Cos'è la Computer Vision?

Definizioni.

Rispondere a questa domanda è meno semplice di quanto si possa pensare. Dare una definizione a una materia in continua evoluzione, che presenta aspetti variegati e permette approcci diversificati, è a dir poco arduo. Perciò è meglio riportarne più d'una, nella speranza di trasmettere non solo delle spiegazioni formali, ma un'intuizione. In effetti, comunque, una definizione traccia un limite di separazione tra ciò che è compreso e ciò che viene escluso ma ciò, specie durate la fase di ricerca, può risultare più uno svantaggio che un vantaggio.

È utile fare una premessa di carattere terminologico: i termini "computer vision" e "machine vision" vengono usati spesso con lo stesso senso, anche se il secondo si riferisce più ad applicazioni industriali; il termine viene tradotto in italiano con "visione artificiale". La preferenza da me accordata a computer vision è in ragione della sua diffusione in ambito della comunità scientifica internazionale e, in secondo luogo, poiché si riferisce alla generalità della materia[1].

David Forsyth e Jean Ponce, affermano: "Noi vediamo la computer vision – o semplicemente 'vision', scusandoci con chi studia la visione umana o animale – come un'impresa che usa metodi statistici per districarsi nell'immensità di dati utilizzando dei modelli costruiti con l'aiuto della geometria, della fisica e delle teorie dell'apprendimento. Perciò, dal nostro punto di vista, la visione si basa sulla comprensione dei principi della camera oscura[2] e dei processi fisici della formazione dell'immagine, per ottenere semplici inferenze dai valori di singoli pixel, combinare l'informazione disponibile da molte immagini in un'unità coerente, imporre ordine su gruppi di pixel per separarli gli uni dagli altri o inferire informazioni sulla forma, e riconoscere oggetti usando informazioni geometriche o tecniche probabilistiche."[3].

Linda G. Shapiro e George C. Stockman la definiscono come "il campo dove informazioni significative devono essere ricavate/estratte automaticamente dalle immagini[4]" con l'obbiettivo di "prendere decisioni utili riguardanti oggetti fisici reali e scene basate sulle immagini percepite"[5].

Emanuele Trucco e Alessandro Verri preferiscono porsi delle domande più precise: "Che tipo di problemi stiamo affrontando? E come programmiamo di risolverli?". Il problema dovrebbe essere "ricavare/elaborare le proprietà di un mondo 3D da una o più immagini digitali", basandosi su

1 Cfr Linda G. Shapiro, George C. Stockman, *Computer Vision,* New Jersey, Prentice Hall, 2001 p. 1 si veda anche http://it.wikipedia.org/wiki/Computer_vision 20 gennaio 2007
2 O, comunque, di tutti gli strumenti in grado di acquisire delle immagini: macchine fotografiche, cineprese, telecamere, scanner, ecc (n.d.t.)
3 Cfr David A. Forsyth, Jean Ponce, *Computer Vision: A Modern Approach,* Op. Cit., p. xvii
4 Cfr Linda G. Shapiro, George C. Stockman, *Computer Vision,* Op. Cit., p. xvii
5 Ibidem p.1

proprietà geometriche (es.: forma e posizione di oggetti solidi) e dinamiche (es.: la velocità degli oggetti). Per loro, essenzialmente, il computer dovrebbe interpretare le immagini[6].

Ramesh Jain, Rangachar Kasturi e Brian G. Schunck riprendono il concetto *"A machine vision system recovers useful information about a scene from its two-dimensional projections"*[7] ma poi aggiungono che "Decidere richiede sempre la conoscenza dell'applicazione o dello scopo. Come vedremo, a ogni stadio in *machine vision* le decisioni devono essere prese dal sistema. L'enfasi sui sistemi di *machine vision* si è massimizzato sull'automazione delle operazioni di ogni livello, e questi sistemi dovrebbero utilizzare la conoscenza per riuscirvi. Questa conoscenza include modelli delle caratteristiche, della formazione dell'immagine, modelli di oggetti, e le relazioni fra oggetti."[8].

Avvincente è quanto riportato da Microsoft sulla sezione del sito dedicato alla ricerca sulle tecnologie visive: "Lo scopo della ricerca sulla computer vision è di dotare il computer dell'abilità di comprendere le immagini ferme e in movimento. Anche se noi, come esseri umani, possiamo dare un senso alle fotografie e ai video, per un computer esse sono solo una matrice di numeri rappresentante la luminanza e il colore di un pixel. Come possiamo ottenere da questa matrice di numeri la comprensione che c'è una ragazza che sta giocando a pallone davanti all'edificio? Quanto alto sta gettando la palla? Qual'è la struttura di fondo della casa? Questi sono gli argomenti che ci interessano!"[9].

Roberto Marangoni e Marco Geddo, da cui ho preso spunto, affermano: "Possiamo ora pensare alla nostra immagine non più come a una collezione di pixel tra loro scorrelati, ma come a un insieme di *oggetti* ciascuno con proprietà *misurabili* quali per esempio la lunghezza, l'area, il livello medio di luminosità, ecc.

Questo schema di rappresentazione costituisce il primo passo verso la "Computer Vision" (visione artificiale)[...].

L'immagine deve quindi esser pensata come un messaggio scritto in un linguaggio nel quale gli oggetti identificati sono le parole e le relazioni sono i connettivi."[10].

Ognuna di queste definizioni ha degli aspetti in comune con le altre, e ognuna focalizza un aspetto particolare della materia. Esse sono comunque, in qualche modo incomplete, in quanto si limitano a descrivere un procedimento (ricavare/estrarre informazioni dalle immagini), o uno scopo (prendere decisioni utili riguardanti oggetti fisici reali e scene basate sulle immagini percepite). Possono solo descrivere in quanto si basano su un linguaggio che possiamo definire, sulle orme di Francesco

---

6 Emanuele Trucco, Alessandro Verri, *Introductory Techniques for 3-D Computer Vision,* Op. Cit, p. 1
7 Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 1
8 Ibidem pp. 5-6
9 http://research.microsoft.com/vision/#research_groups 30 dicembre 2006
10 Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., pp. 74-75

Antinucci, logico-simbolico. A questo linguaggio se ne può contrapporre uno di tipo percettivo-motorio, ossia non basato sull'utilizzo di simboli, astrazioni e di una logica sottostante, ma sul coinvolgimento dei sensi percettivi e sul movimento.

Utilizzare il linguaggio percettivo-motorio per definire la computer vision appare difficile, eppure, forse, il cinema di fantascienza l'ha già fatto. Ad esempio, nel film "Terminator" di James Cameron[11], divenuto un "cult" per gli appassionati del genere. Ciò che qui interessa non sono l'ambientazione o gli artifici usati dal regista per creare suspense e paura o reazioni emotive, ma quelli utilizzati per rendere credibile l'interpretazione di Schwarzenegger-Terminator. Dopo ben 35 minuti e 42 secondi, durante i quali si vede Schwarzenegger coinvolto in fenomeni quantomeno anomali (l'arrivo in questo tempo con lampi d'energia, essere colpito a morte e rialzarsi, ecc) e compiere azioni violente (l'uccisione a sangue freddo di più persone), Cameron suggerisce la vera natura del personaggio: un robot. L'aspetto interessante e però costituito dal "come" lo fa, e cioè con una soggettiva che porta lo spettatore ad identificarsi con il Terminator, a vedere come il Terminator.

Il regista utilizza la soggettiva per cinque volte nell'arco del film. Esaminando nel dettaglio le relative scene, si possono cogliere i messaggi impliciti utili all'analisi:

- la prima soggettiva s'incontra nella scena in cui il Terminator insegue i due personaggi principali, appena incontrati nella discoteca. Con essa Cameron fa capire che il Terminator è un robot, ponendo così le basi per l'attendibilità delle dichiarazioni di Michael Biehn, l'interprete di Kyle Reese (l'altro essere venuto dal futuro);
- nella scena girata alla stazione di polizia, la soggettiva riafferma la natura inumana della macchina e, a seguire, la sua presunta superiorità nei confronti dell'uomo. Da notare che il robot elimina la luce prodotta dall'impianto elettrico in quanto a lui non necessaria;
- nella stanza affittata all'hotel, la soggettiva mostra uno schermo dove appare un menù di scelta tra risposte alternative usate per rispondere al proprietario;
- nel residence dove gli altri due personaggi hanno trascorso la notte l'uso della soggettiva non appare del tutto motivato, se non per indicare l'indifferenza della macchina nei confronti degli esseri viventi (si pensi al cane e all'altro inquilino);
- infine, lo spettatore si immedesima nel Terminator quando questi sale sul camion per innestare la marcia. In tale occasione, il regista voleva probabilmente sottolineare la

11 *"The Terminator"*, di James Cameron, USA, 1984.

differenza di questa macchina rispetto alle altre di cui si è soliti servirsi. Questo robot è in grado di apprendere il funzionamento di altre macchine, quindi di utilizzarle.

La soggettiva "inumana" si percepisce dal fatto che la visione è resa in rosso e nero, quasi fosse a infrarossi. A questa sono aggiunte, di volta in volta, delle video-grafie riproducenti simboli di puntamento, o coordinate incomprensibili (almeno per gli umani; in effetti si tratta di un linguaggio numerico, tipico delle macchine), o la riproduzione del meccanismo del cambio... insomma degli elementi che "svelano" come la macchina "pensa".

Purtroppo in questo elaborato non si può trascendere dal linguaggio logico-simbolico, ma si spera che la memoria sia di aiuto per richiamare le scene sopra menzionate.

Come si può notare, l'opera di Cameron richiama più o meno esplicitamente alcune delle definizioni esposte dando, nel contempo, sia la misura di come la visione possa essere ritenuta una funzione semplice, ma si riveli un fenomeno complesso. Il film mostra quelle che sono le fasi del processo visivo, ossia l'acquisizione di un'immagine e la sua interpretazione/elaborazione al fine di agire.

Legami tra tecnologia e arte.

Certamente vi sono altre opere, più o meno conosciute, che potrebbero essere portate ad esempio per definire/trasmettere l'essenza della computer vision. Questa considerazione permette di riprendere l'argomento della supposta reciproca influenza tra arte e tecnologia. In via esplicita corre l'obbligo di citare, ancora una volta, Asimov: "Non si deve, tuttavia, sottovalutare interamente l'influenza dei racconti di fantascienza. All'inizio degli anni '50, uno studente della Columbia University, Joseph F. Engelberger, lesse *Io, Robot* e come risultato, fu contagiato da una passione, che durò tutta la vita, per il lavoro sui robot.

Nel 1956 Engelberger incontrò George C. Devol Jr., che, due anni prima, aveva ottenuto il primo brevetto per un robot industriale. Egli chiamò il suo sistema di controllo e di memoria tramite calcolatore *universal automation* (automazione universale) o, abbreviato, *unimation*[12].

Engelberger e Devol fondarono insieme la Unimation Inc. e, in seguito, Devol ottenne da trenta a quaranta brevetti. [...]

Fu solo con la comparsa dei <<microchip>> che i robot progettati dalla Unimation acquistarono un interesse commerciale. Ben presto la Unimation divenne la più importante e la più redditizia fabbrica di robot nel mondo.

Iniziò così l'era del *robot industriale*. [...]"[13].

12 Per una cronologia degli avvenimenti più importanti si veda http://trueforce.com/Articles/Robot_History.htm 2 gennaio 2006.

13 Isaac Asimov, *Il libro di Biologia,* Op. cit. pp. 386-387

Discipline correlate.

È opportuno, dopo aver considerato il rapporto tra arte e tecnologia e, in precedenza, le relazioni tra biologia, psicologia e computer vision, esaminare gli stretti collegamenti tra quest'ultima e altre discipline informatiche che, comunque, si differenziano in alcuni aspetti. Ecco le principali:

1. Elaborazione dell'immagine.

   Lo scopo principale di questa materia è la trasformazione di immagini in altre immagini; il recupero delle informazioni è lasciato all'uomo. Quest'area s'interessa del miglioramento dell'immagini, degli algoritmi di compressione, della correzione delle immagini sfuocate. Da sottolineare, invece, che la computer vision si occupa del recupero delle informazioni in modo automatico, o con un minimo intervento umano: quindi essenzialmente della "comprensione" dell'immagine. Gli algoritmi utilizzati per l'elaborazione delle immagini vengono utilizzati nella computer vision per evidenziare particolari informazioni ed eliminare il rumore[14].

2. Computer grafica.

   Si occupa della creazione di immagini partendo da primitive geometriche quali linee, cerchi o superfici a forma libera. Queste tecniche giocano un ruolo significativo nella visualizzazione e nella realtà virtuale. La computer vision ha come scopo il problema inverso: misurare/stimare le primitive geometriche e altre caratteristiche dall'immagine. La computer grafica è quindi la sintesi delle immagini, mentre la computer vision ne rappresenta l'analisi. In passato queste due aree non erano molto collegate, ma ultimamente lo sono sempre più. La computer vision utilizza la rappresentazione di curve e superfici, e diverse altre tecniche della computer grafica. Viceversa, quest'ultima utilizza diverse tecniche della computer vision per fornire al computer i modelli necessari alla creazione di immagini realistiche. I due campi sono resi contigui dalla visualizzazione e dalla realtà virtuale[15].

3. Pattern recognition (riconoscimento di modelli).

   Questa disciplina si occupa della classificazione di dati numerici e simbolici. Molte tecniche statistiche e sintattiche sono state sviluppate per la classificazione dei modelli. Esse sono importanti nella computer vision per il riconoscimento degli oggetti. Infatti, molte applicazioni industriali si basano sulla *pattern recognition*. Ma la computer vision normalmente richiede ulteriori tecniche[16]. Infatti molti metodi sviluppati in passato erano adatti per oggetti 2D o 3D posizionati secondo alcuni vincoli, ma del tutto insoddisfacenti

14 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 4
15 Ibidem
16 Ibidem p. 4-5

per un ambiente 3D generico[17].

4. Intelligenza artificiale.

   Si occupa della progettazione di sistemi intelligenti e dello studio degli aspetti computazionali dell'intelligenza. L'intelligenza artificiale è utilizzata, nella fase successiva all'estrazione delle caratteristiche dall'immagine, per l'analisi delle scene da seguirsi attraverso l'elaborazione di rappresentazioni simboliche dei contenuti delle stesse. Si possono considerare tre livelli del processo di visione posto in essere dall'intelligenza artificiale: percezione, cognizione e azione. La percezione traduce i segnali esterni in simboli, la cognizione manipola i simboli, l'azione traduce i simboli in segnali che portano dei cambiamenti all'esterno (nel mondo). Molte tecniche dell'intelligenza artificiale svolgono un ruolo importante in vari aspetti della computer vision: infatti quest'ultima viene spesso considerata un sottoinsieme della prima.

   La progettazione e l'analisi di reti neurali sono campi in continuo fermento. L'impiego di reti neurali per risolvere problemi relativi alla visione artificiale cresce costantemente. Non vi sono però tecniche consolidate applicabili alla computer vision, ragion per cui se ne farà solo limitato cenno nel proseguo della tesi[18].

5. Fotogrammetria.

   Si occupa di ricavare misurazioni affidabili e accurate da immagini remote. Questa disciplina ha un collegamento più flebile con la computer vision rispetto alle precedenti. La fotogrammetria si differenzia dalla computer vision per gli alti livelli di accuratezza ricercati. Inoltre, non tutta la computer vision si occupa di misurazione.

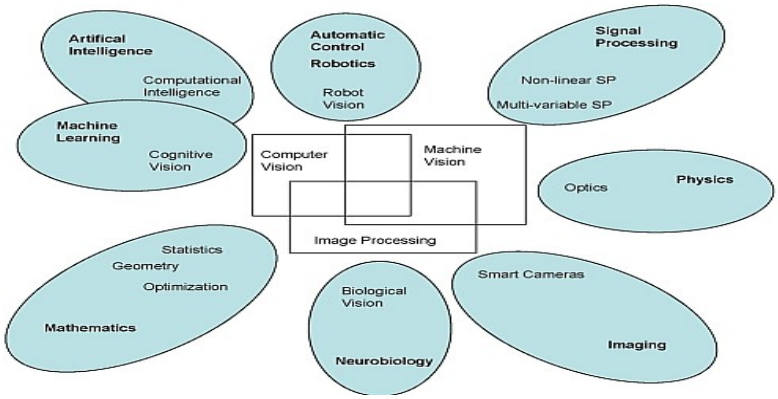

Figura 18	Relazione tra computer vision e altri campi di ricerca http://en.wikipedia.org/wiki/Computer_vision 5 febbraio 2007

17 Cfr. Emanuele Trucco, Alessandro Verri, *Introductory Techniques for 3-D Computer Vision,* Op. Cit, p. 3
18 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 5

Le immagini digitali

Questo paragrafo ha lo stesso titolo del testo di Marangoni e Geddo, ma non ne vuole essere un omaggio. Esso è invece la base d'obbligo della computer vision e di tutte le operazioni sulle immagini che si possono eseguire con il calcolatore.

Sono necessarie alcune considerazioni, già in precedenza esposte, ma che è utile richiamare: le immagini rivestono un ruolo centrale in gran parte delle attività umane. La comunicazione delle informazioni e la trasmissione di concetti attraverso le immagini sono universalmente presenti nell'umano agire. Ecco le due ragioni essenziali:

- l'uomo usa la vista per osservare e analizzare l'ambiente che lo circonda: tende quindi a rappresentare per immagini ciò che vede;
- l'immagine sintetizza una grande quantità di informazioni.

Si pensi alla cartina di una città (o una mappa qualsiasi): è un'immagine che mostra in modo semplice, completo e sintetico le informazioni relative alla rete stradale: le parole richiederebbero più tempo e spazio[19].

E' possibile introdurre una distinzione tra immagini reali e artificiali/sintetiche. Nelle prime rientrano quelle ottenute attraverso strumenti come macchine fotografiche, telecamere, scanner……. insomma strumenti di acquisizione di una scena reale. Le seconde sono quelle costruite dall'uomo, principalmente attraverso il calcolatore, che non necessariamente hanno un corrispettivo reale.

Cos'è un'immagine?

Nel capitolo dedicato alla luce e nel paragrafo "Formazione delle immagini" sono stati esposti alcuni processi fisici e geometrici coinvolti con la sua formazione. Ma, essenzialmente, cos'è un'immagine? In generale, è possibile definire un'immagine come *"un'area con una determinata distribuzione di colori"*[20] . In tale definizione è possibile riconoscere sia le immagini retiniche (o in generale quelle attribuibili a esseri viventi), sia le immagini ottenute con strumenti analogici (es: quadri, le classiche foto, i film su pellicola, ecc) nonché le immagini digitali (es: immagini digitali sintetiche, fotografie digitali, ecc).

19 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., pp. 1-2
20 Ibidem p. 4

Cosa sono le immagini digitali?

Che cosa distingue un'immagine digitale dalle altre? Tautologicamente si potrebbe dire: è digitale! Ma cosa significa?

Con digitale si vuole esprimere il concetto di un qualcosa che può essere compreso da un calcolatore, cioè una rappresentazione numerica[21]. Digitale deriva da digit che in inglese significa cifra; a sua volta digit deriva dal latino digitus che significa dito. In definitiva, è digitale ciò che è rappresentato con i numeri che si contano, appunto, con le dita[22].

