Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 21-51
http://rcci.uci.cu

Tipo de artículo: Artículo de revisión
Temática: Reconocimiento de patrones
Recibido: 29/07/2014 | Aceptado: 9/10/2014

Modelos de representación de características para la clasificación de acciones humanas en video: estado del arte

Features representation models for human actions classification in video: state of art

Ruber Hernández García [1*], Edel García Reyes [2], Julián Ramos Cózar [3], Nicolás Guil Mata [3]

[1] Dpto. Señales Digitales. Centro de Desarrollo GEYSED. Universidad de las Ciencias Informáticas, Carretera a San Antonio de los Baños, km 2 ½, Torrens, Boyeros, La Habana, Cuba. CP.: 19370.
[2] Centro de Aplicaciones de Tecnologías de Avanzadas (CENATAV). 7ma A #21406 e/ 214 y 216, Rpto. Siboney, Playa, La Habana, Cuba.
[3] Dpto. Arquitectura de Computadores. Universidad de Málaga. Bulevar Louis Pasteur #35, Campus de Teatinos, 29071, Málaga, España.

[*] Autor para correspondencia: rhernandezg@uci.cu

Resumen

La clasificación de acciones humanas en video es un área del conocimiento muy activa en la comunidad científica de la visión por computador. El objetivo de este campo de investigación es clasificar automáticamente acciones humanas a partir de los fotogramas que componen una secuencia de video, utilizando para ello técnicas de reconocimiento de patrones. El rendimiento de los métodos de reconocimiento de patrones depende en gran medida de la representación de los datos utilizada. Por esta razón, se centra la atención en el análisis del estado del arte referente a los modelos de representación de la información visual para la clasificación de acciones humanas en videos. El presente trabajo tiene como objetivo examinar desde un enfoque crítico las diferentes aproximaciones reportadas, así como los referentes teóricos de la temática tratada. A partir del estudio realizado se logró concluir que la aplicación de técnicas de selección de características, el uso de modelos relacionales y la obtención de una representación basada en n-gramas visuales, figuran como alternativas interesantes a incorporar como parte de los modelos de representación de características para la clasificación de acciones humanas.

Palabras clave: clasificación de acciones humanas, representación de características, selección de características, vocabularios visuales.

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 21-51
http://rcci.uci.cu

Abstract

Human actions classification in video is a very active investigation area in computer vision. The objective of this research area is to classify automatically human actions from the frames that make up a video sequence, using pattern recognition techniques. The performance of pattern recognition methods is heavily dependent on the choice of data representation on which they are applied. For this reason, this paper focuses on the analysis of the state of the art concerning the representation models of visual information for human actions classification. This paper aims to critically analyze the different approaches reported and their theoretical aspects. Finally, the study concluded that the application of features selection techniques, the use of relational models and obtaining representation based on visual n-grams shown as interesting alternatives to incorporate as part of representation models for human actions classification.

Keywords: *features representation, features selection, human actions classification, visual vocabularies.*

Introducción

La clasificación y recuperación de contenido en imágenes y videos de acuerdo a su semántica es uno de los desafíos actuales de la visión por computadora. En particular, el reconocimiento de acciones humanas[1] (*Human Actions Recognition*, HAR) en video es un área del conocimiento muy activa en este campo de investigación. En los últimos diez años la literatura recoge numerosos enfoques que permiten clasificar acciones humanas en videos (Poppe, 2010; Weinland *et al.*, 2010; Aggarwal Ryoo, 2011; Chaaraoui *et al.*, 2012), tanto en entornos controlados como reales. En gran parte, este auge se debe a sus disímiles aplicaciones en la educación, el entretenimiento, la video-vigilancia, la interacción hombre-máquina, entre otras (Bregonzio, 2011, Chakraborty, 2012).

De manera general, la clasificación automática de acciones humanas en video se compone de cuatro etapas fundamentales, Figura 1: (1) el pre-procesamiento del video, (2) la representación de la información visual, (3) el aprendizaje automático y (4) la clasificación (Turaga *et al.*, 2008; Poppe, 2010). Además se pueden incluir otros sub-procesos que pueden ejecutarse con el objetivo de incrementar la efectividad de los resultados. En general, los aportes que han impulsado el estado del arte se concentran en la segunda y cuarta etapa.

[1] Conocido igualmente como clasificación de acciones humanas.

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 21-51
http://rcci.uci.cu

Según (Bengio *et al.*, 2013), el rendimiento de los métodos de aprendizaje automático depende en gran medida de la representación de los datos utilizada. Por esta razón, el presente trabajo centra su atención en el análisis del estado del arte referente a los modelos de representación de la información visual de videos, a la luz de su aplicación en la clasificación de acciones humanas. Se profundiza en las técnicas más relevantes de la bibliografía, con el objetivo de estudiar los precedentes para el desarrollo de un modelo de representación que pueda hacer frente a las limitaciones actuales.

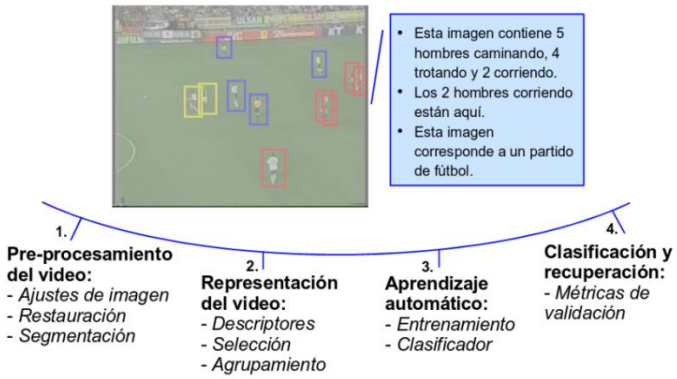

Figura 1. Proceso general de clasificación de acciones humanas en video.[2]

El resto del trabajo está estructurado de la siguiente manera. Primeramente se exponen conceptos fundamentales asociados a la representación de la información visual de las acciones humanas en videos. En las dos secciones siguientes se analizan los referentes teóricos relacionados con las técnicas de representación de características y generación de vocabularios visuales. Posteriormente, se presentan los principales enfoques empleados para la clasificación de acciones humanas en videos. Finalmente, se examinan las principales alternativas de solución en la discusión y se ofrecen las conclusiones.

[2] Fuente: elaboración propia.

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 21-51
http://rcci.uci.cu

Desarrollo

Taxonomía de las acciones humanas para su reconocimiento

Debido a los diferentes niveles de abstracción de las acciones humanas y los diversos términos usados en la literatura, es importante tratar la taxonomía de estas. La Figura 2 muestra ejemplos de acciones de algunas de las bases de datos más usadas.

Bobick (1997) emplea una clasificación para el reconocimiento de movimiento, actividad y acción; asociadas a tareas de bajo, medio y alto nivel computacional. Moeslund y colaboradores (2006) sugiere que en general los movimientos humanos pueden dividirse en tres niveles: primitiva, acción y actividad. Otros proponen incluir el nivel de situación (Gonzàlez *et al.*, 2002) o usar una jerarquía de primitivas de acciones (Jenkins y Mataric, 2002). Por su parte, Chaaraoui y colaboradores (2012) estudian las técnicas de visión por computadora aplicadas al análisis del comportamiento humano, estableciendo una jerarquía de cuatro niveles asociada al grado semántico y duración del movimiento: movimiento, acción, actividad y comportamiento.

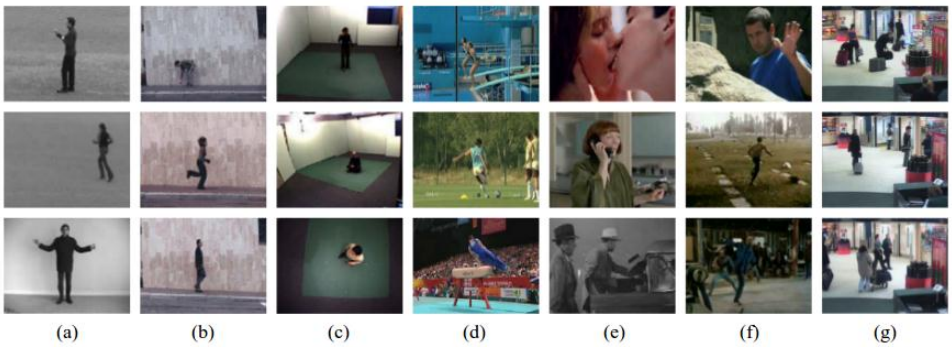

Figura 2. Ejemplos de fotogramas de acciones humanas de bases de datos simples: (a) KTH, (b) Weizmann, y reales: (c) UCF Sports, (d) UCF YouTube, (e) Hollywood, (f) HMDB51. Se puede apreciar la diversidad de los tipos de acciones.[3]

Las anteriores clasificaciones no tienen en cuenta la complejidad de las acciones por sí solas, debido a que existen acciones que son ejecutadas por una sola persona (acción simple; p.ej. caminar, correr, saltar), mientras otras requieren de la interacción de dos o más individuos para poder llevarse a cabo (acción compuesta; p.ej. besar, saludar). De igual forma es posible introducir un nivel de abstracción para la combinación de actividades en las que están involucrados

[3] Fuente: elaboración propia.

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 21-51
http://rcci.uci.cu

varios individuos (evento; p.ej. un juego de béisbol, un accidente de tráfico). Es importante establecer estas diferencias, no solo por sus especificaciones semánticas sino por la complejidad requerida para llevar a cabo su reconocimiento.

De esta manera, se propone una taxonomía de acciones de forma jerárquica, Figura 3, capaz de describir los diferentes niveles de abstracción que se pueden tener en cuenta para el reconocimiento de acciones humanas. Esta conjuga los principales elementos de los trabajos analizados e incorpora aquellos que se consideran necesarios para lograr una mejor descripción de las acciones. Es posible apreciar que tanto el tiempo involucrado en la acción como la complejidad semántica de la misma aumentan en los niveles superiores de la pirámide.

