TESI DI LAUREA

INTERPRETAZIONE AUTOMATICA DI SEQUENZE D'IMMAGINI PER LA VIDEOCOMUNICAZIONE

Alberto Avanzi

matr. 628700

Relatore: Chiar.ma Prof.ssa Giuseppina GINI

Milano, 3 dicembre 2003

Dedica...

Ringraziamenti

Ringrazio innanzitutto la Prof.ssa Gini, relatrice del presente lavoro, per le gentilezza e la disponibilità che mi ha sempre dimostrato in questi ultimi mesi. I suoi preziosi consigli sono stati fondamentali per la felice redazione di questo documento.

Il presente lavoro è frutto di mesi di ricerca nei laboratori Orion dell'INRIA di Sophia Antipolis. Ringrazio in primis François Brémond, per la pazienza, la passione e il senso critico con i quali ha animato le nostre numerose, interminabili discussioni di lavoro (e non solo). Gran parte del successo di questo progetto di ricerca è merito suo. La mia riconoscenza va poi alla direttrice dell'équipe Orion, M.me Thonnat, per la competenza con la quale mi ha diretto e consigliato nel corso di tutta l'attività. Ringrazio poi i membri dell'équipe Orion, e in modo particolarmente sentito M.me Cortell, per l'insostituibile attività di supporto, svolta con encomiabile efficenza e immancabile sorriso.

Sebbene non abbiano contribuito in alcun modo a questo progetto, non posso dimenticare due scapestrati compagni d'avventure, l'ing. Armando Beffani e l'ing. Marco Chierchia. La nostra avventura francese, senza la quale nulla di tutto questo sarebbe arrivato a compimento, cominciò coraggiosamente tre anni fa. Per Marco non si è mai conclusa, per Armando è "in pausa" e per me... si vedrà. Li ringrazio di cuore per aver allietato con la loro compagnia e la loro presenza due magnifici anni di studi.

Ringrazio poi l'ing. Guido Pagana, compagno di indimenticabili stagioni in Costa Azzurra. Per la sua generosità, la sua pazzia, sempre più incontenibile, e la sua ... auto.

Infine, un sorriso particolare, anche se ne riceverà infiniti altri in futuro, a Damiana. Per aver voluto dividere con me, tra le tante altre cose, anche innumerevoli serate di studi tristi e silenziosi.

Indice

Elenco delle figure

Capitolo 1

Un contributo originale all'interpretazione automatica di sequenze video

1.1 Inquadramento

1.1.1 Introduzione

Il soggetto di questo studio è l'interpretazione, tramite un sistema informatico autonomo, di sequenze temporali d'immagini (un altro termine utilizzato è interpretazione dinamica di scene). Un sistema autonomo deve essere capace di apprendere e comprendere il mondo che lo circonda grazie a sensori che, nel nostro caso, sono rappresentati da videocamere. I dati percepiti tramite le sequenze d'immagini devono essere trasformati in una rappresentazione interna della scena che permetta al sistema autonomo di spiegare questi dati sulla base di alcuni scenari e di attribuire loro un senso. Questo passaggio dalla percezione all'attribuzione di un significato costituisce il processo d'interpretazione.

Il soggetto è ambizioso, perché mira alla descrizione astratta delle attività che si svolgono in una determinata scena e cerca di spiegare le ragioni di queste attività. Bisogna tuttavia considerare che il contesto della scena è conosciuto; detto contesto comprende in particolare le descrizioni dell'ambiente e dei comportamenti che ci si attende possano essere tenuti.

Nella prima parte di questo capitolo si delineeranno i tratti caratteristici del lavoro di tesi, la problematica in esame, la soluzione ideata e le tecniche utilizzate per implementarla. Un particolare accento verrà posto sulle caratteristiche di originalità che contraddistinguono la soluzione proposta.

La seconda parte del capitolo sarà dedicata alla presentazione dell'organizzazione di questo documento, unita alla sommaria descrizione del contenuto di ogni capitolo.

Il presente lavoro di ricerca si è concluso con la stesura di un articolo, presentato alla conferenza internazionale ICIP 2001 (*International Conference in Image Processing*, http://icip01.ics.forth.gr/, Tessalonica, 7-10 ottobre 2001). Il testo integrale dell'articolo è riportato nell'appendice B.

1.1.2 Videoconferenza e "MediaSpaces"

Attualmente la ricerca nel dominio dell'interpretazione automatica di sequenze video offre numerosissime applicazioni in svariati campi. Nel capitolo 2 (cf il paragrafo 2.1.1 a pagina 19) verranno presentate alcune di queste possibilità applicative. Il presente lavoro

di ricerca è un primo tentativo di offrire una soluzione ad una problematica aperta nel dominio della videoconferenza e dei MediaSpaces.

Attualmente il termine *videoconferenza* non è caratterizzato da una definizione univoca, essendo spesso funzione della piattaforma particolare che la realizza. Possiamo tuttavia affermare che alcune sue caratteristiche universalmente riconosciute della videoconferenza sono:

- la trasmissione di un flusso audio e video avviene da un sito A ad un sito B, in generale geograficamente distanti;
- la natura della trasmissione è bidirezionale; cioè A invia dati audiovisivi a B e B a sua volta ne invia ad A;
- la trasmissione è temporalmente limitata ad un certo periodo di tempo di durata variabile;
- la piattaforma utilizzata per rendere possibile la videoconferenza effettua il prelievo dei dati (tramite videocamera/e e microfono/i), il loro trattamento, la trasmissione, la ricezione e la visualizzazione/riproduzione degli stessi. Eventuali estensioni del servizio sono naturalmente possibili, ma si configurano principalmente come *servizi accessori*.

A fianco della videoconferenza, la ricerca si orienta verso piattaforme in grado di fornire un servizio dalle caratteristiche differenti (cf [1], [2], [3], [4]):

- la connessione è estesa ad un numero maggiore di siti, geograficamente separati; a priori questo numero è indeterminato;
- il legame video diventa permanente, 24 ore su 24;
- la connessione permanente non prevede legame audio;
- a causa della natura permanente della connessione, e del fatto che questa è estesa a più utilizzatori, si è costretti a occupare una banda passante minore; di conseguenza la frequenza di trasmissione delle immagini è molto minore rispetto alla videoconferenza;
- è possibile, per due utilizzatori A e B del sistema, passare in modalità *videoconferenza*; in questo caso il sistema ripristina le caratteristiche tipiche di questa modalità di connessione, attivando il legame audio e incrementando il numero di immagini trasmesse per secondo (limitatamente ai due utilizzatori A e B).

Si indica questa seconda forma di connessione video permanente utilizzando il termine inglese di *MediaSpace*, introdotto agli inizi degli anni '90 per indicare qualsiasi mezzo (mediatico) capace di fornire un supporto alla coesione di équipes distribuite su siti geograficamente separati (cf [5]). Il legame video rafforza il senso di appartenenza ad una stesso gruppo e facilita la comunicazione informale, intesa come quell'insieme di scambi relazionali gestiti, per lo più, dal caso (per esempio gli incontri casuali in corridoio o sulle scale, oppure lo scambio di saluti attraverso la porta dell'ufficio).

In una tipica piattaforma MediaSpace come Cavecat (cf [6]), Cruiser (cf [7]) o Montage (cf [5]), l'utilizzatore del sistema può "gettare un colpo d'occhio" in un ufficio lontano (grazie alla connessione video permanente), iniziare una videotelefonata o comunque mantenere un legame video con i collaboratori distanti.

La figura 1.1 illustra un esempio di una tipica piattaforma MediaSpace: CoMeDi (cf [8]).

Figura 1.1: *Una vista della tipica schermata di una piattaforma MediaSpace (in questo caso CoMeDi, piattaforma sviluppata dall'équipe IIHM dell'università Joseph Fourier di Grenoble). Si può notare come ogni utente distante sia raffigurato tramite le immagini trasmesse dalla videocamera personale (eventualmente filtrate, come nel caso dell'utente in basso a sinistra).*

1.1.3 Nuovi requisiti introdotti da un MediaSpace

La caratteristica fondamentale che distingue un MediaSpace da una videoconferenza è il fatto che il legame video sia attivo 24 ore su 24. In altri termini, gli utilizzatori del servizio sono ripresi da una (o più) videocamere, e le loro immagini trasmesse agli altri utilizzatori 24 ore su 24. Semplificando all'estremo, l'idea è di offrire un supporto visivo *continuo nel tempo* che possa essere, in un certo senso, l'alternativa alle ore di lavoro trascorse in uno stesso ufficio.

Questa caratteristica porta alla nascita di importanti problemi di tutela della privacy, e di *protezione dello spazio privato*. Con questo termine vogliamo indicare la tendenza, più o meno marcata, di ogni persona, ad identificare alcuni spazi del proprio ambiente lavorativo come "personali", e, in quanto tali, da proteggere in modo più o meno geloso dalle attenzioni altrui. Tali spazi possono essere largamente variabili da persona a persona. Per fare alcuni esempi, essi potrebbero essere: la scrivania di lavoro, lo schermo del computer, i cassetti, gli schedari, anche la propria sedia. In generale essi si identificano con lo spazio fisico che circonda la postazione di lavoro.

Affinché una piattaforma MediaSpace sia accettata ed utilizzata come uno strumento di lavoro, è necessario garantire sicuramente:

1. i servizi tipici di un MediaSpace (vedi paragrafo 1.1.2);

2. un servizio efficace e completo di *protezione dello spazio privato*;

3. la trasparenza di questo servizio;

Le prime esperienze nell'ambito dei MediaSpace portarono a piattaforme software capaci di soddisfare essenzialmente il primo requisito (cf [9], [5]). Il servizio di protezione dello spazio privato era gestito in modo binario: o l'utente accettava di trasmettere le proprie immagini, oppure poteva disattivare la videocamera.

Successivi esperimenti di piattaforme MediaSpace furono caratterizzati da sistemi anche raffinati di protezione dello spazio privato. Ingegnose e molteplici erano le soluzioni offerte all'utente affinché egli potesse adattare le immagini trasmesse alle proprie esigenze. Le protezione attivabili erano abbastanza modulabili: dall'oscuramento del flusso di immagini, alla parziale mascheratura (tramite filtri di distorsione), a filtri ancora più intelligenti capaci di "cancellare" selettivamente alcune parti delle immagini trasmesse (per esempio il volto di un eventuale visitatore).

I periodi di sperimentazione successivi, incentrati sulla raccolta delle impressioni d'uso da parte dell'utente finale, rivelarono come la mancanza di trasparenza a livello della protezione dello spazio privato comportasse il rifiuto della piattaforma da parte dell'utente, che ne percepiva l'utilizzo più come un carico di lavoro aggiuntivo che come una facilitazione offertagli (cf [6], [10], [11]).

Immediatamente si introdusse la possibilità, per ogni utente del sistema, di rendere noto agli altri utilizzatori il proprio *livello di disponibilità*, in forme e modalità variabili nelle varie implementazioni MediaSpace. Questo livello di disponibilità è utile per agevolare gli altri utilizzatori nella scelta del momento più opportuno per una videoconferenza o una telefonata. L'inconveniente maggiore restava appunto la mancanza di trasparenza: le piattaforme, in quanto non intelligenti, erano incapaci di applicare automaticamente sia le protezioni disponibili che i cambiamenti nel livello di disponibilità; essi restavano quindi totalmente a carico dell'utente finale. Ne conseguiva un effettivo e reale carico di lavoro aggiuntivo per l'utente, che oltre alla quotidiana routine lavorativa doveva preoccuparsi anche di adattare manualmente la piattaforma MediaSpace alle proprie esigenze.

Il presente lavoro di tesi rappresenta un primo tentativo, in quest'ambito di ricerca, di soluzione a questo genere di tematiche: **ideare, sviluppare ed implementare tecniche capaci di dotare una piattaforma MediaSpace della capacità di proteggere automaticamente lo spazio privato dell'utente.**

1.2 Il lavoro di tesi

1.2.1 Caratteristiche generali

Nell'ambito delle piattaforme MediaSpace si è quindi focalizzata l'attenzione sui problemi di protezione automatica dello spazio privato. Se si vuole rendere una piattaforma MediaSpace capace di agire in modo autonomo (ed intelligente) sia sul flusso di immagini che sta trasmettendo, sia sull'indicatore del livello di disponibilità, in modo da adattarli alle esigenze dell'utente, bisogna affiancare a questa piattaforma dei programmi capaci di interpretare le immagini filmate, in modo da comprendere, anche solo a livello rudimentale, che cosa stia succedendo nella scena. Sulla base di queste informazioni, comunicate alla piattaforma MediaSpace, questa sarà in grado di decidere come operare. Lo schema della figura 1.2 mostra l'architettura di principio di una piattaforma classica e le variazioni introdotte al fine rendere la piattaforma intelligente.

Nella figura (a) distinguiamo, oltre all'hardware e al software che costituiscono la piattaforma stessa, uno (o più) dispositivi di acquisizione delle immagini (videocamere), un dispositivo di visualizzazione delle immagini degli altri utenti del sistema, l'interfaccia con la rete di trasmissione dati e un dispositivo di input, che schematizza la necessità da parte dell'utente di intervenire sul sistema, cambiandone adeguatamente il comportamento al variare delle situazioni. In base alle scelte dell'utente, le immagini che saranno trasmesse agli altri utilizzatori potranno essere in chiaro, oppure modificate dagli algoritmi di protezione dello spazio privato implementati.

Nella figura (b) notiamo come, rispetto alla precedente, si sia sostituito il dispositivo di comando da parte dell'utente con un sistema dedicato capace di portare a termine l'interpretazione delle immagini filmate da una videocamera aggiuntiva. Grazie alle informazioni che questo sistema può fornire alla piattaforma MediaSpace, questa sarà in grado di proteggere automaticamente lo spazio privato dell'utente senza che egli debba intervenire personalmente.

1.2.2 Prerequisiti e specifiche

Il lavoro di tesi si concentra sull'ideazione, progettazione e realizzazione di un prototipo della piattaforma di interpretazione delle immagini filmate. Questo prototipo farà uso di alcune routines e moduli già esistenti (per esempio il modulo di interpretazione), così come il modulo di acquisizione delle immagini. Le altre parti della catena d'elaborazione sono invece originali e create ex-novo sulla base delle esigenze del progetto e dello stato dell'arte in materia Si vedano ulteriori particolari nel paragrafo 1.2.4.

Per limitare la difficoltà del problema, si sono fatte alcune ipotesi a priori:

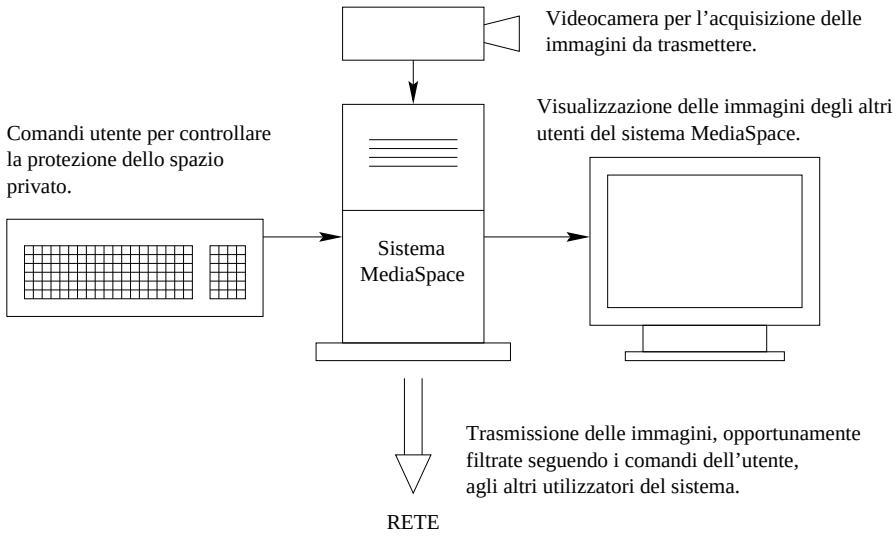
1.2.2.1 L'oggetto del riconoscimento: gli scenari

La piattaforma sarà in grado di riconoscere, con una ragionevole probabilità di errore, alcuni scenari predefiniti. Con il termine *scenario* si intende la descrizione, espressa utilizzando un formalismo adeguato, dell'insieme degli eventi, uniti agli adeguati vincoli temporali, che permettono di esprimere univocamente il verificarsi di una situazione ben determinata (si veda anche il paragrafo 3.2.3.3 per un esempio di scenario).

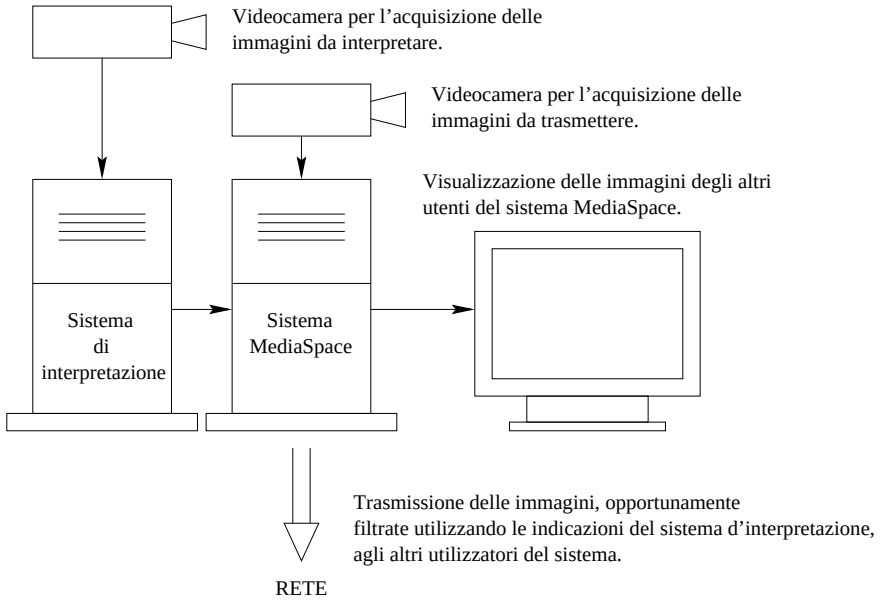
Questa ipotesi a priori ci permette di limitare lo spazio degli scenari da riconoscere ai più probabili ed attendibili in una determinata situazione (per esempio, la vita quotidiana in un ufficio). Da notare che l'ipotesi in sè non è limitativa nei confronti delle possibilità interpretative del sistema che stiamo definendo. Si vedrà nel seguito che il riconoscimento di uno scenario si basa sul riconoscimento di

- determinati eventi;
- adeguati vincoli temporali (nella loro forma più generale essi esprimono l'ordine temporale di accadimento dei differenti eventi).

La definizione di un nuovo scenario da riconoscere comporta quindi, essenzialmente, la capacità di riconoscere determinati eventi (essendo infatti elementare effettuare l'ordinamento temporale degli stessi). Questi eventi, caratterizzanti il nuovo scenario, possono essere standard, cioé il sistema è già in grado di riconoscerli, oppure richiedere routines apposite, create ex-novo, capaci di portare al loro riconoscimento. Queste ultime saranno, in ogni caso, delle applicazioni diverse di tecniche e procedure già implementate e sperimentate per il riconoscimento degli eventi standard, e quindi, pur necessitando di una fase



(a) Lo schema architetturale MediaSpace standard



(b) Lo schema architetturale MediaSpace intelligente

Figura 1.2: *Nella figura (a) possiamo osservare l'architettura di principio di una piattaforma MediaSpace standard. La figura (b) mostra le modifiche introdotte per rendere la piattaforma intelligente.*

aggiuntiva di programmazione (quindi di un'estensione delle capacità del sistema), faranno ricorso comunque a metodologie già formalizzate.

Risulta quindi evidente che la vera difficoltà del progetto consiste nell'ideare e nell'implementare un insieme di tecniche di analisi ed interpretazione di sequenze di immagini capaci di fornire, potenzialmente, gli strumenti e le informazioni necessari per riconoscere uno spettro il più possibile ampio di eventi. In altri termini, il sistema deve essere il me-

no specializzato possibile dal punto di vista della varietà degli eventi riconoscibili, perché proprio dalla completezza di quest'ultimo insieme risulteranno la potenza e l'efficacia della piattaforma stessa (se non dal punto di vista pratico, sicuramente in termini potenziali).

Nel capitolo 3.2.3.3 si daranno alcuni esempi di scenari riconosciuti dal sistema, nonché la loro formalizzazione dal punto di vista degli eventi.

1.2.2.2 Le informazioni a priori: la descrizione della scena

Come si vedrà più in dettaglio nei capitoli 2 e 3, le tecniche di interpretazione di sequenze d'immagini più promettenti, al giorno d'oggi, fanno uso di una descrizione geometrica dell'ambiente nel quale si effettua l'interpretazione stessa (cf [12], [13]). Tale descrizione ha molteplici finalità, prima fra tutte effettuare con buona precisione la collocazione spaziale degli oggetti in movimento (una descrizione completa delle altre finalità di tale descrizione sarà presentata nei succitati capitoli). Gli algoritmi interpretativi ideati e implementati nel presente lavoro di tesi si basano tutti sulla conoscenza della geometria tridimensionale del luogo geometrico ove si conduce l'interpetazione stessa.

In aggiunta a ciò, data l'applicazione MediaSpace di tali algoritmi, si è introdotta l'ulteriore ipotesi che il luogo sia un ufficio "classico" (cioé non "open-space": deve essere delimitato da pareti che si elevano dal pavimento fino al soffitto) e non uno spazio aperto (come in altre applicazioni d'interpretazione).

1.2.3 Descrizione qualitativa del problema

Una formulazione descrittiva del problema per il quale è stata ideata ed implementata una possibile soluzione, che tenga conto degli obiettivi e delle ipotesi a priori che si sono introdotte, è la seguente:

> *Si dispone di una piattaforma MediaSpace "classica" (nel senso che possiede le caratteristiche descritte al paragrafo 1.1.2), alla quale si vuole integrare un modulo di interpretazione automatica di sequenze video. Lo scopo di tale integrazione è comandare in modo appropriato la piattaforma MediaSpace in modo che essa possa effettuare le due operazioni seguenti senza intervento da parte dell'utente:*
>
> - *filtrare, dove opportuno, e secondo modalità scelte una volta per tutte dall'utente, le immagini diffuse agli altri utilizzatori MediaSpace in base a ciò che avviene nelle immagini filmate;*
> - *adattare automaticamente l'indicatore del livello di disponibilità dell'utente in base alle regole che egli ha fissato una volta per tutte e a ciò che accade nella scena.*
>
> *Si suppone che la piattaforma MediaSpace sia in grado di ricevere questo tipo di comandi (tramite un protocollo stabilito). L'interpretazione effettuata consiste nel riconoscimento di un certo numero di scenari predefiniti e nel conseguente comando della piattaforma MediaSpace. Il luogo dell'interpretazione è un ufficio, il cui spazio tridimensionale (chiuso, salvo le normali porte e finestre) si suppone completamente conosciuto (nel senso che si dispone di una descrizione tridimensionale adeguata). L'acquisizione delle immagini è effettuata tramite una videocamera, indipendente da quella dedicata all'acquisizione delle immagini da trasmettere via MediaSpace. Gli algoritmi di elaborazione saranno eseguiti da un comune elaboratore, che si interfaccerà via rete alle*

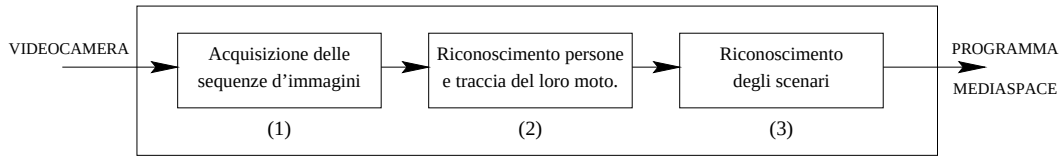


Figura 1.3: *Una prima decomposizione funzionale dell'SDI.*

eventuali periferiche in uso e ad altri elaboratori (per esempio, quello su cui è in esecuzione il pacchetto MediaSpace).

1.2.4 Architettura del sistema

Partendo da questi requisiti di massima, la descrizione funzionale del sistema di interpretazione (indicato con l'acronimo *SDI* nel seguito) è stata decomposta approssimativamente nel modo illustrato dalla figura 1.3:

1. Un primo modulo (o serie di moduli) si occupa dell'acquisizione delle sequenze d'immagini; le immagini, filmate da una videocamera ad una frequenza fissa prestabilita (circa 5 immagini per secondo) sono memorizzate nella memoria di massa dell'elaboratore e da qui prelevate dai moduli seguenti al fine di elaborarle;

2. Questa seconda parte della catena d'elaborazione si occupa di riconoscere la presenza di esseri umani in ognuna delle immagini filmate (nel gergo specialistico anglosassone questa fase è indicata con il termine "human detection") e, in relazione alle immagini già elaborate ed alla posizione che uno stesso individuo vi occupava, ricostruire il movimento della persona, al fine di comprendere che cosa stia facendo ("tracking"). Notare che con il termine *inseguimento di un oggetto in movimento* non si indica semplicemente una serie di coordinate (X,Y,Z) al variare del tempo, ma anche le eventuali interazioni della persona con altre persone, con altri oggetti o con zone particolari dello spazio geometrico. In altre parole, una serie di eventi che si rivelano a priori interessanti per la successiva fase interpretativa;

3. Infine, la terza parte dell'SDI è dedicata al riconoscimento degli scenari. Data la conoscenza degli scenari e dei movimenti delle persone presenti nella scena, questo modulo è capace di riconoscere quali degli scenari predefiniti si siano verificati. In seguito questo modulo si occupa di comunicare l'avvenuto riconoscimento dello scenario alla piattaforma MediaSpace; un'altra possibilità è che il modulo comunichi direttamente alla piattaforma non tanto l'avvenuto riconoscimento, quanto i comandi da attuare sul flusso delle immagini (la scelta varia a a seconda di dove si decide di implementare il modulo che calcola le corrispondenze *scenario-comando*).

Il modulo (1) è relativamente semplice, fa uso di tecniche standard e l'implementazione che ne è stata fatta sarà oggetto solo di una rapida presentazione nei capitoli che seguono. Il modulo (3), al contrario, può essere oggetto di interessanti sviluppi, in quanto è un modulo "decisionale", il cui studio costituisce una branca importante dell'intelligenza artificiale. Tuttavia questi sviluppi esulano dal soggetto della nostra ricerca; i capitoli seguenti presenteranno comunque una descrizione sommaria dell'algoritmo, relativamente semplice, utilizzato per implementare questo modulo. L'algoritmo alla base del modulo di interpretazione non è stato oggetto di ricerca (non fa quindi parte del presente lavoro di tesi, che si arresta alla determinazione dei movimenti e alla comunicazione degli stessi, in un formato opportuno, al modulo d'interpretazione); si è fatto uso di un'unità già

collaudata e utilizzata all'interno dell'équipe Orion, concepita e realizzata da N.Rota e M. Thonnat(cf a tal proposito [14]). Tale unità verrà comunque sinteticamente descritta nel seguito.

Scendendo un poco di più nel dettaglio, la figura 1.4 mostra un ulteriore suddivisione funzionale del modulo numero (2), che rappresenta il nucleo del lavoro di tesi.

1. Una prima unità funzionale effettua la *segmentazione* delle immagini, cioè il confronto tra ogni immagine ricevuta ed un'immagine di riferimento, presa a scena vuota (cioè senza persone). Questo confronto porta ad individuare le regioni differenti tra le due immagini. Queste regioni vengono quindi elaborate al fine di stabilire se sono originate da un movimento (altre potrebbero essere originate da cambiamenti di luminosità, oppure semplicemente da rumore). Le regioni riconosciute come *in movimento* vengono trasmesse all'unità funzionale successiva.

2. Questa riceve in ingresso le immagini da elaborare e le regioni in movimento calcolate dal modulo precedente. Analizzando le caratteristiche cromatiche delle porzioni di immagini corrispondenti alle regioni in movimento, essa stabilisce se ogni regione contiene o meno della pelle umana, cioè se essa può corrispondere ad una parte di individuo (il viso, la testa, le braccia...). Questa unità comunica alla successiva le regioni in movimento ricevute in ingresso, opportunamente completate dall'informazione sulla presenza/assenza di pelle umana.

3. Utilizzando la lista delle regioni in movimento dell'immagine attuale e delle T precedenti (cf il paragrafo 1.2.5.2 a proposito del parametro T), e la conoscenza della geometria tridimensionale della scena (*contesto 3D*) l'algoritmo di classificazione etichetta ogni regione in movimento cercando di comprendere che cosa essa rappresenti. Le differenti possibilità sono:

   individuo : la regione in movimento corrisponde ad un individuo o ad una parte di esso;

   individuo occultato : la regione in movimento corrisponde ad un individuo, parzialmente occultato da un oggetto della scena o da un'altra persona;

   gruppo di persone : la regione in movimento corrisponde a più persone vicine;

   oggetto dell'arredo : uno degli oggetti presenti nella descrizione dello spazio tridimensionale (es: sedie, porta, ...) che è stato spostato rispetto all'immagine di riferimento;

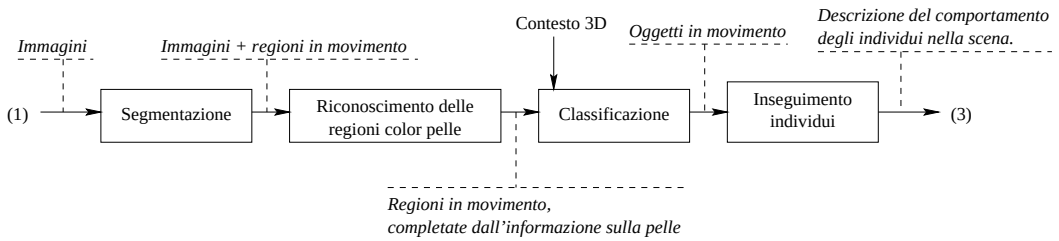


Figura 1.4: *La figura mostra un'ulteriore decomposizione funzionale del modulo (2) mostrato nella figura precedente. Le scritte in corsivo e le relative linee tratteggiate indicano i dati che vengono scambiati tra due moduli, mentre su ogni modulo è riportata la funzione svolta.*

rumore : del rumore nell'immagine viene erroneamente rilevato come una zona in movimento;

indeterminato : si ricorre a questa etichetta quando le informazioni delle quali si dispone sono insufficienti per classificare correttamente la regione in movimento.

Il contesto 3D (vedi anche l'appendice A) si compone di informazioni sulla geometria tridimensionale della scena (muri, pareti, oggetti fisici presenti all'interno) e di informazioni di tipo semantico (per esempio, ogni oggetto ha un nome che permette di individuarlo univocamente. Inoltre il pavimento è suddiviso in differenti aree bidimensionali, ognuna con un nome ben preciso, che permettono di situare con precisione l'individuo). Una informazione di fondamentale importanza facente parte del contesto 3D è la definizione delle cosiddette *aree di ingresso/uscita*, cioè le zone da cui si entra/esce dall'ufficio (situate quindi in corrispondenza delle porte). Le regioni in movimento, una volta opportunamente classificate, prendono il nome di *oggetti in movimento* e vengono comunicate al modulo successivo.

4. Infine, questo modulo ricostruisce i legami temporali tra i differenti oggetti in movimento presenti in immagini successive. Lo scopo è di inseguire gli oggetti in movimento (nella fattispecie gli individui presenti nella scena) (*tracking*). La descrizione del movimento costituisce la base della fase interpretativa successiva.

Gli algoritmi che implementano ognuna di queste unità funzionali sono illustrati nei capitoli seguenti. Nel seguito si anticiperà brevemente la descrizione dell'algoritmo di inseguimento degli oggetti, che costituisce la parte di maggior originalità del presente lavoro di tesi.

1.2.5 L'innovazione: il modulo di inseguimento degli oggetti

1.2.5.1 Introduzione

Lo scopo dell'inseguimento degli oggetti è comprendere l'interazione degli individui presenti nella scena con gli oggetti del contesto (sedie, computers, scrivanie). Al fine di effettuare l'interpretazione del comportamento di un individuo, infatti, è basilare situarne i movimenti nello spazio tridimensionale e riconoscere eventuali interazioni con oggetti di particolare significato (cf [12], [13]). Se, per esempio, disponessimo della descrizione del movimento seguente:

1. Un individuo apre la porta dell'ufficio (interazione con la porta);
2. Lo stesso individuo entra nell'ufficio;
3. Si porta in prossimità del tavolo di lavoro che sappiamo appartenere all'utente A;
4. Si siede sulla sedia che è posizionata dietro al tavolo di lavoro dell'utente A;;
5. interagisce con il computer dell'utente A;
6. il sistema informatico registra un login a nome dell'utente A.

La conclusione immediata che si potrebbe trarne (l'interpretazione, dunque) è che l'utente A (identificato grazie al fatto che si è seduto sulla sua sedia e lavora al suo computer) sta lavorando al proprio computer. Il login nel sistema (informazione che, a priori, potrebbe non essere disponibile) conferma l'attribuzione dell'identità di "utente A" all'individuo

che è entrato nella stanza. A prescindere dalla precisa identità dell'individuo che si muove nell'ufficio (particolare che solleva problematiche di sicurezza che esulano completamente dal presente dominio di ricerca), si è comunque in grado di dire **che un individuo è entrato nell'ufficio e si è messo a lavorare al computer.**

Ciò che permette di arrivare a queste conclusioni sono i movimenti dell'individuo e le interazioni che esso stabilisce. Entrambe queste classi di informazioni faranno quindi parte della nostra descrizione del movimento.

1.2.5.2 Caratteristiche generali

Si è ritenuto di fondamentale importanza, per ideare un algoritmo robusto ed affidabile, elaborare globalmente i dati appartenenti ad una finestra temporale di larghezza T; in altri termini, prendendo in considerazioni tutte le T immagini intercorse tra l'istante $t_f - T$ e l'istante t_f che corrisponde all'immagine in elaborazione, l'algoritmo è capace di interpretare il comportamento degli individui presenti nella sala all'istante $t_f - T$ e contemporaneamente di fare delle ipotesi sui "futuri" spostamenti di questi individui all'interno dell'intervallo $[t_f - T, T]$. Questo ritardo T che intercorre tra il momento dell'acquisizione dei dati e il momento in cui si rende disponibile l'interpretazione offre vantaggi ed inconvenienti, che saranno discussi successivamente nel paragrafo 6.2 a pagina 72.

I dati fondamentali gestiti dall'algoritmo di inseguimento sono dunque:

- gli oggetti in movimento rilevati in ognuna delle T immagini dell'intervallo $[t_f - T, T]$;
- le traiettorie possibili degli individui nell'intervallo $[t_f - T, T]$;
- gli individui effettivamente presenti nell'ufficio all'istante $t_f - T$ e dei quali l'algoritmo fornisce la cronologia degli spostamenti sino all'istante $t_f - T$;

Si veda anche il grafico 1.5 per un'illustrazione grafica di queste entità.

Riguardo alle due entità *traiettoria* e *individuo*, la differenza tra le due è che la traiettoria è la struttura dati che descrive un possibile spostamento nell'intervallo $[t_f - T, T]$ di una persona fisica, mentre l'individuo è la struttura dati corrispondente ad una reale persona che si sta muovendo nella scena filmata, e di cui l'algoritmo fornisce la cronologia del movimento fino all'istante $t_f - T$. Ovviamente ad ogni individuo l'algoritmo associerà una ed una sola probabile traiettoria.

L'algoritmo ideato per l'inseguimento degli oggetti in movimento si basa inoltre su:

- la conoscenza del modello tridimensionale dello spazio ove si svolge l'azione (contesto 3D) (cf anche la figura 1.6);
- un modello dell'individuo, inteso come insieme delle caratteristiche che deve possedere un oggetto in movimento per rappresentare, nella scena, l'immagine di una persona fisica;
- un modello di traiettoria, inteso come insieme delle caratteristiche che deve avere una qualunque serie temporale di oggetti in movimento per poter rappresentare una reale traiettoria fisica di una persona.

1.2.5.3 L'algoritmo

I passi logici nei quali si suddivide l'algoritmo sono dunque i seguenti:

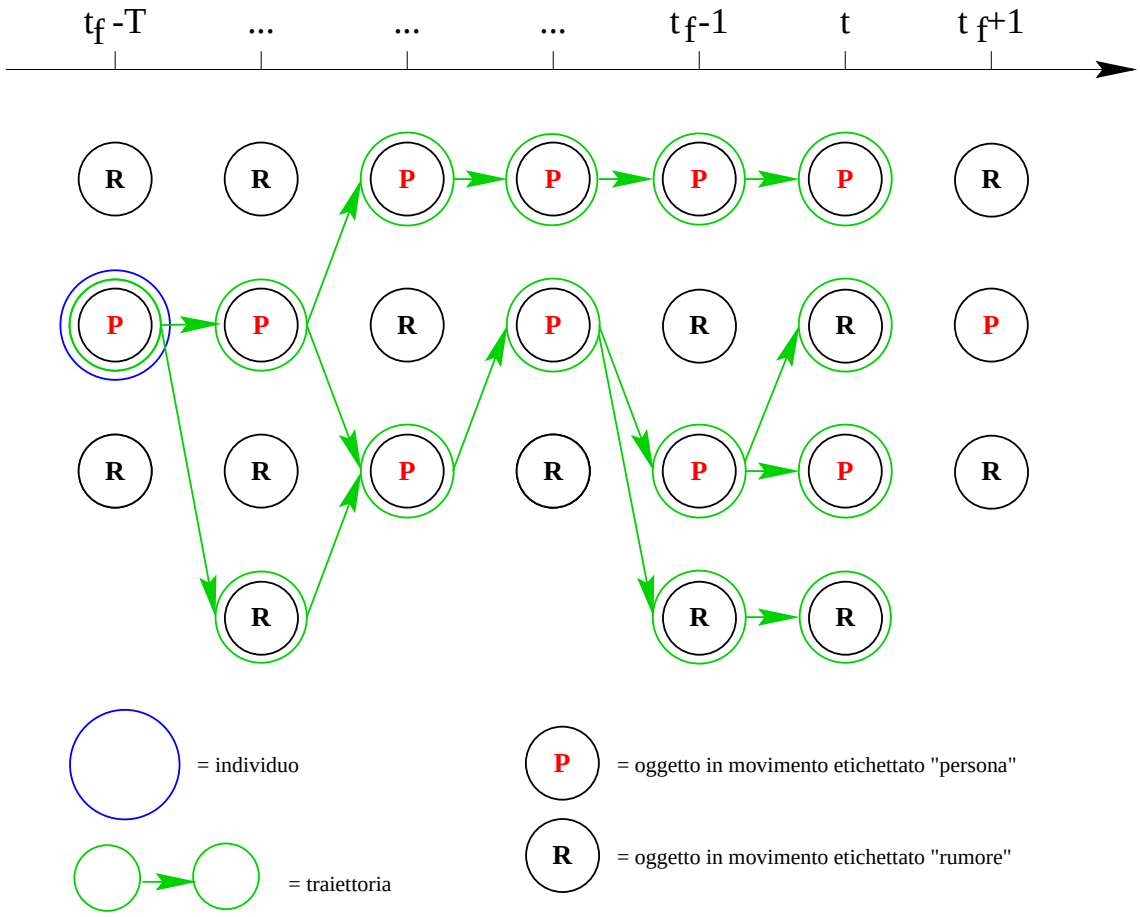


Figura 1.5: *Questa figura rappresenta il grafo degli oggetti in movimento come si presenta prima dell'elaborazione di quelli rilevati all'istante* $t_f + 1$. *In ascissa sono riportati i differenti istanti, mentre in ordinata gli oggetti in movimento rilevati in ogni immagine dell'intervallo di tempo. La finestra temporale considerata ha dimensione* T. *L'ultima colonna a destra mostra gli oggetti in movimento rilevati all'istante* $t_f + 1$ *e non ancora elaborati. In questo esempio si nota un individuo e una serie di traiettorie (7) a lui associate. Il ritardo* T *permette di scegliere la traiettorie che, tra queste, meglio corrisponde all'individuo.*

1. Dato l'insieme delle traiettorie definite all'istante $t_f - 1$ e i nuovi oggetti in movimento rilevati nell'immagine corrente (al tempo t_f), ogni traiettoria viene estesa[1] utilizzando uno dei nuovi oggetti in movimento disponibili; questa fase corrisponde alla fase logica di calcolo dei possibili aggiornamenti dei movimenti (traiettorie) delle persone nell'ufficio.

2. Se uno o più oggetti in movimento al tempo t_f necessitano la creazione di una o più nuove traiettorie, queste vengono create utilizzando tali oggetti in movimento; questa fase corrisponde alla inizializzazione di una nuova traiettoria per una persona fisica che sta entrando nell'ufficio;

3. Dato l'insieme delle nuove traiettorie (estese + create) e l'insieme degli individui

[1]Nel seguito utilizzeremo il termine *estendere una traiettoria* nel senso di aggiornare (update) i dati della struttura *traiettoria* utilizzando i nuovi oggetti in movimento

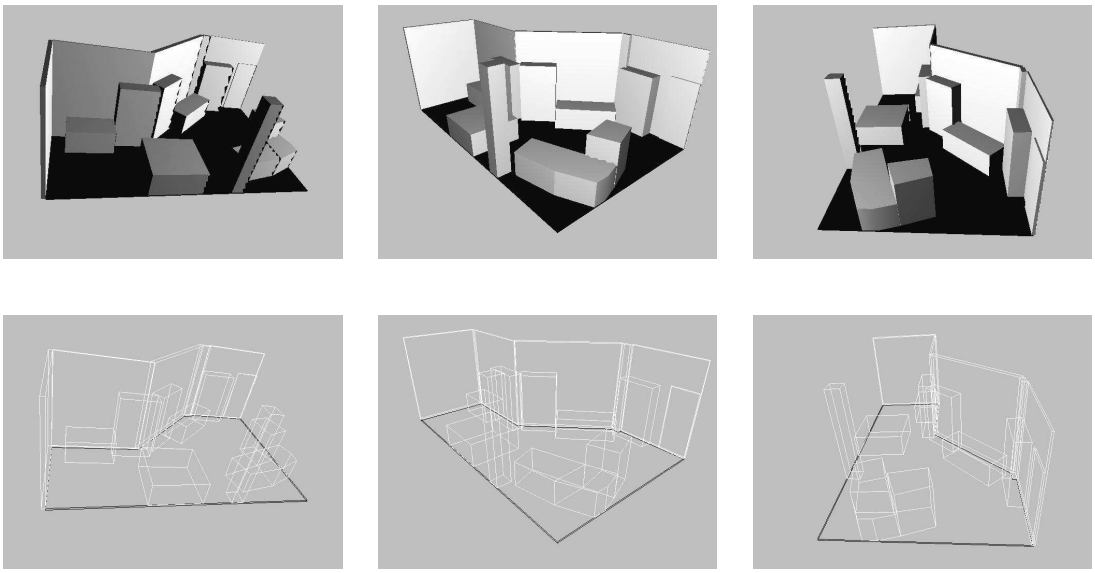


Figura 1.6: *La figura mostra 3 ricostruzioni differenti in formato VRML del contesto 3D. Ogni vista è presentata sia in volumi solidi (prime tre immagini in alto), sia in trasparenza (ultime tre immagini in basso).*

esistenti, si aggiorna[2] ogni individuo utilizzando le nuove traiettorie; questa fase corrisponde all'assegnazione logica della miglior traiettoria ad ogni individuo che si sta muovendo nell'ufficio;

4. Se una o più traiettorie richiedono la creazione di uno o più nuovi individui, questi vengono creati;

5. Se uno o più individui risultano aver lasciato la stanza, questi vengono distrutti;

6. Se una o più traiettorie si rivelano "errate", esse vengono cancellate.

1.2.5.4 Principali caratteristiche dell'algoritmo

L'algoritmo ideato per il modulo di inseguimento degli oggetti fa uso dei due sopraccitati modelli: traiettoria e individuo.

Una traiettoria P_i è composta da una sequenza, ad istanti di tempo successivi, di oggetti in movimento $M_i(t), t \in [t_f - T; t_f]$. Per essere definita *traiettoria*, la sequenza P_i deve soddisfare due condizioni:

- $(a_1)\ \forall\ t\ \exists!\ M(t)$
- $(a_2)\ dist_{2D}(M(t); M(t+1)) \leq D_{max}$.

La condizione (a_1) ha come scopo di limitare il numero di traiettorie calcolate nel caso sia stato rilevato un numero eccessivo di oggetti in movimento (per esempio quando più oggetti in movimento rappresentano parti differenti di una stessa persona). (a_2) evita la creazione di traiettorie con discontinuità spaziali importanti.

[2]Nel seguito utilizzeremo il termine *aggiornare un individuo* nel senso di aggiornare (update) i dati della struttura *individuo* utilizzando una opportuna traiettoria

Un coefficiente di qualità Q_i caratterizza ogni traiettoria P_i; esso è calcolato prendendo in considerazione tre aspetti differenti:

- la coerenza tra ogni $M_i(t)$ di P_i e il modello dell'essere umano, dal punto di vista delle dimensioni tridimensionali e dell'etichetta dell'oggetto in movimento (*individuo* piuttosto che *gruppo*);
- la coerenza temporale tra i differenti oggetti in movimento $M_i(t)$ che costituiscono P_i (calcolata dal punto di vista delle dimensioni tridimensionali e delle caratteristiche dello spostamento).
- la distanza bidimensionale (nel piano dell'immagine, quindi) tra $M_i(t)$ e $M_i(t+1)$, confrontata con il modello del movimento umano.

Una nuova traiettoria è creata ogniqualvolta un oggetto mobile è rilevato nella zona di ingresso/uscita. Questa zona (bidimensionale, sul piano dell'immagine) è calcolata utilizzando il modello 3D della scena e corrisponde approssimativamente all'immagine della porta dell'ufficio.

Una traiettoria è aggiornata ad ogni istante di tempo utilizzando gli oggetti in movimento rilevati all'istante t_f ed è cancellata quando si verificano due condizioni: nessun nuovo oggetto in movimento è rilevato (e non è quindi più possibile aggiornarla) e gli ultimi oggetti in movimento che la costituiscono si trovano nella zona di ingresso/uscita (tipica situazione che si verifica quando una persona è uscita dall'ufficio).

Un individuo I_j è la struttura che rappresenta il dato in uscita del modulo di inseguimento. Ogni individuo è associato ad una serie di traiettorie, che rappresentano tutti i suoi possibili percorsi futuri. Una nuova struttura individuo viene creata quando si ha la sicurezza che essa corrisponda ad una persona reale (cioè quando la traiettoria P_i, non ancora associata con nessun individuo, soddisfa le quattro condizioni seguenti):

- P_i comincia nell'area di ingresso/uscita;
- finisce al di fuori di quest'area;
- è composta da T oggetti in movimento;
- non si sovrappone ad una seconda traiettoria P_j utilizzata da un altro individuo.

Lo scopo di queste quattro condizioni è di evitare la creazione di un nuovo individuo a causa di una traiettoria affetta da errori di rilevamento (per esempio dovuti alle ombre). Il ritardo temporale T è fondamentale per avere la sicurezza che ogni individuo sia associato alla traiettoria corretta.

Un individuo è aggiornato utilizzando la traiettoria che gli si addice maggiormente. Esso è cancellato quando l'ultimo oggetto in movimento che lo compone si trova nella zona di ingresso/uscita e quando nessuno dei nuovi oggetti in movimento che sono rilevati gli corrispondono.

1.2.5.5 Principali problemi incontrati e soluzioni offerte

Una primo problema è l'esplosione combinatoria del numero di traiettorie calcolate che si otterrebbe se si tentasse di prendere in considerazione l'estensione di ogni traiettoria esistente con ogni nuovo oggetto in movimento rilevato. La soluzione ideata consiste nel dividere l'insieme delle traiettorie esistenti all'istante t_f in $N+1$ sottoinsiemi S_j, dove

N è pari al numero di individui esistenti al tempo $t_f - T$. Ogni sottoinsieme S_j contiene tutte le traiettorie associate all'individuo I_j. Un sottoinsieme aggiuntivo S_0 contiene le traiettorie (incomplete) corrispondenti ad una persona che sta entrando nell'ufficio (il cui corrispondente individuo non è quindi stato ancora creato).

Si è stabilito che ogni sottoinsieme S_j riceva un numero di possibilità d'aggiornamento pari a N_{ext} (dove N_{ext} è un parametro dell'algoritmo); in altri termini, ogni sottoinsieme S_j può dare origine a massimo N_{ext} nuove traiettorie (ognuna ottenuta estendendo una traiettoria di S_j con un nuovo oggetto in movimento); una volta effettuati tutti gli aggiornamenti, le vecchie traiettorie vengono eliminate per essere sostituite dalle nuove, aggiornate. Ogni traiettoria P_{ji} (appartenente al sottoinsieme S_j) riceve un numero di estensioni N_{ji} proporzionale al coefficiente di qualità Q_{ji} che la contraddistingue rispetto alla qualità totale Q_j del sottoinsieme S_j:

$$N_{ji} = N_{ext}\frac{Q_{ji}}{Q_j}, \quad con \quad Q_j = \sum_{k \in S_j} Q_{jk}.$$

In ogni caso il massimo numero di estensioni concesso ad ogni traiettoria è pari al numero di nuovi oggetti in movimento rilevati O_{max}. Se $N_{ji} < O_{max}$ allora solo gli N_{ji} oggetti in movimento che offrono le estensioni con coefficiente di qualità più elevato sono utilizzati per estendere la traiettoria P_i.

