Dal Giornale di bordo di Visione Computazionale a.a. 2009–2010

Carlo Colombo

1° luglio 2016

01/V260210/01–02. Visione $\longleftrightarrow$ descrizione di una scena (geometria, riconoscimento) a partire da una o più immagini. Nome del corso: visione *computazionale*. Altre denominazioni (ma con delle sfumature diverse di significato): computer vision, visione artificiale. Aneddoto di Minsky: storicamente la visione nasce per risolvere il problema del riconoscimento. Poi però gli argomenti geometrici hanno preso il sopravvento. A cavallo del millennio sono apparsi tre libri di testo sulla geometria della visione da viste multiple. La geometria costituirà l'ossatura del nostro corso, proprio perché è stata più studiata, ed è ormai stata formulata una teoria matematica generale. La ricerca in CV si concentra oggi più sul riconoscimento che sulla geometria. La teoria geometrica ricorda quella delle reti logiche combinatorie: dice cosa è possibile fare con i dati in assenza di memoria, ossia quando non si mette a frutto l'esperienza visiva. Esempio tipico: applicando un metodo geometrico di ricostruzione 3D a una coppia di immagini ottenute fotografando la fotografia di una scena 3D (ad es., di una piazza circondata di palazzi e gremita di persone) da due punti di vista differenti, tutto ciò che si ottiene è la superficie piana della fotografia. Un osservatore umano invece è perfettamente in grado di inferire, dalla foto, informazioni 3D sulla struttura geometrica della scena fotografata. L'informazione geometrica è ricavata non grazie alla visione binoculare (che, anzi, non aiuta affatto in questi casi, essendo preferibile l'uso di un solo occhio alla maniera dei pittori—ricordare anche lo strumento di visione stereoscopica in voga nell'800 in cui si enfatizzava la resa tridimensionale di una foto semplicemente facendo sì che le immagini sulle due retine fossero identiche, cioè prive di disparità), ma grazie all'esperienza visiva. La memoria visiva opera in modo simile allo stato interno di una macchina a stati (rete sequenziale): l'output non dipende dal solo dato in ingresso (nel nostro caso, la fotografia della piazza), ma da un complesso meccanismo interno di interpretazione, frutto dell'attività cerebrale e della classificazione di innumerevoli input visivi precedenti. La ricostruzione 3D di una scena da immagine singola è, per l'appunto, un argomento di ricerca molto popolare negli ultimi tempi. Peraltro, la visione geometrica è ormai matura per le applicazioni commerciali. Grandi aziende quali Microsoft, Google, Toshiba, HP, Samsung, Panasonic, etc. hanno dipartimenti di computer vision che producono sia ricerca di base che applicata. Esempio di applicazione moderna della computer vision: trasformare un film girato con una sola telecamera in un film stereo, fruibile in 3D (molti film, ad es. "Avatar", sono girati

con telecamere stereo, ossia con una coppia di telecamere che producono immagini lievemente differenti della stessa scena). Si tratta di ricostruire, a partire dalla sequenza di input, le caratteristiche ottiche della telecamera, il suo movimento, la geometria 3D della scena, e il movimento degli oggetti al suo interno. Tutti questi dati verranno poi utilizzati per sintetizzare una coppia di sequenze video, come se queste fossero state acquisite con una telecamera stereo.

La visione è un potente strumento di percezione dell'ambiente circostante, ed è a questo scopo impiegata dagli animali come dai robot. Essa permetere di chiudere il loop azione-percezione, fondamentale per condurre a buon fine un task sensomotorio. Esempi: percorrere un corridoio rimanendo equidistanti dalle due pareti, mettere il cappuccio a una penna, afferrare un oggetto posto su un tavolo. Nella *robot vision*, le telecamere hanno la funzione di occhi, e il computer a cui sono collegate quella di cervello. Esistono anche sistemi di guida automatica (AGV), in cui i comandi di guida (acceleratore, freno, sterzo) sono impartiti ai motori in base agli input provenienti da una o più telecamere. Altro ambito nel quale la visione è usata per chiudere il loop percezione-azione: projector-camera systems (o smart projectors). In tali sistemi, l'output di un videoproiettore è riguardato come un'azione sull'ambiente (ne modifica infatti le caratteristiche fotometriche); l'uso di una telecamera che osserva il risultato della videoproiezione può servire ad es. per misurare le deformazioni (geometriche e cromatiche) delle immagini videoproiettate. Tali deformazioni, dovute all'interazione della luce proiettata con la superficie di proiezione (che può essere ad es. non planare, e/o non bianca) possono essere compensate mediante un'opportuna pre-distorsione delle immagini, prima della loro proiezione. I sistemi di visione artificiali hanno peraltro delle finalità differenti, in genere, da quelli naturali (si pensi anche alle differenze costruttive e funzionali tra un aeromodello ed un uccello). Ad esempio, un sistema di visione industriale deve avere in genere alte prestazioni in termini di accuratezza di misurazione e ripetibilità. Peraltro, non è richiesto che sia flessibile ed adattabile a mutamenti delle condizioni ambientali (ad es., di illuminazione).