En la Tabla 1 se resumen las diferentes clasificaciones consideradas en la taxonomía adoptada, estableciendo su correspondiente descripción, intervalo de tiempo y ejemplos. Mientras que las primitivas de movimiento son muy limitadas y específicas para describir un movimiento simple, las acciones proveen una representación compacta y detallada de la dinámica humana. Por el contrario, los niveles superiores de clasificación consisten en conjuntos de acciones ordenados semánticamente, por lo que su reconocimiento requiere en primer lugar de la clasificación de las acciones que los componen. Además, las acciones presentan una resolución espacial y temporal adecuada respecto al resto, lo que hace mucho más factible su procesamiento automático. Por esta razón, las acciones deben ser interpretadas como la muestra unitaria de la vida humana.

Figura 3. Taxonomía jerárquica de acciones humanas. Se adapta la pirámide semántica de la clasificación de (Chaaraoui *et al.*, 2012), añadiendo los nuevos niveles propuestos.[4]

Tabla 1. Clasificación de las acciones humanas teniendo en cuenta su complejidad semántica.

Clasificación	**Descripción**	**Intervalo de tiempo**	**Ejemplos**

[4] Fuente: elaboración propia basada en la Figura 1 de (Chaaraoui et al., 2012).

Movimiento	Primitiva de acción que representa un cambio de pose o lugar entre fotogramas consecutivos. (Bobbick, 1997; Gonzàlez *et al.*, 2002).	fotogramas	subir un brazo, abrir la mano, mover la cabeza
Acción	Conjunto o repetición de primitivas de acciones que tienen un significado semántico como un todo, pueden estar relacionadas con objetos. (Moeslund *et al.*, 2006)	segundos o minutos	caminar, beber, saludar, besar, montar bicicleta
Acción simple	Acción que es ejecutada sin la interacción de una segunda persona.	segundos o minutos	caminar, correr, beber
Acción compuesta	Acción que requiere la interacción de dos o más individuos para poder ejecutarse.	segundos o minutos	besar, saludar
Actividad	Secuencia de acciones en un orden determinado. (Gonzàlez *et al.*, 2002)	minutos	cocinar, tomar una ducha
Situación	Actividad que adquiere significado en dependencia del contexto. (Gonzàlez *et al.*, 2002)	minutos	pedir ayuda o pedir un pase en el fútbol
Evento	Combinación de actividades en las que están involucrados varios individuos.	horas	juego de béisbol
Comportamiento	Nivel de mayor complejidad semántica, combina todos los anteriores en largos períodos de tiempo. (Chaaraoui *et al.*, 2012)	horas, días, semanas	modo de vida, hábito personal

Representación de la información visual

El proceso de representación de características visuales del video incluye diferentes sub-procesos que permiten transformar la información visual a un espacio vectorial adecuado para su posterior clasificación. A continuación se tratan tres de los sub-procesos fundamentales: la extracción de características, la representación de estas a partir de diferentes enfoques relacionales y la selección de las características de mayor poder discriminatorio.

Proceso de extracción de características

La extracción de características de bajo nivel a partir de los fotogramas del video – como el color, la textura y el flujo óptico – resulta la etapa básica para la representación de la información visual a un espacio multidimensional de rasgos de un descriptor. La representación multidimensional obtenida es más compacta, descriptiva y factible para ser utilizada por técnicas de aprendizaje automático (Bishop, 2006). Esto permite diseñar un modelo capaz de etiquetar las secuencias de videos en diferentes clases de acuerdo a su contenido.

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 21-51
http://rcci.uci.cu

En general, una función de extracción de características recibe un objeto como entrada y devuelve una representación matemática del mismo en un espacio vectorial donde se conserven las similitudes inherentes a dichos objetos. Este mismo enfoque también se aplica a las imágenes. En cuanto al video, la representación se basa principalmente en extraer descriptores a los fotogramas que lo componen.

De esta manera, dado un video *V* el proceso de extracción de características se puede definir como sigue:

DEFINICIÓN 1. *Una función de extracción de características F tal que:* $F(V,d) : V \rightarrow \mathcal{F}_d^N$*, donde la información visual contenida en V es transformada al espacio vectorial* $\mathcal{F}_d^N = \{F_1^d, \dots, F_N^d\}$*, siendo* $F_n^d = (f_1, \dots, f_D)^T$ *un vector de características de dimensión D asociado a un descriptor d.*

En la definición anterior el superíndice *T* denota la traspuesta del vector de características F, por lo que F^T es un vector fila. De este modo, es posible combinar las características en una matriz F en la que la *n*-ésima fila corresponde al vector de características F_n.

Los descriptores ofrecen una medida cuantitativa de la información visual. Según Díaz-Espinosa (2010) pueden clasificarse a partir de la información que resumen en locales y globales. Mientras que en cuanto a la dimensión de la información representada en espaciales, temporales y espacio-temporales. Ambos conjuntos de clasificación no son excluyentes. Este tipo de clasificación es extendida a los modelos de representación de acuerdo al tipo de descriptor utilizado. A continuación se resumen las características de cada tipo (Díaz-Espinosa, 2010).

- Locales: extraen una serie de puntos o regiones con información relevante a partir de los cuales se calcula un descriptor de información.
- Globales: efectúan un resumen de la información contenida en un fotograma.
- Espaciales: se aplica la función de extracción de características sobre la información contenida en cada fotograma por separado, es decir sobre las dimensiones espaciales del video.
- Temporales: extraen características utilizando la dimensión temporal del video. Para ello se puede hacer el seguimiento temporal de puntos obtenidos con un descriptor local. La idea es que la extracción de características no considere un solo fotograma, sino un sub-conjunto de ellos.
- Espacio-temporales: son una combinación de los dos anteriores, donde se calcula el descriptor a partir de una representación local teniendo en cuenta las tres dimensiones del video.

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 21-51
http://rcci.uci.cu

Las representaciones globales están dirigidas a enfoques de reconocimiento basados en toda la imagen o regiones que encierren el objetivo de reconocimiento en su totalidad, Figura 4 (a, b, c). Dichas aproximaciones son más apropiadas para aprovechar características globales de la estructura del objeto de clasificación, pero son más sensibles a oclusiones parciales y cambios de perspectiva. Por esta razón, generalmente incluyen una etapa de pre-procesamiento de la información para segmentar la región de interés para la clasificación. Estos métodos presentan altos porcientos de precisión cuando procesan secuencias de video con fondos simples y estáticos. Sin embargo, el pre-procesamiento de la imagen para segmentarla requiere de un costo computacional adicional. Además, para el tratamiento de secuencias de video de condiciones complejas – tales como múltiples perspectivas, movimientos de cámara o fondos dinámicos – se requiere de un proceso de calibración manual (Li *et al.*, 2014).

En contraposición, las representaciones basadas en descriptores locales son ampliamente aceptadas para el reconocimiento de objetivos específicos. Este tipo de representaciones han hecho posible desarrollar enfoques de reconocimiento robustos y eficientes ante una amplia variedad de condiciones de perspectivas y oclusiones (Grauman y Leibe, 2011, p. 9).

Las representaciones espacio-temporales se basan en la extracción de características locales y funcionan mejor que en ambas dimensiones por separado, Figura 4 (d, e, f). Este tipo de enfoque codifica los cambios de la información en ambas dimensiones y provee descripciones generalizables y robustas para la clasificación de acciones humanas. Su bajo costo computacional posibilitan su aplicación en sistemas HAR, siendo las más extendidas en este campo de investigación (Li *et al.*, 2014). Teniendo esto en cuenta el presente trabajo enfoca su estudio en este tipo de representaciones.

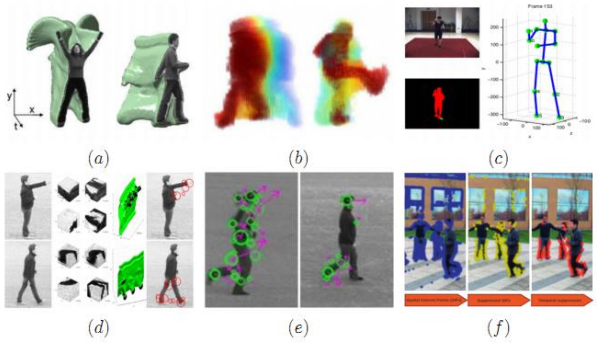

Figura 4. Ejemplos de diferentes enfoques de representaciones de acuerdo a la información codificada. Globales: (a) Volumen espacio-temporal de siluetas (Gorelick *et al.*, 2007), (b) Volumen histórico del movimiento (Weinland *et al.*, 2006), (c) Modelo

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 21-51
http://rcci.uci.cu

3D de la pose (Gong *et al.*, 2012). Locales: (d) Puntos de interés espacio-temporales (Laptev y Lindeberg, 2003), (e) MoSIFT (Chen y Hauptmann, 2009), (f) Puntos de interés espacio-temporales selectivos (Chakraborty, 2012).

Principales características de las representaciones locales

Las funciones de extracción de características locales se conocen como detectores de puntos de interés. Estos emplean diferentes criterios para encontrar las regiones donde ocurren cambios significativos de la información combinando las tres dimensiones del video. Normalmente este proceso de búsqueda se realiza utilizando diferentes escalas, tanto espaciales como temporales. Las representaciones espacio-temporales se basan en este principio, obteniendo información de ambas dimensiones.

