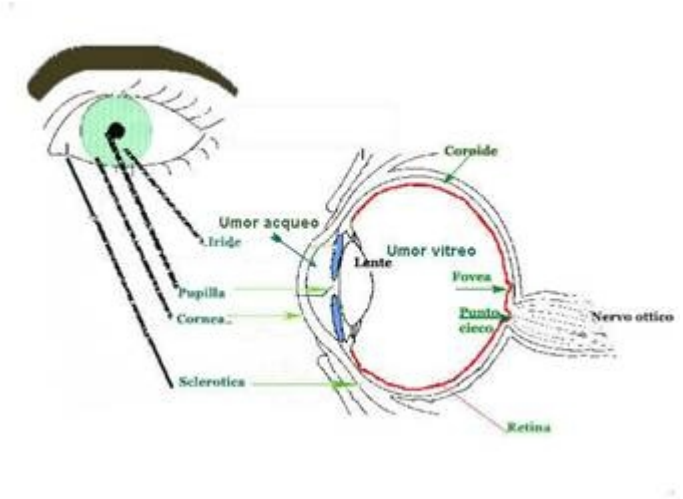

DAVID MARR

ED IO

Differenze e similitudini fra le nostre teorie della visione

^^^^^^^^^^

[Premessa]

Alessandra dopo aver letto qualcosa sul mio sito mi scrive per sapere come si riesca ad isolare un oggetto dall'ambiente, comprensivo di altri oggetti, in cui esso è immerso. Nella visione artificiale questa operazione si chiama (spesso) segmentazione e non è ne semplice ne immediata. Alessandra è una persona colta ma non specialista in materia per cui, nel risponderle, tenterò di evidenziare i concetti, anche sviluppando una breve storia della materia, senza cadere in tecnicismi. Il lettore che volesse approfondire le questioni può iniziare dai libri che consiglio.
In questo articolo rispondo ad Alessandra che ***la segmentazione non è essenziale ai fini del riconoscimento delle forme***. La mia posizione va contro le idee di ***David Marr*** che, esposte negli ultimi anni settanta, continuano a dominare il panorama della visone artificiale. Egli infatti, secondo me, ***è il campione della segmentazione. Accanto alla sua formulazione illustro la mia che distingue fra modello geometrico proprio della percezione visiva e attività cerebrale di elaborazione della forma.*** Le due funzioni, una conscia e una inconscia, sono in parte sovrapposte e ciò provoca confusione.

Sarò grato al lettore che mi vorrà far conoscere le sue opinioni.

La segmentazione e il riconoscimento nella visione artificiale

[Tecniche di segmentazione] - [Visione di basso livello] - [Visione di alto livello]

Ciò premesso, consideriamo, per esempio, una mano su un tovaglia bianca e la più semplice delle procedure di segmentazione: la sogliatura. Si riprende con una telecamera la scena e la si divide in tanti quadratini per ognuno dei quali si rileva il valore della luminosità. Questa tabella di quadratini si chiama matrice di pixels e si considerano appartenenti all'oggetto quei pixels della matrice sotto un certo valore (la soglia), perché la tovaglia bianca genera pixels con alto valore, riflettendo la luce meglio della mano.

Tuttavia il problema non è affatto risolto perché, per esempio, se la mano getta un'ombra sulla tovaglia, l'ombra sarà considerata formante l'oggetto. Passando dalla sogliatura all'estrazione dei contorni, come è avvenuto nella storia della visione artificiale, si va incontro ad altre difficoltà, che non si risolvono usando i colori, la riflettanza delle superfici o il movimento dell'oggetto. Inoltre i risultati di queste varie metodologie, che si chiamano "visione di basso livello", sono in parte in accordo e in parte contradditori fra loro, quindi occorrono scelte intelligenti per discernere i risultati giusti da quelli sbagliati: tali procedimenti vanno sotto il nome di "visione ad alto livello". Un libro tuttora valido per approfondire la visione a basso livello è il testo "Digital picture processing" di Rosenfeld e Kak del 1976.

[Allineamento]

Essi si concretizzano nel confrontare dei modelli precedentemente memorizzati con l'immagine attuale (la mano) tenendo solo quei pixels che corrispondono e togliendo il resto (per esempio, l'ombra nella sogliatura).

Per questo si può ricorrere all'allineamento, che consiste nel deformare dei campioni primitivi, conservati, in memoria fino a far loro assumere la forma dell'oggetto da riconoscere. L'oggetto è riconosciuto con il nome del campione che si è adattato ad esso con la minor deformazione. Il metodo tuttavia non funziona innanzi tutto perchè occorre definire matematicamente che cosa si intende per deformazione (infatti poi la si deve realizzare con delle formule). Precisato il tipo di deformazione, se non si riesce a riconoscere (intendo ad attribuire lo stesso nome), per esempio, due lettere alfabetiche A di fogge tipografiche diverse, i rimedi sono due: si aumentano i campioni o si introducono approssimazioni. Se gli oggetti da riconoscere sono molti, nel primo caso il processo di riconoscimento diventa lunghissimo e improponibile, nel secondo caso porta spessissimo a risultati sbagliati.

