Ciencias de la Información Vol. 43, No. 2, mayo - agosto, pp. 21 - 27, 2012

Descriptores temporales-espaciales en la detección automática de información audiovisual

Yanio Hernández Heredia
José Ortiz Rojas
Ruber Hernández García
José María González Linares

La presente investigación tiene como objetivo proponer un método para la detección automática de puntos de interés en contenido audiovisual utilizando la información temporal y espacial. Este método es una alternativa novedosa que pueden aplicar los medios de comunicación en sus centros de documentación audiovisual, con el fin de lograr una localización del contenido audiovisual en videos de forma automática, lo que facilita la gestión de información. Igualmente es crucial su aplicación en sistemas para la búsqueda y catalogación de medias y sistemas que realicen gestión documental y archivística. En este trabajo se propone el uso de diferentes tipos de descriptores para la creación de vocabularios para tareas de detección de objetos diferentes en movimientos y acciones. El método supone que las clases de objetos o acciones son desconocidas por adelantado y explota las propiedades temporales y espaciales de los videos para la creación de un vocabulario que describe estas clases. Las características del espacio y el tiempo se han convertido recientemente en una representación popular de los vídeos para el reconocimiento de acciones y la detección objetos. El nuevo método presentado se compara con propuestas actuales para situaciones similares, obteniendo mejores resultados en la precisión y el rechazo de objetos o acciones falsas.

***Palabras clave**: detección de objetos, segmentación de videos, vocabulario, contenido audiovisual*

RESUMEN

ABSTRACT

This research aims at proposing a method for the automatic detection of points of interest in audio-visual content using temporary and spatial information. This method is a novel alternative that can be used by the media in their audio-visual documentation centers, in order to automatically locate the audio-visual content in videos, making information management easier. Likewise, its application is paramount in searching and cataloging systems for documents and files management. This paper proposes the use of different types of descriptors for creating vocabularies for object detection tasks in movements and actions. The method assumes that the classes of objects or actions are unknown and use time and space properties of videos in order to create a vocabulary to describe these classes. The characteristics of space and time have recently become a popular representation of videos for the recognition of actions and objects detection. The new method proposed is compared with current proposals for similar situations, obtaining better results in accuracy and rejection of false objects or actions.

***Keywords**: Object detection, video segmentation, vocabulary, audio-visual content*

Introducción

El creciente desarrollo de las tecnologías propicia que diariamente se incremente el número de personas e instituciones que las utilizan para interactuar por diversos fines, lo que sitúa al mundo hoy en medio de la llamada Sociedad de la Información. Esta acelerada tendencia de intercambio de información, fotos, videos hace cada vez más trascendental la gestión

de la información y el conocimiento.

En tal sentido, para los medios de comunicación se hace muy necesaria la utilización de centros para la documentación audiovisual con el fin de preservar y resguardar la memoria audiovisual del medio que lo produce, y a su vez el que permite procesar, recuperar y difundir eficientemente la información que sobre cualquier soporte físico disponga en su fondo documental (Echenagusía, 2008). Una de las principales dificultades que afrontan estos centros, para la gestión y catalogación de medias, es la detección automática de objetos.

Existen varias técnicas en la literatura mundial para la detección de objetos en videos e imágenes (Yingzi, 2004) (Alfredo, 2008), sin embargo no suelen tener buenos resultados cuando se utilizan en aplicaciones reales de análisis de video, como la video vigilancia o el monitoreo de señales de televisión. La mayor parte de las aproximaciones (Laptev, 2009) (R. Cipolla, 2008) segmentan los fotogramas extraídos por las cámaras como primer paso, identificando por un lado el fondo[1] de la escena y por otro lado, el primer plano[2] compuesto por objetos en movimiento. Luego se incluyen algoritmos de seguimiento para analizar la evolución de los objetos. Los objetos detectados mediante estas técnicas son representados como blobs[3] que identifican el área de la imagen ocupada por el objeto.

El reconocimiento de estos objetos, requiere el uso de técnicas avanzadas que combinan tres elementos esenciales para optimizar los resultados previstos en entornos que no sean controlados; teniendo en cuenta cambios de perspectivas, iluminación y colores, así como errores en la imagen que puedan aparecer, introducidas por la codificación de los videos:

- La selección adecuada de las características que se van a usar para representar el objeto.