Tomando como base la Definición 1, se define una función de extracción de características espacio-temporales como sigue:

DEFINICIÓN 2. *Una función de extracción de características espacio-temporales* F^{ST} *tal que:* $F^{ST}(V,d) : V \rightarrow \mathcal{F}_{ST}{}_d^N$, *donde la información visual contenida en V es transformada al espacio vectorial* $\mathcal{F}_{ST}{}_d^N = \{F_1^d, \dots, F_N^d\}$, *siendo* F_n^d *un vector de características de la forma* $F_n^d = (f_{n,loc}, f_{n,desc})^T$, *donde* $f_{n,loc} = \{x, y, t, \sigma, \tau\}$ *son las características espacio-temporales del punto de interés en la escala espacial* σ *y temporal* τ; $f_{n,desc} = (f_1, \dots, f_D)^T$ *es la representación vectorial de dimensión D asociada al descriptor d calculado alrededor del punto de interés.*

Cuando se trata de una secuencia de video los detectores se aplican sobre los fotogramas que la componen. De esta manera, se detectan cientos de puntos de interés y a su vez son representados por un vector de características multidimensional, que por lo general sobrepasa el centenar de componentes. El resultado de la representación es un espacio vectorial de una alta cardinalidad.

Como describe Bishop (2006, pp. 33-38), ante la alta dimensionalidad de los descriptores se pueden presentar problemas como la maldición de la dimensión y el sobre ajuste de los clasificadores. Además, aumenta el costo computacional, la redundancia de los datos y el ruido introducido por la detección de puntos de interés que se encuentran en el fondo de la imagen.

Una manera efectiva de mitigar algunas de las limitantes anteriores es representar los descriptores locales en un espacio vectorial de menor cardinalidad. Para este propósito el método de representación más extendido en la bibliografía es el conocido como BoVW (Grauman y Leibe, 2011, p. 27). Sus excelentes resultados para tareas de reconocimiento han

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 21-51
http://rcci.uci.cu

sido demostrados en numerosos estudios previos (Uijlings *et al.*, 2009; Jégou *et al.*, 2010; Kong *et al.*, 2011). Usando este enfoque el espacio multidimensional de los descriptores es mapeado a un vocabulario visual. Así, la representación del video se basa en un histograma de ocurrencia de las palabras visuales que representan los descriptores locales. De esta manera se reduce la alta dimensionalidad del espacio de características a un solo vector.

La Figura 5 muestra el esquema general del modelo BoVW para la representación de videos. Como resultado de este método de representación se obtiene un vocabulario visual $W = \{w_1, \dots, w_k\}$ compuesto por *k* palabras visuales, donde $k = |W|$ define el tamaño del vocabulario. Como parte del proceso de cuantización cada video es representado como una bolsa de palabras $V = \{w_{F_1}, \dots, w_{F_N}\}$, donde w_{F_1} es la palabra visual asignada a la característica local F_1. Por último el histograma de ocurrencia de las palabras visuales pasa a representar el video $V = \{h_1, \dots, h_k\}$, donde h_i codifica la ocurrencia de la palabra visual w_i en el video *V*.

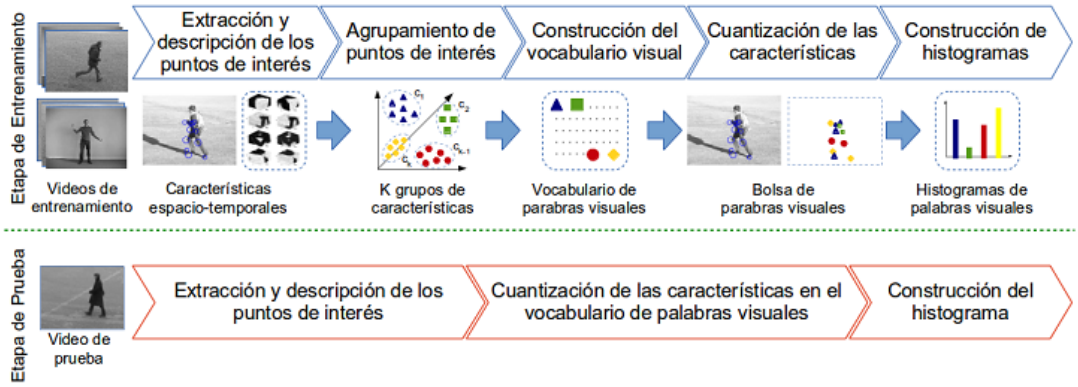

Figura 5. Esquema del modelo BoVW para la representación de videos.[5]

Modelos de representación relacionales

En su mayoría las representaciones locales presentan una limitación común, dada porque no tienen en cuenta la disposición estructural a partir de las relaciones espacio-temporales entre los descriptores. La disposición estructural de las características aporta una información adicional a la representación. Además resulta una forma de hacer frente a variaciones de la imagen debido al ruido, oclusiones parciales y cambios de perspectivas (Poppe, 2010). De esta manera es posible obtener una representación más general de la imagen a partir de las características locales.

[5] Fuente: elaboración propia basada en la Figura 3.1 de (Liu et al., 2009).

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 21-51
http://rcci.uci.cu

En el espacio 3D de las características espacio-temporales del video las relaciones pueden ser modeladas empleando diferentes enfoques: topológicas espaciales (Morales-González y Reyes, 2010; Acosta-Mendoza *et al.*, 2012), de adyacencia (Özdemir y Aksoy, 2010), temporales (Ryoo y Aggarwal, 2009; Gaur *et al.*, 2011), de proximidad (Ryoo y Aggarwal, 2009; Ta *et al.*, 2010; Zhang *et al.*, 2011), de similitud (Çeliktutan *et al.*, 2012) o como combinación de algunas de las anteriores (Ben Aoun *et al.*, 2014). De todas, las que requieren menor costo computacional son las de proximidad.

Las relaciones de proximidad básicamente establecen si un punto de interés está cerca de otro basándose en un umbral espacio-temporal $\mathcal{T}(\tau_{sp}, \tau_{tp})$. Tomando como base la definición de Ta y colaboradores (2006) para un descriptor en pareja, se define la relación de proximidad espacio-temporal entre dos características locales como sigue:

DEFINICIÓN 3. *Sean* $F_i^d = (f_{i,loc}, f_{i,desc})^T$ *y* $F_j^d = (f_{j,loc}, f_{j,desc})^T$ *dos vectores de características espacio-temporales, la relación de proximidad* $R(F_i^d, F_j^d)$ *es satisfecha si se cumplen las siguientes condiciones:*

a) $d_{sp}(f_{i,loc}, f_{j,loc}) \leq \tau_{sp}$
b) $d_{tp}(f_{i,loc}, f_{j,loc}) \leq \tau_{tp}$

donde $d_{sp}(.,.)$ *y* $d_{tp}(.,.)$ *son funciones de distancia espacial y temporal, respectivamente.*

Un enfoque básico para preservar las relaciones de proximidad entre las características locales son las representaciones basadas en mallas (Laptev *et al.*, 2008; Bregonzio, 2011). Sin embargo, generalmente resultan representaciones redundantes y contienen características poco informativas. Como otra alternativa es posible explotar las correlaciones entre los descriptores locales para la construcción de otros descriptores de mayor nivel de abstracción (Poppe2010). En su mayoría, estos métodos se basan en la creación de una matriz de correlación de los descriptores locales (Scovanner *et al.*, 2007; Liu *et al.*, 2008; Kim y Cipolla, 2009).

Aunque los enfoques anteriores han intentado incorporar más detalles de las relaciones entre los descriptores, el modelo estructural de los mismos no es tenido en cuenta. En este sentido las representaciones basadas en partes juegan un papel fundamental para preservar la información estructural de las características (Grauman y Leibe, 2011).

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 21-51
http://rcci.uci.cu

Son varios los modelos de este tipo recogidos por la literatura, Figura 6. De todos, el modelo BoVW es el más simple[6], debido a que no codifica ninguna relación de los descriptores. A partir de este es posible crear modelos que expresen las relaciones entre los descriptores. Estos van desde los más complejos – como los tipo constelación que engloban todas las relaciones posibles – hasta los más simples – como los tipo estrella o esparcido flexible – que posibilitan reducir el costo computacional de la representación.

Estos métodos han dado origen a las representaciones basadas en grafos, que permiten la creación de modelos de representación estructurales. Además, los grafos presentan propiedades de invariabilidad posibilitando que la representación de la imagen se mantenga igual ante determinadas transformaciones tales como traslación o rotación (Bunke, 2000). Precisamente estas propiedades han posibilitado el desarrollo de diversas aproximaciones aplicadas a la representación del video (Ta *et al.*, 2010; Gaur *et al.*, 2011; Çeliktutan *et al.*, 2012; Ben Aoun *et al.*, 2014).

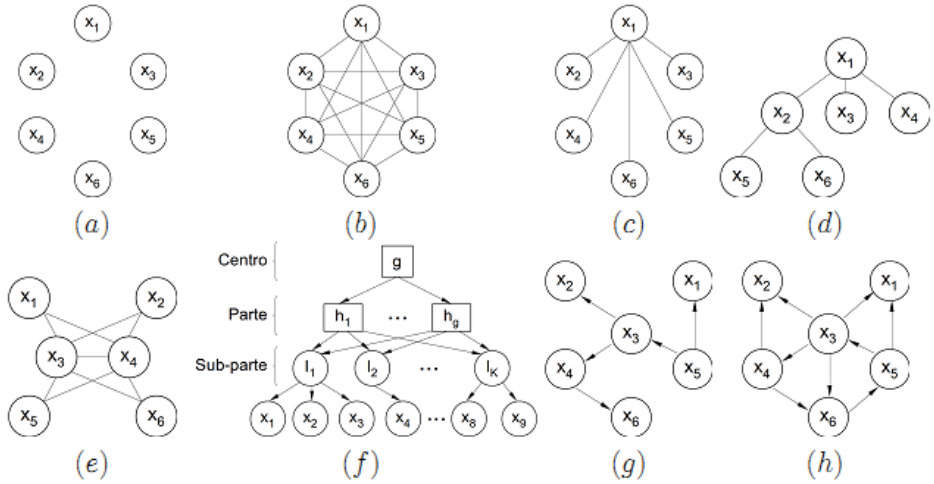

Figura 6. Visión general de las representaciones basadas en partes más populares recogidas en la literatura: (a) BoVW (Csurka *et al.*, 2004), (b) Constelación (Fergus *et al.*, 2003), (c) Estrella (Leibe *et al.*, 2008), (d) Árbol (Felzenszwalb y Huttenlocher, 2005), (e) Abanico (Crandall *et al.*, 2005), (f) Jerárquica (Bouchard y Triggs, 2005), (g) Esparcida flexible (Carneiro y Lowe, 2006).[7]

[6] Aunque no representa un modelo relacional, se incorpora como base del resto de los enfoques.
[7] Fuente: elaboración propia basada en la Figura 8.5 de (Grauman y Leibe, 2011).