Sensori visivi: occhio (visione biologica), macchina fotografica o telecamera (visione artificiale), quadro (arte). Differenze tra occhio umano e telecamera: enormi. La telecamera genera immagini di forma rettangolare e ha in genere pixel di forma quadrata (il quadrato è una delle forme geometriche più astratte che esistano, ed è molto più comune nei manufatti umani che non in natura – il primo esemplare noto di quadrato disegnato da un essere umano – 17000 a.C. – si trova all'interno della grotta di Lascaux in Francia), identici per dimensione. La retina umana ha una simmetria radiale, e pixel di dimensioni crescenti man mano che ci si allontana dalla fovea (piccola zona ad alta risoluzione, situata al centro della retina). Si ha una visione con buon dettaglio soltanto in corrispondenza della fovea; la periferia visiva genera immagini molto grossolane ed indistinte degli oggetti, ma ha lo scopo di dirigere le successive foveazioni (i movimenti oculari corrispondenti sono detti saccadi) su regioni del campo visivo di possibile interesse. La visione umana richiede che il sensore possa muoversi. L'interesse (contenuto informativo) è legato ad es. ad alto contrasto, differenze cromatiche, movimento (negli animali: possibili

predatori o prede). Le immagini sono ricostruite dal cervello, che organizza spazialmente e temporalmente gli input successivi provenienti dalla fovea. La visione umana è sempre a fuoco, a differenza di quanto accade con la fotografia (e con i quadri impressionisti). Il meccanismo dell'attenzione e la stessa geometria spaziovariante del sensore fanno sì da ridurre notevolmente la quantità di calcoli svolti dal sistema visivo. È dunque consigliabile che anche i sistemi artificiali incorporino dei meccanismi di riduzione del flusso di informazione (soprattutto quando è richiesto un funzionamento in tempo reale, come accade spesso in robotica).

Visione computazionale $\longleftrightarrow$ aspetti matematici indipendenti dall'implementazione hardware/software. Tali aspetti matematici risaltano con molta evidenza nell'analisi delle illusioni visive. Queste sono legate spesso ad una interpretazione errata dei dati di input, che non sono in quantità sufficiente per forzare una soluzione percettiva univoca. L'interpretazione errata da parte del cervello è dovuta all'applicazione di "regole", desunte dall'esperienza visiva in contesti reali, che mal si applicano ai contesti sintetici che tipicamente caratterizzano le illusioni. (Per questo motivo, anche gli algoritmi di computer vision sono spesso messi alla prova più da una scena sintetica, che da una reale!) Esempi di regole utilizzate dal sistema visivo: i movimenti lenti sono più probabili di quelli veloci ("ruota della diligenza"), la luce proviene con maggior probabilità dall'alto che dal basso ("concavo o convesso?" – trucco usato nella elaborazione di immagini per rendere tridimensionali oggetti 2D (embossing)). È interessante come il cervello operi come una "macchina per interpretazioni", alla costante ricerca di un significato (pensare alla forma apparente delle nuvole), e pervenga sempre e comunque ad una soluzione, quella considerata la più plausibile, anche quando dovrebbe sospendere il giudizio per mancanza di dati e/o equazioni. Altri esempi tipici: l'"insegna del barbiere" e i "cerchi rotanti". Nel primo caso, una rotazione intorno ad un asse viene interpretata erroneamente come una traslazione lungo lo stesso asse; nel secondo, un movimento di traslazione perpendicolare ad un piano dà luogo ad un movimento apparente di rotazione nel piano stesso.

