UABDIVULGA
BARCELONA RECERCA I INNOVACIÓ

01/2011

Comprensión y descripción automática de vídeo contenidos

Los últimos avances en comunicación digital han favorecido una presencia masiva de tecnologías de vídeo en entornos multimedia y de videovigilancia, donde cada vez más se exigen métodos de análisis automático de contenidos. Esta tesis propone una perspectiva ontológica para automatizar el reconocimiento de acontecimientos de interés a secuencias de vídeo y su descripción lingüística. Se plantean tres retos básicos: (i) localizar regiones de interés en las escenas, (ii) razonar sobre la información visual obtenida, (iii) implementar interfaces de comunicación avanzada con el usuario.

La información digital cada vez está más ligada a nuestras rutinas diarias. En potenciar este tipo de contenidos, el vídeo ha convertido en una herramienta privilegiada para la comunicación, como lo demuestra el crecimiento exponencial de multimedia social (YouTube, Dailymotion, Metacafe), o la incrementada presencia de sistemas de videovigilancia en todo el mundo. Esta alza espectacular del vídeo deriva nuevas necesidades tecnológicas: pensemos, por ejemplo, que el volumen diario de vídeos en portales sociales hace imposible que sus gestores puedan etiquetarlos uno a uno de forma cuidadosa, así mismo, nuestras limitaciones de atención innatas impiden a los operarios de videovigilancia poder examinar los numerosas grabaciones en tiempo

real. Aquí surge la idea de desarrollar sistemas informáticos que realicen estas tareas de forma automática, mediante la visión por computador.

En esta tesis se persigue reconocer y describir automáticamente acontecimientos significativos en secuencias de vídeo: peatones o vehículos en entornos de tráfico urbano, acciones señaladas en eventos deportivos, comportamientos de usuarios de transporte público o situaciones de personas con necesidades de atención especial, por ejemplo. Entre las tareas se incluye el diseño de interfaces de comunicación lingüística, para transmitir las interpretaciones del sistema a usuarios finales de forma natural y multilingüe, y permitirles así buscar o manipular fácilmente los contenidos de las secuencias. Las contribuciones se organizan en tres bloques principales:

1. Reconocer automáticamente regiones de interés funcional de una escena, a partir del movimiento observado. Aprender en qué zonas las personas y los vehículos suelen entrar o salir, cruzar o interactuar con objetos es fundamental para identificar comportamientos complejos, como riesgos de atropello, caídas de personas mayores o usos abusivos de instalaciones públicas. Nuestro método actualiza modelos probabilísticos locales caracterizando prototipo de las regiones de interés a partir de las trayectorias capturadas, y obtiene regiones coherentes mediante interpolación geodésica y campos aleatorios de Markov (MRF).

2. Construir modelos semánticos para interpretar situaciones y comportamientos complejos a partir de información visual. Los sistemas de visión capturan datos geométricos a lo largo del tiempo (posiciones, orientaciones, velocidades), que hay que calificar conjuntamente para deducir qué acontecimientos suceden en la escena. Para ello, utilizamos mecanismos de lógica difusa y árboles de grafos de situación (SGT) para crear modelos de comportamiento humano, y ontologías para representar el conocimiento semántico obtenido.

3. Diseñar interfaces avanzadas de comunicación con usuarios finales. Describir detalladamente o bien resumir los acontecimientos más importantes, en 6 lenguas diferentes; generar animaciones virtuales de acciones observadas o simulaciones de situaciones posibles, o hacer que el ordenador responda coherentemente a cualquier pregunta que se tenga sobre el contenido de los vídeos. Todas estas aplicaciones, basadas en los resultados anteriores, se han hecho posibles mediante ingeniería ontológica, gráficos por computador y técnicas de lingüística computacional, como representación del discurso (DRT) o parsing ontológico.

El sistema propuesto se ha evaluado experimentalmente para cada uno de los procesos implicados, comparando los resultados con otras técnicas del estado del arte y con resultados aportados por voluntarios. Se han utilizado bases de datos públicas de dominios urbanos, interiores y deportivos, y cámaras web públicas. El sistema ha contribuido a la implementación de un sistema prototipo que se actualmente se encuentra en pleno funcionamiento en el Centro de Visión por Computador.

Esta investigación se ha llevado a cabo por los investigadores Carlos Fernández, Pablo Baiget, Jordi González y Xavier Roca, del grupo de evaluación de secuencias de imágenes (ISE Lab) del Centro de Visión por Computador, y ha sido parcialmente financiada por los proyectos del Fondo Europeo: IST-027110 (HERMES) y IST-045547 (VIDI-video) y del Ministerio de Educación y Ciencia: TIN2006-14606 y CONSOLIDER-INGENIO 2010: MIPRCV (CSD2007-00018).

Carles Fernández
perno@cvc.uab.es

Referencias

"Understanding Image Sequences: the Role of Ontologies in Cognitive Vision". Tesis doctoral defendida por Carlos Fernández el 2 de julio de 2010 a las 12h, en la Sala de Actos del Centre de Visió per Computador. Director: Jordi González i Sabaté.

View low-bandwidth version