Considerando la Definición 2, un video *V* puede ser representado de forma alternativa como $V = \{F_1^d, \dots, F_N^d\}$. Siendo así, una representación basada en grafo para un video puede definirse como sigue:

DEFINICIÓN 4. *Sea* $V = \{F_1^d, \dots, F_N^d\}$ *un video representado por N vectores de características locales y* $W = \{w_1, \dots, w_k\}$ *el vocabulario visual de k palabras visuales obtenido por el modelo BoVW. Un video se representa como un grafo* $G = (P, E, W, \mathfrak{B})$*, donde* $P = \{p_1, \dots, p_N\}$ *es el conjunto de vértices*[8] *que representa los N puntos de interés que forman el video,* $E = \{\{u, v\} \mid u, v \in P, u \neq v, \exists R(F_u^d, F_v^d)\}$ *es el conjunto de aristas, W son las etiquetas de los vértices que coinciden con el vocabulario visual y* $\mathfrak{B} : P \rightarrow W$ *es la función de etiquetado para la asignación de etiquetas a los vértices.*

Las representaciones basadas en grafos ofrecen una aplicación importante para el reconocimiento de patrones. El descubrimiento de patrones frecuentes – especialmente la minería de subgrafos frecuentes – ha permitido el desarrollo de diversas técnicas con aplicación en diferentes dominios de la ciencia (Acosta-Mendoza *et al.*, 2012).

Las representaciones basadas en grafos preservan la limitante asociada a la alta dimensionalidad de las representaciones locales. La cantidad de vértices del grafo generado es igual a la cantidad de puntos de interés que representan el video. Mientras que las aristas son generadas para todas aquellas características locales que cumplen con una relación de proximidad. Debido a esto han sido propuestos determinados enfoques orientados a disminuir la complejidad de la tarea de similitud entre subgrafos frecuentes (Öezdemir y Aksoy, 2010; Ben Aoun *et al.*, 2014).

En particular, Özdemir y Aksoy (2010) proponen una representación intermedia que combina el poder estructural de los grafos con la eficiencia del modelo BoVW. Esta aproximación representa cada imagen con un histograma de los subgrafos frecuentes presentes en el grafo correspondiente a la misma. De esta manera se logra obtener un modelo que reduce el costo computacional de la representación basada en grafo. Por sus características este tipo de enfoque resulta interesante para aplicar a la representación de la información visual del video.

Selección de características

En principio, lo ideal para la resolución de un problema de clasificación es disponer de la máxima información posible. Sin embargo, como ha sido tratado antes, el rendimiento de los algoritmos de aprendizaje se puede deteriorar ante la

[8] En la bibliografía normalmente se denota como V, en este caso se denota como P para diferenciar de la notación del video.

abundancia de información (Bishop, 2006, p. 33). Además, la presencia de características irrelevantes y redundantes – p. ej. las ubicadas en el fondo de la imagen – genera información ruidosa en las representaciones de mayor nivel. Esto implica también un procesamiento adicional e innecesario en las etapas superiores de representación.

Teniendo esto en cuenta, la selección de características resulta una etapa fundamental en cualquier proceso de representación. El objetivo de la selección de características es reducir la dimensión de los datos mediante la eliminación de aquellas características que son ruidosas, redundantes o irrelevantes para el problema de clasificación (Bonev, 2010).

La aplicación de técnicas de selección de características aplicadas a tareas de reconocimiento de video de acuerdo a su contenido – en especial la clasificación de acciones humanas – exige métodos diferentes a los tradicionales. Particularmente, centrándose en la similitud visual que existe entre algunos tipos de acciones, resulta difícil distinguir aquellas características que son más discriminatorias, debido a que estas clases comparten determinadas primitivas de acciones. Por lo que generalmente, los métodos tradicionales de selección de características resultan ineficientes por sí solos. Siendo así, los aportes existentes en este sentido emplean métodos con fines específicos diseñados para este campo de investigación. Por ejemplo, un requerimiento específico consiste en eliminar las características locales ubicadas en el fondo y seleccionar aquellas asociadas a la figura del cuerpo humano.

Varios enfoques han sido propuestos para reducir la influencia de las características localizadas en el fondo de la imagen. Liu y colaboradores (2009) proponen una técnica basada en el conocido algoritmo *PageRank* (PR) (Brin y Page, 1998) para ordenar y seleccionar las características de mayor relevancia. Los resultados experimentales muestran la eficacia de la técnica PR en escenarios de una sola persona, donde la mayoría de los puntos de interés en el fondo son eliminados. Sin embargo, ante la presencia de varias personas el método pierde efectividad (Bregonzio, 2011).

Por su parte, Gilbert y colaboradores (2009) agrupan las características de forma jerárquica para producir un conjunto de características compuestas. Logran identificar los conjuntos que aparecen con mayor frecuencia en determinadas secuencias de acciones usando la técnica de minería de datos *APriori* (Agrawal y Srikant, 1994). De esta manera, detectan de forma simultánea las configuraciones de características localizadas en diferentes posiciones de la acción o que representan diferentes movimientos. No obstante, el elevado costo computacional de esta aproximación es su principal limitante, por lo que solo es factible para pequeñas bases de datos.

Con el objetivo de disminuir el costo computacional de la selección de características en entornos no controlados, Bregonzio (2011) desarrolla un método de selección de características basado en la técnica *Multi-Class Delta Latent*

Dirichlet Allocation[9] (MC-ΔLDA) (Andrzejewski *et al.*, 2007). La idea consiste en seleccionar las características de forma colaborativa a partir de los patrones compartidos entre las diferentes clases de acciones. Esto implica que el método no es capaz de identificar los elementos distintivos de una acción, por lo que resulta menos efectivo para identificar determinadas acciones que sean similares.

A diferencia de los métodos anteriores, que intentan obtener las características más relevantes a partir de medidas de relevancia, Chakraborty (2012) presentan un enfoque bastante novedoso. Su aproximación se basa en la detección de puntos de interés de una manera selectiva, aplicando una máscara de supresión circundante (*Surround Suppression Mask*, SSM) (Grigorescu *et al.*, 2004) combinada con restricciones locales y temporales. Este método permite eliminar satisfactoriamente los puntos de interés localizados en el fondo y obtiene características locales repetibles, estables y distintivas para el cuerpo humano.

Generación de vocabularios visuales

El modelo BoVW ha sido ampliamente usado para la clasificación de acciones humanas (Liu *et al.*, 2009; Ullah *et al.*, 2010; Kong *et al.*, 2011; Chakraborty, 2012; Hernandez-Heredia, 2013). Mediante este método las secuencias de acciones son representadas por un histograma de la ocurrencia de las palabras visuales en el video, las cuales son obtenidas por el agrupamiento de las características locales. Algunas de las limitantes de este método están asociadas fundamentalmente al proceso de generación del vocabulario visual (Zhang *et al.*, 2011; Yang *et al.*, 2012; Cózar *et al.*, 2012).

En el agrupamiento de las características locales se realiza usando *K-means* u otro método no supervisado. Desafortunadamente, estos métodos no son capaces de obtener un vocabulario con adecuado poder discriminatorio debido a que son generadas palabras innecesarias y no descriptivas (Zhang *et al.*, 2011; Cózar *et al.*, 2012). El poder descriptivo de los vocabularios visuales es influenciado por el tamaño del mismo. En principio, entre más palabras visuales son generadas el rendimiento es mejor. Sin embargo, el rendimiento es saturado cuando el número de palabras visuales sobrepasa cierto nivel (Liu *et al.*, 2008b). Otra consecuencia de usar más palabras visuales es que el número de estas que describen cada categoría disminuye (Cózar *et al.*, 2012), Figura 7.

[9] Se usa la terminología en inglés debido a no haber una traducción exacta al español.

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 21-51
http://rcci.uci.cu

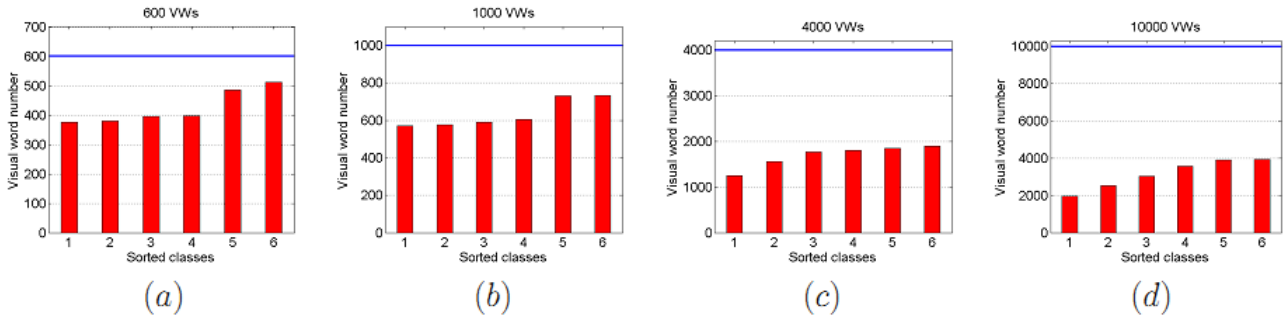

Figura 7. Número de palabras visuales usadas en la descripción de cada categoría de la base de datos KTH, para diferentes tamaños de vocabularios del descriptor STIP: (a) 600, (b) 1000, (c) 4000, (d) 10000. Cada clase es descrita solo con una porción del vocabulario total. (Cózar *et al.*, 2012)

Las ideas anteriores sugieren que eliminar palabras visuales del vocabulario posibilita incrementar la eficiencia de la clasificación. No obstante, determinar el tamaño óptimo de un vocabulario visual es una tarea aún no resuelta y es uno de los problemas más tratados del modelo BoVW. Según Zhao e Ip (2013), la base de un vocabulario visual compacto es que contenga un número mínimo de palabras visuales similares, tanto desde el punto de vista semántico como visual. Siendo así, para que el tamaño del vocabulario sea óptimo cada palabra visual debe corresponder a un significado o concepto único. Además debe considerarse que el rendimiento del proceso de clasificación sea satisfactorio.