02/L010310/03–04. Libri di testo: Hartley & Zisserman (libro azzurro), Ma et al. (libro giallo), Trucco & Verri (libro verde), Fusiello. Altro materiale utile: appunti 2007–2008 (E. Agostini), "trattatelli" su argomenti specifici nella homepage del corso.

Integrazioni alla lezione precedente:

- Trasformazione log-polare, piano corticale e risparmio di memoria nei sensori antropomorfi.
- Ambiguità della visione: ricostruire il 3D da singola immagine è un problema di ottica inversa (esempio della dama con l'ombrellino).
- La proiezione è centrale: è lì che il pittore pone l'occhio (disegno di Dürer).

Visione nel continuo e nel discreto. Continuo: stime dense, metodi computazionalmente onerosi (richiedono spesso un approccio multirisoluzione, o piramidale),

baseline stretta (o "narrow": matching facile, ma stime rumorose). Discreto: stime sparse, pre-selezione dei punti di interesse (es. corners, descritti poi attraverso SIFT) secondo un paradigma pseudo-attentivo, matching difficile ("wide" baseline). Uso di GPU per esecuzione di calcoli paralleli su tutti i pixel dell'immagine (ad es., filtraggi, disparità stereo, optical flow). Problema dell'apertura: non si può conoscere la velocità perpendicolare alla direzione del gradiente nell'immagine. Brightness constancy equation. Optical flow constraint. Il gradiente spazio-temporale si calcola dalle immagini, mentre u e v sono incognite. Proposta di un elaborato/tesi sulla matematica delle illusioni visive. Autori consigliati: Todor Georgiev (Adobe), Steve Zucker (Harvard University).

03/M020310/05–06. Motion field vs optical flow. Si tratta di due campi vettoriali 2D. L'optical flow (OF) è un'approssimazione del motion field (MF) calcolabile a partire dai dati immagine. Il MF è invece un'entità matematica, definita come la proiezione sul piano immagine del campo di moto 3D relativo tra sensore e scena. Sulle differenze tra OF e MF, vedere anhe l'articolo di Verri & Poggio, PAMI 1989. Lettura di parte del capitolo 4 del libro giallo: matching tra immagini e calcolo dell'optical flow. Il moello più generale di trasformazione globale di un'immagine nell'altra ha la stessa complessità della scena 3D che le ha generate ("preimmagine"). Ai fini del calcolo, ci si limita a trasformazioni del piano nel piano locali, cioè valide per intorni immagine, che possono essere modellate in forma parametrica. Esempi: traslazione (moto uniforme: 2 gdl), affinità (6 gdl). Brightness constancy equation: modella le superfici come lambertiane. Per il problema dell'apertura, l'OF in un punto non può essere calcolato usando solo informazioni (stime del gradiente spazio-temporali) relative a quel punto. Si accumulano allora informazioni relative a tutto l'intorno del punto di valutazione, si impone un modello parametrico, si definisce una funzione errore, e la si minimizza trovando i parametri incogniti. Ad es., nel caso di moto uniforme i due parametri incogniti (le componenti u e v dell'OF) si possono trovare con i minimi quadrati risolvendo un sistema lineare. La matrice dei coefficienti `G` di questo sistema lineare può però non avere rango 2, rendendo così impossibile la determinazione del flusso nel punto scelto. Nel caso di rango 1, siamo in presenza del problema dell'apertura (esiste una sola direzione di gradiente in tutto l'intorno, e di conseguenza le rette di vincolo sono tutte parallele tra loro, e non si inersecando nello spazio delle velocità); nel caso di rango 0 (matrice nulla), l'intorno ha una intensità uniforme, e dunque siamo in presenza del blank wall problem. La stessa matrice `G` è utilizzata per la selezione dei punti di interesse (corners) nel matching discreto.

04/V050310/07–08. Programma MATLAB `OpticFlow` per la stima del flusso ottico. Si tratta di un'elaborazione del programma `SampleFlow` fornito dagli autori del libro giallo. Pre-filtraggio lineare (smoothing) con kernel gaussiano. Differenza tra convoluzione (tipica del filtraggio) e correlazione (tipica del template matching): nel primo caso il kernel viene ribaltato prima di fare la somma pesata. Soglia