- La representación compacta de estas características mediante descriptores.

- La construcción adecuada de un modelo del objeto que permita asimilar, convenientemente, cambios de forma en el mismo, cambios de iluminación, rotaciones, escalados y transformaciones de perspectiva, así como que sea robusto a errores y artefactos que aparecen al codificar videos.

Utilizar además la información temporal de las escenas (descripción de las acciones de la escena, movimiento, cambios de toma, tiros de cámara), unido a la información espacial (relación entre los elementos de la escena, próximo a, sobre qué), permite enriquecer los descriptores con un etiquetado semántico de la información del objeto.

El trabajo presentado en este documento evita problemas anteriores encontrados en la literatura por el procesamiento previo del video y una correcta selección de descriptores para diferentes tareas. Por lo tanto, fotogramas que contienen objetos ruidosos son rechazados con seguridad, sin comprometer la precisión de la técnica general. Entonces, las secuencias con objetos similares en el espacio son entrenadas con las palabras correctas en un clúster y obtener así la mejor clasificación posible. Como resultado de esta identificación, se detecta la posición temporal de los objetos en el video. El resto del documento está organizado como sigue: sección 2 introduce una novedosa técnica para la detección de objetos mediante la formación de vocabularios de descriptores espacio-temporales. En la sección 3 la técnica se ha probado con un conjunto de videos y se compara con otras. Por último, en la sección 4 son presentadas las conclusiones y el trabajo futuro.

Metodología

La captura de características del espacio y el tiempo describe las formas y el movimiento en un video, además proporciona una representación independiente de los acontecimientos con respecto a cambios espacio-temporales y cambios de escalas, diferencias de fondos y movimientos múltiples en una escena (ver figura 1). Estas características suelen ser extraídas directamente del video y por lo tanto evitan los posibles errores de un método de pre-procesamiento, como la segmentación de movimiento y de seguimiento.

La representación, la detección y el aprendizaje son los principales problemas que deben abordarse en el diseño de un sistema visual para el reconocimiento categorías de objetos. El desafío de la detección es la definición de métricas y algoritmos que sean adecuados para hacer coincidir los modelos a las imágenes en presencia de la oclusión.

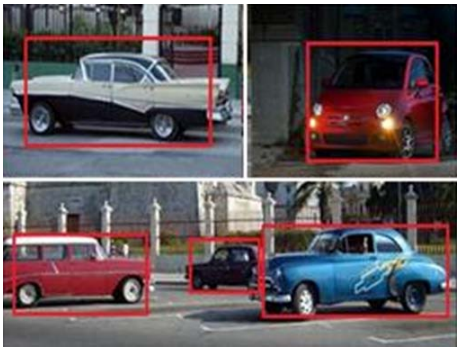

Figura 1. Imágenes con objetos tipo carro detectados en diferentes perspectivas. a) ambiente controlado. b) Carro frontal y atenuación de sombras por oscuridad. c) Varios carros en diferentes tamaños y posiciones.

Varios autores (I. Laptev M. M., 2008); (R. Cipolla, 2008) centran su atención en los pasos para eliminar las diferencias de fondo y normalizar los ejemplos de entrenamiento; el reconocimiento a menudo procede por una exhaustiva búsqueda de posición de la imagen. Los enfoques probabilísticos (Michael C. Burl, 2001), con modelos aleatorios donde se combinan varias piezas, producen principios y métodos eficientes de detección. El autor (M. Weber, 2002) propone un algoritmo con una máxima de verosimilitud de aprendizaje no supervisado para diferentes categorías de objetos que es un ejemplo de lo planteado anteriormente.