Teniendo en cuenta esto se han desarrollado diferentes técnicas con el objetivo de seleccionar las palabras visuales de mayor relevancia. En (Zhang *et al.*, 2011) se propone un ordenamiento de las palabras visuales basado en el algoritmo PR para seleccionar las palabras más descriptivas y en este orden reducir el vocabulario. Esta misma técnica es aplicada al reconocimiento de acciones en (Cózar *et al.*, 2012), demostrando que una reducción del vocabulario puede mejorar la efectividad de la clasificación de acciones.

Liu y Shah (2008) adoptan una búsqueda tipo *greedy* sobre la pérdida de la Información Mutua (*Mutual Information*, MI) entre las palabras visuales para determinar el tamaño óptimo del vocabulario. Su trabajo está inspirado en el agrupamiento de las palabras visuales de acuerdo a la información mutua. Usando esta misma métrica de co-ocurrencia Chakraborty (2012) propone un método de reducción pero basado en la técnica de Agrupamiento Aglomerado de Información (*Agglomerative Information Bottleneck*, AIB). Otro de los trabajos enfocados en la construcción de un vocabulario óptimo es (Zhao e Ip, 2013). Para esto proponen una representación de alto nivel que denominan Descriptor Semántico Aproximado a partir de la asociación semántica de las palabras visuales. Esta relación es formulada utilizando el grado de co-ocurrencia de las parejas de palabras visuales basado en la medida conocida como *Pearson*

Product Moment[10] (PPM), en lugar de MI. Finalmente, el vocabulario es generado por un agrupamiento espectral de un grafo de las relaciones semánticas entre los descriptores.

La generalidad de los enfoques planteados se basa en realizar primeramente un agrupamiento de las características de bajo nivel usando una técnica común como *K-means*. De esta manera crean vocabularios visuales de un tamaño relativamente grande para luego aplicar una técnica de reducción y obtener un vocabulario mucho más compacto y discriminatorio. Sin embargo, los métodos de reducción propuestos siguen requiriendo de un valor umbral para seleccionar las palabras y determinar el tamaño final del vocabulario (Liu y Shah, 2008; Zhang *et al.,* 2011; Cózar *et al.,* 2012; Chakraborty, 2012).

Otro de los problemas del modelo BoVW reportado en la bibliografía está relacionado con la pérdida de las restricciones espaciales y temporales. Precisamente esta es una de las razones más importantes del limitado poder discriminatorio de este método. El uso de modelos relacionales – en particular mediante la co-ocurrencia de las palabras visuales – han dado lugar a nuevas representaciones de mayor nivel de abstracción (Poppe, 2010).

En este sentido, se pueden presentar dos tipos de relaciones entre las palabras visuales: las semánticas y las de proximidad (Li *et al.,* 2011). Las primeras se ven reflejadas por la similitud conceptual de las palabras visuales. Esto se debe a que las palabras visuales son obtenidas a partir del agrupamiento de características de bajo nivel en términos de su apariencia visual. Como resultado diferentes palabras visuales pueden corresponder al mismo concepto – p. ej. una misma parte del cuerpo o un mismo tipo de movimiento – por lo que se dice que están conceptualmente relacionadas. Por otra parte, la distribución estructural de las palabras visuales se pueden expresar como relaciones espacio-temporales de proximidad. En el caso de no tener en cuenta estas relaciones las correspondientes características son codificadas por separado y esto reportaría una pérdida de información relevante para el proceso de clasificación.

Varios han sido los enfoques propuestos con el objetivo de incorporar información relacional al modelo BoVW. Liu y colaboradores (2009) proponen el uso de las palabras visuales descriptivas a partir de la construcción de correlogramas. Mientras que en los trabajos (Chen y Hauptmann, 2009; Zhang *et al.,* 2011) se obtienen frases visuales tomando la co-ocurrencia de dos palabras visuales. Sin embargo estas aproximaciones solo tienen en cuenta las relaciones entre los pares de palabras, obviando estructuras mucho más complejas que pueden brindar más información.

[10] Se usa la terminología en inglés debido a no haber una traducción exacta al español.

Las representaciones basadas en n-gramas[11] son ampliamente usadas en el campo del procesamiento de lenguaje natural (Lopez-Monroy *et al.*, 2013), en particular en la minería de texto y la recuperación de información (Bekkerman y Allan, 2004). Este modelo es propuesto por Li y colaboradores (2011) para la obtención de n-gramas visuales a partir de una exploración de la vecindad de cada palabra visual, para finalmente representar cada imagen con el histograma de ocurrencia de estos. López-Monroy y colaboradores (2013) aplican este mismo enfoque en la clasificación de imágenes histopatológicas, desarrollando una extensión del modelo BoVW denominado "bolsa de n-gramas visuales" (*Bag-of-Visual-Ngrams*, BoVN). Otros trabajos como (Quack *et al.*, 2007; Yuan *et al.*, 2007) usan técnicas de minería de datos para crear frases visuales de diferentes longitudes con el propósito de conseguir más información relevante. Sin embargo, resultan ineficientes por sus altos costos computacionales.

Representación de la información visual de acciones humanas

Típicamente, la representación de la información visual de las acciones tiene dos requerimientos esenciales. Primero, la representación necesita ser invariante a las diversas transformaciones de la imagen y variabilidad de la ejecución de las acciones. En segundo lugar, la representación debe ser suficientemente discriminatoria para el proceso de clasificación. La interpretación práctica de estas restricciones es que una representación ideal para la clasificación de acciones debe ser invariante a los cambios de apariencia de la persona, modificaciones del entorno – como el fondo o la perspectiva – y velocidad de ejecución de la acción (Poppe, 2010).

El enfoque general de los detectores de puntos de interés, para la clasificación de acciones humanas en video, se basa en la selección de localizaciones en el video donde se maximiza una función específica de prominencia y su representación usando el modelo BoVW (Liu y Shah, 2008; Laptev *et al.*, 2008; Ullah *et al.*, 2010; Kong *et al.*, 2011; Chakraborty, 2012; Zhao e Ip, 2013; Ben Aoun *et al.*, 2014). Diferentes autores han desarrollado bastas revisiones sobre los detectores y descriptores espacio-temporales (Mikolajczyk y Schmid, 2005; Wang *et al.*, 2009; Shao y Mattivi, 2010). En la

Tabla 2. Análisis de las principales publicaciones relacionadas con la temática en los últimos 5 años. se muestra un análisis de las principales publicaciones de los últimos 5 años, distinguiendo el uso de los diferentes enfoques de representación, así como los principales autores y espacios de publicación. A continuación se describen los enfoques locales utilizados para la clasificación de acciones humanas en el video.

[11] Se puede describir un n-grama de grado n como una secuencia de n palabras.

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 21-51
http://rcci.uci.cu

Tabla 2. Análisis de las principales publicaciones relacionadas con la temática en los últimos 5 años.

Representación de la información visual para la clasificación de acciones humanas				
Tipos de representación		**Publicaciones**		
Globales	**Locales**	**Autores**	**Revistas**	**Eventos**
Volumen espacio-temporal: 12	Puntos de interés: 143	Globales: Blank, M.; Gorelick, L.; Weinland, D.; Gong, W., Mori, G.; Bobick, A.F.; Wang, L.; Wang, Y. Locales: Schüldt, C.; Laptev, I.; Dollár, P.; Niebles, J. C.; Ikizler, N; Scovanner, P.; Shah, M.; Wong, S.; Gilbert, A.; Kläser, A.; Marszalek, M.; Schmid, C.; Liu, J.; Wang, H.; Bregonzio, M.; Gong, S.; Chakraborty, B; Moeslund, T.B.	Pattern Recognition; Computer Vision and Image Understanding; IEEE Transactions on Pattern Analysis and Machine Intelligence; Pattern Recognition Letters; International Journal on Computer Vision; Image and Vision Computing	International Conference on Computer Vision; International Conference on Computer Vision and Pattern Recognition; British Machine Vision Conference; International Conference on Pattern Recognition
Volumen histórico: 5	MoSIFT: 28			
Modelo de la pose:19	Puntos selectivos: 19			
Otros: 7	Otros: 21			
Total: 43	Total: 211			

Detectores de puntos de interés espacio-temporales

Laptev y Lindeberg (2003) fueron los primeros que propusieron un detector de características locales – conocido como Harris3D – basado en una extensión espacio-temporal del detector Harris (Harris y Stephens, 1988). Los puntos de interés espacio-temporales (*Spatio-Temporal Interest Points*, STIP) se determinan a partir de los máximos locales de los valores característicos de una matriz espacio-temporal de segundo orden para cada punto del video, a partir de escalas espaciales y temporales independientes. La importancia de usar escalas para el espacio y el tiempo por separado está dada porque en general estas magnitudes son independientes para las acciones. La escalas pueden ser seleccionadas automáticamente (Laptev, 2005) y ser adaptables a la velocidad de los eventos para compensar las variaciones de la imagen (Laptev *et al.*, 2007; Laptev y Lindeberg, 2004b).