Per immagine digitale s'intende l'immagine che è stata ***digitalizzata***, ossia rappresentata in numeri. Come avviene questo processo? Il primo passo è dividere il soggetto in unità distinte. L'immagine è divisa in una griglia di elementi della figura, detti anche pixel (contrazione della locuzione inglese picture element); ne si fa, tecnicamente, un *campionamento spaziale*. Il dettaglio raggiungibile e la complessità della griglia (ossia la sua *risoluzione*) variano a seconda di quanto è sofisticato il sistema di acquisizione.

Successivamente avviene la *quantizzazione cromatica,* ossia viene assegnato a ciascun pixel un valore numerico relativo che ne rappresenta il colore medio[23].

Queste matrici numeriche bidimensionali possono a loro volta dare luogo a due tipi d'immagine:

1. Immagini a ***modulazione d'intensità o luminanza***: le normali fotografie che codificano l'intensità della luce, acquisite tramite i normali sensori alla luce (misurano la quantità di luce che si imprime sul sensore).
2. Immagini ***spaziali***: codificano la forma e la distanza (stimano direttamente le strutture 3D della scena osservata attraverso varie tecniche). Sono ottenute con l'utilizzo di sonar o scanner laser.

A seconda della natura dell'immagine i numeri possono quindi rappresentare valori diversi, quali l'intensità della luce, le distanze o altre quantità fisiche. La prima

21 Lev Manovic, *Il linguaggio dei nuovi media*, trad. it. Roberto Merlini, Milano, Fres, 2004 p. 46 e seguenti, (ed. originale *The Language of New Media,* 2001)

22 http://it.wikipedia.org/wiki/Digitale_(informatica) 4 gennaio 2007

23 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., p. 5 e Ben Long, *Fotografia Digitale. Il Manuale. 3A ed.,* trad. it. Riccardo Mori, Milano, Apogeo, 2005, pp. 6-7, (ed. originale *Complete Digital Photography,* 2002)

considerazione che si può trarre è che la relazione tra l'immagine e il mondo rappresentato dipende dal processo di acquisizione e quindi dal sensore utilizzato. La seconda considerazione riguarda il fatto che ogni informazione contenuta nell'immagine deve essere ricavata da una matrice numerica[24].

Quest'ultima affermazione mi consente d'introdurre un'ulteriore distinzione tra le immagini digitali, che potremmo ricondurre a differenze logiche o, meglio, di memorizzazione. Possiamo infatti distinguere tra immagini ***raster/bitmap*** e immagini ***vettoriali***. Le prime, definite da una griglia (mappa di bit,) sono quelle più utilizzate. Le seconde sono definite quasi per contrapposizione alle prime, ossia dall'assenza della mappa e si caratterizzano per essere formate da curve o tracciati definiti da entità matematiche dette vettori[25], insomma da *primitive geometriche* che rappresentano l'immagine[26]. Se è semplice passare da un'immagine vettoriale ad una bitmap attraverso un processo di "rasterizzazione" (d'obbligo in fase di stampa e per la visualizzazione a schermo), il processo inverso è così arduo da essere ritenuto "praticamente impossibile"[27]. Questo tipo di problematica è simile a quella già discussa nel paragrafo "Formazione delle immagini", ossia la ricostruzione da un'immagine 2D presente nella retina all'ambiente 3D che l'ha generata: questa può essere un'applicazione della computer vision[28].

Sempre per quanto riguarda la memorizzazione/archiviazione delle immagini è da tener presente che nel corso del tempo si sono proposti e affermati vari formati (gif, jpg, png, ecc), diversificati per meccanismi di compressione e tipologia di informazioni considerate.

Alcune nozioni per l'acquisizione

Immagini a modulazione d'intensità/luminanza

È necessario richiamare alcuni concetti relativi alla luce e alla formazione dell'immagine approfondendo, brevemente, l'aspetto computazionale.

---

24 Cfr. Emanuele Trucco, Alessandro Verri, *Introductory Techniques for 3-D Computer Vision,* Op. Cit, p. 16
25 Cfr. Pizzirani A., *Teoria e tecnica di elaborazione delle immagini*, pp. 14-17
26 Cfr. http://www.dmi.unict.it/~gdiblasi/didattica/ixg/lezione1.pdf p. 104 7 gennaio 2007
27 Cfr. Marco Pedroni, Giorgio Poletti, *U.D. Fondamenti di Informatica*, Op. Cit, p. 137
28 Cfr. http://www.dmi.unict.it/~gdiblasi/didattica/ixg/lezione1.pdf p. 1112 7 gennaio 2007

Considerando l'immagine come un'area con una determinata distribuzione di colori, è interessante comprendere come questi vengano disposti al fine di fornire un significato[29]. Riprendendo il concetto di "formazione delle immagini ottiche" come la trasformazione "di un mondo che ha tre dimensioni spaziali in rappresentazioni bidimensionali", si hanno sia le regole per la disposizione dei punti nell'area, sia le informazioni necessarie alla sua interpretazione. Le regole di disposizione sono principalmente quelle della geometria e dell'ottica[30]. Per l'interpretazione, è l'intuito che suggerisce si tratti di immagini reali, scaturite dalla riflessione della luce sulle superfici degli oggetti.

Per la luce è sufficiente ricordare che, normalmente, essa si diffonde nell'etere in linea retta, cambiando direzione quando i quanti rimbalzano sulle superfici o le attraversano.

I meccanismi simili alla camera oscura utilizzano questa proprietà della luce per ottenere immagini bidimensionali da proiezioni di ambienti tridimensionali. I punti dell'ambiente si posizionano sul piano dell'immagine (la parete interna alla camera opposta a quella forata) in base alla proiezione prospettica o, con determinati vincoli, alla prospettiva debole. Le "macchine fotografiche" (in senso lato), successive alla camera oscura utilizzano un obiettivo per focalizzare la luce sul piano dell'immagine. Qui, nelle macchine digitali, viene posizionato il CCD (*Charge Coupled Device,* dispositivo ad accoppiamento di carica) che si occupa di *campionare* la luce e di convertirla in segnali elettrici. Sono segnali molto deboli, che devono essere amplificati prima di essere inviati a un convertitore analogico-digitale (*frame-grabber)* che li trasforma in numeri. Questi vengono poi elaborati da un processore e archiviati[31].

Il CCD è un chip di silicio ricoperto da una serie di piccoli elettrodi chiamati *photo-site* (foto-elementi). Sistemati in una griglia, esiste un photo-site per ogni pixel di un'immagine. È il numero di photosite che determina la risoluzione di un CCD[32].

Campionamento

La risoluzione di un'immagine indica il suo grado di qualità. Generalmente il termine è

29 Cfr. Marvin Minsky, *Significato e definizione,* in *Mente umana, mente artificiale,* a cura di Riccardo Valle, Op. Cit. pp. 253-259
30 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 6-10
31 Cfr. Ben Long, *Fotografia Digitale. Il Manuale. 3A ed.,* Op. Cit., pp. 19-20
32 Ibidem p. 23

riferito alle immagini digitali, ma anche una fotografia analogica ha una sua risoluzione. La risoluzione è la misura della densità dei pixel, cioè dei puntini elementari che formano l'immagine. Tale densità è riferita all'unità di lunghezza, di solito al pollice (ppi, pixel per inch o dpi, dot per inch). Per alcuni dispositivi, la densità dei pixel differisce nelle due dimensioni (per esempio negli scanner d'immagini); in tali casi è necessario indicare sia la risoluzione orizzontale sia quella verticale. Uno schermo di computer ha valori di risoluzione di 72 dpi per il Machintosh, e 96 per i pc. Le attuali stampanti casalinghe permettono di stampare immagini con risoluzioni di alcune centinaia di dpi. La risoluzione equivalente di una normale pellicola fotografica è di 3-4.000 dpi[33].

Le dimensioni assolute dell'immagine bitmap dipendono dal numero di pixel che la compongono lungo l'altezza e la larghezza.

Dimensione e risoluzione dei pixel sono entità inversamente proporzionali: tenendo costante la dimensione (larghezza e altezza), un'immagine ad alta risoluzione avrà più pixel, ma più piccoli, rispetto ad un'immagine con risoluzione inferiore[34].

Tornando al processo di acquisizione, la superficie del CCD deve essere, per captare la luce, caricata di elettroni. Quando viene colpita dalla luce, gli elettroni si agglomerano sopra la griglia di photo-site. Maggiore è la luce, maggiore sarà il numero di elettroni agglomerati sul photo-site. Successivamente all'esposizione, la macchina deve semplicemente misurare la quantità di carica a ogni photo-site, stabilendo così quanta luce ha inciso in ogni punto determinato. Questa matrice d'incidenza è poi passata al convertitore analogico-digitale[35], che la trasformerà in numeri.

Il termine "dispositivo di accoppiamento di carica" deriva da come le cariche dei singoli photo-site vengono interpretate dalla macchina: successivamente all'esposizione, le cariche della prima fila di photo-site sono trasferite al dispositivo/registro di uscita (read out register), dove vengono amplificate per poi essere inviate al convertitore analogico-digitale. Ogni fila di cariche è elettricamente unita alla fila successiva in modo che, dopo che una fila è stata processata (letta e cancellata), le successive si muovono posizionandosi sullo spazio appena lasciato libero. Una volta processate (e cancellate)

33 http://it.wikipedia.org/wiki/Risoluzione 11 gennaio 2007
34 Cfr. Pizzirani A., *Teoria e tecnica di elaborazione delle immagini*, p. 14
35 Cfr. Ben Long, *Fotografia Digitale. Il Manuale. 3A ed.,* Op. Cit., p. 23

tutte le cariche, il CCD si ricarica ed è pronto a ripetere il processo[36].

I photo-site reagiscono solo alla quantità di luce ricevuta, rimanendo indifferenti al relativo colore. Per ottenere una percezione al colore è necessario operare un filtraggio Red Green Blue (RGB)[37]. Con il metodo introdotto da Maxwell ogni photo-site viene colorato da un filtro, rosso, verde o blu. Questa combinazione è detta *allineamento (array) di filtri a colori* e segue uno schema abbastanza standard per la maggior parte dei CCD. Con questi filtri il CCD può produrre immagini distinte per ogni colore, ma incomplete in quanto mancanti dei pixel coperti dagli altri filtri; per ottenere l'immagine completa, le immagini dei vari colori devono essere interpolate fra loro. Questo è quanto avviene su macchine a singolo CCD, tuttavia si trovano macchine che ne hanno uno per ogni colore[38].

Dal CCD si ottiene un segnale elettrico continuo, il segnale video, che viene elaborato dal frame grabber o, meglio, digitalizzato. Il risultato è una matrice bidimensionale rettangolare di righe e colonne, contenente valori interi, archiviata in memoria[39].

Come detto, l'operazione di digitalizzazione avviene in due passaggi: campionamento e quantizzazione. La prima, appena trattata, inizia nella matrice del CCD e suddivide la scena 3D nella matrice bidimensionale; da questa fase dipende la risoluzione dell'immagine e, di conseguenza, la sua qualità.

Quantizzazione

L'operazione di quantizzazione cromatica consiste nell'assegnare a ciascun pixel uno o più valori numerici che ne definiscano il colore. L'operazione non è semplice in quanto, come già visto, dipende non solo dalla luce, ma anche dalla superficie dell'oggetto e dall'osservatore. Esso è infatti un campo di ricerca aperto della psicofisica. Tuttavia, come accennato poc'anzi, Maxwell risolse, nel 1869, il problema. Egli agì per approssimazione, cercando di ridurre le infinite variazioni dello spettro a tre colori primari: il rosso, il verde e il blu[40].

---

36 Ibidem p. 24
37 *Red, Green, Blue,* i tre colori in cui viene scomposta la luce nello spazio colore additivo.
38 Cfr. Ben Long, *Fotografia Digitale. Il Manuale. 3A ed.,* Op. Cit., p. 26
39 Cfr. Emanuele Trucco, Alessandro Verri, *Introductory Techniques for 3-D Computer Vision,* Op. Cit, p. 29
40 Cfr. Ben Long, *Fotografia Digitale. Il Manuale. 3A ed.,* Op. Cit., p. 20

Questa scomposizione è ancor oggi utilizzata per i tubi catodici delle televisioni e dei monitor, mentre per la stampa si usa la rappresentazione CMY (ciano, magenta e giallo) cui si aggiunge il nero per risolvere alcuni problemi d'assorbimento della carta e per migliorare la resa cromatica.

La scala d'intensità di ciascuno dei colori primari non può essere casuale; l'intervallo, tra il valore minimo ed il valore massimo dell'intensità di ogni colore è ripartito in un certo numero di livelli. Più sono i livelli, maggiore è la precisione nel rendere il colore; normalmente si utilizzano 256 livelli per ogni colore primario. Il numero di livelli è, di solito, espresso in bit, mentre, il termine *banda* indica che si tratta di informazioni relative al colore. Un'immagine a 256 livelli di colore in RGB viene spesso definita come a 24 bit per pixel, 8 per ogni colore primario[41].

Le immagini in bianco e nero sono un tipo particolare di quantizzazione, in quanto a ogni pixel corrisponde un valore numerico che fornisce informazioni solamente sulla sua luminosità. Queste immagini si dicono a *livelli di grigio* e sono particolarmente utili al riconoscimento delle forme geometriche: per questo sono spesso prese come base per la computer vision[42].

Le *immagini binarie* sono un caso particolare delle immagini in bianco e nero, in quanto i valori d'intensità associati a ciascun pixel sono solo due. Le sfumature sono ottenute con una tecnica particolare definita *dithering,* che sfrutta un'illusione ottica[43].

Il rumore

Quanto finora scritto è basato sull'assunto che l'immagine digitale sia "perfetta", ossia non presenti difetti quali riflessi, tonalità di colore errate, pixel contigui con colori contrastanti, ecc. Generalizzando, si hanno dei pixel con valori nell'immagine che non sono quelli attesi, poiché si sono corrotti durante la fase di acquisizione, quindi non sono utili allo scopo dell'elaborazione. Come conseguenza, i valori dei pixel di due immagini della stessa scena, ottenuti con la stessa macchina e con le stesse condizioni di luce non sono mai esattamente gli stessi. Queste fluttuazioni producono degli errori nei calcoli basati sui valori dei pixel; è quindi necessario considerare la quantità di rumore

41 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., p. 8
42 Ibidem p. 10-11
43 Ibidem

presente nell'immagine in modo da non inferire conclusioni errate (un po' come accade per le illusioni nella visione umana).

Immagini spaziali o di profondità[44]

In molte applicazioni si utilizza la visione per stimare le distanze; ad esempio, per tenere i veicoli lontani dagli ostacoli, controllare la produzione di oggetti ,e più in generale rilevare la forma dalle superfici.

Un'immagine che misura l'intensità della luce si dimostra carente per rilevare le distanze, in quanto i valori dei pixel sono collegati alle superfici geometriche solo indirettamente (ossia, dipendono sia dalle proprietà geometriche e fisiche di queste, sia dalle condizioni d'illuminazione).

Tuttavia è possibile rilevare direttamente la forma degli oggetti attraverso dei sensori che misurano direttamente le distanze spaziali, ottenendo delle immagini spaziali, ossia delle matrici i cui pixel esprimono la distanza tra un quadro di riferimento e un punto visibile della scena. Viene così riprodotta direttamente la struttura 3D della scena.

Per quanto riguarda la rappresentazione, le immagini spaziali possono essere rese in tre modi distinti: come un'immagine d'intensità, nella forma di una "nuvola di punti" o in forma r_{ij}. La prima forma corrisponde alle immagini finora considerate. La seconda è una lista di coordinate 3D con un dato riferimento (es: un piano ,ecc): essa non richiede un ordine preciso e suggerisce direttamente l'idea della forma. L'ultima è una matrice che riporta i valori relativi alla profondità dei singoli punti lungo le direzioni degli assi *x y* dell'immagine, rendendo esplicite le informazioni spaziali.

Per quanto riguarda i sensori, essi possono misurare la profondità di un singolo punto, la distanza e la forma dei profili delle superfici o superfici complete. Si possono distinguere in attivi e passivi. I primi proiettano energia (es.: un tipo di luce, impulsi sonar) sulla scena e ne rilevano la posizione per ottenere la misura; o utilizzano l'effetto di cambiamenti controllati di alcuni parametri del sensore (es.: la messa a fuoco). I secondi si affidano solamente alla misurazione intensità luminosa dell'immagine per rilevare la profondità.

I sensori utilizzano vari principi fisici di funzionamento come i radar e i sonar,

44 Cfr. Emanuele Trucco, Alessandro Verri, *Introductory Techniques for 3-D Computer Vision,* Op. Cit, pp. 40-47

interferometria Moiré, la messa a fuoco/sfocatura attiva e la triangolazione[45].

Quest'ultima è particolarmente interessante in quanto utilizza le macchine a rilevazione d'intensità (fotografiche, telecamere, ecc) partendo, quindi, da una conoscenza già acquisita. Questi sensori forniscono mappe di coordinate 3D accurate e dense, sono semplici da comprendere e costruire, nonché molto diffusi. La rilevazione dell'immagine scaturisce dall'interazione di un proiettore con una telecamera. Il primo fornisce sia la luce (il mezzo) che il piano di luce (coordinate geometriche) di riferimento, mentre la seconda effettua le rilevazioni[46].

Strumenti di acquisizione

Dopo aver esaminato la complessità e la bellezza dell'occhio umano è quasi triste constatare i limiti degli strumenti di acquisizione oggi disponibili. Probabilmente per questo motivo Paul Sajda, della Columbia University, cerca di interfacciare le capacità di acquisizione delle immagini del sistema visivo umano con alcuni algoritmi della computer vision[47]. Di seguito non si tratteranno queste sperimentazioni d'avanguardia, ma ci si limiterà a riportare alcuni strumenti di uso comune specificandone, brevemente, le caratteristiche.

Scanner

Sono stati, probabilmente, fra i primi strumenti utilizzati per la digitalizzazione delle immagini: il loro processo di acquisizione parte infatti da immagini già stampate su carta o da negativi.

Vi sono due tecnologie di scansione, quella a CCD in parte già vista, e quella PMT (*Photo Multiplier Tubes*). Entrambe convertono diversi livelli di luminosità in segnali elettrici a variazione continue (analogici), i quali passano al convertitore A/D (analogico/digitale) che effettua i processi di campionatura (suddivisione del segnale in quantità discrete) e quantizzazione (ossia l'attribuzione di un'etichetta numerica alle quantità discrete ottenute)..

---

45 Ibidem p.42

46 Cfr Linda G. Shapiro, George C. Stockman, *Computer Vision,* Op. Cit., pp. 47-49

47 http://newton.bme.columbia.edu/mainTemplate.htm?liinc_news.htm 13 gennaio 2007 http://punto-informatico.it/p.aspx?id=1568530&r=PI 13 gennaio 2007 http://www.wired.com/news/technology/medtech/0,71364-0.html 13 gennaio 2007

La tecnologia PMT permette di riprodurre una gamma cromatica più vasta; tuttavia essa è più costosa, richiede un più ampio livello professionale e un elevato controllo, e accetta solo originali flessibili[48].

