

Propuesta de Tesina para la obtención del grado  
Licenciado en Ciencias de la Computación  
Detección de eventos en videos: análisis de partidos  
de Rugby

9 de diciembre de 2015

**Postulante:** Martín Reixach, Lucía Norma

**Director:** Granitto, Pablo

**Codirector:** Larese, Mónica

## 1. Motivación

La producción de material audiovisual ha sufrido importantes cambios en las últimas décadas. La reducción de costos junto con la aparición de nuevas tecnologías permitieron masificar el acceso a medios de grabación y almacenamiento, modificando rotundamente la forma en que se trabajaba con videos y generando tanto nuevas herramientas como desafíos y necesidades.

Con todos estos cambios, las cámaras de video que originalmente habían sido concebidas para la producción de contenidos de entretenimientos pasaron a formar parte de la vida cotidiana.

Ya sea a través de videograbadoras o usando cámaras integradas en otros dispositivos como celulares, computadoras o cámaras de fotos, los videos comenzaron a utilizarse con nuevos fines, incluídos monitoreo, control de seguridad, comunicación, medicina, almacenamiento de recuerdos personales, y educación entre otros.

Ésto trajo aparejado un incremento exponencial en la cantidad de información generada día a día, haciendo inviable el reconocimiento manual de las imágenes obtenidas. Como consecuencia de ésto, los esfuerzos por automatizar la detección de eventos en videos crecieron, ya que abrían un abanico de

nuevas posibles aplicaciones, tanto para videos obtenidos previamente como para el análisis en tiempo real.

En este contexto, el estudio del movimiento humano tuvo un lugar muy importante. Sin embargo, esta tarea no es simple ya que hay importantes disparidades en los movimientos, variaciones entre los distintos escenarios y grandes diferencias entre individuos. Para superar estas dificultades, se ha trabajado desde distintos enfoques y con distintos algoritmos, siempre buscando el balance entre precisión y costo computacional.

Los primeros trabajos fueron realizados con videos especialmente creados para este fin, donde había un único individuo realizando movimientos repetitivos frente a la cámara y el fondo era prácticamente liso para que no interfiriera en el análisis, como KTH \*\*\*. Una vez que hubo más avances, se empezaron a estudiar conjuntos de videos más complejos, donde entraban en juego nuevos desafíos, como cambios de cámara, escenarios cambiantes, multiples objetos por cuadro, diferencias de luz y escala, entre otros. Algunos de ellos fueron \*\*\*.

Los resultados obtenidos en estos trabajos confirman los avances que hubo en las áreas de reconocimiento y clasificación en las últimas décadas.

## **2. Objetivo General**

En una primera instancia, este trabajo busca estudiar técnicas y métodos que permitan aplicar el modelo de *Bag of Words* al reconocimiento y la clasificación de videos. Una vez que se hayan adquirido estos conocimientos, el propósito es aplicar estas herramientas al análisis de imágenes de video correspondientes a la práctica del Rugby.

El objetivo de todo ésto es automatizar la detección de ciertas escenas de juego que se dan a lo largo de los partidos de Rugby, permitiendo a entrenadores y jugadores poder concentrarse en momentos significativos del juego sin necesidad de estudiar las cintas completas.

## **3. Fundamentos y estado del conocimiento sobre el tema**

La clasificación de imágenes o videos en categorías semánticas es un problema de interés tanto para la comunidad científica como para la industria. La detección de diferentes tipos de escenas por lo general se basa en vectores de características que describen el color y la textura de las imágenes entre otras propiedades visuales.

Hace ya más de una década, comenzó a verse una tendencia a usar *keypoints* y puntos de interés local en la recuperación y clasificación de la información contenida en las imágenes. \*\*\*\*\* Los *keypoints* son zonas destacadas de las imágenes, que contienen abundante información local acerca de la misma, los cuales pueden ser identificados usando diferentes detectores \*\*\* y representados por diversos descriptores.\*\*\*

Una vez que los *keypoints* son obtenidos, éstos son distribuidos en una gran cantidad de *clusters*, asignando a un mismo *cluster* aquellos de características similares. Cada *cluster* es considerado una *visual word* que representa el patrón local específico compartido por todos los *keypoints* de ese *cluster*, esto permite obtener un vocabulario de *visual words* que describe todos los patrones locales de las imágenes. A partir de esto, una imagen puede representarse como una *Bag of Visual Words*, un vector que contiene la cantidad de veces que cada *visual word* aparece en la imagen, el cual es usado como vector de características durante la clasificación.