La figura 2 muestra el procedimiento a seguir con los videos de entrada en el framework propuesto para la detección de los objetos

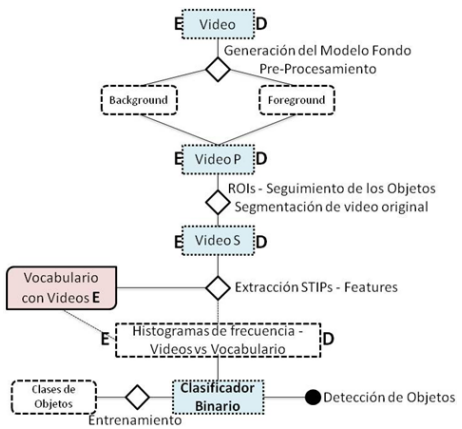

Figura 2. Diagrama del modelo propuesto. E-Videos de Entrenamiento. D – Videos donde se encuentran los objetos que se quieren detectar.

[1] Contenido estático donde se encuentra la información complementaria del primer plano.

[2] Contiene la mayor cantidad de información que identifica la secuencia de video.

[3] Manchas sobre objetos detectados para el tracking en sistemas de video vigilancia.

Para el modelo se utilizan dos grupos de clases de videos, las secuencias de entrenamiento y las de detección. Se aplica un pre procesamiento para separar el fondo de la imagen donde se encuentran los objetos a clasificar, y así ganar precisión y tiempo en el procesamiento principal posterior de la extracción de STIPs[4] y descriptores.

Cuando se tienen los vectores de descriptores de cada video, se utilizan los de entrenamiento para crear un vocabulario con un Kmean, y luego los histogramas de frecuencia de la aparición de los videos (entrenamiento y detección) con respecto al vocabulario, son los dos grupos de datos que necesita el clasificador binario para entrenar y detectar las posibles clases de objetos que se tienen y desean.

A. Extracción de STIPs y descriptores

Después de obtener las áreas de interés de los videos que se analizan durante el pre procesamiento donde se encuentran los objetos a seguir, se pasa a obtener los puntos de interés que caracterizan y definen a los objetos. En la actualidad existen varios algoritmos para detectar puntos de interés, algunas variantes basadas en técnicas para detectar puntos en imágenes, como Harris (I. Laptev T. , 2004) o Hessian (G. Willems, 2008) y otros que utilizan el espacio y el tiempo directamente para detectar puntos de interés en secuencias de video, como el detector Cuboid (P. Dollar, 2005).

Para modelar una secuencia de imágenes espacio-temporal f, se construye su representación lineal escala-espacial por la convolución[5] de con un Kernel anisotropico gausiano con distintas varianzas espaciales y temporales τ_l^2.

$$L(\cdot;\sigma_l^2,\tau_l^2) = g(\cdot;\sigma_l^2,\tau_l^2) * f(\cdot) \qquad (1)$$

Luego se considera una matriz de segundo orden espacio temporal 3 x 3, compuesta por el promedio de las primeras derivadas espacio-temporales de las funciones gaussianas de peso.

$$\mu = g(\cdot;\sigma_l^2,\tau_l^2) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix} \quad (2)$$

Por último para detectar los puntos de interés, se buscan regiones en la función f que tengan valores propios significantes $\lambda_1\lambda_2\lambda_3$ de μ.

$$H = det(\mu) - k * trace^3(\mu)$$

$$H = \lambda_1\lambda_2\lambda_3 - K(\lambda_1 + \lambda_2 + \lambda_3)^3 \quad (3)$$

Cuando se tienen los puntos de interés que mejor describen los videos, es necesario extraer los descriptores con los pixeles obtenidos por cada imagen de la forma (x, y, t) que mejores resultados puedan arrojar según los tipos de objetos en movimiento que se desean obtener.

Luego que se obtienen los puntos de interés que mejor describen las escenas de los videos, se extraen descriptores asociados a los pixeles en el tiempo. Según el diccionario de la lengua española el término descriptor define el contenido de un documento y permite localizarlo en un archivo manual o informatizado. En la informática, en las temáticas relacionadas con las medias, los descriptores visuales se refirieren a los descriptores de audio y a los de video.

A continuación se presentan algunas propiedades a tener presente en el momento que se desea seleccionar un descriptor:

Simplicidad: El descriptor debería representar las características extraídas de la imagen de manera clara y sencilla para permitir una fácil interpretación de su contenido.

Repetibilidad: El descriptor generado a partir de una imagen debe ser independiente del momento en el que se genere.

Diferenciabilidad: Dada una imagen, el descriptor generado debe poseer alto grado de discriminación respecto de otras imágenes y al mismo tiempo contener información que permita establecer una relación entre imágenes similares.