Fotocamera Digitale

È il dispositivo più comune e flessibile per l'acquisizione di immagini. Utilizza normalmente la tecnologia CCD. È molto simile alla normale macchina fotografica; la differenza sta nel fatto che la pellicola viene sostituita con un sottile strato di celle allo stato solido, necessarie alla conversione dell'energia luminosa in cariche elettriche. Dalla matrice del CCD dipende direttamente la risoluzione e, in parte, la dimensione dell'immagine.

I sensori possono essere anche di tipo CMOS (*Complimentary Metal Oxide Semiconductors*), ovvero semiconduttori complementari a ossido di metallo. Questi sono normalmente più economici dei sensori CCD in quanto costruiti con tecnologie comuni ai diffusi chip per calcolatori. La differenza si nota anche nei consumi: i sensori CMOS sembrano infatti essere meno esigenti, assicurando una vita più lunga alle batterie della macchina e meno problemi di surriscaldamento. Permettono anche di integrare più funzioni nello stesso chip agevolando, quindi, la costruzione di macchine più compatte[49].

Le fotocamere di ultima generazione possono essere collegate direttamente a un computer (normalmente via USB o standard IEEE 1394) sia per l'archiviazione delle immagini, sia per il controllo dello scatto o delle impostazioni della macchina.

Alcuni apparecchi offrono anche caratteristiche avanzate quali l'autofocus, pre-impostazioni per situazioni o effetti particolari, ecc.

Telecamere

Le telecamere, aventi come obbiettivo la creazione di immagini visionabili dall'uomo, registrano sequenze d'immagini con una frequenza di 30 fotogrammi al secondo, permettendo la rappresentazione del movimento degli oggetti. Si ha quindi una dimensione aggiunta alle coordinate spaziali contenute nella singola immagine. Per

48 Cfr. Pizzirani A., *Teoria e tecnica di elaborazione delle immagini*, p. 5
49 Cfr. Ben Long, *Fotografia Digitale. Il Manuale. 3A ed.,* Op. Cit., pp. 30

permettere una percezione armoniosa dell'immagine si utilizzano 60 mezzi fotogrammi al secondo; se ogni singola immagine può essere considerata come una successione di righe, questi mezzi fotogrammi sono dati dagli insiemi delle righe dispari e pari che scorrono in successione alternata.

Le telecamere che creano immagini per un utilizzo automatico possono registrare immagini a qualsiasi velocità e non necessitano di utilizzare la tecnica dei mezzi fotogrammi.

La tecnologia a CCD utilizzata nelle telecamere per la computer vision ha qualche volta sofferto l'utilizzo di standard video ottimizzati per l'uomo. Un primo problema è dato dalla complessità dovuta al fatto che righe pari e dispari dell'immagine sono intrecciate, processo non necessario per la macchina. Secondo, molte matrici di CCD hanno un rapporto 4:3 tra ampiezza e altezza, come la maggioranza dei video per l'uomo. Pixel quadrati e un rapporto a 1 tra le dimensioni è più favorevole alla computer vision in quanto facilitano l'elaborazione. Il mercato ha guidato i costruttori verso standard umani: gli sviluppatori di sistemi visivi automatici hanno perciò dovuto adattarsi o sostenere costi superiori per strumenti prodotti su misura in quantità limitate[50].

50 Cfr Linda G. Shapiro, George C. Stockman, *Computer Vision,* Op. Cit., p. 26

Elaborazione automatica dell'immagine

Dopo aver cercato di definire in cosa consiste la ricerca/tecnologia della computer vision e l'immagine digitale, nonché alcuni strumenti per la relativa acquisizione, si può parlare di automatizzazione del processo di elaborazione. A tal fine è importante comprendere la relazione tra la geometria della formazione dell'immagine e la rappresentazione delle immagini nel computer. Deve essere chiaro il collegamento tra la notazione matematica utilizzata nello sviluppo degli algoritmi della computer vision e quella algoritmica utilizzata nei programmi[51].

Come detto un pixel è un campione dell'intensità dell'immagine quantizzato a un valore intero e l'immagine è una matrice bidimensionale di pixel. Gli indici [*i,j*] di un pixel sono valori interi che specificano le righe e le colonne della matrice. Il pixel [0,0] è posizionato nell'angolo in alto a sinistra dell'immagine. I valori dell'indice *i* si susseguono dall'alto verso il basso, mentre quelli di *j* si dirigono da sinistra verso destra. Questo tipo di notazione corrisponde strettamente alla sintassi della matrice utilizzata nei programmi. La posizione dei punti sul piano dell'immagine ha come coordinate *x* e *y*. La coordinata *y* (ordinate) corrisponde alla direzione verticale, la *x* (ascisse) a quella orizzontale. L'asse delle *y* va dal basso verso l'alto, quella delle *x* *sinistra* verso destra. Quindi i valori riportati dalla matrice degli indici *i* e *j* sono in ordine rovescio rispetto ai valori riportati dalla matrice delle coordinate relative alle posizioni *x* e *y*. È necessario quindi eseguire degli algoritmi per passare da un sistema di coordinate all'altro[52].

In un sistema per la formazione dell'immagine, ogni pixel occupa un'area definita sul piano dell'immagine. Le posizioni sul piano dell'immagine possono essere quindi rappresentate da frazioni di pixel. La matrice dei pixel del software corrisponde alla griglia di posizioni sul piano dell'immagine da cui è ottenuta[53].

Questa premessa ci permette di comprendere quali elaborazioni possono essere compiute sull'immagine, o meglio sulla matrice di valori che la rappresenta. Tuttavia è necessario considerare il livello, o meglio la posizione, su cui si opera. Quindi occorre considerare ogni algoritmo in base alle trasformazioni che pone in essere, a quali sono i

51 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., pp. 12-13
52 Ibidem
53 Ibidem

dati richiesti e i risultati forniti. Certamente l'elaborazione si svolge sull'immagine, ma come risultato si hanno dei simboli rappresentanti, per esempio, l'identità e la posizione di un oggetto. La quantità di dati elaborati da un sistema visivo è enorme, quindi è richiesto hardware dalle capacità adeguate. Negli ultimi anni si sono sperimentate anche architetture particolari, come le accennate reti neurali. Analizzare quali siano le caratteristiche delle operazioni significa, implicitamente, valutare le possibili richieste computazionali[54].

A seconda di quali dati vengono trattati possiamo distinguere i seguenti livelli[55]:

1. A livello dei punti (puntuale)

   Alcune operazioni si basano solamente su un punto dell'immagine. Un esempio è l'operazione di soglia.

   Questa operazione è efficientemente implementata con una *look-up table.*

2. A livello locale

   L'intensità dei punti nell'immagine risultato dipende non solo da un singolo punto dell'immagine di partenza, ma anche da quelli che gli sono adiacenti/vicini. La ridistribuzione dei punti (*smoothing)* e la rilevazione dei contorni sono operazioni locali. Per questo motivo sono adatti a queste operazioni sistemi in grado di effettuare calcoli matriciali o quelli *Single Instruction, Multiple Data* (SIMD). In pratica, sono operazioni facilmente implementate su elaboratori paralleli ed eseguite in tempo reale.

3. A livello globale

   Il risultato dipende dall'intera immagine, e può essere una nuova immagine o un'interpretazione simbolica della prima. Un istogramma dei valori d'intensità o una trasformata di Fourier sono esempi di operazioni globali.

   La complessità dei processi globali rallenta l'elaborazione nei sistemi visivi; purtroppo molte operazioni sono nella loro natura globali.

4. A livello oggetto

   Questo livello è il più specifico della computer vision, dato che i precedenti sono la

54 Ibidem
55 Ibidem pp. 14-17

base anche per altre materie come l'elaborazione dell'immagine. Le dimensioni, l'intensità media, la forma, e altre caratteristiche dell'oggetto devono essere valutate perché il sistema le riconosca. Al fine di determinare queste proprietà vengono effettuate delle operazioni solamente sui pixel appartenenti all'oggetto. Ma il punto centrale è: cos'è un oggetto? Come si può rilevare?

Gli oggetti sono normalmente definiti dal loro particolare contesto. In effetti, molte operazioni in computer vision sono svolte per trovare la posizione di un oggetto nell'immagine. Tuttavia, definire cos'è un oggetto è come trovarsi nella condizione del "gatto che si morde la coda". Per valutare le caratteristiche dell'oggetto abbiamo bisogno di sapere quali punti appartengono a tale oggetto, ma per identificarli è necessario sapere da quali caratteristiche sono contraddistinti. Sono stati fatti molti sforzi per distinguere le figure dallo sfondo, o raggruppare i punti in oggetti.

Questi livelli richiamano le teorie psicologiche della visione e la loro evoluzione. La teoria strutturalista è collegabile al primo livello, ossia la visione come percezione di punti. La Gestalt e la teoria ecologica di Gibson ai tre successivi. Per quanto riguarda il costruttivismo, esso suppone anche una "comprensione" dell'immagine che non è strettamente collegabile a questi livelli, se non come fase successiva, quella definita da Marr come "The Category-Based stage".

Marr sosteneva che si può decomporre il livello algoritmico in quattro stadi di elaborazione/percezione dell'immagine retinica, di complessità crescente: *image-based, surface-based, object-based e category-based.*

Questi livelli possono essere utilizzati sia per l'analisi del sistema visivo umano che di uno artificiale in quanto l'immagine retinica può essere considerata alla stregua di un'immagine digitale: i singoli recettori possono essere infatti considerati, anche se in via molto approssimativa, come singoli pixel.

Ogni stadio è definito in base al tipo di rappresentazione che fornisce e dai processi necessari ad elaborarla partendo dalla precedente rappresentazione. Nel dettaglio si ha[56]:

1. Image-based.

   È lo stadio immediatamente successivo alle percezione ottica dell'immagine. Si

56 Stephen E. Palmer, *Vision Science - Photons to Phenomenology,* Op. cit. pp. 87-92

occupa di rilevare i bordi e le linee, collegandoli tra loro in modo più globale, unendo le immagini stereoscopiche, definendo regioni bidimensionali nell'immagine e rilevando altre caratteristiche di base, quali i termini delle linee e le cosiddette “bolle”. Le caratteristiche bidimensionali delle immagini descrivono la loro struttura e organizzazione prima di essere interpretate come proprietà di scene tridimensionali.

2. Surface-based.

   Il secondo stadio riguarda le modalità di recupero delle proprietà intrinseche delle superfici visibili nell'ambiente esterno che possono aver prodotto le caratteristiche riscontrate nello stadio precedente. La fondamentale differenza tra questi sta nel fatto che il secondo rappresenta le informazioni sull'ambiente in termini di disposizioni nello spazio tridimensionale delle superfici visibili, mentre il primo si riferisce alle caratteristiche dell'immagine in termini di modelli ipotetici bidimensionali.

   Il concetto di una rappresentazione esplicita basata sulle superfici, introdotto inizialmente da Gibson, divenne popolare quando fu quantitativamente formulato dai teorici della computer vision e implementato in simulazioni funzionanti sull'elaboratore. Marr, Barrow e Tennenbaum fecero delle rappresentazioni basate sulle superfici quasi nello stesso periodo (1978), e descrissero degli algoritmi che potevano ricavarle realmente da immagini in scala di grigio.

   Marr chiamò la sua rappresentazione *“2.5-D sketch”* per enfatizzare il fatto che si colloca tra la struttura 2-D basata sull'immagine e quella 3-D ottenuta dalla rappresentazione basata sugli oggetti.

3. Object-based.

   È in questo stadio che le rappresentazioni includono realmente informazioni tridimensionali. Affinché il sistema visivo sia in grado di gestirle, è necessario si facciano ulteriori assunzioni implicite, riguardanti la natura del mondo visivo, poiché le inferenze attuate si basano anche su superfici non visibili o parti di superfici. Questo stadio è chiamato *Object-based* in quanto il considerare superfici non viste comporta una rappresentazione esplicita degli oggetti presenti nell'ambiente. Questo può essere ottenuto in due modi:

a. Con la semplice estensione della rappresentazione basata sulle superfici fino a includere quelle non viste all'interno di uno spazio tridimensionale.

b. Concependo gli oggetti come entità intrinsecamente tridimensionali rappresentati, a loro volta, da disposizioni di alcuni insiemi di forme primitive tridimensionali. Questo approccio può essere definito volumetrico, in quanto gli oggetti sono rappresentati come volumi di particolare forma.

4. *Category-based.*

Se il fine della percezione è fornire all'organismo accurate informazioni sull'ambiente per aiutarlo a sopravvivere e riprodursi, l'ultimo stadio della percezione deve riguardare la riacquisizione delle proprietà funzionali degli oggetti: cosa apportano all'organismo (sollievo, benessere, nonché desideri, scopi e motivazioni).

Questo processo è denominato *category-based stage* in quanto è convinzione ampiamente condivisa che le proprietà funzionali degli oggetti siano accessibili tramite un processo di categorizzazione.

La categorizzazione (o *"pattern recognition"* riconoscimento di modelli) è un metodo per desumere le funzioni evolutive rilevanti che propone il coinvolgimento di due operazioni:

a. Il sistema visivo classifica un oggetto come appartenente a un insieme di categorie conosciute basandosi sulle sue proprietà visibili: forma, dimensioni, colore e posizione.

b. L'identificazione così ottenuta permette di accedere alle conoscenze acquisite riguardanti la tipologia dell'oggetto, tra cui le sue funzioni e le aspettative sul suo comportamento. Ad esempio, una tazza è considerata utile per contenere liquidi e, quindi, per bere.

I vantaggi di un simile approccio dovrebbero essere evidenti. Tuttavia esiste un approccio alternativo. Il sistema visivo è in grado di percepire le funzioni dell'oggetto più o meno direttamente, registrandole/acquisendole dalle sue caratteristiche visibili senza una preventiva categorizzazione. I primi a sostenere questa ipotesi furono i teorici della Gestalt, parlando di *caratteri fisiognomici*.

Gibson concordò con questa impostazione, espandendo la sua teoria della percezione diretta all'inclusione delle funzioni.

È possibile, e anche molto probabile, che entrambi i processi siano utilizzati nella percezione delle funzioni degli oggetti. Alcuni come sedie e tazze hanno delle caratteristiche funzionali così intrinsecamente connesse al loro aspetto che non necessitano del processo di categorizzazione per comprenderne l'utilizzo. Altri, come calcolatori e telefoni, hanno delle funzioni così slegate dalla loro forma che non possono fare a meno del procedimento di categorizzazione.

Questi quattro stadi del processo visivo rappresentano l'ipotesi più plausibile sulla struttura della percezione visiva. Essi sono elencati nell'ordine in cui dovrebbero essere logicamente svolti, tanto che ognuno di essi deve essere completato prima che il successivo abbia luogo. Come accennato, le connessioni nel cervello suggeriscono invece l'opposto, ossia dei meccanismi di continui feedback.

Nel prossimo capitolo si analizzeranno nel dettaglio le operazioni basilari della computer vision, associate ai vari livelli.

Come funziona? - Tecnologie implementate

Al fine di offrire le basi su cui poter elaborare analisi e considerazioni si riportano di seguito alcune delle tecniche alla base della computer vision e dell'elaborazione dell'immagine, rilevandone le analogie con la visione umana.

Informazioni dall'immagine

L'immagine è ricca d’informazioni: per questo essa dev'essere interpretata e le relative informazioni estratte. Uno dei fini della computer vision (non il solo) è eseguire questo compito automaticamente.

Il primo passaggio, eseguito dai sensori, è trasformare l'immagine ottica in una rappresentazione numerica. Questo “semplice” passaggio permette già di avere una conoscenza più approfondita dell'immagine stessa, ossia i valori di luminosità di ogni pixel. Non solo, permette anche di conoscere come i valori dei pixel si ripetono nello spazio.

Nel proseguo dell’elaborato si considereranno le sole immagini binarie (in bianco e

nero) o a livelli di grigio, le prime utilizzate dai ricercatori[57]. La procedura di trasformazione delle immagine in bianco e nero può essere comunque facilmente estesa alle immagini a colori.

L'istogramma dei livelli di grigio

È una delle modalità più semplici per rappresentare il contenuto informativo di un'immagine. Per costruirlo si supponga di avere un'immagine a 256 livelli di grigi: in ascissa vengono riportati i valori possibili di ogni pixel, da 0 (corrispondente al nero) a 255 (corrispondente al bianco); in ordinata la quantità (il numero) di pixel che corrisponde a ogni singolo valore di grigio[58].

L'istogramma ottenuto riassume la distribuzione dell'informazione cromatica presente nell'immagine, divenendo un valido strumento per determinate operazioni sulla stessa, quali l'appartenenza di un pixel all'oggetto, l'eliminazione di alcuni disturbi, l'esaltazione di particolari, ecc[59].

Per inciso, l'istogramma fornisce informazioni *quantitative* sulla distribuzione dei livelli di grigio nell'immagine tralasciando, tuttavia, ogni informazione sulla distribuzione spaziale[60].

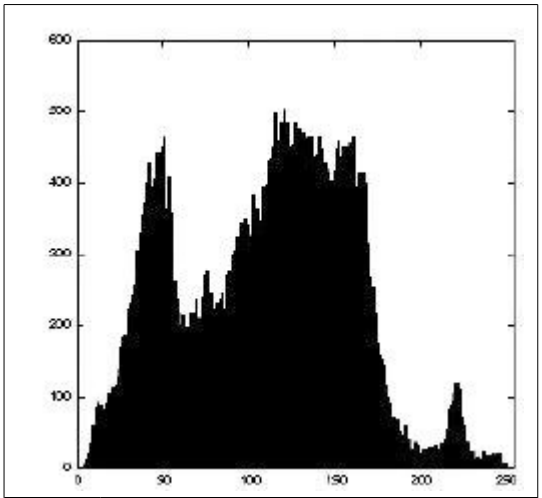

Figura 19	Esempio d'Istogramma dei livelli di grigio.

57 Cfr. Stephen E. Palmer, *Vision Science - Photons to Phenomenology,* Op. cit. p. 60 e Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 25
58 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., p. 43
59 Ibidem
60 Ibidem p. 45

Spettro di frequenza spaziale

Il concetto di frequenza è già stato introdotto trattando la luce come grandezza fisica, in particolare definendola una caratteristica dell'onda elettromagnetica. Si tratta di una grandezza unidimensionale (il numero di oscillazioni compiute in un dato periodo di tempo), mentre le immagini, e di conseguenza le matrici o *reticoli* che le rappresentano, sono bidimensionali, hanno cioè righe e colonne. Sorgono quindi alcuni quesiti:

1. *Com'è possibile utilizzare la frequenza per ottenere le immagini/reticoli?*

   Se si considera una riga di pixel, ognuno con un dato valore, e la si replica per le successive, si ottiene un'immagine: abbiamo un segnale bidimensionale come replica di un segnale unidimensionale. Questa è la forma bidimensionale più elementare che si possa avere[61].

2. *Ma cosa misura la frequenza?*

   Si supponga che un primo reticolo sia costruito con una riga per metà da pixel bianchi e per il resto neri, mentre il secondo reticolo da ¼ di bianchi, ¼ da neri, ¼ bianchi e ¼ neri, e i successivi con quantità via via dimezzate.

   Possiamo considerare il bianco e il nero come le due opposte creste dell'onda e la frequenza come la misura delle variazioni che avvengono in ciascun reticolo. Per i suddetti reticoli, avremo quindi una frequenza di 2, 4, 8, ecc.