Dollár y colaboradores (2005) argumentan que en determinados casos los puntos de interés obtenidos por Harris3D resultan poco frecuentes, sobre todo cuando no existen suficientes movimientos característicos. En este caso se dice que se genera una representación de STIP esparcida. Con el objetivo de mejorar este problema ellos proponen el detector Cuboid. Este método emplea un kernel espacial gaussiano de suavizado y filtros temporales de Gabor. Sin embargo, este método presenta como aspecto negativo que los parámetros de escala espacial y temporal son definidos manualmente y permanecen fijos, por lo que no es invariante a la escala.

Un enfoque basado en el flujo óptico es el propuesto por Chen y Hauptmann (2009). El detector MoSIFT primero aplica el algoritmo SIFT (Lowe, 2004) para encontrar los componentes visualmente distintivos en el dominio espacial. Este método tiene el inconveniente de no ser invariable a la escala temporal. Otro trabajo que propone un detector de puntos de interés basado en el movimiento es (Li *et al.,* 2014), a partir de un filtrado multi-dirección de la energía de movimiento. A pesar que el movimiento es una magnitud significativa para la clasificación de acciones, estos métodos son totalmente dependientes de esta para la detección de los puntos de interés, por lo que presentan limitantes para su aplicación en entornos no controlados.

En los métodos anteriores las secuencias son representadas por los puntos de inicio y parada del movimiento. En contraposición a esto, Oikonomopoulos y colaboradores (2006) proponen un detector de puntos prominentes que se basa en los picos de variación de la acción, como pueden ser los bordes de un objeto en movimiento. En su trabajo presentan una extensión espacio-temporal del detector de regiones prominentes (Kadir y Brady, 2000) usando la entropía.

Un enfoque similar al de (Oikonomopoulos *et al.,* 2006) – pero basado en la medida de prominencia Hessian (Lindeberg, 1998) – resulta el detector Hessian3D propuesto por Willems y colaboradores (2008). Este detector mide la prominencia usando el determinante de la matriz 3D de Hessian. Los puntos obtenidos de manera densa son invariantes de escala y su procesamiento es computacionalmente más eficiente que (Oikonomopoulos *et al.,* 2006).

A pesar de los resultados prometedores reportados por los enfoques descritos, los métodos anteriores presentan algunas limitantes. Estas aproximaciones resultan vulnerables a los movimientos de cámara y los fondos no homogéneos (Chakraborty *et al.,* 2012), condiciones presentes en los entornos reales de aplicación. Como resultado de esto la estabilidad de los puntos de interés es variable ante estas condiciones, encontrándose muchos puntos ubicados en el fondo o partes no significativas a la acción en cuestión (Chakraborty, 2012), lo que decrementa los resultados de la clasificación.

Para superar estos problemas, dos direcciones principales se han seguido. Algunos métodos como (Wong y Cipolla, 2007; Gilbert *et al.,* 2009; Bregonzio, 2011) aplican diferentes vías de obtener los STIP. Wong y colaboradores (2007) proponen una estructura de información global para detectar los puntos de interés. Una modificación del detector Cuboid es aplicada en (Bregonzio, 2011) a partir de un filtro Gabor bidimensional (2D). Estos métodos funcionan satisfactoriamente en bases de datos simples, pero no son suficientemente robustos para entornos no controlados. Por

su parte, Gilbert y colaboradores (2009) usan características locales densas que son espacial y temporalmente agrupadas mediante un proceso jerárquico. Un enfoque similar es usado en (Wang *et al.*, 2009) para conformar un detector Denso muestreando bloques de video en posiciones y escalas regulares. Aunque este método reporta buenos resultados para entornos no controlados, se introduce demasiado ruido por la presencia de puntos que no son significativos para la ejecución de la acción.

Otras aproximaciones como (Liu *et al.*, 2009b; Ullah *et al.*, 2010; Bregonzio, 2011; Hernandez-Heredia, 2013) primero aplican un detector de STIP y luego usan diferentes heurísticas para seleccionar los puntos de interés más significativos. El principal inconveniente de estos métodos radica en la aplicación de diferentes técnicas de pre/post-procesamiento – como la segmentación, las ROI y el seguimiento – que incrementan el costo computacional de la representación. Sin embargo, a pesar de incluir estos tipos de sub-procesos no logran mejorar significativamente los resultados de clasificación en bases de datos complejas.

Basado en las limitantes anteriores, Chakraborty (2012) proponen una técnica de detección de puntos de interés selectivos (*Selective Spatio-Temporal Interest Points*, SSTIP). Con este objetivo se aplica una SSM junto a restricciones espaciales y temporales. No obstante, presenta como desventaja que no tiene en cuenta la información temporal para determinar las regiones donde ocurren cambios significativos asociados a la ejecución de la acción, debido a que los puntos de interés son detectados solo en la dimensión espacial de los fotogramas.

Descriptores espacio-temporales

De conjunto a los detectores de STIP se han propuesto diversos descriptores espacio-temporales que capturan las características de la imagen alrededor de los puntos de interés. Uno de los primeros trabajos de descriptores locales para el video es el desarrollado por Laptev y Lindeberg (2004). Ellos comparan diferentes tipos de descriptores y reportan los mejores resultados para los basados en histogramas del flujo óptico y los gradientes espacio-temporales.

Por su parte, Dollár y colaboradores (2005) evalúan diferentes descriptores espacio-temporales basados en el brillo, el gradiente y el flujo óptico. Ellos probaron varias combinaciones: la concatenación de los valores de los píxeles, una malla de histogramas locales y un histograma global. De todas las variantes, la concatenación del gradiente reporta el mejor rendimiento.

Los descriptores HOG y HOF son presentados por Laptev y colaboradores (2008). Para caracterizar localmente el movimiento y la apariencia los autores combinan el histograma del gradiente orientado en el espacio (*Histogram of*

Oriented Spatial Gradients, HOG) y el histograma del flujo óptico (*Histogram of Optical Flow*, HOF). Los histogramas son acumulados en la vecindad espacio-temporal de los puntos de interés detectados, para finalmente ser concatenados y dar lugar al descriptor HOG/HOF. Otro trabajo que usa los descriptores HOG y HOF es (Chen y Hauptmann, 2009). El descriptor MoSIFT crea un solo vector de características a partir de una fusión de estos descriptores.

Una extensión del descriptor de imagen SIFT (Lowe, 2004) al espacio 3D del video es propuesta en (Scovanner *et al.*, 2007). El descriptor 3D-SIFT en esencia es similar a su antecesor, exceptuando que se calcula la dirección del gradiente para cada punto en las tres dimensiones del video. Otra generalización del descriptor SIFT es propuesta por Kläser *et al.*, 2008). Conocido como HOG3D, se basa en histogramas orientados del gradiente en las tres dimensiones, donde los gradientes se calculan usando una representación integral del video.

En (Bay *et al.*, 2006) se presenta el descriptor SURF como una variante parcialmente inspirada en SIFT. Willems y colaboradores (2008) proponen la extensión ESURF aplicada al video. En este método se determina un volumen alrededor de los puntos de interés, el cual es dividido en celdas y cada una es representada por un vector de sumas ponderadas usando las funciones *Haar-wavelets* a lo largo de las tres dimensiones.

Otro descriptor muy popular es el conocido como N-jets (Laptev *et al.*, 2007), aunque no es muy usado para el reconocimiento de acciones. Consiste en un conjunto de derivadas parciales de una función hasta el orden N y es comúnmente calculada a partir de una representación espacial. Esencialmente describe el movimiento alrededor del punto de interés debido a que sus dos primeros niveles representan la velocidad y la aceleración. Este descriptor es usado por (Chakraborty, 2012) para caracterizar los (SSTIP), sin embargo resulta la variante menos efectiva del estado del arte para la descripción de acciones en el video (Laptev y Lindeberg, 2004; Mikolajczyk y Schmid, 2005).

Discusión

Los enfoques basados en representaciones locales basan sus ventajas en encontrar estructuras locales de la imagen. De esta forma es posible codificar la información visual en un descriptor que sea invariante a transformaciones de la imagen – tales como traslación, rotación, escalado o deformaciones – cambios de perspectiva y presencia de ruido. A partir del uso de detectores de puntos de interés se obtienen las porciones de la imagen que contienen información distintiva – como pueden ser las esquinas o porciones con cambios de movimiento – que puedan ser fácilmente localizables bajo estas disímiles condiciones. De esta manera se consigue un conjunto amplio de características locales que capturan la esencia de la imagen.

No obstante, las representaciones locales presentan limitantes asociadas a su alta dimensionalidad. Con el objetivo de hacer frente a estas, el modelo BoVW resulta una alternativa eficiente para representar las características locales a un espacio vectorial de menor cardinalidad. Este método se caracteriza por su sencillez y capacidad para agrupar los conceptos semánticamente más significativos, elemento que lo ha convertido en un referente de solución a este problema. Por estos motivos ha sido ampliamente utilizado en la clasificación de acciones humanas en video.

A pesar de la eficiencia del modelo BoVW, varios resultados experimentales reportados en la literatura muestran que las palabras visuales no son tan expresivas como las textuales (Zhang *et al.*, 2011; Yang *et al.*, 2012). Esto se debe fundamentalmente a que el agrupamiento es un proceso no supervisado y usualmente genera palabras descriptivas y no descriptivas. En este sentido, Zhang y colaboradores (2011) describen dos problemas del modelo BoVW que provocan su limitado poder discriminatorio:

- El primero está dado porque las palabras visuales no tienen en cuenta la información del contexto. Esto provoca que la información semántica descrita por las relaciones entre las palabras visuales se pierda y el resultado de la clasificación sea erróneo.
- El segundo se debe a que la mayoría de las técnicas de generación del vocabulario visual emplean una métrica de distancia general, tales como la distancia euclidiana o la norma L1 (Duda *et al.*, 2001). Debido a esto muchas características locales con información semántica similar pueden ser representadas por palabras visuales diferentes y viceversa. Esto hace que se generen palabras visuales innecesarias y no descriptivas que generan ruido durante el proceso de clasificación.