Un secondo problema è il calcolo della migliore associazione tra ogni individuo ed una delle traiettorie. Il problema principale è di evitare alcune situazioni di errore tipiche, per esempio:

- associare a due individui, corrispondenti a due persone che nella realtà si sono incrociate, la stessa traiettoria; un tipico errore potrebbe essere, una volta che le due persone si separano nella realtà, associare entrambi gli individui alla stessa traiettoria (una delle due); ne risulterebbe che una persona fisica è "seguita" tramite due strutture *individuo* mentre l'altra è "persa";

- associare ad ognuno dei due individui, corrispondenti a due persone che nella realtà si sono incrociate, la traiettoria dell'altro; l'errore è che le traiettorie, anch'esse intersecantisi, sono facilmente scambiabili l'una con l'altra. Ne deriverebbe un errore di identità sulle due persone fisiche.

Per evitare questi problemi, l'algoritmo calcola quattro coefficienti che esprimono le mutue relazioni tra traiettorie ed individui, tra traiettoria e traiettoria e tra individuo ed individuo:

- D_1: la traiettoria P_k e l'individuo I_q sono **associati** se hanno in comune lo stesso oggetto in movimento all'istante $t - T$. Le traiettorie associate all'individuo I_q rappresentano tutte e sole le sue possibilità di spostamento (all'interno dell'intervallo temporale $[t_f - T; t]$). Per ogni possibile associazione $(I_q;P_k)$ viene quindi calcolato un coefficiente di compatibilità $D_1(I_q; P_k)$ che prende in considerazione la compatibilità tra le dimensioni e la direzione di I_q e le dimensioni/direzioni delle varie P_k.

- D_2: due traiettorie **si sovrappongono** l'una con l'altra se esse hanno in comune almeno un oggetto in movimento. Un coefficiente $D_2(P_q; P_r)$ è utilizzato per quantificare questa sovrapposizione; esso decresce esponenzialmente al crescere del numero

di oggetti in movimento comuni:

$$D_2(P_q; P_r) = e^{-\left(\frac{N_{ov}}{K}\right)^2}$$

dove N_{ov} è il numero di oggetti in movimento (che non siano però classificati come *gruppo* [3]) comuni a P_q e P_r e K è un coefficiente di normalizzazione.

- D_3: due individui si **incrociano** se una qualsiasi delle traiettorie associate al primo ed una qualsiasi di quelle associate al secondo si sovrappongono; questa informazione assume valori binari: 1 = sovrapposizione, 0 = nessuna sovrapposizione;
- D_4: una coppia individuo-traiettoria è classificata *persona* o *gruppo* a seconda della classe degli oggetti in movimento che costituiscono la traiettoria e a seconda che l'individuo sia *isolato* (cioè se l'oggetto in movimento che lo costituisce non è comune a nessun altro individuo).

Dopo aver calcolato queste relazioni, l'algoritmo come prima cosa aggiorna gli individui che non ne incrociano nessun altro (cioé con $D_3 = 0$) utilizzando, per ognuno d'essi, la traiettoria che offre il miglior coefficiente di compatibilità. Quindi, per calcolare in modo globale ed ottimale l'associazione tra le traiettorie e i restanti individui, l'algoritmo calcola gli insiemi V_h degli individui che si incrociano l'uno con l'altro ($D_3 = 1$). All'interno di ognuno di questi insiemi V_h l'algoritmo sceglie la coppia $I_j - P_k$ con il miglior coefficiente di compatibilità (D_1). Questa coppia è aggiornata e, se l'individuo è classificato come *persona* ($D_4 = persona$), il coefficiente di compatibilità delle altre coppie $I_l - P_m$ è diminuito se la traiettoria P_m si sovrappone a P_k (D_2):

$$D_{1new}(I_l; P_m) = D_{1old}(I_l; P_m) \cdot D_2(P_k; P_m).$$

Infine l'algoritmo seleziona la seconda miglior coppia $I_{j'} - P_{k'}$ con il secondo miglior coefficiente di compatibilità (D_1). L'algoritmo è iterato finché, per ogni insieme V_h, tutte le coppie dell'insieme (contenente gli individui che si incrociano) sono state processate.

Una terza problematica è la gestione dei mancati rilevamenti: si ha un mancato rilevamento quando l'oggetto mobile corrispondente alla persona reale filmata non è stato rilevato. Per risolvere questo problema si è introdotto l'oggetto mobile virtuale M_{lost}. Ogniqualvolta l'algoritmo effettua l'estensione delle traiettorie, questo oggetto mobile virtuale è considerato al fine di calcolare una possibile estensione aggiuntiva. In questo caso, il coefficiente di qualità $Q_{i,lost}$ della traiettoria è pari ad una costante numerica che rappresenta la mancanza d'informazione rappresentata dall'oggetto mobile virtuale M_{lost} (in quanto virtuale, appunto, questo oggetto mobile non apporta nessun tipo di informazione). La traiettoria estesa tramite M_{lost} ha un coefficiente di qualità più basso di quello di qualsiasi altra traiettoria P_i estesa utilizzando un vero oggetto in movimento; tuttavia, in caso di mancato rilevamento dell'oggetto in movimento, questa traiettoria può ottenere il coefficiente di qualità più elevato.

1.3 Cenni tecnici d'implementazione

Il sistema presentato è stato studiato, ideato e realizzato nei laboratori del centro di ricerca INRIA, a Sophia Antipolis (Francia). L'appendice C presenterà rapidamente questa struttura.

[3] Si considera corretto il caso in cui un oggetto in movimento etichettato *gruppo* sia comune a due traiettorie, in quanto la definizione stessa prevede che un oggetto mobile *gruppo* sia l'immagine di più persone contemporaneamente.

La piattaforma di interpretazione automatica è stata sviluppata per la maggior parte in linguaggio C, su piattaforme UNIX/Linux. Nella fattispecie, il programma è stato installato e testato su una piattaforma con le seguenti caratteristiche:

- Architettura Intel Pentium III 600 Mhx, biprocessore con 256 Mb di RAM;
- Sistema operativo Linux RedHat 6.1;
- Videocamera grandangolo 180ř per la cattura delle immagini da elaborare.

Pur disponendo di un'installazione funzionante del pacchetto MediaSpace CoMeDi (sviluppato dall'équipe IIHM dell'istituto Joseph Fourier di Grenoble, cf [8]), non si è potuto pilotare direttamente tale piattaforma in quanto essa non era predisposta per interfacciarsi con un modulo di comando esterno. Per rendere "visibili" i risultati dell'interpretazione, si è quindi deciso di creare un modulo aggiuntivo capace di tradurre in linguaggio VRML 1.0 lo scenario 3D, le persone presenti nell'ufficio e un indicatore del livello di disponibilità (sotto la forma di un cubo di diversi colori) capace di indicare l'avvenuto riconoscimento dello scenario e il conseguente cambiamento del livello di disponibilità. I risultati ottenuti saranno illustrati nel capitolo 7.

1.4 Struttura di questo documento

Capitolo 1 Si introduce il concetto di *MediaSpace* sottolineandone le caratteristiche principali e le differenze rispetto al concetto di *videoconferenza*. Si delinea la storia delle piattaforme MediaSpace, dalla loro comparsa fino ai giorni nostri e si forniscono alcuni riferimenti bibliografici per approfondire l'argomento. Viene quindi delineato il lavoro di tesi nelle sue caratteristiche generali, descrivendo il problema per il quale si è studiata una soluzione. In seguito si dettaglia la struttura della piattaforma proposta, spiegando il ruolo assolto da ogni sottounità funzionale e descrivendo le interfacce tra le diverse sottounità. Si pone quindi l'accento sul modulo di inseguimento degli oggetti in movimento, in quanto caratterizzato da maggior originalità e dal maggior grado di innovazione. Viene illustrato l'algoritmo che lo costituisce. Si completa la descrizione passando in rassegna alcuni dei problemi incontrati ed introducendo per ciascuno d'essi la soluzione offerta.Infine il capitolo descrive alcune caratteristiche tecniche dell'ambiente di sviluppo utilizzato e dei mezzi tecnici a disposizione.

Capitolo 2 Si descrive nella sua generalità la ricerca sull'interpretazione automatica di sequenze video. Vengono illustrate alcune delle principali applicazioni pratiche di tale ricerca, così come le problematiche sollevate dalle nuove scoperte. Si presenta quindi lo stato dell'arte in questo settore e si danno alcuni brevi cenni sulle tecniche che maggiormente si stanno affermando come robuste e suscettibili di un'applicazione industriale. Alla luce dello stato dell'arte, si delineano infine le caratteristiche principali che caratterizzeranno i diversi moduli della piattaforma realizzata.

Capitolo 3 Si introduce nel dettaglio il concetto fondamentale di *base del contesto*, illustrandone la funzione, le caratteristiche e l'implementazione che ne è stata data. In seguito il capitolo analizza l'architettura generale della piattaforma sviluppata, prendendo in considerazione la ripartizione delle differenti fasi d'elaborazione e l'assegnazione di ognuna di queste ad una sottounità funzionale ben precisa. Viene

fornita una descrizione di ognuna delle operazioni che è necessario svolgere per poter interpretare una sequenza d'immagini video. Viene infine brevemente descritto il funzionamento dei moduli facenti parte della piattaforma ma che non sono stati ideati nell'ambito del lavoro di tesi.

Capitoli 4-5-6 Si descrive più approfonditamente una delle tre sottounità funzionali della piattaforma ideate e realizzate nell'ambito del lavoro di tesi. I capitoli analizzano gli algoritmi implementati e descrivono come si è offerta una soluzione a svariati problemi incontrati. Nella fattispecie il **Capitolo 4** si occupa del modulo di classificazione e fusione delle regioni in movimento, il **Capitolo 5** del modulo di riconoscimento della pelle umana e il **Capitolo 6** del modulo di inseguimento degli oggetti in movimento.

Capitolo 7 Si presentano i risultati ottenuti e un'analisi critica degli stessi, sottolineando i punti di forza e di debolezza degli algoritmi ideati.

Capitolo 8 Si presenta una serie di prospettive che si aprono per un eventuale proseguimento della ricerca in questa direzione. Offre una riflessione, sorretta dall'esperienza di 12 mesi di ricerca e sviluppo in questo settore, volta ad indicare alcune linee guida per il prosieguo.

Appendice A Si riporta il glossario dei termini utilizzati in tutto il documento, inserito al fine di permettere di reperire velocemente una definizione o puntualizzare un concetto.

Appendice B Si riporta il testo integrale dell'articolo che sarà presentato alla conferenza internazionale ICIP2001 nell'ottobre 2001 e che sarà pubblicato nello stesso periodo nei *Proceedings* della conferenza.

Appendice C Si presenta succintamente il centro di ricerca INRIA: la storia, l'organizzazione, i settori di ricerca nei quali è impegnato. Viene quindi presentata l'équipe ORION, in seno alla quale è stato sviluppato il presente lavoro: soggetti di ricerca, personale e competenze.

Capitolo 2

L'interpretazione automatica delle sequenze d'immagini

2.1 Cenni introduttivi

Nei paragrafi seguenti si vedrà come alla base dell'interpretazione automatica di sequenze video ci sia uno stretto accoppiamento tra tecniche di visione e tecniche di ragionamento astratto. Questo accoppiamento caratterizza il dominio dell'interpretazione automatica di sequenze video come un soggetto di ricerca di prim'ordine. Ciononostante esso è relativamente recente, essendo apparso nei primi anni '80 in Europa (specialmente in Germania, Inghilterra e Francia) e solo successivamente anche negli Stati Uniti e in Asia.

2.1.1 Alcune applicazioni

Nella sezione seguente si passeranno in rassegna, in modo non esaustivo, alcune grandi applicazioni di questo campo di ricerca. In funzione dei differenti domini applicativi, si descriveranno gli obbiettivi che un sistema d'interpretazione deve raggiungere e le caratteristiche principali del sistema stesso.

Porre l'accento sui differenti domini applicativi è molto importante in quanto permette d'illustrare le prospettive di sviluppo futuro del settore di ricerca, così come di stimare con più esattezza il contributo pratico e direttamente utilizzabile del presente lavoro.

Comune a tutte le applicazioni che saranno presentate è l'obbiettivo: studiare il comportamento degli oggetti in movimento presenti all'interno di una scena. Si tratta dunque di realizzare un sistema, chiamato *sistema d'interpretazione*, capace di analizzare in modo autonomo una scena data a partire da sequenze d'immagini. Le motivazioni di ordine applicativo sono molteplici e riguardano differenti domini:

- **Videosorveglianza del traffico stradale**: le applicazioni di questo dominio sono facilmente convertibili in pacchetti operativi e commercializzabili, soprattutto se confrontate con quelle di altri domini. I veicoli sono oggetti in movimento facilmente identificabili e la gamma dei loro comportamenti è limitata dall'ambiente stesso. Questo settore offre spunto a diversi tipi di applicazioni:
  - Rilevamento d'incidenti: il sistema d'interpretazione previene le situazioni pericolose su una rete stradale, per esempio la *circolazione contromano di un veicolo*, rileva gli incidenti e attiva gli opportuni allarmi. Da tempo esistono dei sistemi commerciali di rilevamento d'incidenti già disponibili all'uso. Altri

sistemi più ambiziosi, soprattutto nell'ambito del riconoscimento dei comportamenti, sono ancora in fase di ricerca e sviluppo. Per esempio, B. Neumann e la sua équipe hanno sviluppato il sistema "Naos", il quale, a partire da una sequenza d'immagini, descrive una scena di traffico ad un ascoltatore che si trova nell'impossibilità di osservarla (cf [15]). L'obbiettivo di Naos è di dare una descrizione in linguaggio naturale delle attività che si stanno svolgendo nella scena. Allo stesso modo, H. Nagel e la sua équipe hanno sviluppato un sistema, "Epex", dedicato all'interpretazione di scene di traffico a partire da una sequenza d'immagini (cf [16]). Infine, il progetto europeo VIEWS aveva come obbiettivo la videosorveglianza di scene esterne in tempo reale, a partire dall'analisi di sequenze d'immagini (cf [17]). Due applicazioni pilota sono state sviluppate in seno a questo progetto: la prima gestisce il traffico aereo a terra, in un aeroporto, mentre la seconda rileva le situazioni di potenziale incidente in una scena di traffico.

  - Rotonde intelligenti: il sistema d'interpretazione ottimizza la circolazione e il passaggio dei pedoni, controllando i semafori di circolazione di diverse rotonde. Alcuni sistemi di rotonde intelligenti sono già operativi e cominciano ad essere commercializzati, come per esempio quello sviluppato da S.Sellam e la sua équipe (cf [18]).
  - Sorveglianza di zone specifiche: il sistema sorveglia delle zone frequentate da veicoli, per esempio le barriere dei caselli o le stazioni di servizio per rilevare dei comportamenti anormali dei veicoli. H. Nagel e i suoi collaboratori hanno per esempio lavorato su un sistema di sorveglianza delle stazioni di servizio (cf [19]). Alcuni sistemi di sorveglianza delle barriere dei caselli sono già stati commercializzati.
  - Navigazione in una scena di traffico: il sistema è imbarcato in un veicolo. Deve rilevare i veicoli circostanti e analizzare il loro comportamento. Questi sistemi di navigazione sono spesso limitati alla sorveglianza dei veicoli circostanti (cf [20]).
  - Sorveglianza aerea: il sistema sorveglia delle zone sensibili, quali una strada, un campo di battaglia, al fine di proteggere gli oggetti di valore (per esempio un convoglio in movimento), o di rilevare dei comportamenti particolari. Questi sistemi di sorveglianza aerea sono ancora in fase di ricerca e sviluppo.

- **Videosorveglianza di attività umane**: in queste applicazioni un sistema d'interpretazione ha come obbiettivo di sorvegliare delle zone caratterizzate da comportamenti specifici, come le stazioni della metropolitana, i parcheggi, i supermercati, gli aeroporti, le banche o le zone pedonali. Lo scopo è di rilevare i comportamenti insoliti degli individui che si muovono in queste zone e di prevenire i comportamenti pericolosi, quali per esempio gli atti di vandalismo, di vendita di stupefacenti, di scippo, d'aggressione o di terrorismo. Per esempio, l'analisi del comportamento dei gruppi, o delle bande, permette di prevenire determinati atti di aggressione. Questo tipo di applicazioni sono in pieno sviluppo e numerosi sistemi di sorveglianza sono allo studio nel mondo industriale ed accademico. Per esempio, il progetto europeo Esprit HPCN PASSWORDS è sfociato nella definizione e prototipazione di piattaforme di sorveglianza delle stazioni della metropolitana, dei parcheggi e dei supermercati tramite una sola videocamera (cf [21]). Allo stesso modo, il progetto Perception aveva come obiettivo la sorveglianza di parcheggi, utilizzando però più videocamere (cf [22]).

Altre applicazioni simili studiano l'analisi statistica dei comportamenti, il cui obiettivo è di contare il numero degli individui e di determinare il loro flusso di circolazione. Tali sistemi sono già operativi (cf [23]). Applicazioni riguardanti l'analisi del comportamento degli animali, come quelle portate avanti da D. Hogg sugli stormi d'uccelli, sono allo studio.

- **Analisi di scene sportive**: in queste applicazioni un sistema d'interpretazione analizza i comportamenti degli sportivi. Queste applicazioni differiscono dai casi precedentemente esposti per la presenza di un ambiente particolare (es: campo di gioco) e per il numero limitato di comportamenti. Gli sport interessati sono principalmente il calcio (cf [24]), il tennis, il basket e il football americano (cf [25]). Questo tipo di sistemi ha per obiettivo di aiutare gli allenatori nell'analisi del gioco e di fornire delle statistiche sulle tattiche utilizzate. Alcuni di questi sistemi sono già commercializzati o sono in via di commercializzazione, anche se altri richiedono ancora correzioni manuali. Una variante a questo tipo di applicazioni è stata abbordata dal progetto VITRA (VIsual TRAnslator). L'obbiettivo di questo progetto è lo sviluppo di un sistema capace di spiegare, tramite un dialogo con l'utilizzatore, il contenuto di una sequenza d'immagini. VITRA in particolare ha portato alla creazione del sistema Soccer. Soccer analizza e commenta simultaneamente in tedesco delle corte sequenze di calcio, come in una radiocronaca (cf [26]).

- **Analisi dei gesti**: in questo tipo di applicazioni la videocamera è fissa, in prossimità dell'individuo, posta frontalmente. In più, i sistemi d'interpretazione spesso prendono in considerazione non più di un individuo. Questi sistemi hanno come scopo la comprensione della gestualità dell'individuo, al fine di permettere la comunicazione. Per esempio, certi sistemi hanno come obbiettivo la lettura delle labbra o la comprensione dei linguaggio dei segni (per i sordomuti). Questi sistemi cominciano ad essere affidabili, anche se sono ancora in fase di messa a punto. Un'altra applicazione riguarda gli apparecchi automatici intelligenti (per esempio distributori automatici di biglietti, punti d'informazione automatica). Il sistema deve reagire in funzione del comportamento dell'utilizzatore, e comprendere se è soddisfatto. Il sistema è allora in grado di valutare il successo del servizio proposto. Un'applicazione simile riguarda i locali intelligenti. Un locale intelligente è una stanza munita di sensori capaci di rispondere alle aspettative di un utilizzatore senza che questi abbia bisogno di far uso di dispositivi particolari (per esempio una tastiera) (cf [27]). Per esempio, A. Bobick ha costruito una camera per i bambini, capace di raccontare loro una storia utilizzando effetti speciali e reagendo ai loro comportamenti. Allo stesso modo l'autore ha sviluppato un sistema di controllo per l'abitacolo di un veicolo, al fine di anticipare il comportamento del conducente. In [28] il sistema d'interpretazione controlla le videocamere di uno studio televisivo, obbedendo agli ordini del regista.In campo sportivo, altri sistemi d'interpretazione hanno come obiettivo l'insegnamento dei gesti atletici. Per esempio, in [29] gli autori presentano un sistema d'interpretazione capace di insegnare il Thai-chi. In [30] il sistema riconosce i passi di una ballerina classica. Altre applicazioni riguardano i videogiochi e la realtà virtuale. Per esempio A. Pentland utilizza una videocamera e uno schermo gigante che visualizza ciò che deve vedere l'utilizzatore. Il sistema fa allora evolvere il mondo rappresentato sullo schermo occupandosi anche delle interazioni con l'utilizzatore, basandosi sugli scenari di alcuni videogiochi, come per esempio "Doom". Infine altre applicazioni riguardano le riparazione specializzata di macchine e la realtà virtuale. Per esempio in [31] un operaio munito di occhiali ripara una stampante e ottiene la visualizzazione, sugli

occhiali, dei nomi delle differenti parti che sta smontando nonché della procedura di riparazione.

- **Analisi delle scene in robotica**: in questo tipo d'applicazione, l'obbiettivo è, per un robot mobile, riuscire a comprendere l'ambiente nel quale si sta spostando. Per esempio nel progetto SKIDS (cf [32]) un robot mobile munito di numerosi sensori è capace di analizzare delle sequenze d'immagini. Applicazioni più ambiziose consistono nel permettere la cooperazione di più robots. Per esempio, una coppa del mondo di calcio tra robot, "Robot Cup", è organizzata ogni anno tra differenti centri di ricerca e università.

Questa enumerazione non vuole essere esaustiva, ha piuttosto come scopo di fissare le idee sulle possibilità d'applicazione d'un sistema d'interpretazione. Una prima caratteristica di queste applicazioni è il trattamento in tempo reale dei dati (cioè il trattamento in continua dei dati, alla frequenza con la quale questi si rendono disponibili). Ciononostante, un buon numero di queste applicazioni possono essere effettuate con un trattamento differito delle sequenze d'immagini. Per esempio, nell'analisi delle scene sportive l'allenatore non ha bisogno di una analisi in diretta delle performances dei suoi giocatori. In questo tipo di applicazione una correzione manuale dell'analisi del comportamento è allora possibile, e permetterebbe ad un'azienda di commercializzare l'analisi di una sequenza d'immagini senza tuttavia che il sistema di cui fa uso sia totalmente autonomo.

La problematica dell'intepretazione in differita si ricongiunge a quella dell'indicizzazione tramite contenuto delle sequenze d'immagini. In effetti, nelle basi di dati che contengono grandi volumi di sequenze video, è spesso interessante ricercare automaticamente una sequenza, a partire da una descrizione delle attività che si svolgono nella sequenza stessa. Il problema si riconduce allora al poter disporre di un riconoscimento automatico di scenari sufficientemente preciso e informativo da poter discriminare la sequenza all'interno dell'intera base di dati. In videosorveglianza, per esempio, gli operatori conservano in generale un gran numero di sequenze. Nel momento in cui si ha l'esigenza di ritrovare una sequenza particolare, si è obbligati a passare in rassegna manualmente tutta la base di dati.

In tutte queste applicazioni si possono distinguere differenti gradi di realizzazione. In campo industriale i sistemi d'interpretazione hanno come obiettivo principale la robustezza. Questi sistemi si limitano allora spesso ad un rilevamento degli oggetti in movimento. Dal punto di vista accademico, i sistemi d'interpretazione hanno come obiettivo l'analisi dei comportamenti complessi. Nonostante siano utilizzabili principalmente su sequenze particolari, meno "difficili" (per esempio, girate ad hoc in laboratorio), questi sistemi permettono di delimitare lo spazio degli scenari possibili e, di conseguenza, di portare a termine anche la fase d'interpretazione delle differenti sequenze.

2.1.2 Problematiche sollevate

Da un punto di vista etico l'interpretazione automatica di sequenze video è un'attività nobile, nonostante la cattiva fama di cui gode da parte del grande pubblico, soprattutto quando declinata nelle sue applicazioni di videosorveglianza. Nella maggior parte dei casi, scopo primario della videosorveglianza in particolare (e comunque dell'interpretazione automatica) è affiancare l'operatore umano in compiti ripetitivi e noiosi di esame prolungato di decine di schermi diversi. Il tentativo è quello di permettere all'operatore umano di distogliere l'attenzione dalle immagini trasmesse e di affidare al sistema automatico il compito di richiamare questa attenzione, nel momento opportuno, sulla precisa sequenza che potrebbe richiedere un intervento oppure una validazione dell'interpretazione effettuata (per fare solo un esempio, un allarme potrebbe richiamare l'attenzione dell'operatore

su una presunta scena di vandalismo o di violenza). Sotto quest'ottica appare evidente quanto importante ed utile possa essere l'utilizzo di potenti sistemi di interpretazione per affiancare l'operatore umano, e quale miglioramento ne possa derivare per le condizioni sia di lavoro, che di efficienza dell'operatore stesso. Inoltre lo scopo primario della videosorveglianza non è la repressione, bensì piuttosto la sicurezza e la protezione degli individui contro le situazioni di pericolo o di violenza. Per esempio, determinati sistemi di videosorveglianza salvano già delle vite umane sulle autostrade allertando i soccorsi in caso di incidente e permettendo loro di portare rapidamente assistenza agli incidentati.

Nel nostro caso particolare, vista l'applicazione in chiave "mediaspace" del nostro sistema d'interpretazione, i problemi di tutela della privacy che si pongono sono tutt'altro che trascurabili, anzi giocano un ruolo fondamentale nel determinare l'accettazione del sistema stesso da parte dell'utente. È per questo che la gestione ottimale di opportune routines di protezione della privacy e dello spazio privato è diventata uno dei requisiti base del progetto del sistema.

Come per ogni strumento potente, l'utilizzo di un sistema di interpretazione automatica per scopi deviati può avere conseguenze estremamente negative. Ciononostante sarebbe un peccato decidere di fare a meno del fuoco per paura di potersi scottare.

2.2 Precedenti esperienze

In questa sezione presenteremo una rapida rassegna delle principali architetture di sistemi di elaborazione esistenti, in modo da poter riconoscere i punti comuni e le eventuali divergenze. Si vedrà che la catena di operazioni che vengono eseguite sul flusso dati (in origine la sequenza d'immagini) è approssimativamente comune a tutte queste realizzazioni.

Tutti i sistemi presentati in questa sezione sono considerati *completi*. Intendiamo con il termine *completo* un sistema capace di ricevere in ingresso delle sequenze di immagini reali (in opposizione alle immagini di sintesi) e che produce come risultato in uscita un'analisi dei comportamenti non elementari. Alla luce di questa definizione un gran numero di sistemi non si possono considerare completi, nella misura in cui questi sistemi pongono l'accento sulla determinazione del movimento e delle azioni elementari e non trattano il problema del riconoscimento degli scenari non elementari.

Nel seguito presentiamo cinque sistemi completi:

- Il progetto Vitra ha sviluppato il sistema d'interpretazione "Soccer" (cf [33], [34]). Questo progetto è stato condotto in Germania all'università di Karlsruhe e nella ditta Saarbrücken GmbH. "Soccer" è composto da una base contestuale che comprende l'ambiente statico e da tre moduli principali, associati a due moduli complementari. I tre moduli principali realizzano la catena completa d'interpretazione, dal trattamento delle immagini fino all'analisi delle attività. Il primo modulo, "Action", calcola la posizione e la velocità di ogni oggetto in movimento. Il secondo modulo, "Events recognition" produce un insieme di proposizioni corrispondenti alle relazioni spazio-temporali relative agli oggetti in movimento, quali *il giocatore A passa la palla al giocatore B*. Il terzo modulo "Selection and Generation" sceglie gli eventi pertinenti e genera delle frasi in tedesco che descrivono la scena. I due moduli complementari hanno come obiettivo il miglioramento della selezione degli eventi pertinenti. Il primo modulo, "Replay" (cf [35]) riconosce i piani e le strategie degli oggetti mobili e filtra gli eventi riconosciuti, che sono i dati d'ingresso del modulo di selezione e generazione. Il secondo modulo complementare, "Antlima" (cf [36]), rappresenta lo stato mentale di un ascoltatore virtuale e permette di selezionare gli avvenimenti più pertinenti per l'utilizzatore.

Il sistema "Soccer" risulta così scomponibile in due parti. La prima, costituita dal modulo "Action", rileva, riconosce e insegue[1] gli oggetti in movimento. La seconda parte, costituita dai restanti quattro moduli, riconosce le azioni e genera le frasi che descrivono la scena. La seconda parte risulta composta da un gran numero di moduli, dovuto ai numerosi collaboratori che parteciparono al progetto e al loro interesse primo, che era la generazione di frasi in linguaggio naturale.

- H. Nagel e la sua équipe hanno sviluppato in Germania (all'ITTB Karlsruhe e all'università di Karlsruhe) un sistema capace di riconoscere il comportamento di un veicolo in una stazione di servizio (cf [19]). Questo sistema è composto da due parti, una che si occupa della visione e l'altra dell'interpretazione. La parte di visione è costituita da un insieme di metodi sofisticati che permettono di rilevare e di seguire i veicoli a partire dalla loro traccia video. La parte d'interpretazione, più sommaria, analizza il comportamento dei veicoli, riconoscendo scenari quali *parcheggiare per fare rifornimento*. La disproporzione tra queste due componenti è dovuta primariamente alle competenze di questa équipe, che proviene da un laboratorio di visione e secondariamente dalla volontà di realizzare innanzitutto un sistema di visione performante e solo in secondo luogo di abbordare il problema dell'interpretazione.

- Il progetto europeo VIEWS è stato condotto da più di 17 collaboratori, dall'ITTB Fraunhofer (Atlas Elektronik GmbH), al GEC Hirst Research Center, al GEC Marconi Research Center, al Marconi Command and Control Systems, all'università di Reading, al Queen Mary and Westfield College e all'FTC (Framentec Cognitech). Questo progetto ha portato allo sviluppo di un sistema d'interpretazione costituito da due parti: la componente di percezione (visione) e la parte interpretazione (cf [17]). La parte di percezione è costituita da tre moduli principali. Il primo rileva le regioni in movimento e concentra l'attenzione del sistema sulle regione di interesse dell'immagine. Il secondo modulo segue le regioni in movimento interessanti rilevate e il terzo modulo identifica gli oggetti in movimento grazie ad una metodologia di classificazione. La parte percettiva segnala anche al modulo d'interpretazione i cambiamenti intervenuti nello stato della scena. Il modulo d'interpretazione è organizzato su tre livelli logici: eventi, comportamento e dinamica. Gli eventi sono dei cambiamenti considerati come interessanti (cioè riconosciuti come appartenenti ad una classe d'eventi interessanti per l'utilizzatore). I comportamenti sono delle sequenze interessanti d'avvenimenti. Le dinamiche sono degli insiemi interessanti di comportamenti, caratterizzati dall'intervento di molteplici oggetti mobili allo stesso tempo. Ogni livello possiede tre compiti principali: classificare una entità (per esempio un evento), verificare la sua coerenza con i risultati già ottenuti e predire l'entità successiva (per esempio il prossimo avvenimento). Il livello delle dinamiche non è stato portato a completamento. Questo sistema si distingue per il suo elevato numero di moduli e di funzionalità, numero che si spiega, in parte, con la quantità importante di partners di questo progetto.

- Il progetto europeo Esprit HPCN PASSWORDS è stato portato avanti da più di 12 collaboratori, su una finestra di 3 anni, all'università di Genova (DIBE), al Research Center CRIF, all'INRIA Sophia Antipolis e al Sepa (Centro di Ricerche Fiat). Questo progetto ha sviluppato un sistema d'interpretazione composto da tre moduli (cf [37]). Il primo modulo rileva le regioni in movimento, il secondo modulo segue le regioni rilevate e il terzo modulo identifica gli oggetti in movimento e analizza il

[1] Con il termine *inseguire un oggetto* si intende l'azione espressa dal termine inglese "to track".

loro comportamento. Ogni modulo è stato sviluppato da équipe partner differenti. Il sistema risulta quindi dalla disposizione in cascata dei tre moduli: visione a basso livello, visione intermedia ed interpretazione.

- Il progetto Perception è stato condotto da più di 12 collaboratori in Francia al CERT DERA, al CERT DERI, all'ONERA DES-SIA e all'ONERA DES-STD. Questo progetto ha sviluppato un sistema che comprende una base del contesto, due moduli principali e un modulo complementare (cf [22]). Il primo modulo principale "Trattamento numerico" è il modulo di visione; rileva, riconosce e segue gli oggetti in movimento. Il secondo modulo "Trattamento simbolico" è il modulo d'interpretazione; riconosce gli scenari relativi alle attività degli oggetti mobili. Il modulo complementare "Gestione della percezione" gestisce le risorse del sistema, che comprendono i sensori e il trattamento percettivo. Questo sistema è così globalmente costituito da una parte di visione e da una di interpretazione.

Questi sistemi sono stati tutti sviluppati da équipes numerose, in seno a progetti di ricerca di lungo periodo (da 3 a 10 anni). Si basano tutti sulla stessa successione di operazioni: rilevamento delle regioni in movimento, identificazione degli oggetti mobili e inseguimento di quelli interessanti e infine analisi del loro comportamento. In più, questi sistemi si caratterizzano, in generale, per una separazione marcata in due parti, che raggruppano i moduli di visione e i moduli di interpretazione, senza che ci sia nessuna reale cooperazione tra le due parti. I moduli di visione raggruppano le routines di rilevamento e inseguimento delle regioni in movimento. I moduli d'interpretazione contengono le routines di riconoscimento degli scenari elementari e complessi. Se la parte visione è sufficientemente performante, essa contiene anche le routines d'identificazione degli oggetti in movimento. In caso contrario questa routine appartiene alla parte interpretativa. In questo secondo caso tutti gli oggetti sono inseguiti e solo successivamente si indaga sulla natura di ognuno di essi. La principale differenza tra questi sistemi d'interpretazione proviene dalla proporzione tra i moduli di visione e quelli d'interpretazione. Questa differenza di proporzione (che si riscontrerà anche nel presente lavoro di tesi) è dovuta essenzialmente alla provenienza delle équipe che hanno concepito i sistemi (più orientate verso la visione) è dall'altro lato agli obiettivi che ogni équipe si è posta.

2.3 Lo stato dell'arte

In quest'ultima sezione presenteremo rapidamente lo stato dell'arte riguardo ai tre grandi blocchi funzionali che sono stati progettati ex-novo per questo progetto (e che costituiscono dunque il vero nucleo del lavoro di tesi). Essi sono:

- il modulo di riconoscimento della pelle umana;
- il modulo di classificazione;
- il modulo di inseguimento degli oggetti mobili.

2.3.1 Il riconoscimento della pelle umana

Il riconoscimento della pelle che è stato implementato si basa sull'analisi cromatica delle immagini. Negli ultimi anni, un numero sempre crescente di ricerche sono state indirizzate al problema specifico del riconoscimento dei volti umani basato sul colore della pelle (cf [38], [39],[40],[41],[42]). Fondamentalmente le ricerche si sono indirizzate al riconoscimento

di pixels di pelle umana in immagini a colori e alla differenziazione tra i pixels di pelle e i pixels non di pelle utilizzando vari modelli statistici del colore: in [43] e [44] modelli cromatici ad hoc vengono utilizzati come un primo filtro per analizzare vasti database di immagini; altri ricercatori hanno usato modelli cromatici del color pelle come il modello Gaussiano semplice ([42] e [39]), il modello "Gaussian mixture density" ([45] e [46]) o il modello ad istogrammi ([47] e [48]). Recentemente M.J.Jones e J.M.Rehg hanno implementato un'analisi completa e dettagliata dei modelli cromatici sia del color pelle che di altri colori, utilizzando un vasto database di pixels sia di pelle che non di pelle costruito manualmente a partire dalle immagini disponibili nel web. Lo studio compara le performances dei modelli ad istogrammi e dei modelli "Gaussian mixture density" nello spazio cromatico classico RGB a 24 bits. Le conclusioni a cui i due studiosi sono giunti è la marcata superiorità dei modelli a base di istogrammi, per lo spazio cromatico scelto, nell'individuare i pixels di pelle (cf [49]). Nello stesso articolo la distribuzione dei pixels sia color pelle che non di color pelle può ancora essere modelizzata nello spazio 3D RGB non normalizzato, malgrado gli effetti di variazione di luminosità, a priori sconosciuti, grazie alle notevoli dimensioni del database di pixel di riferimento utilizzati. Nella maggior parte degli esperimenti i pixels di pelle di riferimento provengono da un gruppo limitato di persone; ugualmente limitate sono le differenti condizioni di illuminazione utilizzate. Una discreta robustezza rispetto ai cambiamenti di illuminazione si può ottenere se lo spazio colorimetrico utilizzato separa efficacemente le componenti di crominanza e di luminosità nelle immagini. Ciò richiede una riduzione dimensionale dello spazio cromatico, ottenibile tramite una opportuna (sia lineare che non lineare) trasformazione dalle variabili 3D RGB ad uno spazio bidimensionale di crominanza (associato ad un terzo indice di luminosità). Lo spazio (r, g) normalizzato è stato spesso utilizzato per il rilevamento dei volti (cf [38],[40],[47], [50]) soprattutto perché esso consente una riduzione della sensibilità rispetto ai cambiamenti d'illuminazione. Altri spazi di crominanza-luminosità studiati sono lo spazio HSV (o HSI) (cf [51], [50], [52]). La scelta di un opportuno spazio cromatico è un punto essenziale, in quanto la forma delle distribuzioni del color pelle e dei colori non di pelle dipende dallo spazio di crominanza scelto.

2.3.1.1 Caratteristiche del modulo progettato

Alla luce dello stato dell'arte in materia di riconoscimento di pelle e dei vincoli del presente progetto, si sono fissate le caratteristiche di massima del modulo di riconoscimento della pelle implementato:

- l'algoritmo utilizzato è a base statistica e il modello cromatico del color pelle è codificato tramite istogrammi. Il motivo delle due scelte è la grande diffusione in letteratura di questo tipo di tecniche (e la conseguente grande disponibilità di progetti di riferimento e di documenti in proposito). Inoltre la tecnica a base di istogrammi è risultata nettamente migliore di altre, a parità di condizioni ([49]). Infine dal punto di vista algoritmico le difficoltà offerte sono modeste.

- Lo spazio cromatico sarà opportunamente scelto sulla base di test e prove valutative. Le possibilità considerate saranno tuttavia orientate verso gli spazi RGB e derivati, per ragioni legate esclusivamente al tempo di calcolo aggiuntivo che sarebbe necessario per trasformare immagini codificate nello spazio RGB ad altri spazi (per esempio HSV).

- Si porrà particolare attenzione a progettare un algoritmo robusto dal punto di vista della sensibilità alle condizioni d'illuminazione. Infatti l'applicazione per la quale

si progetta il sistema d'interpretazione è tale da dover lavorare nelle più svariate condizioni di luminosità.

- A causa del tempo di sviluppo dedicato a questo modulo, piuttosto ridotto, si limiterà il più possibile la complessità del modello del color pelle utilizzato. Questo potrebbe voler dire limitare le tipologie di pelle prese in considerazione.

2.3.2 La classificazione delle regioni in movimento

Il modulo di classificazione svolge il compito di associare ad ogni regione in movimento individuata nell'immagine un'etichetta che ne definisca, nel modo più preciso e rigoroso possibile, la natura. In altri termini, si vuole rendere disponibile ai moduli successivi (ed essenzialmente al modulo di inseguimento degli oggetti mobili) l'informazione sulla natura dell'oggetto in movimento che stanno trattando: *individuo* piuttosto che *parte di un individuo* (in caso di occultazione) o *veicolo*, *porta*, *rumore*... Si tratta dunque di raffinare, o stabilire, la decisione sull'interesse delle particolari regioni in movimento rilevate nell'immagine. Questa fase pone il problema di stabilire su quali regioni in movimento si deve concentrare l'attenzione, o, altrimenti detto, quali sono le regioni in movimento utili per la comprensione di ciò che accade nella scena e quali meno. Da qui la difficoltà: suddividere in categorie precise le zone in movimento. L'analisi del problema della classificazione può essere trattato partendo dal punto di vista della natura degli oggetti rilevati.

Gli oggetti da classificare sono spesso caratterizzati dal loro potenziale di deformazione intrinseco, o altrimenti detto, dalla loro rigidità. Gli *oggetti rigidi* sono gli oggetti la cui forma non varia al variare del tempo, solo la proiezione nel piano dell'immagine può cambiare. Dunque un oggetto rigido conserva sempre la propria forma; eventuali differenze sono sempre giustificabili alla luce di rotazioni, traslazioni o omotetie. Di conseguenza i metodi di classificazione che lavorano sugli oggetti rigidi spesso cercano di stimare i parametri di tali trasformazioni, al fine di istanziare correttamente il modello dell'oggetto rigido nell'immagine corrente. La seconda classe di oggetti da classificare sono i cosiddetti *oggetti non rigidi*, la cui forma non può essere considerata costante né derivata da una forma precedente tramite trasformazione geometrica omotetica. In ogni caso, il processo di classificazione si basa quasi sempre sul tentativo di mettere in corrispondenza uno dei modelli di cui si dispone (e che corrispondono ognuno ad una possibilità di classificazione, cioé ad una classe di oggetti in movimento) con la zona in movimento da classificare. La corrispondenza migliore determina la classificazione utilizzata.

In base a che cosa si descriverà tramite il modello, le tecniche di classificazione si distinguono in base all'utilizzo di *modelli di come può apparire l'oggetto nell'immagine* piuttosto che di *modelli dell'oggetto reale*. Esaminiamo ognuna delle due famiglie e le metodologie di classificazione relative studiate fino ad oggi.

2.3.2.1 Modelli di come può apparire l'oggetto

Questa famiglia di metodologie, basata sull'aspetto dell'oggetto nell'immagine, è spesso utilizzata qualora sia difficile qualificare l'oggetto per dei motivi di rumore particolarmente pronunciato o quando gli oggetti sono troppo complicati o troppo diversi. In questa famiglia l'oggetto della modellizzazione è dunque *l'immagine* dell'oggetto.

In [53] il modulo di segmentazione e rilevazione delle regioni in movimento fornisce un risultato caratterizzato da un rumore molto forte, che non permette di istanziare con precisione un modello complesso. L'elaborazione resta quindi confinata al piano dell'immagine (anziché al volume della scena). In [24] gli autori utilizzano ancora un modello dell'aspetto

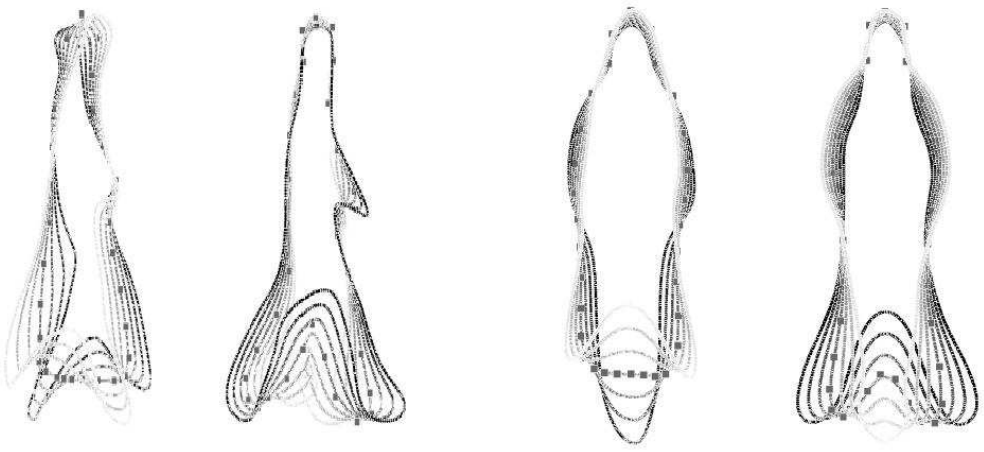


Figura 2.1: *Il modello d'aspetto per un oggetto non rigido secondo Hogg.*

dell'oggetto. Il campo d'applicazione è l'analisi di scene sportive (calcio). Invece che un modello d'aspetto preciso, gli autori scelgono una decisione basata sull'analisi dei colori. In [54] gli autori giustificano il loro approccio sulla base dell'impossibilità di definire un modello d'oggetto per il particolare soggetto in esame. Baumberg e Hogg ([55]) definiscono un modello d'aspetto per oggetti non rigidi complessi al fine di analizzare il movimento umano. Gli autori definiscono un modello 2D di forma grazie all'estrazione di un vettore di punti del contorno della proiezione dell'individuo sul piano dell'immagine, basandosi a tal fine su un modello di distribuzione chiamato PDM ("Point Distribution Model"), derivato dagli studi di Cootes ([56]). L'idea è di misurare la differenza dx tra un modello qualunque e il modello medio, per esempio $dx = x - x_{medio}$, dove x è un vettore di punti ottenuti e x_{medio} è il vettore medio. L'analisi del vettore dx restituisce il modo di variazione più significativo del modello. La forma del contorno è quindi ottimizzata utilizzando delle b-splines cubiche i cui punti di definizione sono rappresentati dai punti del contorno definiti precedentemente. I risultati sono valutati alla luce della cifra di merito "compattezza del modello" (cf la figura 2.1).

Il vantaggio dell'uso di modelli d'oggetto basati sull'aspetto è di poterli istanziare molto presto nella catena d'elaborazione, subito dopo la fase di rilevamento. Il difetto principale è che così facendo non si tengono in conto né l'aspetto volumetrico né le caratteristiche intrinseche dell'oggetto stesso.

2.3.2.2 Modelli d'oggetti reali

Questi ultimi sono definiti tramite le caratteristiche dell'oggetto stesso. In questa famiglia, cioé, ciò che è modellizzato è *l'oggetto*. I metodi che ne fanno uso ricorrono molto spesso alle proprietà naturali degli oggetti. Per esempio oggetti rigidi come i veicoli possono essere descritti sulla base di semplici considerazioni volumetriche (la disposizione spaziale dei differenti "volumi" che caratterizzano il veicolo stesso). In [57] l'oggetto consiste in un poliedro 3D che modellizza un veicolo tramite dodici gradi di libertà. L'insieme dei segmenti del contorno ottenuti nella fase di rilevamento viene allora messo in relazione con gli spigoli del modello poliedrico grazie ad un processo iterativo mutuato dalle ricerche di Lowe (cf [58]). L'idea è di minimizzare una cifra di merito costituita da una particolare distanza (distanza di Mahalanobis) per ogni spigolo del poliedro.

La maggior parte degli oggetti naturali sono di tipo non rigido: vegetali, animali e esseri umani. Il trattamento di tali oggetti richiede l'uso di modelli e tecniche particolari. Inoltre il problema legato alla classificazione delle persone è doppio: da un lato, per una determinata persona, le caratteristiche che la descrivono non sono costanti al variare del tempo. Ciò perché esiste una forte variazione delle forme durante lo spostamento (oscillazione delle braccia e delle gambe), unita ad una variazione del colore. In secondo luogo le stesse caratteristiche morfologiche variano da persona a persona, il che comporta dei problemi a livello di generalità del modello utilizzato.

In [59] l'autore utilizza un modello di persona la cui struttura è composta da 6 segmenti (due gambe, due braccia, il torso e la testa). In [60] il numero di segmenti si eleva a 17. Un modello volumetrico di quattordici cilindri è invece utilizzato in [61]. Il vantaggio nell'uso di modelli degli oggetti reali è l'indipendenza che ne deriva tra gli algoritmi di classificazione e i vincoli del modello. L'inconveniente è che si allontana il momento della classificazione dell'oggetto dalla fase di rilevamento delle regioni in movimento.

2.3.2.3 Caratteristiche del modulo di classificazione progettato

Alla luce dello stato dell'arte in termini di classificazione e delle esigenze del presente progetto, si sono stabilite le caratteristiche seguenti per il modulo di classificazione:

- l'algoritmo di classificazione è a base di modelli, nella fattispecie di modelli dei differenti oggetti in movimento che possono generare regioni in movimento nell'immagine. Più precisamente, il modello di oggetto in movimento descrive l'oggetto stesso essenzialmente in termini di dimensioni e forma, che saranno confrontati con quelli evinti dalla regione in movimento corrispondente. Eventuali altre caratteristiche possono completare il modello.

- L'algoritmo di classificazione non è specializzato per l'applicazione "MediaSpace" ma, facendo uso di basi di contesto opportune, potrà essere utilizzato anche in altre applicazioni (questo vincolo è dettato soprattutto dall'esigenza di riutilizzabilità del modulo in altre applicazioni similari).

- L'algoritmo di classificazione gestisce le occultazioni degli individui a causa degli ogetti del contesto. A tal fine opportune informazioni sono disponibili nella base del contesto (cf anche il paragrafo 3.1.2).

- Infine, l'algoritmo implementa anche una fase di fusione tra le regioni, nella quale si cerca di ricostituire, da sottoregioni eventualmente presenti, la regione logica più opportuna.

2.3.3 L'inseguimento degli oggetti in movimento

I problemi principali dell'inseguimento degli oggetti in movimento sono da un lato la ricostruzione e la stima del movimento dell'oggetto seguito, che permette di predire la sua nuova posizione e dall'altro lato la gestione delle cosiddette *associazioni ambigue* tra gli oggetti inseguiti fino all'istante $t-1$ e quelli (nuovi) presenti all'istante t. Il valore di qualsiasi algoritmo di inseguimento di oggetti, la sua robustezza, è fortemente legata all'efficacia con la quale esso tratta le associazioni ambigue.

L'inseguimento degli oggetti è un passo chiave del processo globale d'interpretazione delle sequenze d'immagini, nella misura in cui la perdita della continuità spazio-temporale di un oggetto inseguito (cioè quando il programma, improvvisamente, commette un errore

nell'inseguire un oggetto in movimento) è capace di mettere in scacco tutta la piattaforma interpretativa.

In letteratura troviamo tre grandi famiglie di metodi di inseguimento degli oggetti, che si differenziano in base al tipo di oggetto inseguito.

2.3.3.1 Inseguimento di oggetti rigidi

Numerosi studi hanno esaminato il problema di inseguire un oggetto rigido (veicoli, robot: oggetti la cui forma non cambia se non per trasformazione omotetica), permettendo la messa a punto di metodologie robuste ed affidabili. Principalmente queste metodologie sono tre:

- Un primo metodo consiste del rilevare delle primitive particolari dell'oggetto in movimento (per esempio spigoli e angoli) e a seguire queste primitive d'immagine in immagine ([62]). Questa metodologia si applica tuttavia solamente su oggetti molto particolari, caratterizzati da numerose primitive che siano facili da rilevare e che si possano organizzare a formare un modello dell'oggetto.