I reticoli con frequenze così regolari sono composti da linee verticali; per averne di diversi è necessario considerare tutti i loro parametri[62]:

1. ***Ampiezza*** (o contrasto): si riferisce alla differenza di luminosità tra la parte più chiara e quella più scura, che corrisponde alla differenza d'altezza dell'onda tra il picco e la valle sotto il profilo della luminanza.
2. ***Orientamento***: si riferisce all'inclinazione delle barre bianche o nere; è espressa in gradi partendo dalla posizione verticale e procedendo in senso orario.
3. ***Fase***: si riferisce alla posizione della curva in riferimento a un punto determinato.
4. ***Frequenza spaziale***: si riferisce all'ampiezza delle righe (i reticoli a bassa frequenza hanno linee spesse, quelli ad alta frequenza linee fini).

61 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., p. 45
62 Cfr. Stephen E. Palmer, *Vision Science - Photons to Phenomenology,* Op. cit. p. 160

Modificando “a piacere” questi parametri si possono ottenere i reticoli più vari, in pratica corrispondenti a tutte le immagini acquisibili dall'occhio umano.

Per visualizzare il contenuto in frequenze di un reticolo qualsiasi è possibile utilizzare un istogramma simile a quello usato per la quantizzazione dei livelli di grigio: nelle ascisse vi sono i valori delle frequenze spaziali presenti nel reticolo, nelle ordinate, è indicata la percentuale di ciascun pixel dell'immagine in corrispondenza a ciascun valore di frequenza. Questo istogramma può essere considerato, in via euristica, lo spettro di frequenza del reticolo[63].

Poter manipolare lo spettro di frequenza permette di intervenire sulle proprietà spaziali dei pixel, ossia di aumentare o diminuire lo spessore delle linee che costituiscono una delle primitive dell'immagine. Come questo sia possibile si vedrà analizzando le operazioni globali sull'immagine; ora, invece si considereranno le operazioni sui singoli punti[64].

Operazioni puntuali

Una delle più semplici elaborazioni eseguibili su un'immagine è la sostituzione dei pixel corrispondenti a un determinato valore con pixel di altro valore. Un esempio è la sostituzione di un valore di grigio dell'immagine di partenza con un altro nell'immagine derivata. La variazione sarà tanto più evidente quanto più alto sarà il numero di pixel coinvolti o la differenza tra il valore di partenza e quello di sostituzione[65].

Un’elaborazione puntuale è detta omogenea se il risultato dipende solo dal valore del pixel cui è applicata; per la stesso motivo vengono anche dette manipolazioni della scala dei grigi o dei colori. Se il risultato dell’elaborazione dipende anche dalla posizione del pixel nell’immagine, si parla di elaborazioni puntuali non omogenee.

Ecco alcune tipiche elaborazioni puntuali:

1. Aggiunta o sottrazione di una costante a tutti i pixel (per compensare sotto o sovraesposizioni).
2. Clipping (ritagliare determinate aree per renderle visibili all'osservatore o stabilirne

63 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., p. 47
64 Ibidem p. 48
65 Ibidem p. 39

l'appartenenza a una regione)[66].

3. Espansione del contrasto.
4. Inversione della scala dei grigi(negativo).
5. Modifica (equalizzazione o specifica) dell'istogramma.
6. Presentazione in colore falso.
7. Soglia.

Quest'ultima elaborazione puntuale, di seguito spiegata, è tra quelle più spesso utilizzate.

L'operazione di soglia

L'operazione di soglia è comune in psicofisica e nell'analisi dei segnali. Consiste essenzialmente nello stabilire un valore di riferimento per procedere, in successione, al confronto con altri valori al fine di eseguire una determinata azione. Nel nostro caso, la soglia viene confrontata con i valori dei pixel componenti l'immagine e l'azione consiste nella loro sostituzione con grandezze diverse a seconda siano inferiori o superiori al valore di riferimento[67].

In questo caso, un esempio in formula può forse essere più chiaro.

$$P(m,n) = \begin{matrix} 255 \; se \; p(m,n) \geq k \\ 0 \; se \; p(m,n) < k \end{matrix}$$

Se con *k* si indica un valore di soglia nei livelli di grigio dell'immagine, i valori del pixel dell'immagine elaborata [*P(m,n)]* saranno zero se il valore di partenza [*p(m,n)*] è inferiore a *k*, 255 se superiore. Come si può dedurre, questa operazione è utile per trasformare un'immagine a colori o a scala di grigi in un'immagine binaria. È un primo metodo per poter distinguere i valori d'illuminazione dei pixel che appartengono ad un oggetto rispetto a quelli appartenenti allo sfondo, ossia per eseguire una segmentazione/suddivisione dell'immagine.

La segmentazione è una procedura per suddividere l'immagine in sotto-immagini, denominate regioni che, a loro volta, rappresentano oggetti o parti di oggetti.

66 http://it.wikipedia.org/wiki/Clipping e http://en.wikipedia.org/wiki/Clipping 20 gennaio 2007
67 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., p. 39

L'operazione di soglia e la segmentazione sono, per le immagini binarie, sinonimi[68].

L'immagine binaria ottenuta permette di elaborare le proprietà geometriche e topologiche degli oggetti per poi utilizzarli: il valore di soglia deve essere perciò scelto in relazione all'illuminazione e all'indice di riflessione degli oggetti. Questo implica una conoscenza del dominio sotto analisi, in quanto lo stesso valore di soglia applicato ad un'altra immagine potrebbe non funzionare. La scelta della soglia ottimale, se fatta dall'uomo è spesso frutto dell'esperienza. In alcuni casi all'avvio del sistema si fanno alcuni tentativi per determinarne interattivamente il valore adatto[69].

La determinazione automatica della soglia ottimale è il primo passo per l'analisi dell'immagine con i sistemi di computer vision. Molte tecniche sono state sviluppate per utilizzare la distribuzione della luminosità nell'immagine e la conoscenza degli oggetti d'interesse per selezionare automaticamente l'appropriato valore di soglia.

Si può notare il parallelismo tra quest'operazione e la funzione svolta dai bastoncelli nell'occhio umano: infatti questi si occupano di rilevare la luminosità, individuando solo movimenti e ombre. Sono considerati da Gregory come la forma più primitiva di visione[70], è forse per questo che non abbiamo difficoltà a comprendere le immagini binarie, come le linee tratteggiate o *silhouettes*.

Operazioni locali

Si parla di elaborazioni locali quando le trasformazioni dipendono non solo da un singolo punto dell'immagine di partenza, ma anche da quelli ad esso adiacenti/vicini, ossia da quello che viene definito l'"*intorno*" del pixel. Avremo un intorno di 3'3 quando il pixel in posizione centrale, identificato con *p(m,n)*, è circondato da un quadrato di lato 3 pixel; un intorno di 5'5 quando il lato è di 5 pixel, e così via. Vengono scelti intorni di lato dispari per fare in modo che il pixel considerato sia esattamente al centro dell'intorno. Vi possono essere eccezioni per alcune operazioni particolari. Il lato è scelto in base al tipo di filtraggio da compiere, con preferenza di intorni piccoli per ridurre la quantità di calcoli da effettuare[71].

---

68 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 29
69 Ibidem p. 27
70 Cfr Richard L. Gregory, *Occhio e cervello – La psicologia del vedere,* Op. cit., p. 79
71 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., pp. 40-41

p(m-1,n-1)	p(m,n-1)	p(m+1,n-1)
p(m-1,n)	p(m,n)	p(m+1,n)
p(m-1,n+1)	p(m,n+1)	p(m+1,n+1)

Figura 20	Esemplificazione di una maschera/filtro, in giallo il pixel centrale, in blu l'intorno; nello specifico si tratta di un filtro di media, che assegna, cioè, al pixel centrale la media calcolata sui valori dell'intorno.

Le operazioni locali possono essere di tipo aritmetico o logico[72].

Operazioni aritmetiche

Sono tra le più utilizzate nell'elaborazione delle immagini a basso livello (riduzione del rumore, miglioramento di qualità, estrazione di contorni, ecc...).

Il risultato di un'operazione locale è sempre un'immagine, che ha perso però delle informazioni, dato che il livello di grigio iniziale del pixel, sottoposto a elaborazione locale, non può più essere recuperato.

Queste operazioni sono eseguite a livello locale perché applicare un filtro a tutti i pixel di una immagine è una operazione che richiede lunghi tempi di elaborazione: si pensi che, per applicare una maschera 3x3 ad una immagine da 512x512 richiede 9 moltiplicazioni e 8 addizioni in ogni posizione, per un totale di 2.359.296 moltiplicazioni e 2.097.152 addizioni. Negli anni passati, per superare questi problemi è stato sviluppato hardware dedicato capace di effettuare operazioni logico-aritmetiche in parallelo, a velocità di video-frame (consentendo quindi l'elaborazione in tempo reale di filmati). Oggigiorno, anche elaboratori di fascia medio-bassa, visto il generale aumento delle performance computazionali, possono essere usati allo scopo.

Operazioni logiche

Le operazioni logiche usate nell'elaborazione delle immagini sono quelle booleane: ***AND***, ***OR***, ***NOT***.

72 http://jada1.unime.it/visilab/AppuntiCV/operazioni.htm 20 gennaio 2007

Ogni altra operazione logica può essere ricavata tramite una loro combinazione: è per questo che esse sono definite complete.

Si possono applicare esclusivamente alle immagini binarie e tra pixel omologhi (cioè aventi le stesse coordinate) di due immagini.

Sono utili per le operazioni di estrazione di caratteristiche, analisi della forma, applicazione di maschere, ecc...

Filtri

L'operazione di soglia è un esempio di filtro che può essere applicato alle immagini. I filtri sono operazioni che possono essere eseguite sia punto per punto, sia locali[73].

Il termine filtro prende origine dall'analisi dei segnali, dove le operazioni in generale vengono definite *filtraggi*, mentre *filtro* indica l'operazione specifica messa in atto.

Essi possono essere utilizzati per migliorare l'immagine, sia sopprimendo il rumore, sia mettendo in risalto alcune caratteristiche. Per riuscire in tale compito essi agiscono sulla parte alta dello spettro di frequenza dell'immagine, dove si posizionano i punti isolati o le piccole macchie.

Per eliminare il rumore si utilizzano dei filtri passa-basso o il motion blur (nel caso in cui l'operatore si sia mosso durante lo scatto) o del rumore statico per sequenze d'immagini in movimento.

Filtri passa-basso – riduzione del rumore

Essi eliminano le alte frequenze lasciando intatte le basse. Poiché i punti di alta frequenza spaziale coincidono anche con i contorni degli oggetti, se le alte frequenze vengono eliminate "indiscriminatamente" questi risultano sfocati, e l'immagine risulta poco nitida.

I filtri passa-basso rimuovono alcuni tipi di rumore comuni, come[74]:

1. *Sale e pepe*: si tratta di punti variamente dispersi nell'immagine con luminosità bianca e nera.
2. Rumore da *impulsi*: si tratta di punti bianchi sparsi per l'immagine.

73 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., p. 40
74 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 118

3. Rumore *gaussiano*: diversamente dai precedenti il rumore è dovuto a una distribuzione normale o gaussiana delle variazioni di luminosità.

I filtri passa-basso più utilizzati sono:

1. Filtro di media locale

   Si esegue assegnando la media dell'intorno di un pixel nell'immagine di origine al corrispondente nell'immagine derivata.

   Si tratta di un filtro passa basso in quanto prendendo i valori medi dei livelli di grigio, automaticamente si riducono le brusche variazioni di luminosità; tuttavia, se l'intorno è troppo ampio, vengono sfuocati i contorni.

   Il contributo al valore del pixel derivato può venire in modo uguale da tutti i pixel dell'intorno oppure, se ve n'è necessità, in modo differenziato/pesato. L'importante, se non si vogliono alterare i valori globali dell'immagine, è che la somma dei valori dell'intorno sia pari a 1.

Figura 21	Esempi di filtri locali passa-basso, a sinistra il filtro di media non pesato; al centro un filtro di media con maggiore importanza al pixel centrale; a destra il filtro attribuisce via via meno valore al pixel centrale, a quelli che sono in verticale e orizzontale per finire su quelli delle diagonali.

2. Filtro di mediana.

   È un modo per ottenere l'eliminazione del rumore ad altissime frequenze lasciando quasi inalterata l'immagine.

   Il filtro considera l'intorno del pixel disponendoli in ordine crescente o decrescente. Il valore centrale della lista (la mediana) sarà il valore del pixel nell'immagine di destinazione. Questo filtro non sostituisce il valore medio ma, più semplicemente, "sopprime" i valori troppo scostanti dalla media dell'intorno.

   Il filtro di mediana è più accurato del precedente filtro di media locale. Può essere ancora migliorato, applicandolo solo se la differenza tra il valore della mediana dell'intorno del pixel e quello di partenza è inferiore ad una determinata soglia. In

questo modo il filtro è più selettivo, elimina solo il rumore e non incide su altre caratteristiche dell'immagine[75].

3. Filtri Gaussiani – *Smoothing*

   Sono una classe di filtri per la distribuzione lineare i cui pesi sono scelti in base alla funzione di Gauss. Essi sono ottimi filtri per la rimozione del normale rumore, assimilabili ai filtri di media, ma con alcune proprietà particolari come il fatto che il peso decresce verso zero all'allontanarsi dall'origine (ciò significa che i valori dell'immagine vicini alla posizione centrale sono più importanti dei valori più remoti)[76].

Filtri passa-alto – aumento del dettaglio

I filtri possono essere utilizzati anche per evidenziare i contorni degli oggetti, perciò aumentare il contenuto delle alte frequenze dello spettro e diminuire le basse.

L'uso di filtri passa-alto per evidenziare i particolari fini dell'immagine è detto *sharpening* (rafforzamento), in quanto aumenta il contrasto medio dell'immagine.

-1	-1	-1
-1	+9	-1
-1	-1	-1

0	-1	0
-1	+5	-1
0	-1	0

1	-2	1
-2	+5	-2
1	-2	1

Figura 22 — Esempi di filtri passa alto per l'esaltazione di contorni

Rilevazione dei bordi

I primi stadi del processo visivo identificano le caratteristiche delle immagini che sono rilevanti per stimare la struttura e le proprietà degli oggetti della scena. I bordi, una di queste caratteristiche, sono variazioni locali significative nell'immagine e quindi estremamente utili per la sua analisi. Normalmente si trovano tra i margini/limiti di due regioni diverse. Riuscire ad individuarli è il primo passo per poter estrarre ulteriori informazioni dall'immagine, e associarle ad oggetti effettivamente presenti nella scena[77].

Può essere una base di partenza il considerare che i punti che definiscono un bordo sono quelli posizionati nella zona delle alte frequenze spaziali e che un filtro passa-alto è in

75 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., pp. 56-57
76 Cfr Linda G. Shapiro, George C. Stockman, *Computer Vision,* Op. Cit., p. 151
77 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 140

grado di evidenziarli. Ma per estrarre un bordo è necessario individuare solo i pixel che appartengono allo stesso, eliminando gli altri: costruire cioè un'immagine binaria con tutti i pixel appartenenti al contorno aventi valore 255 e gli altri 0. Operando solo sulle frequenze sarebbe sufficiente eliminare quelle basse e mantenere quelle alte. Spesso, però, si preferisce utilizzare dei filtri direttamente sull'immagine, o meglio sul dominio spaziale[78].

Gli algoritmi che si occupano della rilevazione dei bordi sviluppano solitamente le seguenti fasi[79]:

1. ***Filtraggio:*** dato che il calcolo dei gradienti si basa sui valori d'illuminazione, solo due punti sono soggetti a rumore o altri accidenti nei calcoli discreti; i filtri sono quindi utilizzati per migliorare l'esecuzione del rilevatore di contorni con riferimento al rumore. Tuttavia è necessario valutare la rilevanza dei bordi e la quantità di rumore. Un filtro troppo forte riduce spesso anche i margini.
2. ***Miglioramento:*** per facilitare la rilevazione dei bordi è essenziale determinare i cambiamenti d'intensità vicino al punto. È necessario evidenziare i pixel dove vi sono significativi cambi nei valore locali d'intensità, e questo normalmente avviene valutando l'ampiezza del gradiente.
3. ***Rilevamento:*** dato che molti punti dell'immagine hanno valori diversi da zero, e non tutti sono bordi, è necessario individuare un metodo per individuare quali punti sono contorni. A tal fine si usano, solitamente, i criteri di soglia.
4. ***Localizzazione:*** è un algoritmo incluso solo in alcuni rilevatori. Si tratta di valutare la posizione del bordo basandosi sulla risoluzione, richiesta da alcune applicazioni, dei sub-pixel. Può essere rilevato l'orientamento del bordo.

I filtri normalmente utilizzati per rilevare i bordi sono i seguenti[80]:

1. Filtri gradiente

   Prendono il nome dall'operatore gradiente che, nelle funzioni continue, esprime per ogni punto, la variazione della funzione. Poiché le immagini digitali sono discrete, e non continue, è necessario applicare gli operatori discreti analoghi ad esso, ovvero

78 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., p. 64
79 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., pp. 145-146
80 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., pp. 64-68

il rapporto incrementale.

Basandosi sul fatto che i bordi risultano spesso in brusche variazioni della luminosità, essi cercano di massimizzare le differenze esistenti tra pixel consecutivi, in modo da assegnare al pixel dell'immagine trasformata un valore tanto maggiore, quanto più ampia è la differenza tra i pixel considerati nell'immagine d'origine. La *direzione* del filtro dipende dai pixel considerati, come è possibile vedere in figura 23.

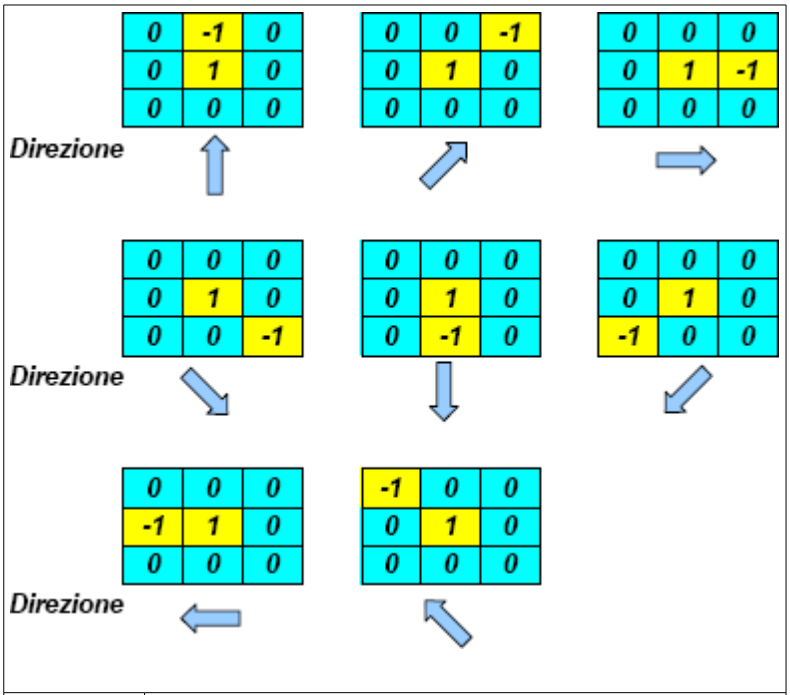

Figura 23	Possibili direzioni del filtro gradiente di lato 3: nord, nord-est, est, sud-est, sud, sud-ovest, ovest, nord-ovest.