Invarianza: Cuando existen deformaciones en la representación de dos imágenes, es deseable que los descriptores que las representan aporten la robustez necesaria para poder relacionarlas aún bajo diferentes transformaciones.

Los descriptores de video son funciones matemáticas que reducen la información contenida en el video, manteniendo el máximo de información relevante y generalmente se expresan mediante vectores sobre algún espacio matemático particular. Los descriptores de contenido de bajo nivel hacen referencia a la información más básica del material audiovisual. En esta investigación se utilizan estos descriptores, por ser los que brindan la información necesario para clasificar diferentes tipos de clases, teniendo cuenta las propiedades mencionadas anteriormente. Estas informaciones pueden ser características visuales como el color, la textura, la forma o el movimiento del contenido asociado a las imágenes. En la actualidad los descriptores de tiempo, asociados a la información cíclica de algunos objetos, han revolucionado la selección y extracción de descriptores para la gestión audiovisual; hasta hace muy poco solamente se utilizaban descriptores espaciales.

Para los resultados de este artículo se hicieron diferentes pruebas con algunos descriptores espacio temporales y según las bases de datos utilizadas[6] y las variaciones para la obtención del vocabulario, el descriptor de mejores resultados fue el MoSIFT[7], variante de flujo óptico del SIFT. También se hicieron pruebas con descriptores como eSURF, Histograma Orientado del Gradiente (HOG) e Histograma de Flujo Óptico (HOF), así como una variante de un vector con la unión HOG/HOF propuesta en (I. Laptev M. M., 2008).

B. Creación del Vocabulario y los Histogramas de Frecuencia

El objetivo principal de este artículo es la creación de un Vocabulario con las palabras precisas que al ser comparada con secuencias de video, de cómo salida el histograma que mejor lo describa.

[4] Puntos de interés en un video o una imagen.

[5] Operador matemático que transforma dos funciones f y g en una tercera función que en cierto sentido representa la magnitud en la que se superponen y una versión trasladada e invertida de g.

[6] Colecciones de videos que se utilizan para probar algoritmos para el procesamiento automático de archivos audiovisuales.

[7] http://lastlaugh.inf.cs.cmu.edu/libscom/downloads.htm

Los vocabularios que en ingles son conocidos como Bag of Words (BoW), es una técnica usada en varios campos como Procesamiento de Lenguaje Natural, Recuperación de la Información o Análisis de Patrones (Wallach, 2006), consistente en la representación de un documento mediante un conjunto no ordenado con las frecuencias de aparición de las palabras de un diccionario contenidas en dicho documento.

Para la creación del vocabulario se utilizan las características (pueden ser todas o una cantidad que de los mejores resultados según pruebas) extraídas de los videos en este caso de aplicación. Con estas características se crea un clúster de entrenamiento (conjunto de datos paralelizados) que agrupa los descriptores similares para obtener el nombrado vocabulario. Para el modelo se utiliza un Kmean[8] en la creación del vocabulario con una cantidad de palabras o contador del clúster, igual a la siguiente fórmula, que durante todas las pruebas dio los mejores resultados, sin sobrecargar ni dejar por debajo cada palabra respecto a la cantidad de descriptores que la conforman, así como 6 ejecuciones del Kmean para un clúster más efectivo:

$$BOWKMeansTrainer\ (Contador\_Cluster\ (Palabras) = Cant\_Descriptores * 0.04) \quad (4)$$

$$Cant_{Descriptores} = 0.3 * Total_{Descriptores})$$

Para mejorar los resultados se pueden aplicar algunas extensiones como por ejemplo:

- Eliminar palabras visuales demasiado comunes (Stop Word Removal)

- Selección de las palabras visuales más informativas basándose en la frecuencia de aparición en todos los documentos, o la correlación entre una palabra y una clase de documentos (usando estadísticas x^2, ganancia de información o información mutua).

