Reconocimiento de Acciones en Videos de Tenis usando Flujo Óptico y CRF

José Francisco Manera[†‡], Jonathan Vainstein[†‡], Claudio Delrieux[†], y Ana Maguitman[‡]

josemanera@gmail.com, jjvainstein@gmail.com, cad@uns.edu.ar, agm@cs.uns.edu.ar
† Laboratorio de Ciencias de las Imágenes (IIIE - CONICET)
Departamento de Ingeniería Eléctrica y de Computadoras (DIEC)

‡ Grupo de Investigación en Administración de Conocimiento y Recuperación de Información - LIDIA
Departamento de Ciencias e Ingeniería de la Computación (DCIC)

Universidad Nacional del Sur (UNS)
Av. Alem 1253, (B8000CBP), Bahía Blanca, Argentina
Tel: (0291) 459-5135 / Fax: (0291) 459-5136

Resumen

El objetivo del Reconocimiento de Acciones (*Action Recognition*) es el análisis e interpretación automatizados de eventos particulares en secuencias de video. Esta área está siendo ampliamente investigada en diferentes dominios tales como videos de seguridad, interacción humano-computadora, monitoreo de pacientes y recuperación de video, entre otros, dadas las importantes aplicaciones que pueden desarrollarse, y la proliferación de cámaras y videos de seguridad y monitoreo en la actualidad. El objetivo de este proyecto es la identificación automática de acciones en secuencia de videos, utilizando *Conditional Random Fields* (CRFs). Como caso de estudio se utilizan videos de partidos de tenis para la identificación de golpes. Se abordan tres desafíos, el *tracking*, la representación del movimiento del jugador y el reconocimiento de acciones.

Palabras clave: Reconocimiento de Acciones, Tracking, Conditional Random Fields, Flujo Óptico

1. Contexto

El presente proyecto se da en el marco de la colaboración conjunta desarrollada por el Laboratorio de Ciencias de las Imágenes (IIIE-CONICET http://www.imaglabs.org) y el Grupo de Investigación en Administración de Conocimiento y Recuperación de Información (http://ir.cs.uns.edu.ar), pertenecientes a la Universidad Nacional del Sur. Esta línea de investigación se lleva a cabo dentro del ámbito del LCI, y está asociada a los siguientes proyectos de investigación:

- Procesamiento inteligente de imágenes. PGI 24/K047 (SECyT-UNS). Director: Claudio Delrieux.
- PICT Start-Up 2442/2010 (ANPCyT). Director Claudio Delriuex.
- Soporte inteligente para el acceso a información Contextualizada en entornos centralizados y distribuidos. PIP: 11220090100863 (CONICET). Director: Ana G. Maguitman.

2. Introducción

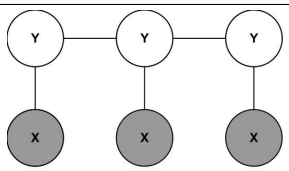
En los últimos años se ha producido un gran crecimiento en la disponibilidad de información multimedia, lo que ha motivado en gran medida el desarrollo del reconocimiento de acciones en video dentro del área de visión por computadora. En función del gran volumen de información se hace necesario desarrollar sistemas automáticos o semi-automáticos que permitan el etiquetado de acciones en video

con diferentes aplicaciones, como por ejemplo la detección de acciones sospechosas en cámaras de vigilancia [17], detección de incidentes de tránsito [9] y búsqueda de acciones en videos deportivos [21].

Este trabajo aborda la temática del análisis de secuencias de video de partidos de tenis. Este tipo de análisis ha sido ya estudiado por diversos autores. En [11] se desarrolla un sistema de anotación automático de acciones en partidos de tenis basado en transiciones de siluetas. En [5] se utiliza un descriptor de movimiento basado en el flujo óptico espacio-temporal junto con un clasificador de vecino más cercano para llevar a cabo la categorización de acciones. En [20] se utiliza el flujo óptico como descriptor de los movimientos de los jugadores y emplea Support Vector Machines para entrenar al clasificador en donde la información del flujo óptico es utilizada como dato de entrada del mismo.