- Un secondo metodo consiste nel mettere in corrispondenza la regione in movimento 2D rilevata con un modello geometrico 3D dell'oggetto mobile. Per questo si stima il centro e la direzione del movimento della regione e poi si mettono in corrispondenza, per esempio, i segmenti di destra della regione con quelli del modello. Questa corrispondenza è spesso delicata a causa del rumore che affligge le zone in movimento rilevate. In [63] gli autori calcolano un grado di coerenza della corrispondenza effettuata, valutando le distanze, le differenze d'orientamento e di lunghezza dei segmenti giustapposti. In [57] gli autori utilizzano un modello geometrico a parametri che può essere adattato al modello. Utilizzano inoltre un modello d'illuminazione per calcolare l'ombra al suolo degli oggetti in movimento (spesso rilevata come facente parte della regione in movimento).

- Il terzo metodo consiste nell'inseguire i contorni delle regioni in movimento grazie ad un modello deformabile dei contorni. Per esempio, in [64] la regione in movimento viene approssimata grazie ad un poligono convesso e si inseguono i punti di questo poligono. A tal fine si cerca una corrispondenza globale tra il poligono precedente e quello che è appena stato rilevato, calcolando la distanza tra i due. L'interesse di questo metodo è che esso prende in considerazione le occultazioni parziali delle regioni mobili. La perdita di una parte del poligono è compensata dalla presenza dell'altra parte. In [65] gli autori procedono allo stesso modo, ma utilizzando delle curve b-splines cubiche al posto dei poligoni. In [66] gli autori propongono un trattamento specifico delle situazioni di occultazione parziale. Questo metodo d'inseguimento dei contorni possiede principalmente due inconvenienti: l'inizializzazione del primo contorno e la gestione delle occultazioni complete. Ciononostante questo metodo può essere applicato anche per gli oggetti non rigidi.

L'inseguimento di un oggetto mobile rigido si caratterizza anche per l'uso di un filtro di Kalman per stimare la nuova posizione dell'oggetto mobile ([20], [63]). Il filtro di Kalman fornisce la stima dello stato di un insieme di misure tenendo conto del rumore ([67]). Il filtraggio consiste in due operazioni: (1) aggiornamento del filtro (predizione dello stato all'istante successivo e calcolo dell'errore di predizione commesso); (2) correzione del filtro tenendo conto delle nuove misure di cui è stata pesata (cioé filtrata) l'influenza utilizzando il modello del rumore. L'utilizzo del filtro di Kalman necessita dunque della

definizione di un modello di movimento. In [57] gli autori utilizzano un movimento a tre gradi di libertà del centro di gravità dell'oggetto, accoppiato ad un movimento di rotazione angolare e di scivolamento. L'efficacia del filtro dipende essenzialmente dalla corrispondenza tra i modelli utilizzati e la realtà. Il problema di modellizzazione è delicato, visto che i movimenti degli oggetti mobili reali seguono raramente delle leggi facilmente modellizzabili.

2.3.3.2 Inseguimento di oggetti non rigidi

Gli oggetti mobili non rigidi (per esempio gli esseri umani) non possiedono un modello geometrico preciso della loro forma. Quindi, contrariamente al caso dell'inseguimento degli oggetti rigidi, non si possono utilizzare degli algoritmi d'inseguimento basati su tali modelli. Questi metodi utilizzano invece dei modelli dinamici o temporanei degli oggetti mobili. In letteratura si ritrovano essenzialmente tre metodi di inseguimento di oggetti in movimento non rigidi:

- Il primo metodo utilizza dei modelli deformabili del contorno, come nel caso dell'inseguimento di oggetti rigidi. La differenza tra i due metodi risiede nell'uso di modelli di deformazione preferenziale. Per esempio in [55] vengono utilizzate delle curve cubiche b-splines per rappresentare i modelli del contorno. A seconda del modello del contorno, si utilizza un insieme di vettori di deformazione che permettono di deformare, in una certa direzione, la porzione di curva compresa tra due punti di definizione della b-spline. In questo modo si può inseguire un uomo che sta camminando e riconoscere la sua modalità di spostamento (per esempio egli cammina) e la direzione del suo movimento. L'interesse di questo metodo risiede nella capacità di modificare in itinere i modelli del contorno e le deformazioni associate. I punti deboli sono la dipendenza del modello rispetto all'angolo con cui la videocamera filma la scena (per esempio, è necessario definire due diversi modelli per un uomo che cammina nella direzione della videocamera o spostandosi ortogonalmente) e la difficoltà di gestione delle situazioni di occultazione.

- Il secondo metodo utilizza un modello temporaneo delle regioni in movimento che rappresentano l'oggetto sull'immagine ([54]). Questo modello temporaneo è definito dall'intensità dall'insieme dei pixels che appartengono alla regione in movimento. Ad ogni nuova immagine questo modello temporaneo viene comparato con le nuove intensità della regione in movimento rilevata. Una volta che si è stabilita la corrispondenza con una regione in movimento, si aggiorna il modello temporaneo utilizzando i valori d'intensità di questa nuova regione. Esistono diverse varianti di quest'algoritmo. Per esempio, in [24] gli autori utilizzano anche un istogramma dei colori per trattare i problemi di occultazione dinamica tra oggetti in movimento di colori diversi. Questo metodo è utilizzato in particolare per seguire i giocatori di calcio. In [25] gli autori includono nel modello temporaneo degli oggetti in movimento anche una piccola striscia del contorno della regione stessa (quindi una piccola parte dell'immagine di ciò che circonda l'oggetto mobile), per evitare di perdere una parte dell'oggetto da seguire. Essi utilizzano inoltre un modello del fondo della scena che permette loro di eliminare gli oggetti statici dell'ambiente che potrebbero mischiarsi agli oggetti mobili. In questo modo si ottiene un inseguimento più affidabile. In [27] e [68] gli autori associano ad ogni punto del modello temporaneo la distribuzione dell'intensità del colore (nello spazio cromatico YUV) e la distribuzione spaziale nel piano dell'immagine secondo le coordinate (x, y). Inoltre modellizzano ogni punto

del fondo della scena. Queste distribuzioni sono utilizzate per predire le evoluzioni dei modelli (in particolare il movimento degli oggetti) e per compensare le variazioni d'illuminazione e la presenza d'ombre. Modellizzando le differenti regioni di colore di un individuo essi arrivano ad inseguire precisamente i suoi movimenti (per esempio il movimento delle mani).

- Il terzo metodo è stato proposto dall'équipe di P.Huttenlocker (cf [69]). Si utilizza ancora un modello temporaneo delle regioni in movimento che rappresenta l'oggetto, ma questo modello conserva gli spigoli delle regioni in movimento piuttosto che l'intensità di ogni pixel. Inoltre questo metodo cerca innanzitutto di aggiustare il modello temporaneo dell'oggetto in movimento adattandolo alla regione mobile appena rilevata, quindi calcola la distanza tra i due insiemi di spigoli autorizzando una corrispondenza parziale tra i due insiemi. La percentuale autorizzata di corrispondenza è uno dei parametri di questo metodo. Questa tolleranza permette di trattare i casi d'occultazione parziale degli oggetti in movimento. Il metodo si caratterizza anche per l'uso di un modello originale di movimento degli oggetti. A tal fine viene calcolata la classe delle trasformazioni (essenzialmente traslazioni per motivazioni legate al tempo di calcolo) che permettono di minimizzare la distanza tra il modello temporaneo di oggetto in movimento e le regioni della nuova immagine. Si determina così la regione in movimento che corrisponde all'oggetto e il movimento stesso dell'oggetto recuperando la classe delle trasformazioni applicate. Questo metodo è utilizzabile quale che sia l'ampiezza dello spostamento dell'oggetto in movimento. La sola restrizione è relativa al cambiamento del modello temporaneo, che non deve essere eccessiva.

La maggior parte di questi metodi si caratterizzano per l'uso di un filtro di Kalman come modello del movimento, per predire la nuova posizione dell'oggetto nell'immagine successiva. Si caratterizzano anche per i forti vincoli restrittivi imposti al fine di permettere la creazione di inseguimenti affidabili ed efficaci (un esempio di vincolo è che la videocamera sia in posizione frontale rispetto all'oggetto che si sta spostando, oppure che si faccia uso di più videocamere, o ancora che gli oggetti in movimento spicchino bene rispetto allo sfondo della scena). Questi vincoli si giustificano in parte con l'intrinseca difficoltà del problema dell'inseguimento degli oggetti ed inoltre con la specificità delle applicazioni per le quali i vari sistemi sono stati declinati.

Dei metodi passati in rassegna, solo il metodo proposto da P.Huttenlocker integra, tra i suoi principi, un trattamento delle occultazioni parziali (il caso delle occultazioni dinamiche può essere trattato sotto certe condizioni aggiuntive). Ciononostante si può criticare a questo metodo a base di modelli temporanei di spigoli la non utilizzazione di nessuna cronologia del modello del movimento. In effetti esso autorizza ogni tipo di spostamento dell'oggetto in movimento inseguito, ammesso che il suo modello temporaneo di forma non ne risulti radicalmente modificato. Questo vincolo impedisce di trattare efficacemente i problemi di occultazione totale. Inoltre il metodo necessita immagini ad elevata risoluzione.

2.3.3.3 Inseguimento di oggetti senza modello

I metodi per inseguire oggetti senza l'uso di un modello consistono nell'inseguire degli oggetti in movimento dei quali non si conoscono che le coordinate posizionali. I metodi precedenti d'inseguimento utilizzano un modello degli oggetti da inseguire al fine di evitare associazioni ambigue. Di conseguenza i metodi d'inseguimento che non si basano

su di un modello d'oggetto si caratterizzano per un'analisi sofisticata delle associazioni ambigue. Questi metodi non utilizzano proprietà legate al trattamento delle immagini e all'inseguimento di primitive (per esempio spigoli ed angoli). Essi sono particolarmente utilizzati nelle applicazioni radar. Possono anche essere sfruttati per inseguire degli oggetti che hanno un modello nel caso il rilevamento sia stato di cattiva qualità (in generale nelle applicazioni di videosorveglianza). Questi metodi d'inseguimento si applicano allo stesso modo quando si fa uso di molteplici sensori e alcuni di questi non sono in grado di rilevare altro che la posizione degli oggetti in movimento.

In letteratura si trovano essenzialmente quattro metodi d'inseguimento non a base di modelli:

- Il metodo del filtro d'associazione dati a probabilità congiunta ("Joint Probabilistic Data Association Filter" (JPDAF) (cf [67]) tratta efficacemente il problema dell'associazione dei dati grazie ai filtri di Kalman. Quando un oggetto "bersaglio"[2] corrisponde a più punti rilevati nella nuova immagine, il filtro associato combina l'influenza dell'insieme dei punti corrispondenti, utilizzando delle probabilità condizionate. Così facendo si considerano contemporaneamente le informazioni relative a più punti. In questo caso si utilizza un filtro dello stesso tipo di quello associato a un oggetto bersaglio che sia in corrispondenza con un solo punto nella nuova immagine. Questo problema di coerenza è infatti il punto debole di quest'ultimo metodo (cf [70]).

- Il metodo delle ipotesi di inseguimento multiplo, "Multiple Hypothesis Tracking" (MHT) tratta ancora i casi di associazione ambigua. Questo metodo conserva tutte le informazioni relative alle associazioni tra i bersagli e i nuovi punti misurati, quindi attende di ricevere le nuove informazioni per eliminare le associazioni incoerenti. Queste associazioni sono definite come delle ipotesi e corrispondono ognuna ad un filtro di Kalman distinto. Questo metodo mantiene così in parallelo più mondi concorrenti (insiemi d'ipotesi incompatibili). Per evitare un'esplosione combinatoria, questo metodo da un lato non conserva che le ipotesi d'associazione più coerenti e dall'altro lato fissa una soglia di "memoria" da conservare (generalmente estesa alle ultime 2 o 3 immagini) delle informazioni utilizzate per il calcolo delle associazioni. La critica principale mossa a questo metodo è la complessità di realizzazione. Ciononostante, in [71] si propone un'implementazione efficace dell'algoritmo MHT.

- Il metodo di strategia di ricerca per fasci, ("Beam Search Strategy") (cf [70]) duplica fisicamente tutti i bersagli (e i filtri di Kalman associati) corrispondenti a delle associazioni ambigue. Esso continua quindi l'inseguimento dei bersagli per entrambe le copie, fino a quando la loro pista (la traiettoria) diviene incoerente o termina in modo inusuale. Per stabilire la coerenza di una pista questo metodo utilizza essenzialmente la stima dell'errore associato al filtro e la durata d'esistenza del bersaglio. Esso esamina così molteplici ipotesi senza mantenere dei legami di concorrenza tra di esse. Ciò permette di ottenere delle implementazioni semplici del metodo, ma può porre dei problemi di coerenza. Per esempio, se si desidera studiare il comportamento di un bersaglio, è necessario poter discriminare un bersaglio originato da una duplicazione da uno realmente inseguito.

- Il metodo del filtraggio alla Kalman distribuito (cf [72]) ("Decentralized Kalman Filter", DKF) permette di combinare più filtri di Kalman per aumentare la robustezza

[2] Con il termine "bersaglio" si indicano i nuovi oggetti in movimento da mettere in relazione con i precedenti.

dell'inseguimento. Ogni filtro è associato in modo indipendente a un sensore e a un modulo d'inseguimento e realizza le proprie predizioni delle nuove posizioni degli oggetti mobili. In seguito, una seconda tappa combina queste predizioni per ottenere un inseguimento globale degli oggetti in movimento. Le ambiguità sono così risolte grazie all'utilizzo di più sorgenti d'informazione. L'interesse di questo metodo risiede nella sua robustezza. Per esempio, una implementazione efficace di questo metodo è stata proposta nell'ambito del progetto europeo Esprit SKIDS. Il sistema aveva come scopo l'inseguimento di più individui che si incrociavano con dei robot in una stanza. Quattro videocamere erano disposte ai quattro angoli della stanza e delle barriere ottiche erano piazzate nei luoghi più delicati, in particolare in prossimità delle zone di ingresso/uscita. L'insieme di questo sistema ha permesso di inseguire efficacemente degli individui su lunghe sequenze d'immagini. Allo stesso modo, in [20] l'autore propone un metodo multi-sensore simile, sviluppato in seno al progetto Eureka Prometheus, per l'equipaggiamento di un veicolo da strada. Questo metodo differisce dal precedente principalmente per il tipo di filtro di Kalman utilizzato. Esso utilizza una modellizzazione precisa del movimento dei veicoli, dato sì che questo movimento è più facilmente modellizzabile di quello degli individui. Visto che questo modello è non lineare, il metodo utilizza dei filtri di Kalman estesi (più sofisticati e che possono effettuare un filtraggio non lineare).

I metodi di inseguimento di oggetti mobili senza modello sono molto generici, perché pongono pochi vincoli nel loro utilizzo. Essi possono così essere utilizzati al fine d'inseguire degli oggetti mobili non rigidi. Ciononostante, dato che essi non possiedono informazioni sugli oggetti da seguire, essendo privi di modello, possono portare a numerose situazioni ambigue e anche alla perdita del bersaglio, quando queste situazioni diventano troppo difficili da gestire.

2.3.3.4 Caratteristiche del modulo d'inseguimento progettato

Come già effettuato precedentemente per i moduli di riconoscimento della pelle e di classificazione, alla luce delle tecniche sperimentate dalla ricerca e delle esigenze del presente progetto, elenchiamo le caratteristiche di massima che risultano per il modulo di inseguimento:

- Nonostante i bersagli con i quali ci si confronta siano tipicamente persone, si è deciso di far uso di un algoritmo a base di modelli. Questo perché la fase di classificazione stessa fa già uso di numerosi modelli per classificare gli oggetti in movimento, e la cosa è possibile per i motivi esposti nel paragrafo 2.3.2.3. L'algoritmo di inseguimento si serve di un modello di individuo, per validare, in un certo senso, le scelte fatte dal modulo di classificazione; inoltre si serve di un modello di traiettoria al fine di valutare la bontà dell'associazione "oggetto in movimento precedente - nuovo bersaglio".

- L'algoritmo non prevede nessun filtro di Kalman che stimi la nuova posizione dell'oggetto in movimento. Questo perché per la natura dello spazio nel quale si effettua l'interpretazione (un ufficio, che si presta molto bene a cambiamenti repentini di direzione nonché a frequenti occultamenti parziali o totali), un filtraggio alla Kalman si rivelerebbe delicato ad implementarsi e, nella maggior parte dei casi, fornirebbe predizioni poco corrette.

- L'algoritmo prevede una sofisticata gestione delle occultazioni parziali/totali, all'ordine del giorno nel nostro tipo d'applicazione.

Gli ulteriori dettagli tecnici via via stabiliti sono descritti nel capitolo 6.

Nei capitoli seguenti si descriveranno nel dettaglio i moduli che costituiscono la piattaforma di interpretazione implementata, soffermandosi particolarmente su quelli ideati e realizzati nell'ambito di questo lavoro di tesi.

Capitolo 3

La definizione del contesto e l'architettura generale

Questo capitolo è suddiviso in due parti fondamentali: una prima sezione riguarda la definizione di contesto e la descrizione del contesto utilizzato per l'applicazione all'oggetto della presente trattazione. Una seconda sezione descriverà poi l'architettura generale del sistema, in termini di moduli funzionali e di interfacce di connessione. Le funzionalità svolte da ognuno dei tre moduli principali (classificazione, riconoscimento di pelle e inseguimento degli oggetti in movimento) saranno oggetto dei capitoli successivi

3.1 Il contesto

La base del contesto comprende l'insieme delle informazioni contestuali relative all'ambiente della scena e utilizzate dal processo d'interpretazione delle sequenze video. Il contesto costituisce un insieme essenziale d'informazioni per il sistema d'interpretazione. In più, queste informazioni sono spesso indispensabili per poter risolvere certe situazioni particolarmente delicate. Essendo il contesto funzione dell'applicazione, i problemi principali connessi al suo utilizzo sono innanzitutto la definizione precisa della nozione di contesto e secondariamente di rappresentare queste informazioni al fine di facilitare la loro acquisizione e il loro utilizzo.

Da un punto di vista generale, la comunità scientifica riconosce l'importanza del contesto e la sua influenza sulla qualità dei risultati del processo d'interpretazione. Per contro, la maniera nella quale si può far uso delle informazioni contestuali resta un problema, dovuto in particolare alla difficoltà di formalizzare questa nozione.

3.1.1 Definizione di contesto

La definizione del contesto di un processo dipende dalla natura del processo stesso. Per H.Nagel (cf [16]) il contesto del processo di riconoscimento di azioni è una struttura complessa che comprende delle descrizioni generiche dello spazio, l'evoluzione temporale delle strutture e la supposta intenzione dell'azione. Per T.Strat (cf [73]) il contesto del processo d'analisi d'immagini statiche è, nel suo senso più largo, qualunque informazione che può influenzare la maniera in cui la scena è percepita. In modo più generale, un processo utilizza tre tipi d'informazione: le conoscenze principali, le informazioni contestuali e le informazioni fattuali. Le **conoscenze principali** sono sempre valide, sono direttamente connesse agli obbiettivi del processo e fanno spesso parte di un modello ben definito. Se mancasse una di queste conoscenze, il processo non sarebbe più in grado d'inferire alcun

risultato. Le **informazioni contestuali** dipendono dall'applicazione, ma restano costanti lungo tutto il trattamento. Sono informazioni secondarie, che possono diventare essenziali per risolvere delle situazioni particolari. Le **informazioni fattuali** dipendono dalla stato d'esecuzione del processo. La loro durata d'esistenza è spesso limitata. Corrispondono ai dati d'ingresso e ai dati calcolati.

Proponiamo allora di adottare, come definzione delle informazioni contestuali di un processo, la definizione proposta da F.Brémond (cf [53]). Le informazioni che si definiscono **contestuali** di un processo sono le informazioni che verificano le due condizioni:

1. i loro valori restano costanti durante l'esecuzione del processo;

2. i loro valori variano qualora il processo sia utilizzato per un'altra applicazione;

Le difficoltà d formalizzare il contesto proviene da un lato dalla dipendenza di queste informazioni dal dominio d'applicazione e dall'altro lato dalla vaga frontiera che separa il contesto dagli altri tipi d'informazione.

La rappresentazione della base di contesto che si è deciso di utilizzare per la presente applicazione ha le caratteristiche seguenti:

- È centrata sulla descrizione geometrica di un solo luogo (la scena dove si conduce l'interpretazione); ogni elemento geometrico di questo luogo sarà descritto tramite un formalismo invariante al variare dell'elemento. Questo per permettere ad ogni modulo della catena d'interpretazione di disporre della base del contesto.

- Le informazioni contestuali sono rappresentate sotto forma simbolica; questo per permettere il riutilizzo della base del contesto al variare dell'implementazione della piattaforma d'interpretazione. Si deve infatti tener presente che l'acquisizione del contesto è un compito impegnativo e temporalmente oneroso. Si cerca quindi di garantire la massima riutilizzabilità della base del contesto creata.

3.1.2 Descrizione della base del contesto creata

La base contestuale creata contiene le informazioni geometriche 3D così come informazioni di tipo semantico sulla scena. Le informazioni geometriche sono innanzitutto una pianta 3D della scena sotto forma di suddivisione dello spazio in zone 3D poligonali. Per esempio la pianta dell'ufficio della presente applicazione è stata suddivisa in 11 zone, che definiscono in modo più specifico delle sottoaree di interesse (per esempio le scrivanie) (cf la figura 3.1). Le informazioni geometriche contengono allo stesso modo una descrizione 3D degli oggetti del mobilio come le scrivanie, gli armadi e le porte.

Ad ogni zona e ad ogni oggetto del mobilio si associa una informazione di tipo semantico che permette al sistema d'interpretazione di migliorare il trattamento effettuato. Per esempio si associa alle aree corrispondenti al pavimento l'informazione che le regioni in movimento che vi si trovano possono essere delle persone.

Nell'implementazione ideata, la base di contesto contiene cinque elementi principali relativi alla descrizione geometrica dello spazio:

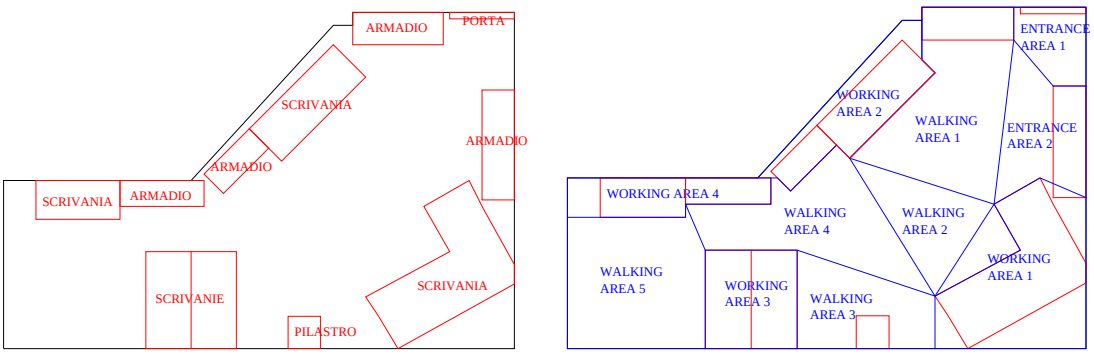
- un insieme di zone poligonali (per esempio la zona d'ingresso, quella d'uscita,...). Ogni zona è un poligono ed è definita tramite una serie di punti 3D appartenenti al poligono stesso. Per esempio la zona d'ingresso ("Entrance area 1" nella figura 3.1.b) è definita nel modo seguente:

zone(name(entrance_zone),
outline(
520, 0, 0,
520, 100, 0,
500, 100, 0,
500, 0, 0),
reference(plane(ground)),
function([“])
)

- un insieme di aree semantiche. Queste aree permettono di raggruppare un insieme di zone connesse accomunate dalla stessa semantica (per esempio l’area “accessible_area” permette di accomunare tutte le zone calpestabili dell’ufficio). Essa risulta così definita:

area(name(accessible_area),
visible(zone(entrance_zone),
zone(entrance_zone_2),
zone(walking_zone_1),
zone(walking_zone_2),
zone(walking_zone_3),
zone(walking_zone_4),
zone(walking_zone_5))
)

- un insieme d’oggetti del mobilio (per esempio armadi, scrivanie, la porta, pilastri, muri...). Ognuno di questi oggetti, schematizzato da un parallelepipedo retto, è definito tramite la propria base (un poligono 3D) e l’altezza. Per esempio un armadio risulta così definito:



(a) Oggetti del mobilio dell’ufficio

(b) Suddivisione dello spazio in zone poligonali

Figura 3.1: *La struttura geometrica della scena nel caso dell’ufficio. Si evidenzia, nella figura (b), la decomposizione in zone effettuata.*

```
object(name(bookcase_1),
        reference(plane(ground),
        shape(block),
        base(
                520, 124, 0,
                520, 244, 0,
                478, 244, 0,
                478, 124, 0),
        height(200),
        labelled([fragile]),
        proximity(100.0),
        normal_time(30.0)
)
```

- una matrice di calibrazione della scena. Questa matrice 4x3 permette di ottenere, con una discreta approssimazione, le coordinate 3D di un punto nello spazio della scena conoscendo le sue coordinate 2D nel piano dell'immagine. È infatti di fondamentale importanza, per ogni sistema d'interpretazione di sequenze video, poter calcolare le relazioni spaziali sussistenti tra i differenti oggetti in movimento così come tra gli oggetti in movimento e gli elementi del mobilio.

  Questa matrice è costruita utilizzando dei punti dei quali si conoscono sia le coordinate 3D che le coordinate 2D. In linea teorica sono necessari solo 6 punti per ricavare i coefficienti della matrice, praticamente se ne usa una trentina e i coefficienti sono stimati ricorrendo al metodo dei minimi quadrati. La matrice è così definita:

```
matrix(
        1.29133,      -0.506317,    -0.799604,    237.786,
        -0.189471,    -0.108496,    -1.55317,     527.943,
        0.00205786,   0.00170457,   -0.00307316,  1
)
```

- Un insieme di piani (nel nostro caso solo il piano del suolo). Questa informazione è necessaria per calcolare le coordinate 3D degli oggetti in movimento a partire dalle coordinate 2D nell'immagine. Per calcolare questa trasformazione è necessario conoscere un piano 3D contenente il punto inferiore della zona in movimento. Per questo nel contesto si associa ad ogni zona a quale piano essa appartiene (per esempio il piano del suolo). Per calcolare la posizione 3D di una persona (nella scena) si fa allora l'ipotesi che il punto più basso della zona mobile corrispondente appartenga al piano del suolo.

  Un piano è definito nel modo seguente:

```
plane(name(ground),
        orientation(horizontal),
        equation(0, 0, 1, 0)
)
```

Agli oggetti così definiti nel contesto sono associate due importanti informazioni di tipo semantico, che saranno poi utilizzate dal modulo di classificazione (cf 5):

- la prima descrive la possibilità (eventuale) di movimento da parte dell'oggetto in questione (per esempio, nel caso questo fosse una sedia, ci si aspetta che essa possa essere spostata). Questa possibilità di movimento è codificata tramite una proprietà dell'oggetto:
  - *moving(stationnary)*: anche se l'oggetto è localmente mobile (per esempio una porta, la sua posizione resta comunque fissa;
  - *moving(mobile)*: la posizione dell'oggetto può cambiare.
- la seconda informazione descrive la capacità dell'oggetto di essere responsabile di occultazioni di individui, e ne specifica la probabile natura:
  - *occlusion(bottom)*: l'oggetto può occultare una persona in basso;
  - *occlusion(top)*: l'oggetto può occultare una persona in alto;
  - *occlusion(partial)*: l'oggetto può occultare parzialmente una persona (per esempio a destra piuttosto che a sinistra);
  - *occlusion(complete)*: l'oggetto può occultare completamente una persona.

Infine, completa la descrizione del contesto l'elenco delle classi di oggetti in movimento caratteristici della specifica applicazione. Come si vedrà più nel dettaglio nel capitolo 5, per un'esigenza di riutilizzazione del modulo, questo è stato concepito con carattere generale, capace cioè di svolgere il proprio compito di classificazione su scene diverse, ognuna con un proprio contesto particolare. Ovviamente, non tutte le scene sono caratterizzate dalla presenza degli stessi oggetti in movimento: la scena "stazione del metro" contempla anche la classe di oggetti in movimento "veicolo" (il treno) mentre in un'applicazione come la presente questa classe non avrebbe senso.

Al fine di poter rendere disponibile al modulo di classificazione l'elenco delle classi di oggetti in movimento sui quali si troverà a lavorare, quest'elenco completa la descrizione del contesto, per esempio:

expected_object(name(person),
w3d_min(35),
w3d_max(100),
h3d_min(80),
h3d_max(200),
found_in([platform,tracks]))

descrive la classe oggetto in movimento "persona", descrivendone alcune caratteristiche ed elencando anche le zone della scena ove ci si aspetta di trovare questo oggetto (per maggiori particolari si rimanda al capitolo 5).

3.2 Architettura del sistema d'interpretazione progettato

Riproponiamo in questo capitolo i due schemi già introdotti nel capitolo di presentazione del presente lavoro. Dettaglieremo nel seguito le parti finora tralasciate e daremo una descrizione funzionale dei moduli ai quali non sarà dedicato un opportuno capitolo nel seguito.

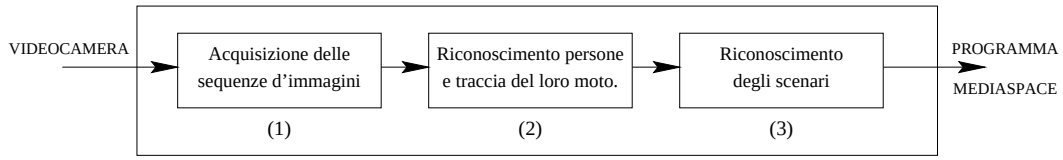


Figura 3.2: *Una prima decomposizione funzionale dell'SDI.*

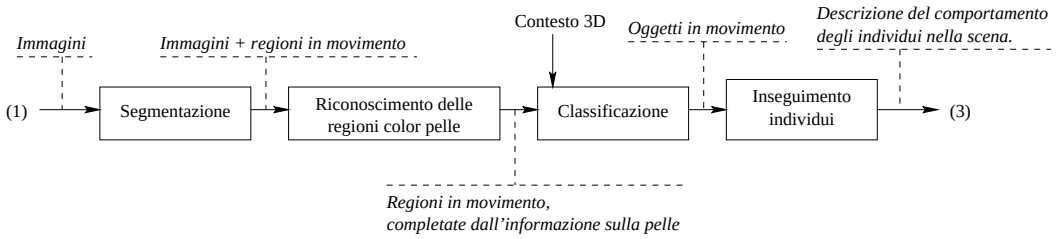


Figura 3.3: *La figura mostra un'ulteriore decomposizione funzionale del modulo (2) mostrato nella figura precedente. Le scritte in corsivo e le relative linee tratteggiate indicano i dati che vengono scambiati tra due moduli, mentre su ogni modulo è riportata la funzione svolta.*

3.2.1 Modulo (1): l'acquisizione delle immagini

Come già spiegato nel capitolo 1, questo modulo è assolutamente standard. Si occupa di interfacciare una videocamera a colori con un PC, occupandosi di:

1. acquisire le immagini alla cadenza massima consentita dall'hardware (nel nostro caso, per immagini RGB a 24 bits, 512x384, circa 5 immagini per secondo);

2. registrare le immagini su disco, in files formato PPM (RGB a 24 bits) oppure JPEG (RGB a 24 bit compresso con perdita di informazione).

Le immagini vengono poi lette da disco ed elaborate dalla piattaforma di interpretazione. Lo spazio disco costituisce una sorta di "spazio tampone" dove immagazzinare le immagini in attesa dell'elaborazione.

3.2.2 Modulo (2): Segmentazione, rilevamento di pelle, classificazione e inseguimento

Questo modulo, vero nucleo centrale del presente lavoro di tesi, nelle sue parti "Riconoscimento delle regioni color pelle", "Classificazione" e "Creazione delle cronologia" è dettagliato nei capitoli successivi (rispettivamente i capitoli 4, 5, 6). Nel seguito si dettaglierà esclusivamente il modulo di segmentazione.

3.2.2.1 La segmentazione

Con il termine *segmentazione* si intente il rilevamento dei singoli pixels "in movimento" di un'immagine. Questo processo di rilevamento consiste quasi sempre nel confronto tra due immagini, una rappresentate il *prima* e l'altra il *dopo* del fenomeno di movimento. In letteratura vi sono principalmente due varianti di algoritmi di segmentazione: alcuni utilizzano una immagine "vuota" (cioè senza oggetti in movimento al suo interno) come immagine di riferimento, e calcolano in seguito la differenza tra ogni nuova immagine da elaborare e questa immagine di riferimento, rilevandone i pixel differenti. Altri algoritmi

calcolano invece questa differenza tra l'immagine attualmente in elaborazione e l'immagine immediatamente precedentente. A seconda del tipo di algoritmo adottato, cambiano radicalmente i risultati ottenuti a livello di rilevamento dei pixels in movimento e, di conseguenza, anche la catena di elaborazione a valle di questo modulo.

Nel nostro caso si è optato per una algoritmo appartenente alla prima famiglia: i pixels in movimento sono dunque calcolati per differenza rispetto ad un'immagine di riferimento vuota. Indicando con

$$\vec{p}_{rif}(i,j) = \begin{bmatrix} R_{rif} \\ G_{rif} \\ B_{rif} \end{bmatrix} \quad e \quad \vec{p}_{el}(i,j) = \begin{bmatrix} R_{el} \\ G_{el} \\ B_{el} \end{bmatrix}$$

le coordinate cromatiche RGB dei pixel di posizione (i,j) nell'immagine di riferimento $(\vec{p}_{rif}(i,j))$ e nell'immagine da elaborare $(\vec{p}_{el}(i,j))$, l'operazione di segmentazione, in prima approssimazione, consiste nel calcolo di

$$\vec{p}_{diff}(i,j) = Thr(s; Abs(\vec{p}_{rif}(i,j) - \vec{p}_{el}(i,j)))$$

dove $\vec{p}_{diff}(i,j)$ indica il pixel risultato dell'immagine di segmentazione, $Thr(s;\cdot)$ indica l'operazione binario di soglia (=1 se il secondo argomento è $\geq s$, 0 altrimenti) e $Abs(\cdot)$ indica l'operatore valore assoluto.

L'immagine differenza contiene quindi il pixel (i,j) colorato (bianco) se questo pixel ha variato in modo sufficientemente pronunciato le proprie coordinate cromatiche nel passaggio dall'immagine di riferimento alla nuova immagine.

La tecnica appena esposta rappresenta un classico nel dominio dell'elaborazione basso livello delle immagini. I vantaggi sono la semplicità d'implementazione e la discreta robustezza. Lo svantaggio principale è l'elevata sensibilità al rumore. Le immagini differenza che si ottengono applicando brutalmente la tecnica esposta sono affette da un forte rumore, dovuto essenzialmente alle variazioni di luminosità e ai riflessi.

Al fine di limitare in modo drastico questo rumore di fondo, l'algoritmo di segmentazione è stato raffinato aggiungendo una routine di aggiornamento dei pixels dell'immagine di sfondo $(\vec{p}_{rif}(i,j))$ in base ai nuovi pixels delle immagini da elaborare. Il principio è di modificare leggermente le coordinate cromatiche di ogni pixel dell'immagine di sfondo secondo le coordinate cromatiche dei nuovi pixels dell'immagine da elaborare:

$$\vec{p'}_{rif}(i,j) = \alpha \cdot \vec{p}_{rif}(i,j) + (1-\alpha) \cdot \vec{p}_{el}(i,j)$$

dove il parametro α assume valori variabili tra 0.9 e 0.99 a seconda della velocità di aggiornamento scelta.

Questo raffinamento dell'algoritmo di segmentazione consente di ridurre il rumore di fondo che affligge l'immagine differenza. Purtroppo presenta un inconveniente altrettanto fastidioso: essendo l'aggiornamento dello sfondo esteso in modo indiscriminato a tutta l'immagine, se un oggetto in movimento restasse nella stessa posizione per un certo periodo di tempo (decine di secondi), finirebbe per essere integrato nello sfondo, e quindi non darebbe più origine ad alcuna zona di differenza. Il risultato sarebbe il mancato rilevamento di un oggetto in movimento presente nell'immagine. Si rivela quindi necessario introdurre un meccanismo di discriminazione tra i pixels da aggiornare e quelli da non aggiornare; come si desidera, infatti, che un pixel soggetto alle normali variazioni di luminosità tipiche di una lampada al neon o dell'illuminazione naturale venga aggiornato per tenere conto di questi fenomeni (limitando quindi il rumore che esse introdurrebbero altrimenti nell'immagine differenza), così vogliamo anche che i pixels appartenenti agli oggetti in movimento non

aggiornino i corrispondenti pixels di riferimento, per evitare il fenomeno di "integrazione nello sfondo" degli oggetti in movimento.

L'ulteriore raffinamento introdotto nell'algoritmo di segmentazione consiste allora a limitare l'aggiornamento alle zone dell'immagine non coperte dalla precedente immagine differenza. In altri termini, l'algoritmo utilizza l'immagine differenza calcolata nel ciclo precedente come "maschera" per decidere quali pixels dello sfondo aggiornare e quali no. Le regioni indicate nell'immagine differenza come in movimento (quindi i cui pixels sono colorati) non saranno aggiornate (in quanto corrispondenti ad oggetti in movimento, appunto), mentre i pixels "spenti" dell'immagine di riferimento verranno aggiornati. Il fatto di utilizzare l'immagine differenza calcolata all'istante precedente per determinare l'aggiornamento all'istante attuale, sebbene concettualmente poco soddisfacente, è un fenomeno del tutto trascurabile, in quanto elaborando 3 immagini al secondo, le variazioni alle quali sono soggette le regioni in movimento in 300 ms sono assolutamente trascurabili.

Una volta ottenuta l'immagine differenza, che consta quindi di un insieme di pixels colorati, è necessario astratte informazioni più generali e facilmente manipolabili. Questo processo consiste nel passare dal singolo pixel (inteso come "quanto" d'informazione) alle regioni in movimento, che raggruppano pixels contigui costituendo una unità logica ad livello di astrazione superiore. Fondamentalmente, questo processo è organizzato nei passi seguenti:

- partendo dai pixels colorati, si determinano le zone colorate, intese come insieme di pixels immediatamente a contatto l'uno con l'altro;
- di ognuna di queste zone si calcolano centro di gravità, numero di pixel componenti e rettangolo circoscrivente;
- si esamina la possibilità di fondere ogni zona con altre a lei più prossime. Se le distanze tra le due (tra i bordi dei rettangoli circoscriventi) sono piccole confrontate alle dimensioni medie (media di altezza e larghezza del rettangolo circoscrivente) della più grande delle due, si effettua la fusione, altrimenti no. Lo scopo di questo passaggio è di raggruppare più aree limitrofe (ma non a contatto) in una di dimensioni superiori. Si è volutamente limitata la complessità di questo algoritmo di fusione, che non risulta molto sofisticato, in quanto un secondo, più raffinato e performante, è implementato a livello del modulo di classificazione (cf il paragrafo 5).

La figura 3.4 mostra un esempio di come appaiono le regioni in movimento all'uscita dal modulo di segmentazione.

3.2.3 Modulo (3): l'interpretazione e il riconoscimento degli scenari

In questa sezione si descriverà il principio di funzionamento del modulo di interpretazione. Tale modulo è stato ideato e sviluppato interamente da N.Rota, in quanto esula dal soggetto del presente lavoro di ricerca. A tal proposito si vedano anche [14] e [74].

Il modulo di interpretazione rappresenta l'ultimo stadio dell'intera piattaforma. Esso riceve i dati d'ingresso sia dal modulo d'inseguimento, sia dalla base del contesto. Tali dati rappresentano, ad ogni istante di tempo t, la descrizione degli oggetti (in movimento e non) presenti nell'immagine trattata. Tale descrizione comprende quindi essenzialmente due categorie di oggetti:

- la descrizione di ogni oggetto del contesto presente nella scena; tale descrizione, costante al variare delle immagini, viene prelevata dal modulo di interpretazione direttamente dalla base del contesto;

(a) Immagine di sfondo (b) Immagine da elaborare (c) Regioni in movimento

Figura 3.4: *La figura mostra le regioni in movimento (c) ottenute dall'algoritmo di segmentazione, elaborando l'immagine (b) ed utilizzando come immagine di riferimento l'immagine (a).*

- la descrizione degli individui in movimento nella scena; il modulo d'inseguimento fornisce ad ogni t la lista delle persone inseguite nella stanza, la loro posizione, le loro dimensioni e la loro velocità.

3.2.3.1 Gli eventi

Il problema del riconoscimento degli scenari può essere ricondotto al riconoscimento di una serie di eventi ordinati temporalmente. Un *evento* è una proprietà spazio-temporale che rappresenta un significativo cambiamento nello stato di una scena. Eventi tipici sono "entrare", "cominciare a correre", "alzarsi" o "uscire". L'algoritmo per il riconoscimento degli eventi è così strutturato: un evento viene riconosciuto se, fissando uno stato, il valore di questo stato cambia tra un'immagine I_0 ed un'immagine I_n. Per esempio, se la persona A si trova, nell'immagine all'istante t, lontana dalla macchina del caffé (la posizione della macchina del caffé, nonché il fatto che l'oggetto in questione sia una "macchina del caffé" sono informazioni che provengono dalla base del contesto; la posizione della persona A proviene invece dal modulo d'inseguimento), mentre all'istante $t + d$ si trova vicina alla macchina del caffé, l'evento "la persona A si è avvicinata alla macchina del caffé" viene riconosciuto. Al fine di poter riconoscere un numero sufficiente di eventi, suscettibili di descrivere svariati scenari, è quindi necessario disporre di una descrizione sufficientemente accurata in termini simbolici della scena. Tale descrizione è costituita appunto dal contesto e dagli individui inseguiti dal modulo d'inseguimento. Disponendo di questa descrizione simbolica della scena, per ogni immagine è allora sufficiente comparare le due descrizioni corrispondenti per scoprire i cambiamenti intervenuti, cioé gli eventi che si sono verificati.

L'autore di [14] definisce un modello rigoroso di stato, che non verrà dettagliato nel seguito. Per ogni immagine, il modulo di interpretazione istanzia un insieme di stati predefiniti, basandosi sugli oggetti e sugli individui presenti nella scena. Questo insieme di stati predefiniti costituisce così una descrizione della scena stessa. Il riconoscimento degli eventi è effettuato confrontando i differenti insiemi di eventi istanziati nei differenti istanti di tempo. Lo stato il cui valore simbolico è cambiato genera un nuovo evento.

Gli otto stati predefiniti utilizzati dal modulo d'interpretazione sono i seguenti:

- *Posizione(individuo$_i$)* $\in$ {*sdraiato, inginocchiato, in piedi*}: il valore simbolico dello stato si ottiene comparando le dimensioni dell'oggetto in movimento "individuo" con le dimensioni standard definite nel modello di essere umano (appartenente alla base del contesto);

- *Direzione(individuo$_i$)* $\in$ {*verso destra, verso sinistra, verso la videocamera, in direzione opposta alla videocamera*}: il valore simbolico è ottenuto dalle informazioni di velocità e di direzione associate all'individuo;
- *Velocità(individuo$_i$)* $\in$ {*fermo, in cammino, di corsa*};
- *Localizzazione(individuo$_i$, area$_j$)* $\in$ {*all'esterno, all'interno*};
- *Prossimità(individuo$_i$, oggetto$_j$)* $\in$ {*vicino, lontano*};
- *Localizzazione_relativa(individuo$_i$, individuo$_j$)* $\in$ {*vicino, lontano*}, $(i \neq j)$;
- *Posizione_relativa(individuo$_i$, oggetto$_j$)* $\in$ {*seduto, nessuna*};
- *Movimento_relativo(individuo$_i$, individuo$_j$)* $\in$ {*insieme, nessuno*}, $(i \neq j)$.

Per esempio, lo stato *Movimento_relativo(individuo$_i$, individuo$_j$)* è definito misurando l'angolo tra i vettori velocità di *individuo$_i$* e *individuo$_j$* e la mutua distanza tra i due *individuo$_i$* e *individuo$_j$* stessi. Se i vettori velocità hanno approssimativamente la stessa direzione (l'angolo tra i due deve essere compreso tra 315 gradi e 45 gradi) e se la distanza è piccola (inferiore a 200 cm) allora si considera che le due persone stiano camminando insieme.

I cambiamenti dello stato *Posizione(individuo$_i$)* generano gli eventi *l'individuo$_i$ cade*, *l'individuo$_i$ si inginocchia* e *l'individuo$_i$ si alza*.

I cambiamenti dello stato *Direzione(individuo$_i$)* generano gli eventi *l'individuo$_i$ si sposta verso destra*, *l'individuo$_i$ si sposta verso sinistra*, *l'individuo$_i$ si allontana* e *l'individuo$_i$ si avvicina*.

I cambiamenti dello stato *Velocità(individuo$_i$)* generano gli eventi *l'individuo$_i$ si ferma*, *l'individuo$_i$ corre* e *l'individuo$_i$ cammina*.

I cambiamenti dello stato *Localizzazione(individuo$_i$, area$_j$)* generano gli eventi *l'individuo$_i$ esce dall'area$_j$* e *l'individuo$_i$ entra nell'area$_j$*.

I cambiamenti dello stato *Prossimità(individuo$_i$, oggetto$_j$)* generano gli eventi *l'individuo$_i$ si avvicina all'oggetto$_j$* e *l'individuo$_i$ si allontana dall'oggetto$_j$*.

I cambiamenti dello stato *Localizzazione_relativa(individuo$_i$, individuo$_j$)* generano gli eventi *l'individuo$_i$ si avvicina all'individuo$_j$* e *l'individuo$_i$ si allontana dall'individuo$_j$*.

I cambiamenti dello stato *Posizione_relativa(individuo$_i$, oggetto$_j$)* generano l'evento *l'individuo$_i$ si siede sull'oggetto$_j$*.

I cambiamenti dello stato *Movimento_relativo(individuo$_i$, individuo$_j$)* generano l'evento *l'individuo$_i$ e l'individuo$_j$ camminano affiancati*.

3.2.3.2 Gli scenari

Il problema finale è il riconoscimento incrementale di un certo numero di scenari predefiniti. Uno scenario è definito come *un insieme di eventi sottostanti a determinati vincoli*. Per esempio, lo scenario "due persone si incontrano alla macchina del caffé" può essere descritto tramite quattro eventi: "una persona si avvicina alla macchina del caffé", "questa persona si ferma", "una seconda persona entra nell'area denominata zona caffé" e "la seconda persona si avvicina alla prima".

Riconoscere uno scenario implica il saper riconoscere tutti gli eventi che lo compongono e saper verificare i vincoli tra i differenti eventi. I vincoli possono essere temporali, spaziali, logici o algebrici. Uno scenario può quindi essere:

- totalmente riconosciuto, quando tutti gli eventi sono stati riconosciuti e i vincoli sono rispettati;
- parzialmente riconosciuto, quanto un sottoinsieme S di tutti gli eventi è stato riconosciuto e i vincoli inerenti gli elementi $\in S$ sono soddisfatti;
- non riconosciuto, quando nessun evento è (ancora) riconosciuto.

L'algoritmo di riconoscimento degli scenari si compie in due passi: primo, immagine dopo immagine si generano gli eventi corrispondenti ai cambiamenti nel valore logico di uno stato e in seguito si istanziano i modelli di scenario predefiniti utilizzando questi eventi. Ciò equivale a dire che il riconoscimento degli scenari corrisponde all'aggiornamento di un insieme di scenari parzialmente riconosciuti.

3.2.3.3 Il modello di scenario utilizzato

Uno scenario si_t, dove i è l'identificatore associato allo scenario e t l'istante di riconoscimento, si compone di quattro parti: **Events**, **Constraints**, **Conditions** e **Success**:

- la parte **Events** consiste in un insieme di eventi $\{e_1, \ldots, e_i, \ldots, e_n\}$ necessari per il riconoscimento dello scenario. Ogni evento e_i è associato alla variabile t_i che rappresenta l'istante in cui si è verificato e_i. Due sono le categorie di eventi in questa sezione: gli eventi positivi e quelli negativi. I positivi devono verificarsi affinché lo scenario sia totalmente riconosciuto mentre i negativi *non* devono verificarsi;
- la parte **Constraints** consiste in un insieme di vincoli temporali $\{c_1, \ldots, c_i, \ldots, c_n\}$ descritti come una serie di disuguaglianze lineari di primo grado;
- la parte **Conditions** consiste in un insieme di vincoli non temporali $\{k_1, \ldots, k_i, \ldots, k_n\}$ riguardanti gli oggetti che generano gli eventi. Essi forzano un attributo dell'oggetto ad assumere un valore predefinito. L'attributo può essere simbolico (nome, funzione, etc.) o numerico (altezza, dimensione, velocità, etc.);
- la parte **Success** consiste in un insieme di parole chiave $\{f_1, \ldots, f_i, \ldots, f_n\}$ che indicano il tipo di azione associata al riconoscimento dello scenario. Questa parte viene utilizzata in caso di completo riconoscimento dello scenario. Esistono due tipi di azione: esterna od interna. L'azione esterna può essere utilizzata, per esempio, per inviare un allarme ad un operatore umano (o per comandare un modulo esterno quale una piattaforma MediaSpace), mentre l'azione interna è utilizzata per generare un evento che indica l'avvenuto riconoscimento dello scenario.

La figura 3.5 mostra un esempio di scenario "due persone si incontrano alla macchina del caffé".

