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

Abril de 2016

Postulante: Martín Reixach, Lucía

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, educación y almacenamiento de recuerdos personales 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 el análisis offline como 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 los 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 [16]. 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 dinámicos, diferencias de luz y escala y multiples objetos por cuadro entre otros, como los trabajos realizados en la Universidad de Castilla-La Mancha [14, 7].

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 contendida en la imágenes [6, 11]. 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 respresentados 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*, ésto permite

obtener un vocabulario de *visual words* que describe todos los patrones locales de las imágenes. A partir de ésto, una imagen puede respresentarse 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.

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

Los *keypoints* suelen encontrarse en los ángulos y bordes de los objetos presentes en las imágenes. Uno de los métodos más utilizados para su detección es la Diferencia de Gaussianas (DoG), sin embargo existen otros como Harris Corner Detector [8], Fast Hessian Detector [3], AGAST [13] y multi-scale AGAST [10].

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 [12], SURF [3], BRISK [10] y FREAK [1]. Por lo general, cada descriptor se utiliza acompañado de un detector específico. En lo referente a videos, suelen utilizarse descriptores que se basan en los de imágenes e incorporan el factor temporal, como MoSIFT [5] y MoFREAK [17].

Una vez que se completa el proceso de obtención de información de los objetos a clasificar y se construye la *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). La *performance* del SVM puede variar en gran medida de acuerdo al *kernel* elegido, algunos de los más populares son: Histogram Intersection Kernel (HIK) [2], Radial Basis Function (RBF) y Chi-Squared, el cual es una variante de RBF [17, 14].

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 [16] y HMDB51 [9] [5, 17] y otros con clips pertenecientes a películas [4], deportes [14, 7] y cámaras de vigilancia [5]. 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.

Trabajaremos con un dataset compuesto por numerosos clips de corta duración, los cuales fueron obtenidos a partir de filmaciones de partidos de Rugby. Los videos varían mucho entre sí, ya que se utilizaron tanto filmaciones amateur como transmiciones de televisión. Todos los videos se encuentran en formato .mp4 y tienen entre 24 y 30 FPS. Debido a la variedad en su origen, el tamaño

y el aspecto varía entre ellos.

Éstos clips se clasifican en tres clases diferentes: *line*, *scrum* y *juego*. Nuestro objetivo es poder automatizar su reconocimiento y clasificación.

Para lograrlo nos basaremos en el modelo de *Bag-of-Words*. Estudiaremos distintos detectores y descriptores con el fin de encontrar los más adecuados para el dataset con el que estamos trabajando. También analisaremos y compararemos el costo computacional de ellos.

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. *4 semanas*
- Implementación de los algoritmos necesarios. *3 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. *3 semanas*
- Síntesis de los resultados obtenidos y escritura de la tesina. *4 semanas*

El trabajo se realizará durante aproximadamente 3 meses con una dedicación de 30 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] Annalisa Barla, Rancesca Odone y Alessnndro Verr. "Histogram intersection kernel for image classification". En: *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*. Vol. 3. IEEE. 2003, págs. III-513.

[3] Herbert Bay, Tinne Tuytelaars y Luc Van Gool. "Surf: Speeded up robust features". En: *Computer vision–ECCV 2006*. Springer, 2006, págs. 404-417.

[4] 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.

[5] 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).

[6] Gabriella Csurka y col. "Visual categorization with bags of keypoints". En: *Workshop on statistical learning in computer vision, ECCV*. Vol. 1. 1-22. Prague. 2004, págs. 1-2.

[7] O. Deniz y col. "Fast Violence Detection in Video". En: *The 9th International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, Portugal*. 2014.

[8] Chris Harris y Mike Stephens. "A combined corner and edge detector." En: *Alvey vision conference*. Vol. 15. Citeseer. 1988, pág. 50.

[9] 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.

[10] Stefan Leutenegger, Margarita Chli y Roland Yves Siegwart. "BRISK: Binary robust invariant scalable keypoints". En: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE. 2011, págs. 2548-2555.

[11] Ana Paula Brandão Lopes y col. "Action recognition in videos: from motion capture labs to the web". En: *arXiv preprint arXiv:1006.3506* (2010).

[12] David G Lowe. "Distinctive image features from scale-invariant keypoints". En: *International journal of computer vision* 60.2 (2004), págs. 91-110.

[13] Elmar Mair y col. "Adaptive and generic corner detection based on the accelerated segment test". En: *Computer Vision–ECCV 2010*. Springer, 2010, págs. 183-196.

[14] 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.

[15] Ronald Poppe. "A survey on vision-based human action recognition". En: *Image and vision computing* 28.6 (2010), págs. 976-990.

[16] Christian Schüldt, 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.

[17] Chris Whiten, Robert Laganiere 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.

[18] Jun Yang y col. "Evaluating bag-of-visual-words representations in scene classification". En: *Proceedings of the international workshop on Workshop on multimedia information retrieval*. ACM. 2007, págs. 197-206.