- Utilizar información espacial teniendo en cuenta la posición del descriptor dentro de la imagen (restricciones geométricas)

- Utilizar bi-gramas visuales que indiquen la proximidad espacial de dos palabras distintas (usando histogramas de co-ocurrencia) (Wallach, 2006)

Con el vocabulario salvado y entrenado con el clúster que se realiza, se crean los histogramas de frecuencia de cada uno de los videos que se utilizan en el modelo (entrenamiento y detección). Para lograr estos histogramas, que no son más que vectores con dimensión igual al número de palabras que indican en cada posición cuanto se parece ese video al clúster, se aplica un emparejamiento jerárquico entre descriptores y el vocabulario. Para esta prueba se utilizó el «Radius-Match» que encuentra al mejor emparejamiento para cada descriptor de consulta que tenga menor distancia dado un umbral, este paso asegura eliminar del vector, ocurrencias muy alejadas del clúster con el Descriptor-Matcher BruteForce.

A. Detección de los objetos en una secuencia de video

Cuando se tienen los dos conjuntos de histogramas de frecuencia de los videos de entrenamiento y detección, se pasa a entrenar un clasificador supervisado que sirva para tener etiquetadas las clases y realizar la mejor detección posible.

Para este paso en el modelo se podría haber utilizado como clasificador los arboles binarios de búsqueda, con muy buenos resultados en (Lempitsky, 2009) o máquinas de puntos bayesianos (Herbrich, 2001), sin embargo los datos que se tienen en los vectores son Floats de 32 bits muy regulares, y es mucho más fácil de implementar una máquina de soporte vectorial (SVM), muy utilizada en la visión por computador por sus diversas formas de aplicación y variantes de implementación.

Las máquinas de soporte vectorial o máquinas de vectores de soporte son un conjunto de algoritmos de aprendizaje supervisado equipo en los laboratorios AT&T (Dragonfly Interactive, 2008).

En la actualidad es muy común el uso de herramientas, librerías o aplicaciones con soporte para diversos lenguajes de programación o con propias interfaces a partir de ficheros con los datos de entrenamiento y regresión (Joachims, 2008). Son muy fáciles de utilizar y dan a los usuarios muchas opciones. Para las pruebas realizadas con el framework, se utilizó la librería LibSVM (Lin, 2010), con la modalidad clase múltiple uno contra todos para el entrenamiento y un Kernel pre-calculado de la forma:

$$K(H_i,H_j) = \exp\left(-\frac{1}{2A}\sum_{n=1}^{V}\frac{(h_{in}-h_{jn})^2}{(h_{in}+h_{jn})}\right) \quad (5)$$

Donde H_i y H_j son los histogramas de frecuencia y V el tamaño del vocabulario.

Resultados experimentales

Para probar el método, se utilizaron bases de datos de videos de acciones, ya que nuestra técnica está diseñada para escenas y videos reales, objetos en imágenes no es nuestra intención, como se ha explicado los descriptores que construyen el vocabulario dependen de la información temporal y por tanto se necesitan objetos que se van desplazando en el tiempo.

Primero se probó con la base de datos KTH[9], muy utilizada para pruebas iníciales de un método por ser muy sencilla, contiene clases de acciones de personas en ambientes controlados y con pequeñas resoluciones.

Para analizar la relación entre los descriptores y cantidad de palabras del vocabulario las

Tabla 1. Precisión promedio de varios métodos utilizando combinaciones de detectores/ descriptores en la base de datos KTH.

	HOG/HOF	HOG3D	MoSIFT	Cuboids	ESURF
Harris3D	**91.8%**	89.0%	-	-	-
Cuboid	88.7%	**90.0%**	-	89.1%	-
Hessian	**88.7%**	88.3%	-	-	81.4%
MoSIFT	89.5%	84.28%	**95.83%**	-	-
Vharris[10]	**92.13%**	-	91.7%	-	-

[8] *Método de análisis de clústeres que tiene como objetivo la partición de n observaciones en k grupos en los que cada observación pertenece al grupo más cercano de la media de todos los clúster.*

[9] *http://www.nada.kth.se/cvap/actions/*

[10] *El Método propuesto. Vocabulario con Detector Harris3D y descriptores HOG/HOF y MoSIFT*

siguientes gráficas muestran como varia la precisión de acierto de las clases en la base de dato KTH, así como la confusión entre ellas por similitud de los descriptores ya que son acciones muy parecidas.