Dato un insieme di scenari $\{s_{1,t-1}, \ldots, s_{i,t-1}, s_{i+1,0}, \ldots, s_{k,0}\}$ composto da scenari parzialmente riconosciuti al tempo $t-1$ e scenari non ancora riconosciuti ed inoltre un insieme di eventi $\{e_{1,t}, \ldots, e_{n,t}\}$ riconosciuti all'istante t, il principio del riconoscimento degli scenari si basa su due punti:

- per ogni $s_{i,t-k} \in \{s_{1,t-1}, \ldots, s_{i,t-1}, s_{i+1,0}, \ldots, s_{k,0}\}$, per ogni e_i di $s_{i,t-k}$, se l'evento e_i coincide con uno degli eventi $e_{1,t}, \ldots, e_{n,t}$ e verifica i vincoli temporali $c_1, \ldots, c_i, \ldots, c_m$ nonché i vincoli non temporali $k_1, \ldots, k_i, \ldots, k_p$, creiamo $s_{i,t}$. Ne risulta un un nuovo insieme di scenari $\{s_{1,t}, \ldots, s_{l,t}\}$;

Scenario{**Name**("Two persons meet at the coffee machine"),
Events(occur(t_1, *moves close to*(p_1 : Person, e_1 : Equipement)),
occur(t_2, *stops*(p_1 : Person)),
occur(t_3, *enters*(p_2 : Person, a_1 : Area)),
occur(t_4, *moves close to*(p_1 : Person, p_2 : Person))),
Contraints(t1 $\leq$ t2, t3 $\leq$ t4),
Conditions(name(e_1, "coffee machine"),
name(a_1, "Coffee area"),
name(e3, "Guest's chair")),
Success(alarm(h_1, "and", h_2, "meet at", e_1)),
}

Figura 3.5: *Esempio di scenario basato sul riconoscimento di quattro eventi.*

- gli scenari $s_{i,t}$ non validi vengono rimossi da $\{s_{1,t}, \ldots, s_{l,t}\}$ se:
  - un evento negativo e_k di $s_{i,t}$ si è verificato;
  - uno dei vincoli $c_1, \ldots, c_i, \ldots, c_m$ è incompatibile con $s_{i,t}$;
  - tutti gli e_k positivi di $s_{i,t}$ sono stati istanziati. In questo caso $s_{i,t}$ è stato completamente riconosciuto e l'insieme di azioni $\{f_1, \ldots, f_i, \ldots, f_n\}$ viene eseguito.

Nei tre capitoli successivi si descriveranno nel dettaglio i moduli ideati e realizzati nell'ambito di questo lavoro di tesi.

Capitolo 4

Il riconoscimento della pelle

4.1 Importanza del riconoscimento delle regioni color pelle

Una delle tecniche utilizzate per classificare una regione in movimento è rilevare se questa possieda o meno, al suo interno, delle parti rappresentanti pelle umana. Ciò si riconduce a rilevare quali delle regioni in movimento siano anche *regioni color pelle* (nel seguito queste saranno indicate con l'abbreviazione "rcp"). Lo scopo è calcolare, utilizzando le informazioni cromatiche contenute nell'immagine, la probabilità che un pixel abbia un colore corrispondente al colore della pelle umana. Questa informazione permette di completare la descrizione di ogni regione mobile con un'etichetta "contiene una rcp" (o "non contiene regioni color pelle"); questa etichetta facilita la classificazione della regione mobile.

4.1.1 I problemi principali della classificazione

Uno dei problemi principali che appaiono in fase di classificazione delle regioni in movimento è saper distinguere la natura di ogni diversa regione: l'esperienza mostra che se si filma il movimento di un essere umano tramite una videocamera e successivamente si analizzano le differenze tra queste immagini ed un'immagine di riferimento, presa a scena vuota (immagine dello sfondo), le aree di differenza individuate (regioni in movimento) sono essenzialmente originate da tre fenomeni differenti:

- l'essere umano che si è spostato;
- il rumore (per esempio l'ombra dell'essere umano);
- gli oggetti fisici presenti nella scena e che sono stati spostati (sedie, porte,...).

La figura 4.1 mostra un'esempio di queste diverse situazioni.

La classificazione si basa, tra l'altro, sulla corrispondenza tra una regione in movimento ed un modello, ricavato da una base di conoscenze. L'operazione di classificazione diventa particolarmente difficile quando:

- la corrispondenza tra la regione in movimento e il corrispondente modello è approssimativa. Per esempio: la regione mobile corrisponde ad un essere umano, ma essendo occultata da un'oggetto della scena, le sue dimensioni non corrispondono più al modello o corrispondono addirittura ad un modello diverso (per esempio un oggetto della scena);

Figura 4.1: *Un esempio delle differenti regioni in movimento calcolate per differenza rispetto ad un'immagine di riferimento.*

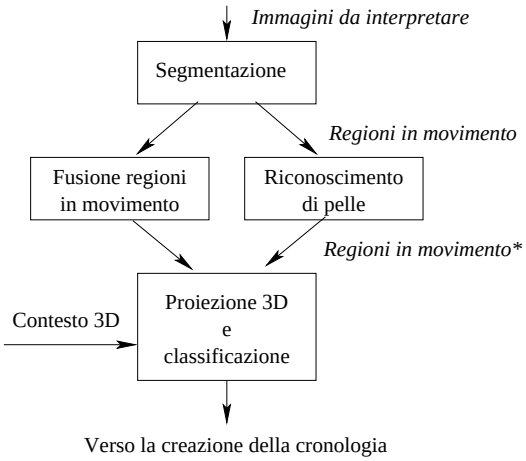


Figura 4.2: *Il modulo di rilevamento della pelle all'interno dell'architettura generale.*

- la regione in movimento corrisponde a più modelli contemporaneamente (per esempio, un'ombra può corrispondere sia a del rumore, sia ad una persona).

4.1.2 Vantaggi del riconoscimento delle regioni color pelle (rcp)

Il modulo di rilevamento delle rcp fornisce un'informazione capace di migliorare la precisione del modulo di classificazione. La presenza/assenza del color pelle costituisce uno dei criteri utilizzati dall'algoritmo di classificazione. Il modulo analizza le componenti cromatiche di ogni pixel dell'immagine d'ingresso e, utilizzando una biblioteca di pixel color pelle, effettua una analisi statistica che restituisce, come risultato intermedio, la probabilità che il pixel considerato abbia il colore della pelle. Questa probabilità è quindi elaborata per estrarne un valore binario che indica la presenza/assenza del color pelle nella regione trattata (processo di astrazione dell'informazione).

Lo schema 4.2 ripropone il ruolo di questo modulo all'interno dell'architettura generale della piattaforma d'interpretazione.

L'informazione sulla presenza/assenza di pelle fornita da questo modulo aiuta a classificare correttamente le differenti regioni in movimento, soprattutto nei casi incerti elencati al paragrafo 4.1.1:

- la presenza del color pelle indica una regione in movimento corrispondente ad un essere umano piuttosto che ad un oggetto della scena;
- la presenza (assenza) del color pelle permette d'individuare correttamente l'essere umano (la sua ombra).

4.2 L'algoritmo utilizzato

In base alle considerazioni espresse precedentemente (cf. paragrafo 2.3.1.1 a pagina 26), l'algoritmo progettato è di tipo statistico a base di istogrammi.

4.2.1 Il principio di funzionamento

Gli algoritmi a base statistica utilizzano un modello (statistico, appunto) del *colore pelle*. In alcune varianti si utilizzano perfino due modelli, uno relativo al color pelle e l'altro relativo ai colori non di pelle (la definizione di questi ultimi varia sensibilmente da sistema a sistema, in generale, comunque, i due modelli non sono complementari rispetto all'intero spazio cromatico nel quale si lavora). I modelli sono codificati utilizzando degli istogrammi multidimensionali (in prima approssimazione, una dimensione per variabile cromatica utilizzata. Nel caso RGB, per esempio, gli istogrammi sono tridimensionali): il principio è quindi organizzare in classi, a seconda del valore delle loro coordinate cromatiche, i pixels di pelle (non pelle) utilizzati per costruire il modello.

Sotto opportune condizioni, specificate nel seguito, il modello così costruito rappresenta la distribuzione statistica (ricostruita a partire da un vasto campione della popolazione che si vuole studiare, i pixels di pelle, per esempio) del color pelle (non pelle) nelle variabili cromatiche scelte.

Una volta costruito il modello, si può assegnare ad ogni pixel di pelle in studio un valore di probabilità (la probabilità che le coordinate cromatiche di questo pixel corrispondano a quelle di un pixel di pelle) ottenuto ricavando il valore della distribuzione statistica precedentemente costruita nel punto indirizzato dalle coordinate cromatiche del pixel all'esame.

4.2.2 Introduzione sul formato di codifica delle immagini bitmap PPM (RGB)

Le immagini utilizzate sono codificate in formato bitmap RGB a 32 bits (8 bits per canale cromatico, 4 canali cromatici), alla risoluzione di 512x384 pixels. In pratica, il colore di ogni pixel è espresso utilizzando quattro numeri di 8 bits ciascuno, che corrispondono all'intensità della componente cromatica rossa (R), verde (G), blu (B) e del canale alfa (α). Il canale alfa non sarà mai preso in considerazione nel seguito, in quanto non fornisce un'informazione cromatica, bensì di trasparenza.

I tre canali, in quanto codificati su 8 bits, restituiscono ciascuno un valore $v \in [0; 255]$. Il numero massimo di colori che possono essere utilizzati è quindi

$$N_{col} = 256 \cdot 256 \cdot 256 = 2^8 \cdot 2^8 \cdot 2^8 = 2^{24} = 16\,777\,216$$

4.2.3 Il primo algoritmo: spazio cromatico RGB e istogramma del color pelle

Per sapere se un pixel possiede il color pelle si confronta il colore di questo pixel con il modello cromatico del color pelle, indicato con $M(r, g, b)$. M è la distribuzione statistica sperimentale del color pelle, ottenuta a partire da $N = 352000$ pixels rappresentativi del vero color pelle.

Gli N pixels sono stati ricavati da una biblioteca d'immagini contenenti esclusivamente delle immagini di pelle. Il modello M è espresso sotto forma di una matrice, contenute alla posizione (i, j, k) il rapporto tra il numero di pixels aventi componente cromatica $(r = i, g = j, b = k)$ (trovati tra tutti gli N pixels color pelle contenuti nella biblioteca) e il numero N totale dei pixels di pelle. Questa matrice rappresenta una distribuzione di probabilità, visto che i suoi valori $\in [0; 1]$ e che la somma di tutti questi valori è pari ad 1.

Per un pixel dato di coordinate cromatiche (r_p, g_p, b_p), la probabilità che questo abbia il color pelle si ottiene leggendo il valore contenuto in $M(r = r_p, g = g_p, b = b_p)$.

L'istogramma M può raggiungere al massimo dimensioni $D_R \cdot D_G \cdot D_B = 255^3$, nel caso si considerino 256 classi per ogni variabile cromatica (quindi un numero di classi pari alla risoluzione massima ottenibile per ogni asse cromatico). In questo caso, per ottenere una distribuzione statistica soddisfacente, sarebbe necessario utilizzare un numero enorme di pixels campione aventi color pelle. Disponendo solamente di $N = 352000$ pixels nella libraria utilizzata, si sono fatte due ipotesi per poter ridurre le dimensioni di questa matrice:

1. *Che ognuna delle variabili R, G e B abbia distribuzione statistica indipendente dalle altre.* Indicando con p_X la distribuzione di probabilità della variabile X, si ha:

   $$p_{R,G,B} = p_R \cdot p_G \cdot p_B$$

   Questa relazione di indipendenza è verificata solo in prima approssimazione; si potrebbe scegliere un diverso spazio dei colori (invece che quello nelle tre variabili R,G,B) nel quale le variabili cromatiche scelte siano realmente statisticamente indipendenti (per esempio lo spazio HSV, Hue, Saturation, Value); tuttavia il passaggio dallo spazio RGB (nel quale sono codificate le nostre immagini) allo spazio HSV si rivela costoso dal punto di vista del tempo di calcolo necessario;

2. ogni componente cromatica (R, G, B) del color pelle abbia una distribuzione statistica *gaussiana* di media μ e varianza σ. Per ogni componente il valore di μ è a priori sconosciuto, ma calcolabile. Indicheremo con μ_R il valor medio della componente cromatica rossa (R) (dei pixels aventi color pelle), μ_G la media del canale G e μ_B quella del canale B.

Al fine di verificare la correttezza di tali ipotesi, si sono calcolate le distribuzioni cromatiche di ogni componente (R, G, B) in funzione dei pixels (N) rappresentativi del colore della pelle. Sul grafico 4.3 si può verificare come queste distribuzioni approssimino soddisfacentemente delle gaussiane. Inoltre la figura 4.4 mostra che la distribuzione congiunta dei canali R e G, per B qualunque, corrisponde con buona approssimazione ad una gaussiana.

Le due ipotesi introdotte giustificano dal punto di vista teorico la riduzione delle dimensioni della matrice M (cf anche [75]); inoltre, una serie di test ci hanno portato a scegliere come dimensioni finali $D_R = 20, D_G = 20, D_B = 5$. La dimensione scelta per il canale B è inferiore alle altre due perché, come confermano numerosi studi (cf per esempio [76]), il canale B è meno informativo rispetto ai canali blu e verde.

Il primo algoritmo lavora nello spazio cromatico RGB illustrato. Per esso sono state utilizzate le dimensioni $D_R = 20, D_G = 20, D_B = 5$.

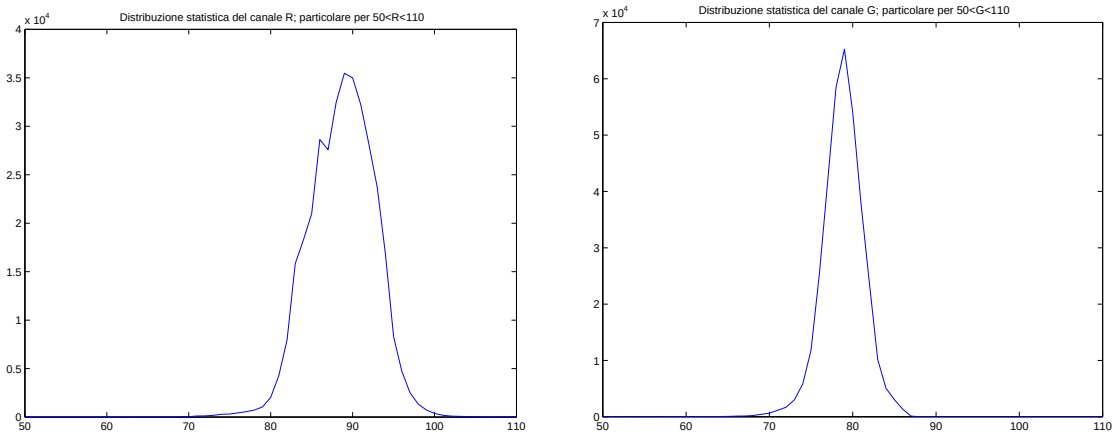


Figura 4.3: *I due grafici rappresentano gli istogrammi dei canali R e G, calcolati utilizzando 352000 pixels color pelle di individui caucasici. Dal grafico si deduce che* $\mu_R \simeq 89$ *e che* $\mu_G \simeq 80$. *Allo stesso modo,* $\sigma_R \simeq 4$ *e* $\sigma_G \simeq 3$.

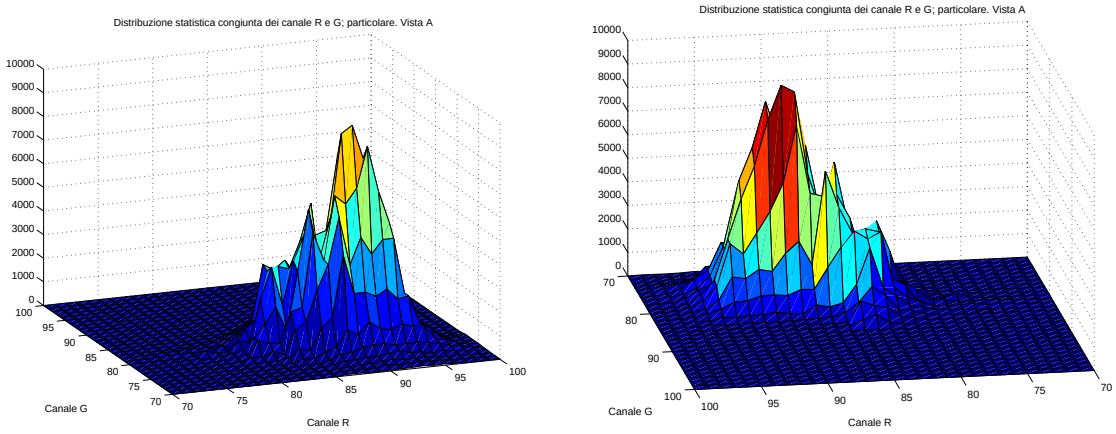


Figura 4.4: *Due punti di vista differenti di un particolare della curva di distribuzione congiunta dei due canali R e G (per B qualunque). La distribuzione è calcolata utilizzando 352000 pixels color pelle di individui caucasici.*

4.2.4 Seconda versione dell'algoritmo: il rapporto degli istogrammi

Lo scopo di questa seconda variante è migliorare il rapporto $\frac{segnale}{rumore}$ ottenuto. In effetti, i risultati offerti dalla prima variante si rivelano, da questo punto di vista, mediocri. I problemi di rumore vengono in gran parte dalla qualità delle immagini che costituiscono la biblioteca di pixels color pelle, e sulle quali è stato costruita la matrice-istogramma M. L'esperienza mostra che una parte dell'immagine, percepita dall'osservatore come di colore (X_R, X_G, X_B), risulta in realtà dalla giustapposizione di pixels di colori differenti che, messi l'uno a fianco dell'altro, danno, in media, il colore percepito. Questo fenomeno è conosciuto con il termine inglese di *dithering*; si tratta dello stesso meccanismo che porta a vedere come "grigia" una pagina bianca riempita di punti neri molto fini. La figura 4.5 mostra un esempio di questo fenomeno.

La conseguenza più indesiderata del fenomeno di dithering è che la distribuzione M, che è calcolata considerando dei pixels color pelle (indicati sinteticamente con "pcp" nel seguito, in analogia con rcp, regioni color pelle), contiene anche dei pixels corrispondenti

Figura 4.5: *L'ingrandimento di una immagine proveniente da una videocamera fino a poter distinguere i singoli pixels mostra la presenza del fenomeno di dithering. A fianco di pixels che hanno il reale colore della pelle si osserva la presenza di pixels di colori differenti.*

a rumore (verdi, grigi, marroni), i quali provocano un'attenuazione del contrasto tra il color pelle e il rumore di fondo. In questo modo si è introdotto del rumore direttamente nella distribuzione, rumore il cui effetto risulta nell'attribuzione di valori di probabilità (di aver il color pelle) elevati anche a pixels che di questo colore non sono (cf la figura 4.6 a pagina 55).

Una soluzione consiste nell'utilizzare videocamere di qualità elevata, che riducono il fenomeno di dithering. Una seconda possibilità è quantificare il rumore di fondo tramite un secondo istogramma: si indicherà il primo istogramma con la notazione M_{pelle}, e il secondo con M_{sfondo}.

Questo secondo istogramma ha le stesse dimensioni del primo, ma è calcolato utilizzando tutti i pixels di un'immagine di "sfondo" che non contiene nessun pixel color pelle. Entrambi gli istogrammi risultano allora calcolati utilizzando pixels corrispondenti al rumore (il primo a causa del fenomeno di dithering e il secondo perché così l'abbiamo costruito); il primo, tuttavia, sarà il solo a contenere anche pixels corrispondenti al color pelle.

Dividendo ogni elemento di M_{pelle} per l'elemento corrispondente di M_{sfondo} si ottiene un istogramma rapporto ($M_{rapporto}$) il cui rapporto $\frac{segnale}{rumore}$ è migliore rispetto al caso precedente. In contropartita, questo algoritmo richiede l'utilizzo di una immagine di sfondo, presa a scena vuota e senza che questa contenga zone di colore cromaticamente prossime al color pelle.

4.2.5 Terzo algoritmo: il rapporto degli istogrammi r,g,l (luminosità)

Nonostante il secondo algoritmo appena presentato, resta un secondo problema: la sensibilità alle condizioni d'illuminazione.

La *luminosità* di un pixel RGB di coordinate cromatiche (r, g, b) è definita come la grandezza

$$l = r + g + b.$$

Essa varia tra 0 (pixel "spento", nessuna intensità luminosa, colore nero) e 765 (pixel bianco, luminosità massima). Le variazioni di luminosità provocano un fenomeno di "deriva" delle componenti cromatiche di ogni pixel all'interno dello spazio tridimensionale RGB. Visto che la distribuzione M dei pixels è calcolata a luminosità costante, questo fenomeno rende il riconoscimento dei pcp tanto più difficile quanto più la differenza di luminosità tra le due immagini (di riferimento e in elaborazione) è importante.

Alcune soluzioni possibili per ovviare a questo inconveniente sono:

1. tener conto del fenomeno calcolando la distribuzione M utilizzando più immagini di pelle riprese in condizioni d'illuminazione differenti; questa soluzione è piuttosto onerosa nell'implementazione (cf il paragrafo 4.4.1 a pagina 57).

2. normalizzare la distribuzione M rispetto alla variabile luminosità.

È stata scelta la seconda soluzione, sostituendo alla distribuzione M (nelle tre variabili r, g et b) la distribuzione M'; M' è ottenuta tramite il cambiamento di variabili seguente:

$$\begin{cases} r' & \rightarrow & \dfrac{r}{r+b+g+1} \\ g' & \rightarrow & \dfrac{g}{r+g+b+1} \\ l & \rightarrow & r+g+b+1 \end{cases}$$

Si può notare che:

1. la variabile r è normalizzata utilizzando il valore di luminosità ($l = r+g+b$)[1];

2. allo stesso modo, la variabile g è normalizzata utilizzando il valore di luminosità;

3. la variabile b è stata sostituita dalla luminosità stessa.

Si sarebbe potuto aggiungere la luminosità comme quarta variabile, e passare dunque ad una matrice M a quattro dimensioni. Per evitare di aumentare la complessità dell'algoritmo si è preferito eliminare il canale blu, meno informativo rispetto al rosso e al verde (cf. [76]) sostituendolo con la luminosità.

La matrice M risulta così essere semidiagonale inferiore, visto che è impossibile ottenere $\frac{r}{r+g+b+1} > 0.5$ e, allo stesso tempo, $\frac{g}{r+g+b+1} > 0.5$([2]).

In seguito si calcola il rapporto dei due istogrammi M_{pelle} et M_{sfondo}. M_{pelle} è relativo alle immagini della biblioteca di pelle e M_{sfondo} all'immagine di sfondo, senza regioni di pelle. Utilizzando l'istogramma rapporto si ottiene un rapporto $\frac{segnale}{rumore}$ migliore (cf l'immagine 4.6).

4.3 Le tre tecniche a confronto

Ciascuna delle tre tecniche offre determinati vantaggi ed inconvenienti. La tabella 4.1 riassume le differenti prestazioni delle tre varianti, secondo tre punti di vista diversi: la sensibilità al rumore, la sensibilità alle variazioni di luminosità e la quantità di errori prodotti.

A tal scopo si sono definite le due categorie d'errori seguenti:

- falso rilevamento (falso positivo): quando si rileva il pixel P_a come avente il color pelle quando in effetti non l'ha;

- mancato rilevamento (vero negativo): quando non si è rilevato il pixel P_a come avente il color pelle quando in effetti l'aveva.

(a) Immagine originale

(b) Risultato dell'algoritmo 1

(c) Risultato dell'algoritmo 2

(d) Risultato dell'algoritmo 3

Figura 4.6: *Una serie di immagini con i risultati ottenuti tramite le varianti 1,2 e 3 dell'algoritmo. Il colore rosso indica una probabilità elevata di possedere il color pelle, il bianco una probabilità nulla.*

Il rumore si manifesta essenzialmente sotto la forma di errori di falso rilevamento, mentre gli errori di mancato rilevamento sono all'origine della mancanza di sensibilità del modulo.

Tutti e tre gli algoritmi sono caratterizzati dalla stessa complessità algoritmica; i numeri 2 e 3 richiedono una fase d'inizializzazione più onerosa, visto che è necessario calcolare 3 istogrammi invece che uno solo. Tuttavia si può trascurare questo aspetto, visto che esso incide solamente sulla fase d'inizializzazione dell'algoritmo. I parametri che entrano in gioco sono:

- n_p, il numero dei pcp contenuti nella biblioteca di pelle (352000); la costruzione della matrice M richiede $3n_p$ operazioni, visto che è necessario accedere alle tre componenti cromatiche di ogni pixel;

[1] Si è aggiunto 1 al valore di luminosità per evitare un denominatore nullo in caso di pixel nero ($r = 0, g = 0, b = 0$).

[2] Ciò implicherebbe infatti $\frac{r+g}{r+g+b+1} > 1$, che è assurdo.

	Sensibilità al rumore	**Sensibilità alla luminosità**	**Frequenza degli errori di mancato rilevamento**
Algoritmo 1	Elevata	Elevata	Ridotta
Algoritmo 2	Ridotta	Elevata	Media
Algoritmo 3	Ridotta	Ridotta	Elevata

Tabella 4.1: *Tabella riassuntiva delle caratteristiche delle tre varianti.*

- n_i, il numero di pixels presenti nelle immagini da elaborare (196608);
- n_s, il numero di pixels presenti nell'immagine di sfondo (che ha le stesse dimensioni delle immagini da elaborare);
- N, il numero di immagini elaborate (500 o più);
- D_x, D_y, D_z, le dimensioni della matrice M;

La complessità della fase d'inizializzazione è quindi pari a:

$$C_{init} = 3n_p + D_x \cdot D_y \cdot D_z$$

per l'algoritmo 1 e

$$C_{init} = 3n_p + 3n_s + 2 \cdot D_x \cdot D_y \cdot D_z$$

per le altre due varianti.

Per contro la complessità del trattamento di una immagine è, per tutti e tre gli algoritmi, pari a:

$$C_{image} = 3n_i \cdot N$$

.

Si nota che, in controparte ai vantaggi ottenuti sotto la forma di riduzione del rumore e di minore sensibilità alle condizioni di illuminazione, il terzo algoritmo è più sensibile agli errori di mancato rilevamento. Il fenomeno non ci sorprende, visto che le tre versioni impongono dei vincoli via via più stringenti per poter riconoscere un pixel come avente il color pelle. Tuttavia la figura 4.6 mostra che le regioni di pelle sono correttamente rilevate, anche se la loro taglia è ridotta.

Sulla base delle verifiche sperimentali e in considerazione del fatto che la nostra applicazione si rivela molto sensibile agli errori di falso rilevamento (che comportano l'errato rilevamento di individui), è stata dunque scelta ed integrata alla piattaforma di interpretazione la terza versione dell'algoritmo.

4.4 L'implementazione

Il modulo che si occupa del rilevamento delle rcp è implementato tramite una biblioteca di funzioni in linguaggio C e un biblioteca di immagini di pelle umana. La biblioteca di funzioni contiene tutti gli algoritmi necessari alla creazione, al calcolo e alla manipolazione degli istogrammi e delle immagini.

La biblioteca di immagini di pelle umana è invece utilizzata per la costruzione degli istogrammi di riferimento per il color pelle.

4.4.1 La biblioteca di immagini di pelle

La biblioteca delle immagini della pelle è composta dall'insieme delle immagini contenenti i pixels rappresentativi del color pelle, e che sono utilizzati per la costruzione della matrice M.

Secondo quanto è stato precedentemente esposto, la terza versione dell'algoritmo si rivela più robusta rispetto ai cambiamenti d'illuminazione tra le immagini di riferimento (contenenti i pixels color pelle) e l'immagine da analizzare. Questa caratteristica ci permette di limitare la biblioteca delle immagini di pelle a un insieme di immagini filmate in qualunque condizione d'illuminazione (e non obbligatoriamente corrispondenti alle condizioni che si avranno nelle immagini da analizzare). Dal punto di vista del tipo di pelle raffigurata nelle immagini, è necessario introdurre alcune considerazioni.

L'attribuzione a un pixel di coordinate cromatiche $P_a(r_a, g_a, b_a)$ di una probabilità p che P_a rappresenti un pixel di pelle (pcp) si basa essenzialmente sul fatto che le coordinate cromatiche di P_a siano "prossime" alle coordinate cromatiche d'un pcp di pelle, statisticamente codificate tramite la matrice M. Più la gaussiana rappresentata tramite M è caratterizzata da una varianza elevata, meno la probabilità p sarà discriminante (i valori saranno più o meno simili per la maggioranza dei pixels). Quindi, se si costruisce una biblioteca d'immagini di color pelle contenente un gran numero di pixels aventi dei colori di pelle differenti (pelle di individui di colore, oppure caucasici, o orientali...), ciò che si otterrà, analizzando una stessa immagine, sarà il riconoscimento di un maggior numero di pixels come aventi il color pelle (e, conseguentemente, un aumento degli errori di falso rilevamento, quindi un aumento del rumore). In modo del tutto simmetrico, limitando i tipi di color pelle utilizzati per il calcolo della biblioteca (per esempio, semplicemente pelle di individui caucasici), la probabilità di rilevare dei pcp di tipo differente diminuisce sensibilmente (e di conseguenza aumenta il numero degli errori di mancato rilevamento, specialmente per gli individui il cui colore di pelle non è stato considerato all'atto della creazione della biblioteca).

Di conseguenza la qualità dei risultati è fortemente legata ai tipi di color pelle utilizzati per costruire la biblioteca. Algoritmi più sofisticati di quello implementato sono parzialmente capaci di ridurre questa dipendenza. Ciononostante i risultati ottenuti dall'algoritmo utilizzato sono soddisfacenti, di conseguenza l'algoritmo è stato adottato senza intraprendere ulteriori ricerche.

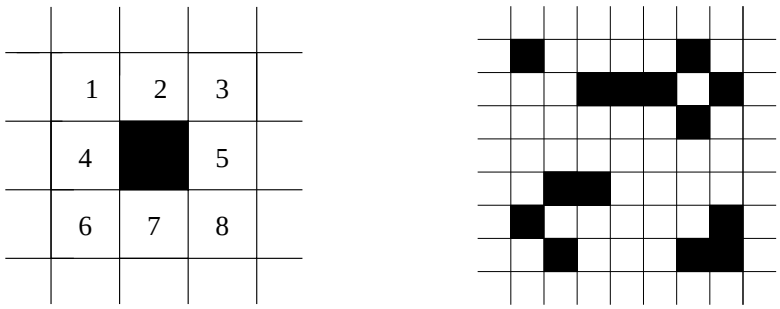
In più, essendo maggiormente interessati a limitare gli errori di falso rilevamento rispetto agli errori di mancato rilevamento, la biblioteca di immagini color pelle è costruita utilizzando solamente immagini di pelle corrispondenti ad individui caucasici.

4.4.2 Calcolo della probabilità che un pixel abbia il color pelle

Una volta costruito l'istogramma M, utilizzando la biblioteca di immagini color pelle descritta al paragrafo 4.4.1, si associa ad ogni pixel dell'immagine da analizzare un valore di probabilità che esso abbia il color pelle. Si indica questo valore con p_p.

Conoscendo le coordinate di ogni pixel da analizzare all'interno dello spazio cromatico scelto (nel nostro caso, r, g, l), il valore della distribuzione $M(r, g, l)$ è letto direttamente nella matrice tridimensionale M alla posizione (r, g, l). È necessario porre l'accento sul fatto che, dal punto di vista statistico, tutti i vincoli necessari affinché questo valore sia realmente una probabilità sono soddisfatti:

$$0 \leq p_p \leq 1;$$



(a) Gli otto pixel adiacenti al pixel centrale

(b) Quattro esempi di regione connessa

Figura 4.7: *La figura (a) mostra la definizione di adiacenza "a 8 pixels": gli otto pixels direttamente a contatto con un pixel scelto sono definiti "a lui adiacenti". La figura (b) mostra quattro regioni connesse secondo la definizione introdotta. Una definizione differente, per esempio l'adiacenza "a 4 pixels" avrebbe determinato un numero differente di regioni connesse (dieci in questo caso).*

$$\sum_{r=0}^{D_r}\sum_{g=0}^{D_g}\sum_{l=0}^{D_l} p_p(r,g,l) = 1$$

.

Una lettura dell'istogramma M all'indirizzo (r, g, l) permette quindi di associare in modo immediato ad ogni pixel il suo valore di probabilità.

4.5 Costruzione delle regioni connesse color pelle

Dopo aver attribuito ad ogni pixel dell'immagine da analizzare la sua probabilità p_p, si astrae questa informazione a livello della regione in movimento. Scopo di questo processo d'astrazione è di riassumere tutta l'informazione ottenuta sotto una forma più compatta, sintetica e più facile da gestire. Per interpretare una scena filmata è infatti inutile disporre di un'informazione che associa ad ogni pixel dell'immagine, ad ogni istante, la sua probabilità p_p di rappresentare l'immagine di un pixel di pelle. Al contrario vogliamo stabilire se una regione in movimento, identificata sull'immagine, contiene una o più rcp, e calcolarne la posizione e le dimensioni. Come la regione in movimento rappresenta un'entità appartenente ad un livello di astrazione superiore rispetto ai pixels che risultano differenti tra due immagini, è necessario compiere lo stesso procedimento d'astrazione sui pcp.

Questo processo è organizzato nel modo seguente:

1. Utilizzando un valore soglia, fissato in base ad osservazioni sperimentali, la probabilità p_p, (rappresentata da valori reali compresi tra 0 e 1) è trasformata in un'informazione di tipo booleano: 1 equivale a segnalare la presenza di un pcp, 0 segnala l'assenza.

2. si costruisce una tabella di regioni connesse color pelle. Una regione connessa color pelle (indicata $\mathcal{R}_i$) è definita come un insieme di pcp adiacenti (cf la figura 4.7). Questa regione è caratterizzata tramite:

   - il rettangolo che la circoscrive;

- la densità media di probabilità di $\mathcal{R}_i$, calcolata sui pixels di color pelle;
- il numero di pixels color pelle che compongono $\mathcal{R}_i$;

L'algoritmo che permette di calcolare le regioni connesse è uno dei più onerosi, dal punto di vista computazionale, dell'intero modulo; infatti è necessario esaminare diverse volte ogni pixel dell'immagine. L'approccio che è stato utilizzato consiste nel calcolare queste regioni connesse in un solo ciclo immagine (cioé percorrendo una sola volta l'immagine), al fine di ridurre il tempo di calcolo.

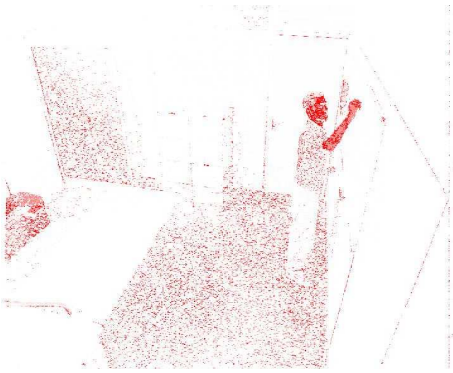
3. Applicando successive tecniche di filtraggio (per dimensione, per densità, per posizione...) si cerca di eliminare in massima misura il rumore residuo e si conservano solo le regioni più significative;

4. Si esamina, per ogni regione filtrata $\mathcal{R}_i$, la possibilita di fonderla con le regioni più prossime, al fine di ottenere regioni connesse più grandi e più significative. Per esempio, se il viso di una persona risultasse scomposto in tre o quattro rcp, sarebbe interessante riunirle in una unica regione connessa color pelle. Per effettuare questa fusione, si calcolano le caratteristiche geometriche e morfologiche delle nuove regioni, ottenute per fusione da due regioni vicine. Se le caratteristiche di quest'ultima sono migliori (rispetto ad alcune cifre di merito quali la densità, per esempio), si effettua la fusione, altrimenti no.

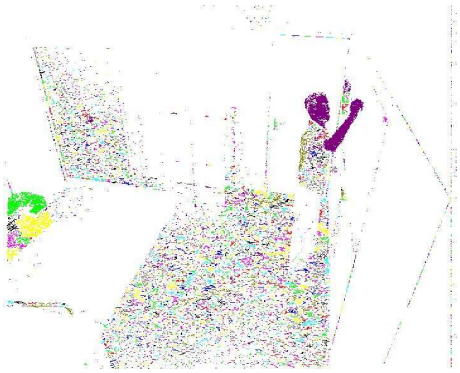
5. Le regioni che restano sono le rcp (regioni color pelle) ricercate. Si crea una variabile (appartenente ad un tipo di dato astratto opportunamente progettato) nella quale si registrano tutte le informazioni geometriche e morfologiche di ogni rcp; questa variabile viene quindi integrata nella descrizione (tramite un secondo tipo di dato astratto) della regione mobile ed è utilizzata dal modulo successivo (che si occupa della classificazione delle regioni in movimento) per determinare se questa regione mobile è un individuo (sono quindi presenti al suo interno delle rcp) piuttosto che del rumore (assenza di rcp).

Nel capitolo successivo si vedrà come l'informazione sulla presenza o assenza, in una regione in movimento, di zone di pelle costituisce uno dei criteri utilizzati dal modulo di classificazione al fine di etichettare le differenti regioni in movimento (passando quindi dalle *regioni in movimento* agli *oggetti in movimento*).

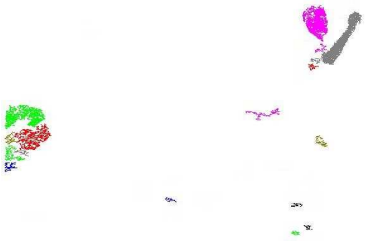
(a) L'immagine da interpretare



(b) La probabilità di ogni pixel di avere il color pelle



(c) Le regioni connesse color pelle: tutti i pixels di una stessa regione sono rappresentati con lo stesso colore



(d) Le regioni connesse color pelle che hanno superato il filtraggio

Figura 4.8: *La figura (a) mostra l'immagine di partenza da interpretare; la figura (b) mostra le rcp caratterizzate da una densità più o meno elevata (il colore bianco rappresenta un pixel con probabilità nulla di aver il color pelle, il rosso una probabilità crescente con l'intensité del pixel.). La figura (c) mostra le regioni connesse color pelle calcolate: un gran numero di regioni sono causate dal rumore e qualche grande regione corrisponde alle reali regioni di pelle nell'immagine. Si può notare un errore commesso, a sinistra nell'immagine, dove una scatola in cartone è caratterizzata da componenti cromatiche quasi identiche a quelle del colore della pelle. Infine, l'immagine (d) mostra come, dopo il filtraggio e la fusione delle regioni limitrofe, le regioni connesse color pelle interessanti siano conservate (insieme ad alcune zone di rumore e al macroscopico errore dovuto alla scatola.*

Capitolo 5

Il modulo di classificazione e fusione

In questo capitolo si descriverà il modulo di classificazione. Questo modulo si interfaccia con l'unità di segmentazione, dalla quale riceve i dati in ingresso (le *regioni in movimento*) e con il modulo di inseguimento, al quale fornisce i dati calcolati (cioé gli *oggetti in movimento*). La figura 5.1 illustra la posizione di questo modulo all'interno dell'architettura del sistema.

5.1 Introduzione

Come già puntualizzato precedentemente (cf il paragrafo 2.3.2.3 a pagina 29), il modulo di classificazione assolve essenzialmente a due compiti:

- assegna ad ogni regione in movimento, ricevuta in ingresso, un'etichetta che ne specifica la classe; in altre parole, opera una suddivisione in classi predeterminate delle regioni in movimento, precisandone la natura;
- secondariamente, esamina la possibilità di una *fusione* tra due o più regioni in movimento, allo scopo di ottenerne una di dimensioni maggiori e logicamente più corretta.

Entrambe queste operazioni saranno dettagliate nelle sezioni seguenti.

5.2 L'importanza del contesto

Fondamentale per poter svolgere la routine di classificazione è disporre di una serie di informazioni a priori di varia natura. Tali informazioni (dettagliate nel paragrafo 3.1.2) costituiscono la base del contesto specifica all'applicazione MediaSpace. Si vedrà infatti nel seguito che sia il modulo di classificazione, sia il formalismo utilizzato per creare la base del contesto sono stati ideati in modo da essere "general purpose", cioè riutilizzabili per altre applicazioni simili (sempre di interpretazione automatica di sequenze video, ma per esempio declinate più in chiave di videosorveglianza rispetto alla nostra applicazione). A seconda dell'applicazione specifica, quindi, il sistema farà uso di una base del contesto particolare.

Nel nostro caso, le informazioni che il modulo di classificazione preleva dalla base del contesto sono essenzialmente di vari tipi:

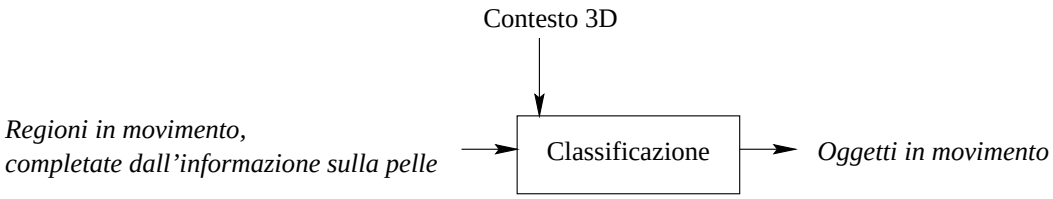


Figura 5.1: *La figura mostra il ruolo del modulo di classificazione e fusione all'interno dell'archiettura del sistema di interpretazione. In corsivo sono riportati i dati scambiati dai vari moduli.*

- di tipo geometrico:
  - la descrizione 3D della scena; il modulo di classificazione necessita di tale descrizione al fine di calcolare le possibili occultazioni degli oggetti in movimento dovute ad elementi dell'arredo;
  - la matrice di calibrazione, utilizzata per ottenere le coordinate 2D di un punto nel piano dell'immagine, conoscendone la posizione 3D nella scena, e viceversa. Questa matrice è fondamentale per situare tridimensionalmente nello spazio della scena un oggetto in movimento (di cui si conosce la posizione 2D nel piano dell'immagine), al fine di studiarne la posizione in riferimento agli elementi dell'arredo;
- di tipo semantico:
  - ogni elemento 3D costituente la scena è associato ad informazioni capaci di specificarne, per esempio, la capacità di generare occultazioni;
  - i modelli degli oggetti in movimento rappresentativi di ogni classe (modelli utilizzati nel processo di classificazione) (cf anche i paragrafi 2.3.2.1 e seguenti);
  - le classi di oggetti riconosciuti nella specifica applicazioni. Per esempio, l'applicazione MediaSpace non prevede alcuna classe "veicolo", utilizzata invece in un'applicazione di videosorveglianza stradale o di una stazione metropolitana. Tale classe non compare quindi nella base del contesto dell'applicazione MediaSpace.

La figura 5.2 mostra un esempio di classificazione nell'applicazione MediaSpace.

Colore del rettangolo circoscrivente	Classe corrispondente	Colore del rettangolo circoscrivente	Classe corrispondente
	individuo		oggetto dell'arredo
	individuo occultato		veicolo
	gruppo d'individui		indeterminato
	folla		rumore

Tabella 5.1: *Il colore del rettangolo circoscrivente indica la classe dell'oggetto in movimento rappresentato nelle diverse immagini.*

(a) Immagine da elaborare

(b) Le regioni in movimento

(c) Gli oggetti in movimento

Figura 5.2: *La figura mostra un esempio di classificazione nell'ambito dell'applicazione MediaSpace. La figura (a) mostra l'immagine d'origine, da elaborare. La figura (b) mostra i dati d'ingresso per il modulo di classificazione, cioé le regioni in movimento. La figura (c) mostra il risultato dell'algoritmo di classificazione: il rettangolo circoscrivente assume un valore diverso per ogni classe individuata (si faccia anche riferimento alla tabella 5.1). In questo caso i due rettangoli rossi indicano che entrambi gli oggetti in movimento appartengono alla classe individuo.*

5.3 La classificazione

In questa sezione si descriverà l'algoritmo utilizzato al fine di classificare le differenti regioni in movimento.

5.3.1 Le classi utilizzate

Il modulo di classificazione gestisce otto classi in totale. A seconda dell'applicazione considerata alcune di queste potrebbero non essere utilizzate (per esempio, per l'applicazione MediaSpace le classi utilizzate sono solo 6, evidenziate in corsivo grassetto):

- ***Individuo***; la regione in movimento corrisponde all'immagine di una persona.
- ***Individuo occultato***; la regione in movimento corrisponde all'immagine di una persona, parzialmente occultata da un ostacolo (tipicamente un oggetto dell'arredo).
- ***Gruppo di persone***; spesso le persone che si spostano in gruppo, o affiancate, risultano nell'immagine come una sola zona in movimento, che viene così classificata "gruppo di persone".
- *Folla*; utilizzata per classificare grosse regioni in movimento originate da movimenti di molte persone. Un tipico esempio di situazione che porta al riconoscimento di un oggetto in movimento di tipo "folla" è l'uscita in massa delle persone dai vagoni della metropolitana, in un'applicazione di videosorveglianza. Questa classe non viene utilizzata nell'applicazione MediaSpace.
- ***Oggetto dell'arredo***; gli oggetti dell'arredo, quali una sedia o una porta. Ovviamente solo gli oggetti suscettibili di spostamenti rientrano in questa classe, essendo gli unici all'origine di una zona in movimento.
- *Veicolo*; la metropolitana nel caso di videosorveglianza nelle stazioni metropolitane. Un generico veicolo in applicazioni di videosorveglianza di traffico (non sono mai state

trattate dal presente modulo di classificazione). Questa classe non viene utilizzata nell'applicazione MediaSpace.

- ***Rumore***; il rumore è in questo caso un fenomeno ottico che può provenire da diverse sorgenti, quali per esempio le ombre degli oggetti, il riflesso degli oggetti sulle superfici riflettenti, i cambiamenti d'illuminazione o gli spostamenti della videocamera.

- ***Indeterminato***; è la classe per default quando le informazioni di cui il modulo dispone sono insufficienti per la classificazione.

Il modello di ciascuna di queste classi fa parte della base del contesto. Detto modello consta essenzialmente della descrizione della forma e delle dimensioni dell'oggetto, ed eventualmente di dove ci si aspetta di trovarlo. Per esempio, il modello *individuo* per l'applicazione di videosorveglianza della metropolitana è il seguente:

expected_object(name(person),
w3d_min(35),
w3d_max(100),
h3d_min(80),
h3d_max(200),
found_in([platform,tracks]))

5.3.2 L'algoritmo di classificazione

L'algoritmo di classificazione associa ad ogni classe dei *criteri di appartenenza* di una regione in movimento a quella classe. Sia r una regione in movimento che deve essere classificata, sia m la classe (a priori una qualsiasi delle classi disponibili) della quale si vuole calcolare il grado di appartenenza di r; il grado di appartenenza di r alla classe m viene indicato con $D(r, m)$ ed è calcolato nel modo seguente:

$$D(r,m) = \frac{\sum_{i=1}^{N_m} w_{m,i} C_i(r,m)}{\sum_{i=1}^{N_m} w_{m,i}}$$

$$D(r,m), C_i(r,m) \in [0,100]$$

dove C_i è l'i-esimo criterio utilizzato per quantificare l'appartenenza, N_m è il numero di criteri utilizzati per la classe m e quindi $C_i(r, m)$ è l'i-esimo criterio di appartenenza della regione in movimento r alla classe m.

L'appartenenza ad una classe viene quindi calcolata come media pesata sui differenti criteri della classe stessa. I $w_{m.i}$ sono i pesi di ogni criterio, parametri fissi che non variano al variare dell'applicazione.

5.3.2.1 I criteri della classe "veicolo"

Per questa classi, i criteri utilizzati sono quattro:

- $C_{H_{3D}}$: cioé l'altezza (in cm) della regione in movimento nello spazio tridimensionale della scena;

- $C_{W_{3D}}$: cioé la larghezza (in cm) della regione in movimento nello spazio tridimensionale della scena;

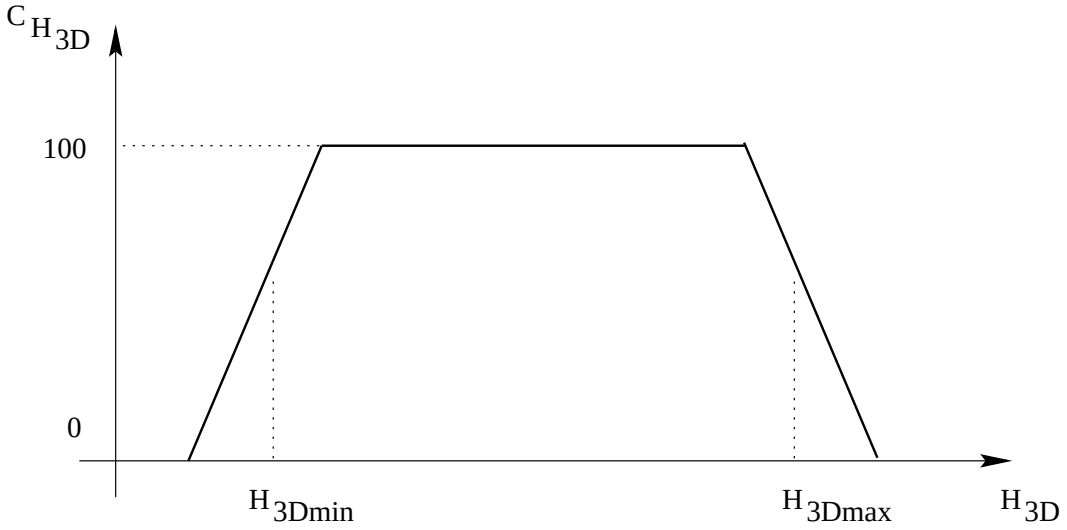


Figura 5.3: *Esempio di funzione per il calcolo di* $C_{H_{3D}}$

- $C_{\frac{H_{3D}}{W_{3D}}}$: il rapporto tra l'altezza e la larghezza della regione in movimento nello spazio tridimensionale della scena;
- C_{po}: la posizione nella scena.

