Comprensió i descripció automàtica de vídeo continguts

01/2011 - **Telecomunicacions, Electrònica i Informàtica.** **Els darrers avenços en comunicació digital han afavorit una presència massiva de tecnologies de vídeo en entorns multimèdia i de videovigilància, on cada cop més s'exigeixen mètodes d'anàlisi automàtica de continguts. Aquesta tesi proposa una perspectiva ontològica per automatitzar el reconeixement d'esdeveniments d'interès a seqüències de vídeo i la seva descripció lingüística. Es plantegen tres reptes bàsics: (i) localitzar regions d'interès a les escenes; (ii) raonar sobre la informació visual obtinguda; (iii) implementar interfícies de comunicació avançada amb l'usuari.**

El sistema analitza informació gràfica i permet la comunicació amb l'usuari sobre el contingut visual.

La informació digital cada cop està més lligada a les nostres rutines diàries. En potenciar aquest tipus de continguts, el vídeo ha esdevingut una eina privilegiada per a la comunicació, com ho demostra el creixement exponencial de multimèdia social (YouTube, Dailymotion, Metacafe), o la incrementada presència de sistemes de videovigilància arreu del món. Aquesta alça espectacular del vídeo deriva noves necessitats tecnològiques: pensem, per exemple, que el volum diari de vídeos en portals socials fa impossible que els seus gestors puguin etiquetar-los un a un de forma acurada; així mateix, les nostres limitacions d'atenció innates impedeixen als operaris de videovigilància poder examinar els nombrosos enregistraments en temps real. Aquí sorgeix la idea de desenvolupar sistemes informàtics que realitzin aquestes tasques de forma automàtica, per mitjà de la visió per computador.
En aquesta tesi es persegueix reconèixer i descriure automàticament esdeveniments significatius en seqüències de vídeo: vianants o vehicles en entorns de trànsit urbà, accions assenyalades en esdeveniments esportius, comportaments d'usuaris de transport públic o situacions de persones amb necessitats d'atenció especial, per exemple. Entre les tasques s'inclou el disseny d'interfícies de comunicació lingüística, per a transmetre les interpretacions del sistema a usuaris finals de forma natural i multilingüe, i permetre'ls així cercar o manipular fàcilment els continguts de les seqüències. Les contribucions s'organitzen en tres blocs principals:
1. Reconèixer automàticament regions d'interès funcional d'una escena, a partir del moviment observat. Aprendre en quines zones les persones i els vehicles acostumen a entrar o sortir, creuar o interactuar amb objectes és fonamental per a identificar comportaments complexes, com riscos d'atropellament, caigudes de gent gran o usos abusius d'instal·lacions públiques. El nostre mètode actualitza models probabilístics locals caracteritzant prototipus de les regions d'interès a partir de les trajectòries capturades, i obté regions coherents mitjançant interpol·lació geodèsica i camps aleatoris de Markov (MRF).

2. Construir models semàntics per interpretar-ne situacions i comportaments complexes a partir d'informació visual. Els sistemes de visió capturen dades geomètriques al llarg del temps (posicions, orientacions, velocitats), que cal qualificar conjuntament per a deduir quins esdeveniments succeeixen a l'escena. Per a fer-ho, utilitzem mecanismes de lògica difusa i arbres de grafs de situació (SGT) per a crear models de comportament humà, i ontologies per a representar-ne el coneixement semàntic obtingut.

3. Dissenyar interfícies avançades de comunicació amb usuaris finals. Descriure detalladament o bé resumir els esdeveniments més importants, en 6 llengües diferents; generar animacions virtuals d'accions observades o simulacions de situacions possibles; o fer que l'ordinador respongui coherentment a qualsevol pregunta que es tingui sobre el contingut dels vídeos. Totes aquestes aplicacions, basades en els resultats anteriors, s'han fet possibles mitjançant enginyeria ontològica, gràfics per computador i tècniques de lingüística computacional, com ara representació del discurs (DRT) o pàrsing ontològic.
El sistema proposat s'ha avaluat experimentalment per a cadascun dels processos implicats, comparant els resultats amb altres tècniques de l'estat de l'art i amb resultats aportats per voluntaris. S'han emprat bases de dades públiques de dominis

urbans, interiors i esportius, i càmeres web públiques. El sistema ha contribuït a la implementació d'un sistema prototipus que es actualment es troba en ple funcionament al Centre de Visió per Computador.
Aquesta recerca s'ha dut a terme pels investigadors Carles Fernández, Pau Baiget, Jordi Gonzàlez i Xavier Roca, del grup d'avaluació de seqüències d'imatges (ISE Lab) del Centre de Visió per Computador, i ha estat parcialment finançada pels projectes del Fons Europeu: IST-027110 (HERMES) i IST-045547 (VIDI-video) i del Ministeri d'Educació i Ciència: TIN2006-14606 i CONSOLIDER-INGENIO 2010: MIPRCV (CSD2007-00018).

Carles Fernández

Centre de Visió per Computador

"Understanding Image Sequences: the Role of Ontologies in Cognitive Vision". Tesi doctoral defensada per Carles Fernández el 2 de juliol de 2010 a les 12h, a la Sala d'Actes del Centre de Visió per Computador. Director: Jordi Gonzàlez i Sabaté.