sul rango di `G`: non filtra a sufficienza i punti inaffidabili, per cui il flusso viene denso, ma "sporco". La soglia sul minimo autovalore di `G` funziona meglio (è, meno estremizzato e dunque meno selettivo, lo stesso criterio usato per decidere se un punto è di corner) ma produce un flusso meno denso. Immagine di "reliability": fornisce in ogni punto il valore del minimo autovettore della `G` (questa matrice ha autovalori non negativi, e gode delle stesse proprietà della matrice dei momenti di massa del II ordine, che fornisce semiassi ed orientazione dell'ellisse di inerzia per una figura piana qualsiasi). *Regolarizzazione* del flusso: è volta al riempimento di "buchi" di stima, e al miglioramento della stima di vettori di flusso che deviano sensibilmente dall'andamento locale del campo. Filtro mediano vettoriale: è non lineare, ed efficace in presenza di grossi errori di stima sparsi sull'immagine (abbiamo visto anche il filtro mediano scalare e l'effetto della sua applicazione nel restauro di un'immagine sporcata con rumore sale & pepe).

05/L080310/09–10. Espressione del motion field nel caso in cui la scena è fissa e il sensore si muove (moto rigido: 6 gdl). Modello semplificato di sensore: punto principale, lunghezza focale (quest'ultimo parametro è differente dalla focale degli obiettivi fotografici). Il nostro modello semplificato non comprende le lenti, ma tratta il sensore come una camera oscura, o *pinhole camera*. La presenza di lenti limita la profondità di campo e la messa a fuoco. Il nostro sensore mette a fuoco dappertutto. Field of view (FOV) e zoom. Ulteriore semplificazione del modello (che diventa ad un solo parametro: f): punto principale posto nell'origine del riferimento di camera. Oltre che dalla lunghezza focale f, il MF nel punto (x, y) dipende dalla struttura della scena (rappresentata dalla profondità $Z(x, y)$, uno scalare), e dal moto 3D relativo ($\mathbf{T}$,$\boldsymbol{\Omega}$). Il problema della stima del moto 3D e della struttura a partire dal flusso ottico si chiama *structure from motion* (3D structure from 2D motion, SFM), ma oggi si usa più spesso l'espressione *structure and motion*, intendendo che dalle immagini si può ricavare anche il moto 3D. Osservazioni:

- La struttura è legata al solo moto di traslazione e non alla rotazione. Quindi per percepire la struttura della scena non basta ruotare il sensore attorno al centro ottico: è necessario che il centro ottico venga spostato. È quello che fanno i piccioni, quando muovono avanti e indietro la testa.

- *Speed-scale ambiguity*: siccome il MF dipende dal rapporto $\mathbf{T}/Z$, oggetti lontani possono dare luogo al medesimo MF di oggetti vicini, se il modulo della traslazione è opportunamente scalato. Ne consegue che dall'osservazione del solo OF (ossia, senza informazioni addizionali sulla scena) non è possibile stabilire la distanza reale degli oggetti, né il modulo della loro velocità relativa di traslazione 3D.

- Oggetti a profondità diversa hanno una velocità apparente sull'immagine tanto più grande quanto più sono vicini. Questo fenomeno, illustrato da Hermann von Helmholtz nel 1866, prende il nome di *motion parallax*. La computer

graphics fa uso di questo effetto per generare l'impressione di tridimensionalità nell'animazione di scene, ad es. nei videogiochi ("parallax scrolling").

- Nel caso di pura traslazione, gli oggetti molto lontani ($Z \to \infty$) non hanno moto apparente sul piano immagine. Differenza tra il moto della luna e di un lampione.

06/M090310/11–12. SFM nel continuo. Lettura dell'articolo "Interpretation of a moving retinal image" di H.C. Longuet-Higgins e K. Prazdny (1980). Lo abbiamo rivisitato usando una notazione più moderna. Anzitutto, il motion field viene decomposto nelle sue componenti *polare* $\mathbf{v}_{\mathrm{T}}(x, y)$ ed *assiale* $\mathbf{v}_{\Omega}(x, y)$, legate rispettivamente al moto di traslazione ($\mathbf{T}$) e a quello di rotazione ($\mathbf{\Omega}$). Tali componenti hanno ciascuna una singolarità (punto in cui il MF vale zero) nel punto immagine dove il vettore di moto "buca" il piano immagine. Nel caso della componente polare, tutti i vettori di moto giacciono su linee rette che formano un fascio con centro nella singolarità. Tale singolarità si chiama *focus of expansion* (FOE) o *focus of contraction* (FOC), a seconda che i vettori divergano dal polo (avvicinamento alla scena; tipico esempio: atterraggio di un aereo) o convergano verso di esso (allontanamento dalla scena; tipico esempio: treno che lascia la stazione). Le direzioni del campo polare sono prefissate, ma il modulo dei vettori può cambiare a seconda della forma della scena. Questa informazione è usata ad es. dagli uccelli durante il volo. La componente assiale ha linee di campo di forma ellittica, che divengono circonferenze se la rotazione avviene attorno all'asse ottico.