I criteri $C_{H_{3D}}$, $C_{W_{3D}}$ e $C_{\frac{H_{3D}}{W_{3D}}}$ sono calcolati tramite una funzione per intervalli, definita utilizzando i valori massimi e minimi del criterio (cf anche la figura 5.3). Questi valori provengono direttamente dalla base del contesto utilizzata, di conseguenza sono variabili al variare dell'applicazione.

Per il criterio C_{po} la scelta è di tipo binario: se la base della regione in movimento cade nella regione semantica specificata dal modello, allora $C_{po} = 100$, altrimenti $C_{po} = 0$.

Di conseguenza il grado di appartenenza delle regioni in movimento a questa classe viene calcolato nel modo seguente:

$$D(r,m) = \frac{w_{H_{3D}} \cdot C_{H_{3D}} + w_{W_{3D}} \cdot C_{W_{3D}} + w_{\frac{H_{3D}}{W_{3D}}} \cdot C_{\frac{H_{3D}}{W_{3D}}} + w_{po} \cdot C_{po}}{w_{H_{3D}} + w_{W_{3D}} + w_{\frac{H_{3D}}{W_{3D}}} + w_{po}}$$

dove $w_{H_{3D}} = 0.4$, $w_{W_{3D}} = 0.3$, $w_{\frac{H_{3D}}{W_{3D}}} = 0.15$ e $w_{po} = 0.15$; questi valori sono stati fissati sperimentalmente in base ai risultati ottenuti.

5.3.2.2 I criteri delle classi "individuo", "gruppo di persone" e "folla"

Rispetto alla classe "veicolo", queste classi utilizzano sostanzialmente gli stessi criteri, con l'aggiunta del criterio sulla presenza/assenza di zone di pelle nella regione in movimento. Il principio base è identico, cioé il calcolo della media pesata di cinque criteri. Il criterio di pelle è di tipo binario: $C_{pelle} = 0$ in caso di assenza di regioni di pelle, mentre $C_{pelle} = 100$ in caso di presenza di zone di pelle. I pesi dei diversi criteri risultano così modificati: $w_{H_{3D}} = 0.2$, $w_{W_{3D}} = 0.10$, $w_{\frac{H_{3D}}{W_{3D}}} = 0.1$, $w_{po} = 0.1$ mentre $w_{pelle} = 0.5$;

5.3.2.3 I criteri della classe "individuo occultato"

Nel caso più generale, le occultazioni possono essere dovute a due fenomeni differenti: un individuo è parzialmente occultato da un oggetto dell'arredo che si trova interposto tra l'individuo stesso e la videocamera, oppure una parte dell'individuo esce dal campo visuale della videocamera.

In realtà esiste un terzo tipo di occultamento, molto frequente, ed è l'occultamento dovuto ad un secondo oggetto in movimento che si frappone tra la videocamera e il primo oim. Questo tipo di occultazione, tuttavia, richiede un trattamento molto più sofisticato per poter essere gestita correttamente. Inoltre le ripercussioni di questo tipo di occultazione sono limitate dal fatto che due oggetti in movimento sovrapposti vengono percepiti come una sola regione in movimento, e di conseguenza classificati come "gruppo di individui". È poi il modulo di inseguimento a riconoscere e gestire correttamente questa classe di oim. A livello di classificazione, quindi, si gestiscono solamente i primi due tipi di occultazione. La figura 5.4 mostra due esempi di questi tipi di occultazione.

L'occultazione ha come conseguenza immediata la perdita dell'informazione sulla dimensione dell'oggetto in movimento. In altri termini, è impossibile definire un modello di "oggetto in movimento parzialmente occultato" specificandone le dimensioni minime. L'unica informazione disponibile è che le dimensioni massime non possono superare quelle di un individuo non parzialmente occultato. Ciononostante, poiché il fenomeno di occultazione interessa spesso la base della regione in movimento, la matrice di calibrazione non è più in grado di stimare correttamente le dimensioni 3D dell'oggetto in movimento. Di conseguenza si è deciso di non far uso del criterio sulle dimensioni per calcolare il grado di appartenenza di un oggetto alla classe "individuo occultato".

Per il primo tipo di occultazione (dovuta ad un oggetto dell'arredo) si utilizza allora un criterio che prende in considerazione la posizione dell'oggetto in movimento rispetto agli oggetti dell'arredo. Il grado di appartenenza in questo caso è:

> *Se un oggetto dell'arredamento si trova tra la regione in movimento e la videocamera, e se il tipo di occultazione provocata dall'oggetto (ricavata dalla base del contesto) è compatibile con la mutua posizione della regione in movimento e dell'oggetto dell'arredo, allora:*
>
> $$C_{1a} = 100$$
>
> *altrimenti*
>
> $$C_{1a} = 0.$$

Per il secondo caso di occultazione (limiti del campo visuale della videocamera), si calcola la posizione della regione in movimento rispetto ai bordi dell'immagine. In questo caso, il grado di appartenenza è allora:

> *Se la regione in movimento si trova ai bordi dell'immagine,*
>
> $$C_{1b} = 100$$
>
> *altrimenti*
>
> $$C_{1b} = 0.$$

In entrambi i casi si considera poi il criterio che tiene conto della presenza/assenza di pelle:

> *Se la regione in movimento contiene regioni di pelle,*
>
> $$C_2 = 100$$
>
> *altrimenti*
>
> $$C_2 = 0.$$

Figura 5.4: *Le due immagini in alto mostrano un esempio di occultazione dovuta ai limiti di campo di visione della videocamera. Le due immagini in basso un esempio di occultazione dovuto ad un oggetto dell'arredo (un cestino dei rifiuti).*

Il grado finale di appartenenza della regione r alla classe "individuo occultato" è:

$$D(r,''individuo\ occultato'') = \frac{MAX(C_{1a}, C_{1b}) + C_2}{2}.$$

5.3.2.4 I criteri della classe "oggetto dell'arredo"

Per questa classe si fa ancora uso di un unico criterio C_1, che calcola la posizione della regione in movimento rispetto agli oggetti dell'arredo che sono suscettibili di spostamenti. Per esempio, una regione in movimento rilevata nella zona d'apertura di una porta è suscettibile appartenere alla classe "oggetto dell'arredo" (per esempio potrebbe essere originata dalla porta stessa). La figura 5.5 illustra un esempio di questa situazione.

Il grado di appartenenza è calcolato come percentuale di sovrapposizione tra la regione in movimento (il rettangolo circoscrivente) e la proiezione sul piano dell'immagine degli oggetti dell'arredo suscettibili di spostarsi:

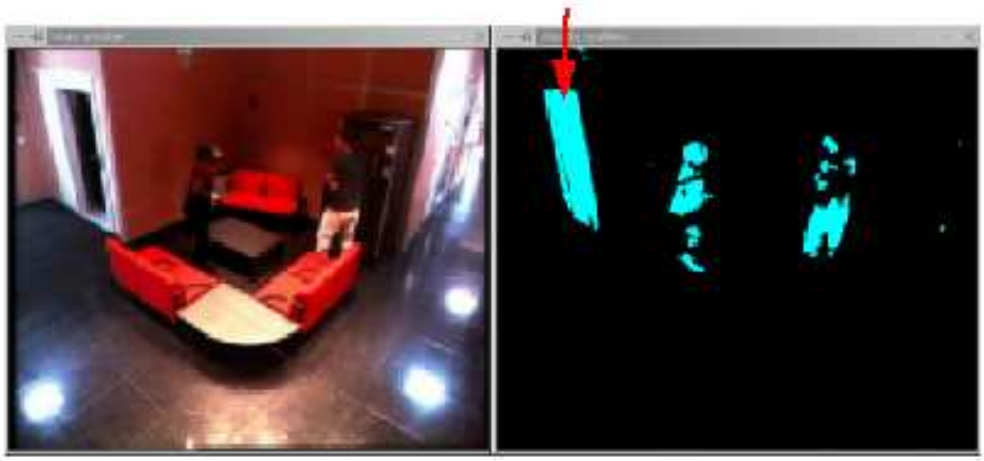


Figura 5.5: *Una regione in movimento di tipo "oggetto dell'arredo". La regione in movimento indicata sull'immagine si trova nella zona d'apertura della porta dell'ascensore.*

$$D = MAX\left(\frac{surf(intrsc(rett_{regione\ in\ movimento}, rett_{oggetto\ k}))}{surf(rett_{regione\ in\ movimento})}\right) \cdot 100.$$

dove $surf(\cdot)$ indica la superficie dell'argomento nel piano 2D dell'immagine, $intrsc$ indica l'intersezione dei due argomenti e $rett_i$ indica il rettangolo circoscrivente la regione in movimento i.

5.3.2.5 I criteri della classe "rumore"

La classificazione ottimale delle regioni in movimento originate da fenomeni di rumore richiede trattamenti sofisticati, spesso non attuabili con una semplice impostazione a base di modelli. Questo perché spesso il rumore, per sua natura, si presenta sotto forma di regioni in movimento di qualsiasi forma e dimensione, difficilmente raggruppabili, quindi, in un modello rappresentativo. Nel nostro caso ci si è limitati a trattare due tipi ben precisi di rumore, tralasciando casi più complicati che avrebbero richiesto tempi di sviluppo ben più importanti. I due casi trattati sono le regioni in movimento che si trovano al di fuori delle cosiddette *ROI* (Region Of Interest, regioni di interesse) e i rumori causati, al bordo dell'immagine, dagli spostamenti della videocamera.

Per ROI si intende una regione, definita nella base del contesto, nella quale si concentra l'attenzione del modulo di classificazione, che tralascia così tutte le regioni in movimento che non vi si trovano, classificandole automaticamente come "rumore". Questa possibilità di selezionare una zona della scena come "degna d'interesse" è molto utile in alcune applicazioni (per esempio la videosorveglianza delle stazioni della metropolitana), dove si possono così tralasciare spazi privi d'interesse (per esempio la banchina opposta). In altre applicazioni, per esempio l'applicazione MediaSpace, questa possibilità si rivela priva di interesse e non viene sfruttata.

I criteri di appartenenza di una regione in movimento alla classe "rumore" sono allora sostanzialmente due, entrambi con valori di tipo binario: 0 o 100. Se la regione in movimento si trova al di fuori delle ROI, oppure il suo centro di gravità si trova sufficientemente vicino al bordo dell'immagine, il criterio di appartenenza assume valore 100, altrimenti assume valore 0.

5.3.3 L'attribuzione della classe

Una volta ottenuti i gradi di appartenenza di ogni regione in movimento ad ogni possibile classe, si cerca per ogni regione in movimento qual é il grado di appartenenza con valore massimo, per esempio $D_{max} = D(r, classe_j)$. Se D_{max} è superiore ad una valore soglia, specifico della classe j cui si riferisce D_{max}, la classe j viene assegnata alla regione in movimento. Se invece D_{max} è inferiore al valore soglia della classe j, si procede allo stesso modo con il secondo miglior grado di appartenenza. Se nessuno dei D è superiore al rispettivo valore soglia, la classe utilizzata è "indeterminato".

In pratica, si sono sperimentalmente fissati tutti i valori soglia a 85.

5.4 La fusione

In tutte le applicazioni del modulo di classificazione, l'accento è stato posto sul riconoscimento degli individui. Di conseguenza la fase di fusione riguarda essenzialmente gli oggetti in movimento classificati "individuo". L'ultima fase dell'algoritmo consiste quindi nel ricostituire in un unico oim diversi oim che possono rappresentare parti di una persona (per esempio i due oim corrispondenti al corpo e alle gambe di una stessa persona).

Per selezionare gli oim da raggruppare si utilizzano due criteri:

- C_{dis}: la distanza 2D tra i bordi dei due oggetti in movimento;
- C_{fus}: il grado di appartenenza alla classe "individuo", calcolato secondo il procedimento già descritto, dell'oim risultante dalla fusione di due oim.

Per ogni coppia di oggetti in movimento $(r_i, r_j), i, j \in [1, n]$, dove n è il numero di oggetti in movimento presenti nell'immagine in elaborazione, viene calcolato un *coefficiente di raggruppamento*, $C_r(r_i, r_j)$ pari a:

$$C_r(r_i, r_j) = \frac{w_{dis} C_{dis}(r_i, r_j) + w_{fus} C_{fus}(r_i, r_j)}{w_{dis} + w_{fus}}$$

dove i due pesi w_{dis} e w_{fus} sono stati posti entrambi $= 0.5$ in base ai risultati sperimentali.

Se questo coefficiente di raggruppamento si rivela superiore ad un valore soglia ($s_{raggruppamento} = 60$) i due oggetti in movimento vengono fusi in un unico, appartenente alla classe "individuo".

Il criterio di distanza C_{dis} è calcolato tramite la funzione definita per intervalli e rappresentata in figura 5.6.

Per il criterio C_{fus} si utilizza invece l'algoritmo seguente:

r_f = *fusione*(r_i, r_j);
Se $D(r_f, '' individuo'') \geq MAX(D(r_i, '' individuo''), D(r_j, '' individuo''))$
    allora $C_{fus} = D(r_f, '' individuo'')$;
altrimenti
    $C_{fus} = 0$;

5.4.1 L'algoritmo di fusione degli oggetti in movimento

L'algoritmo di fusione si basa sulla costruzione di una tabella, contenente per ogni coppia di oggetti in movimento (r_i, r_j) il valore del coefficiente di raggruppamento:

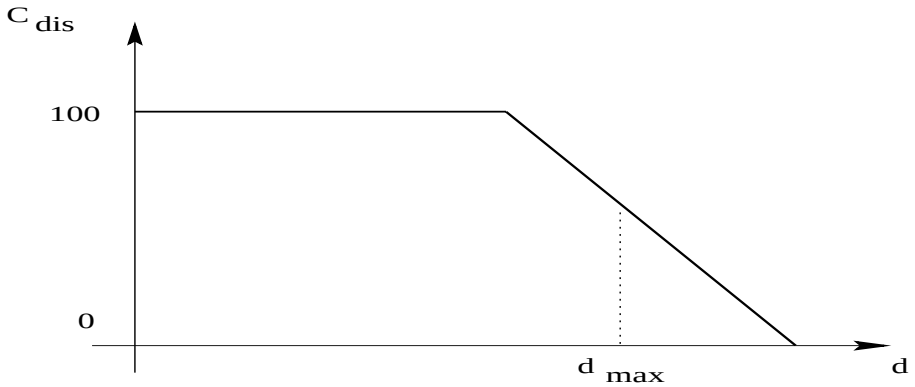


Figura 5.6: *La funzione utilizzata per il calcolo di* C_{dis}.

	oim 1	$\cdots$	*oim n*
oim 1	$C_r(r_1, r_1)$	$\cdots$	$C_r(r_1, r_n)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
oim n	$C_r(r_n, r_1)$	$\cdots$	$C_r(r_n, r_n)$

L'algoritmo di raggruppamento è quindi il seguente:

(1) Costruzione della tabella dei coefficienti di raggruppamento
(2) fine := 0
(3) Finché (fine $\neq$ *0)*
(4) Scelta del coefficiente massimo della tabella $C_{r_{max}} = C_r(r_i, r_j)$
(5) Se $C_{r_{max}} > s_{raggruppamento}$
(5.1) r_i *e* r_j *vengono sostituiti da* $r_f = fusione(r_i, r_j)$
(5.2) Aggiornamento della tabella dei coefficienti di raggruppamento;
(6) Altrimenti
fine := 1;

L'immagine 5.7 mostra invece un risultato dell'algoritmo di fusione.

Nel capitolo successivo si vedrà come gli oggetti in movimento classificati e raggruppati dal presente modulo saranno oggetto, se individui, dell'inseguimento da parte dell'opportuno modulo, in modo da poter ricostruire il loro moto nell'arco di una sequenza di immagini.

Figura 5.7: *La figura a sinistra mostra le differenti regioni in movimento nelle quali sono scomposte le immagini delle due persone. La figura di destra mostra il risultato ottenuto dopo l'algoritmo di fusione: si hanno solo due oggetti in movimento corrispondenti alle due figure.*

Capitolo 6

L'algoritmo di inseguimento

6.1 Introduzione

Il modulo di inseguimento spazio-temporale degli individui riceve in ingresso gli oggetti in movimento (indicati nel seguito con *oim*) e restituisce come uscita degli *individui*. Un *individuo* è una struttura dati che descrive una persona che si sta muovendo nella scena, a partire dal momento del suo ingresso nell'ufficio fino al momento in cui ne esce (cf il paragrafo 6.3.1). Questa persona origina, nelle immagini della sequenza video, una serie di oggetti in movimento; per semplificare il problema, per il momento si ipotizzarà che ogni persona ripresa dalla videocamera corrisponda ad un unico oim in ogni immagine (in altri termini, stiamo ipotizzando che il modulo di fusione esegua perfettamente il suo compito senza possibilità di errore) (la persona A corrisponde dunque all'oggetto in movimento a_1 all'istante t, a_2 all'istante $t+1$ mentre la persona B corrisponde agli oggetti in movimento b_1 a t, b_2 a $t+1$ e così via). Lo scopo di questo modulo è mettere in relazione tra di loro gli oggetti in movimento a_i e, a loro volta, i b_i, creando quindi come dato d'uscita le strutture *individui* corrispondenti alle persone A e B.

Gli individui sono le descrizioni necessarie (dati d'ingresso) per il riconoscimento degli scenari. La figura 6.1 mostra il modulo d'inseguimento nel sistema d'interpretazione.

6.2 L'approccio scelto: il ritardo T

L'approccio scelto per la realizzazione del modulo d'inseguimento si basa su di un'ipotesi: inseguire gli individui con un ritardo T rispetto ai dati d'ingresso. Questo tempo T di differenza tra gli ingressi e le uscite è utilizzato per consolidare, confermare o correggere, utilizzando i dati ricevuti tra $t-T$ e t , le scelte effettuate agli istanti precedenti. Ciò permette di correggere le decisioni che, corrette nell'immediato, si rivelano false con l'evolvere della situazione.

La capacità di modificare le scelte non corrette cresce al crescere di T; parallelamente cresce il ritardo con cui è possibile ottenere in uscita gli individui, e quindi anche il ritardo tra l'istante in cui si verifica un evento e l'interpretazione di questo evento; questo inconveniente si rivela più o meno fastidioso a seconda delle diverse applicazioni: nel caso della videosorveglianza applicata al rilevamento di atti vandalici, o al rilevamento di situazioni di pericolo (videosorveglianza delle stazioni della metropolitana), è importante che il ritardo T resti limitato, per poter intervenire rapidamente; l'applicazione MediaSpace pone invece meno vincoli, dato sì che un ritardo più pronunciato è ugualmente accettabile.

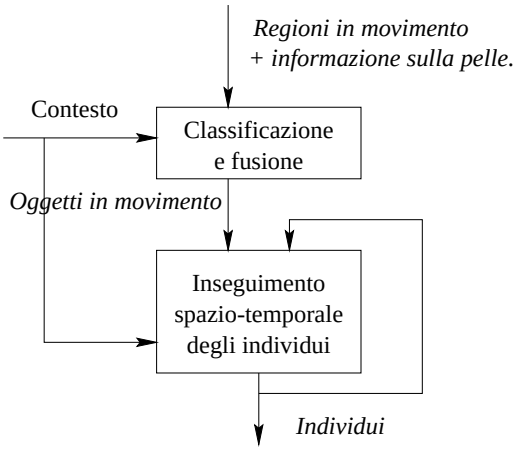


Figura 6.1: *Il modulo d'inseguimento nel sistema d'interpretazione.*

6.3 Il principio

6.3.1 Le strutture dati utilizzate

L'algoritmo gestisce due strutture dati principali:

- le traiettorie;
- gli individui.

6.3.1.1 Le traiettorie

Una traiettoria corrisponde ad un possibile percorso di un individuo nella scena. Per ogni individuo possono esistere molteplici traiettorie (cioé differenti cammini sul grafo degli oggetti in movimento). Una traiettoria può anche corrispondere ad un rumore che si ripete nella sequenza. Lo scopo dell'algoritmo è scegliere per ogni individuo presente nella scena la miglior traiettoria possibile (cioé l'effettiva traiettoria descritta dall'individuo nel suo movimento).

La struttura traiettoria è interna all'algoritmo. Essa si compone essenzialmente di una serie di oggetti in movimento ad istanti di tempo successivi. La figura 1.5 illustra la relazione tra oggetti in movimento, traiettorie ed individui; la figura 6.4 mostra una traiettorie sull'immagine. Qualunque serie temporale di oggetti in movimento che soddisfa i due vincoli seguenti costituisce una traiettoria:

1. ad ogni istante, ogni traiettoria contiene al massimo un oim. Per limitare la complessità dell'algoritmo di calcolo delle traiettorie, si ipotizza che una persona origini al massimo un oggetto in movimento in ogni immagine della sequenza. Questa ipotesi è in effetti falsa, a causa degli errori del modulo di classificazione e fusione; ciononostante questi errori non sono troppo fastidiosi, visto che è comunque possibile associare all'individuo, all'istante t, solamente uno degli oggetti in movimento che costituiscono la sua immagine (e che sono stati rilevati), portando comunque a termine in modo corretto l'inseguimento spazio-temporale; la figura 6.3 mostra un esempio di questa situazione;

2. due oggetti in movimento temporalmente successivi devono essere compatibili dal punto di vista spaziale (cioé non possono essere troppo lontani l'uno dall'altro).

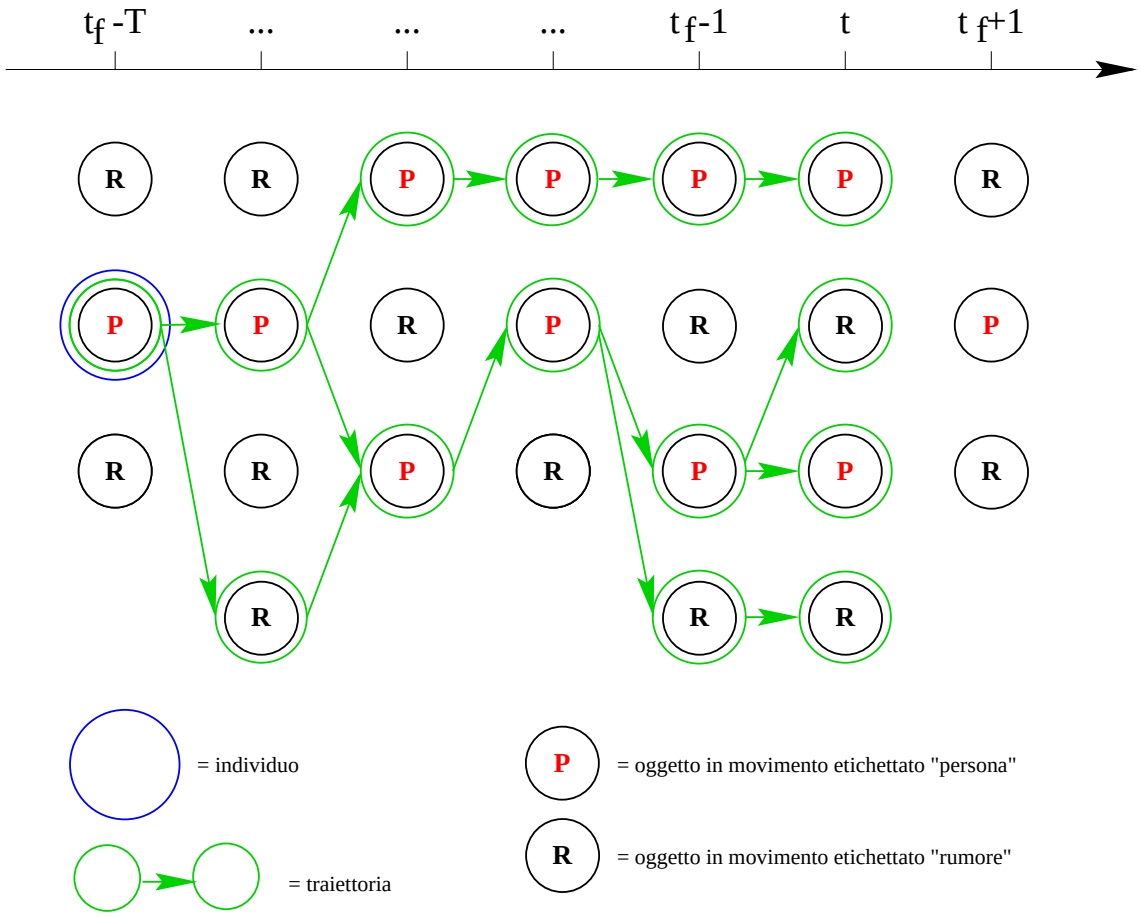


Figura 6.2: *Questa figura rappresenta il grafo degli oggetti in movimento come si presenta prima dell'elaborazione degli oggetti mobili rilevati all'istante* $t+1$*. In ascissa sono riportati i diversi istanti, in ordinata gli oggetti in movimento rilevati nell'immagine ad ogni istante. La lunghezza della finestra temporale considerata è* T *e l'ultima colonna a destra mostra gli oggetti in movimento rilevati all'istante* $t+1$ *e che devono ancora essere elaborati. In questo esempio si nota un individuo ed una serie di traiettorie (7) a lui associate. Il ritardo* T *permette di scegliere la traiettoria ideale per l'individuo, in quanto esse sono completamente determinate e conosciute su tutta la lunghezza* T*. Alcune traiettorie comprendono degli oggetti in movimento classificati "rumore", mentre altre sono costituite solamente da oggetti in movimento classificati "persona".*

Lo scopo del calcolo delle traiettorie è di considerare tutti gli spostamenti possibili di una persona durante l'intervallo temporale T considerato. Si potrebbe calcolare a priori tutte le combinazioni degli N oggetti in movimento presenti ad ogni istante t nella finestra temporale. Questo calcolo è di complessità $O(N,T) = N^T$ ed inoltre la maggior parte di queste combinazioni non ha alcun significato fisico, visto che rappresentano delle traiettorie che legano oggetti in movimento in parti opposte dell'immagine. Da questa considerazione deriva l'introduzione del vincolo sulla distanza tra due oggetti in movimento successivi.

Si utilizza un *coefficiente di qualità* per scegliere quale traiettoria (nel senso di struttura dati, o anche di ipotetico percorso) corrisponde meglio al percorso (reale) di un individuo. Questo coefficiente permette di classificare le traiettorie ognuna nei confronti delle altre

(a) Gli oggetti in movimento

(b) Gli individui (rettangolo blu)

Figura 6.3: *La persona sulla destra è rilevata tramite tre diversi oggetti in movimento (figura (a)). La struttura "individuo", all'uscita dal modulo d'inseguimento, (figura (b)) considera un solo oggetto in movimento (quello con la miglior corrispondenza con il modello d'essere umano) e la persona viene inseguita correttamente.*

(cf il paragrafo 6.5.2.2) ed esprime:

- l'aderenza di ogni oim che costituisce la traiettoria al modello di essere umano. Il modello di essere umano è descritto tramite:
  - un etichetta equivalente all'etichetta apposta sugli oim dal modulo di classificazione;
  - le dimensioni medie del rettangolo circoscrivente l'oim;
  - il rapporto medio tra l'altezza e la larghezza di questo rettangolo.

(a) L'oggetto in movimento

(b) La traiettoria, successione nel tempo di oggetti in movimento

Figura 6.4: *La figura (a) mostra una situazione tipica di creazione di una traiettoria: una persona è appena entrata nell'ufficio. La figura (b) mostra un esempio di traiettoria associata (sono rappresentati solo alcuni oggetti in movimento).*

Queste sono quindi le grandezze sulle quali si misura l'aderenza dell'oggetto in movimento al modello di essere umano;

- la reciproca compatibilità dei differenti oim costituenti la traiettoria, dal punto di vista delle dimensioni del rettangolo circoscrivente (in altri termini, si vuole quantificare l'omogeneità di dimensioni degli oggetti mobili che costituiscono una traiettoria);
- l'aderenza della traiettoria al modello di movimento, inteso come successione temporale di oim. Il modello di movimento è definito tramite:
  - una direzione media (a priori qualsiasi, poi inizializzata sulla base degli oim);
  - una distanza media tra due oim temporalmente successivi.

Le informazioni contenute nella struttura *traiettoria* sono quindi:

- un numero che identifica in modo univoco ogni traiettoria;
- una lista di lunghezza T contenente gli oim che compongono la traiettoria;
- l'istante t_s di creazione;
- l'istante t_c dell'ultimo aggiornamento;
- l'altezza e la larghezza medie degli oim che compongono la traiettoria (espresse in pixels);
- la distanza media tra due oim successivi (in pixels);
- la direzione media (in gradi, misurata sul piano dell'immagine) della traiettoria;
- il coefficiente di qualità Q della traiettoria;
- una prima etichetta che può assumere i valori: GRUPPO, SEMIGRUPPO, INDIVIDUO o INDETERMINATO. Questa etichetta specifica che tipi di oim costituiscono, in media, la traiettoria: gli oim possono rappresentare una sola persona (INDIVIDUO) o più persone (GRUPPO). In base al numero di oggetti in movimento appartenenti a queste due classi la traiettoria può essere di tipo GRUPPO, SEMIGRUPPO e via dicendo (cf il paragrafo 6.5.2.4 per ulteriori dettagli);
- l'indicazione del tipo di movimento descritto dalla traiettoria: ENTRANTE (nell'ufficio), INTERNO, USCENTE (dall'ufficio), USCITO, INDETERMINATO;
- gli individui ritenuti compatibili con questa traiettoria (cf il paragrafo 6.3.1.3);
- le altre traiettorie che incrociano la traiettoria presente, oppure che le si sovrappongono. Due traiettorie si incrociano se condividono un medesimo oggetto in movimento ad un istante di tempo t. Esse si sovrappongono se gli oggetti in movimento condivisi sono più di uno, ad istanti di tempo differenti.

6.3.1.2 Gli individui

Gli individui sono strutture dati corrispondenti a persone reali che si stanno muovendo nella scena; queste strutture dati sono il risultato dell'algoritmo di inseguimento. Un individuo è creato quando è disponibile una traiettoria *completa* (quindi composta da T oim) di qualità sufficientemente elevata (cf il paragrafo 6.5.3). L'individuo è quindi aggiornato utilizzando ad ogni istante t la miglior traiettoria tra quelle a lui associate (cf il paragrafo 6.5.4), fino ad essere distrutto quando la traiettoria migliore a lui associata si esaurisce (cioé quando la persona fisica che corrisponde alla struttura individuo e che ne è all'origine è uscita dalla scena). Le informazioni contenute nella struttura individuo sono:

- un numero che l'identifica in modo univoco;
- il primo oggetto in movimento (il più vecchio, relativo all'istante $t - T$) della traiettoria che è stata utilizzata per aggiornarlo;
- la traiettoria che è stata utilizzata per aggiornarlo (cf il paragrafo 6.5.4);
- l'istante t_s di creazione;
- l'istante t_c dell'ultimo aggiornamento;
- l'altezza e la larghezza medie degli oim associati (in pixels);
- la distanza media tra due oim successivi (in pixels);
- la direzione media (in gradi, misurata sul piano dell'immagine) della traiettoria;
- lo stato dell'individuo: INSEGUITO (dal modulo), USCITO (dalla scena), ENTRANTE, RAGGRUPPATO (con un altro individuo), SOSPESO (quando il modulo non è più in grado di inseguire correttamente la persona) (cf il paragrafo 6.5.2.4);
- le traiettorie associate all'individuo (cf il paragrafo 6.3.1.3);
- gli altri individui associati al presente (cf il paragrafo 6.3.1.3);

6.3.1.3 Le relazioni tra piste ed individui: la compatibilità e l'associazione

Definiamo in questa sezione due concetti che saranno utilizzati nel seguito; una traiettoria si dice *associata* ad un individuo se l'oim dell'individuo è anche l'oim più vecchio che costituisce la traiettoria; in altri termini, si tratta delle traiettorie che rappresentano percorsi futuri possibili dell'individuo. Per esempio, nella figura 6.7 la traiettoria A è associata all'individuo rappresentato. Simmetricamente, l'individuo si dice allora *compatible* con la traiettoria in questione. Nella stessa immagine, l'individuo è compatibile con la traiettoria A. Utilizzando una formula:

$$\text{individuo} \underset{\text{associazione}}{\overset{\text{compatibilità}}{\rightleftharpoons}} \text{traiettoria}$$

Allo stesso modo si definisce l'associazione **tra due individui**: l'individuo A ha come miglior traiettoria P_a; l'individuo B ha come miglior traiettoria P_b; se P_a e P_b si incrociano, o si sovrappongono, A e B si dicono allora *associati*.

6.4 Le problematiche dell'inseguimento

Numerose situazioni, tipiche delle attività umane che si svolgono in un ufficio, sono in grado, se non correttamente gestite, di provocare degli errori di inseguimento. Nella sezione che segue si descriverà qualche tipico caso d'errore al quale vanno incontro i moduli d'inseguimento meno sofisticati. Nella sezione 6.5 si spiegheranno le tecniche utilizzate per guadagnare in robustezza a fronte di queste situazioni.

6.4.1 L'occultazione dinamica (incrocio di persone)

Quando due strutture *individuo* si incrociano nella sequenza, le due traiettorie utilizzate dagli individui condividono un certo numero di oim, prima di separarsi nuovamente (il numero di questi oggetti in movimento condivisi è proporzionale alla durata del fenomeno di incrociamento). Una delle due traiettorie ha un coefficiente di qualità più elevato dell'altra. Di conseguenza, le due strutture individuo associate alle due traiettorie possono essere aggiornate utilizzando entrambe la traiettoria con il coefficiente di qualità più elevato; il risultato è che i due individui inseguono entrambi una delle persone fisiche, mentre l'altra non risulta inseguita da nessun individuo. Questa situazione corrisponde ad un errore di scelta (**scelta della traiettoria errata**); l'immagine 6.5 illustra questa situazione.

L'errore ha come conseguenza immediata che nel momento in cui una delle due persone esce dalla scena, il numero di individui inseguiti è in ogni caso scorretto (ricordiamo che agli occhi del modulo di inseguimento l'esistenza di una struttura individuo è associata alla presenza di una corrispondente persona fisica nella scena. Nel nostro caso avremmo o due o nessun individuo, corrispondenti o a due persone nell'ufficio, o a nessuna; situazione palesemente errata, visto che in scena resta una ed una sola persona).

Un secondo possibile errore, meno importante rispetto alla scelta di una traiettoria errata, è lo **scambio di traiettorie**; la struttura individuo A è associata alla traiettoria

(a) La situazione prima dell'errore

(b) Scelta della traiettoria errata: la stessa persona per due individui

Figura 6.5: *Le due strutture individuo inseguono correttamente le due persone nella figura (a); la figura (b) mostra che i due individui hanno scelto entrambi la traiettoria che si prolunga sulla persona di sinistra (cf anche l'ingrandimento nell'angolo in basso a destra della figura (b), che mostra i numeri 1 e 2 sovrapposti). Ne risulta che la persona di destra non è più inseguita da nessuna struttura individuo.*

(a) La situazione prima dell'errore (b) Scelta della traiettoria scorretta: la sedia

Figura 6.6: *le due strutture individuo seguono correttamente le due persone nella figura (a); la figura (b) mostra una traiettoria corrispondente allo spostamento di una sedia rispetto all'immagine di sfondo. La struttura individuo di destra è portata a scegliere questa traiettoria. Ne risulta che la persona di destra non è più inseguita da nessuna struttura individuo, mentre la sedia è associata ad un individuo.*

1, che, per esempio, si dirige verso destra. La struttura individuo B è associata alla traiettoria 2, che si dirige verso sinistra. In seguito ad un incrocio, nella realtà, delle due persone corrispondenti agli individui A e B, anche le piste 1 e 2 si incrociano e può succedere che A scelga la traiettoria 2 mentre B sceglie la traiettoria 1. Questo secondo errore è meno importante del primo, dato che non affligge il numero di individui gestiti dal modulo d'inseguimento. Tuttavia questo errore può essere la causa di problemi a livello del riconoscimento degli scenari, in quanto l'errore comporta uno scambio d'identità a livello degli individui A e B.

6.4.2 Lo spostamento di un oggetto dell'arredo

Qualora un oggetto dell'arredo (una porta, una sedia...) venga spostato, la differenza introdotta rispetto all'immagine di riferimento è interpretata come un oggetto in movimento, che viene rilevato nella posizione in cui si trova l'oggetto spostato. Questo oim può dare origine ad una traiettoria con un buon coefficiente di qualità (perché le dimensioni dell'oggetto non cambiano da un'immagine all'altra e possono corrispondere alla dimensioni del modello di essere umano); l'errore consiste allora nell'assegnare questa traiettoria ad uno degli individui esistenti; si avrebbe allora una struttura individuo che insegue un oggetto del mobilio invece di seguire la persona reale. Si tratta di un errore di scelta (**scelta della traiettoria scorretta**); l'immagine 6.6 illustra questa situazione.

6.5 L'algoritmo

L'algoritmo d'inseguimento si divide in cinque grandi routines; ciascuna di esse è descritta in una sezione dedicata. Le operazioni svolte dalle cinque routines sono:

1. l'inizializzazione delle traiettorie;

2. l'aggiornamento delle traiettorie;

3. l'inizializzazione degli individui;

4. l'aggiornamento degli individui;

5. l'eliminazione delle traiettorie (e degli individui).

6.5.1 L'inizializzazione delle traiettorie

Simmetricamente rispetto a ciò che accade nella realtà (cioè che gli utilizzatori di un ufficio entrano ed escono solo dalla(e) porta(e) dell'ufficio stesso), si ipotizza che ogni nuova traiettoria creata debba avere come primo oggetto in movimento un oggetto che si trovi nella zona corrispondente alla porta d'ingresso (l'ufficio utilizzato come scena nei nostri test ha una sola porta). Ciò equivale a dire che ogni traiettoria deve cominciare dalla porta. Di conseguenza si è definita nella base del contesto una "zona d'ingresso/uscita" (indicata *zona IU* nel seguito) che corrisponde alla porta dell'ufficio. Qualora si rilevi un oim nella zona IU, si crea una nuova traiettoria.

È importante notare che esiste un certo numero di situazioni nelle quali la creazione di una traiettoria si rivela non necessaria. Per esempio, quando una persona sta uscendo dalla scena, essa entra nella zona IU, e dà quindi origine ad una nuova traiettoria. Questa traiettoria non è necessaria, perché la persona che sta uscendo ha già una traiettoria (più corretta, tra l'altro) associata. Ciononostante la creazione di questa seconda traiettoria non pregiudica la correttezza delle operazioni successive, in quanto essa non sarà comunque utilizzata.

La politica adottata dall'algoritmo di creazione delle traiettorie è dunque di crearne una nuova per ogni oim rilevato nella zona IU.

6.5.2 L'aggiornamento delle traiettorie

L'operazione di aggiornamento consiste nel prolungare ogni traiettoria esistente all'istante t con gli oggetti in movimento rilevati all'istante $t + 1$. Come spiegato precedentemente, il numero di traiettorie (cf 6.3.1) che sarebbe necessario gestire se si prendesse in considerazione ogni combinazione traiettoria – oggetto in movimento è proibitivo. Al fine di limitare il numero di traiettorie considerate ad ogni istante, si procede nel modo seguente:

1. si fissa un numero massimo di traiettorie gestite dall'algoritmo, N_{max};

2. queste N_{max} possibilità d'estensione vengono divise in E insiemi; l'algoritmo gestisce un insieme per ogni individuo esistente all'istante t, più un insieme aggiuntivo E_e nel caso ci siano delle piste incomplete (cioé formate da meno di T oggetti in movimento, e di conseguenza non ancora associate a nessun individuo). Ogni insieme E_j raggruppa le traiettorie che sono associate all'individuo I_j (cf il paragrafo 6.3.1.3). L'insieme E_e è previsto per assegnare la stessa probabilità anche ad un nuovo individuo che non è stato ancora creato, ma le cui (future) traiettorie devono essere aggiornate;

3. ogni insieme E_j riceve $N_{max}/E = N_j$ possibilità d'estensione. Lo scopo di questa suddivisione delle estensioni possibili è di assegnare ad ogni individuo la stessa probabilità che le traiettorie a lui associate siano correttamente estese;

4. per ogni insieme E_j, ogni traiettoria $P_{ji} \in E_j$ riceve un numero di estensioni possibili N_{ji} proporzionale al proprio coefficiente di qualità Q_{ji} (sulla somma totale dei coefficienti di qualità di tutte le traiettorie dell'insieme E_j):

$$N_{ji} = sup\left(N_j \cdot \frac{Q_{ji}}{\sum_{k=1}^{l} Q_{jk}}\right)$$

dove si è ipotizzato che l'insieme E_j contenga l traiettorie, indicate P_{ji}, con $1 < i < l$. Q_{ji} è il coefficiente di qualità della traiettoria P_{ji}. L'operatore $sup(\cdot)$ arrotonda l'argomento $\in \mathbb{R}$ all'intero superiore.

Inoltre ogni traiettoria può generare al massimo una estensione per ogni nuovo oim; se i nuovi oggetti sono in numero pari a O_{max}, le estensioni possibili per ogni traiettoria sono al massimo O_{max} (cf il paragrafo 6.5.2.1). A seconda del valore di N_{ji}, si possono avere due casi differenti:

   (a) la traiettoria riceve un numero di estensioni possibili $N_{ji} \geq O_{max}$. In questo caso, si considerano tutte e sole le O_{max} estensioni possibili, quindi N_{ji} è uguale a O_{max};

   (b) la traiettoria riceve un numero di estensioni possibili $N_{ji} < O_{max}$. È dunque necessario scegliere N_{ji} estensioni tra le O_{max} possibili. Si estendono allora le N_{ji} traiettorie con gli oggetti in movimento che offrono il miglior coefficiente di qualità.

5. alcune coppie P_j – nuovo oggetto in movimento $O_k(t)$ possono dar luogo ad estensioni con un coefficiente di qualità nullo. In effetti non vengono calcolate che le estensioni aventi significato fisico; la distanza 2D sul piano dell'immagine tra l'oggetto in movimento $O_k(t)$ e l'ultimo oggetto in movimento $O_j(t-1)$ integrato nella traiettoria deve essere inferiore ad un valore soglia:

$$distance(O_k(t); O_j(t-1)) \leq \frac{\overline{H}_M + \overline{W}_M}{2}$$

dove $\overline{H}_M$ e $\overline{W}_M$ sono l'altezza e la larghezza medie degli oim che compongono la traiettoria.

6. l'operatore $sup(\cdot)$ implica che ogni traiettoria riceva come minimo una possibilità d'estensione. Così facendo, se si sommano tutte le possibilità d'estensione attribuite ad ogni traiettoria si supera la quantità massima ammessa:

$$\sum_{j=1}^{E} \sum_{i=1}^{l_j} N_{ji} > N_{max}.$$

La soluzione consiste allora nell'estendere, per ogni insieme E_j, solo le traiettorie P_{ji} che offrono il miglior coefficiente di qualità, fino a raggiungere le N_{ji} estensioni permesse. Così facendo è impossibile superare le N_{max} estensioni possibili.

6.5.2.1 L'estensione tramite l'oggetto in movimento "non rilevato"

Per correggere eventuali errori di non rilevamento (si ha un errore di non rilevamento quando l'oim corrispondente ad una persona fisica non è rilevato), si è prevista la possibilità di estendere le traiettorie utilizzando un oim virtuale, chiamato "non rilevato". Questo oggetto in movimento virtuale non ha alcun significato fisico, poiché esso non esiste sull'immagine. Esso rimpiazza l'autentico oggetto in movimento che avrebbe dovuto essere rilevato, ma che non lo è stato a causa di un errore.

Questo oim "non rilevato" va ad aggiungersi agli O_{max} oggetti in movimento rilevati al tempo $t+1$ e rappresenta una ulteriore possibilità d'estensione per ogni traiettoria esistente. Il numero massimo d'estensioni per ogni traiettoria s'innalza così a $O_{max}+1$.

6.5.2.2 Il coefficiente di qualità di una traiettoria

Il coefficiente di qualità di una traiettoria è la grandezza che permette, in ogni istante, di decidere quale traiettoria sarà scelta per un individuo tra tutte quelle a lui associate. Il coefficiente di qualità esprime:

- l'aderenza dell'oggetto in movimento k rilevato all'istante $t+1$ ($M_k(t+1)$) al modello di essere umano. Questa aderenza è calcolata sotto la forma di un coefficiente $C_1 \in [lim_{inf} = 0.5; 1]$; il calcolo di questo coefficiente si basa su tre differenti valori, ognuno dei quali esprime la compatibilità tra una delle caratteristiche del modello e la corrispondente caratteristica dell'oggetto in movimento:
  1. le dimensioni (altezza e larghezza) di $M_k(t+1)$ (questa compatibilità è calcolata sotto forma di un coefficiente C_{11});
  2. il rapporto altezza/larghezza di $M_k(t+1)$ (C_{12});
  3. la classe dell'oggetto in movimento $M_k(t+1)$ (cf il paragrafo 5) (C_{13});
- la compatibilità tra l'oim $M_k(t+1)$ e gli oggetti ($M_i(\leq t)$) che compongono la traiettoria P_i; il coefficiente che esprime questa compatibilità è $C_2 \in [lim_{inf} = 0.5; 1]$. Esso prende in considerazione:
  1. la compatibilità tra l'altezza e la larghezza di $M_k(t+1)$ e le medie di altezza e larghezza degli oim che compongono la traiettoria (sotto forma di un coefficiente C_{21});
  2. la direzione definita da $M_k(t+1)$ e $M_i(t)$ rispetto alla direzione media definita dagli altri $M_i(< t)$ (C_{22});
- la distanza tra i due oim $M_k(t+1)$ e $M_i(t)$ (cf il paragrafo 6.5.2), sotto forma di un coefficiente $C_3 \in [0; 1]$;
- la qualità della traiettoria P_i all'istante t prima dell'estensione, $Q_i(t)$.

Il coefficiente di qualità (Q_{new}) della traiettoria dopo l'estensione (P_{new}) vale:

$$Q_{new} = C_1 \cdot C_2 \cdot C_3 \cdot Q_i$$

Ogni coefficiente C_i (salvo C_3) è ottenuto come media pesata dei diversi sotto-criteri:

$$C_l = \sum_{k=1}^{N_{crit}} w_{lk} \cdot C_{lk}, \quad w_{lk} \in [0; 1]$$

I coefficienti C_{lk} e C_3 sono calcolati considerando una gaussiana (normalizzata per avere il valore 1 nel punto di media) centrata sul valore atteso (in corrispondenza del quale si ottiene quindi il valore massimo, 1). Nel caso dei C_{1k} la varianza è calcolata a partire da verifiche sperimentali (in generale è posta uguale alla metà della media), mentre per i C_{2k} la varianza è proporzionale alla media delle grandezze caratteristiche degli oim che costituiscono la traiettoria (e dunque varia al variare delle caratteristiche della traiettoria).

Infine, per il coefficiente C_{22} (compatibilità dal punto di vista della direzione) la varianza è uguale à 216.2 (calcolato in modo tale che la gaussiana valga $lim_{inf} = 0.5$ per i valori a distanza 180ř dal valor medio).

Per evitare di ridurre troppo la qualità di una traiettoria a causa di un solo cattivo rilevamento (per esempio un oggetto in movimento che non ha le dimensioni corrette, cioé simili a quelle codificate nel modello di essere umano), il coefficiente lim_{inf} fissa a 0.5 il limite inferiore di variazione di C_1 e di C_2, modellizzando con il valore 0.5 la mancanza di informazione. Confrontare a tal proposito il punto 5 del paragrafo 6.5.2.

Nel caso in cui sia impossibile calcolare uno dei coefficienti C_{lk} (per esempio qualora i due oggetti in movimento $M_k(t+1)$ e $M_i(t)$ siano sovrapposti e quindi la direzione del segmento $\overline{M_k(t+1)M_i(t)}$ sia indefinita), il valore di codesto coefficiente è fissato a priori a 0.5. Un caso particolare si ottiene quando si prolunga la traiettoria con un oggetto in movimento "non rilevato"; in questo caso, il prodotto $C_1 \cdot C_2 \cdot C_3$ è uguale ad un coefficiente C_{nd} che rappresenta la frequenza di mancato rilevamento dell'oim sul numero totale dei rilevamenti (sperimentalmente si è misurato $C_{nd} = 10^{-3}$).

6.5.2.3 Limitazione del numero di estensioni per le traiettorie non ottime

L'estensione di una traiettoria di qualità elevata tramite un oggetto in movimento che corrisponde male ad un individuo (per esempio, un oim corrispondente ad un rumore) porta alla creazione di una traiettoria ancora di buona qualità. A causa di questa buona qualità, all'istante successivo essa si vedrà assegnare un numero importante di possibilità d'estensione. Allo stesso tempo il numero di estensioni possibili per la altre traiettorie risulterà ridotto, anche se queste ultime rappresentano spostamenti corretti (questo perché potrebbero avere coefficienti di qualità inferiori a quello della traiettoria "errata").

Per evitare questa situazione si assegna una sola possibilità d'estensione a tutte le traiettorie che si sovrappongono ad una traiettoria *migliore* (cioé una traiettoria che sia associata allo stesso individuo ma che abbia coefficiente di qualità inferiore alle altre). La figura 6.7 illustra questa situazione.