Luego se hicieron pruebas con una base de datos más compleja (Hollywood Dataset[11]) de videos con resoluciones y ambientes reales. Esta base de datos contiene 10 clases distintas con diferentes escenas de películas con gran cantidad de movimiento y diversidad de fondo, la tabla 3 muestra los resultados comparando con los métodos anteriores.

También se realizaron pruebas con objetos convencionales como automóviles y animales en videos normales que fueron perfectamente detectados para una precisión de más de un 80 %.

Aplicaciones en la detección automática de información audiovisual

Los archivos audiovisuales realizan tareas de colección, de gestión, de conservación y de acceso al patrimonio audiovisual del que se ocupen. Para realizar el análisis documental de estas medias que manejan, se realiza un resumen y se indizan, utilizando un lenguaje controlado, lo que permite el acceso al contenido del documento con el objeto de recuperarlo, explotarlo y difundirlo (Gastaminza, 2005).

Para un archivo audiovisual la gestión del patrimonio audiovisual se facilita al realizar el proceso de detección de objetos de forma automática. Este método permite realizar una pre-catalogación de la información que procesan los sistemas de monitoreo, catalogación y búsqueda de información audiovisual. A partir del entrenamiento con las diferentes clases de objetos que se identifiquen, se procesan los videos para detectar información y agregar estos metadatos a los archivos, creando marcas de su posible localización teniendo en cuenta el entrenamiento previo. Este proceso agiliza el trabajo y evita que se cometan errores en la gestión que deben realizar los operadores en estos centros de documentación audiovisual.

Tabla 2. Diferentes pruebas cambiando la relación entre descriptores y clústeres del vocabulario.

Detector	Descriptor	Cantidad Descriptores	Palabras del vocabulario x Cant. Descriptores	Ejecuciones Kmeans	Accuracy
Vharris	**HOG/HOF**	**30%**	**4%**	**6**	**92,13%**
Vharris	HOG/HOF	60%	1%	3	89,50%
Vharris	HOG/HOF	100%	2%	2	75,60%
MoSIFT	MoSIFT	60%	2%	8	83,70%
MoSIFT	MoSIFT	30%	4%	4	89,20%

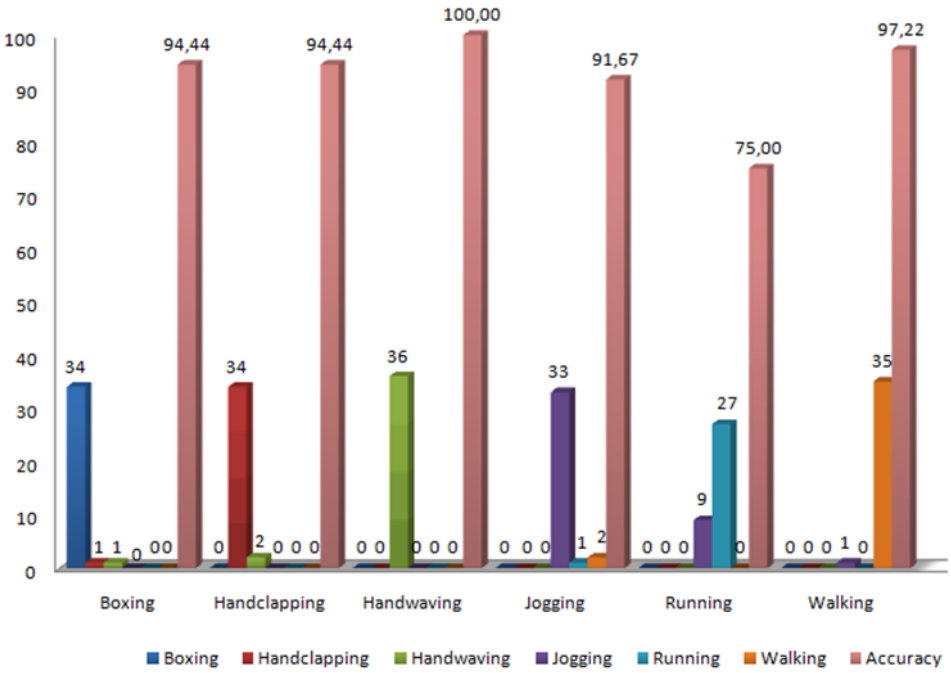

Figura 3. Resultado por clases de la Base de Datos KTH. Boxing-Handclapping-Handwaving-Jogging-Walking.