Da notare che la somma delle maschere dei filtri gradiente o, comunque, dei filtri per l'estrazione di contorni, è pari a zero e non a uno, in quanto si vuole esaltare i contorni. Per questo l'immagine derivata è tutta buia, tranne che per i pixel dei contorni che sono chiari.

Questo tipo di filtri presenta comunque alcuni problemi: essi tendono ad amplificare qualunque variazione brusca della luminosità (quindi di rumore e falsi contorni inclusi) e, peggio, non sempre individuano tutti i contorni, in quanto ignorano spesso le transizioni oggetto/sfondo (o oggetto/oggetto) cui corrispondono limitate variazioni della funzione di luminosità.

2. Filtro laplaciano.

È stato studiato per superare alcuni dei problemi accennati dei filtri gradiente.

Il filtro laplaciano si basa sull'omonimo operatore. Dal punto di vista delle funzioni continue, quest'ultimo è una derivata seconda, ossia misura la variazione della variazione (prima derivata) della funzione di luminosità. Poiché questo filtro si basa sulla rapidità con cui si passa dall'oggetto allo sfondo, esso risulta meno sensibile al rumore e più efficace nell'identificare i contorni. Come il precedente può essere orientato in diverse direzioni, riportate in figura.

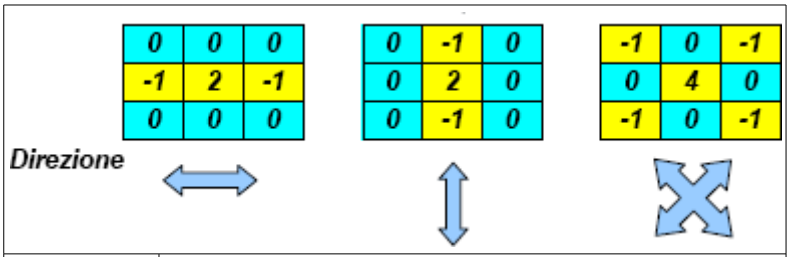

Figura 24	Esempi di filtri laplaciani, in direzione orizzontale, verticale, diagonale.

3. Filtro di Prewitt

Il filtro di Prewitt adotta un approccio di tipo gradiente, ma usa maschere in cui sono contrapposte tra loro delle zone di pixel, anche orientate.

4. Filtro di Sobel

Il filtro di Sobel, di tipo non-lineare, fornisce ottimi risultati. Data la sua complessità, ne si omette la spiegazione dettagliata.

Rilevazione delle linee

Una volta rilevati i bordi è importante riuscire a ricondurli a delle linee. Una linea può essere modellata come due bordi vicini tra loro ma con opposta polarità, separati da una distanza irrilevante. Sebbene sia un'operazione molto semplice, al lato pratico presenta difficoltà non indifferenti, tanto che una soluzione univoca non è stata ancora trovata. L'implementazione in un singolo sistema della rilevazione di bordi e linee sembra ancora un problema insormontabile[81].

Estrazione di contorni

I bordi devono essere in un secondo tempo collegati per rappresentare i confini di una

81 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., pp. 180-181

regione. Questa rappresentazione è detta contorno. I contorni possono essere aperti o chiusi. Quelli chiusi corrispondono ai confini di una regione, e i pixel appartenenti alla regione possono essere individuati da un apposito algoritmo. Un contorno aperto può essere parte del confine di una regione. Quest'ultima può contenere dei salti (ossia vi sono dei punti vuoti nella regione) in quanto il contrasto tra regioni può non essere rilevato a causa della debolezza della sua intensità. Questo in quanto la soglia del rilevatore di bordi può essere stata impostata ad un valore troppo alto oppure il contrasto lungo determinate posizioni di confine può essere troppo debole rispetto a regioni confinanti dell'immagine cosicché le singole soglie non risultino funzionali per tutta l'immagine. Contorni aperti sono presenti anche quando frammenti di linee sono collegati assieme, ad esempio in un dipinto al tratto o in un manoscritto[82].

Un contorno può essere rappresentato sia da una lista ordinata di bordi o da una curva. Una curva è un modello matematico di un contorno. Esempi di curve includono segmenti di linee o funzioni cubiche spline[83]. Ci sono alcuni criteri per una buona rappresentazione dei contorni:

1. Efficienza: il contorno dovrebbe avere una rappresentazione semplice e compatta.
2. Accuratezza: il contorno dovrebbe adattarsi in modo accurato alle caratteristiche dell'immagine.
3. Efficace: il contorno dovrebbe essere appropriato per le operazioni che saranno successivamente effettuate.

L'accuratezza della rappresentazione è determinata dalla forma della curva usata per modellarla, dalla performance dell'algoritmo usato per produrla e dall'accuratezza del rilevatore di bordi. La più semplice rappresentazione di un contorno è una lista ordinata dei suoi bordi. Questa è tanto più accurata quanto più è la stima della posizione dei bordi, ma risulta essere così la forma meno compatta e può essere inutilizzabile per la successiva analisi dell'immagine. Trovare il modello di curva che meglio si adatta ai bordi ne aumenta l'accuratezza, poiché gli errori nella posizione dei bordi sono ridotti attraverso la media, e l'efficienza ne è accresciuta attraverso una più appropriata e più compatta rappresentazione per le successive operazioni. Ad esempio, un insieme di

82 Ibidem pp. 186-187
83 http://it.wikipedia.org/wiki/Interpolazione_spline 28 gennaio 2007

bordi che sono vicini ad una linea possono essere rappresentati in modo più efficiente adattando questa ai bordi. Così facendo si semplificano i calcoli successivi quali: determinare l'orientamento o la lunghezza della linea, accrescerne l'accuratezza (l'errore quadratico medio tra la linea stimata e la linea "vera" sarà minore dell'errore tra la linea "vera" e ognuno dei bordi)[84].

Vi sono anche altri tipi di approcci all'estrazione di contorni, suggeriti da alcune scoperte della psicofisica moderna, che si basano sulle trasformate di Fourier. Il procedimento è più complesso dell'individuazione delle differenze tra i livelli di grigio. Dato che questo metodo valuta l'intera immagine e solo successivamente individua dei punti localizzati, lo si esaminerà trattando le operazioni globali.

Texture

Le textures ("tessiture") giocano un ruolo importante in molti sistemi di computer vision, come l'ispezioni di superfici, la classificazione di scene, la determinazione dell'orientamento e della forma delle superfici[85].

La texture è caratterizzata dalla distribuzione spaziale di livelli di grigio in una zona. Esse non possono essere perciò definite in un punto. La risoluzione alla quale un'immagine è osservata determina come la texture è percepita. Ad esempio, osservando un'immagine di un pavimento piastrellato da grande distanza si possono notare le strutture formate dalle disposizioni delle piastrelle, mentre sfuggono i disegni, formati dalle singole unità. Quando la stessa scena è osservata a distanza ravvicinata, così che solo alcune piastrelle rientrino nel campo visivo, si percepiscono gli intrecci formati dalla disposizione dei disegni dettagliati che compongono ogni piastrella[86].

Si può quindi definire una texture come una ripetizione di un modello locale di variazione d'intensità, che è troppo debole per essere distinto, alla risoluzione da cui lo si osserva, come oggetto separato. Un insieme di pixel connessi che soddisfino una data proprietà di grigio che si ripete in una data regione dell'immagine costituisce quindi una zona di texture. Un testo su foglio bianco può essere considerato una texture. Ogni carattere è dato da un insieme di pixel connessi e che hanno un determinato valore di grigio o colore. Disponendo i caratteri in linea e le linee in sequenza, come elementi

84 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., pp. 186-187
85 Ibidem p. 234
86 Ibidem

della pagina, si ha una texture ordinata[87].

Nell'analisi delle textures si devono affrontare principalmente tre problemi[88]:

1. Classificazione

   Il problema riguarda l'identificazione della texture in esame da un insieme dato di classi. L'algoritmo d'analisi ricava le caratteristiche distintive da ogni area per favorire la classificazione del modello. Ad esempio, da delle foto aeree si possono estrarre i modelli per i terreni agricoli, le foreste, le zone urbane, ecc. Per semplicità spesso si assume che i margini tra le varie regioni siano già stati determinati. Nella classificazione delle strutture, i metodi statistici sono usati in modo estensivo.

2. Segmentazione

   È il problema opposto al precedente. Se prima si rilevavano le caratteristiche di una regione per classificarla, qui si segmenta/suddivide automaticamente l'immagine determinando, sempre automaticamente, i margini che dividono le varie regioni di texture.

3. Rilevazione della forma dalle textures.

   Grazie ad alcune variazioni sul piano dell'immagine riguardanti le proprietà delle textures, come densità, grandezza, e orientamento, è possibile ricavare informazioni sulla forma e sull'orientamento delle superfici.

Riassumendo, una texture si può definire o come una serie di elementi (*texel* = TEXture Element) uniti tra loro in modo regolare e ripetuto o come la misurazione della disposizione quantitativa dell'intensità in una data regione.

Nel primo caso, considerato come un approccio strutturale, per costruire e analizzare la texture è sufficiente individuare la forma degli elementi e le relazioni. È il metodo utilizzato nella maggior parte delle immagini artificiali.

Il secondo caso, denominato approccio statistico, risulta più interessante, in quanto applicabile alle immagini reali. Infatti, quantità numeriche o statistiche che descrivono una texture possono essere calcolate partendo da un'immagine a scala di grigi (o a colori) autonomamente. Questo approccio, forse meno intuitivo, è efficiente in termini

87 Ibidem p. 235
88 Ibidem p. 235-236

di calcolo, e può andar bene sia per la segmentazione sia per la classificazione delle textures[89].

Operazioni globali

Si è già accennato come, in generale, sia possibile acquisire delle informazioni sull'immagine considerata attraverso l'istogramma dei livelli di grigio (o comunque, dei colori) e lo spettro di frequenze spaziali. Questi strumenti possono essere utilizzati per modificare l'immagine, come spiegato nei paragrafi seguenti.

Operazioni con l'istogramma d'intensità

L'istogramma d'intensità indica la distribuzione quantitativa dei livelli di grigio presenti nell'immagine. Queste informazioni possono essere utilizzate per determinare automaticamente il valore di soglia da applicare a una determinata immagine, e quindi alla segmentazione della stessa[90].

Come già accennato, per riuscire in questo compito l'algoritmo di segmentazione deve avere conoscenza degli oggetti presenti nella scena, del loro possibile utilizzo e dell'ambiente circostante. Questa conoscenza in particolare può riguardare:

1. L'indice di rifrazione (luminosità caratteristica) degli oggetti.
2. Le dimensioni degli oggetti.
3. La frazione dell'immagine occupata dagli oggetti.
4. Il numero di tipi differenti d'oggetti appartenenti all'immagine.

Un algoritmo che utilizzi questo tipo di conoscenza e arrivi a determinare autonomamente il corretto valore di soglia può essere definito automatico[91].

Ecco alcuni algoritmi applicati ad oggetti che, per semplicità, si suppongono scuri perché illuminati da una luce di fondo. In questo modo si possono attribuire agli oggetti i valori di grigio sotto una certa soglia, allo sfondo quelli sopra la soglia stessa[92]:

1. Metodo *P-Tile* (percentuale di copertura).

89 Cfr Linda G. Shapiro, George C. Stockman, *Computer Vision,* Op. Cit., p. 213-214
90 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 77
91 Ibidem
92 Ibidem pp. 78-85

È forse il metodo più semplice, in quanto parte dalla conoscenza dell'area o delle dimensioni degli oggetti per stabilire la soglia da applicare all'immagine. Si suppone che in una data applicazione gli oggetti occupino una determinata percentuale dell'immagine. In questo modo si può suddividere l'istogramma dell'immagine di partenza e scegliere una o più soglie per assegnare determinate percentuali di pixel agli oggetti.

Questo metodo ha un uso chiaramente limitato. Solo poche applicazioni, come i lettori di pagine, permettono di applicarlo alla generalità dei casi.

2. Metodo della moda.

   Se gli oggetti dell'immagine hanno gli stessi valori di grigi, diversi dallo sfondo, e non vi è rumore, si può assumere che l'istogramma presenti due picchi separati, quindi la soglia può essere individuata in un punto qualsiasi tra questi due valori. Nella pratica questo non succede, in quanto, per vari fattori sono presenti molti valori intermedi di grigio, con la conseguenza che non si avranno picchi separati ma delle creste d'onda, è possibile, tuttavia collocare la soglia nella valle tra due creste dell'istogramma.

   Il problema d'individuare picchi e valli non è insignificante e molti sono i tentativi di risolverlo. Uno è di ignorare i picchi locali, considerando solo quelli che si trovano a certe distanze. La rilevazione dei picchi si basa sull'altezza degli stessi e sulla profondità delle valli, mentre la distanza tra valli e picchi è ignorata.

   Questo approccio può essere generalizzato alle immagini aventi molti oggetti con diversi valori medi di grigio.

3. Selezione interattiva della soglia

   Si stabilisce una soglia iniziale che viene successivamente affinata. Ci si aspetta che dalla soglia applicata alla prima immagine se ne ricavi una che possa essere usata per individuare una soglia migliore. Dall'algoritmo di modifica dipende il successo di questo approccio.

4. Soglia adattativa.

   Se l'illuminazione della scena è ineguale, lo schema precedente può non essere adatto. La mancata uniformità può essere dovuta a delle ombre o alla direzione

dell'illuminazione. In questi casi gli stessi valori di soglia non possono essere usati per tutta l'immagine.

L'approccio adattativo analizza l'istogramma dell'intera immagine, per poi suddividerla in sotto immagini. Di ognuna di queste si calcola poi la soglia, basandosi sui relativi istogrammi. La segmentazione finale dell'immagine scaturirà dalla riunione del tutto.

5. Soglie variabili.

   Sempre nel caso di illuminazione ineguale, si possono approssimare i valori d'intensità dell'immagine attraverso semplici funzioni piane o biquadratiche. La funzione adatta è in gran parte determinata dai valori di grigio dello sfondo. L'istogramma e la soglia possono essere valutati in relazione al livello base dato dalla funzione appropriata.

6. Doppia soglia.

   In molte applicazioni, i valori di grigi appartenenti agli oggetti sono conosciuti. Vi possono infatti essere valori di grigio aggiunti appartenenti sia agli oggetti che allo sfondo. In questi casi si può utilizzare una prima soglia per ricavare gli oggetti principali e un secondo metodo per migliorarli. Spesso, quest'ultimo è una seconda soglia. Con essa si dovrebbero individuare i pixel che hanno un vicino sicuramente appartenente all'oggetto oppure, usando le caratteristiche di luminosità rilevate con l'istogramma, individuare i punti da includere nella regione dell'oggetto.

   Questo algoritmo implementa i principi di somiglianza e prossimità spaziale. I pixel dei margini hanno valori vicini ai pixel "*core"* già appartenenti all'oggetto, dato che i due insiemi di pixel sono adiacenti nell'istogramma, e sono anche vicini nello spazio in quanto confinanti.

La limitazione più pesante incontrata dal metodo di segmentazione basato sull'istogramma è la perdita delle informazioni relative alla posizione dei valori d'intensità dell'immagine. Molte immagini con diversa distribuzione spaziale possono avere istogrammi d'intensità simili. La natura globale dell'istogramma limita la sua applicabilità a scene complesse. Esso non evidenzia il fatto importante che punti dello

stesso oggetto solitamente sono vicini in quanto appartenenti alla stessa superficie[93]. Per questo è necessario operare sullo spettro di frequenza spaziale.

Un ulteriore utilizzo dell'istogramma è la ridistribuzione dei valori di grigio in quelle immagini che hanno i valori d'intensità posizionati in un raggio ridotto, ad esempio quelle a basso contrasto. Questa operazione è definita equalizzazione dell'istogramma e permette di aumentare il contrasto e spesso, di conseguenza, la qualità dell'immagine[94].

La trasformata di Fourier

La trasformata di Fourier permette di modificare le immagini agendo sullo spettro di frequenza. Questo è possibile in quanto l'immagine è riconducibile ad un segnale.

Un segnale è definito come una successione di valori discreti o continui derivati dalla misurazione della variazione di una qualche grandezza nel tempo e nello spazio. Ricordando come si è ottenuto il reticolo da una singola linea, si può dire che un'immagine è un segnale bidimensionale discreto derivato dal campionamento spaziale della funzione di luminosità[95].

Al fine di analizzare il contenuto in frequenze spaziali dell'immagine è necessario utilizzare la trasformata di Fourier discreta, ossia generare un segnale bidimensionale discreto derivato. Il segnale così ottenuto sarà composto da numeri complessi, dove nella parte reale sono contenute le informazioni relative alla frequenza, mentre nella parte immaginaria quelle relative alla fase[96].

Il modulo (valore assoluto) di questa funzione bidimensionale discreta è, quindi, una funzione bidimensionale reale, che rappresenta il contenuto delle frequenze spaziali dell'immagine di partenza. E' possibile utilizzare questa definizione di spettro di frequenza e sostituirla alla precedente, ottenuta con un metodo euristico[97]. Dato che il modulo è un segnale, esso può a sua volta essere rappresentato da un'immagine e,

93 Ibidem p. 86
94 Ibidem p. 112
95 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., p. 48
96 Ibidem
97 L'euristica (dal greco ευρίσκω, heurísko, letteralmente "scopro" o "trovo") è una parte dell'epistemologia e del metodo scientifico.
È quella parte della ricerca il cui compito è di favorire l'accesso a nuovi sviluppi teorici o scoperte empiriche. Si definisce infatti procedimento euristico un metodo di approccio alla soluzione dei problemi che non segue un chiaro percorso, ma si affida all'intuito e allo stato temporaneo delle circostanze, al fine di generare nuova conoscenza. È opposto al procedimento algoritmico.
http://it.wikipedia.org/wiki/Euristica 25 gennaio 2007

quindi, manipolato con un qualunque programma di elaborazione delle immagini, agendo direttamente sui valori dei pixel[98].

Grazie alla trasformata di Fourier si possono eseguire importanti operazioni sull'immagine quali:

1. Rimuovere il rumore ad alte frequenze dall'immagine.
2. Estrarre le caratteristiche strutturali (*texture*) che possono essere utilizzate per individuare la tipologia di oggetti presenti in una regione dell'immagine.
3. Compressione delle immagini.

Il primo punto è già stato in parte trattato nel paragrafo relativo ai filtri.

Il secondo è un approccio alternativo all'estrazione dei contorni e di altre caratteristiche dell'immagine suggerito dalle teorie della psicofisica. Consiste, molto semplicemente, nello scomporre l'immagine in serie di Fourier (l'immagine viene rappresentata attraverso la somma di funzioni periodiche di seno e coseno) e nel costruire la funzione rilevando i punti in cui le fasi delle armoniche concordano. Secondo queste ricerche, i contorni dovrebbero coincidere con i punti in cui la concordanza di fase ha un massimo locale. Questo approccio, nonostante la complessità, sembra essere quello più vicino alla visione umana, dato che produce illusioni simili.

Proprio per questo, e si è già al terzo punto, la trasformata di Fourier è utile per comprimere le immagini. In effetti, lo standard JPEG si basa su di essa, permettendo di considerare solo i dettagli dell'immagine importanti per la visione umana e diminuendo, parallelamente, lo spazio occupato in memoria[99].

