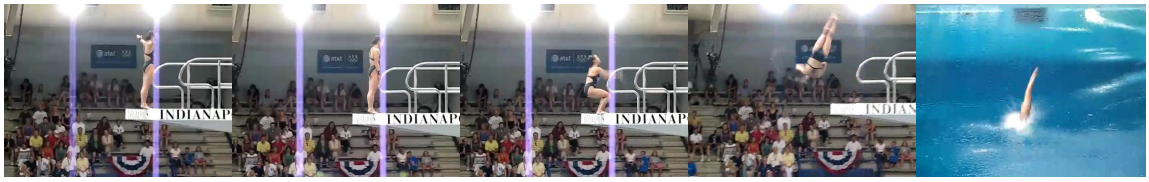
(b) *Frames* de la clase *long jump*.

(c) *Frames* de la clase *triple jump*.

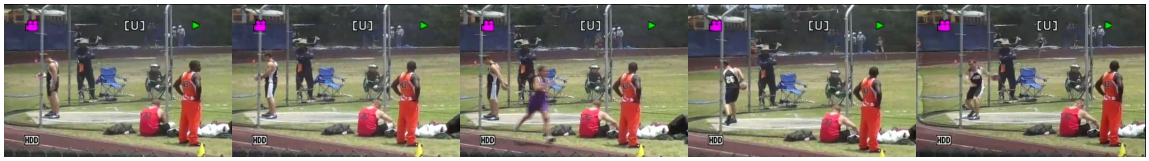
Figura 5.8. Los videos de las clases *high jump*, *long jump* y *triple jump* son visualmente similares y difíciles de distinguir para el ojo humano, y por consiguiente son las que presentan mayor grado de confusión para el modelo.

En contraste, para las clases *basketball lay-up*, *platform 10m* y *discus throw* presentan patrones de movimientos únicos y son fáciles de distinguir para el ojo humano y para el modelo. Esto se obsevra en la matriz de confusión que clasifica correctamente estas instancias. La figura 5.9 muestra *frames* de dichas instancias.

(a) *Frames* de la clase *basketball lay-up*.



(b) *Frames* de la clase *platform 10m*.



(c) *Frames* de la clase *discus throw*.

Figura 5.9. Los videos de las clases *basketball lay-up*, *platform 10m* y *discus throw* presentan patrones únicos y son fáciles de clasificar para el modelo.

5.3.3. Visualización de palabras visuales destacadas

A modo de ejemplo, las figuras 5.10 y 5.11 muestran las palabras más usadas y, con mayor puntaje de clasificación, en las clases *biking* y *tennis*, respectivamente. En cada canal, los descriptores se especializan y representan diferentes aspectos del video. Las figuras muestran que, los descriptores del primer nivel se especializan en el movimiento circular de los pies en el canal *HOF*, en el manubrio en el canal *HOG*, en el casco del ciclista en el canal *MBHx*, en las ruedas en el canal *MBHy* y en el movimiento del fondo en el canal *Trajectory*. Como se ha analizado anteriormente, los descriptores de trayectorias incorporan ruido al modelo, y en este caso se comprueba dado que se enfocan en el fondo del video.

Análogamente, para el segundo nivel sucede algo similar. Cada canal muestra diferentes patrones que permiten la correcta clasificación de la acción del video. En la figura 5.11,

los canales *HOF*, *MBHx*, *MBHy* y *Trajectory* muestran como áreas relevantes aquellas relacionadas al movimiento del jugador y la raqueta, dichas áreas se observan en los canales *HOF*, *MBHx*, *MBHx* y *Trajectory*, mientras que el canal *HOG* representa el movimiento de la parte superior de las piernas.

Tabla 5.7. Precisión en set de datos *UCF11* y *Olympic Sports* variando los tamaños k_1 y k_2 de los diccionarios.

		K-means		*Sparse coding*	
		$k_2 = 250$	$k_2 = 500$	$k_2 = 250$	$k_2 = 500$
UCF11	$k_1 = 250$	74,6 %	74,8 %	88,4 %	90,3 %
	$k_1 = 500$	73,7 %	87,2 %	82,4 %	82,4 %
Olympic Sports	$k_1 = 250$	55,5 %	67,9 %	58,7 %	60,6 %
	$k_1 = 500$	55,4 %	64,1 %	62,1 %	64,6 %

Tabla 5.8. Precisión en set de datos *UCF11* y *Olympic Sports* concatenando descriptores del nivel 1 y 2 para diferentes valores de k_1 y k_2.