Tabla 3. Precisión promedio de varios métodos utilizando combinaciones de detectores/ descriptores en la base de datos Hollywood.

	HOG/HOF	HOG3D	MoSIFT	Cuboids	ESURF
Harris3D	**45.2%**	43.7%	-	-	-
Cuboid	**46.2%**	45.7%	-	45.0%	-
Hessian	**46.0%**	41.3%	-	-	38.2%
Vharris[12]	47.9%	-	**49.2%**	-	-

En los sistemas de video vigilancia, la aplicación en la visualización de las cámaras de este método, permite automatizar el envío de alarmas ante situaciones previamente entrenadas de situaciones o escenas de video. Un objeto abandonado, un hecho de vandalismo, una persona buscada por las autoridades, son algunas de las aplicaciones que se les pueden agregar a los visores de los sistemas de vigilancia inteligente basados en cámaras IP. En la actualidad estos sistemas son llamados sistemas de video vigilancia de tercera generación y no se conciben aplicaciones de alto impacto en la seguridad ciudadana si no cuentan con detección automática como uno de sus componentes.

[11] *http://www.irisa.fr/vista/Equipe/People/Laptev/download.html*

[12] *El Método propuesto. Vocabulario con Detector Harris3D y descriptores HOG/HOF y MoSIFT*

Conclusiones

En este artículo, se ha propuesto un método para el aprendizaje de la estructura espacial y temporal de una categoría de objeto visual para reconocer nuevos objetos de esa categoría, localizarlos en escenas reales y automáticamente obtener los segmentos del fondo. Se han proporcionado algoritmos eficientes para cada uno de los pasos y se ha evaluado el rendimiento resultante de reconocimiento en varios conjuntos de datos. Los resultados muestran que el método funciona bien en categorías diferentes de objetos a diferentes escalas y logra un buen rendimiento de segmentación y detección de objetos en difíciles escenas reales de sistemas de video vigilancia o monitoreo y catalogación de medias.

Una contribución importante es la integración de una segmentación de los videos con la selección adecuada de descriptores que puedan caracterizar lo mejor posible a las clases de objetos, así como la creación de un vocabulario con las palabras exactas para el emparejamiento adecuado con los videos de entrenamiento y los videos de prueba. Así, la fase inicial de reconocimiento no sólo inicializa el proceso de segmentación con la ubicación de un objeto posible, sino también proporciona una estimación de las mediciones locales y de su influencia sobre la hipótesis del objeto.

Este mecanismo constituye una opción aplicable para la detección de información audiovisual de forma automática en videos reales de video vigilancia o monitoreo de señales televisivas y mejora resultados actuales de modelos similares, constituyendo una técnica más precisa que las basadas en características espaciales de las imágenes solamente. Esta aproximación es lo suficientemente flexible como para poder combinar la información de los descriptores según el tipo de videos con la cantidad de palabras del vocabulario y el tipo de emparejamiento a utilizar. El tiempo de ejecución del enfoque resultante depende principalmente de tres factores: modelo de complejidad (variación de los objetos con respecto al fondo), tamaño de los videos analizados (dimensiones) y el rango de la escala de búsqueda seleccionado.

La aplicación sobre los sistemas de catalogación y gestión archivística, usados en videotecas o televisoras y la influencia en la automatización de parte de la gestión audiovisual al detectar información importante y agregarla como metadatos automáticamente, acelerando la búsqueda parcial o total de datos en los videos, es un resultado alcanzable con este método.

Con un sistema de búsqueda referenciada, se acelera considerablemente la gestión de la información audiovisual en este tipo de sistemas, pues datos generados automáticamente en una primera pasada se agregan a la base de datos de videos, por ejemplo se podrían buscar personas corriendo o realizando acciones específicas en videos para estudios asociados a temáticas de comportamientos de la población en gran cantidad de datos audiovisuales.