Tra le operazioni che possono essere compiute in quello che possiamo definire lo spazio di frequenza (l'immagine rappresentante la funzione bidimensionale reale) vi è la convoluzione[100]. Questa operazione consiste nel far scorrere una maschera (un'altra matrice contenente dei valori definiti *pesi*) sull'immagine, centrandola, in sequenza, su ciascun pixel; calcolare il prodotto tra i valori dei pixel e i pesi relativi alle varie posizioni e, infine, sommare tutti i pixel considerati[101]. Normalmente tale operazione è

98 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., p. 49
99 Ibidem p. 71
100 Ibidem p. 49
101 Ibidem p. 42

impiegata in modo analogo ai filtri, con la differenza che si considera l'intera immagine. Essa è importante in quanto, per una delle proprietà della trasformata di Fourier, se la si applica al dominio spaziale è come si eseguisse il prodotto delle trasformate nel dominio di frequenza. L'equivalenza delle due operazioni è garantita dal fatto che la trasformata di Fourier è invertibile (*antitrasformata* di Fourier). Calcolare un prodotto è molto più efficiente a livello computazionale che calcolare una convoluzione, utilizzando la trasformata di Fourier per passare da un dominio all'altro si possono diminuire i tempi di elaborazione.

Oltre alle operazioni che modificano l'intera immagine, ve ne sono alcune che operano a livello globale, piuttosto che locale o puntuale, per estrarre informazioni riguardanti le superfici, le dimensioni, ecc

Estrazione di superfici dalle ombre

E' possibile costruire una mappa dell'indice di riflessione dove venga registrata la luminosità dei pixel in funzione dell'orientamento della superficie su cui giacciono i punti nella scena reale. In tal modo, avendo illuminazioni fisse durante la formazione dell'immagine e conoscendo l'indice di riflessione della superficie, si possono tradurre le variazioni dell'orientamento in variazioni d'illuminazione dell'immagine stessa[102].

Ponendo alcuni vincoli, ad esempio che si tratti di superfici lisce, è successivamente possibile risalire alle loro forme. Questo costituisce comunque uno svantaggio, in quanto le superfici reali non sono sempre lisce[103].

Stereo-Fotometrica

Il metodo della stereo-fotometrica registra immagini multiple (almeno due, di norma tre) di un oggetto illuminato in sequenza, da differenti sorgenti di luce[104]. Questo al fine di ricostruire più parti di superfici e misurarne le dimensioni delle forme-oggetti inferendone la profondità. Per riuscire in quest'operazione si stabiliscono dei vincoli relativi all'illuminazione della scena[105] o alla fissità degli oggetti.

---

102 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 269
103 Ibidem
104 Cfr Linda G. Shapiro, George C. Stockman, *Computer Vision,* Op. Cit., pp. 471-472
105 Cfr David A. Forsyth, Jean Ponce, *Computer Vision: A Modern Approach,* Op. Cit., pp. 81-82

Uno dei vantaggi di questo metodo consiste nel fatto che i punti delle immagini sono perfettamente registrati l'uno con l'altro: non hanno cioè problemi di corrispondenza, in quanto sia la fotocamera sia la scena sono fissi. D'altra parte, questo costituisce anche uno svantaggio, in quanto condizioni così rigide sono difficili da realizzare e/o controllare. Il metodo rileva inoltre la profondità solo in modo indiretto.

La profondità

E' risaputo che nella rappresentazione 2D della scena 3D si perde la dimensione della profondità e che il processo inverso costituisce un problema non immediatamente risolvibile. Calcolare la distanza relativa di vari punti della scena dal punto di acquisizione è uno dei compiti più importanti per un sistema di computer vision. Uno dei metodi più utilizzati per riuscirvi è acquisire una copia di immagini utilizzando due macchine fotografiche di cui si conosca la distanza che le divide, in pratica simulando una stereo-visione. In alternativa si può utilizzare una telecamera, spostandola secondo necessità[106].

Esistono anche le immagini spaziali (*range image*), in cui la profondità è rilevata direttamente grazie a sensori che si basano sui principi del radar e della triangolazione.

Profondità dalle immagini d'intensità

Nel caso si voglia simulare la stereo-visione (ossia acquisendo due immagini, una a destra, l'altra a sinistra), è sufficiente, per comprendere come i punti della scena 3D siano localizzati nello spazio, utilizzare la geometria e l'algebra. La misurazione avviene attraverso la tecnica della triangolazione, misurando la parallasse tra due immagini ottenute riprendendo l'oggetto da due posizioni diverse, separate da una distanza sufficiente[107].

Il metodo della parallasse è descritto molto semplicemente da Asimov nel suo libro di Fisica: "Un metodo per calcolare le distanze cosmiche è quello basato sulla *parallasse.* Il significato di questo termine è facile da spiegarsi: mettete un dito alla distanza di una decina di centimetri dagli occhi e guardatelo prima con l'occhio sinistro, poi con il destro; vedrete il vostro dito spostarsi rispetto allo sfondo, perché avete cambiato il

106 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 289
107 Cfr Linda G. Shapiro, George C. Stockman, *Computer Vision,* Op. Cit., p. 397

vostro punto di vista. Se ora ripetete lo stesso procedimento tenendo il dito più lontano, per esempio alla distanza del braccio teso, esso si sposterà ancora rispetto allo sfondo, ma meno di prima; l'entità dello spostamento può quindi servire a determinare la distanza del dito dai vostri occhi.

Naturalmente, se un oggetto dista una ventina di metri il cambiamento di posizione quando lo si guarda con l'uno o l'altro occhio comincia a essere troppo piccolo per essere misurato; si deve avere una <<linea di base>> maggiore della distanza tra i due occhi. Per ottenere uno spostamento maggiore del punto di vista basterà guardare l'oggetto prescelto da una determinata posizione, poi spostarsi, per esempio di qualche metro a destra, e guardare di nuovo: ora la parallasse è sufficiente per essere facilmente misurata e si può determinare la distanza. È proprio a questo metodo che si ricorre per determinare l'ampiezza di un fiume o di un burrone."[108].

La profondità di vari punti della scena può essere quindi recuperata conoscendo la distanza (disparità) tra punti corrispondenti nelle diverse immagini[109].

È da notare che a causa della natura discreta delle immagini digitali, i valori della disparità sono numeri interi, a meno che particolari algoritmi siano usati per aumentarne l'accuratezza. Solitamente, però, si preferisce aumentare la distanza della <<linea di base>> così da aumentare anche la disparità. Questa soluzione introduce, a sua volta, atre problematiche relative, ad esempio, all'ampiezza del campo visibile o alle distorsioni introdotte dalla prospettiva[110].

Questa tecnica dà per implicito che si possa identificare le coppie di punti congiunti nelle immagini stereo. Tuttavia questo non è così semplice, tanto da costituire il famoso problema di *corrispondenza*: per ogni punto dell'immagine di sinistra, si trovi il corrispondente in quella di destra. Le soluzioni finora trovate si fondano sulla rilevazione di alcune caratteristiche per identificare i singoli punti, come l'appartenenza a bordi o regioni già individuate[111].

Oltre a quello della parallasse ci sono anche altri metodi basati su tecniche alternative che possono fornire alcune informazioni/suggerimenti relativi alla profondità.

108 Isaac Asimov, *Il libro di Fisica*, trad. it. di Carla Sborgi, Milano, Arnoldo Mondadori, 1986 pp. 24-25, (ed. originale *Asimov's New Guide to Science*,1984)
109 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 291
110 Ibidem
111 Ibidem p. 293

L'interposizione è, forse, una delle informazioni nascoste più importanti: gli oggetti che sono più vicini occludono parti degli oggetti che sono lontani; il riconoscimento delle occlusioni fornisce una profondità relativa (es.:una persona davanti a un muro è più vicina al sensore del muro; un uomo dietro un'auto è più lontano dell'auto)[112].

Anche le dimensioni relative degli oggetti sono importanti: un'auto lontana apparirà più piccola e lenta di un'auto vicina.

Il punto di vista da cui si osserva è rilevante per la profondità: si pensi ad una porta aperta che proietta nella retina la forma di un trapezio e non di un rettangolo; il bordo più lontano appare più corto rispetto al più vicino, a causa dell'effetto scorcio (*foreshortening).*

Le textures delle superfici, come detto, cambiano in relazione sia alla distanza dall'osservatore sia al loro orientamento.

Sono stati implementati, negli anni '80, vari metodi per usufruire di queste informazioni/suggerimenti. Tali metodi, tutti indiretti, acquisiscono informazioni sia sulla forma sia sulla profondità: si parla infatti, in generale, di *shape from X*, e nello specifico, di *shape from shading* (forma dalle ombre), *shape from texture*, *shape from focus* (forme dalla messa a fuoco), *shape from motion* (forma dal movimento), nonché di *photometric stereo* (stereo-fotometrica).

Si approfondirà ora l'estrazione di profondità dalla messa a fuoco, rimandando l'estrazione di forme dal movimento all'operazione sugli oggetti. Gli altri metodi sono stati già considerati.

Grazie al fatto che i sistemi ottici hanno una profondità di campo finita, solo gli oggetti a una distanza appropriata appaiono a fuoco nell'immagine, mentre altri risultano confusi/indistinti. Alcuni algoritmi sono stati realizzati per sfruttare quest'effetto. L'immagine è modellata come una convoluzione di immagini a fuoco con una funzione di diffusione del punto[113] determinata in ragione dei parametri della fotocamera e della distanza degli oggetti dalla stessa. La profondità è recuperata dalla stima della confusione nell'immagine e utilizzando una conosciuta o stimata funzione di diffusione lineare. Questa ricostruzione presenta dei problemi di calcolo; tuttavia è utile nelle

112 Cfr Linda G. Shapiro, George C. Stockman, *Computer Vision,* Op. Cit., p. 40
113 http://en.wikipedia.org/wiki/Point_spread_function 1 febbraio 2007

applicazioni che richiedono informazioni dettagliate sulla profondità[114].

La profondità nelle immagini spaziali

Come trattato in precedenza, le immagini spaziali (o mappe di profondità) sono ottenute da apparecchi che misurano la distanza di ogni punto della scena all'interno dell'angolo visivo e li registrano come una funzione bidimensionale.

Due metodi molto usati per la formazione delle immagini spaziali sono la triangolazione e il radar (*Radio Detection And Ranging*)[115].

I sistemi ad illuminazione strutturata, usati in modo estensivo in computer vision, si basano sulla triangolazione per calcolare la profondità.

I sistemi radar per la formazione delle immagini usano rilevatori (*finder*) acustici e laser (*Light Amplification by the Stimulated Emission of Radiation,* amplificazione della luce attraverso l'emissione di radiazioni[116]) per ottenere le mappe di profondità[117].

Sistemi di Visione Attiva[118]

Considerati le due precedenti tipologie d'immagini e i relativi sistemi d'acquisizione, si può introdurre una riflessione sulle loro caratteristiche.

Molti sistemi hanno delle caratteristiche fissate, che includono sia sensori passivi come fotocamere e telecamere, sia attivi come rilevatori laser di profondità.

Tuttavia si sta sempre più diffondendo l'idea che sistemi di visione attiva, di natura contrapposta ai precedenti, dove i parametri e le caratteristiche di acquisizione sono dinamicamente controllati da sistemi d'interpretazione della scena, siano cruciali per la percezione della stessa. Questo è ciò che accade negli esseri viventi, dove i dati sono acquisiti in modo attivo.

I sistemi di visione attiva possono impiegare sensori attivi o passivi. Tuttavia, in questo caso, gli stadi parametrici dei sensori, come messa a fuoco, apertura, vergenza[119] e

114 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 300
115 http://it.wikipedia.org/wiki/Radar 1 febbraio 2007
116 Per una comprensione del funzionamento del laser si veda Isaac Asimov, *Il libro di Fisica*, Op. Cit., pp. 492-498
117 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 300
118 Ibidem p. 305
119 Il movimento simultaneo di entrambi gli occhi nei sensi opposti (convergenza e divergenza) per ottenere o effettuare la singola visione binoculare. http://en.wikipedia.org/wiki/Vergence 1 febbraio

illuminazione sono controllati per acquisire i dati utili a semplificare il compito d'interpretazione della scena stessa.

La visione attiva è essenzialmente un processo d'acquisizione intelligente dei dati, controllato dai parametri misurati e stimati, nonché dai possibili errori provenienti dalla scena.

Una precisa definizione di questi parametri dipendenti dalla scena e dal contesto richiede una completa comprensione non solo delle proprietà della formazione delle immagini e dei sistemi di elaborazione, ma anche delle loro interdipendenze.

Operazioni a livello di oggetti

Operare a livello di oggetti significa anzitutto identificarli. Le operazioni precedenti (puntuali, locali, globali), reiterabili, dovrebbero aver contribuito a migliorare l'immagine, ad evidenziarne alcune caratteristiche e a conoscerne alcune informazioni. Tuttavia, prima di riuscire a riconoscere gli oggetti contenuti in una scena è necessario "interpretare" queste informazioni. Ad esempio, selezionando e unendo i bordi al fine di ottenere i contorni, ed eseguendo una segmentazione più accurata dell'immagine. Proprio la segmentazione è strettamente collegata al riconoscimento degli oggetti: senza almeno un parziale riconoscimento degli oggetti, essa non può essere eseguita, e senza di essa gli oggetti non possono essere riconosciuti.**Segmentazione**

Si è già accennato a questo processo esaminando l'operazione di soglia, l'estrazione di bordi, la rilevazione delle textures e le operazioni con l'istogramma d'intensità; è comunque opportuno richiamarne alcuni aspetti.

Con segmentazione si indica la procedura che consente di suddividere, in base a criteri predefiniti, un'immagine in aree considerate omogenee. Per certi versi questa procedura funziona in modo opposto all'estrazione dei contorni: quest'ultima ricerca delle discontinuità nella funzione di luminosità, mentre le tecniche di segmentazione ricercano pixel con valori d'intensità simili. La maggior parte delle tecniche utilizzate, sebbene presentino differenze anche notevoli, si basano su una filosofia comune detta *region growing*[120].

2007

120 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., p. 70

L'operazione di *region growing* procede come segue: dato un pixel di partenza, lo si etichetta (ossia si procede al *labeling:* al pixel viene assegnato un codice interno completamente slegato dalla funzione di luminosità) e si copia tale etichetta su tutti i pixel dell'intorno, sempre che non differiscano oltre una determinata soglia dal valore del pixel di riferimento. Questa procedura è reiterata per tutti i pixel considerati e s'interrompe quando le differenze di soglia non consentano di proseguire in nessuna direzione. Regioni contrassegnate con la stessa etichetta hanno una buona probabilità di appartenere al medesimo oggetto, o quantomeno a parti di un medesimo oggetto, fornendone quindi una descrizione[121]. Errori di segmentazione possono portare ad una non perfetta corrispondenza tra regioni e oggetti, ragion per cui per una corretta interpretazione dell'immagine è necessaria una conoscenza specifica degli oggetti stessi.

Da quanto detto si desume che i più importanti principi della segmentazione sono due: la somiglianza dei valori e la vicinanza spaziale. Due pixel possono appartenere alla stessa regione se hanno caratteristiche di luminosità simili o se sono vicini tra loro[122].

Superfici[123]

Le operazioni sulle superfici, come quelle sul movimento, sono a cavallo tra quelle globali e quelle sugli oggetti. Per poter risalire agli oggetti è infatti necessario considerare l'intera scena. Per questa ragione i due problemi principali relativi alle superfici nella computer vision riguardano la loro identificazione e la successiva segmentazione. Le superfici devono essere ricostruite dalle misurazioni della profondità, tenendo conto che vi possono essere oggetti a sé stanti. Poi esse vengono segmentate in vari tipi per permetterne il riconoscimento degli oggetti e per una loro migliore considerazione.

Molti oggetti 3D, specialmente manufatti, possono essere facilmente descritti in termini di forma e posizione delle superfici di cui sono costituiti. La descrizione delle superfici è utilizzata per la classificazione degli oggetti, la stima della posizione, l'ingegneria inversa, ed è onnipresente in computer graphics.

---

121 Ibidem p. 71
122 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 74
123 Ibidem p. 365

Il movimento

Finora si sono considerate le elaborazioni su singole immagini, o al limite su due immagini acquisite simultaneamente. Introducendo il movimento si amplia la prospettiva in quanto l'elaborazione dell'immagine abbraccia la dimensione temporale. Più precisamente, si può disporre delle informazioni visive che possono essere estratte da variazioni spaziali e temporali presenti in una sequenza d'immagini[124].

La dimensione temporale nel processo visivo è importante principalmente per due ragioni. In primo luogo, l'apparente movimento degli oggetti sul piano dell'immagine dà una forte indicazione per comprendere struttura e movimento tridimensionale. Secondo, i sistemi visivi biologici utilizzano il movimento per inferire le proprietà dell'ambiente 3D con una bassa conoscenza *a priori* dello stesso[125].

Il dato di partenza per un sistema di analisi di una scena in movimento è una sequenza di fotogrammi presa da un mondo in continua evoluzione. Anche la telecamera che riprende può essere in movimento. Ogni fotogramma rappresenta un'immagine della scena in un particolare istante. I cambiamenti, della scena possono essere dovuti al movimento della telecamera, allo spostamento degli oggetti, a variazioni d'illuminazione, o a quelli di struttura, dimensioni o forma degli oggetti[126].

Normalmente si presume che le variazioni della scena siano dovute a spostamenti della telecamera o degli oggetti, e che queste siano rigide o quasi-rigide; altre variazioni non sono ammesse[127].

Il sistema deve rilevare i cambiamenti, determinare le caratteristiche del movimento dell'osservatore e degli oggetti, caratterizzare il movimento attraverso astrazioni di alto livello, ricostruire la struttura degli oggetti e riconoscere gli oggetti in movimento[128].

In applicazioni quali *video editing* e *video database* potrebbe essere richiesto di rilevare delle *macro* variazioni nella sequenza. Queste variazioni suddivideranno il segmento in parti tra loro collegate in quanto simili per tipo di movimenti della telecamera e tipo di

124 Cfr. Emanuele Trucco, Alessandro Verri, *Introductory Techniques for 3-D Computer Vision,* Op. Cit, p. 178
125 Ibidem
126 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 406
127 Ibidem
128 Ibidem

scene nella sequenza[129].

Per quanto riguarda le relazioni tra telecamera (o punto d'osservazione), oggetti e ambiente, si possono presentare le seguenti situazioni, ognuna delle quali richiede una diversa tecnica d'analisi[130]:

Camera fissa, un singolo oggetto fisso, sfondo fisso.

Inserita per completezza, è una semplice scena statica, praticamente una foto su cui applicare le tecniche/algoritmi già esposti.

Camera fissa, un singolo oggetto in movimento, sfondo fisso.

L'oggetto in movimento sullo sfondo comporta dei movimenti dei pixel nell'immagine associati all'oggetto. La rilevazione di questi pixel può svelare la forma dell'oggetto così come la sua velocità e percorso. Questo tipo di sensori è normalmente utilizzato per la sicurezza e la sorveglianza[131].

Camera fissa, più oggetti in movimento, sfondo fisso.