		K-means		*Sparse coding*	
		$k_2 = 250$	$k_2 = 500$	$k_2 = 250$	$k_2 = 500$
UCF11	$k_1 = 250$	90,4 %	89,6 %	89,6 %	89,6 %
	$k_1 = 500$	90,9 %	90,0 %	83,7 %	88,1 %
Olympic Sports	$k_1 = 250$	72,1 %	72,4 %	68,2 %	71,0 %
	$k_1 = 500$	72,5 %	74,5 %	65,8 %	75,5 %

(a) HOF

(b) HOG

(c) MBHx

(d) MBHy

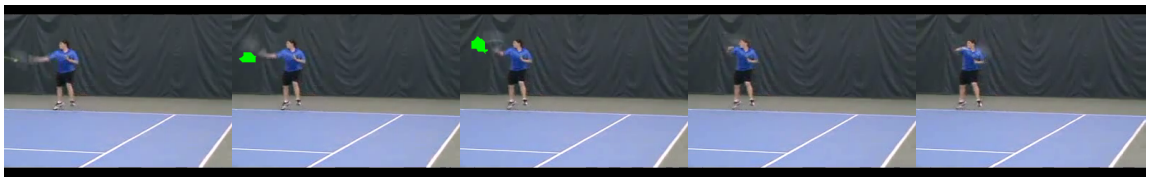
(e) Traj

Figura 5.10. Visualización de las palabras del primer nivel más comunes y con mayor puntaje de clasificación utilizando *k-means* para la clase *biking* del set de datos *UCF*. Se observa que los diferentes canales de información capturan diferentes patrones, y que los descriptores de trayectoria se enfocan en el fondo incorporando ruido al modelo.

(a) HOF

(b) HOG

(c) MBHx



(d) MBHy

(e) Traj

Figura 5.11. Visualización de las palabras más comunes del segundo nivel de jerarquía y con mayor puntaje de clasificación utilizando *k-means* para la clase *tenis* del set de datos *UCF*. Se observa que el segundo nivel de jerarquía captura patrones de movimientos propios de cada acción, como por ejemplo el movimiento de la raqueta, la parte superior de las piernas, etc. En este caso, las trayectorias dejan de ser ruido y forman patrones de movimientos complejos.

Tabla 5.9. Comparación del trabajo propuesto con métodos del estado del arte.

Método	*UCF11*	*Olympic Sports*
Wang *et. al.* (H. Wang y col. 2011)	85,4 %	77,2 %
Taralova *et. al.* (Taralova, De la Torre y Hebert 2014)	83,8 %	-
Gaidon *et. al.* (Adrien Gaidon, Zaid Harchaoui y Cordelia Schmid 2012)	-	**82,7 %**
Utilizando todos los canales de información		
Nivel 1 + 2 [*k-means*]	90,9 %	74,5 %
Nivel 1 + 2 [*sparse coding*]	89,6 %	71,0 %
Utilizando los mejores parámetros		
Nivel 1 + 2 [*k-means*]	**95,1 %**	75,1 %
Nivel 1 + 2 [*sparse coding*]	92,7 %	75,5 %

Tabla 5.10. Mejores resultados en precisión en los set de datos *UCF11* y *Olympic Sports*, para ambos niveles de la jerarquía, como también concatenados.

Modelo	*UCF11*		*Olympic Sports*	
	K-means	*Sparse Coding*	*K-means*	*Sparse Coding*
Precisión				
Nivel 1	89,69 %	83,33 %	73,11 %	71,10 %
Nivel 2	87,27 %	90,30 %	67,98 %	64,60 %
Nivel 1 + Nivel 2	90,90 %	89,69 %	74,51 %	75,57 %
Mejor Conf.	**95,15 %**	**95,15 %**	**76,68 %**	75,57 %

Tabla 5.11. Configuración óptima del modelo para cada set de dato. Cada configuración corresponde a la elección de los mejores parámetros para el modelo, tomando en cuenta la cantidad de palabras visuales k_1 y k_2, como también la mejor combinación de canales de información.

