Università degli Studi di Salerno

Dottorato di Ricerca in Ingegneria dell'Informazione
XII Ciclo – Nuova Serie

université de Caen Basse-Normandie

Doctorat de l'université de Caen Basse-Normandie

UNIVERSITÀ ITALO FRANCESE

PHD THESIS

Detecting and indexing moving objects for Behavior Analysis by Video and Audio Interpretation

CANDIDATE: **ALESSIA SAGGESE**

SUPERVISORS: **PROF. MARIO VENTO**
PROF. LUC BRUN

COORDINATORS: **PROF. ANGELO MARCELLI**
PROF. FRÉDÉRIC JURIE

Academic Year 2012 – 2013

Negli ultimi decenni abbiamo assistito ad una crescente esigenza di sicurezza in molti ambienti pubblici: secondo uno studio condotto dall'Osservatorio Europeo sulla Sicurezza, la criminalità è in cima alle preoccupazioni della popolazione europea, che richiede insistentemente una maggiore protezione da parte delle forze dell'ordine.
Tale considerazione ha portato alla proliferazione di dispositivi quali telecamere e microfoni, che rappresentano la soluzione ideale visti i costi di manutenzione relativamente bassi, la possibilità di installarli virtualmente ovunque ed infine la capacità di analizzare eventi complessi. La principale limitazione di tali sistemi tradizionali sta nel cosiddetto "problema di sovraccarico psicologico" degli operatori responsabili per la sicurezza, che causa una progressiva decrescita nelle loro capacità di analizzare flussi di dati provenienti da sorgenti multimediali multiple; infatti, come confermato da uno studio condotto dal Security Solution Magazine, "dopo 12 minuti di monitoraggio continuo un operatore perde il 45% delle attività che occorrono nella scena. Dopo 22 minuti di video, più del 95% viene ignorato".
Per i motivi appena menzionati, un sistema di sorveglianza intelligente, capace di attribuire ai segnali audio e video un'interpretazione semantica, si rivelerebbe estremamente utile poiché significherebbe ridurre la distanza che esiste tra la rappresentazione di basso livello in termini di pixel e quella di alto livello di cui invece un operatore vorrebbe disporre, contenente cioè la descrizione della scena in linguaggio naturale.
D'altra parte, un sistema del genere capace di interpretare automaticamente gli eventi che occorrono in una scena potrebbe essere notevolmente utile anche in altri ambiti, e in particolare in applicazioni con fini commerciali. Soprattutto negli ultimi anni, numerosi sono stati i sistemi di business intelligence realizzati al fine di supportare i processi di decision maker e di dare ai dipendenti delle imprese, nonché ai partner e ai fornitori, un facile accesso alle informazioni di cui hanno bisogno per fare in modo efficace il proprio lavoro.
Questa tesi si pone come obiettivo quello di affrontare le questioni sopra menzionate, tanto affascinanti quanto impegnative poiché richiedono diverse conoscenze e competenze, dalla computer vision e la pattern recognition ai database e il machine learning.
Numerose sono le problematiche che rendono tale compito estremamente complesso, tanto che una soluzione definitiva non è ancora stata trovata: innanzitutto, il sistema deve rilevare gli oggetti di interesse che popolano la scena (che siano essi persone, veicoli o animali). È importante sottolineare che tale compito è decisamente complesso: si pensi, ad esempio, ad un albero le cui foglie sono mosse dal vento le quali, seppur in movimento, non possono considerarsi oggetti di interesse.
Tra le principali difficoltà si annovera il problema dell'occlusione, dovuto all'appiattimento prospettico introdotto dall'utilizzo di una singola telecamera: nei sistemi prima menzionati è infatti richiesto che l'oggetto sia inseguito (e quindi che la sua traiettoria sia estratta) anche se questo risulti parzialmente o completamente nascosto da un'altra persona in movimento nella scena o da un oggetto statico: si pensi, ad esempio, ad un muro che lascia solo intravedere una piccola parte di una persona (la testa o le gambe), il che rende questo compito estremamente complesso.
Un'altra operazione decisamente ardua riguarda la gestione delle traiettorie acquisite: in ambienti reali, infatti, è richiesto che bilioni di traiettorie, estratte eventualmente da camere differenti, siano memorizzate e che su tali ampi database gli operatori umani possano sottomettere query complesse, che coinvolgano informazioni tanto spazio-temporali quanto semantiche.
Inoltre, ciascuna traiettoria dovrà essere opportunamente analizzata al fine di interpretare il comportamento dell'oggetto che l'ha generata ed identificare eventuali comportamenti anomali.
Per quanto riguarda la progettazione e l'implementazione di un sistema per il riconoscimento di eventi audio, uno dei principali problemi aperti riguarda invece la caratterizzazione dei differenti eventi di interesse, che deve naturalmente essere effettuata a differenti scale temporali: si confronti, ad esempio, un suono impulsivo come uno sparo, con uno prolungato quale un urlo, che può avere una durata di diversi secondi.
Inoltre, in applicazioni reali può spesso accadere che i suoni di interesse siano sovrapposti da un significativo livello di background, il che rende difficile separare il rumore che deve essere ignorato rispetto ai suoni che devono essere riconosciuti.
A rendere ulteriormente più complicato questo task, bisogna considerare che tutte queste operazioni devono essere eseguite in tempo reale: ciò significa che mentre i segnali audio e video sono acquisiti (di solito con un frame rate di 25 fotogrammi al secondo un bit rate di 128 kbit al secondo) il sistema deve eseguire tutte le operazioni richieste: il rilevamento degli oggetti e dei loro movimenti, la memorizzazione e l'analisi delle loro traiettorie, l'identificazione di eventi audio e infine l'interazione con l'utente. Quest'ultimo aspetto in particolare conferisce a questa tesi un alto valore tecnologico, oltre che un sostanziale avanzamento scientifico rispetto allo stato dell'arte.
Nel seguito, maggiori dettagli relativamente al sistema proposto saranno forniti: il sistema innanzitutto