Il movimento di uno o più oggetti può essere tracciato per ottenere una traiettoria o un percorso dai quali sarà possibile trarre indicazioni sul comportamento dell'oggetto. È il caso di una telecamera usata per analizzare il comportamento di alcune persone che entrano in un edificio per affari o altro lavoro. Diverse telecamere possono essere utilizzate per ottenere diversi punti di vista dello stesso oggetto, permettendo quindi di elaborare percorsi tridimensionali. Possibili applicazioni sono l'analisi del movimento di atleti o di pazienti in riabilitazione. Vi è anche un sistema in via di sviluppo che traccia i movimenti, durante un incontro di tennis, della palla e dei giocatori fornendo l'analisi degli elementi del gioco[132].

Camera in movimento, scena relativamente costante.

Una telecamera in movimento provoca dei cambiamenti nelle immagini dovuti al suo stesso movimento, anche se l'ambiente non cambia. Si può utilizzare questo movimento in modi diversi. Ad esempio si può ottenere una più ampia visione dell'ambiente rispetto all'osservazione da un singolo punto fisso: è il caso di un

129 Ibidem
130 Cfr Linda G. Shapiro, George C. Stockman, *Computer Vision,* Op. Cit., p. 252
131 Ibidem
132 Ibidem

movimento panoramico di macchina. Il movimento della camera può anche fornire informazioni sulla profondità relativa degli oggetti, in quanto le immagini di quelli vicini cambiano più velocemente di quelle relative ai lontani. In terzo luogo, esso può dare la percezione o la misurazione della forma di oggetti 3D vicini: i molteplici punti di vista permettono infatti di effettuare calcoli trigonometrici simili alla visione stereo. Elaborando o analizzando il contenuto di film o video, è spesso importante rilevare la posizione e l'istante in cui la telecamera ha effettuato una panoramica o uno zoom: possiamo così ottenere delle informazioni su come, e in che modo, la scena è vista dal sistema[133].

Camera in movimento, diversi oggetti in movimento.

Questa situazione presenta i problemi relativi al movimento più difficili da risolvere e probabilmente più importanti in quanto riguardano situazioni in cui sono in movimento i sensori ma anche una grande quantità di oggetti nella scena osservata. È il caso di un veicolo che si muove nel traffico di punta, o di alcune telecamere che devono seguire automaticamente degli oggetti in movimento[134].

Da tener presente che una sequenza di fotogrammi offre molte più informazioni per comprendere una scena, ma aumenta in proporzione anche la quantità di dati da elaborare. Applicare quindi tecniche per l'analisi di una scena statica ad ogni fotogramma di una sequenza occorrono elevate capacità di calcolo. Questa esigenza va ad aggiungersi alle difficoltà finora rilevate. Tuttavia, le ricerche finora condotte per l'analisi di scene dinamiche hanno fornito soluzioni che ne facilitano l'estrazione d'informazioni rispetto a quelle statiche[135].

L'analisi di scene dinamiche avviene in tre fasi[136]:

1. Periferica.

   Riguarda l'estrazione d'informazioni approssimative circa l'attività presente in una scena; tali informazioni saranno utilizzate, nelle fasi successive, per decidere quale parte della scena richiede una maggiore attenzione.

2. Attenzione.

133 Ibidem
134 Ibidem
135 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 407
136 Ibidem p. 408

Si focalizza sulle parti individuate dalla fase precedente per estrarre informazioni che possono essere utilizzate per riconoscere gli oggetti, analizzare il loro movimento, stendere una lista cronologica degli eventi presenti nella scena, e così via.

3. Cognitiva.

   Applica alla scena la conoscenza pregressa relativa agli oggetti e ai movimenti con lo scopo di comprendere esattamente quali sono e cosa sta accadendo.

Riconoscimento degli oggetti

Un sistema per il riconoscimento degli oggetti è progettato per ricercare oggetti del mondo reale, basandosi su un'immagine che lo rappresenta e utilizzando dei modelli conosciuti. Per definizione, infatti, il riconoscimento implica che le descrizioni degli oggetti, o modelli, siano già disponibili; non si può riconoscere ciò che non si conosce. Questo compito, semplice e istantaneo per l'uomo, è molto difficile per la macchina[137].

Il problema di riconoscere gli oggetti può essere quindi ricondotto ad un problema di etichettatura/denominazione basato su modelli. Se si ha un'immagine contenente uno o più oggetti, e un insieme di etichette corrispondenti a un insieme di modelli conosciuti dal sistema, questo dovrebbe assegnare ad ogni regione, o gruppo di regioni dell'immagine, l'appropriata etichetta. È evidente l'importanza di suddividere/segmentare correttamente l'immagine al fine di denominare e riconoscere gli oggetti[138].

Il sistema per il riconoscimento dovrebbe contenere i seguenti elementi:

1. Il database dei modelli (detto anche *modelbase*)

   Contenente tutti i modelli conosciuti dal sistema. Questi dipendono dall'approccio utilizzato per il riconoscimento e possono variare da una loro descrizione qualitativa o funzionale ad una precisa informativa geometrica delle superfici. Una caratteristica è un attributo dell'oggetto rilevante per la descrizione e il riconoscimento dell'oggetto in relazione con gli altri. Dimensioni, colori e forme sono caratteristiche normalmente utilizzate.

---

137 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 459
138 Ibidem

Il database è organizzato in modo da eliminare, nella fase d'ipotesi, le informazioni relative ad oggetti che possono generare confusione[139].

2. Il rilevatore di caratteristiche

   Il rilevatore di caratteristiche applica degli operatori all'immagine e identifica la posizione di caratteristiche che possono aiutare a formare alcune ipotesi sugli oggetti. Il sistema utilizza le caratteristiche che sono collegate agli oggetti da analizzare e ai modelli presenti nel database[140].

3. Un generatore d'ipotesi

   Questo componente utilizza le caratteristiche rilevate per inferire quali oggetti hanno buone probabilità di essere presenti nella scena. Questo permette anche d'individuare in quali parti dell'immagine questi sono posizionati e di conseguenza ridurre le aree d'analizzare[141].

4. Un verificatore di ipotesi

   Utilizza i modelli degli oggetti per verificare le ipotesi, quindi elimina le forme/informazioni che possono portare in errore o confondere. Solo le forme, superfici, ecc che con tutta probabilità corrispondono realmente ad oggetti vengono selezionate e considerate come entità uniche[142].

Tutti i sistemi per il riconoscimento degli oggetti utilizzano dei modelli che presentano caratteristiche sia esplicite che implicite, utilizzate dai rilevatori in fase d'analisi. I componenti per la formulazione e la verifica delle ipotesi hanno un'importanza variabile a seconda dell'approccio utilizzato per il riconoscimento. Alcuni utilizzano il generatore d'ipotesi e selezionano gli oggetti solo in base alle loro probabilità di essere presenti nella scena. Un esempio sono i metodi che si basano sulla classificazione dei modelli (*pattern classification*). Dall'altra parte troviamo molti sistemi d'intelligenza artificiale che si concentrano sulla fase di verifica delle corrispondenze tra gli oggetti rilevati e le relative ipotesi.

---

139 Ibidem p. 460-461
140 Ibidem
141 Ibidem
142 Ibidem

Classificazione degli oggetti - Pattern recognition

A questo punto, è bene riassumere brevemente il percorso finora fatto. Si è visto cos'è un'immagine digitale, come acquisirla, quali operazioni eseguire per migliorarla ed estrarne informazioni, e come queste, aggiunte ad altre già conosciute presenti in un database, ci portino ad identificare degli oggetti.

Effettuato il riconoscimento, è possibile ottenere ulteriori informazioni sugli oggetti, quali le dimensioni, l'area, localizzarne il centro, ecc[143]; in aggiunta è possibile operare su di essi, ad esempio raggruppandoli per ottenere oggetti compositi (tecnica utilizzata per comprimere il segnale video da trasmettere a distanza[144]), ma anche per individuare dei volti umani in un'immagine[145].

Tuttavia, aver riconosciuto degli oggetti e possedere alcune informazioni su di essi non è equiparabile alla funzione svolta dalla percezione visiva per l'uomo. Vedere ci pone in relazione con il nostro ambiente, permettendoci non solo di riconoscere, ma anche e soprattutto di utilizzare o, comunque, attribuire una funzionalità a ciò che ci circonda[146]. Semplicemente guardando, individuiamo gli oggetti utili ai nostri scopi. Se così non fosse, se la visione non permettesse di utilizzare ciò che ci circonda, se fosse limitata alla sola percezione di forme, posizioni, orientamenti, colori e ogni altra proprietà fisica, non si potrebbe far altro che navigare in un mondo tridimensionale, evitando di urtare gli oggetto di cui è composto. Nella migliore delle ipotesi si potrà riprodurre qualche oggetto, sempre che si conosca e si disponga del materiale di cui è costituito[147].

Tutti i processi finora esaminati hanno lo scopo ultimo di percepire le funzionalità degli oggetti. È questo il vantaggio evolutivo offerto dalla visione. Per l'uomo, la capacità di utilizzare degli oggetti costituisce un tema complesso[148], soggetto a continue disamine, riguardante anche la struttura culturale e interpersonale della società[149]. Trattasi comunque di un argomento che va oltre gli obiettivi iniziali di questo elaborato, e che quindi non si espone per lasciare invece spazio al modo in cui è possibile percepire le funzioni. Esistono essenzialmente due tipi di approccio, già accennati: una è la

143 Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., p. 72
144 Ibidem p. 86
145 Cfr David A. Forsyth, Jean Ponce, *Computer Vision: A Modern Approach,* Op. Cit., p. 537
146 Cfr. Stephen E. Palmer, *Vision Science - Photons to Phenomenology,* Op. cit. p. 409
147 Ibidem
148 Tanto da essere stato individuato come una anello evolutivo, si parla infatti di *Homo faber*
149 Cfr. Stephen E. Palmer, *Vision Science - Photons to Phenomenology,* Op. cit. p. 409

percezione diretta postulata prima dalla *Gestalt,* ed estesa alle funzioni da James J. Gibson (che introduce il concetto di *affordances*)[150], l'altro è quello della categorizzazione.

Probabilmente l'uomo utilizza entrambi questi metodi per percepire le funzionalità degli oggetti, tanto che alcuni ricercatori presumono vi siano,a tal fine, parti distinte del cervello[151]; nella computer vision prevale, al momento, il secondo.

Per la percezione diretta è sufficiente ricordare che essa si basa sull'assunto che le funzioni di un oggetto siano ricavabili dalla suo forma o struttura, senza il bisogno di averne una precedente esperienza o memoria[152].

La categorizzazione è un processo più complesso, in quanto richiede una prima percezione delle proprietà intrinseche di un oggetto al fine di determinarne l'appartenenza a una certa classe e, successivamente, il richiamo dalla memoria delle funzioni ad essa collegate. Da notare che non ci sono limiti teorici alla memorizzazione di funzioni legate ad ogni oggetto; tuttavia, probabilmente ciò non è fisicamente conveniente, ed è forse per questo che gli esseri umani tendono a dimenticare ciò che non usano[153].

All'interno della macro categoria della classificazione è possibile avere diversi approcci, fra questi la *Pattern recognition* e le reti neurali[154].

Il termine pattern recognition può creare confusione in quanto si riferisce a una materia quasi autonoma, che si occupa di immagini 2D; esso comprende varie tecniche d'analisi, che possono essere utilizzate non solo per il riconoscimento e la classificazione degli oggetti, ma anche per la successiva verifica di corrispondenza[155]. La verifica differisce dal riconoscimento e classificazione in quanto si occupa del confronto di un singolo oggetto rilevato con uno o al massimo due modelli dati da immagini o classificazioni[156].

I problemi più rilevanti per la pattern recognition sono, da un lato, la costruzione delle

---

150 Il termine è difficilmente traducibile in quanto coniato da Gibson stesso, sta ad indicare ciò che un oggetto permette, quindi, con una forzatura si potrebbe rendere con “permissioni” o “possibilità/funzionalità”

151 Cfr. Stephen E. Palmer, *Vision Science - Photons to Phenomenology,* Op. cit. p. 413 e 412

152 Ibidem p. 410

153Ibidem p. 413

154Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 474

155Cfr Linda G. Shapiro, George C. Stockman, *Computer Vision,* Op. Cit., pp. 92-93

156Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 481

classi di modelli e, dall'altro, i metodi per collegarvi i potenziali oggetti (tratti dall'insieme di caratteristiche rilevate).

Il primo problema è stato finora risolto con la creazione di database che abbracciano conoscenze specifiche. L'esempio classico è un programma di riconoscimento testi, la cui conoscenza/database/memoria riguarda le caratteristiche dei caratteri dell'alfabeto[157]. Il problema di espandere le classi di modelli a più materie è più complesso di quanto possa sembrare, tanto che alcuni ritengono possa essere superato solo riproducendo la struttura fisica del cervello umano, ossia con le reti neurali, permettendo alle macchine di apprendere autonomamente. Nella *machine learning*, le operazioni avvengono in modo *unsupervised*, ossia senza l'intervento umano, e il sistema determina autonomamente sia il numero sia la struttura delle classi[158].

Il collegamento tra oggetti e modelli presenta delle difficoltà legate alla somiglianza e/o complessità degli elementi, nonché alla possibilità di sovrapposizioni. Per questo sono stati implementati degli algoritmi complessi che per brevità si citano solamente: classificatori che utilizzano i vicini più prossimi, classificatori di Bayes, classificatori che utilizzano la classe media più vicina, tecniche strutturali, classificatori che utilizzano una struttura decisionale ad albero o dati multidimensionali.

Come anticipato, anche le reti neurali possono essere usate per implementare una classificazione di modelli e oggetti. La loro convenienza risiede nel fatto che possono utilizzare anche classi di contorni non lineari[159] per ripartire l'insieme di caratteristiche rilevate. I contorni sono ottenuti sottoponendo la rete ad apprendimento/training. Durante questa fase, vengono mostrati al sistema molti esempi di oggetti da riconoscere. Se gli esempi dell'insieme di prova sono selezionati accuratamente, al fine di prevedere tutti i possibili oggetti che si troveranno nella fase di riconoscimento, la rete potrà apprendere la classificazione dei contorni nelle sue caratteristiche spaziali. Nella fase di riconoscimento, essa si comporterà come gli altri algoritmi di classificazione. I vantaggi, rispetto a questi, sono la capacità di considerare classi di bordi non lineari, la

157Cfr. Roberto Marangoni, Marco Geddo, *Le Immagini digitali,* Op. cit., p. 76 e Cfr Linda G. Shapiro, George C. Stockman, *Computer Vision,* Op. Cit., p. 98-100

158Cfr Linda G. Shapiro, George C. Stockman, *Computer Vision,* Op. Cit., pp. 119-126

159 In matematica un sistema è non lineare (si usa anche nonlineare) quando il suo comportamento non è riconducibile alla semplice somma delle parti che lo descrivono o di loro multipli. Questo significa che in un sistema lineare è possibile fare delle assunzioni e delle approssimazioni, mentre, per un sistema non lineare, questo non è possibile. http://en.wikipedia.org/wiki/Nonlinear 11 febbraio 2007

possibilità di apprendimento e l'auto-apprendimento della rete. Tuttavia, vi sono dei limiti che riguardano l'impossibilità d'inserire/trasferire la conoscenza già acquisita relativa ad un determinato dominio, operazione semplice nei normali computer, e le difficoltà di controllo/correzione (*debugging*) delle performance del sistema[160].

160 Cfr. Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision,* Op. Cit., p. 479

Relazioni con l'intelligenza artificiale

L'interrogativo di fondo posto all'inizio è se la computer vision, o meglio la riproduzione della percezione in generale, debba essere un passaggio obbligato per replicare/costruire la vita. Dovrebbe essere oramai chiaro che per ora non esiste una risposta che possa essere certa, soddisfacente e condivisa. Inoltre, il quesito stesso è per molti versi troppo sintetico e necessita di essere interpretato. È ora necessario precisare il significato di vita artificiale.

La vita artificiale (detta anche *Alife o alife*) è sia una materia di ricerca sia una forma d'arte che esamina i sistemi collegati alla vita, ai suoi processi ed evoluzioni. Questo avviene attraverso simulazioni che utilizzano modelli computazionali, robotici e biochimici (che vengono, rispettivamente denominati, approcci "*soft*", "*hard*", e "*wet*"). Dato che la simulazione al computer è per il momento prevalente, parlando di vita artificiale ci si riferisce spesso a questa[161]. Il termine è stato coniato da Christopher Langton verso la fine degli anni '80, quando ha tenuto la prima "Conferenza Internazionale sulla Sintesi e Simulazione dei Sistemi Viventi" (anche nota come Artificial Life I) presso il Laboratorio Nazionale di Los Alamos, nel 1987[162].

La vita artificiale, quindi, non si pone semplicemente lo scopo di ricreare la vita biologica (esperimento riuscito a Stanley Lloyd Miller nel 1952[163]), ma lo studio, la riproduzione, la sintesi di interi sistemi viventi[164].

Riuscire a riprodurre un sistema "vivente", sebbene possa fornire interessanti informazioni sull'evoluzione umana, non dimostra però come si sia potuta sviluppare l'intelligenza. Questa è uno dei fattori, ma non l'unico, che ha permesso all'uomo di "dominare" l'ambiente-mondo che lo circonda[165]. La presenza di un essere intelligente

---

161 http://en.wikipedia.org/wiki/Artificial_life#_ref-1 11 febbraio 2007

162 http://it.wikipedia.org/wiki/Vita_artificiale 11 febbraio 2007

163 Cfr. Isaac Asimov, *Civiltà extraterrestri*, trad. it. Paola Cusumano, Massimo Parizzi, Milano, Arnoldo Mondadori, 1986 p. 166, (ed. originale *Extraterrestrial Civilizations,* 1979), oppure dello stesso autore *Il libro di Biologia,* Op. cit. p. 154 e in rete http://www.cosediscienza.it/bio/07_vita.htm 11 febbraio 2007

164 È questa l'approccio forte alla vita artificiale, sostenuto tra gli altri da Tom Ray, che ritiene la vita sia un'astrazione indipendente dal substrato materiale. Tom Ray dimostra questi suoi convincimenti con un programma di simulazione denominato Tierra. http://en.wikipedia.org/wiki/Artificial_life#_ref-1 11 febbraio 2007 http://www.his.atr.jp/~ray/ 11 febbraio 2007 http://www.his.atr.jp/~ray/tierra/index.html 11 febbraio 2007

165 Cfr. Isaac Asimov, *Civiltà extraterrestri*, Op. Cit., p. 1191-194

influenza, notevolmente, la configurazione complessiva di un sistema/ambiente vivente.

È bene precisare che l'intelligenza non porta solo vantaggi. Essendo legata al volume e alla struttura del cervello, e questo a quella del corpo, se ne trae che un organismo dev'essere relativamente grande. Questo comporta che la sua presenza nell'ambiente non possa superare un certo numero. Inoltre, per avere un effettivo vantaggio sulle altre specie, tale specie deve vivere abbastanza a lungo (altrimenti non riuscirebbe ad apprendere abbastanza), e riprodursi quindi con una certa lentezza[166]. Se ne può dedurre che la realizzazione di sistemi per l'intelligenza artificiale, capaci di memorizzare ingenti quantità di dati per tempi incalcolabili, può apportare dei vantaggi non indifferenti.