Los CRFs son un modelo probabilístico discriminativo para el etiquetado de secuencias. Este modelo condiciona las probabilidades a la secuencia de observaciones, lo cual evita computar las probabildiades para cada posible observación de la secuencia. En lugar de depender de las probabilidades conjuntas $P(X, Y)$, los CRFs específican la probabilidad de secuencias de etiquetas posibles dada la observación $P(Y|X)$. Un modelo gráfico típico de CRF es ilustrado en la Fig. 1, donde X e Y refieren a las observaciones y las secuencias de etiquetas respectivamente.

Figura 1 Modelo gráfico de un CRF. Los nodos rotulados con X corresponden a observaciones y los rotulados con Y a etiquetas.



Los CRFs han sido aplicados a una variedad de dominios, tales como procesamiento de lenguaje natural [15, 16, 10, 4], bioinformática [14, 18] y visión por computadora. En esta última área algunos autores han utilizado CRFs para el etiquetado de imágenes [7] y para el reconocimiento de objetos. En [13] se utilizan CRFs para determinar partes características de un objeto. En particular, para el caso de reconocimientos de acciones en secuencias de video hay diversos trabajos sobre detección de acciones en video deportivos [8, 20].

El objetivo de este trabajo consiste en diseñar e implementar un sistema de identificación de acciones, que utilice procesamiento de video y reconocimiento basado en CRF. Como caso de entrenamiento, se propone entrenar al sistema para clasificar clips de videos de tenis, según el tipo de golpe llevado a cabo en en cada video.

3. Líneas de investigación y desarrollo

La línea de investigación se centra en el diseño e implementación de un sistema de reconocimiento de acciones en videos de tenis, más concretamente se avanzará en el estudio de técnicas de procesamiento de video y de CRFs con el objetivo de desarrollar y mejorar nuevas técnicas en ambas áreas. Existen varios desafios a ser aboradados en el desarrollo de esta línea de investigación: tracking, extracción de *features*, clasificación y reconocimiento de patrones.

Se han abordado dos desafíos complementarios: la representación del movimiento del jugador y el reconocimiento de acciones. Para el primero se propone la utilización del flujo óptico [6] para modelar los patrones de movimiento del jugador en el campo de juego. Para el reconocimiento de acciones se propone el uso de CRFs utilizando la información obtenida a partir del flujo óptico de los sucesivos *frames* como atributos de entrada del clasificador.

4. Propuesta y Metodología

Como en todo proceso de clasificación supervisada, se pueden describir claramente dos grandes etapas. La primera consiste en la construcción de un modelo de clasificación a partir del entrenamiento realizado con un conjunto de muestras clasificadas de forma supervisada, y una segunda etapa en la cual se lleva a cabo la clasificación de una nueva instancia, empleando para ello el modelo previamente obtenido.

4.1. Entrenamiento

La fase de entrenamiento se compone de un pipeline cuyas etapas son: tracking, extracción de features y construcción del clasificador. Los videos utili-

zados como entrada para esta etapa corresponden a capturas de video de dominio público realizadas con una cámara oblicua. Estos videos fueron previamente clasificados en forma supervisada en dos clases: *drives* de derecha y *drives* de izquierda. Para esta etapa se utiliza la biblioteca OpenCV.

La primera etapa tiene como objetivo llevar a cabo el tracking del jugador que se encuentra de espaldas a la cámara, seleccionándolo de manera supervisada en el frame 0 del video. A partir de la selección de esta región de interés se genera un modelo del jugador que se compone de dos histogramas. El primero consiste de los valores de luminancia correspondientes a los pixeles de la ropa del jugador. Este histograma se calcula a partir de la imagen obtenida al aplicar una máscara que elimina los pixels que no corresponden a la ropa del jugador. El segundo histograma se obtiene a partir de la imagen que resulta de aplicarle una máscara que elimina los pixels que no corresponden a la piel del jugador. Para este histograma se utiliza el canal Hue del espacio cromático HSV [19]. En el frame 1 se toma la misma posición de la región de interés correspondiente al frame 0, y para cada uno de los histogramas antes descriptos se aplica el siguiente algoritmo [1]:

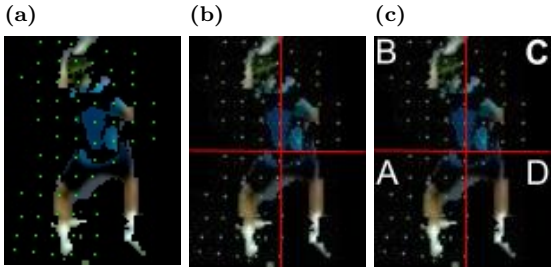
1. Para cada pixel de la imagen se toma su valor y se busca el recipiente (*bin*) correspondiente en el histograma.
2. Se toma el valor asociado al bin seleccionado.
3. Se almacena el valor del bin en una nueva imagen.

Los valores almacenados en cada imagen de salida representan la probabilidad de que un pixel en la imagen de entrada pertenezca a la zona de interés representada por el histograma usado (en este caso piel y color de la ropa respectivamente).

Luego se suman las imágenes obtenidas y su resultado -junto a la región de interés del frame anterior- son utilizadas como entradas para el algoritmo de Meanshift [3]. Como resultado de dicho algoritmo se obtiene una nueva región de interés que corresponde a la posición del jugador en el frame actual. Este proceso se repite para todos los frames del video.

Luego, para describir de manera robusta y discriminativa cada una de las clases de golpes se utiliza el flujo óptico (Fig. 2a), el cual es calculado utilizando el algoritmo Gunnar Farnebäck [6]. La matriz de desplazamientos obtenida se divide en cuatro regiones (Fig. 2b) para luego generar una etiqueta para cada parte que representa la variación del flujo óptico en cada región (Fig. 2c). Este último paso tiene como objetivo discretizar los datos del flujo óptico a fin de flexibilizar la coincidencia de atributos en las etapas de construcción y validación del clasificador.

Figura 2 Flujo Óptico



Finalmente, para llevar a cabo la construcción del clasificador que nos permita realizar la tarea de reconocer los dos tipos de golpes propuestos en este trabajo se utiliza Conditional Random Fields. Este clasificador es entrenado utilizando el conjunto de entrenamiento obtenido en los pasos anteriores. Para las etapas de entrenamiento y clasificación se utiliza la herramienta CRFSuite [12]. La entrada a dicha herramienta está compuesta por un archivo de texto plano, en el cual cada clip de video usado para la etapa de entrenamiento es representado por tantas líneas como frames tenga el mismo. A su vez, cada línea está compuesta por 5 columnas. La primera corresponde a la etiqueta del frame (es decir, a qué tipo de golpe corresponde e frame), y las restantes 4 columnas corresponden a la representación del flujo óptico dado en este trabajo.

4.2. Clasificación y Validación

Esta etapa consiste en la clasificación de un video de entrada en una de las clases definidas en el proceso de entrenamiento. El video de entrada es procesado empleando el tracking y la extracción de features desarrollados anteriormente. Estos característicos son formateados de forma adecuada para ser utilizados como entrada al clasificador, resultando como salida la clase a la cual pertenece el video de entrada.

Para determinar la efectividad de los métodos propuestos se utilizan las medidas de precisión, recall y F1 [2].

5. Resultados y Objetivos

5.1. Resultados Preliminares

Se llevó a cabo una primera prueba que consistió en entrenar un clasificador utilizando un conjunto de 20 clips de video (10 clips representando drives de izquierda, y 10 clips de drives de derecha). Luego se validó dicho clasificador utilizando 22 clips de video (11 clips de cada tipo de golpe). La matriz de confusión se encuentra en la Tabla 1. A partir de los datos de la matriz de confusión se pueden inferir los valores de precisión: 86.36 %, recall: 85.57 % y F1: 85.3 %.

		Clase predicha DI	DD
Clase real	DI	0.73	0.27
	DD	0	1

Tabla 1: Matriz de confusión. DI: Drive izquierda. DD: Drive derecha.

5.2. Trabajo Futuro

Como trabajo futuro se espera avanzar y mejorar en los siguientes aspectos: tracking, extracción de features para el modelado de las acciones, representación de los datos de entrada para CRFs, implementación propia de CRFs y reconocer una mayor cantidad de golpes.