	Método	Variante	k_1	k_2	Traj	HOG	HOF	MBHx	MBHy	Precisión
UCF11	*k-means*	Nivel 1 + Nivel 2	500	250	-	✓	✓	✓	✓	95,1 %
	sparse coding	Nivel 1	500	-	-	✓	✓	✓	✓	95,1 %
Olympic Sports	*k-means*	Nivel 1	500	-	✓	✓	✓	✓	✓	76,6 %
	sparse coding	Nivel 1 + Nivel 2	250	250	-	✓	✓	✓	✓	75,5 %

CAPÍTULO 6. CONCLUSIONES Y TRABAJOS FUTUROS

En esta tesis se propone el uso de un modelo basado en representaciones intermedias mediante el uso de diccionarios jerárquicos para la clasificación de acciones en videos. Dicho modelo fue implementado sobre dos bases de datos utilizando técnicas de *k-means* y *sparse coding*. Los resultados experimentales muestran que al utilizar los descriptores del primer y segundo nivel, de forma conjunta, el modelo adquiere un alto grado de robusteza frente al cambio en los tamaños de los diccionarios. Los resultados de la tabla 5.7 muestra que al variar los tamaños de los diccionarios utilizando los descriptores del primer y segundo nivel de forma independiente, la métrica de precisión varía considerablemente, pero al utilizar los descriptores de manera conjuntamente, concatenados, la precisión es poco sensible ante cambios en los tamaños de los diccionarios, como se muestra en la tabla 5.8.

No obstante lo anterior, los mejores resultados fueron por lo general obtenidos utilizando un solo nivel de jerarquía, afinando los tipos de descriptores que se utilizaron, como también la cantidad de palabras visuales. Este fenómenos se observó en ambos set de datos y utilizando ambos métodos de aprendizaje.

Se observó empíricamente que el modelo tiene mejor rendimiento utilizando *k-means*, en vez de *sparse coding*, para ambos set de datos. Por otro lado, se comprobó que el uso de un segundo nivel de jerarquía incorpora información discriminativa que mejora el desempeño. Sin embargo, el uso de dicha información es complementaria a un modelo de un nivel y debe ser usado de forma conjunta. Esto indica que el primer nivel captura información de bajo nivel, mientras que el segundo nivel detecta patrones de acciones de más alto nivel, por ejemplo, movimientos de brazos, torsos y piernas propios de cada clase de acciones.

Una de las ventajas del modelo propuesto, es que es agnóstico al uso de los descriptores base que se utilizan. También se puede observar e interpretar las representaciones

generadas en cada nivel de la jerarquía como fue descrito en el capítulo 5, donde se muestra cómo el segundo nivel de jerarquía captura movimientos humanos propios de cada acción mediante la codificación de volúmenes espacio temporales determinados por un segmentador jerárquico aplicado en videos. Dichos volúmenes muestran movimientos de torsos, brazos, piernas, etc.

El método propuesto obtuvo resultados del estado del arte para el set de datos *UCF11* con una precisión de $95,15\,\%$, mejorando en $9,75$ puntos porcentuales con respecto a estos. En el set de datos *Olympic Sports*, se obtuvo una precisión de $76,68\,\%$, 6 puntos porcentuales por debajo de los resultados en el estado del arte. Este último resultado se entiende al considerar que el método propuesto incorpora información espacial con ventanas de tiempo fijas, dividiendo cada video en 3 ventanas, y luego concatenando los vectores característicos, mientras que el estado del arte utiliza información temporal jerárquica, detectando los patrones de cada acción en el tiempo.

Con el propósito de mejorar el rendimiento del modelo, se pueden tomar varias acciones. Una de ellas es incorporar información del clasificador en la representación intermedia, generando un entrenamiento *end-to-end*, con el propósito de mejorar las representaciones. Resulta interesante también explorar la incorporación de regularizadores del tipo *Group Sparsity*, los que obligarían al modelo aprender simultáneamente conjunto de palabras visuales especializadas para cada clase. Con el objeto de, se podría trabajar con diferentes técnicas en la generación de los volúmenes espacio-temporales.

Una de las fortalezas del modelo propuesto es ser fácilmente extendible a un mayor numero de niveles de abstracción, acercándolo a la linea de aprendizaje profundo. Por supuesto, esta es una linea de trabajo futuro a explorar.

REFERENCIAS

Aggarwal, Jake K y Michael S Ryoo: *Human activity analysis: A review*. En: *ACM Computing Surveys (CSUR)* 43.3 (2011), pág. 16.

Alfaro, A., D. Mery y A. Soto: "Action Recognition in Video Using Sparse Coding and Relative Features". En: *CVPR*. 2016. URL: http://saturno.ing.puc.cl/media/papers_alvaro/FinalVersion-Anali-CVPR-2016.pdf.

Boureau, Y-Lan y col.: "Learning mid-level features for recognition". En: *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*. IEEE, 2010, págs. 2559-2566. DOI: http://dx.doi.org/10.1109/CVPR.2010.5539963.