Estas ideas resaltan aún más la importancia de aplicar técnicas para seleccionar las palabras visuales de mayor relevancia y a la vez reducir el tamaño del vocabulario. De esta forma se podrá obtener una representación de la información visual con mayor poder discriminatorio, lo que a su vez posibilitará lograr mejores resultados de clasificación. Además, su empleo puede aportar algunas de las siguientes ventajas:

- La eficiencia (en tiempo y/o en espacio) de la mayoría de los algoritmos de aprendizaje depende del número de características empleado. Por tanto, seleccionando un conjunto de características más pequeño el algoritmo funcionaría más rápido y/o con menor consumo de memoria u otros recursos.
- Mejora en los resultados obtenidos: algunos de los algoritmos de aprendizaje, que trabajan muy bien con pocas características relevantes, ante la abundancia de información pueden tratar de usar características irrelevantes

y ser confundidos por las mismas, ofreciendo resultados peores. Así que la selección de características puede ayudar a obtener mejores resultados indicando qué características son más adecuadas para la clasificación.

- Reducción de los recursos necesarios para el almacenamiento y transmisión de la información de las características no seleccionadas.

Otra de las limitantes del modelo BoVW se debe a la pérdida de las relaciones espacio-temporales de los descriptores. Como ha sido analizado, las representaciones basadas en partes permiten conservar las estructuras relacionales de las características locales. En particular, las representaciones basadas en grafos posibilitan hacer frente a esta problemática, pero su aplicación en el dominio del video presenta un elevado costo computacional. Esto pudiera compensarse aplicando un modelo esparcido flexible que limite el grado de los vértices. A su vez, este tipo de representación puede posibilitar la obtención de n-gramas visuales a partir de los subgrafos frecuentes que aparezcan en los videos. Esto sin dudas permitirá aumentar el grado de abstracción de la representación.

Finalmente, se puede afirmar que aunque la bibliografía consultada reporta diferentes aproximaciones con el objetivo de obtener mejores representaciones de la información visual, los enfoques existentes presentan aún determinadas limitantes y aún se muestran ineficientes para su aplicación en la clasificación de acciones humanas. Esto ha llevado a la comunidad científica, en los últimos años, a buscar nuevas alternativas para mejorar los resultados de la clasificación, tales como aumentar el grado de abstracción de las características bases (Wang *et al.*, 2013, Guthier *et al.*, 2014), nuevos métodos de codificación de las mismas (Oneata *et al.*, 2013; Cai *et al.*, 2014) o concentrarse en las técnicas de clasificación (Ji *et al.*, 2013, Tran *et al.*, 2014).

Conclusiones

La gran cantidad y actualidad de propuestas de técnicas de representación de la información visual disponibles en la literatura muestra que esta temática constituye un campo de investigación muy activo. El estudio realizado ratifica que, las representaciones locales se caracterizan por su efectividad y eficiencia en los resultados que ofrecen. No obstante, este tipo de representación presenta determinadas limitaciones – sobre todo relacionadas con la pérdida de la información del contexto – que limitan su poder discriminatorio. Por lo que es necesario la creación y actualización de modelos de representación de la información visual con vistas a mejorar los resultados de clasificación de acciones humanas en video.

Por estas razones, la creación de métodos de representación que tengan en cuenta la estructuración relacional de las características locales y logren seleccionar las de mayor poder discriminatorio, constituye un reto vigente para la comunidad científica en este campo de investigación. En este sentido, la incorporación de técnicas de selección de características en el proceso de representación de la información visual es vital para garantizar el poder discriminatorio de las mismas y al mismo tiempo reducir la dimensionalidad de los datos, su aplicación es posible en los diferentes sub-procesos de representación en pos de incrementar la eficacia y eficiencia de las etapas posteriores de entrenamiento y clasificación. Por otra parte, el uso de modelos relacionales han dado lugar a nuevas representaciones de mayor nivel de abstracción, su empleo como parte del modelo BoVW posibilita que se tengan en cuenta las restricciones espaciales y temporales, en este sentido los n-gramas visuales permiten conservar las estructuraciones semánticas y contextuales de las palabras visuales con mayor precisión.

Referencias

- ACOSTA-MENDOZA, N.; GAGO-ALONSO, A., et al. Frequent approximate subgraphs as features for graph-based image classification. Knowledge-Based Systems, 2012, 27: p. 381-392.
- AGGARWAL, J.; RYOO, M. Human activity analysis: A review. ACM Computing Surveys (CSUR), 2011, 43 (3): p. 1-43.
- AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. En: International Conference on Very Large Data Bases (VLDB), 1994, p. 487-499.
- ANDRZEJEWSKI, D.; MULHERN, A., et al. Statistical debugging using latent topic models. En: Machine Learning: ECML 2007: Springer, 2007, p. 6-17.
- BAY, H.; TUYTELAARS, T.; GOOL, L. Surf: speeded up robust features. En: European Conference on Computer Vision (ECCV'06): Springer, 2006, p. 404-417.
- BEKKERMAN, R.; ALLAN, J. Using bigrams in text categorization. Technical Report, Department of Computer Science, University of Massachusetts, 2004.
- BEN AOUN, N.; MEJDOUB, M., et al. Graph-based approach for human action recognition using spatio-temporal features. Journal of Visual Communication and Image Representation, 2014, 25 (2): p. 329-338.
- BENGIO, Y.; COURVILLE, A., et al. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2013, 35 (8): p. 1798-1828.
- BISHOP, C. M. Pattern Recognition and Machine Learning. Jordan, M.; Kleinberg, J.; Schölkopf, B (editores). New York, Springer, 2006.

- BOBICK, A. F. Movement, activity and action: the role of knowledge in the perception of motion. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 1997, 352 (1358): p. 1257-1265.
- BONEV, B. I. Feature Selection based on Information Theory. Ph.D. Thesis, Universidad de Alicante, 2010.
- BOUCHARD, G.; TRIGGS, B. Hierarchical part-based visual object categorization. En: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR): IEEE, 2005, p. 710-715.
- BREGONZIO, M. Representation and Recognition of Human Action in Video. Ph.D. Thesis, Queen Mary University of London, 2011.
- BUNKE, H. Graph matching: Theoretical foundations, algorithms, and applications. En: Proceeding of Vision Interface, 2000, p. 82-88.
- CAI, Z.; WANG, L., et al. Multi-View Super Vector for Action Recognition. En: IEEE Conference on Computer Vision and Pattern Recognition (CVPR): IEEE, 2014, p. 596-603.
- CARNEIRO, G.; LOWE, D. Sparse flexible models of local features. En: Proceedings of the European Conference on Computer Vision (ECCV'06): Springer, 2006, p. 29-43.
- CHAARAOUI, A. A.; CLIMENT-PÉREZ, P., et al. A review on vision techniques applied to Human Behaviour Analysis for Ambient-Assisted Living. Expert Systems with Applications, 2012, 39 (12): p. 10873-10888.
- CHAKRABORTY, B. Model free approach to human action recognition. Ph.D. Thesis, Universitat Autònoma de Barcelona, 2012.
- CHEN, M. Y.; HAUPTMANN, A. MoSIFT: Recognizing Human Actions in Surveillance Videos. Research Showcase, Computer Science Department, School of Computer Science, Carnegie Mellon University, 2009.
- CÓZAR, J. R.; HERNÁNDEZ, R., et al. Reducing Vocabulary Size in Human Action Classification. En: M. Graña, C. Toro, J. Posada, R. J. Howlett, L. C. Jain (editores). Frontiers in Artificial Intelligence and Applications: IOS Press, 2012, 243: p. 1712-1719.
- CRANDALL, D.; FELZENSZWALB, P., et al. Spatial priors for part-based recognition using statistical models. En: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR): IEEE, 2005, p. 10-17.
- CSURKA, G.; BRAY, C., et al. Visual categorization with bags of keypoints. En: Workshop on Statistical Learning in Computer Vision, in conjunction with ECCV, 2004, p. 1-2.
- DIAZ-ESPINOSA, D. A. Implementación y comparación de descriptores para búsqueda en video. Master Thesis, Universidad de Chile, 2010.

- DOLLÁR, P.; RABAUD, V., et al. Behavior Recognition via Sparse Spatio-Temporal Features. 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, (VS-PETS'05), 2005, (p. 65-72).
- DUDA, R. O.; HART, P. E., et al. Pattern Classification, 2nd edition. Willey-Interscience, New York, 2001.
- FELZENSZWALB, P.; HUTTENLOCHER, D. Pictorial structures for object recognition. International Journal of Computer Vision, 2005, 61 (1): p. 55-79.
- FERGUS, R.; ZISSERMAN, A., et al. Object class recognition by unsupervised scale-invariant learning. En: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: IEEE, 2003, p. II-264-II-271.
- GAUR, U.; ZHU, Y., et al. A "String of Feature Graphs" Model for Recognition of Complex Activities in Natural Videos. En: IEEE International Conference on Computer Vision (ICCV): IEEE, 2011, p. 2595-2602.
- GILBERT, A.; ILLINGWORTH, J., et al. Fast Realistic Multi-Action Recognition using Mined Dense Spatio-temporal Features. En: IEEE 12th International Conference on Computer Vision (ICCV): IEEE, 2009, p. 925-931.
- GONG, W., GONZÀLEZ, J., et al. Human Action Recognition based on Estimated Weak Poses. EURASIP Journal on Advances in Signal Processing, 2012, 2012 (1): p. 1-14.
- GONZÀLEZ, J.; VARONA, J., et al. aSpaces: Action Spaces for Recognition and Synthesis of Human Actions. En: F. J. Perales, E. R. Hancock (editores). Articulated Motion and Deformable Objects (AMDO): Springer, 2002, LNCS 2492: p. 189-200.
- GORELICK, L.; BLANK, M., et al. Actions as space–time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2007, 29 (12): p. 2247-2253.
- GRAUMAN, K.; LEIBE, B. Visual Object Recognition. Brachman, Ronald J.; Dietterich, Thomas G (editores). Morgan & Claypool Publishers, New York, 2011.
- GRIGORESCU, C.; PETKOV, N., et al. Contour and boundary detection improved by surround suppression of texture edges. Image and Vision Computing, 2004, 22 (8): p. 609-622.
- GUTHIER, T.; ŠOŠIĆ, A., et al. sNN-LDS: Spatio-temporal Non-negative Sparse Coding for Human Action Recognition. En: Artificial Neural Networks and Machine Learning - ICANN 2014: Springer, 2014, p. 185-192.
- HARRIS, C.; STEPHENS, M. A combined corner and edge detector. En: Alvey Vision Conference, 1988, p.147-151.
- HERNANDEZ-HEREDIA, Y. Modelo para la detección y reconocimiento de acciones humanas en videos a partir de descriptores espacio-temporales. Ph.D. Thesis, Universidad de las Ciencias Informáticas, 2013.