analizza il video e ne estrae le traiettorie degli oggetti che popolano la scena (tracking): è importante sottolineare che la traiettoria è una feature estremamente discriminante, in quanto il movimento di oggetti in una scena non è casuale, bensì ha una struttura di base che può essere sfruttata per costruire dei modelli di comportamento. Le principali novità dell'algoritmo di tracking proposto riguardano i seguenti aspetti: in primo luogo, l'intera storia di ciascun oggetto è analizzata utilizzando un automa a stati finiti; in secondo luogo, l'aggiornamento delle informazioni relative a ciascun oggetto viene gestito grazie ad un innovativo approccio graph-based. Infine, le occlusioni sono opportunamente gestite inseguendo in modo diverso singoli oggetti e gruppi di oggetti. L'algoritmo di tracking proposto è stato valutato nell'ambito di una competizione internazionale (PETS 2013) che ha vantato la partecipazione di un numero elevato di gruppi di ricerca (più di trenta) e si è posizionato nelle prime posizioni in tutti gli indici considerati.
Una volta estratte, questa grande quantità di traiettorie deve essere indicizzata e propriamente memorizzata al fine di migliorare le prestazioni del sistema durante la successiva fase di retrieval: la novità principale di questo modulo riguarda l'evoluzione di soluzioni off-the-shelf, ed in particolare di PostGis (l'estensione spaziale del tradizionale database relazionale PostGres) al fine di analizzare traiettorie, la cui gestione si rileva decisamente complessa vista la loro natura spazio-temporale.
In generale, uno dei principali vantaggi del metodo proposto consiste nel fatto che l'operatore umano può interagire con il sistema in diversi modi: in primo luogo, l'utente viene informato non appena un comportamento anomalo è identificato; nel contesto della videosorveglianza intelligente, un comportamento anomalo può essere associato a comportamenti pericolosi, poiché si riferisce a quegli oggetti che disegnano nella scena una traiettoria mai vista dal sistema. È evidente che il sistema deve essere abbastanza robusto per fronteggiare quegli errori che di solito si verificano durante la fase di tracking (ad esempio le traiettorie spezzate); inoltre, è richiesto un elevato livello di generalizzazione al fine di evitare di classificare erroneamente come anormali quelle traiettorie normali che ricorrono solo raramente. Al fine di far fronte alle questioni di cui sopra, la somiglianza tra traiettorie viene valutata attraverso una metrica basata su string kernel, definita appositamente per risolvere questo problema.
Laddove le informazioni estratte dal video non siano sufficienti o non sufficientemente affidabili, il sistema proposto è arricchito da un modulo incaricato di riconoscere eventi audio, come spari, urla o vetri rotti. È opportuno sottolineare che l'integrazione tra le informazioni audio e video attribuisce un significativo valore aggiunto per il sistema proposto, essendo un aspetto completamente nuovo nel campo dell'analisi video e audio.
Un'altra importante possibilità che viene data all'operatore riguarda la possibilità di estrarre in modo automatico i percorsi tipici degli utenti, senza necessità di conoscere il particolare scenario (clustering): ogni percorso può rappresentare, a seconda del particolare dominio applicativo, gli interessi comuni dei clienti in un centro commerciale o un comportamento normale, relativo tanto a una persona che si sposta in una stazione ferroviaria quanto a dei veicoli che attraversano un'autostrada. L'algoritmo di clustering proposto, basato su una struttura ad albero, sfrutta la metrica somiglianza sopra definita al fine di svolgere il suo compito. Questo metodo è stato sperimentato su diversi dataset standard ed è stato confrontato con altri metodi allo stato dell'arte, confermando la sua applicabilità in scenari reali e affollati.
D'altra parte, l'operatore può sottomettere diverse tipologie di query al framework proposto, andando a personalizzare i parametri solo a query-time, ovvero nel momento in cui la query è pensata: esempi di query che possono essere sottomesse al sistema riguardano gli oggetti che attraversano una determinata area in un dato intervallo di tempo (dynamic spatio temporal query) o gli oggetti che seguono una particolare traiettoria, simile a quella disegnata a mano dall'utente (query by sketch).
Ciascuno dei moduli proposti e sopra menzionati è stato testato sia su dataset standard che in ambienti reali e i promettenti risultati ottenuti confermano tanto l'avanzamento rispetto allo stato dell'arte, quanto la loro applicabilità in scenari reali.