Campbell, L. W. y A. F. Bobick: "Recognition of human body motion using phase space constraints". En: *Proceedings of IEEE International Conference on Computer Vision*. Jun. de 1995, págs. 624-630. DOI: 10.1109/ICCV.1995.466880.

Chaudhry, R. y col.: "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions". En: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Jun. de 2009, págs. 1932-1939. DOI: 10.1109/CVPR.2009.5206821.

Dalal, Navneet, Bill Triggs y Cordelia Schmid: "Human Detection Using Oriented Histograms of Flow and Appearance". En: *Proceedings of the 9th European Conference on Computer Vision - Volume Part II*. ECCV'06. Graz, Austria: Springer-Verlag, 2006, págs. 428-441. ISBN: 3-540-33834-9, 978-3-540-33834-5. DOI: 10.1007/11744047_33. URL: http://dx.doi.org/10.1007/11744047_33.

— "Human detection using oriented histograms of flow and appearance". En: *European conference on computer vision*. Springer. 2006, págs. 428-441.

Dollar, P. y col.: “Behavior recognition via sparse spatio-temporal features”. En: *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. Oct. de 2005, págs. 65-72. DOI: 10.1109/VSPETS.2005.1570899.

Elhamifar, E., G. Sapiro y R. Vidal: “See all by looking at a few: Sparse modeling for finding representative objects”. En: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. Jun. de 2012. DOI: 10.1109/CVPR.2012.6247852.

Felzenszwalb, Pedro F. y Daniel P. Huttenlocher: *Efficient Graph-Based Image Segmentation*. En: *Int. J. Comput. Vision* 59.2 (sep. de 2004), págs. 167-181. ISSN: 0920-5691. DOI: 10.1023/B:VISI.0000022288.19776.77. URL: http://dx.doi.org/10.1023/B:VISI.0000022288.19776.77.

Gaidon, A., Z. Harchaoui y C. Schmid: “Actom sequence models for efficient action detection”. En: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. Jun. de 2011, págs. 3201-3208. DOI: 10.1109/CVPR.2011.5995646.

Gaidon, Adrien, Zaid Harchaoui y Cordelia Schmid: “Recognizing activities with cluster-trees of tracklets”. En: *BMVC 2012 - British Machine Vision Conference*. Ed. por Richard Bowden, John P. Collomosse y Krystian Mikolajczyk. Guildford, United Kingdom: BMVA Press, sep. de 2012, págs. 30.1-30.13. DOI: 10.5244/C.26.30. URL: https://hal.inria.fr/hal-00722955.

Grundmann, Matthias y col.: *Efficient Hierarchical Graph Based Video Segmentation*. En: *IEEE CVPR* (2010).

Guha, T. y R. K. Ward: *Learning Sparse Representations for Human Action Recognition*. En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.8 (ago. de 2012), págs. 1576-1588. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2011.253.

Hu, N. y col.: “Learning latent structure for activity recognition”. En: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. Mayo de 2014, págs. 1048-1053. DOI: 10.1109/ICRA.2014.6906983.

Huang, Ke y Selin Aviyente: “Sparse representation for signal classification”. En: *NIPS*. Vol. 19. 2006, págs. 609-616.

Klaser, Alexander, Marcin Marszałek y Cordelia Schmid: “A spatio-temporal descriptor based on 3d-gradients”. En: *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association. 2008, págs. 275-1.

Laptev, Ivan: *On Space-Time Interest Points*. En: *Int. J. Comput. Vision* 64.2-3 (sep. de 2005), págs. 107-123. ISSN: 0920-5691. DOI: 10.1007/s11263-005-1838-7. URL: http://dx.doi.org/10.1007/s11263-005-1838-7.

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

Lee, Honglak y col.: *Efficient sparse coding algorithms*. En: *Advances in neural information processing systems* 19 (2007), pág. 801.

Leung, Thomas y Jitendra Malik: *Representing and Recognizing the Visual Appearance of Materials Using Three-dimensional Textons*. En: *Int. J. Comput. Vision* 43.1 (jun. de 2001), págs. 29-44. ISSN: 0920-5691. DOI: 10.1023/A:1011126920638. URL: http://dx.doi.org/10.1023/A:1011126920638.

Liu, Jingen, Jiebo Luo y Mubarak Shah: *Recognizing Realistic Actions from Videos ïn the Wild"*. En: *IEEE International Conference on Computer Vision and Pattern Recognition* (2009). URL: http://server.cs.ucf.edu/~vision/projects/liujg/realistic_action_recognition.html.