Esta representación es análoga a la de Bag of Words utilizada en textos para describir tanto la forma como la semántica. Esto permite que muchas de las técnicas ya desarrolladas y usadas para el análisis de textos hayan podido ser aplicadas al trabajo con imágenes.

Uno de los métodos más utilizados para la detección de *keypoints* es la Diferencia Gaussiana (DoG) .... Los *keypoints* suelen encontrarse en los ángulos y bordes de los objetos presentes en las imágenes. BRISK ver

En cuanto a los descriptores, a lo largo de los años se han publicado numerosos desarrollos. Para trabajar con imágenes, alguno de los más utilizados son SIFT[6] y FREAK [1]. En lo referente a videos, suelen utilizarse descriptores que se basan en los de imágenes e incorporan el factor temporal, como MoSIFT [3] y MoFREAK [9].

Una vez que se completa el proceso de obtención de información de los objetos a clasificar y construimos nuestro *Bag of Visual Words*, se llega a la etapa de clasificación propiamente dicha. Para completar esta última etapa uno de los métodos más utilizado actualmente es *Support Vector Machine* (SVM).

Se han hecho diversas pruebas en videos utilizando la idea de *Bag of Visual Words*, algunas con datasets de videos específicos, con movimientos y escenas acotados, creados para el uso académico como los datasets KTH y HMDB51 [8, 5, 9] y otros con clips pertenecientes a películas y deportes [2, 7, 4]. En ambos casos, los resultados fueron alentadores, mostrando el potencial de esta manera de abordar la detección y clasificación.

## 4. Objetivos específicos

Proponemos desarrollar una herramienta que nos permita identificar acciones específicas en videos correspondientes a partidos de Rugby. Nos interesa diferenciar *lines* y *scrums* del resto de los *momentos de juego*.

Para lograrlo nos basaremos en el método de *Bag-of-Words* junto a descriptores de imágenes y videos (en este trabajo el audio del video no será tenido en cuenta).

También se analizará la *performance* de los métodos de detección usados y se compararán los resultados obtenidos con los alcanzados utilizando otros métodos.

## 5. Metodología y Plan de Trabajo

Para alcanzar los objetivos propuestos proponemos las siguientes tareas (programa tentativo de trabajo):

- Estudio de los métodos más usados para análisis de videos e imágenes. *6 semanas*
- Implementación de los algoritmos necesarios. *4 semanas*  
La implementación se hará de manera progresiva, incorporando las herramientas que sean necesarias para obtener información más clara.
- Evaluación del funcionamiento de los métodos sobre los videos de nuestra base de datos y ajuste de los parámetros según sea necesario. *4 semanas*
- Síntesis de los resultados obtenidos y escritura de la tesina. *6 semanas*

El trabajo se realizará durante aproximadamente 5 meses con una dedicación de más de 25 hs. semanales.

## Referencias

- [1] Alexandre Alahi, Raphael Ortiz y Pierre Vandergheynst. “Freak: Fast retina keypoint”. En: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, págs. 510-517.
- [2] Liang-Hua Chen y col. “Violence detection in movies”. En: *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on*. IEEE. 2011, págs. 119-124.

- [3] MY Chen y A Hauptmann. “MoSIFT: Recognizing Human Actions in Surveillance Videos. Research Showcase”. En: *Computer Science Department, School of Computer Science, Carnegie Mellon University* (2009).
- [4] O Deniz y col. “Fast Violence Detection in Video”. En: *The 9th International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, Portugal*. 2014.
- [5] Hildegard Kuehne y col. “HMDB: a large video database for human motion recognition”. En: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE. 2011, págs. 2556-2563.
- [6] David G Lowe. “Distinctive image features from scale-invariant keypoints”. En: *International journal of computer vision* 60.2 (2004), págs. 91-110.
- [7] Enrique Bermejo Nievas y col. “Violence detection in video using computer vision techniques”. En: *Computer Analysis of Images and Patterns*. Springer. 2011, págs. 332-339.
- [8] Christian Schödl, Ivan Laptev y Barbara Caputo. “Recognizing human actions: a local SVM approach”. En: *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. Vol. 3. IEEE. 2004, págs. 32-36.
- [9] Chris Whiten, Robert Laganieri y G-A Bilodeau. “Efficient action recognition with MoFREAK”. En: *Computer and Robot Vision (CRV), 2013 IEEE International Conference on*. IEEE. 2013, págs. 319-325.