Alessia Saggese(*)

Semantica e analisi video: tecnologia del futuro, oggi

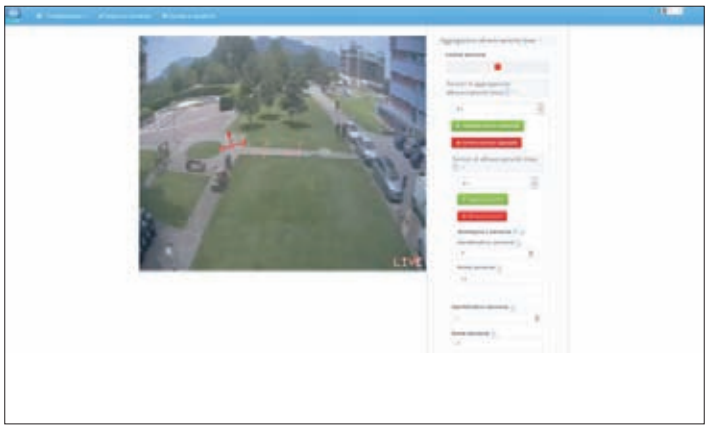
Nel 2014 si dice che il numero di telecamere di sicurezza installate in giro per il mondo si aggirasse intorno ai 245 milioni di unità (praticamente una telecamera per meno di 30 persone), con 413 petabytes di dati raccolti ogni giorno. Questa era la più grande produzione di big data disponibile nel mondo. Ad oggi, dopo pochi anni, la situazione non è cambiata; il numero di camere e quindi la quantità di dati da gestire non ha fatto altro che aumentare. E contestualmente è aumentata la necessità di analizzare in modo automatico questa grande quantità di dati raccolti. Come? Attraverso l'impiego di avanzati algoritmi di analisi video.

(*) Account Manager @ A.I. Tech **www.aitech.vision**

Ma come funzionano tali algoritmi? Questo è un argomento già trattato nelle precedenti lezioni di questo "corso di formazione" sulle pagine di a&s Italy, ma rivediamolo brevemente: le due fasi su cui si basano la maggior parte degli algoritmi di analisi video (pensati per acquisire il flusso video da telecamere fisse) sono la fase di "detection", in cui - a partire dall'immagine (e quindi dall'insieme di pixel che compongono l'immagine) - si estraggono gli oggetti in movimento in ciascun fotogramma; e la fase di "tracking", in cui, a partire dagli oggetti in movimento, se ne estraggono le traiettorie. Insomma, un approccio bottom-up (dal basso verso l'alto): dai pixel dell'immagine fino ad arrivare alle traiettorie degli oggetti in movimento. Tutto ciò senza alcuna esplicita conoscenza circa il contesto. Questo, inevitabilmente, crea errori quali "falsi allarmi", ossia oggetti spuri come ad esempio ombre o alberi in movimento erroneamente considerati oggetti di interesse; oppure "miss", ossia oggetti in movimento che non sono invece considerati in alcun modo dall'algoritmo di analisi video. Tutte problematiche, queste, che potrebbero essere invece risolte utilizzando un approccio knowledge driven, guidato ossia dalla conoscenza piuttosto che dai dati.

QUALE TIPO DI CONOSCENZA?

Ci sono tre sorgenti di conoscenza che possono essere utilizzate: (1) la conoscenza relativa alla *scena osservata*: quali sono i punti di ingresso (e di uscita) delle persone che entrano nella scena inquadrata dalla telecamera? Ci sono oggetti all'interno della scena che potrebbero causare dei falsi allarmi (ad esempio un albero in movimento)? Ci sono degli oggetti che potrebbero nascondere il passaggio delle persone, in modo parziale o totale (ad esempio un palo piuttosto che un muretto dietro ai quali è consentito il passaggio delle persone)?



Un esempio di analisi video che sfrutta la conoscenza

(2) La seconda sorgente di conoscenza che può essere utilizzata è quella legata alla *tipologia di utenti* che popolano la scena. Ci aspettiamo che vi siano veicoli e persone? Oppure solo persone e animali? (3) Infine, vi è sicuramente la conoscenza relativa alle *possibili interazioni tra gli oggetti* che si muovono all'interno della scena.