Definire esattamente cosa è l'intelligenza non è semplice: la si potrebbe descrivere come l'insieme dato dalle funzioni conoscitive, adattative e immaginative, in possesso dell'uomo e di alcuni animali, grazie ai loro cervelli. L'intelligenza è quindi riconducibile alle capacità di ragionare, apprendere, risolvere problemi, comprendere le idee e il linguaggio[167]. Per quanto riguarda le macchine, sapere se saranno in grado di compiere operazioni simili (in breve, di pensare) è un quesito cui molti hanno cercato, invano, di dare una risposta. Allan Turing ha bypassato il problema proponendo un test: si tratta di un gioco che consiste, in versione semplificata, nel tentativo di una macchina, di convincere un intervistatore umano che essa è umana. Il test è importante, non per il risultato, ma in quanto propone la possibilità di etichettare come intelligente una macchina in base alle capacità ad essa richieste. Può essere una proposta opinabile, tuttavia essa fornisce un riferimento per valutare il grado d'intelligenza raggiunto dalle macchine[168].

A questo punto è possibile riformulare la domanda iniziale: la computer vision è un passo obbligato per l'intelligenza artificiale? In questo caso, la risposta è certamente affermativa. Sono molti a sostenere questa ipotesi, a volte considerandola scontata. Percepire significa infatti apprendere con la mente, ed è quindi indispensabile per conoscere. È per questo che lo studio della percezione è divenuto un importante ambito di ricerca, pur differendo da quello classico della risoluzione di problemi (*problem*

166 Ibidem p. 192
167 Cfr Nils J. Nilsson, *Intelligenza artificiale,* Op. Cit., p. 21 si veda anche http://it.wikipedia.org/wiki/Intelligenza_(psicologia) 11 febbraio 2007
168 Ibidem p. 24-25

solving), del "ragionamento". Quest'ultimo tipo di ricerca richiede sia fornita, per procedere alla formulazione di una risposta, un'ampia conoscenza della materia: il problema deve essere ben posto[169], in quanto la risposta dev'essere unica e ben definita. La percezione, invece, è caratterizzata da problemi mal posti dove le informazioni per decidere sono insufficienti, e si può giungere a soluzioni diverse e non ben definite (es.: illusioni)[170].

Se la computer vision è importante per la ricerca sull'intelligenza artificiale, è vero anche l'inverso, ossia vi è un rapporto di interdipendenza. La percezione permette la conoscenza e questa, a sua volta, consente la percezione. Infatti, l'utilizzo di una conoscenza *a priori* permette di restringere la classe di soluzioni possibili di un problema mal posto (cercando di ricondurlo ad un problema ben posto)[171]. In modo più formale, è possibile dire che l'approccio non è solo di tipo bottom-up, come avviene per la vita artificiale, ma anche di tipo top-down, in quanto devono essere utilizzate delle conoscenze già acquisite (assunzioni nascoste) e vi devono essere dei meccanismi di continuo feed-back[172]. È grazie all'utilizzo dell'intelligenza artificiale che diviene possibile, automaticamente, trasformare una rappresentazione numerica dell'immagine in una rappresentazione simbolica, attraverso i passaggi intermedi d'identificazione e descrizione degli oggetti[173]. Questo è particolarmente importante se si vogliono implementare sistemi di visione attiva.

È necessario precisare che le assunzioni utilizzate per implementare l'algoritmo possono risultare, in casi specifici, false e provocare quindi delle *illusioni ottiche*[174].

Prima di passare alle possibili applicazioni, è importante sottolineare che ai livelli più alti dell'elaborazione delle immagini, dove le operazioni utilizzano simboli, vi è un avvicinamento sostanziale all'altro ramo dell'intelligenza artificiale, il *problem solving*: in entrambi i casi la ricerca è focalizzata sulla riduzione delle possibili soluzioni di problemi utilizzando la conoscenza delle probabili cause di un determinato

169 La distinzione tra problemi ben e mal posti è quella definita da Hadamard, Cfr. Tomaso Poggio, *Visione: l'altra faccia dell'Intelligenza Artificiale,* in *Mente umana, mente artificiale,* a cura di Riccardo Valle, Op. Cit. p. 283

170 Cfr. Tomaso Poggio, *Visione: l'altra faccia dell'Intelligenza Artificiale,* in *Mente umana, mente artificiale,* a cura di Riccardo Valle, Op. Cit. pp. 278-279

171 Ibidem pp. 283-184

172 Ibidem p. 282

173 Ibidem p. 292

174 Ibidem p. 290

evento/fenomeno. Questo non significa, comunque, che i metodi di risoluzione e il tipo di conoscenza siano gli stessi, né che una soluzione matematica possa sempre esistere[175].

175Ibidem pp. 293-294

Possibili applicazioni

Questa tecnologia è applicabile a un così ampio numero di settori che approfondirne solo alcuni equivarrebbe a fornirne un'idea distorta. Per questo motivo si riporta un elenco, reperito in rete[176], dove le imprese che hanno già sviluppato e commercializzato alcuni prodotti sono raggruppate per categorie. Nelle tabelle che seguono compare il nome dell'azienda e la relativa localizzazione, quindi il sito web di riferimento e in fine una descrizione del prodotto.

Assistenza alla guida di veicoli

Iteris (Anaheim, California).
www.iteris.com
Sistemi di partenza intelligenti per auto e camion, per monitorare la posizione su strada (installati, al 2005, in ben 10000 mezzi). Produce anche sistemi per il monitoraggio del traffico.

MobilEye (Jerusalem, Israel)
www.mobileye.com
Realizza sistemi di visione per avvertire gli automobilisti di pericoli, permettendo di controllare la velocità, fornendo, quindi assistenza alla guida.

Smart Eye (Göteborg, Sweden).
www.smarteye.se
Sistemi per monitorare gli occhi del guidatore, rilevando sonnolenza o disattenzione.
Assistenza/Gestione del traffico

Appian Technology (Bourne End, Buckinghamshire, UK).
www.appian-tech.com
Sistemi per leggere le targhe dei veicoli.

AutoVu (Montreal, Canada).
www.autovu.com/website/indexEng.html
Sistemi per leggere le targhe dei veicoli.

Image Sensing Systems (St. Paul, Minnesota).
www.imagesensing.com
Hanno creato il sistema Autoscope che, utilizzando delle telecamere ai lati della carreggiata, permette il controllo/la gestione in tempo reale del traffico.

Cinema e televisione

2D3 (Oxford, UK).
www.2d3.com

176http://www.cs.ubc.ca/spider/lowe/vision.html 13 febbraio 2007

Sistema per il tracciamento di oggetti in film o video e il rilevamento dei movimenti al fine di permetterne l'elaborazione con la computer graphics.

Hawkeye (Winchester, UK).
www.hawkeyeinnovations.co.uk
Utilizza più telecamere per tracciare i movimenti delle palle da tennis o da cricket, al fine di permettere l'arbitraggio o gli eventuali commenti.

Image Metrics (Manchester, England).
www.image-metrics.com
Un sistema di tracciamento per i volti umani, che può essere utilizzato per la mappatura dei movimenti e delle espressioni facciali, allo scopo di riprodurli artificialmente.

Imagineer Systems (Guildford, UK).
www.imagineersoftware.com
Realizza software per l'industria cinematografica basati sulla computer vision.

Mova (San Francisco, California).
www.mova.com
Fornisce misurazioni tridimensionali e il tracciamento di migliaia di punti sui volti o altre superfici per ottenerne l'animazione di personaggi.

Orad (Kfar Saba, Israel).
www.orad.co.il
Sistemi per creare set virtuali, l'analisi di sport, e altre applicazioni di *augmented-reality* in tempo reale.

PVI (Lawrenceville, New Jersey).
www.pvi.tv
Utilizza le tecniche di computer vision per tracciare i movimenti delle telecamere (panoramiche, inclinazioni, zoom) che riprendono scene reali, e inserirvi, in tempo reale, eventuali spot pubblicitari.

QuesTec (Deer Park, New York).
www.questec.com
Sistemi per riprendere azioni sportive e trasmetterle ad una qualità migliore.

REALVIZ (Sophia Antipolis, France).
www.realviz.com
Sistemi software per la cattura del movimento, il tracciamento delle telecamere, la ricostruzione di panorami, e la costruzione di modelli 3D.

Sport Universal (Nice, France).
www.sport-universal.com
Sistema per il tracciamento dei giocatori e della palla durante le competizioni sportive in tempo reale. Prevede l'assistenza umana.

Sportvision (New York, NY).
www.sportvision.com
Sistemi di visione per migliorare le immagini trasmesse di avvenimenti sportivi.
Sistemi visivi a scopi generici

Cognex (Natick, Massachusetts)
www.cognex.com
È una delle più grandi società che si occupa di computer vision. Sviluppano sistemi con compiti d'ispezione e localizzazione, il conteggio delle persone, ecc

Evolution Robotics (Pasadena, California)
www.evolution.com
Sistemi di visione per il riconoscimento di oggetti e navigazione. Le applicazioni includono: robot mobili, rivendite di spezie, e il riconoscimento utilizzando le telecamere dei cellulari.

Neptec (Ottawa, Canada).
www.neptec.com
Sistemi visivi 3D al laser, per essere utilizzati sugli shuttle spaziali, ma utili anche per altri scopi.

Newton Research Labs (Renton, Washington).
www.newtonlabs.com
Sistemi visivi per il tracciamento ad alta velocità e robot mobili.

Point Grey Research (Vancouver, Canada).
www.ptgrey.com
Sistemi di stereo-visione in tempo reale, di visione sferica; hardware per l'acquisizione delle immagini.

Sarnoff (Princeton, New Jersey).
www.sarnoff.com
Sistemi visivi per il tracciamento, la registrazione, la navigazione, la biometrica, ecc

Seeing Machines (Canberra, Australia).
www.seeingmachines.com
Sistemi di tracciamento dei movimenti della testa e del puntamento dello sguardo (direzione fissa).

SpikeNet (Toulouse, France).
www.spikenet-technology.com
Sistemi di visione capaci di apprendere per essere in grado di riconoscere.

TYZX (Menlo Park, California).
www.tyzx.com
Sistemi di stereo-visione in tempo reale, che utilizzano dei chip su misura per unire velocemente le immagini stereo.

Ricerca delle immagini

LTU Technologies (Paris, France).
www.ltutech.com
Recupero delle immagini basate sul contenuto.

Ojos Inc. (Redwood City, California).
www.ojos-inc.com
Ha sviluppato il sistema Riya, per la ricerca e l'etichettatura delle immagini in un database utilizzando il riconoscimento di testi e volti.

Polar Rose (Malmo, Sweden).
www.polarrose.com
Recupero delle immagini basato sul riconoscimento di volti.

Automazione e ispezione industriale: l'industria dell'automobile

BrainTech (Vancouver, Canada).
www.bnti.com
Sistemi per la visione e la guida di robot nell'industria dell'auto.

CogniTens (Ramat-Hasharon, Israel).
www.cognitens.com
Ha sviluppano un sistema accurato di rilevamento di oggetti 3D principalmente per l'industria automobilistica, ma utilizzabile anche in altri settori. Il sistema utilizza 4 telecamere e un proiettore che emana un fascio di luce al fine di evidenziare e riconoscere le textures presenti nella scena.

Perceptron (Plymouth, Michigan).
www.perceptron.com
Produce sistemi tridimensionali di esplorazione al laser.

Automazione e ispezione industriale: l'industria elettronica

ICOS Vision Systems (Heverlee, Belgium).
www.icos.be
Sistemi per l'ispezione elettronica e per l'assemblaggio di componenti e la fabbricazione di semiconduttori.

KLA-Tencor (San Jose, California).
www.kla-tencor.com
Sistemi per l'ispezione e il controllo di processo nella fabbricazione di semiconduttori.

Orbotech (Yavne, Israel).
www.orbotech.com
Sistemi per l'ispezione automatica di schede a circuiti stampati e di video piatti.

RVSI Inspection (Hauppauge, New York).
www.rvsi.com/rvsi/index.html

Sistemi visivi per l'ispezione elettronica e l'assemblaggio.

Automazione e ispezione industriale: l'industria alimentare e l'agricoltura

Dipix Technologies (Ottawa, Canada).
www.dipix.com
Sistemi visivi per l'industria dei cibi cotti. Il sistema controlla la cottura: i colori, la forma, le dimensioni di pane, dolci, tortillas, ecc

Ellips (Eindhoven, The Netherlands).
www.ellips.nl
Sistemi visivi per l'ispezione e la classificazione di frutta e vegetali.

Automazione e ispezione industriale: la stampa e il tessile

Advanced Vision Technology (Hod Hasharon, Israel).
avt-inc.com
Sistemi per il controllo delle stampe ad alta velocità.

Elbit Vision Systems Ltd. (Yoqneam, Israel).
www.evs.co.il
Sistemi di visione per il controllo dei tessuti e altro.

Mneumonics (Mt. Laurel, New Jersey).
mnemonicsinc.com
Sistemi visivi per il controllo della stampa.

Xiris Automation (Burlington, Ontario, Canada).
www.xiris.com
Sistemi d'ispezione per la stampa e l'imballaggio.

Automazione e ispezione industriale: altri casi

Adept (Livermore, California).
www.adept.com
Robot industriali dotati di visione per il posizionamento di componenti e loro ispezione.

Avalon Vision Solutions (Lithia Springs, Georgia).
www.avalonvisionsolutions.com
Sistemi visivi per l'industria della plastica.

Basler (Ahrensburg, Germany).
www.baslerweb.com
Sistemi di controllo per strumenti ottici, sigillanti, schermi, ecc

Hermary Opto Electronics (Coquitlam, BC, Canada).
www.hermaryopto.com
Sviluppa scanner 3D per segherie e altre applicazioni.

JLI vision (Soborg, Denmark).
www.jli.dk
Sistemi visivi per il controllo industriale nei settori alimentare, della lavorazione del vetro, degli strumenti medici e della lavorazione del ferro.

LMI Technologies (Vancouver, Canada).
www.lmint.com
Sviluppano sistemi di visione 3D al laser per il controllo della produzione di legno, strade, veicoli, ecc

MVTec (Munich, Germany).
www.mvtec.com
Sistemi di controllo e altre applicazioni.

NeuroCheck GmbH (Remseck, Germany).
www.neurocheck.com
Sistemi di controllo per la qualità.

PPT Vision (Eden Prairie, Minnesota).
www.pptvision.com
Sistemi visivi per l'industria farmaceutica, dell'auto, dell'elettronica, ecc

SICK IVP (Linköping, Sweden).
www.ivp.se
Piccole telecamere che utilizzano processori dedicati per applicazioni industriali ad alta velocità.

SIGHTech (San Jose, California).
www.sightech.com
Sistemi visivi in grado di apprendere per il controllo e l'automazione.

Virtek Vision International (Waterloo, Ontario, Canada).
www.virtek.ca
Sistemi di modellazione e controllo basati sul laser.

Wintriss Engineering (San Diego, California).
www.weco.com
Sistemi visivi per il controllo di applicazioni web.

Medicina e biomedicina

Claron Technology (Toronto, Canada).
www.clarontech.com
Utilizza sistemi di stereo-visione in tempo reale per rilevare e tracciare la posizione di marcatori per applicazioni chirurgiche.

CTI Mirada Solutions (Siemens) (Oxford, UK).
www.ctimi.com/portals/ctimi/content/about_mirada.html

Sistemi per l'analisi quantitativa delle immagini mediche, compresa la diagnosi del cancro al seno.

Cynovad (Montreal, Canada).
www.cynovad.com
Sistema per far combaciare il colore prostetico dei denti con il colore naturale dei denti del paziente.

Noesis (St. Laurent, Quebec, Canada).
www.noesisvision.com/index_en.htm
Software per uso biomedico e l'analisi scientifica delle immagini.

TriPath Imaging (Burlington, North Carolina).
www.tripathimaging.com
Sistemi visivi per la rilevazione di macchie sul capezzolo che potrebbero indicare la presenza di cellule anormali.

Sistemi per tracciare i pedoni

Reveal (Auckland, New Zealand).
www.reveal.co.nz
Sistema utilizzato per conteggiare e tracciare i pedoni. Utilizza una telecamera montata sul capo.

Monitoraggio sanitario

Vision IQ (Boulogne-Billancourt, France).
www.vision-iq.com
Il sistema Poseidon monitora le piscine per avvisare di eventuali incidenti e possibili annegamenti.

Sicurezza e biometrica

A4Vision (Sunnyvale, California).
www.a4vision.com
Sistema per l'identificazione di volti che utilizza la luce per la ricostruzione 3D. La compagnia ha creato anche il software per il tracciamento dei volti utilizzato nelle web-cam Logitech.

Activeye (Briarcliff Manor, New York).
www.activeye.com
Sistemi visivi per la sorveglianza, che includono il tracciamento, il monitoraggio degli oggetti e l'analisi dei comportamenti.

Aimetis (Waterloo, Ontario, Canada).
www.aimetis.com
Sistema di sorveglianza intelligente.

Aurora (Northampton, UK).
www.facerec.com

Sistema biometrico per il riconoscimento dei volti.

AuthenTec (Melbourne, Florida).
www.authentec.com
Sistema per il riconoscimento delle impronte digitali basato su nuovo sensore.

Digital Persona (Redwood City, California).
www.digitalpersona.com
Sistemi di riconoscimento delle impronte digitali.

EVITECH (Paris, France).
www.evitech.com
Sistemi di video sorveglianza di dimensioni ridotte.

Equinox (New York, NY).
www.equinoxsensors.com
Sistema di sicurezza che utilizza nuovi sensori, come quelli si basano sulla registrazione degli infrarossi o che sfruttano la luce polarizzata.

Geometrix (San Jose, California).
www.geometrix.com
Riconoscimento di volti che utilizza dati tridimensionali provenienti da immagini stereo.

L-1 Identity Solutions (Stamford, Connecticut).
www.l1id.com
Sistemi di riconoscimento delle impronte, dell'iride e dei volti.

ObjectVideo (Reston, Virginia).
www.objectvideo.com
Prodotti per la video-sorveglianza che consentono il tracciamento, il riconoscimento e la classificazione delle attività.

Vidient (Sunnyvale, California).
www.vidient.com
Sistemi di video sorveglianza con riconoscimento del comportamento.

Modellazione tridimensionale

Creative Dimension Software (Guildford, UK).
www.3dsom.com
Creano modelli 3D da un insieme di immagini.

Eos Systems (Vancouver, Canada).
www.photomodeler.com
PhotoModeler software permette la creazione di modelli tridimensionali che mappano le textures da un ridotto numero di foto. È richiesto un inserimento manuale.

Eyetronics (Leuven, Belgium).

www.eyetronics.com
Produce scanner tridimensionali per il corpo umano utilizzanti luce strutturata.

InSpeck (Quebec City, Canada).
www.inspeck.com
Usa la proiezione della luce per creare un modello tridimensionale di textures del volto o del corpo umano in frazioni di secondi.

Videogiochi

GestureTek (Toronto, Canada).
www.gesturetek.com
Traccia i gesti per giocare, o interagire con il computer.

Reactrix (Redwood City, California).
www.reactrix.com
Pubblicità interattiva per proiettori che tracciano i gesti umani.

Sony EyeToy
www.eyetoy.com
Utilizza la computer vision per tracciare i movimenti delle mani e del corpo dei giocatori per controllare la Playstation. Le vendite nel 2004 hanno superato i 4 milioni di unità.