6. Formación de Recursos Humanos

El equipo de trabajo de esta línea de investigación se encuentra integrado por dos becarios de posgrado que cuentan con una beca interna del Conicet, y los respectivos directores. Por otra parte se cuenta con la colaboración de otros becarios de posgrado del LCI y una vinculación con un grupo de trabajo sobre minería de datos compuesto por doctorandos e investigadores formados.

Como parte de las actividades asociadas al proyecto se realizan cursos de grado y postgrado en Procesamiento de Imágenes, Minería de Datos y Aprendizaje Supervisado.

Referencias

[1] Opencv: Backprojection.

[2] Presición, recall y f1, 2013 (accessed March 9, 2013).

[3] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Kernel-based object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(5):564–575, may 2003.

[4] Aron Culotta, Ron Bekkerman, and Andrew Mccallum. Extracting social networks and contact information from email and the web. In *In Proceedings of CEAS-1*, 2004.

[5] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV'03, pages 726–, Washington, DC, USA, 2003. IEEE Computer Society.

[6] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Josef Bigun and Tomas Gustavsson, editors, *Image Analysis*, volume 2749 of *Lecture Notes in Computer Science*, pages 363–370. Springer Berlin Heidelberg, 2003.

[7] Xuming He, Richard S. Zemel, and Miguel Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition*, CVPR'04, pages 695–703, Washington, DC, USA, 2004. IEEE Computer Society.

[8] Nisha Jain, Santanu Chaudhury, Sumantra Dutta Roy, Prasenjit Mukherjee, Krishanu Seal, and Kumar Talluri. A novel learning-based framework for detecting interesting events in soccer videos. In *Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, ICVGIP'08, pages 119–125, Washington, DC, USA, 2008. IEEE Computer Society.

[9] Shunsuke Kamijo, Yasuyuki Matsushita, Katsushi Ikeuchi, and Masao Sakauchi. Incident detection at intersections utilizing hidden markov model, 1999.

[10] Andrew McCallum. Efficiently inducing features of conditional random fields. In *Proceedings of the Nineteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 403–410, San Francisco, CA, 2003. Morgan Kaufmann.

[11] Hisashi Miyamori and Shu-ichi Iisaku. Video annotation for content-based retrieval using human behavior analysis and domain knowledge. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, FG00, pages 320–, Washington, DC, USA, 2000. IEEE Computer Society.

[12] Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007.

[13] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Conditional random fields for object recognition. In *In NIPS*, pages 1097–1104. MIT Press, 2004.

[14] Kengo Sato and Yasubumi Sakakibara. Rna secondary structural alignment with conditional random fields. In *ECCB/JBI*, page 242, 2005.

[15] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 134–141, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[16] Charles Sutton, Andrew McCallum, and Khashayar Rohanimanesh. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8:693–723, March 2007.

[17] M. Takahashi, M. Naemura, M. Fujii, and S. Satoh. Human action recognition in crowded surveillance video sequences by using features taken from key-point trajectories. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 9–16, June.

[18] Peter Weigele Yan Liu, Jaime Carbonell and Vanathi Gopalakr-ishnan. Protein fold recognition using segmentation conditionalrandom elds (scrfs). *Journal of Computational Biology, 13 (2) 394-406*, 2006.

[19] Benjamin D. Zarit, Boaz J. Super, and Francis K. H. Quek. Comparison of five color models in skin pixel classification. In *In ICCV 99 Int l Workshop on*, pages 58–63, 1999.

[20] Guangyu Zhu, Changsheng Xu, Wen Gao, and Qingming Huang. Action recognition in broadcast tennis video using optical flow and support vector machine. In *Proceedings of the 2006 international conference on Computer Vision in Human-Computer Interaction*, ECCV06, pages 89–98, Berlin, Heidelberg, 2006. Springer-Verlag.

[21] Guangyu Zhu, Changsheng Xu, and Qingming Huang. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. In *in Proc. ACM Multimedia, 2006*, pages 431–440, 2006.