6.5.2.4 L'aggiornamento degli attributi di una traiettoria

Stabilito che una traiettoria P_i deve essere prolungata utilizzando un nuovo oggetto in movimento $M_j(t+1)$, è necessario aggiornare gli attributi di questa nuova traiettoria utilizzando le caratteristiche di $M_j(t+1)$. La procedura è la seguente:

- la nuova traiettoria P_{new}, risultato dell'estensione di P_i, riceve un nuovo identificatore, differente da quello di P_i;
- l'oggetto in movimento $M_j(t+1)$ si aggiunge alla lista degli oggetti precedentemente integrati nella traiettoria. Contemporaneamente è necessario eliminare il più vecchio degli oim della traiettoria, $M_i(t-T)$;
- l'istante t_s di creazione di P_i non cambia per P_{new};

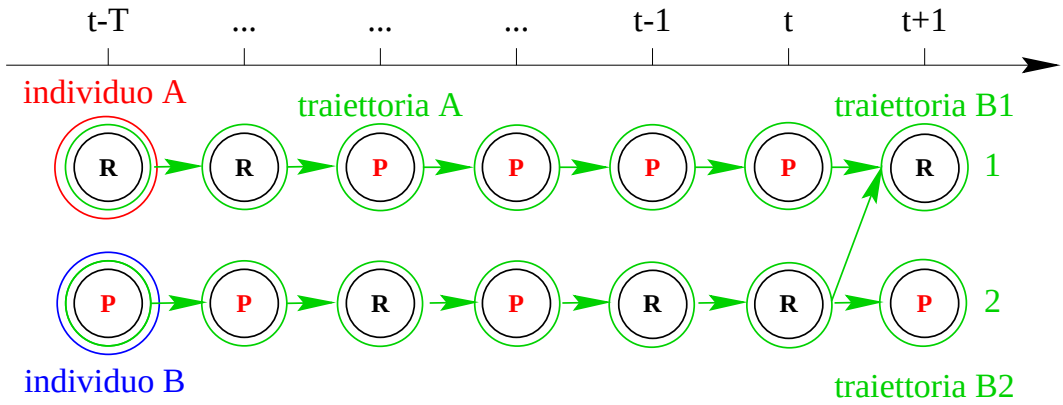


Figura 6.7: *La traiettoria A, rappresentata fino all'istante t, può essere estesa utilizzando i due oggetti in movimento 1 e 2. La traiettoria ha una buona qualità, quindi anche l'estensione A1 (cioé l'estensione di A tramite l'oggetto in movimento 1) sarà caratterizzata da un buon coefficiente di qualità. All'istante successivo si vuole limitare il numero delle estensioni che prolungano la traiettoria A1, visto che A1 è parzialmente sovrapposta ad A2, che possiede coefficiente di qualità migliore e che è associata allo stesso individuo. Riconoscendo questa situazione, l'algoritmo assegna ad A1 una sola possibilità d'estensione.*

- l'istante t_c dell'ultimo aggiornamento è incrementato: diventa $t_c + 1$;
- le medie di altezza e larghezza degli oim sono ricalcolate sostituendo nel calcolo i valori di $M_i(t - T)$ con quelli di $M_j(t + 1)$;
- la distanza tra $M_j(t+1)$ e l'ultimo oim integrato ($M_j(t)$) è utilizzata per aggiornare la media delle distanze tra gli oggetti in movimento successivi;
- se $M_j(t+1)$ e $M_j(t)$ sono sufficientemente lontani (e dunque la loro direzione ha un senso), il vettore individuato dai due centri di gravità di $M_j(t+1)$ e $M_j(t)$ è utilizzato per aggiornare la direzione media della traiettoria P_{new}. Questa direzione media è reinizializzata a 0 ogniqualvolta la distanza tra $M_j(t + 1)$ e $M_j(t)$ è troppo piccola per definire una direzione. Questa situazione è tipica di una persona che arresta il proprio movimento;
- il coefficiente di qualità Q_i diventa Q_{new} (cf il paragrafo 6.5.2.2);
- la classe di $M_j(t+1)$ è presa in considerazione per determinare il tipo di traiettoria; si calcola una media pesata della frequenza delle classi degli oim che compongono la traiettoria; gli oggetti più recenti hanno peso maggiore (il peso è lineare in t, e varia tra 1 per gli oim a $t-T+1$ e T per gli oim a $t+1$). Quindi si assegna alla traiettoria una delle classi:
  - GRUPPO: se la media corrisponde al tipo GRUPPO (ricordiamo che un oim è classificato GRUPPO quando rappresenta più di una persona);
  - SEMIGRUPPO: se la media corrisponde al tipo SEMIGRUPPO (un oim è classificato SEMIGRUPPO quando le sue caratteristiche sono intermedie tra quelle del modello di INDIVIDUO e quelle del modello di GRUPPO);
  - INDIVIDO: se la media è di tipo INDIVIDUO;

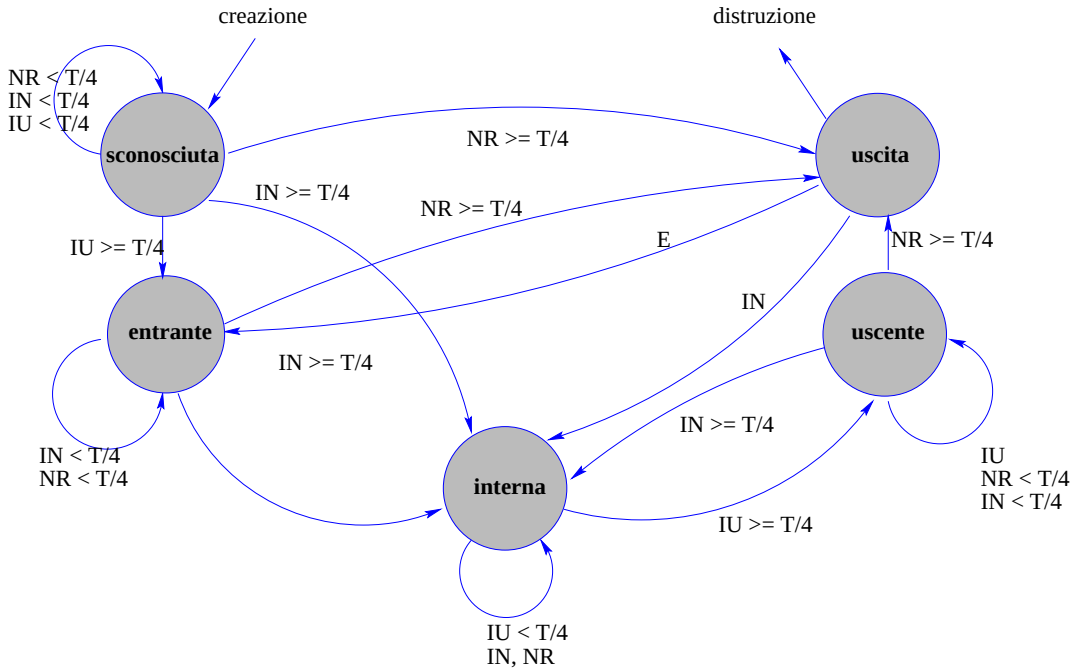


Figura 6.8: *Questa macchina a stati finiti calcola la localizzazione della traiettoria. Cinque sono le localizzazioni possibili: SCONOSCIUTA, ENTRANTE, INTERNA (corrispondente ad una persona che si sta muovendo nella scena), USCENTE (che sta uscendo dall'ufficio), USCITA. I cambiamenti di stato sono calcolati in funzione del numero di volte che i nuovi oggetti in movimento si trovano nella zona IU. Per esempio, I>=3 vuol dire che se i tre ultimi oggetti in movimento si trovano nella zona IU, il cambiamento di stato corrispondente è effettuato. I = numero di oggetti in movimento nella zona IU, IN = numero di oggetti in movimento al difuori della zona IU (cioé all'interno della scena), NR = numero di oggetti in movimento "non rilevati".*

  - INDETERMINATO: se è impossibile stabilirne il tipo (cioè se la media delle classi degli oim ha valore INDETERMINATO).

- La localizzazione della traiettoria è aggiornata considerando le posizioni successive degli oim rispetto alla zona di ingresso/uscita (indicata IU) (infatti gli eventi all'origine dei cambiamenti di stato sono le posizioni dei nuovi oim che prolungano la traiettoria); la figura 6.8 mostra la macchina a stati finiti che calcola la localizzazione della traiettoria;

- gli individui compatibili con questa traiettoria non sono aggiornati. Questa informazione è gestita dalla procedura di aggiornamento delle strutture individuo (cf il paragrafo 6.5.4);

- dallo stesso modo, la lista delle altre traiettorie che incrociano o si sovrappongono a P_{new} non è aggiornata.

Se la traiettoria è estesa utilizzando un oim "non rilevato", alcuni attributi non vengono aggiornati: le dimensioni, la distanza e la direzione della nuova traiettoria.

6.5.3 L'inizializzazione degli individui

La routine d'inizializzazione delle strutture individuo consiste nell'analizzare l'insieme delle traiettorie esistenti; se una di queste traiettorie soddisfa tutte le condizioni necessarie, un nuovo individuo viene creato (questo nuovo individuo sarà quindi associato a questa traiettoria). Le condizioni da verificare sono:

1. la traiettoria deve essere completa (composta cioé da T oim);
2. la traiettoria deve cominciare nella zona IU (ciò equivale ad imporre che il primo oggetto in movimento, il più vecchio, si trovi in questa zona);
3. la traiettoria deve terminare al di fuori della zona IU (il suo oim più recente deve trovarsi al di fuori della zona IU);
4. la traiettoria non si deve sovrapporre (né deve incrociare) una traiettoria scelta come migliore da un individuo.

La condizione 1 è stata introdotta per evitare di creare un individuo senza aver atteso il ritardo di T istanti che assicura che questa struttura corrisponda veramente ad una persona fisica che si sta muovendo nella scena.

La condizione 2 verifica che la traiettoria per la quale si sta creando un individuo rappresenti effettivamente un percorso logico di una persona fisica, che entra dalla porta (parallelamente la traiettoria deve cominciare nella zona IU) (cf il paragrafo 6.5.1).

La condizione 3 è introdotta per evitare di confondere un individuo con un rumore che è rilevato nella zona IU e che vi resta a lungo. Un esempio di situazione di questo tipo è l'apertura/chiusura della porta: essa genera un rumore costante nella zona IU che corrisponde ad un oim. Una traiettoria è allora creata, ma nessun individuo sarà creato.

Infine la condizione 4 è prevista per evitare di creare una nuova struttura individuo quando una traiettoria generata corrisponde ad una persona già rilevata (ed inseguita tramite una struttura individuo), per esempio che si sta muovendo in prossimità della zona IU. È per esempio il caso in cui l'ombra della persona già rilevata si proietta sul muro nella zona IU, generando un oggetto in movimento. Una traiettoria P è creata. Questa traiettoria si mischia quindi agli oggetti in movimento della persona che si trova prossima alla zona IU (ma che ne resta all'esterno) e insegue la persona quando questa si allontana dalla zona IU. Abbiamo allora una traiettoria (P) che comincia nella zona IU, che termina all'esterno e che è completa; ciononostante sarebbe un errore creare un individuo. Il criterio che permette all'algoritmo di riconoscere questa situazione è il fatto che la traiettoria P si sovrapponga ad una traiettoria scelta come *migliore* da un individuo già esistente (quello che segue la persona) (infatti P utilizza gli oim dell'individuo).

6.5.4 L'aggiornamento degli individui

La routine di aggiornamento degli individui consiste nello scegliere quale traiettoria (sviluppantesi tra $[t-T;t]$) è la migliore per ogni individuo esistente all'istante $t-T$. In seguito, l'aggiornamento consiste nell'assegnare all'individuo l'oim più vecchio della traiettoria, $M(t-T)$. La figura 6.9 illustra questa situazione.

L'aggiornamento comincia con il calcolo di quattro coefficienti che esprimono le mutue relazioni tra traiettorie ed individui, tra traiettoria e traiettoria e tra individuo ed individuo:

- D_1: il coefficiente di compatibilità tra I_q e P_k;

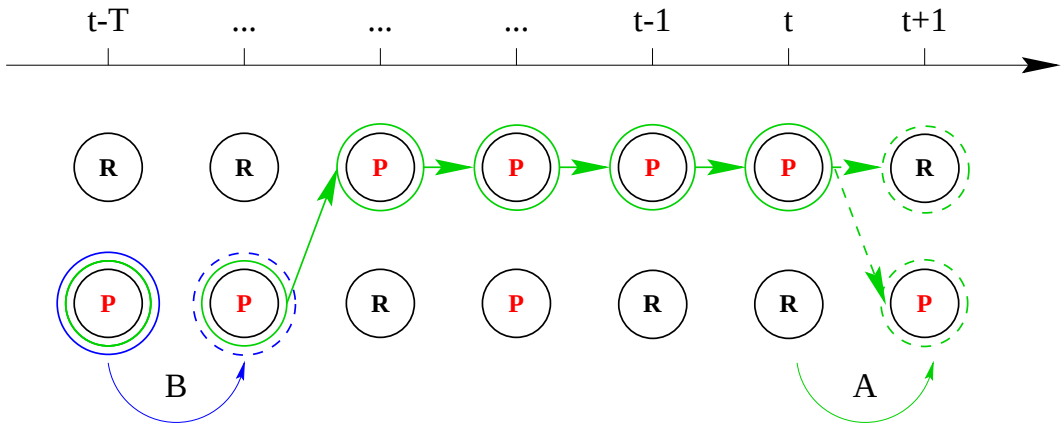


Figura 6.9: *In questa figura si possono notare i diversi istanti di tempo in cui vengono effettuati gli aggiornamenti. La freccia A mostra l'aggiornamento delle traiettorie, agli istanti* t *e* $t+1$*. La freccia B mostra l'aggiornamento degli individui, agli istanti* $t-T$ *e* $t-T+1$*.*

- D_2: $D_2(P_q; P_r)$ è utilizzato per quantificare la sovrapposizione tra le due traiettorie P_q e P_r; esso decresce esponenzialmente al crescere del numero di oggetti in movimento comuni:
$$D_2(P_q; P_r) = e^{-\left(\frac{N_{ov}}{K}\right)^2}$$
dove N_{ov} è il numero di oggetti in movimento (che non siano però classificati come *gruppo* [1]) comuni a P_q e P_r e K è un coefficiente di normalizzazione.
- D_3: $D_3(I_l; I_m)$ assume valori 0 o 1 a seconda che i due individui I_l e I_m siano associati;
- D_4: una coppia individuo-traiettoria è classificata *persona* o *gruppo* a seconda della classe degli oggetti in movimento che costituiscono la traiettoria e a seconda che l'individuo sia *isolato* (cioè se l'oggetto in movimento che lo costituisce non è comune a nessun altro individuo).

Dopo aver calcolato queste relazioni, l'algoritmo come prima cosa aggiorna gli individui che non sono associati a nessun altro (cioé con $D_3 = 0$) utilizzando, per ognuno d'essi, la traiettoria che offre il miglior coefficiente di compatibilità. Quindi, per calcolare in modo globale ed ottimale l'associazione tra le traiettorie e i restanti individui, l'algoritmo calcola gli insiemi V_h degli individui associati ($D_3 = 1$). All'interno di ognuno di questi insiemi V_h l'algoritmo sceglie la coppia $I_j - P_k$ con il miglior coefficiente di compatibilità (D_1). Questa coppia è aggiornata e, se l'individuo è classificato come *persona* ($D_4 = persona$), il coefficiente di compatibilità delle altre coppie $I_l - P_m$ è diminuito se la traiettoria P_m si sovrappone a P_k (D_2):

$$D_{1new}(I_l; P_m) = D_{1old}(I_l; P_m) \cdot D_2(P_k; P_m).$$

Infine l'algoritmo seleziona la seconda miglior coppia $I_{j'} - P_{k'}$ con il secondo miglior coefficiente di compatibilità (D_1). L'algoritmo è iterato finché, per ogni insieme V_h, tutte le coppie dell'insieme (contenente gli individui associati) sono state processate.

[1] Si considera corretto il caso in cui un oggetto in movimento etichettato *gruppo* sia comune a due traiettorie, in quanto la definizione stessa prevede che un oggetto mobile *gruppo* sia l'immagine di più persone contemporaneamente.

6.5.4.1 Il calcolo della compatibilità tra traiettoria ed individuo

Questo calcolo di compatibilità è simile a quello del coefficiente di qualità di una traiettoria. La compatibilità tra l'individuo I_{ji} e la traiettoria P_{ji} si basa su:

- la compatibilità tra l'altezza e la larghezza di I_{ji} (cf il paragrafo 6.3.1.2) e l'altezza e la larghezza dell'oim che costituisce P_{ji} all'istante $t-T+1$ ($M_{ji}(t-T+1)$). Questa compatibilità si esprime sotto la forma di un coefficiente $D_1 \in [lim_{inf}; 1]$;
- la compatibilità tra la direzione definita dai centri di gravità dell'oggetto in movimento di I_{ji} e dell'oggetto in movimento $M_{ji}(t-T+1)$, e la direzione media di I_{ji}. Il coefficiente che esprime questa compatibilità è $D_2 \in [lim_{inf}; 1]$;

Il coefficiente di compatibilità Q_{ji} è ottenuto come media pesata di D_1 e di D_2:

$$Q_{ji} = w_1 \cdot D_1 + w_2 \cdot D_2$$

D_1 e D_2 sono calcolati utilizzando una gaussiana (normalizzata per avere il valore massimo = 1), centrata sul valore medio delle caratteristiche dell'individuo. Per D_1 la varianza è uguale alla media (quindi alla altezza/larghezza di I_{ji}) mentre per D_2 la varianza vale 216 (in modo tale che la gaussiana restituisca $lim_{inf} = 0.5$ per un valore che si trovi a distanza 180 gradi dalla media).

6.5.4.2 Dettaglio sulla routine di scelta della traiettoria con il miglior coefficiente di qualità

Ogni traiettoria P_{ji} contiene un campo ("pmu") che memorizza se il primo oim della traiettoria è già stato "utilizzato" da un'altra traiettoria. Qualora l'oim $M_{ji}(t-T+1)$ sia comune a più traiettorie e qualora una di queste traiettorie (P_{ji}) sia scelta per aggiornare un individuo, il campo pmu è posto = 1 per tutte le altre traiettorie che lo contengono. All'atto della ricerca, per un individuo I_{ji}, tra le piste a lui compatibili, di quella che offre il miglior coefficiente di qualità (P_{mQ}), si possono avere tre casi differenti:

- P_{mQ} ha il campo pmu = 0, cioé l'oim $M_{mQ}(t-T+1)$ non è stato ancora utilizzato da nessun individuo. Questa traiettoria è allora scelta come la migliore;
- P_{mQ} ha pmu = 1 ed esiste, tra le traiettorie compatibili, una traiettoria P_{mQ_2} con coefficiente di qualità inferiore, ma con il campo pmu = 0. Questa seconda traiettoria P_{mQ_2} viene allora scelta come la migliore;
- tutte le traiettorie compatibili hanno pmu = 1. Si sceglie allora tra queste traiettorie quella con coefficiente di qualità più elevato.

Lo scopo di questa procedura è di utilizzare, se possibile, per l'aggiornamento di un altro individuo, un oim diverso da uno già utilizzato. Si cerca così di correggere gli errori di classificazione.

Per esempio, può accadere che un oim che rappresenta una sola persona sia erroneamente classificato di tipo GRUPPO. Se ci si limitasse ad esaminare questa etichetta, se ne dedurrebbe che due individui possono essere estesi utilizzando due traiettorie che contengono entrambe questo oim (l'errore sarebbe allora che due individui inseguirebbero entrambi la stessa persona). Per evitare queste situazioni, si cerca sempre di impedire ad un secondo individuo l'utilizzo di una traiettoria INDIVIDUO già utilizzata, nel caso che questo abbia un'altra ragionevole possibilità di estensione (cioé un'altra traiettoria che utilizza degli oim diversi e che ha un buon coefficiente di compatibilità con l'individuo).

6.5.4.3 L'aggiornamento degli attributi degli individui

Trovato la traiettoria P_{ji} che estende l'individuo I_{ji}, è necessario aggiornare gli attributi dell'individuo utilizzando le informazioni memorizzate nella traiettoria.

La procedura è la seguente:

- l'individuo I_{ji} conserva il proprio identificatore numerico; un individuo, contrariamente ad una traiettoria, conserva sempre lo stesso identificatore;
- l'oim $M_{ji}(t - T + 1)$ sostituisce quello che era associato a I_{ji} all'istante $(t - T)$ $(M_{ji}(t - T))$;
- l'istante t_s di creazione di I_{ji} non cambia;
- l'istante t_c dell'ultimo aggiornamento viene incrementato: diventa $t_c + 1$;
- i valori medi dell'altezza e della larghezza degli oim sono aggiornati utilizzando le formule:

$$\overline{H}(t - T + 1) = \alpha \cdot \overline{H}(t - T) + (1 - \alpha) \cdot H_{M_{ji}(t-T+1)}$$
$$\overline{W}(t - T + 1) = \alpha \cdot \overline{W}(t - T) + (1 - \alpha) \cdot W_{M_{ji}(t-T+1)}$$

- se la distanza tra $M_{ji}(t-T)$ e $M_{ji}(t-T+1)$ è superiore ad un valore soglia (e dunque la direzione individuata dai centri di gravità di $M_{ji}(t - T)$ e $M_{ji}(t - T + 1)$ ha un significato), questa direzione è utilizzata per aggiornare la direzione dell'individuo I_{ji};
- si determina la classe di I_{ji}:
  - si controlla se esiste un altro individuo associato allo stesso oggetto in movimento $M_{ji}(t - T + 1)$; in questo caso, la classe è RAGGRUPPATO;
  - altrimenti, se i due oim $M_{ji}(t-T)$ e $M_{ji}(t-T+1)$ sono "non rilevati", la classe è SOSPESO;
  - altrimenti, se una delle due condizioni seguenti è verificata: (1) $M_{ji}(t-T)$ è "non rilevato" (o non esiste) e $M_{ji}(t-T+1)$ si trova nella zona IU; (2) $M_{ji}(t-T+1)$ e $M_{ji}(t-T)$ sono nella zona IU; (3) $M_{ji}(t-T+1)$ non è nella zona IU mentre $M_{ji}(t-T)$ vi si trova, allora la classe è ENTRANTE.
  - altrimenti, in modo simmetrico, se una delle condizioni seguenti è verificata: (1) $M_{ji}(t-T)$ non è nella zona IU mentre $M_{ji}(t-T+1)$ vi si trova; (2) $M_{ji}(t-T+1)$ è "non rilevato" e $M_{ji}(t-T)$ è nella zona IU, allora la classe è USCENTE.
  - altrimenti, se l'individuo non appartiene a nessuna delle classi precedenti, gli viene assegnata la classe INSEGUITO.

6.5.4.4 Le condizioni che determinano se una traiettoria corrisponde ad una sola persona

Lo scopo di queste condizioni è di verificare se la traiettoria P_{ji} è composta d'oim che rappresentano una sola persona piuttosto che un gruppo di più persone. Nel primo caso, l'oggetto in movimento $M_{ji}(t - T + 1)$ della traiettoria non può essere comune a più individui, ed è dunque necessario impedire agli altri individui di utilizzare le traiettorie

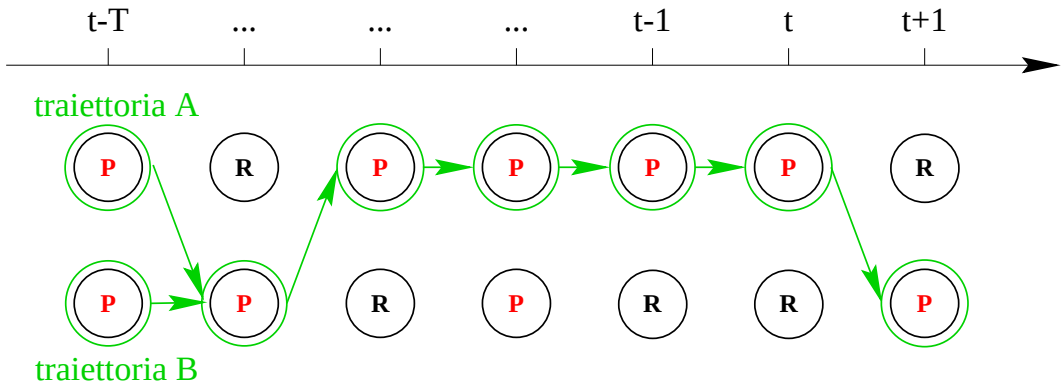


Figura 6.10: *Esempio di situazione nella quale due traiettorie diventano equivalenti: in seguito alla cancellazione degli oggetti in movimento all'istante $t-T$, le due traiettorie A e B hanno gli stessi oim su tutta la finestra temporale T considerata. Esse sono dunque equivalenti. Quella con il coefficiente di qualità più basso è eliminata.*

che hanno lo stesso oggetto in movimento $M_{ji}(t-T+1)$. In caso contrario, ogni scelta è possibile, e dunque nessun vincolo sulla scelta delle traiettorie è posto agli altri individui.

In teoria, l'unica condizione che è necessario verificare è che la classe dell'oim $M_{ji}(t-T+1)$ sia INDIVIDUO, SEMIGRUPPO o GRUPPO. In pratica, la fase di classificazione è affetta da errori, che si vuole, in una certa misura, poter recuperare. Le tre condizioni che, applicate contemporaneamente, determinano quindi se una traiettoria, già associata ad un individuo I_{ji}, descriva in realtà lo spostamento di una sola persona, sono:

- la classe della traiettoria deve essere INDIVIDUO;
- la classe dell'oim $M_{ji}(t-T+1)$ deve essere diversa da GRUPPO e SEMIGRUPPO;
- la classe dell'individuo I_{ji} deve essere diversa da RAGGRUPPATO;

6.5.5 La cancellazione delle piste e degli individui

L'ultima routine da eseguire prima di terminare un ciclo di elaborazione (corrispondente all'elaborazione di una immagine) è di cancellare le traiettorie che, in seguito all'aggiornamento, si trovano completamente sovrapposte tra di loro. Questa situazione è illustrata in figura 6.10.

Quando ciò si verifica, le due traiettorie diventano equivalenti dal punto di vista degli oim che le compongono. Si cancella allora quella con il coefficiente di qualità più basso. In questo modo l'algoritmo deve gestire un numero minore di traiettorie e si è eliminata una ridondanza.

La seconda situazione che richiede la cancellazione di una traiettoria si ha quando questa corrisponde ad una persona che è uscita dalla scena. Dal momento in cui la persona fisica non genera più oim nelle immagini (perché è uscita dal campo della videocamera), la traiettoria è aggiornata utilizzando degli oim "non rilevati". Dopo esser stata aggiornata $T/5$ volte con oim "non rilevati", si sospende l'aggiornamento della traiettoria. Essa è distrutta, insieme all'individuo associato, quando anche l'ultimo oim che essa contiene è stato utilizzato dall'individuo.

Questo capitolo conclude la parte descrittiva dell'architettura e del funzionamento della

piattaforma di interpretazione progettata. Il capitolo successivo illustra alcuni risultati ottenuti, mentre il capitolo 8 vuole essere un bilancio del lavoro svolto nonché gettare un rapido sguardo sulle prospettive di sviluppo e miglioramento esistenti.

Capitolo 7

I risultati ottenuti: analisi critica

7.1 Considerazioni introduttive

In questo capitolo vengono presentati i risultati ottenuti dalla piattaforma d'interpretazione realizzata. Come sempre accade nell'ambito dell'interpretazione automatica di sequenze video, è estremamente difficile fornire un dato quantitativo capace di riassumere in una cifra la bontà di un algoritmo o dell'intero sistema di elaborazione. Se infatti è a priori possibile individuare alcune cifre di merito sulla base delle quali misurare i risultati ottenuti (solo per fare un esempio, tali cifre di merito potrebbero essere: *il numero di situazioni ambigue correttamente interpretate (sul totale delle situazioni ambigue sottoposte all'algoritmo)*, oppure *la capacità di operare su scene filmate le più variegate*, o ancora *la capacità di recuperare o correggere precedenti errori d'interpretazione commessi*), ci si renderebbe però presto conto che la misura delle stesse presenta difficoltà praticamente insormontabili. Per esempio, si dovrebbe poter definire in modo preciso ed univoco una *situazione ambigua*, in secondo luogo si dovrebbe approntare una serie di *situazioni ambigue* di test da sottoporre all'algoritmo, e tutte queste dovrebbero, in teoria, o essere caratterizzate dallo stesso grado di "difficoltà interpretativa", in modo da permettere la misura della cifra di merito come rapporto $\frac{successi}{insuccessi}$, oppure da gradi di "difficoltà interpretativa" diversi, ma quantificabili, in modo da esprimere la misura della cifra di merito come rapporto pesato dei successi sugli insuccessi. In realtà, spesso la differenza tra una "difficoltà interpretativa semplice" e una "complessa" può essere originata da un dettaglio nella sequenza d'ingresso nemmeno percepibile dall'osservatore umano disattento.

Infine, qualora si riesca nell'impresa di approntare questo kit di cifre di merito e di sequenze di test per una data piattaforma in esame, e si volesse procedere ad un confronto tra diverse piattaforme, si farebbe l'amara scoperta che questo kit è, con tutta probabilità, inapplicabile alla seconda piattaforma in esame, in quanto concepito, studiato e creato sulla prima, e che, di conseguenza, esso è completamente inutile al fine di effettuare una comparazione quantitativa dei risultati. Nel caso, più utopistico che reale, che si riesca ad applicare questo kit anche alla seconda piattaforma, esso darebbe, con elevatissima probabilità, risultati assurdi, addirittura in contraddizione con il buon senso dell'osservatore (in altri termini, potrebbero risultare cifre di pessimo valore per una piattaforma che, invece, svolge con affidabilità il proprio lavoro).

Alla base di questo stato delle cose sta una serie vasta e variegata di motivi, la cui analisi esula dalla presente trattazione. Si consideri soltanto che l'interpretazione automatica è una disciplina molto recente, ancora agli albori della propria storia e che, probabilmente, vedrà enormi sviluppi e compirà rilevanti passi in avanti nei prossimi decenni. Allo stato attuale, sebbene i risultati ottenuti siano sempre più incoraggianti e le piattaforme di ela-

borazione sempre più affidabili, non si dispone ancora né di prodotti robusti, né versatili (nel senso di piattaforme che possano essere utilizzate indifferentemente su sequenze video di origine disparata). Le tipiche piattaforme create attualmente, o disponibili nei laboratori di ricerca del settore (e la presente ne è un tipico esempio), nascono per lavorare su sequenze video tutt'altro che generali, al fine di produrre risultati ben precisi ed interpretare comportamenti delineati con precisione in anticipo (al limite, la definizione stessa di tali comportamenti entra a far parte dei requisiti di progetto della piattaforma). Esprimendo il tutto in modo più prosaico, si potrebbe dire che all'interno dell'universo delle situazioni interpretabili, ogni piattaforma creata o in fase di sviluppo ambisce ad uno spazio di lavoro pari ad un piccolo pianeta. A questo punto risulta più facile comprendere l'intrinseca difficoltà insita nella definizione di uno strumento di test quantitativo universale.

Infine, proprio da questo suo carattere di "disciplina giovane" deriva l'impossibilità, per l'interpretazione automatica di sequenze video, di disporre di una teoria della misura affidabile e validata. Si tratta di un semplice problema di priorità: prima di sviluppare una teoria (e degli strumenti) capaci di misurare oggettivamente una realtà, è necessario che tale realtà sia ben definita, conosciuta e padroneggiata con sicurezza.

Tutte queste considerazioni non vogliono comunque essere un invito a sospendere il giudizio in attesa dei futuri passi della ricerca, o a rinunciare ad ogni valutazione dei risultati ottenuti sulla base di una oggettiva non disponibilità di cifre di merito numeriche. Al contrario, vogliono essere un invito a sostituire l'analisi quantitativa dei risultati con una di stampo più qualitativo, sicuramente più imprecisa, ma certamente più versatile e potente in questa fase della ricerca.

7.2 I risultati

Nelle sezioni che seguono saranno quindi presentati alcuni risultati significativi ottenuti. I risultati presentati si riferiscono al modulo di *riconoscimento delle regioni color pelle*, al modulo di *classificazione e fusione* e a quello *d'inseguimento*. Si è scelto di presentare risultati relativi a questi tre moduli in quanto essi sono i più rappresentativi dell'intera piattaforma di elaborazione, e comunque quelli sui quali si è maggiormente concentrato il lavoro di tesi. Infine si presenteranno risultati a livello dell'*interpretazione* delle sequenze, cioé al livello d'astrazione massimo dell'intera piattaforma.

A tal proposito, i risultati sono presentati sotto una forma particolare; al fine di rendere chiaro al lettore il risultato dell'interpretazione, si è completata la piattaforma con un modulo capace di ricostruire in linguaggio VRML la scena filmata interpretata. In tale ricostruzione compaiono (si veda anche la figura 7.1:

- gli oggetti del mobilio, così come muri, pavimenti e porte della scena (in grigio);
- le persone che si muovono nella scena, visualizzate come cilindri di colore differente (il colore è associato all'identità della persona);
- un cubo sospeso al centro della stanza, di colore variabile (verde, arancio o rosso); il colore di questo cubo rappresenta l'effettivo risultato dell'interpretazione, e cambia a seconda dello scenario che si vuole interpretare.

Non bisogna infatti dimenticare che scopo ultimo del nostro sistema è comandare un modulo MediaSpace (quale, per esempio, CoMeDi; cf il paragrafo 1.1.2). Non disponendo di un modulo MediaSpace "comandabile" (capace cioè di adattare il flusso di immagini trasmesse secondo i comandi impartiti da un modulo esterno), si è preferito utilizzare questo

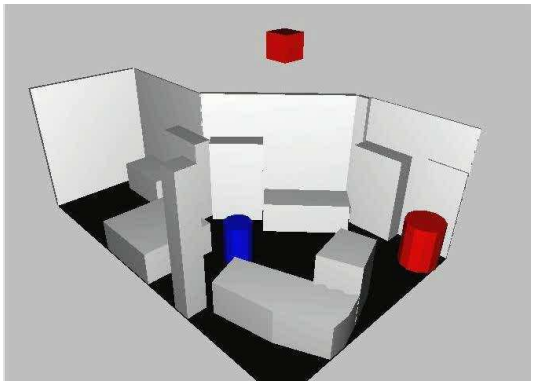


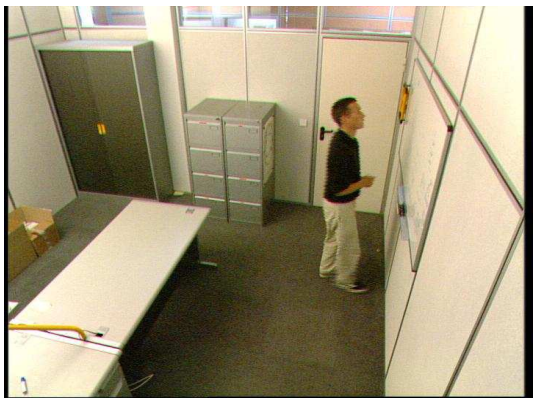
Figura 7.1: *Nell'esempio di ricostruzione VRML mostrato si possono notare i muri e l'arredo della stanza (in grigio), i cilindri che rappresentano gli individui presenti nella scena e inseguiti dal sistema (in blu e rosso) e il cubo colorato.*

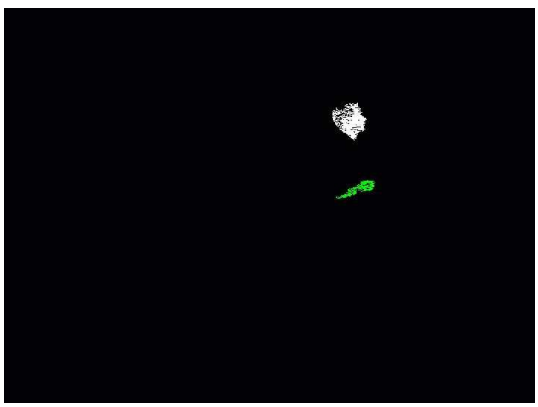
formato VRML al fine di rendere comprensibili al lettore i risultati dell'interpretazione. Esso offre infatti determinati vantaggi:

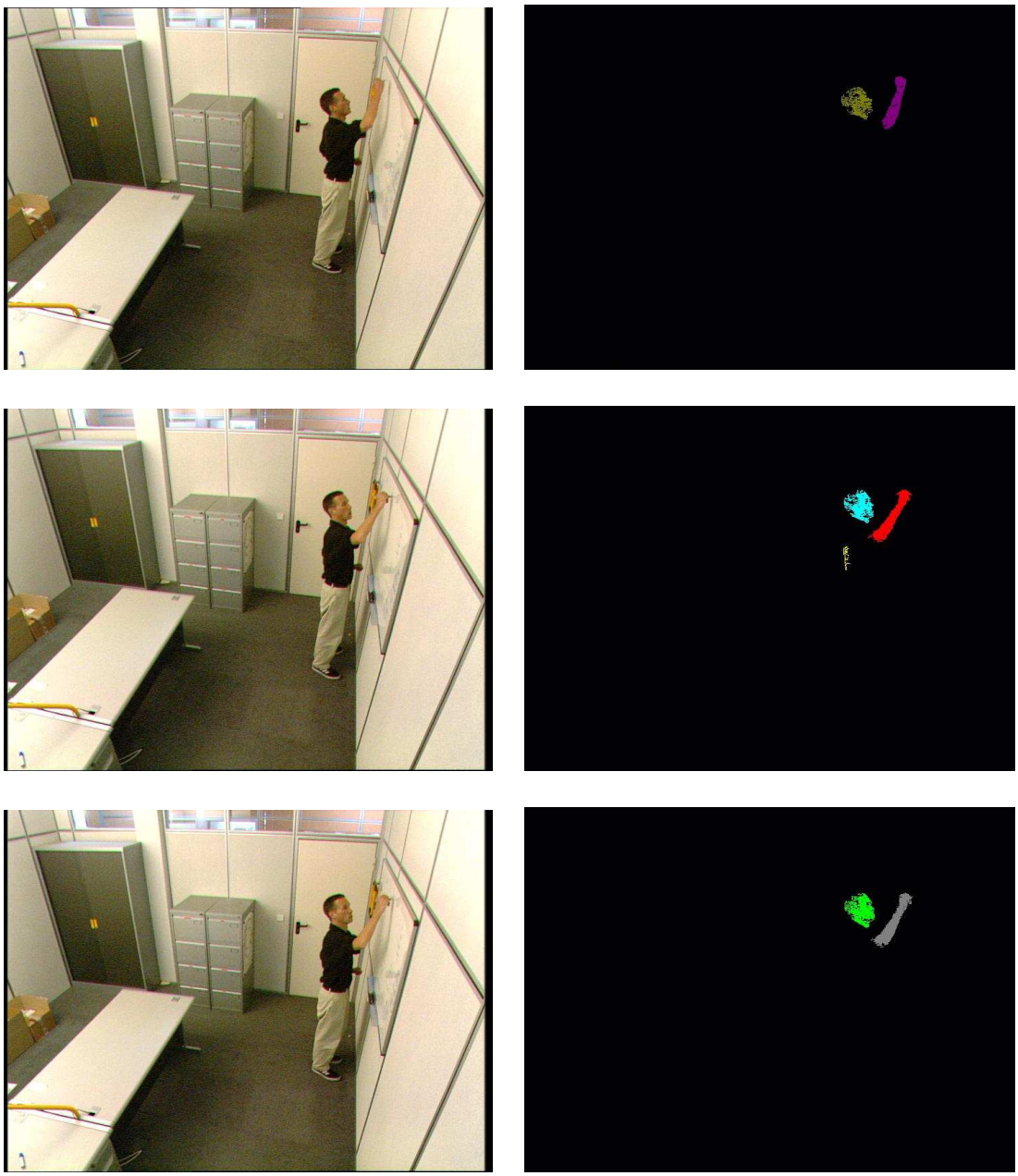
- l'immediatezza della lettura: la semplice struttura geometrica della ricostruzione, unita alla semantica dei differenti colori, permette una immediata comprensione degli avvenimenti;
- l'indipendenza della ricostruzione VRML dal flusso di immagini d'ingresso: la scena è ricostruita utilizzando la base del contesto (per gli oggetti del mobilio e la scena stessa), le descrizioni degli individui offerte dal modulo di inseguimento (per le loro dimensioni, la loro posizione e la loro identità) e il "comando" impartito dal modulo di interpretazione in base allo scenario riconosciuto (per il colore del cubo).

7.2.1 Risultati del modulo di riconoscimento delle regioni color pelle

La serie di immagini che seguono presenta nella colonna di sinistra l'immagine sorgente in esame e nella regione di destra le regioni color pelle individuate dall'algoritmo. Il colore di queste ultime è aleatorio, e non ha alcun significato associato. Questa prima sequenza è una sequenza di test filmata in una scena diversa da quella dell'applicazione MediaSpace.

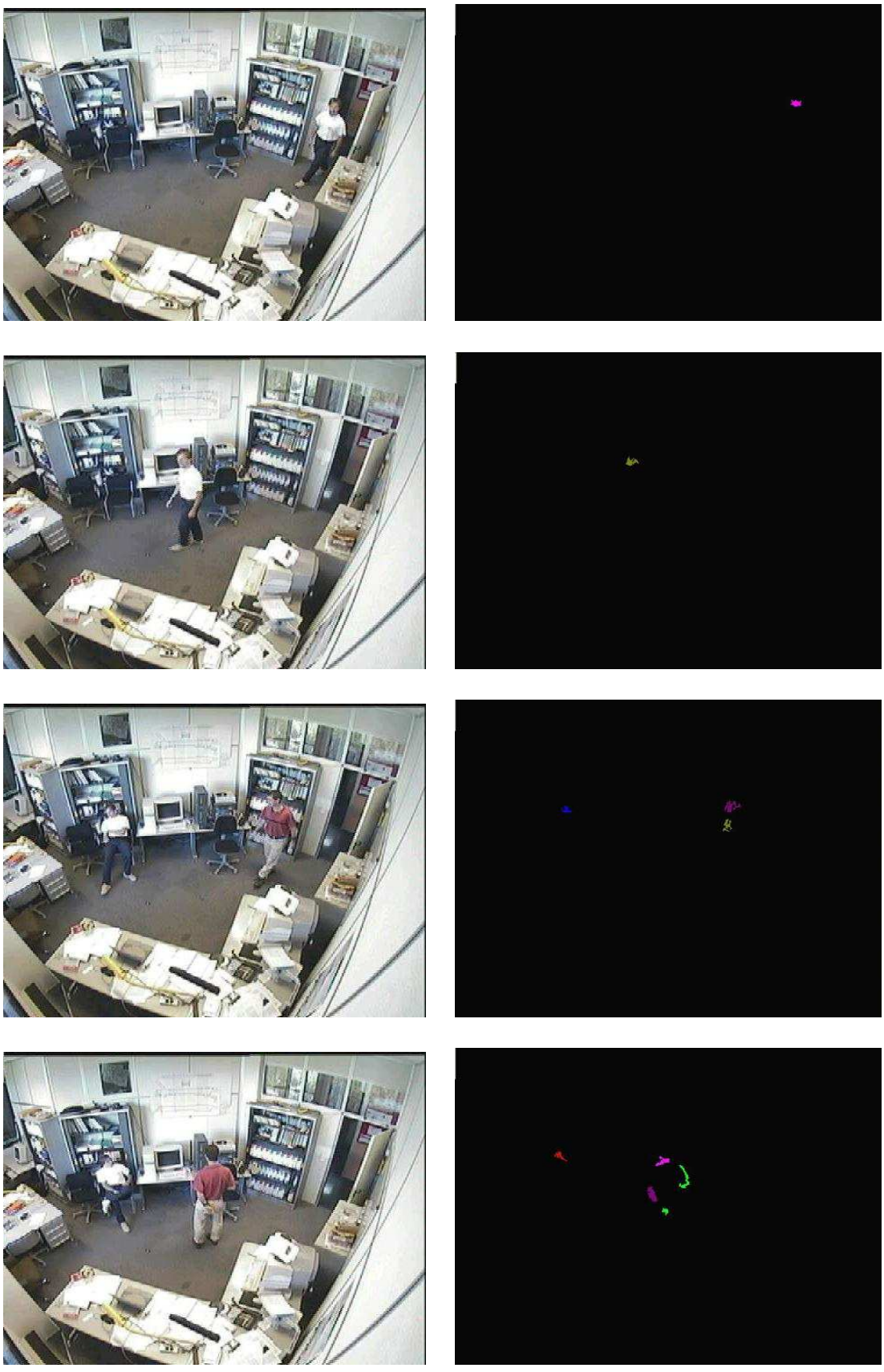






Si può notare, nella terza coppia di immagini, un errore di rilevamento: una regione di pelle è individuata nella zona dell'immagine corrispondente alla schiena della persona.

Le immagini che seguono provengono invece da una sequenza tipica MediaSpace.



I risultati ottenuti dal modulo di riconoscimento delle regioni color pelle sono soddisfacenti: le regioni di pelle sono riconosciute con un tasso di successi superiore al 95%, mentre i falsi riconoscimenti (come osservato nel capitolo 4, meno dannosi dei mancati riconoscimenti) si verificano nel 5-10% dei casi. La qualità dei risultati ottenuti è comunque fortemente influenzata, per la natura dell'algoritmo stesso, dalla qualità della biblioteca di immagini color pelle utilizzata.

7.2.2 Risultati dei moduli di classificazione/fusione e di inseguimento

Le immagini che seguono presentano i risultati ottenuti in due diverse sequenze dal modulo di classificazione e fusione e dal modulo di inseguimento. Le immagini, disposte in ordine lessicografico, sono tratte da sequenze filmate ed interpretate dalla piattaforma realizzata. Per quel che concerne il modulo di classificazione/fusione, i risultati sono rappresentati dal colore del rettangolo più interno che circonda gli oggetti in movimento (individui e non) presenti nella scena. La semantica dei differenti colori è quella illustrata dalla tabella 5.1, qui riportata per facilità di lettura.

Colore del rettangolo circoscrivente	Classe corrispondente	Colore del rettangolo circoscrivente	Classe corrispondente
	individuo		oggetto dell'arredo
	individuo occultato		veicolo
	gruppo d'individui		indeterminato
	folla		rumore

Tabella 7.1: *Il colore del rettangolo circoscrivente indica la classe dell'oggetto in movimento rappresentato nelle diverse immagini.*

Dalle immagini delle sequenze che seguono emerge come il modulo sia capace di riconoscere con soddisfacente esattezza la vera natura dell'oggetto in movimento e come operi, dove necessario, l'opportuna fusione tra le zone in movimento rilevate.

Per quel che concerne invece il modulo d'inseguimento, si faccia riferimento al rettangolo giallo più esterno. La presenza di questo rettangolo attorno all'immagine di una persona indica l'avvenuto riconoscimento da parte del modulo d'inseguimento della presenza di un individuo e indica anche che l'individuo in questione è inseguito dal sistema. Il numero in alto a destra indica l'identità dell'individuo (cioé il sistema etichetta con uno stesso numero due oggetti in movimento presenti in immagini diverse se li riconosce corrispondere alla stessa persona fisica). L'etichetta in basso a sinistra indica invece lo stato d'inseguimento dell'individuo: ENTERING (l'individuo sta entrando nella scena), FOLLOWED (l'individuo si muove nella scena ed è inseguito dal modulo), EXITING (l'individuo sta uscendo dalla scena) etc.

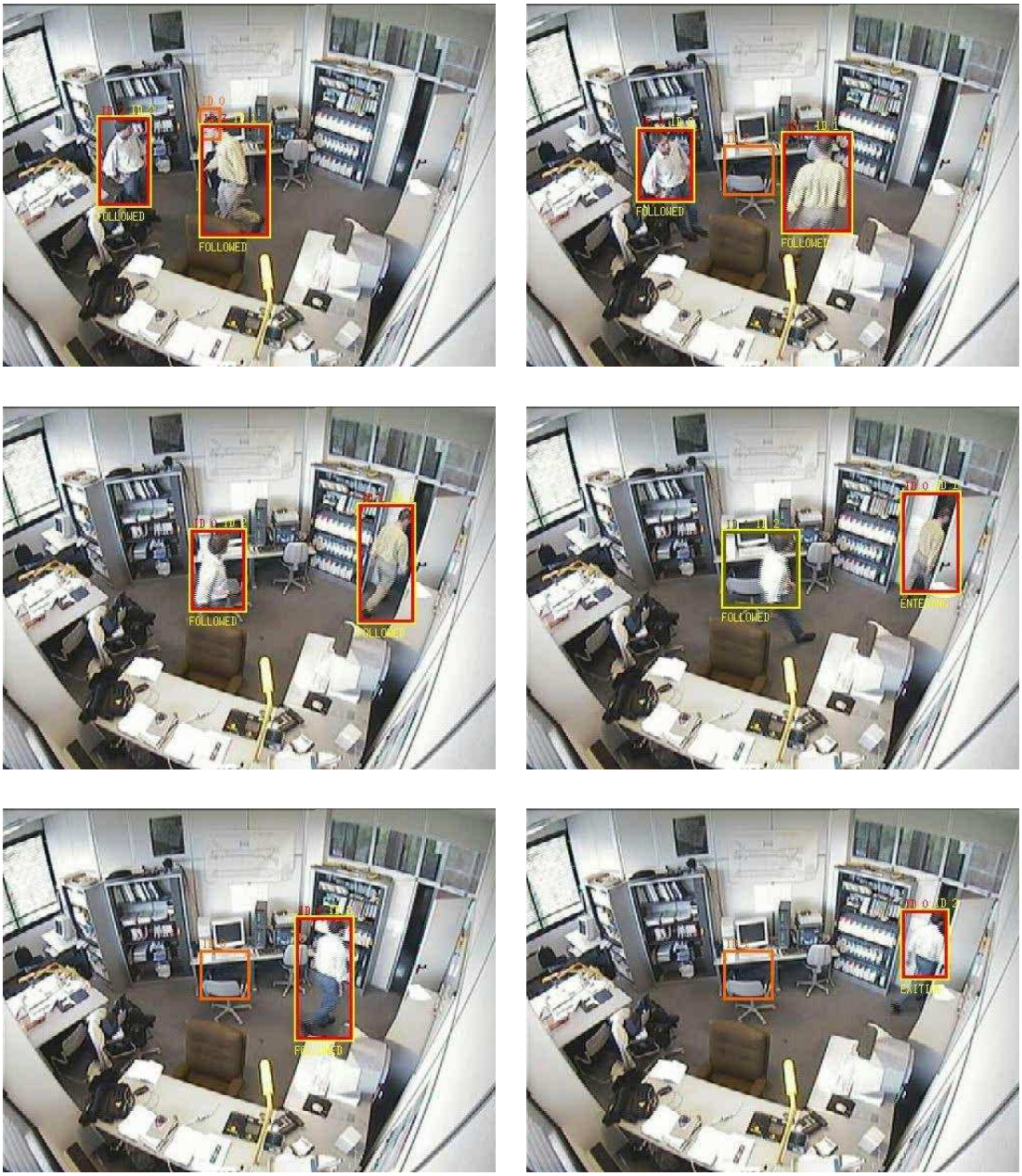
Sequenza 1:


ENTERING
FOLLOWED
FOLLOWED
FOLLOWED
FOLLOWED
FOLLOWED
FOLLOWED

La sequenza 1 mostra un caso di parziale sovrapposizione di due persone. Il sistema riconosce correttamente la situazione ed insegue i due individui senza errori e senza scambi di identità.