Posibles ampliaciones incluyen la integración y la combinación de varios detectores de discriminación multicategorías y la fusión de los descriptores que mejor se adapten a cambios de perspectivas, iluminación y colores aprovechando el tipo de material analizado. Por último, se pudiera también incorporar pruebas con otros clasificadores binarios para el entrenamiento y detección, así como una máquina de soporte vectorial en cascada.

Rerefencias

Alfredo, M. (2008). Vision AIBO. Recuperado el 16 de 01 de 2010, de http://cannes.itam.mx/Alfredo/Espaniol/Cursos/Robotica/Material/VisionAIBO.pdf

C. Schüldt, I. L. (2004). Recognizing human actions: A local SVM approach. ICPR.

Dragonfly Interactive, L. (2008). Nec Laboratories, INC America. (Nec-Labs) Recuperado el 2011, de http://www.nec-labs.com/research/machine/ml_website/person.php?person=vlad

G. Willems, T. T. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. ICCV.

Gool, L. V. (2007). Bag of visual words model: recognizing object categories. England: Oxford University.

Herbrich, R. (2001). Bayes Point Machines. Journal of Machine Learning.

I. Laptev, M. M. (2008). Learning realistic human actions from movies. CVPR.

I. Laptev, T. (2004). Space-time interest points. ICCV.

Joachims, T. (2008). Svm Struct. (Cornell University) Recuperado el 2010, de http://svmlight.joachims.org/svm_struct.html

Laptev, I. (2009). histograms, Improving object detection with boosted. Recuperado el 02 de 2010, de www.elsevier.com/locate/imavis

Lempitsky, J. G. (2009). Class-Specific Hough Forests for Object Detection. Recuperado el 15 de 03 de 2010, de http://www.vision.ee.ethz.ch/~gallju/download/jgall_houghforest_cvpr09.pdf

Lin, C.-C. C.-J. (2010). LIBSVM — A Library for Support Vector Machines. Recuperado el 2011, de http://www.csie.ntu.edu.tw/~cjlin/libsvm/

M. Weber, M. W. (2002). Unsupervised Learning of Models for Recognition. Recuperado el 2011, de http://www.vision.caltech.edu/CNS179/papers/Perona00.pdf

Michael C. Burl, M. W. (2001). A probabilistic approach to object recognition using local photometry and global geometry. Recuperado el 2011, de http://www.springerlink.com/content/n163148470354655/

P. Dollar, V. R. (2005). Behavior recognition via sparse spatio-temporal features. VS-PETS.

R. Cipolla, J. S. (2008). Semantic texton forests forimage categorization and segmentation. Alaska: CVPR.

R. Fergus, P. P. (2004). Object class recognition by unsupervised scale-invariant learning. CVPR, 2004.

Rüping, S. (2010). mySVM. (Technische Universität Dortmund) Recuperado el 2011, de http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html

S. Lazebnik, C. S. (2007). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. CVPR, 2006.

Referencias

Shotton, J. M. (2007). The layout consistent random field for recognizing and segmenting partially occluded objects. CVPR.

Triggs, N. D. (2005). Histograms of oriented gradients for human detection. EUA: IEEE Conference on Computer Vision.

Wallach, H. M. (2006). Topic Modeling: Beyond Bag-of-Words. Cambridge: Cavendish Laboratory, University of Cambridge.

Yingzi, D. (2004). Unsupervised approach to color video thresholding. Recuperado el 12 de 01 de 2 2010, de Optical Engineering, Vol. 43, No 2: http://www.engr.iupui.edu/~yidu/pub.html

Recibido: 13 de diciembre de 2011.
Aprobado en su forma definitiva: 10 de febrero de 2012

MSc. Yanio Hernández Heredia
Universidad de las Ciencias Informáticas (UCI)
País: Cuba
Correo electrónico: <yhernandezh@uci.cu>

Dr.C. José Ortiz Rojas
Universidad de las Ciencias Informáticas (UCI)
País: Cuba
Correo electrónico: <jortiz@uci.cu>

Dr.C. Ruber Hernández García
Universidad de las Ciencias Informáticas (UCI)
País: Cuba
Correo electrónico: <rhernandezg@uci.cu>

Dr.C. José María González Linares
Universidad de Málaga
País: España
Correo electrónico: <jgil@ac.uma.es>