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. Situación del postulante

Actualmente me resta aprobar las materias Sistemas Operativos, Inteligencia Artificial y Compiladores, las que planeo completar durante el primer cuatrimestre del próximo año.

En relación a otras actividades académicas, me estoy desempeñando como docente en las materias Programación I y Programación II.

Proyecto dedicar al menos 25 hs. semanales al desarrollo de esta Tesina.

2. Motivación

El desarrollo de nuevas tecnologías, acompañado de la reducción de costo de las mismas, ha hecho que el acceso a medios para la producción de material audio-visual se haya masificado, cambiando rotundamente la forma en que se trabajaba con videos y generando nuevas herramientas, así como nuevos desafíos y necesidades.

Todo esto ha permitido extender la utilización de cámaras de video a otros ámbitos más allá de la producción de contenidos de entretenimiento, alcanzando distintos objetivos tales como el control de seguridad y el

almacenamiento de recuerdos personales, entre otros. Esto genera grandes volúmenes de datos conservados en formato de video. Si bien los mismos pueden contener muchísima información útil, su análisis y clasificación no es sencilla. En ámbitos profesionales, ya sea para análisis de seguridad o contenidos de entretenimiento, la cantidad de material que se produce ha hecho que sea muy costoso y prácticamente inviable analizarlos manualmente. Por esta razón, se busca automatizar el análisis de los mismos.

A lo largo de los últimos años, se han desarrollado distintas técnicas y enfoques para el análisis de videos. Puesto que éstos no son otra cosa que secuencias de imágenes con sonido incorporado, uno de los enfoques para su análisis se basa en la utilización de los métodos ya desarrollados para imágenes con la incorporación del factor temporal, el cual puede aportar nueva información. Si bien el sonido presente también puede ayudar en el estudio, muchas veces es descartado durante el análisis, basándose éste solamente en los aspectos visuales.

La producción de material audiovisual ha sufrido importantes cambios en las últimas décadas. La reducción de cosotos 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.

Esto trajo aparejado un incremento exponencial en la cantidad de información generada día a día, haciendo cada vez más inviable el análisis 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 diferidas como en tiempo real.

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

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 imágen, diferencias de luz y escala, entre otros. Algunos de ellos fueron ***.

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

EL reconocimiento de acciones humanas basado en visión consiste en el proceso de etiquetado de secuencias con la acción correspondiente. *** En

Un método común y que ha resultado ser eficiente para obtener información a partir de imágenes es la utilización de descriptores. Estos descriptores informan acerca de las características visuales de la imagen, describiendo características elementales como la forma, el color, la textura y la ubicación de elementos dentro de la misma. Para extender esta idea al plano de los videos, lo que se hace es desglosar el video en las imágenes que lo componen (frames), para luego analizar por separado cada una de ellas. Con esta información ya disponible se incorpora el factor temporal, siendo éste el eje en la relación entre los distintos frames, permitiendo, por ejemplo, analizar variaciones de un frame a otro.

3. 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 esos conocimientos, el propósito es aplicar estas herramientas al análisis de imágenes de video correspondientes a la práctica del Rugby.

El objetivo de 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 ver las cintas completas.

4. Fundamentos y estado del conocimiento sobre el tema

La clasificación de imágenes o videos en categorías semánticas en un problema de interés tanto para la actividad 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 entre otras propiedades visuales de las imágenes.

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. ***** 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.

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 y FREAK. En lo referente a videos, suelen utilizarse descriptores que se basan en los de imágenes e incorporan el factor temporal, como MoSIFT y MoFREAK.

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 *** y otros con clips pertenecientes a películas y deportes.*** En ambos casos, los resultados fueron alentadores, mostrando el potencial de esta manera de abordar la detección y clasificación.

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

6. 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] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 510–517. IEEE, 2012.

[2] M. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. research showcase. *Computer Science Department, School of Computer Science, Carnegie Mellon University*, 2009.

[3] O. Deniz, I. Serrano, G. Bueno, and T. Kim. Fast violence detection in video. In *The 9th International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, Portugal*, 2014.

[4] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf. Motion interchange patterns for action recognition in unconstrained videos. In *Computer Vision, ECCV 2012*, pages 256–269. Springer, 2012.

[5] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[6] S. Leutenegger, M. Chli, and R. Y. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2548–2555. IEEE, 2011.

[7] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar. Violence detection in video using computer vision techniques. In *Computer Analysis of Images and Patterns*, pages 332–339. Springer, 2011.

[8] C. Whiten, R. Laganiere, and G.-A. Bilodeau. Efficient action recognition with mofreak. In *Computer and Robot Vision (CRV), 2013 IEEE International Conference on*, pages 319–325. IEEE, 2013.