- JÉGOU, H.; DOUZE, M., et al. Aggregating local descriptors into s compact image representation. En: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'10): IEEE, 2010, p. 3304-3311.
- JENKINS, O. C.; MATARIÇ, M. J. Automated Modularization of Human Motion into Actions and Behaviors. Technical Report, USC Center for Robotics and Embedded Systems, 2002.
- JI, S.; XU, W., et al. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2013, 35 (1): p. 221-231.
- KADIR, T.; BRADY, M. Scale saliency: A novel approach to salient feature and scale selection. En: International Conference on Visual Information Engineering, 2000, p. 25-28.
- KIM, T.-K.; CIPOLLA, R. Canonical correlation analysis of video volume tensors for action categorization and detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2009, 31 (8): p. 1415-1428.
- KLÄSER, A.; MARSZALEK, M., et al. A Spatio-Temporal Descriptor Based on 3D-Gradients. En: British Machine Vision Conference (BMVC'08): British Machine Vision Association, 2008, p. 995-1004.
- KONG, Y.; ZHANG, X., et al. Adaptive learning codebook for action recognition. Pattern Recognition Letters, 2011, 32 (8): p. 1178-1186.
- LAPTEV, I. On Space-Time and Interest Points. International Journal of Computer Vision, 2005, 64 (2/3): p.107-123.
- LAPTEV, I.; LINDEBERG, T. Space-time interest points. En: Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03): IEEE, 2003, 1: p. 432-439.
- LAPTEV, I.; LINDEBERG, T. Local Descriptors for Spatio-Temporal Recognition. En: European Conference on Computer Vision (ECCV'04): Springer, LNCS 3024: 2004.
- LAPTEV, I.; LINDEBERG, T. Velocity adaptation of space-time interest points. En: Proceedings of the 17th International Conference on Pattern Recognition (ICPR): IEEE, 2004b, p. 52-56.
- LAPTEV, I.; CAPUTO, B., et al. Local velocity-adapted motion events for spatio-temporal recognition. Computer Vision and Image Understanding, 2007, 108 (3): p. 207-229.
- LAPTEV, I.; MARSZALEK, M., et al. Learning realistic human actions from movies. En: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'08): IEEE, 2008, p. 1-8.
- LEIBE, B.; LEONARDIS, A., et al. Robust object detection with interleaved categorization and segmentation. International Journal of Computer Vision, 2008, 77 (1-3): p. 259-289.

- LI, C.; SU, B., et al. Human Action and Recognition Using and Multi-Velocity STIPs and Motion Energy and Orientation Histogram. Journal of Information Science and Engineering, 2014, 30 (2): p. 295-312.
- LI, T.; MEI, T., et al. Contextual Bag-of-Words and for Visual and Categorization. IEEE Transactions on Circuits and Systems for Video Technology, 2011, 21 (4): p. 381-392.
- LINDEBERG, T. Feature detection with automatic scale selection. International Journal of Computer Vision, 1998, 30 (2): p. 79-116.
- LIU, D.; HUA, G., et al. Integrated feature selection and higher-order spatial feature extraction for object categorization. En: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'08): IEEE, 2008, p. 1-8.
- LIU, J.; SHAH, M. Learning Human Actions via Information Maximization. En: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'08): IEEE, 2008, p. 1-8.
- LIU, J.; ALI, S., et al. Recognizing human actions using multiple features. En: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'08): IEEE, 2008b, p. 1-8.
- LIU, J.; LUO, J., et al. Recognizing Realistic Actions from Videos in the Wild. En: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'09): IEEE, 2009, p. 1996-2003.
- LIU, J.; YANG, Y., et al. Learning semantic visual vocabularies using diffusion distance. En: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09): IEEE, 2009b, p. 461-468.
- LÓPEZ-MONROY, A. P.; GÓMEZ, M. M., et al. Bag-of-Visual-Ngrams for Histopathology Image Classification. En: IX International Seminar on Medical Information Processing and Analysis: International Society for Optics and Photonics, 2013, p. 89220P-89220P-12.
- LOWE, D. Distinctive image features from scale invariant key points. International Journal of Computer Vision, 2004, 60 (2): p. 91-110.
- MIKOLAJCZYK, K.; SCHMID, C. A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27 (10): p. 1615-1630.
- MOESLUND, T. B.; HILTON, A., et al. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding, 2006, 104 (2): p. 90-126.
- MORALES-GONZÁLEZ, A.; REYES, E. Assessing the role of spatial relations for the object recognition task. En: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: Springer, 2010, p. 549-556.

- OIKONOMOPOULOS, A.; PATRAS, I., et al. Spatiotemporal Salient Points for Visual Recognition of Human Actions. IEEE Transactions on Systems Man and Cybernetics, 2006, 36 (3): p. 710-719.
- ONEATA, D.; VERBEEK, J., et al. Action and Event Recognition with Fisher Vectors on a Compact Feature Set. En: IEEE International Conference on Computer Vision (ICCV): IEEE, 2013, p. 1817-1824.
- ÖZDEMIR, B.; AKSOY, S. Image classification using subgraph histogram representation. En: 20th International Conference on Pattern Recognition (ICPR): IEEE, 2010, p. 1112-1115.
- POPPE, R. A survey on vision-based human action recognition. Image and Vision Computing, 2010, 28 (6): p. 976-990.
- QUACK, T.; FERRARI, V., et al. Efficient mining of frequent and distinctive feature configurations. En: IEEE 11th International Conference on Computer Vision (ICCV'07): IEEE, 2007, p. 1-8.
- REN, H.; MOESLUND, T. B. Action Recognition Using Salient Neighboring Histograms. En: IEEE International Conference on Image Processing (ISIP): IEEE, 2013, p. 2807-2811.
- RYOO, M.; AGGARWAL, J. Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. En: 2009 IEEE 12th International Conference on Computer Vision: IEEE, 2009, p. 1593-1600.
- SCOVANNER, P.; ALI, S., et al. A 3-dimensional SIFT descriptor and its application to action recognition. En: Proceedings of the 15th International Conference on Multimedia: ACM, 2007, p. 357-360.
- SHAO, L.; MATTIVI, R. Feature Detector and Descriptor Evaluation in Human Action Recognition. En: Proceedings of the ACM International Conference on Image and Video Retrieval: ACM, 2010, p. 477-484.
- TA, A.; WOLF, C., et al. Recognizing and localizing individual activities through graph matching. En: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS): IEEE, 2010, p. 196-203.
- TRAN, S.N.; BENETOS, E., et al. Learning motion-difference features using Gaussian restricted Boltzmann machines for efficient human action recognition. En: 2014 International Joint Conference on Neural Networks (IJCNN): IEEE, 2014, p. 2123-2129.
- TURAGA, P.; CHELLAPPA, R., et al. Machine Recognition of Human Activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18 (11): p. 1473-1488.
- UIJLINGS, J.; SMEULDERS, A., et al. Real-time bag of words, approximately. En: Proceedings of the ACM International Conference on Image and Video Retrieval: ACM, 2009, p. 1-8.

- ULLAH, M. M.; PARIZI, S. N., et al. Improving bag-of-features action recognition with non-local cues. En: Procedings of the British Machine Vision Conference (BMVC'10): British Machine Vision Association, 2010, p. 95.1-95.11.
- WANG, H.; KLÄSER, A., et al. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 2013, 103 (1): p. 60-79.
- WANG, H.; ULLAH, M. M., et al. Evaluation of local spatio-temporal features and for action recognition. En: British Machine Vision Conference (BMVC'09): British Machine Vision Association, 2009.
- WEINLAND, D.; RONFARD, R., et al. Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding, 2006, 104 (2-3): p. 249-257.
- WEINLAND, D.; RONFARD, R., et al. A survey of vision-based methods for action representation, segmentation and recognition. Technical Report, INRIA, 2010.
- WILLEMS, G.; TUYTELAARS, T., et al. An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. En: European Conference on Computer Vision (ECCV'08): Springer, 2008, 2: p. 650-663.
- WONG, S.; CIPOLLA, R. Extracting spatiotemporal interest points using global information. En: IEEE 11th International Conference on Computer Vision (ICCV): IEEE, 2007, p. 1-8.
- YANG, Z.; PENG, Y., et al. Visual Vocabulary Optimization with Spatial Context for Image Annotation and Classification. En: K. Schoeffmann et al. (editores). Advances in Multimedia Modeling, MMM 2012, Springer, 2012, LNCS 7131: p. 89-102.
- YUAN, J.; WU, Y., et al. Discovery of Collocation Patterns: from Visual Words to Visual Phrases. En: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07): IEEE, 2007, p. 1-8.
- ZHANG, S.; TIAN, Q., et al. Generating Descriptive and Visual Words and Visual and Phrases for Large-Scale and Image Applications. IEEE Transactions on Image Processing, 2011, 20 (9): p. 2664-2677.
- ZHAO, Q.; IP, H. H. Unsupervised approximate-semantic vocabulary learning for human action and video classification. Pattern Recognition Letters, 2013, 34 (15): p. 1870-1878.