L'articolo si compone di due parti. Nella prima parte, mostra come risolvere il problema della structure from motion usando la parallasse di moto calcolata in almeno due punti immagine. La parallasse istantanea (coincidenza puntuale di due vettori di moto nello stesso punto dell'immagine) è bene approssimata dalla differenza locale di velocità tra due punti appartenenti regioni contigue a moto differente (l'ipotesi è che il moto sia dovuto alla sola telecamera). Ossia, la parallasse si stima nei punti di occlusione/disocclusione visiva. In tali punti, il campo assiale è costante a cavallo della discontinuità, mentre quello polare cambia, in ragione delle differenti profondità. Come le discontinuità di intensità (alto gradiente immagine), anche le discontinuità di moto sono punti ad alta informazione, anche se la stima del moto attorno alle discontinuità è la più critica. Come gli intensity edges, i motion edges si possono calcolare con filtri passa alto, anche se stavolta vettoriali e non scalari (il moto ha due componenti). L'algoritmo si compone dei seguenti passi: (0) sono dati la focale f e il moto sull'immagine $\mathbf{v}(x, y)$; (1) calcolo del polo da due o più punti di motion discontinuity; (2) calcolo della direzione $\mathbf{T}/\|\mathbf{T}\|$; (3) calcolo della rotazione $\mathbf{\Omega}$; (4) calcolo del campo assiale $\mathbf{v}_{\Omega}(x, y)$; (5) calcolo del campo polare come $\mathbf{v}_{\mathrm{T}}(x, y) = \mathbf{v}(x, y) - \mathbf{v}_{\Omega}(x, y)$; (6) calcolo della *scaled depth* $Z(x, y)/T_z$ usando il polo e $\mathbf{v}_{\mathrm{T}}(x, y)$.

La seconda parte dell'articolo mostra come risolvere la structure from motion in zone dell'immagine dove la scena è smooth, facendo uso del flusso ottico e delle sue

derivate fino al II ordine. Ciò significa dover calcolare le derivate terze sull'immagine, il che comporta un'alta sensibilità al rumore. Il problema è risolto facendo uso di sole informazioni puntuali (anche se per il calcolo delle derivate del flusso bisogna considerare intorni di grandi dimensioni). Nel libro giallo (p. 143ss) è mostrato come calcolare il moto 3D (traslazione e rotazione) usando la sola informazione di flusso, ma raccogliendo osservazioni da almeno otto punti nell'immagine. È questa la versione continua del celebre *eight-point algorithm*, che vedremo nel caso del calcolo della geometria stereo da stime sparse.

07/V120310/13–14. Simulazione MATLAB di una telecamera in moto rispetto a un piano. Decomposizioni utili del campo di moto: (1) componenti polare e assiale; (2) invarianti differenziali del I ordine (divergenza, rotazionale, deformazione). Con le sole derivate del I ordine del flusso non si può risolvere il problema della SFM nel punto, ma si può calcolare un bound sulla profondità scalata Z/T_z, che prende il nome di *collision immediacy*. Il suo reciproco è il cosiddetto *time to collision* (tempo all'impatto) $\tau(x, y)$, che rappresenta il tempo necessario a ciascun punto della scena avente immagine (x, y) a raggiungere il piano di telecamera $Z = 0$. (. . .)

08/L150310/15–16. Per un'analisi dettagliata della relazione tra la struttura locale del motion field ed il tempo all'impatto si veda la mia nota "Time to collision from a natural perspective" (1997) disponibile all'indirizzo

`www.dsi.unifi.it/users/CC/Public/rt-199703-11.pdf` .

Bas-relief ambiguity: simile all'ambiguità scale-speed, solo che in questo caso un piano con uno slant grande e un moto trasversale piccolo rispetto al sensore genera lo stesso campo del I ordine di un piano con slant piccolo e velocità trasversale grande. (continua. . .)