E' evidente a questo punto come la conoscenza, nelle direzioni appena menzionate, può aiutare a risolvere i tipici problemi che i tradizionali algoritmi di video analisi "bottom-up" devono affrontare.

IN CHE DIREZIONE SI STA MUOVENDO LA LETTERATURA SCIENTIFICA?

Una risposta nella direzione "della conoscenza" viene dalle cosiddette tecnologie "Semantiche", che stanno appassionando negli ultimi anni la comunità scientifica e che hanno aperto la strada al cosiddetto "Knowledge based Computer Vision". Il paradigma cambia: ci si muove infatti da un approccio bottom-up ad un approccio top-down (dall'alto verso il basso), in cui la conoscenza gioca un ruolo fondamentale al fine migliorare le decisio-

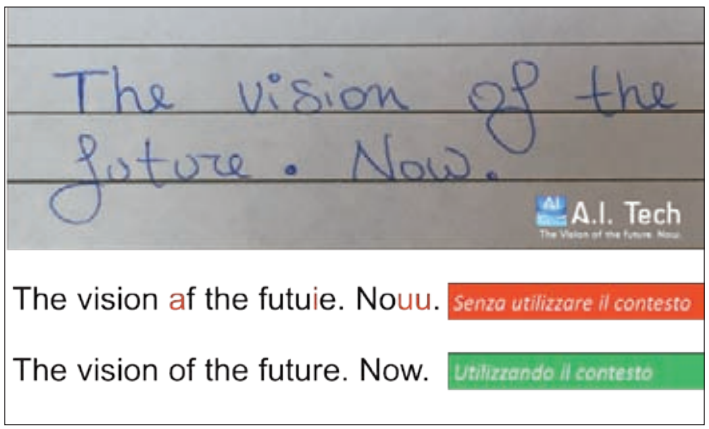


Figura 1 - L'utilizzo della semantica nel riconoscimento dei caratteri

ni prese dai tradizionali approcci basati (bottom up) su pixel. Per chiarire meglio il concetto, facciamo un esempio in un dominio applicativo leggermente differente, ma su cui risulta più semplice fare questo tipo di ragionamenti: quello del riconoscimento dei caratteri. Per anni (o forse per decenni) gli scienziati si sono concentrati sul riconoscimento di un singolo carattere e sul miglioramento degli approcci in tale direzione: trovare la migliore rappresentazione possibile di quel carattere, piuttosto che il miglior paradigma di classificazione per risolvere quello specifico problema. Ma sapete qual è stata la vera svolta? Il momento in cui gli scienziati hanno deciso di utilizzare la conoscenza sul *contesto*.

Piuttosto che riconoscere il singolo carattere, ci si è lasciati infatti aiutare dalla parola a cui tale carattere apparteneva (e cioè dal contesto). Un esempio è riportato in figura: la frase "The vision of the future. Now", scritta a mano, viene riconosciuta come "The vision af the futuie. Nouu": il carattere "O" viene erroneamente riconosciuto come "A" ecc. Ma tali parole esistono nel vocabolario che stiano considerando (ossia nel nostro contesto)? No! Ed è grazie a questa ulteriore informazione (il contesto, e quindi la conoscenza) che riusciamo a riconoscere correttamente la frase "The vision of the future. Now" (**figura 1**). Insomma, approcci bottom-up che si uniscono a quelli top-down. E' stato proprio questo il momento storico in cui i sistemi per il riconoscimento di caratteri hanno subito un notevole incremento delle prestazioni. Ed è proprio questa la molla che sta spingendo la comunità scientifica, così come le aziende più innovative che operano nel settore, ad esplorare il contesto anche nel mondo della visione artificiale. Un esempio? I sistemi anTintrusione di A.I. Tech, AI-Intrusion oppure AI-IntrusionPRO (www.aitech.vision/ai-intrusion), che sfruttano la conoscenza sull'ambiente e sul contesto per migliorare le prestazioni dei tradizionali sistemi di analisi video per l'antintrusione. Provare per credere.