La sequenza 2 illustra un caso ancora più complicato: due persone si incrociano completamente ed inoltre una di esse sposta un oggetto dell'arredo che si trova sulla traiettoria. Tale spostamento genera il riconoscimento di un oggetto in movimento che potrebbe confondere il sistema d'interpretazione.

Sequenza 2:



Anche in questo caso, tuttavia, il sistema insegue correttamente i due individui, senza confonderne le identità e senza considerare l'ggetto dell'arredo spostato come una regione in movimento corrispondente ad un individuo.

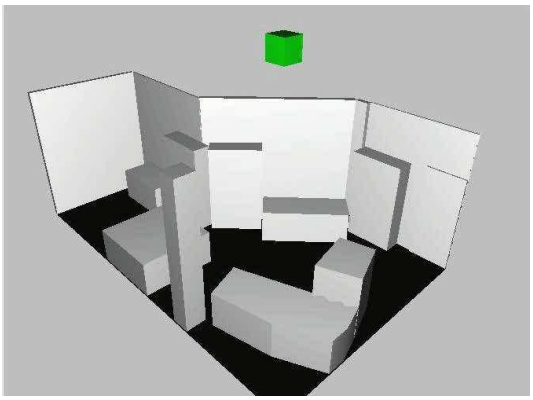
In alcune di queste immagini (per esempio nella sequenza 1, immagini 2 e 7, oppure nella sequenza 2, immagine 10) si possono notare alcuni errori del modulo di classificazione; per l'esattezza, oggetti in movimento classificati "oggetto dell'arredo" sono in realtà individui. L'errore deriva sia dalle dimensioni della regione in movimento in esame (per esempio, nell'immagine 10 della sequenza 2 le dimensioni e le proporzioni della regione in movimento sono più simili a quelle di un oggetto dell'arredo che di una persona), sia dalla zona della scena nella quale essa si trova; per esempio, quando un individuo passa in prossimità della porta è molto facile che la corrispondente regione in movimento sia classificata come "oggetto dell'arredo" in quanto la base del contesto descrive la presenza, in questa posizione,

di un oggetto dell'arredo suscettibile di spostarsi (la porta). Tali errori non inficiano comunque i risultati finali, e gli individui sono correttamente seguiti.

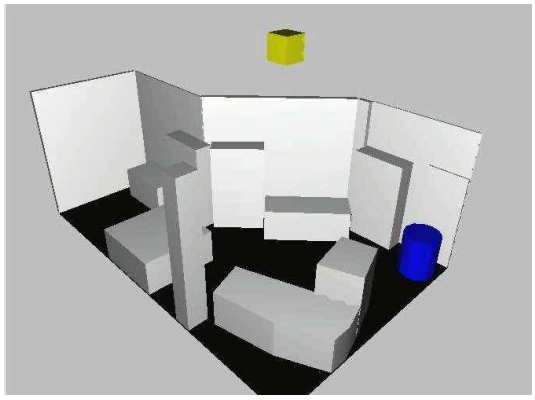
7.2.3 Risultati del modulo d'interpretazione

Le immagini che seguono illustrano i risultati finali offerti dal modulo d'interpretazione sulle due sequenze illustrate precedentemente. La colonna di sinistra mostra le immagini in ingresso, mentre la colonna di destra illustra l'interpretazione VRML di ciò che avviene. Per mostrare le capacità interpretativa del modulo, si è semplicemente deciso di legare il colore del cubo presente nella ricostruzione VRML al numero di persone presenti nella scena: **cubo di colore verde**: nessuna persona in scena (per esempio il comando trasmesso al modulo MediaSpace potrebbe essere "trasmissione delle immagini possibile"); **cubo di colore giallo**: una persona ed una sola presente in scena ("trasmissione possibile, ma stato di allerta. La persona in scena potrebbe essere il proprietario dell'ufficio."); **cubo di colore rosso**: due o più persone presenti nella scena ("trasmissione proibita, o filtrata, in quanto è possibile che sia in corso una riunione"). L'interpretazione condotta è volutamente semplicistica, in quanto il suo scopo non vuole tanto essere poter impartire comandi raffinati e flessibili (tanto più che il modulo MediaSpace del quale si dispone non è comandabile), ma piuttosto dimostrare in potenza quali sono le possibilità del modulo interpretativo. In altri termini, con le informazioni a disposizione del modulo d'interpretazione, schematizzate nella ricostruzione VRML (informazioni riassumibili nella posizione ed identità di ogni persona presente nella scena nonché relazione di prossimità/interazione con gli oggetti dell'arredo quali i computer o i telefoni) si possono predisporre riconoscimenti di scenari ben più complicati e raffinati di questo semplice esempio.

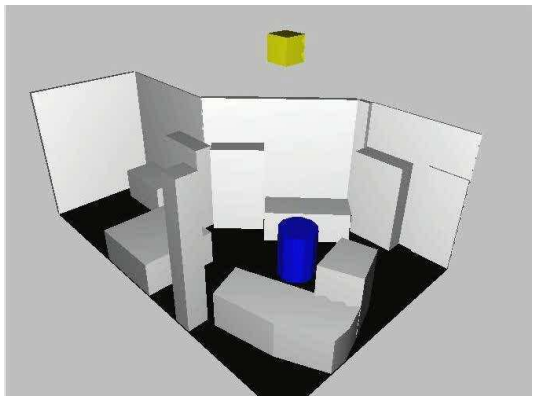
Sequenza 1:



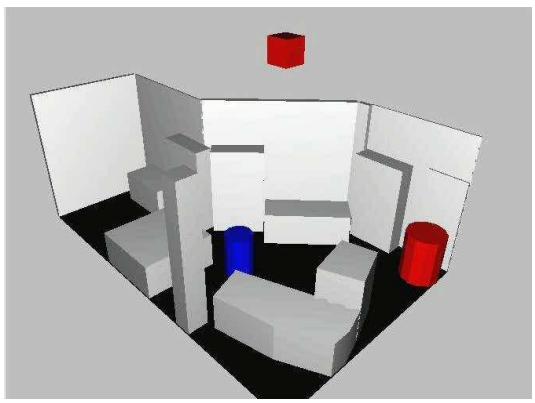

ENTERING



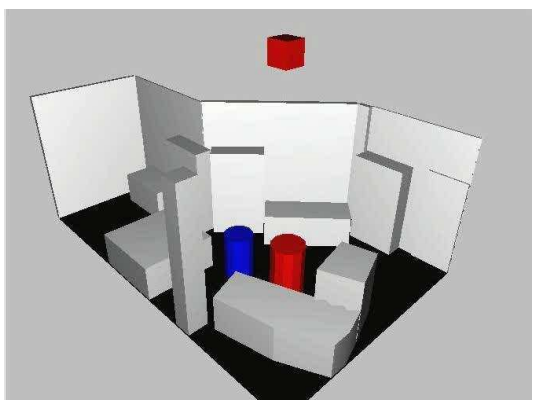

FOLLOWED

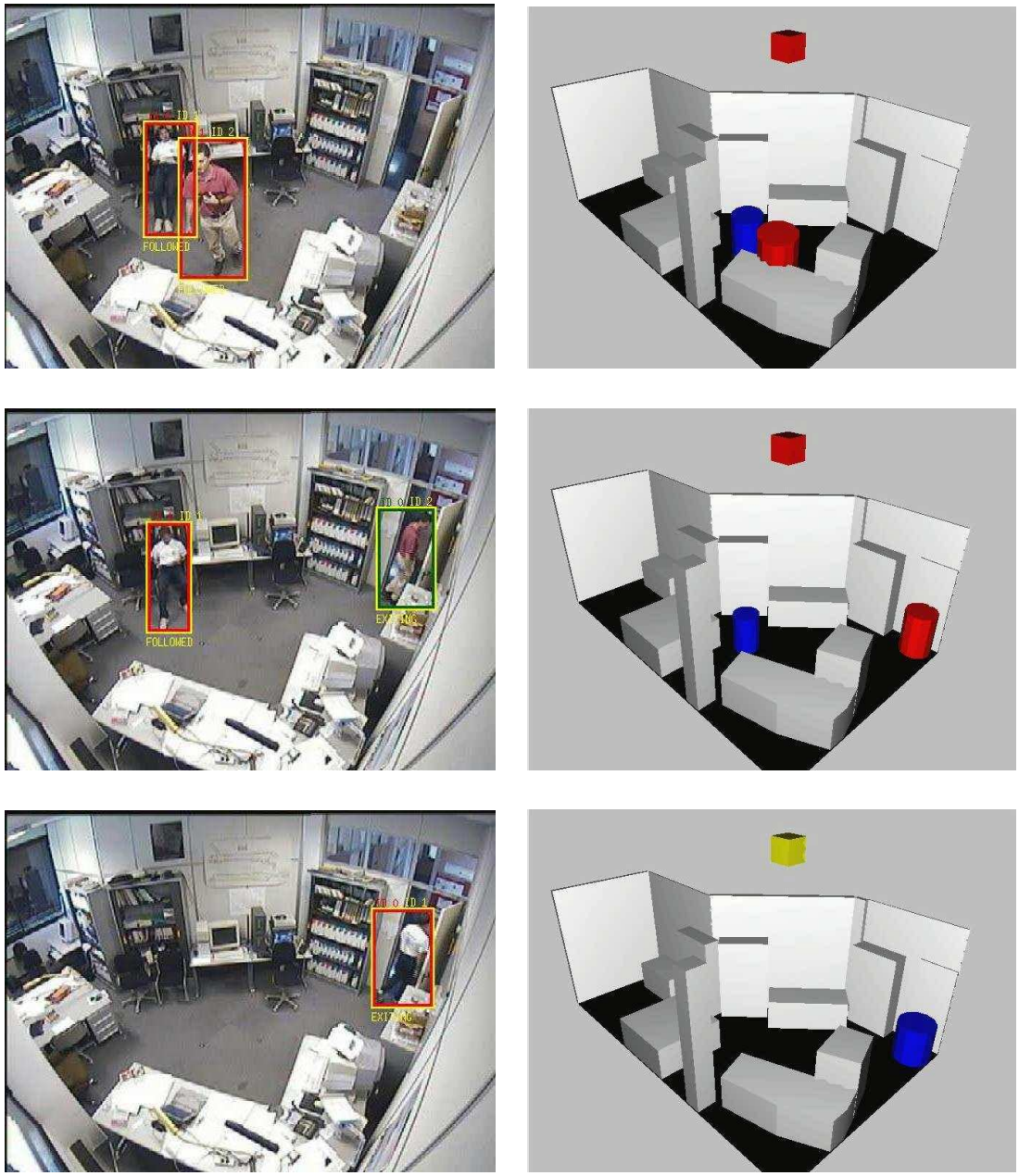



FOLLOWED

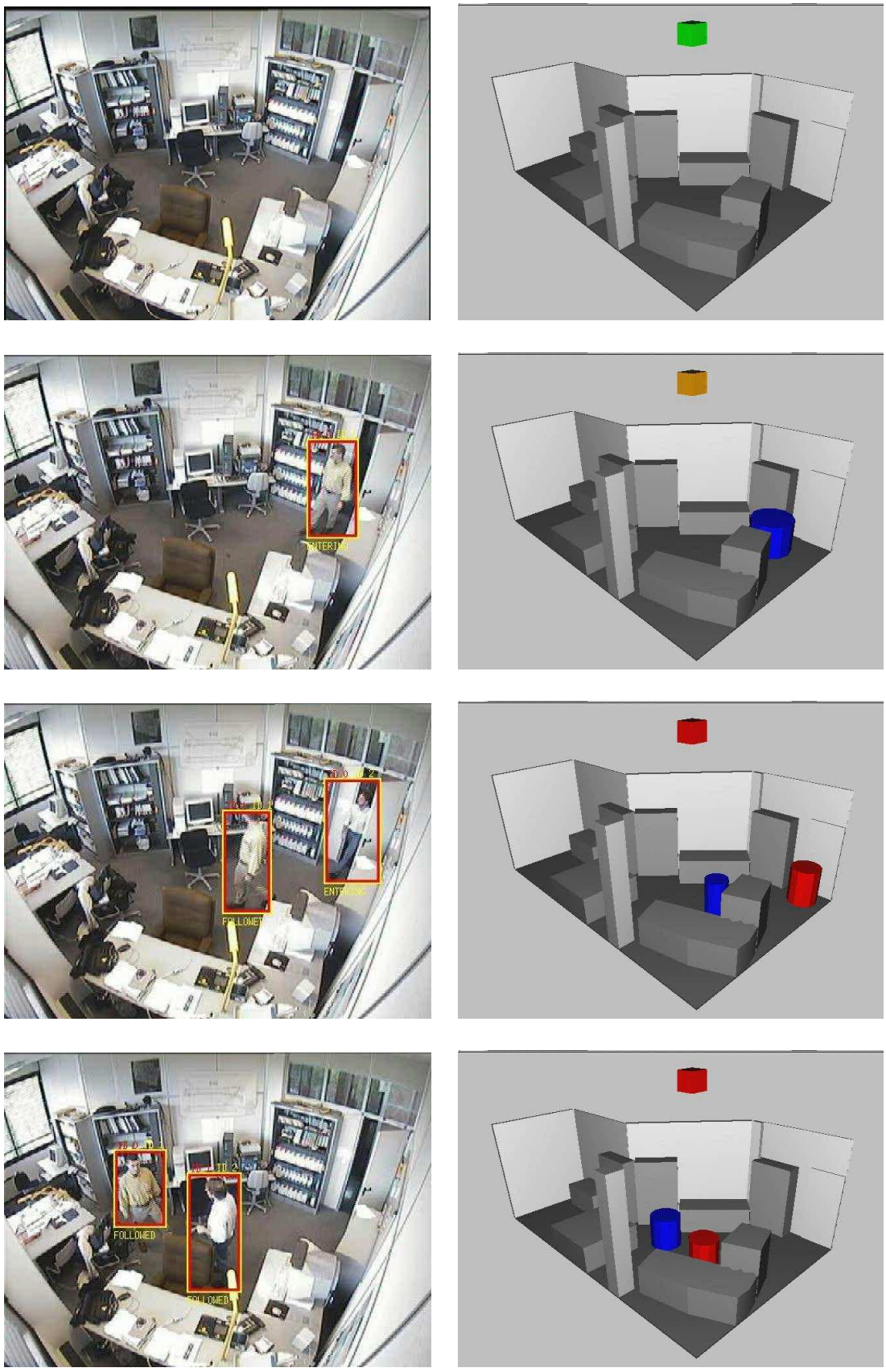


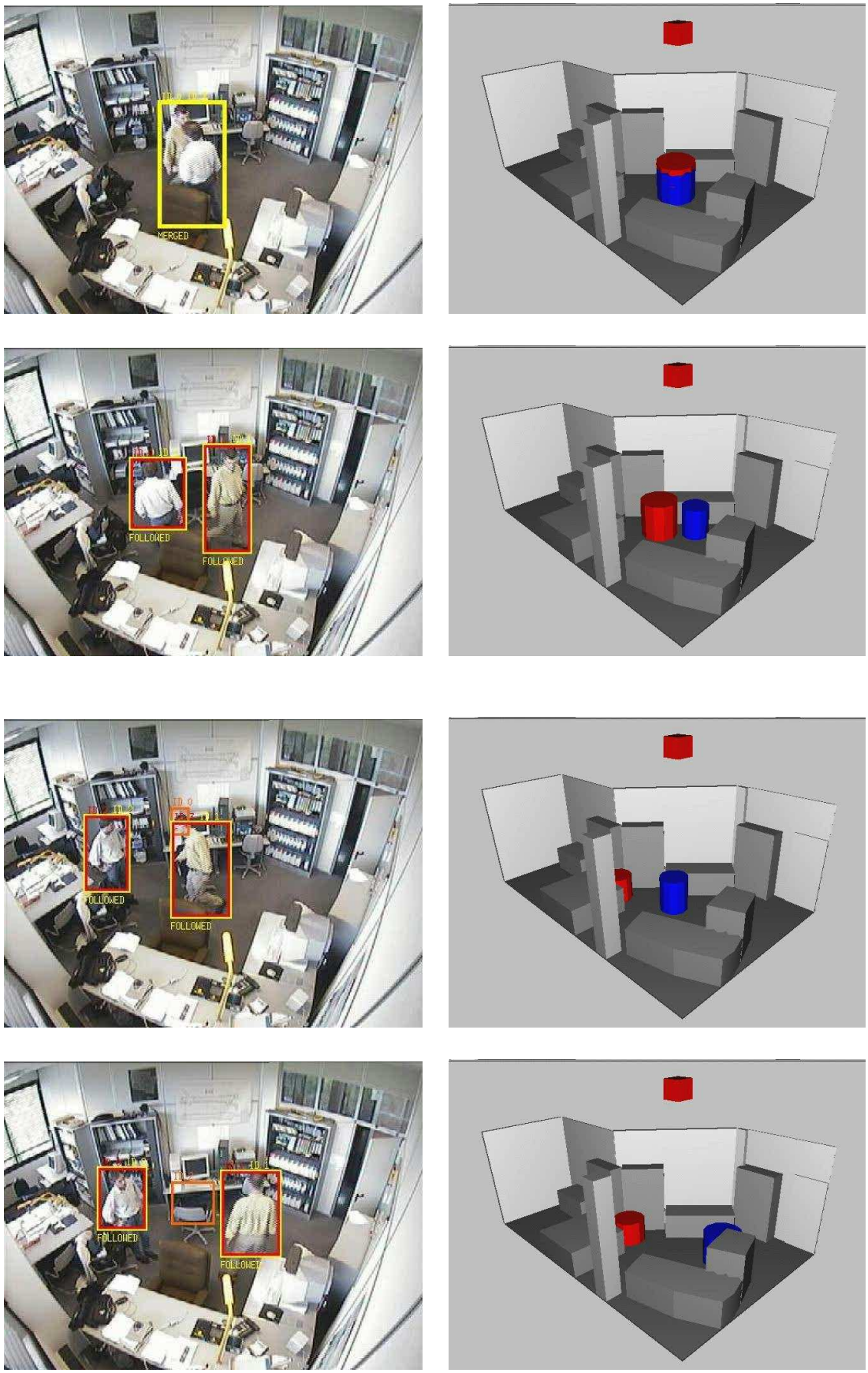

FOLLOWED
FOLLOWED

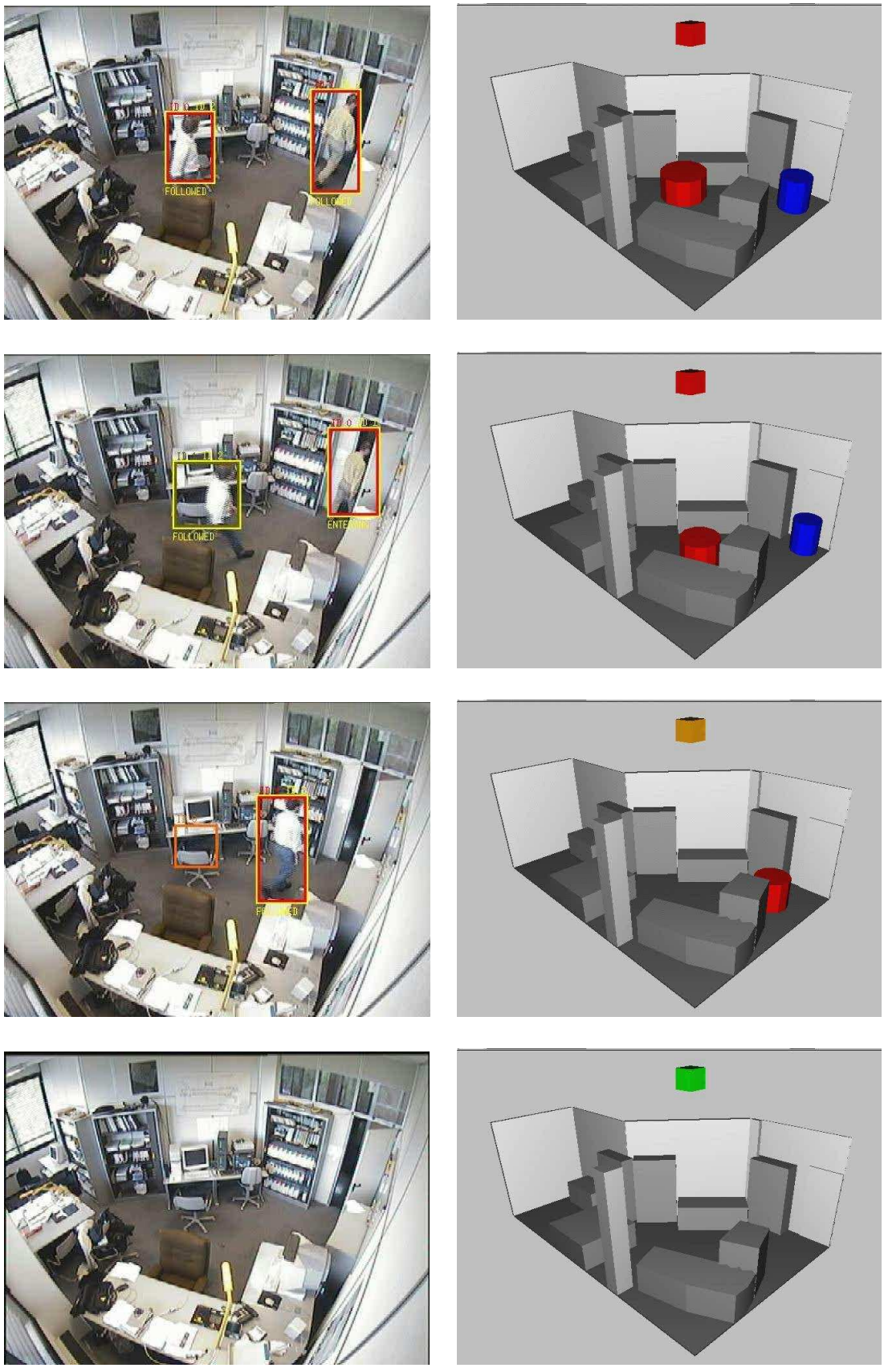




Sequenza 2:

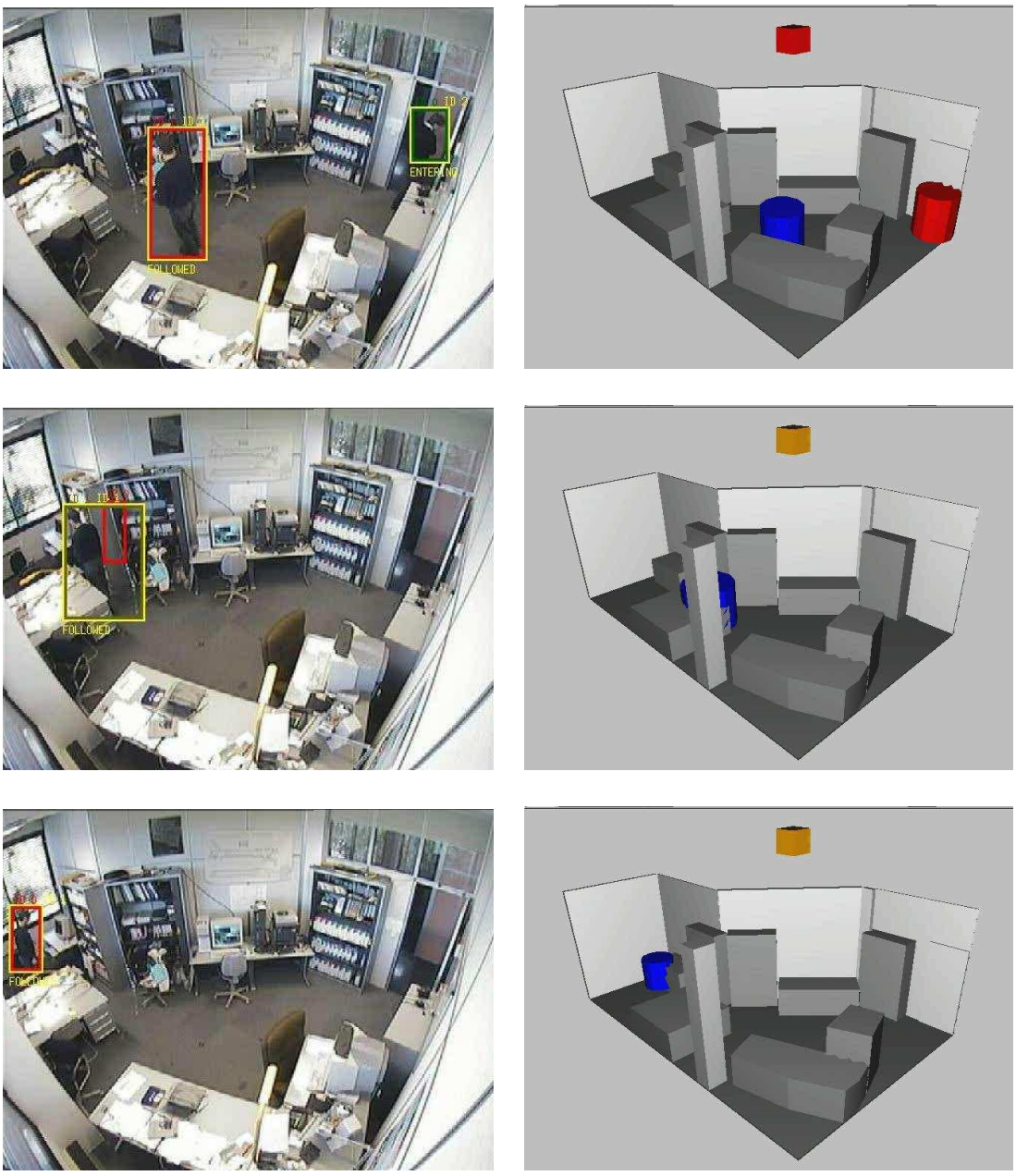

FOLLOWED
FOLLOWED

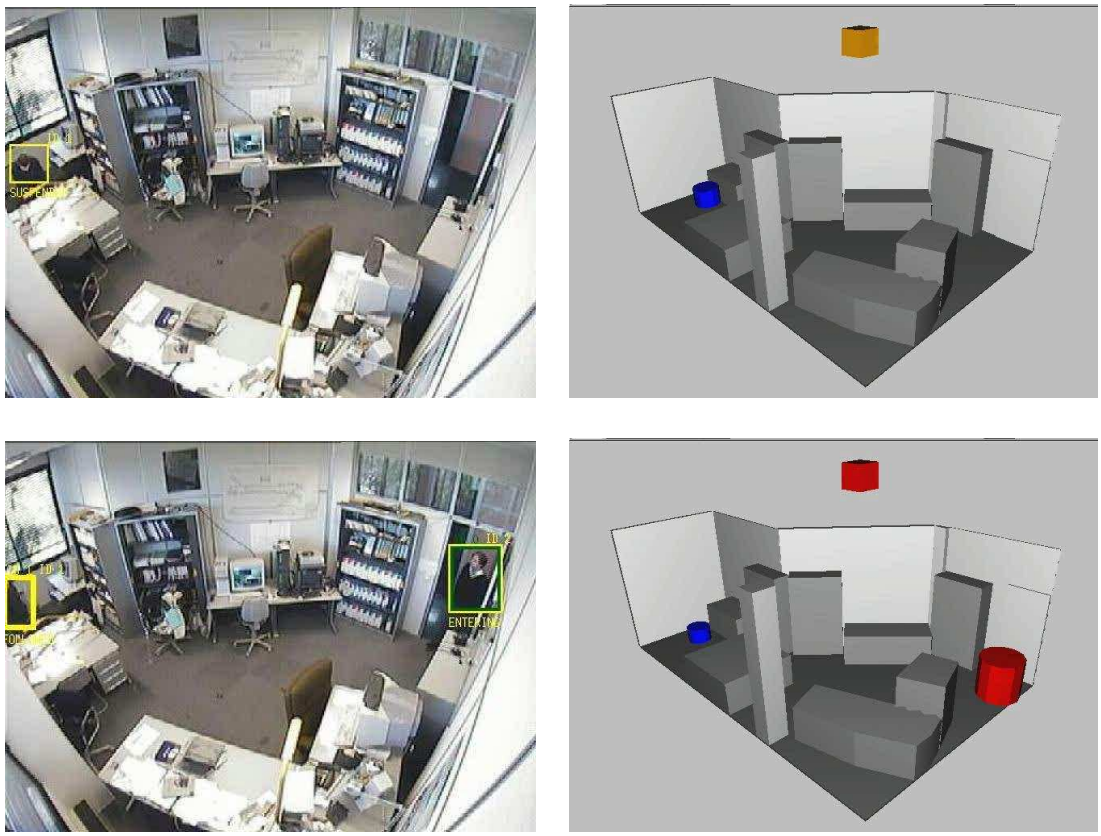

MERGED
FOLLOWED
FOLLOWED
FOLLOWED
FOLLOWED
FOLLOWED
FOLLOWED


FOLLOWED
FOLLOWED
FOLLOWED
ENTE

La terza sequenza illustrata qui di seguito mostra invece come il sistema sia capace di far fronte ad errori di mancato rilevamento delle regioni in movimento corrispondenti agli individui presenti nella scena; la mancanza del rettangolo più interno sull'individuo nell'immagine di sinistra della quarta coppia di immagini indica un mancato rilevamento della regione in movimento corrispondente. Tuttavia, il modulo d'inseguimento, grazie all'utilizzo dell'oggetto in movimento "non rilevato" (cf 6.5.2.1 a pagina 82), è in grado di recuperare questi errori continuando correttamente l'inseguimento (si noti che in corrispondenza dell'utilizzo dell'oggetto in movimento "non rilevato" l'etichetta dell'individuo inseguito diviene SUSPENDED, a significare la particolare situazione nella quale ci si trova).

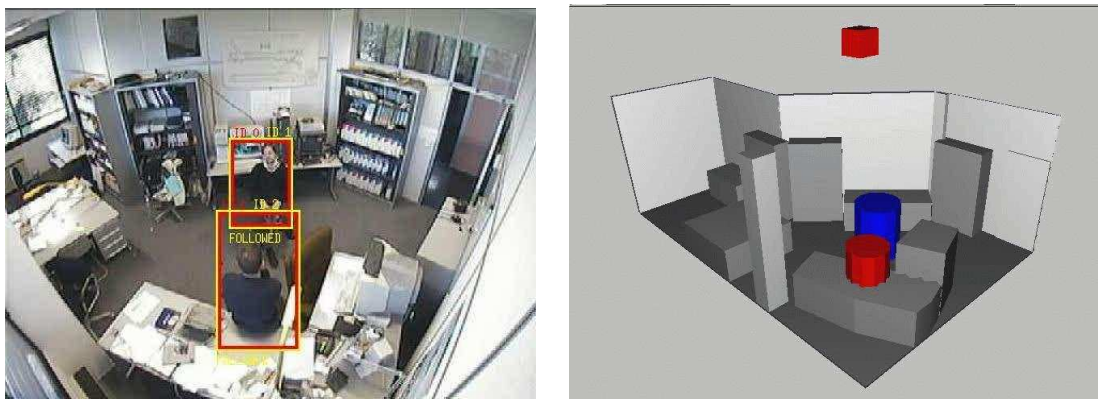
Sequenza 3:

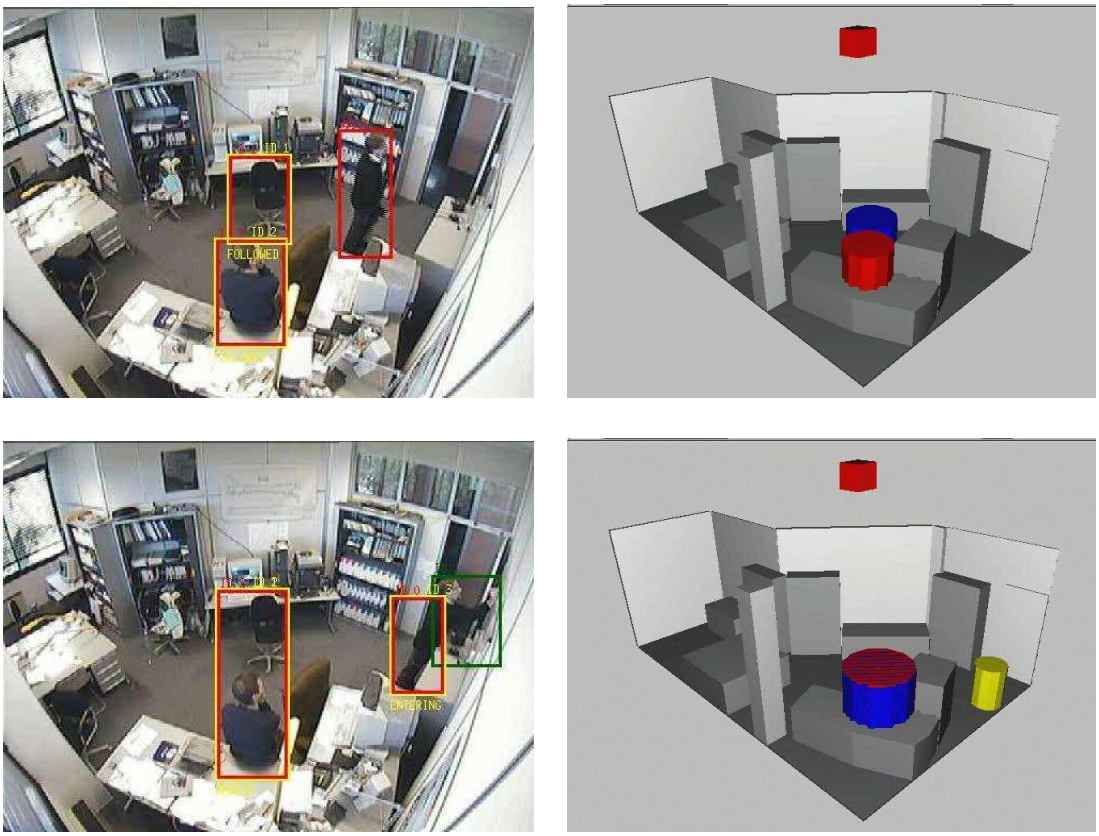




Chiudiamo questo capitolo dei risultati proponendo una sequenza che il sistema non è in grado di interpretare correttamente. Tale sequenza possiede tutti gli elementi capaci di mettere in difficoltà il sistema d'interpretazione, infatti è stata girata al preciso scopo di sondarne i limiti di robustezza. Presentiamo qui di seguito le immagini della sequenza: la colonna di sinistra illustra i risultati dei moduli di classificazione/fusione e del modulo d'inseguimento, mentre la colonna di destra l'interpretazione VRML.

Sequenza 4:





La prima coppia d'immagini mostra la scena interpretata correttamente: due persone discorrono una di fronte all'altra, il modulo d'inseguimento insegue correttamente ognuna delle due ed il modulo di interpretazione ricostruisce correttamente la scena in VRML. Una delle due persone è però seduta sulla sedia da un certo tempo (decine di secondi). Come conseguenza immediata di questo fatto abbiamo che:

- lo spostamento medio della struttura individuo associata a questa persona è praticamente zero, il che implica che la sua traiettoria sarà aggiornata utilizzando preferibilmente oggetti in movimento che si trovano sempre nella stessa posizione (rispetto ad altri in posizioni diverse);
- la sedia è stata spostata, di conseguenza quando la persona si alzerà essa genererà un oggetto in movimento sull'immagine;

Il verificarsi contemporaneo di questi due eventi provoca l'errore d'inseguimento illustrato dalla seconda coppia d'immagini: l'individuo seduto (e che ora si avvicina alla lavagna) viene inseguito in modo errato: il modulo considera l'oggetto in movimento corrispondente alla sedia spostata come il prolungamento più logico della traiettoria associata all'individuo e di conseguenza insegue la sedia invece che l'individuo.

Nella terza coppia di immagini si nota poi un secondo errore, direttamente causato del primo: la persona fisica si è avvicinata alla lavagna, che si trova all'interno della zona di ingresso/uscita. Gli oggetti in movimento da essa generati vengono quindi interpretati come corrispondenti ad una persona che sta entrando nella stanza; non essendoci infatti nessuna struttura individuo associata alla traiettoria (entrante) che lega questi oggetti in movimento, il modulo d'inseguimento crea una struttura individuo e la associa a questa

traiettoria. Il risultato è che, per il modulo d'interpretazione, tre persone sono presenti nella stanza: la due persone entrate precedentemente (una correttamente inseguita, l'altra confusa con la sedia)e la "nuova" terza persona appena entrata (che in realtà è la seconda entrata precedentemente e non più inseguita). Il modulo d'interpretazione a questo punto è incapace di recuperare gli errori commessi (anche perché non è stato progettato per essere in grado di farlo) e di conseguenza tutta l'interpretazione successiva sarà errata.

Al di là di questa particolare sequenza "trappola", la piattaforma d'interpretazione si rivela comunque in grado (come dimostrano le sequenze illustrate precedentemente) di interpretare correttamente anche sequenze considerate "ardue".

Nel prossimo capitolo si esporranno le conclusioni ispirate dai risultati ottenuti, i possibili miglioramenti che si potrebbero introdurre al fine di rendere la piattaforma più robusta ed infine si delineeranno brevemente alcune linee guida per la prosecuzione del progetto.

Capitolo 8

Conclusioni e prospettive

8.1 Bilancio del lavoro svolto

Il presente lavoro di tesi si ispira ad una delle molteplici applicazioni esistenti dell'interpretazione automatica di sequenze video: la possibilità di rendere una piattaforma MediaSpace "intelligente", capace cioè di adattare automaticamente le immagini trasmesse ai desiderata dell'utilizzatore, in modo trasparente per quest'ultimo, e basandosi sull'interpretazione di ciò che sta avvenendo nella scena filmata. Il progetto è sicuramente ambizioso, e il presente lavoro, lungi dal pretendere di essere una soluzione a questo problema, si configura piuttosto come attenta riflessione, alla luce dello stato dell'arte nel dominio dell'interpretazione automatica di sequenze video, su quali potrebbero essere le strategie vincenti per affrontare e risolvere questa problematica.

Partendo appunto dallo stato dell'arte e dai precedenti numerosi lavori in materia di interpretazione automatica, si è stabilita una architettura globale della piattaforma, conformandosi ad alcune scelte dimostratesi vincenti in altre analoghe piattaforme e riservandosi contemporaneamente anche la prerogativa di intervenire sulla struttura, al fine di introdurre innovazioni capaci di rendere il sistema più performante. Nella fattispecie, si è adottata l'architettura abbastanza classica che prevede la seguente macrodecomposizione funzionale:

- segmentazione dell'immagine da analizzare sulla base di un'immagine di riferimento (sfondo);
- raggruppamento logico dei pixels in movimento in regioni in movimento;
- classificazione delle regioni in movimento;
- inseguimento degli individui;
- interpretazione del loro comportamento.

Al fine di rendere più performante l'algoritmo di classificazione, si è però deciso di ricorrere a due soluzioni aggiuntive:

- come dimostrato da precedenti studi, si rivela di fondamentale importanza poter disporre di una conoscenza a priori della scena nella quale si svolge l'azione. Ne consegue l'adozione di una base del contesto, opportunamente progettata e realizzata;

- vista la precisa applicazione MediaSpace, volta ad interpretare soprattutto il comportamento degli esseri umani, si è pensato di introdurre un modulo capace di analizzare le componenti cromatiche dell'immagine al fine di rilevare la presenza di regioni di pelle. Questa informazione si rivela preziosa per il modulo di classificazione, in quanto è in grado di rendere la classificazione stessa molto più robusta

Infine, si è deciso di concentrare particolarmente l'attenzione sul modulo di inseguimento, con il preciso intento di offrire una soluzione innovativa capace di potenziarne l'affidabilità e la robustezza. La correttezza dei risultati offerti da tale modulo si rivela infatti fondamentale per l'interpretazione che segue.

8.1.0.1 La segmentazione

Riguardo al modulo di segmentazione, l'obiettivo ricercato non è tanto progettare un nuovo algoritmo di riconoscimento del movimento, ma di scegliere oculatamente una metodologia esistente, già testata, ed integrarla opportunamente nella piattaforma d'interpretazione. Il metodo di segmentazione adottato deve essere caratterizzato dalle seguenti proprietà:

- **robustezza**: per gestire in maniera ottimale delle scene tratte in condizioni realistiche;
- **rapidità**: per essere utilizzato nel nostro sistema, che tratta da 3 a 5 immagini al secondo, l'algoritmo deve essere poco costoso dal punto di vista delle risorse di calcolo richieste;
- **di facile implementazione**: per poter essere integrato nel sistema d'elaborazione in un tempo limitato. In particolare, l'algoritmo deve essere caratterizzato da una fase d'inizializzazione e di regolazione dei parametri ridotta.

Come mostrato in [77], esistono essenzialmente quattro tecniche classiche per il rilevamento del movimento:

- **L'approccio basato sulla corrispondenza**: esso consiste nell'estrazione delle primitive 2D (contorni, angoli, motivi della texture...), che vengono in seguito messe in corrispondenza con quelle estratte nell'immagine precedente. Questa fase di corrispondenza si basa su ipotesi a priori sulla struttura delle superfici osservate e sulla natura del movimento. Alcuni articoli che trattano di tecniche affini di stima e rappresentazione del movimento sono [78] e [79].

  Vantaggi della metodologia: rapidità di calcolo.

  Svantaggi della metodologia: utilizzo di ipotesi a priori, non sempre disponibili, scarsità delle primitive 2D caratterizzanti un oggetto in movimento, fenomeni d'instabilità.

- **L'approccio differenziale**: questo approccio si basa sull'ipotesi di base dell'invarianza della luminanza di un punto quando questo si sposta nello spazio campionato tramite la sequenza d'immagini (cf [80]) Questa ipotesi porta alla definizione di una equazione differenziale che lega il gradiente spazio-temporale dell'intensità luminosa con il vettore velocità (equazione conosciuta sotto il nome di Equazione di vincolo del movimento apparente):

$$\overrightarrow{\nabla I} \cdot \overrightarrow{v} + \frac{\partial I}{\partial t} = 0$$

dove $\overrightarrow{\nabla I} = \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}\right)$ è il gradiente spaziale dell'intensità luminosa e $\frac{\partial I}{\partial t}$ il gradiente temporale. Questa equazione permette di misurare la componente del vettore velocità parallela al gradiente d'intensità in ogni punto dell'immagine.

Vantaggi della metodologia: rigore e buoni risultati, a patto che alcune ipotesi sulla funzione luminosità siano verificate.

Svantaggi della metodologia: non prende in considerazione i problemi di oscuramento dovuto alle ombre né i cambiamenti di luminosità, necessita di ipotesi stringenti sulla funzione intensità, spesso non verificate in pratica ed una elevata complessità computazionale.

- **L'approccio per differenza rispetto ad una immagine di riferimento**: questo metodo necessita di una immagine di riferimento, corrispondente al fondo della scena, senza oggetti in movimento. Essa consiste nel calcolare la differenza tra l'immagine in esame e l'immagine di riferimento, al fine di ottenere i punti in movimento (suscettibili di variazini tra le due immagini).

  Vantaggi della metodologia: elevata robustezza (a patto di implementare raffinati algoritmi di aggiornamento dell'immagine di riferimento), semplicità d'implementazione, complessità media dal punto di vista computazionale.

  Svantaggi della metodologia: necessita di algoritmi d'aggiornamento dell'immagine di riferimento, altrimenti dà risultati molto rumorosi, è applicabile solo a scene provenienti da una videocamera fissa.

- **Altri approcci**: altri approcci, sono possibili, per esempio il metodo basato sulla trasformata di Fourier, che consiste nell'utilizzare le proprietà dell'evoluzione spazio temporale del segnale nel dominio delle frequenze.

Nel nostro caso si è adottato il classico algoritmo di differenza tra due immagini, opportunamento completato con una fase aggiuntiva di raggruppamento dei pixels in movimento in "regioni in movimento", ciò allo scopo di svincolarsi quanto prima dall'entità "pixel" per poter lavorare ad un livello di astrazione superiore. La scelta è ricaduta su questo modulo in quanto esso si adatta particolarmente alle caratteristiche della piattaforma:

- elevata robustezza;
- non richiede grandi risorse di calcolo, se programmato in modo curato;
- l'applicazione MediaSpace fa uso di una videocamera fissa.

Al fine di ridurre l'inevitabile rumore che affligge i risultati del modulo di segmentazione, si è implementato un sofisticato algoritmo di aggiornamento dell'immagine di sfondo sulla base delle immagini in analisi, capace di compensare le variazioni di luminosità tra differenti immagini. Il modulo offre risultati soddisfacenti, il rapporto segnale/rumore in uscita è decisamente buono e la robustezza rispetto ai cambiamenti di luminosità più che soddisfacente.

Un punto debole evidenziato è il tempo di elaborazione necessario per processare un'immagine, superiore alle aspettative (approssimativamente pari a 100 ms); questo peggioramento rispetto alle previsioni è da imputarsi ad alcuni errori di programmazione, dovuti

essenzialmente all'inesperienza, errori che possono essere completamente rimossi tramite un'accurata riprogrammazione.

8.1.0.2 Il rilevamento delle regioni color pelle

Il riconoscimento della pelle che è stato implementato si basa sull'analisi cromatica delle immagini. Uno studio (cf [49]) condotto da M.J.Jones e J.M.Rehg propone un'analisi completa e dettagliata dei modelli cromatici sia del color pelle che di altri colori. La conclusione a cui i due studiosi giungono è la marcata superiorità dei modelli a base di istogrammi, per lo spazio cromatico scelto, nell'individuare i pixels di pelle. Di conseguenza ci si è indirizzati, per la nostra applicazione, ad un algoritmo di tipo statistico basato sugli istogrammi. Nello stesso articolo si evidenzia come la distribuzione dei pixels sia color pelle che non di color pelle possa ancora essere modelizzata nello spazio 3D RGB non normalizzato, malgrado gli effetti di variazione di luminosità, a priori sconosciuti. Sulla base di questa problematica evidenziata, si è quindi pensato di introdurre una normalizzazione dello spazio cromatico, per tener conto sia degli inconvenienti dovuti al rumore, sia alla luminosità. Ne è nato uno studio comparativo di tre differenti algoritmi:

- una prima versione che utilizza un istogramma del color pelle nello spazio cromatico RGB;
- una seconda versione che utilizza una distribuzione di probabilità calcolata come rapporto di due istogrammi nello spazio RGB, uno relativo al color pelle, l'altro ai colori dell'immagine di riferimento (sfondo). La ricerca di questo secondo istogramma, rispetto al precedente, è dettata dall'esigenza di migliorare il rapporto $\frac{segnale}{rumore}$.
- una terza versione che utilizza il principio dell'algoritmo precedente, calcolando però i due istogrammi nello spazio cromatico normalizzato (r, g, l) $(r = R/l, g = G/l, l = R + G + B + 1$, luminosità).

I risultati dei test condotti ci hanno portato a scegliere la terza versione dell'algoritmo, per l'immunità dimostrata rispetto alle variazioni della luminosità. In letteratura emerge comunque come lo spazio (r, g) normalizzato sia stato spesso utilizzato per il rilevamento dei volti (cf [38],[40],[47],[50]) soprattutto perché esso consente una riduzione della sensibilità rispetto ai cambiamenti d'illuminazione.

Nonostante l'algoritmo utilizzato sia relativamente semplice, la versione implementata si è rivelata robusta ed affidabile. I risultati ottenuti sono decisamente soddisfacenti, e l'informazione sulla presenza/assenza di regioni color pelle ha permesso di aumentare la robustezza del modulo di classificazione successivo. Ancora, un punto debole si rivela l'importante tempo d'elaborazione necessario per ogni immagine.

8.1.0.3 La classificazione/fusione

Il modulo di classificazione svolge il compito di associare ad ogni regione in movimento individuata nell'immagine un'etichetta che ne definisca, nel modo più preciso e rigoroso possibile, la natura. In altri termini, si vuole rendere disponibile ai moduli successivi (ed essenzialmente al modulo di inseguimento degli oggetti mobili) l'informazione sulla natura dell'oggetto in movimento che stanno trattando: *individuo* piuttosto che *parte di un individuo* (in caso di occultazione) o *veicolo*, *porta*, *rumore*... L'analisi del problema della classificazione può essere trattato partendo dal punto di vista della natura degli oggetti rilevati.

La prima scelta fondamentale riguarda l'utilizzo di una delle due categorie di algoritmi seguenti:

- algoritmi basati sui modelli dell'oggetto da riconoscere;
- algoritmi basati sui modelli di come l'oggetto appare nell'immagine.