Mairal, Julien y col.: *Supervised Dictionary Learning*. 2008.

Niebles, Juan Carlos, Chih-Wei Chen y Li Fei-Fei: “Modeling temporal structure of decomposable motion segments for activity classification”. En: *European conference on computer vision*. Springer. 2010, págs. 392-405.

Niyogi, S. A. y E. H. Adelson: “Analyzing and recognizing walking figures in XYT”. En: *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Jun. de 1994, págs. 469-474. DOI: 10.1109/CVPR.1994.323868.

Peng, Xiaojiang, Limin Wang, Yu Qiao y col.: “A Joint Evaluation of Dictionary Learning and Feature Encoding for Action Recognition”. En: *ICPR*. 2014, págs. 2607-2612.

Peng, Xiaojiang, Limin Wang, Xingxing Wang y col.: *Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice*. En: *Computer Vision and Image Understanding* 150 (2016), págs. 109-125. ISSN: 1077-3142. DOI: http://dx.doi.org/10.1016/j.cviu.2016.03.013. URL: http://www.sciencedirect.com/science/article/pii/S1077314216300091.

Poppe, Ronald: *A Survey on Vision-based Human Action Recognition*. En: *Image Vision Comput.* 28.6 (jun. de 2010), págs. 976-990. ISSN: 0262-8856. DOI: 10.1016/j.imavis.2009.11.014. URL: http://dx.doi.org.ezproxy.puc.cl/10.1016/j.imavis.2009.11.014.

Ramirez, I., P. Sprechmann y G. Sapiro: “Classification and clustering via dictionary learning with structured incoherence and shared features”. En: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. Jun. de 2010, págs. 3501-3508. DOI: 10.1109/CVPR.2010.5539964.

Taralova, E., Fernando De la Torre y Martial Hebert: “Motion Words for Video”. En: *European Conference on Computer Vision (ECCV)*. 2014.

Tran, Khai N., Ioannis A. Kakadiaris y Shishir K. Shah: *Modeling Motion of Body Parts for Action Recognition*. 2011.

Wang, Heng y col.: “Action Recognition by Dense Trajectories”. En: *IEEE Conference on Computer Vision & Pattern Recognition*. Colorado Springs, United States, jun. de 2011, págs. 3169-3176. URL: http://hal.inria.fr/inria-00583818/en.

Wang, Y. y G. Mori: *Hidden Part Models for Human Action Recognition: Probabilistic versus Max Margin*. En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.7 (jul. de 2011), págs. 1310-1323. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2010.214.

Willems, Geert, Tinne Tuytelaars y Luc Van Gool: "An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector". En: *Computer Vision – ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II*. Ed. por David Forsyth, Philip Torr y Andrew Zisserman. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, págs. 650-663. ISBN: 978-3-540-88688-4. DOI: 10.1007/978-3-540-88688-4_48. URL: http://dx.doi.org/10.1007/978-3-540-88688-4_48.

Wright, J. y col.: *Robust Face Recognition via Sparse Representation*. En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.2 (feb. de 2009), págs. 210-227. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2008.79.

Yacoob, Yaser y Michael J. Black: *Parameterized Modeling and Recognition of Activities*. En: *Computer Vision and Image Understanding* 73.2 (1999), págs. 232-247. ISSN: 1077-3142. DOI: http://dx.doi.org/10.1006/cviu.1998.0726. URL: http://www.sciencedirect.com/science/article/pii/S1077314298907263.

Yang, Jianchao y col.: "Linear spatial pyramid matching using sparse coding for image classification". En: *in IEEE Conference on Computer Vision and Pattern Recognition(CVPR*. 2009.

Yang, Jun y col.: "Evaluating Bag-of-visual-words Representations in Scene Classification". En: *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*. MIR '07. Augsburg, Bavaria, Germany: ACM, 2007, págs. 197-206. ISBN: 978-1-59593-778-0. DOI: 10.1145/1290082.1290111. URL: http://doi.acm.org/10.1145/1290082.1290111.

Yeffet, Lahav y Lior Wolf: “Local Trinary Patterns for human action recognition.” En: *ICCV*. IEEE Computer Society, 2009, págs. 492-497. ISBN: 978-1-4244-4419-9. URL: `http://dblp.uni-trier.de/db/conf/iccv/iccv2009.html#YeffetW09`.

Ziaeefard, Maryam y Robert Bergevin: *Semantic human activity recognition: a literature review*. En: *Pattern Recognition* 48.8 (2015), págs. 2329-2345.