All'interno della prima categoria, esistono poi metodi basati sulla modelizzazione di oggetti rigidi piuttosto che di oggetti deformabili (non rigidi). Mentre i primi si adattano bene alla classificazione di oggetti come, per esempio, le automobili, i secondi sono più adatti alla descrizione delle persone. Tuttavia entrambe queste possibilità riguardano la prima delle due classi di algoritmi suddette, scelta poco adatta alla nostra applicazione in quanto relativamente onerosa dal punto di vista del tempo di calcolo richiesto. Inoltre la grande varietà di oggetti classificabili (sia persone che oggetti del mobilio che rumore...) renderebbe necessario disporre di accurati e complessi modelli di ognuno di questi oggetti. Ne risulterebbe una accresciuta complessità sia a livello delle conoscenze a priori necessarie (i modelli) che a livello del tempo di calcolo richiesto per istanziare ognuno di questi modelli nell'immagine, cercandone in più i parametri dell'omotetia che lo caratterizza.

Sulla base di quanto detto, si è deciso di utilizzare modelli di come appaiono nelle immagini i differenti oggetti in movimento da classificare. Tali modelli sono molto meno complessi dei precedenti e gli algoritmi che ne fanno uso offrono comunque risultati soddisfacenti.

I modelli sono contenuti nella base del contesto, e come tali fanno parte della conoscenza a priori della quale dispone l'algoritmo. Esso calcola il grado di aderenza di ogni regione in movimento ai differenti modelli disponibili, per concludere assegnando alla regione l'etichetta dell'oggetto che maggiormente le corrisponde.

L'algoritmo gestisce le occlusioni parziali di un oggetto in movimento (tipicamente un essere umano) da parte di un oggetto dell'arredo. A tal fine risulta preziosa la conoscenza a priori della geometria della scena contenuta nella base del contesto. Infine, un algoritmo di fusione si incarica di raggruppare gli oggetti in movimento corrispondenti, nella realtà, ad uno stesso oggetto. Il modulo è capace di classificare correttamente più dell'80% delle regioni in movimento. Dato il principio alla base dell'algoritmo, esso trova difficoltà a classificare correttamente una regione in movimento le cui caratteristiche si discostino molto da quelle del corrispondente modello. Per poter compensare questa lacuna si è progettato il modulo successivo (d'inseguimento) in modo che possa recuperare, anche solo parzialmente, questo tipo di errori.

8.1.1 L'inseguimento

L'algoritmo d'inseguimento progettato deve ottenere, pur rispettando il vincolo d'elaborazione in tempo reale, un inseguimento robusto e simultaneo di diversi oggetti in movimento, che si spostano seguendo traiettorie indipendenti in un ambiente complesso e ingombro di altri oggetti, quale un ufficio. I problemi principali di un algoritmo d'inseguimento sono la stima del movimento dell'oggetto inseguito (il *bersaglio*) e il trattamento delle situazioni ambigue di corrispondenza tra i bersagli già esistenti e le regioni in movimento rilevate nella nuova immagine. Il valore stesso di un algoritmo d'inseguimento, la sua robustezza, è direttamente dipendente dall'efficacia con cui l'algoritmo gestisce le associazioni ambigue.

Il tipo di scene trattate (ufficio) pone diversi problemi a livello di riconoscimento delle regioni in movimento e alla natura stessa delle regioni rilevate. Per esempio i problemi

sono: i riflessi, le ombre sulle pareti, la mancanza di contrasto (per esempio i controluce), i cambiamenti della luminosità e l'aggiornamento dell'immagine di riferimento.

Come condizione preliminare, si sono scartati tutte le medologie esistenti basate sull'inseguimento degli oggetti rigidi. Come già esposto precedentemente, infatti, le persone, oggetto principale dell'inseguimento, non rientrano in questa categoria di oggetti in movimento. Inoltre, si ha l'esigenza di poter inseguire un individuo anche se esso è parzialmente occultato su alcune immagini della sequenza. Ciò porta automaticamente ad eliminare le metodologie d'inseguimento basate sull'utilizzazione di modelli deformabili del contorno. Infatti esse seguono globalmente la forma dell'oggetto in movimento e permettono di compensare l'assenza temporanea di una parte della forma dell'oggetto in movimento con la presenza dell'altra parte. Tuttavia, l'inseguimento della forma dell'oggetto, anche se globale, può essere profondamente disturbata dall'assenza ripetuta di una parte della forma (per esempio nel caso di occultazione prolungata). Infine, le esigenze di tempo di calcolo ridotto e di utilizzo di una sola videocamera portato a scartare i metodi basati sui modelli dinamici e sull'uso di più videocamere.

Sulla base di quanto detto, si sono allora definiti i principi fondamentali alla base dell'algoritmo di inseguimento progettato:

- esso deve essere in grado di trattare in modo completo ed efficace le situazioni ambigue, come sono in grado di fare gli algoritmi d'inseguimento senza modello. A tal fine si decide di far uso sia di un modello di movimento, sulla base del quale stimare i parametri di merito delle traiettorie manipolate, sia di un modello d'individuo (inteso come successione temporale di oggetti in movimento classificati *persona*) sulla base del quale stimare la corretta corrispondenza individuo-traiettoria;

- il modello di oggetto (persona) in movimento utilizzato è un modello di come appare l'individuo nell'immagine, ed è caratterizzato semplicemente dalle proprie dimensioni e dalla propria posizione.

Inoltre, in modo originale rispetto alle piattaforme d'interpretazione esistenti, si è deciso di assumere un ritardo T tra l'immagine in analisi e i risultati in uscita dal modulo d'inseguimento. Questo ritardo T si rivela fondamentale al fine di confermare, rivedere o correggere, alla luce delle nuove informazioni disponibili, le scelte fatte negli istanti precedenti. In altri termini, al fine di aumentare la robustezza della fase di inseguimento si è deciso di progettare un modulo capace di prendere delle decisioni alla luce di tutte le informazioni provenienti dalla conoscenza delle $T - 1$ immagini precedenti.

Tale ipotesi a priori è resa possibile dalla natura dell'applicazione MediaSpace, in grado di ben sopportare un ritardo di alcuni secondi (nel nostro caso $T = 20$ immagini, pari a circa 7 secondi) tra il verificarsi di un evento e la comprensione dello stesso da parte del sistema. In altre applicazioni tale ritardo potrebbe rivelarsi critico.

I requisiti particolari (cioé, in un certo senso, "aggiuntivi" rispetto a quelli richiesti ad un qualsiasi modulo d'inseguimento) sono:

- la capacità di recuperare, almeno parzialmente, gli errori di classificazione commessi dal corrispondente modulo. A tal fine si ricorre a due modelli: di individuo e di traiettoria, ai quali devono conformarsi rispettivamente ogni oggetto in movimento facente parte di una traiettoria ed ogni serie temporale di oggetti in movimento. Questi modelli consentono di rimettere parzialmente in discussione l'etichetta assegnata dal modulo di classificazione alle diverse regioni in movimento;

- la capacità di gestire i mancati rilevamenti, cioè i casi in cui l'oggetto in movimento corrispondente ad un individuo non sia stato rilevato. A tal fine si è definito l'oggetto

in movimento "non rilevato", che rappresenta una possibilità di estensione per ogni traiettoria esistente, e che costituisce un "sostituto" all'oggetto in movimento in tutti i casi in cui questo non sia stato rilevato (quindi non esista);

- la capacità di correggere l'assegnazione di una traiettoria ad un individuo, nel caso quest'ultima si riveli erronea. A tal fine si è espressamente previsto il ritardo T precedentemente descritto nonché la possibilità di riassegnare ad ogni istante ad un individuo una nuova traiettoria in modo quasi indipendente da quella assegnatagli precedentemente;

- la capacità di gestire le sovrapposizioni di individui; per sovrapposizione si intende il rilevamento di un solo oggetto in movimento che rappresenta però più persone. A tal fine si è previsto l'oggetto in movimento di tipo "gruppo", gestito dal modulo di classificazione, nonché speciali algoritmi di trattamento delle traiettorie intersecantesi, capaci di valutare se l'intersezione deriva da un errore di assegnazione dell'oggetto in movimento alla traiettoria (cioè due traiettorie utilizzano erroneamente lo stesso oggetto in movimento non di tipo "gruppo") oppure se l'intersezione modellizza correttamente due persone che si incrociano nella realtà.

Come illustrato dai risultati del capitolo precedente, l'algoritmo è decisamente robusto rispetto a questo tipo di problematiche, ed offre risultati affidabili e corretti nella maggior parte delle situazioni "normali" ed in un buon numero di situazioni "ardue". L'adozione del ritardo T si rivela una scelta vincente per la robustezza dei risultati, in quanto, pur non ricorrendo ad alcun algoritmo di backtracking, il modulo è capace di correggere le scelte errate effettuate. Tale filosofia potrebbe facilmente essere estesa agli algoritmi di inseguimento di altre applicazioni, nell'ipotesi che queste non siano sottoposte a rigidi vincoli riguardo al ritardo massimo con cui devono essere forniti i risultati.

La maggior debolezza del modulo d'inseguimento risiede nell'incapacità di distinguere tra un oggetto in movimento originato da un oggetto dell'arredo che è stato spostato piuttosto che da un individuo che si sta muovendo. Tale incapacità è la causa della maggior parte degli errori d'inseguimento commessi, errori, purtroppo, non recuperabili nemmeno sfruttando la finestra temporale T. Si vedrà nella sezione seguente una tecnica che consente di risolvere questo problema.

L'algoritmo progettato si basa su di un certo numero di parametri, alcuni inseriti nell'algoritmo stesso, altri facenti parte della base del contesto, la cui regolazione è affidata all'operatore umano. In altri termini, alla fase di progetto e realizzazione segue una fase di test nella quale si procede ad una regolazione accurata dei parametri dell'algoritmo, al fine di ottenere un buon funzionamento dello stesso. Di fronte a questa necessità, è interessante chiedersi se non valga la pena di studiare la possibilità di implementare un algoritmo di tipo adattativo, capace di tarare autonomamente i propri parametri di funzionamento in base ai risultati ottenuti e, soprattutto, di variarli in modo dinamico adattandoli alle (nuove) condizioni di un ambiente le cui caratteristiche variano nel tempo. Tale possibilità è stata esaminata approfonditamente all'inizio della fase di progetto. Essa è stata abbandonata per due motivi principali: sebbene sia in grado di fornire risultati più robusti rispetto ad un algoritmo che opera ad "anello aperto", un algoritmo adattativo è decisamente più oneroso sia dal punto di vista del tempo di sviluppo necessario, che dal punto di vista del tempo di elaborazione richiesto. Essendo queste due cifre di merito importanti per il nostro progetto, si è deciso di ripiegare su una algoritmo meno versatile, di tipo non adattativo, appunto. L'eventuale aggiunta di un "anello di retroazione" potrebbe essere una interessante possibilità di sviluppo futuro.

8.1.2 L'interpretazione

Per la piattaforma realizzata è stato utilizzato un modulo d'intepretazione esistente. L'algoritmo si basa sul riconoscimento degli eventi che si verificano nella scena. Questi eventi concorrono poi ad aggiornare gli scenari "parzialmente riconosciuti" gestiti dal sistema. Quando tutti gli eventi che descrivono uno scenario si sono verificati, con il corretto ordinamento temporale, lo scenario è riconosciuto. Il modulo utilizzato è capace di interpretare correttamente gli scenari tipici dell'applicazione MediaSpace.

8.2 Prospettive di miglioramento

In questa sezione, e limitatamente ai tre moduli di classificazione/fusione, di riconoscimento delle regioni color pelle e di inseguimento, si descriveranno le soluzioni applicabili per risolvere i principali problemi individuati. Tali soluzioni possono tracciare le linee guida per un eventuale proseguimento del presente progetto.

8.2.1 La classificazione

Una possibile evoluzione del modulo di classificazione potrebbe essere in una (o più) delle seguenti direzioni:

- al fine di poter correttamente classificare gli oggetti dell'arredo suscettibili di spostarsi (sedie, porte...) si potrebbe introdurre nella base del contesto la proprietà "suscettibile di essere spostato" per ognuno di questi oggetti, ed eventualmente anche la zona nella quale più probabilmente ci si aspetta l'oggetto in questione (per esempio, per la porta il luogo del movimento è circoscritto e invariabile). Il modulo di classificazione potrebbe allora confrontare la posizione di ogni regione in movimento suscettibile di essere un oggetto dell'arredo con la posizione di questi oggetti, al fine di classificarlo come tale. In altri termini, si tratta di completare il modello dell'oggetto dell'arredo con la proprietà relativa alla possibilità di spostarsi. Tale informazione sarebbe di grande utilità per il modulo d'inseguimento, che potrebbe allora gestire correttamente i casi in cui un oggetto in movimento originato da un elemento dell'arredo viene confuso con uno generato da una persona (come evidenziato nell'ultima sequenza presentata nel capitolo precedente).

- Attualmente l'algoritmo elabora ogni immagine in modo indipendente dalle precedenti. In altri termini, i risultati ottenuti nell'immagine precedente non vengono utilizzati in alcun modo nell'analisi dell'immagine corrente. Si potrebbe pensare di estendere l'analisi anche all'asse temporale, confrontando la posizione di ogni nuova regione in movimento da classificare con la posizione degli oggetti in movimento classificati all'istante precedente. Ne deriva immediatamente un guadagno in termini di robustezza e di coerenza temporale tra la classificazione delle differenti immagini.

- Infine, si potrebbe estendere il principio della ricerca del color pelle ad altri colori, caratteristici di alcuni oggetti. Per esempio, si potrebbe cercare di capire se una regione in movimento possiede le caratteristiche cromatiche corrette per essere una sedia o una porta piuttosto che una persona. Tali caratteristiche cromatiche completerebbero allora il modello dei differenti oggetti in movimento e sarebbero quindi contenute nella base del contesto.

8.2.2 Il rilevamento delle regioni color pelle

Nell'ottica di migliorare l'algoritmo di rilevamento delle regioni color pelle, oltre che in vista di un'applicazione dello stesso principio per rilevare differenti colori, le linee guida possono essere:

- un primo tentativo può essere la riduzione del tempo di elaborazione necessario per ogni immagine. In questo senso si può pensare di ridurre il numero degli accessi in lettura e scrittura per ogni pixel dell'immagine. Invece che percorrere l'immagine più volte, effettuando una sola operazione elementare ad ogni passaggio, si potrebbe cercare di fattorizzare queste operazioni in un solo passaggio, guadagnando sul tempo di accesso alla memoria (di massa). Una diversa codifica delle immagini, sotto forma di segmenti invece che di punti, sembra rappresentare una soluzione promettente.

- Sempre allo scopo di ridurre il tempo d'elaborazione, si potrebbe cercare di raggruppare in una stessa procedura due analisi differenti dell'immagine: la ricerca di pixels color pelle e la ricerca di pixels in movimento. In particolare, si potrebbe immaginare di percorrere l'immagine una sola volta, stabilendo se il pixels in analisi è "in movimento" e, in questo caso, analizzandone le componenti cromatiche per stabilire se il suo colore è il color pelle. In questo modo la ricerca del color pelle verrebbe effettuata solo per i pixels che si trovano in movimento, con un evidente risparmio di tempo di calcolo.

- Dal punto di vista della qualità dei risultati ottenuti, una prima possibilità è la ricerca di uno spazio cromatico differente nel quale condurre la ricerca del color pelle. Spazi cromatici diversi potrebbero portare a risultati più affidabili (nell'ipotesi che le caratteristiche del color pelle siano messe maggiormente in evidenza da questo nuovo spazio rispetto allo spazio (r, g, l) utilizzato).

- infine si potrebbe costruire una biblioteca di immagini color pelle più estesa, contenente anche pixels di pelle diversa dalla pelle di ceppo caucasico. Un miglioramento importante, ma sicuramente costoso dal punto di vista del tempo di elaborazione necessario, potrebbe essere l'introduzione di un algoritmo adattativo, che possa modificare il modello del color pelle "di riferimento" sulla base della pelle rilevata in ogni immagine all'analisi. Questa soluzione avrebbe come vantaggio immediato la limitazione dei problemi di cambiamento di luminosità e di pelli differenti (nera/bianca) presenti contemporaneamente nell'immagine..

Al prezzo di qualche modifica, lo stesso algoritmo potrebbe essere utilizzato anche per effettuare l'identificazione delle persone presenti nella scena, sulla base del colore dominante dei loro abiti per quel giorno.

8.2.3 L'inseguimento degli individui

L'algoritmo di inseguimento progettato è stato creato al fine esplicito di inseguire degli individui in movimento all'interno di una scena (tipicamente un ufficio). Ciononostante il principio base può essere esteso anche ad altre applicazioni. Qui di seguito elenchiamo tre direzioni di ricerca percorribili al fine di aumentare ulteriormente la robustezza e l'affidabilità dell'algoritmo.

- L'algoritmo d'inseguimento si basa, tra le altre cose, sui modelli di individuo e di traiettoria. Questi modelli rappresentano delle conoscenze a priori. Potrebbe essere

interessante introdurre un terzo modello che prenda in considerazione gli oggetti dell'arredo che possono spostarsi (come già consigliato per il modulo di classificazione). Questo terzo modello impedirebbe all'algoritmo di scegliere un oggetto mobile di questa classe per prolungare una traiettoria.

- Per risolvere i problemi (principalmente di scambio di identità) derivanti da due o più persone che si incrociano, si rivela necessario poter identificare in modo univoco e preciso ogni individuo presente nella scena. In altri termini, poter risalire all'identità dell'individuo. Questo identificatore (voce? colore degli abiti? login nel sistema informatico?) potrebbe completare il modello di essere umano utilizzato e permetterebbe di scegliere correttamente, ad ogni istante, l'oggetto in movimento corretto corrispondente ad ogni individuo.

- Come già introdotto nella sezione precedente, sarebbe interessante completare l'algoritmo, che attualmente opera ad "anello aperto", con un modulo di retroazione, capace di trasformarlo in un algoritmo adattativo. L'algoritmo sarebbe allora in grado di regolare autonomamente i propri parametri significativi, sollevando dal compito di regolazione l'operatore umano e soprattutto acquisendo la capacità di adattare il proprio funzionamento ad una ambiente le cui caratteristiche variano nel corso del tempo.

8.3 Conclusioni

Alla luce di quanto detto, si comprende come il lavoro svolto costituisca soprattutto uno spunto di riflessione e una fonte di nuove idee per riprendere e continuare la ricerca nel dominio dell'interpretazione automatica di sequenze video. Lo sviluppo di questo progetto è stato sempre governato dal duplice tentativo di utilizzare al meglio tutto ciò che è stato ideato, sviluppato e testato da altri ricercatori del settore, mantenendo però sempre un occhio critico capace di suggerire miglioramenti, cambiamenti o la possibilità di introdurre uno spunto personale. Solo a queste condizioni, infatti, il presente lavoro di tesi si sarebbe potuto considerare un mattoncino utilizzabile per edificare il prossimo, e sicuramente più grandioso, edificio.

Appendice A

Glossario dei termini utilizzati

In questa appendice si elencano i principali termini utilizzati nella presente trattazione.

adiacenza La regola assunta per stabilire quali siano i pixels adiacenti ad un pixel prefissato. Molteplici sono le definizioni di adiacenza possibili, due le principali (adiacenza a 8 pixels e adiacenza a 4 pixels). Nella ricerca oggetto della presente trattazione si è sempre fatto riferimento all'adiacenza a 8 pixels.

adiacenza a 4 pixels Sono definiti come adiacenti al pixel in esame i quattro pixels a contatto con esso nelle quattro direzioni Nord, Est, Sud, Ovest del piano dell'immagine.

adiacenza a 8 pixels Sono definiti come adiacenti al pixel in esame tutti gli otto pixels immediatamente a contatto con esso nelle otto direzioni Nord, Nord-Est, Est, Sud-Est, Sud, Sud-Ovest, Ovest, Nord-Ovest del piano dell'immagine.

aree IU, aree di ingresso/uscita Esse rappresentano le zone tridimensionali, definite nel contesto 3D, dalle quali gli individui possono entrare/uscire dall'ufficio (nel nostro caso, essenzialmente la porta). Esse sono di fondamentale importanza in quanto permettono di individuare una zona sull'immagine corrispondente grossomodo all'immagine della porta (in questa zona ci si aspetta di vedere entrare ed uscire gli individui).

biblioteca di pelle Indica l'insieme delle immagini utilizzate per costruire l'istogramma approssimante $\mathcal{M}$. Tali immagini contengono ovviamente solo pixels di pelle.

coefficiente di qualità (di una traiettoria): cifra di merito associata ad ogni struttura "traiettoria" calcolata dall'algoritmo di inseguimento degli individui. Il coefficiente di qualità esprime l'aderenza della traiettoria sia al modello di movimento umano, sia al modello di essere umano di ogni oggetto in movimento che la compone, sia la coerenza spaziale degli oggetti in movimento che la costituiscono, l'uno rispetto agli altri. È un parametro fondamentale per poter scegliere, tra tutte le traiettorie associate ad ogni individuo, quella che meglio descrive il suo spostamento fisico.

contesto 3D Descrizione geometrica tridimensionale dello spazio fisico che costituisce la scena ove sono riprese le immagini da interpretare: un ufficio nella presente trattazione. La descrizione comprende gli elementi strutturali della scena (muri, colonne, pavimento, soffitto, porte) e gli oggetti presenti in essa (sedie, tavoli, scrivanie, armadi, lavagne, postazioni di lavoro...). Completano il contesto le informazioni che

descrivono l'acquisizione della sequenza delle immagini (per esempio il tipo di videocamera, la sua posizione in un sistema di riferimento cartesiano, la matrice di calibrazione che la caratterizza).

inseguimento degli oggetti in movimento Traduce il termine inglese "tracking"; indica la descrizione spazio temporale (quindi all'interno della scena e nell'intervallo di tempo individuato dalla sequenza di immagini elaborata) dei movimenti di un individuo all'interno della scena.

livello di disponibilità Per l'utilizzatore di una piattaforma mediaspace, è un indicatore della disponibilità personale ad essere interrotto (da telefonate/richieste di videoconferenza) che l'utente stesso (o il sistema intelligente, su programmazione dell'utente) rende noto agli altri utilizzatori del sistema. Ciò permette di avere sempre associata all'immagine di ogni utilizzatore mediaspace un indicatore (pastiglia colorata, stringa di testo...) che ne indica la disponibilità all'interazione. Tipicamente tale indicatore di disponibilità potrebbe assumere il valore minimo (pastiglia rossa, oppure stringa "non disturbare") in caso di riunione, e il valore massimo (pastiglia verde, oppure stringa "utente disponibile") se l'utente svolge attività di routine poco impegnativa.

mediaspace Si intende con mediaspace una connessione video permanente a banda ridotta (tipicamente tra 0.2 e 1 immagine/secondo) capace di legare più utilizzatori su più siti, anche geograficamente separati. Il termine è indifferentemente utilizzato per indicare il concetto stesso come le piattaforme hardware/software che lo implementano ([8]). Il termine è stato introdotto agli inizi degli anni '90 per indicare qualsiasi mezzo (mediatico) per facilitare la comunicazione informale e la consapevolezza di appartenenza ad uno stesso gruppo per individui geograficamente separati.

oggetto in movimento Indica una regione in movimento dopo la fase di classificazione, cioé una regione in movimento corredata da un'etichetta che ne specifica la natura: *individuo, individuo occultato, gruppo di persone, oggetto dell'arredo, rumore, indeterminato*. Dell'oggetto in movimento si seguono gli spostamenti nel tempo e se ne analizza il comportamento. Può essere costituito da una o più regioni in movimento.

pcp, pixel color pelle Indica un pixel dell'immagine che rappresenta un lembo di pelle (cioè un pixel dell'immagine che ha il colore tipico della pelle). Un pcp è l'unità fondamentale alla base di una rcp.

rcp, regione color pelle Indica una regione dell'immagine che rappresenta un lembo di pelle (per esempio un braccio, o la testa di una persona).

regione connessa Un insieme di pixels in un'immagine, accomunati da una caratteristica (per esempio avere probabilità di rappresentare un pixel di pelle superiore a X, oppure essere stati riconosciuti come in movimento) e adiacenti secondo la definizione di adiacenza convenzionalmente scelta.

regione in movimento Una regione connessa in un'immagine che rappresenti qualsiasi cosa sia stata oggetto di una variazione rispetto o all'immagine immediatamente precedente, o ad un'immagine di riferimento. Può corrispondere sia a del rumore, per esempio un riflesso o un'ombra, sia ad un oggetto della scena, per esempio un individuo.

regione in movimento di prima classe Una regione in movimento corrispondente ad uno spostamento di un individuo.

regione in movimento di seconda classe Una regione in movimento corrispondente ad uno spostamento di un oggetto.

regione in movimento di terza classe Una regione identificata come in movimento, ma rappresentante una parte dell'immagine non sede di un effettivo spostamento.

segmentazione Indica l'estrazione, da un'immagine, delle regioni in movimento che essa contiene. Nell'accezione più generale i pixel in movimento che compongono la regione possono essere determinati o per differenza rispetto all'immagine immediatamente precedente nella sequenza video oppure rispetto ad una immagine di riferimento (come nella ricerca oggetto della presente trattazione).

scena Corrisponde al volume 3D filmato dalla videocamera (nel nostro caso l'ufficio), inteso come ambiente statico (gli oggetti 3D che definiscono i volumi dell'ufficio e i rispettivi arredi, così come descritti nel contesto), così come gli oggetti mobili che vi si trovano.

scenario Con questo termine si intende la descrizione, espressa utilizzando un formalismo adeguato, dell'insieme degli eventi, uniti agli adeguati vincoli temporali, che permettono di esprimere univocamente il verificarsi di una situazione ben determinata. Un esempio di situazione potrebbe essere un collega che entra nell'ufficio dell'utente del sistema mediaspace per discutere con lui. Si vorrebbe allora che il sistema di interpretazione comprendesse che ciò che accade nell'ufficio richiede una trasmissione delle immagini parzialmente mascherata (per esempio, perché il collega non ama essere ripreso da una videocamera).

spazio privato Nell'ambito di un utente di una piattaforma mediaspace, si intende con questo termine l'insieme degli oggetti e degli spazi che l'utente stesso percepisce, più o meno consciamente, come personali, e sui quali quindi desidera limitare le intrusioni altrui. La percezione di un oggetto/spazio come "privato" è strettamente personale; un utente può ritenere privati oggetti o spazi che un altro utente non ritiene tali. Un tipico esempio di spazi privati potrebbe annoverare la scrivania di lavoro, tutti gli oggetti che vi si trovano, i cassetti, lo schermo del computer, il computer stesso, la sedia.

Appendice B

Articolo ICIP2001

Nelle pagine che seguono si riporta il testo integrale dell'articolo che sarà presentato alla conferenza internazionale ICIP2001 (International Conference on Image Processing) nell'ottobre 2001, come confermato agli autori in via ufficiale in data 13 aprile 2001 (articolo numero 2437). Il presente articolo sarà pubblicato nello stesso mese nei *Proceedings* relativi a tale conferenza.

Qui la prima pagina dell'articolo...

Qui la seconda...

Qui la terza...

E qui la quarta.

Appendice C

L'INRIA e l'équipe ORION

C.1 L'INRIA

Il progetto di ricerca descritto nel presente documento è stato svolto in seno all'équipe ORION, una delle équipes della sede di Sophia Antipolis dell'INRIA. INRIA è l'acronimo di **I**nstitut **N**ational de **R**echerche en **I**nformatique et en **A**utomatique), Istituto Nazionale di Ricerca in Informatica ed Automatica.

C.1.1 Presentazione generale

Creata nel 1967 à Rocquencourt, nelle vicinanze di Parigi, l'INRIA è un ente pubblico a carattere scientifico e tecnologico, posto sotto la duplice tutela del ministero della ricerca francese e del segretariato di Stato per l'industria.

Secondo lo statuto dell'INRIA, gli scopi di quest'ente di ricerca sono:

- intraprendere ricerca pura ed applicata;
- realizzare sistemi sperimentali;
- organizzare scambi scientifici internazionali;
- garantire la trasmissione e la diffusione delle conoscenze e delle esperienze;
- contribuire alla valorizzazione dei risultati della ricerca;
- contribuire, specialmente nel campo della formazione, a programmi di cooperazione e sviluppo;
- costituire punto di riferimento nella consulenza tecnologica;
- contribuire alla standardizzazione delle tecnologie e metodologie.

C.1.2 Le unità di ricerca

L'INRIA ha cinque sedi, ognuna localizzata in una diversa regione:

- nell'Ile-de-France, a Rocquencourt (dove si trova anche la sede giuridica dell'istituto);
- in Bretagna, a Rennes (questa sede venne creata nel 1980);
- in Provence-Alpes-Côte d'Azur, a Sophia Antipolis (creata nel 1982);

- in Lorraine, a Nancy (creata nel 1984);
- in Rhône-Alpes, a Grenoble (creata nel 1992).

C.2 L'unità di ricerca di Sophia Antipolis

Al centro della più grande tecnopoli europea, la sede di Sophia Antipolis conta su di un potenziale di 400 persone, di cui 140 assunti a tempo indeterminato come funzionari di ricerca. Questa sede vanta numerose collaborazioni con i diversi laboratori dislocati nella regione, così come con il mondo dell'industria, come ben provano le quattro società tecnologiche creata da ex-ricercatori INRIA.

C.2.1 Le équipes della sede di Sophia Antipolis

Il centro di Sophia Antipolis è organizzato secondo quattro grandi temi di ricerca:

1. Reti e sistemi (5 équipes);
2. Ingegneria del software e calcolo simbolico (5 équipes);
3. Relazioni uomo-macchina, immagini, dati e conoscenze (6 équipes);
4. Simulazione ed ottimizzazione di sistemi complessi (8 équipes);

Si elencano brevemente le équipes che fanno parte di ogni tema di ricerca, specificando per ognuna le precise tematiche di ricerca.

C.2.1.1 Primo tema di ricerca: reti e sistemi

- MASCOTTE - Metodi algoritmici, simulazione, tecniche combinatorie e ottimizzazione per le telecomunicazioni;
- MEIJE - Parallelismo, sincronizzazione e calcolo tempo reale;
- MISTRAL - Modellizzazione per l'informatica e per i sistemi di comunicazione: ricerca ed applicazioni software;
- RODEO - Reti ad alta velocità, reti aperte;
- TROPICS - Trasformazioni e tools informatici per il calcolo scientifico.

C.2.1.2 Secondo tema: Ingegneria del software e calcolo simbolico

- CAFE - Calcolo formale ed equazioni;
- LEMME - Software per la matematica;
- OASIS - Oggetti attivi, semantica, internet e sicurezza;
- PRISME - Geometria, algoritmi e robotica;
- SAGA - Sistemi algebrici, geometria e applicazioni;

C.2.1.3 Terzo tema: Relazioni uomo-macchina, immagini, dati e conoscenze

- ACACIA - Acquisizione di conoscenze per l'assistenza alla progettazione tramite interazione tra agenti;
- ARIANA - Problemi inversi nell'osservazione della terra;
- EPIDAURE - Immagini e robotica per la biomedica;
- ORION - Ambienti intelligenti per la soluzione delle problematiche nell'ambito dei sistemi autonomi;
- ROBOTVIS - Visione per mezzo del calcolatore, robotica;

C.2.1.4 Quarto tema: Simulazione ed ottimizzazione di sistemi complessi

- CAIMAN - Calcolo scientifico, modellizzazione e analisi numerica;
- COMORE - Controllo e modellizzazione delle risorse rinnovabili;
- ICARE - Strumentazione, comando e architettura di robot evoluti;
- MIAOU - Matematica e informatica applicate all'automatica e all'ottimizzazione verso l'utilizzatore;
- OMEGA - Metodi numerici probabilistici;
- SINUS - Simulazioni numeriche nell'ambito delle tecnologie ingegneristiche;
- SYSDYS - Sistemi dinamici stocastici.

C.3 ORION - Ambienti intelligenti per la soluzione delle problematiche nell'ambito dei sistemi autonomi

L'équipe ORION è un équipe pluridisciplinare, impegnata nella ricerca nei domini della visione per mezzo del calcolatore, dei sistemi a base di conoscenze e dell'ingegneria del software. Le ricerche portate avanti dall'équipe seguono due direzioni: l'interpretazione automatica di sequenze d'immagini e la riutilizzazione e il pilotaggio automatico di programmi.

C.3.1 Gli assi di ricerca

Questi obbiettivi sono perseguiti sviluppando sia dei linguaggi d'espressione di conoscenze, sia dei meccanismi d'apprendimento e di trattamento di queste conoscenze, il tutto declinato verso classi ben specifiche di problemi.

L'équipe ORION si focalizza sullo studio delle conoscenze che intervengono nelle due tipologie di problemi studiati:

- la conoscenza degli oggetti e degli scenari da manipolare/riconoscere per poter interpretare in via automatica delle sequenze d'immagini;
- la conoscenza dei programmi e del loro utilizzo al fine di realizzarne un pilotaggio automatico.

Più specificatamente l'équipe ricerca nel campo delle tecniche di rappresentazione di conoscenze ibride (a base di frames, di reti semantiche e di regole di produzione), così come le tecniche di pianificazione. Per il raffinamento delle basi di conoscenze, Orion studia delle tecniche di apprendimento simbolico.

Le applicazioni di principale interesse per l'équipe provengono principalmente dal dominio della visione per mezzo del computer.

C.3.2 Relazioni internazionali ed industriali

- Partecipazione al progetto europeo ASTHMA per il riconoscimento automatico delle immagini dei pollini. Partecipanti: la casa farmaceutica Zambon (Italia), ACRI (Francia), le università di Barcellona (Spagna), Cordoba (Spagna), Clermont (Francia), il CHU du Nizza (Francia), il FISBAT (Italia) e il PAMOC (Francia).

- Partecipazione al progetto europeo IST ADVISOR per l'interpretazione video di sequenze d'immagini provenienti dalle stazioni della metropolitana. Partecipanti: Racal Research (Regno Unito), Bull (Francia), The University of Reading (Regno Unito), King's College (Regno Unito) et Vigitec (Belgio).

- Cooperazione con Bull nell'ambito dell'azione di sviluppo Dyade Télescope per la progettazione di un sistema intelligente di videosorveglianza.

- Cooperazione con il progetto Helix (Grenoble) nell'ambito dell'azione di sviluppo Genostar.

- Cooperazione, nell'ambito del progetto Mediaspace, con le équipes Prima (Gravir/INPG Grenoble) e IIHM (Clips/Imag Grenoble).

- Cooperazione con l'Inra/Urih (Sophia Antipolis) nell'ambito del progetto Horticol, Colors INRIA.

- Cooperazione con Action Tick (Inria Sophia), per i progetti Sports et Rainbowă d'I3S (Sophia Antipolis) nell'ambito di Dicosbac, Colors Inria.

- Cooperazione con l'ENSI, Unita di ricerca GRIFT/ASI di Tunisi (Tunisia) nell'ambito delle cooperazioni franco-tunisine Inria/Institut Français de Coopérations.

- Cooperazione con l'università del Maryland (USA).

Bibliografia

[1] W. Gaver, A. Sellen, C. Heath, and P. Luff, "One is not enough: Multiple views on a mediaspace," in *proc. of INTERCHI'93*, 1993, pp. 335–341.

[2] W. Gaver, G. Smets, and K. Overbeeke, "A virtual window on a mediaspace," in *proc. of CHI'95*, 1995, pp. 257–264.

[3] M. Roseman and S. Greenberg, "Groupkit: A groupware toolkit for building real-time conferencing applications," in *proc. of CSCW'92*, 1992, pp. 43–50.

[4] A. Sellen, "The effects of mediating talk with technology," *Human Computer Interaction*, vol. 10(4), pp. 401–444, 1995.

[5] J.C. Tang and M. Rua, "Montage: Providing teleproximity for distributed groups," in *proc. of CHI'94, Conference on Computer Human Interaction*, 1994, pp. 37–43.

[6] M. Mantei, R.M. Backer, A. Sellen, W. Buxton, T. Milligan, and B. Wellman, "Experiences in the use of mediaspace," in *proc. of CHI'91 Conference on Human Factors in Computing Systems*, 1991, pp. 203–208.

[7] R.S. Fish, R.E. Kraut, R.W. Root, and R.E.Rice, "Evaluating video as a technology for informal communications," in *proc. of CHI'92 Conference on Human Factors in Computing Systems*, 1992, pp. 37–47.

[8] J.Coutaz et al., "Early experience with the mediaspace comedi," *Engineering HCI*, 1998.

[9] R. Stults, "Mediaspace," *Rapport technique Xerox PARC*, 1986.

[10] S. Hudson and I. Smith, "Techniques for addressing fundamental privacy and disruption tradeoffs in awareness support systems," in *proc. of CSCW'96*, 1996, pp. 248–257.

[11] A. Lee, A. Girgensohn, and K. Schlueter, "Nynex portholes: Initial user reactions and redesign implications," in *proc. of GROUP'97 International Conference on Supporting Group work*, 1997, pp. 248–257.

[12] F. Brémond and M. Thonnat, "Issues in representing context illustrated by scene interpretation applications," in *proc. of the Int'l and Interdisciplinary Conf. on Modeling and Using Context (CONTEXT-97)*, Rio de Janeiro, Feb. 1997.

[13] F. Brémond and M. Thonnat, "A context representation for surveillance systems," in *Proc. of the Workshop on Conceptual Descriptions from Images at the European Conference on Computer Vision (ECCV)*, Cambridge, April 1996.

[14] M. Thonnat and N. Rota, "Image understanding for visual surveillance application," in *Third international workshop on cooperative distributed vision CDV-WS'99*, UK British Machine Vision Conference, Guildford, Ed., Kyoto, Japan, Nov. 1999, pp. 51–82.

[15] M. Mohnhaupt and B. Neumann, "Understanding object motion: Recognition, learning and spatiotemporal reasoning," Research Report FBI-HH-B-145/90, University of Hamburg, 1990.

[16] H. H. Nagel, "From image sequences towards conceptual descriptions," *Image and Vision Computing*, vol. 6, no. 2, pp. 59–74, May 1988.

[17] D. Corrall, "Deliverable 3: Visual monitoring and surveillance of wide-area outdoor scenes," Tech. Rep., Esprit Project 2152: VIEWS, June 1992.

[18] S. Sellam and A. Boulmakoul, "Intelligent intersection: Artificial intelligence and computer vision techniques for automatic incident detection," *Artif. Intell. Applic. to Traffic Engng*, pp. 189–200, 1994.

[19] H. Nagel, "The representation of situations and their recognition from image sequences," in *Proc. of the RFIA Lyon Villeurbanne*, Nov. 1991, pp. 1221–1229.

[20] D. Hutber, *Suivi multi-capteurs de cibles multiples en vision par ordinateur, appliqué à un véhicule dans un environnement routier*, Ph.D. thesis, I.N.R.I.A., Sophia Antipolis, Nov. 1995, in english.

[21] N. Chleq and M. Thonnat, "Realtime image sequence interpretation for surveillance applications," in *Proc. of the IEEE International Conference on Image Processing, ICIP*, Lausanne, Switzerland, September 1996, pp. 801–804.

[22] C. Castel, L. Chaudron, and C. Tessier, "What is going on? A high level interpretation of sequences of images," in *Proc. of the ECCV'96 workshop on Conceptual Descriptions from Images*, University of Cambridge, Apr. 1996.

[23] A. Sato, K. Mase, A. Tomono, and K. Ishii, "Pedestrian counting system robust against illumination changes," in *Proc. of Visual Communications and Image Processing'93*, Massachusetts, Nov. 1993.

[24] S. Choi, Y. Seo, H. Kim, and K. Hong, "Where are the ball and players? Soccer game analysis with color-based tracking and image mosaik," in *ICIAP'97*, 1997, to appear.

[25] S. Intille and A. Bobick, "Closed-world tracking," in *Proc. of the 5th Int'l Conference on Computer Vision (ICCV)*, Cambridge, MA, June 1995.

[26] G. Herzog, "From visual input to verbal output in the visual translator," Projet VITRA 124, Universität des Saarlandes, Saarbrücken, Germany, 1995.

[27] A. Pentland, "Machine understanding of human action," in *proc. of the 7th Int'l Forum on Frontier of Telecommunication Technology*, Tokyo, Nov. 1995.

[28] A. Bobick and C. Pinharez, "Using approximate models as source of contextual information for vision processing," in *proc. of the IEEE workshop on Context-Based Reasoning (ICCV'95)*, Apr. 1995.

[29] D. Becker and A. Pentland, "Using a virtual environment to teach cancer patients, t'ai chi, relaxation and self-imagery," in *proc. of the ACM Siggraph Symposium on Interactive 3D Graphics*, Providence, RI, USA, 1997.

[30] L. Campbell and A. Bobick, "Recognition of human body motion using phase space constraints," in *proc. of the fifth International Conference on Computer Vision*, Cambridge MA, 1995, pp. 624–630.

[31] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *proc. of CVPR*, Puerto Rico, USA, 1997.

[32] P. Grandjean, *Perception multisensorielle et interprétation de scènes*, Ph.D. thesis, LAAS - Université Paul Sabatier de Toulouse, 1991.

[33] E. André, G. Herzog, and T. Rist, "On the simultaneous interpretation of real world image sequences and their natural language description: the system SOCCER," in *proc. of ECAI*, Munich, Germany, Aug. 1988, pp. 449–454.

[34] G. Herzog, C. Sung, E. André, W. Enkelmann, H. Nagel, T. Rist, W. Wahlster, and G. Zimmermann, "Incremental natural language description of dynamic imagery," Projet VITRA 58, Universität des Saarlandes, Saarbrücken, Germany, 1989.

[35] G. Retz-Schmidt, "Recognizing intentions, interactions, and causes of plan failures," Projet VITRA 77, Universität des Saarlandes, Saarbrücken, Germany, 1991.

[36] J. Schirra and E. Stopp, "Antlima a listener model with mental images," *IJCAI, Chambéry, France*, vol. 1, pp. 175–180, 1993.

[37] M. Bogaert, N. Chleq, P. Cornez, C. Regazzoni, A. Teschioni, and M. Thonnat, "The PASSWORDS project," in *proc. of the Int'l Conf. on Image Processing (ICIP)*, Lausanne (Suisse), Sept. 1996.

[38] S.-H.Kim, N.-K.Kim, S.C.Ahn, and H.-G.Kim, "Object oriented face detection using range and color information," in *Proceedings of the Third International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 76–81.

[39] J-C.Terrillon, M.David, and S.Akamatsu, "Automatic detection of human faces in natural scene images by use of a skin color model and of invariant moments," in *Proceedings of the Third International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 112–117.

[40] Q.B.Sun, W.M.Huang, and J.K.Wu, "Face detection based on color and local simmetry information," in *Proceedings of the Third International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 130–135.

[41] Q.Chen, H.Wu, and M.Yachida, "Face detection by fuzzy pattern matching," in *proc. of the 5th International Conference on Computer Vision*, 1995, pp. 591–596.

[42] J.Yang, W.Lu, and A.Waibel, "Skin color modeling and adaptation," in *Proceedings ACCV*, 1998, pp. 687–694.

[43] D.A.Forsyth, M.Fleck, and C.Bregler, "Finding naked people," in *proc. of 4th ECCV*, 1996, pp. 593–602.

[44] H.Rowley, S.Baluja, and T.Kanade, "Neural network-based face detection," in *proc. of CVPR*, 1996, pp. 203–208.

[45] T.S.Jebara and A.Pentland, "Parametrized structure from motion for 3d adaptative feedback tracking of faces," in *proc. Computer Vision and Pattern Recognition*, 1997, pp. 144–150.

[46] S.J.McKenna, S.Gong, and Y.Raja, "Modeling facial color and identity with gaussian mixture," in *Pattern Recognition*, 1998, pp. 1883–1892.

[47] B.Schiele and A.Waibel, "Gaze tracking based on face-color," in *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, 1995, pp. 344–349.

[48] R.Kjeldsen and J.Kender, "Neural network-based face detection," in *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 312–317.

[49] M.J.Jones and J.M.Rehg, "Statistical color models with applications to skin detection," in *proceedings of Computer Vision and Pattern Recognition*, 1999, pp. 274–280.

[50] H.P.Graf, E.Cosatto, D.Gibbon, M.Kocheisen, and E.Petajan, "Multi-modal sistem for locating heads and faces," in *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 88–93.

[51] K.Sobottka and I.Pitas, "Segmentation and tracking of faces in color images," in *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 236–241.

[52] D.Saxe and R.Foulds, "Toward robust skin identification in video images," in *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 379–384.

[53] F. Brémond, *Image sequence interpretation for video-surveillance applications*, Ph.D. thesis, Institut National de Recherche en Informatique et Automatique (INRIA) in Sophia Antipolis, France, 1997.

[54] J. Woodfill and R. Zabih, "An algorithm for real-time tracking of non-rigid objects," in *proc of AAAI*, Stanford, 1991.

[55] A. Baumberg and D. Hogg, "An adaptive eigenshape model," in *proc. of the British Machine Vision Conference (BMVC)*, Birmingham, Sept. 1995.

[56] T.J.Cootes, C.J.Taylor, D.H.Cooper, and J.Graham, "Training models of shape from sets of examples," in *British Machine Vision Conference*, 1992, pp. 276–285.

[57] D. Koller, K. Daniilidis, and H. Nagel, "Model-based object tracking in monocular image sequences of road traffic scenes," *International Journal of Computer Vision*, vol. 10, no. 3, pp. 257–281, 1993.

[58] D. Lowe, "Three-dimensional object recognition from single two-dimensional images," *Artificial Intelligence*, pp. 355–395, 1988.

[59] K.Atika, "Image sequences analysis of real world human motion," *Pattern Recognition*, vol. 1, pp. 73–83, 1984.

[60] Z.Chen and H.Lee, "Knowledge-guided visual perception of čd gait from a single image sequence," in *IEEE Transaction on systems, man and cybernetic*, 1992, vol. 2, pp. 336–342.

[61] K. Rohr, "Towards model-based recognition of human movements in images sequences," *Computer Vision and Graphism Image Processing (CVGIP)*, vol. 59, pp. 94–115, Jan. 1994.

[62] H. Wang and M. Brady, "Real-time corner detection algorithm for motion estimation," *Image and Vision Computing*, vol. 13, pp. 695–703, Nov. 1995.

[63] L. Du, G. Sullivan, and K. Baker, "Quantitative analysis of the viewpoint consistency constraint in model-based vision," in *Proc. of the International Conference on Computer Vision 93, Berlin, Germany*, May 1993, pp. 632–639.

[64] F. Meyer and P. Bouthemy, "Region-based tracking in an image sequence," in *Proc. of European Conference on Computer Vision (ECCV)*, May 1992, pp. 476–484.

[65] B. Bascle, P. Bouthemy, R. Deriche, and F. Meyer, "Tracking complex primitives in an image sequence," in *proc. of the ICCV'94, Jerusalem*, 1994.

[66] H. Kollnig, H. Nagel, and M. Otte, "Association of motion verbs with vehicle movements extracted from dense optical flow fields," in *Proc. of the ECCV 94, Stockholm, Sweden*, May 1994.

[67] Y. Bar-Shalom and T. Fortmann, *Tracking and data association*, Academic press, London, 1988.

[68] A. Azarbayejani, C. Waren, and A. Pentland, "Real-time 3D tracking of the human body," in *Proc. of IMAGE'COM 96*, Bordeaux, May 1996.

[69] P. Huttenlocker and W. Rucklidge, "Tracking non-rigid objects in complex scenes," in *proc. of Int'l Conf. on Computer Vision (ICCV)*, Berlin, Sept. 1992.

[70] Z. Zhang, "Token tracking in a cluttered scene," Research report 2072, I.N.R.I.A., Sophia Antipolis, Oct. 1993.

[71] I. Cox and S. Hingorani, "An efficient implementation of reid's Multiple Hypothesis Tracking algorithm and its evaluation for the purpose of visual tracking," in *IEEE Transactions on pattern analysis and machine intelligence*, Sept. 1996, vol. 18.

[72] B. Rao, H. Durrant-Whyte, and J. Sheen, "A fully decentralized multi-sensor system for tracking and surveillance," *The International Journal of Robotics Research*, vol. 12, pp. 20–44, Feb. 1993.

[73] T. Strat, "Employing contextual information in computer vision," in *DARPA93*, 1993, pp. 217–229.

[74] N.Rota and M.Thonnat, "Activity recognition from video sequences using declarative models," in *Proceedings of the 14th European Conference on Artificial Intelligence ((ECAI2000)*, Berlin (Germany), aug 2000.

[75] Hsu, Huang, and Wong, "Why discretization works for naive bayesian classifiers," in *Proceedings of the 17th Intn'l Conference on Machine Learning (ICML2000)*, 2000.

[76] J-C.Terrillon, M.N.Shirazi, H.Fukamachi, and S.Akamatsu, "Comparative performances of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images," in *proceedings of 12th Conference on Vision Interface (VI '99)*, 1999, vol. 2, pp. 180–187.

[77] E. François, *Interprétation qualitative du mouvement à partir d'une séquence d'images*, Ph.D. thesis, université de Rennes I, 1991.

[78] R.Lancini, M.Ripamonti, S.Tubaro, and P.Vicari, "Combined motion estimation and image segmentation for accurate representation of motion," in *Picture Coding Symposium, Berlin, Germany*, Sept. 1997.

[79] D.Bagni, R.Lancini, S.Tubaro, and P.Vicari, "Motion estimation using region-based segmentation methods," in *International Workshop on HDTV and the Evolution of Digital Television (HDTV'96)*, Oct. 1996.

[80] P. Bouthémy, "Modèles et méthodes pour l'analyse du mouvement dans une séquence d'images," *Technique et Science Informatique*, vol. 7:6, pp. 527–546, 1988.