Va detto che, nel processo di allineamento, la segmentazione e il riconoscimento vanno di pari passo. Non è che la segmentazione preceda il riconoscimento.

Un miglioramento di tale approccio vuole invece che le procedure di riconoscimento e di segmentazione si richiamino e si affinino ricorsivamente.

[Riconoscimento e segmentazione secondo Marr]

Le sue realizzazioni più in voga sono affinamenti del metodo di David Marr, degli ultimi anni Settanta. Egli ricava dalla scena, usando tutti i metodi detti in principio di questo articolo, una rappresentazione visiva tridimensionale del mondo, ogni oggetto viene poi scomposto in corpi solidi (ad esempio cilindri), prima grandi e, internamente a questi altri piccoli, fino ad approssimarne la forma. In base alla collocazione dei cilindri ed alla loro "gerarchia" la macchina dovrebbe riconoscere l'oggetto. Per rifarmi all'esempio precedente, la mano sarebbe rappresenta ad un cilindro grosso, permettendo così una grossolana segmentazione, poi da sei cilindri: uno grosso, il dorso e cinque piccoli, le dita. Si può convenire che è ben difficile avere molti corpi scomponibili in questo modo, quindi dalla gerarchia e dalla posizione dei cilindri potrebbe essere richiamato univocamente il nome "mano" e la segmentazione di essa. Tuttavia, mi spiace dirlo, anche il metodo di Marr purtroppo non funziona e comporta una pesantezza di calcolo enorme. Le aggiunte di altre caratteristiche utili al riconoscimento, opera dei suoi epigoni, hanno portato ulteriori appesantimenti dei calcoli, senza vantaggi.

Le figure 1,2,3 prese dal libro "Artificial intelligence: an MIT perspective" (Volume 2) MIT press, chiariscono quanto ho detto.

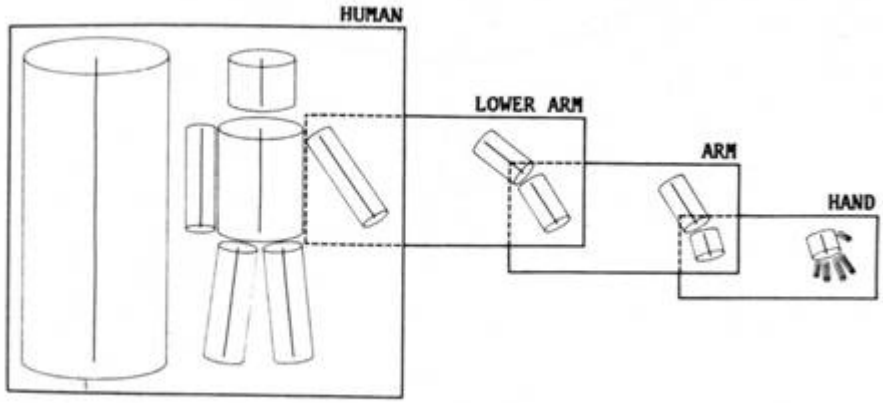

Figura 1

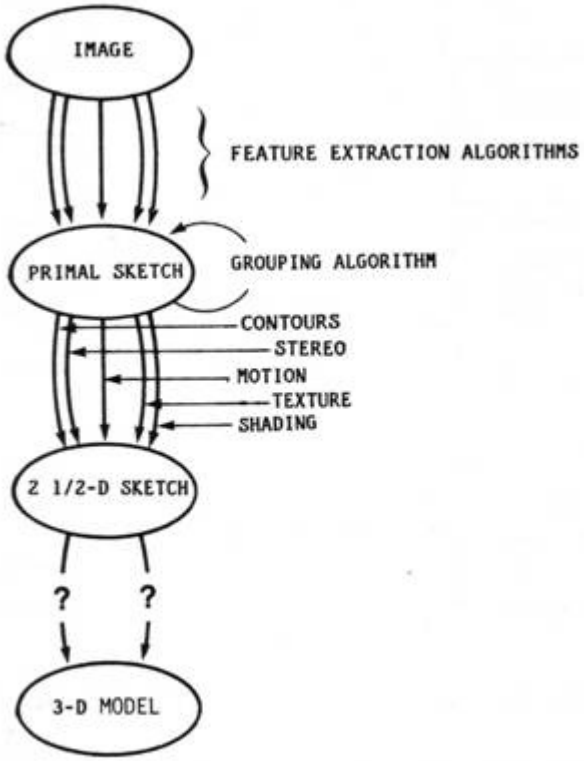

Figura 2

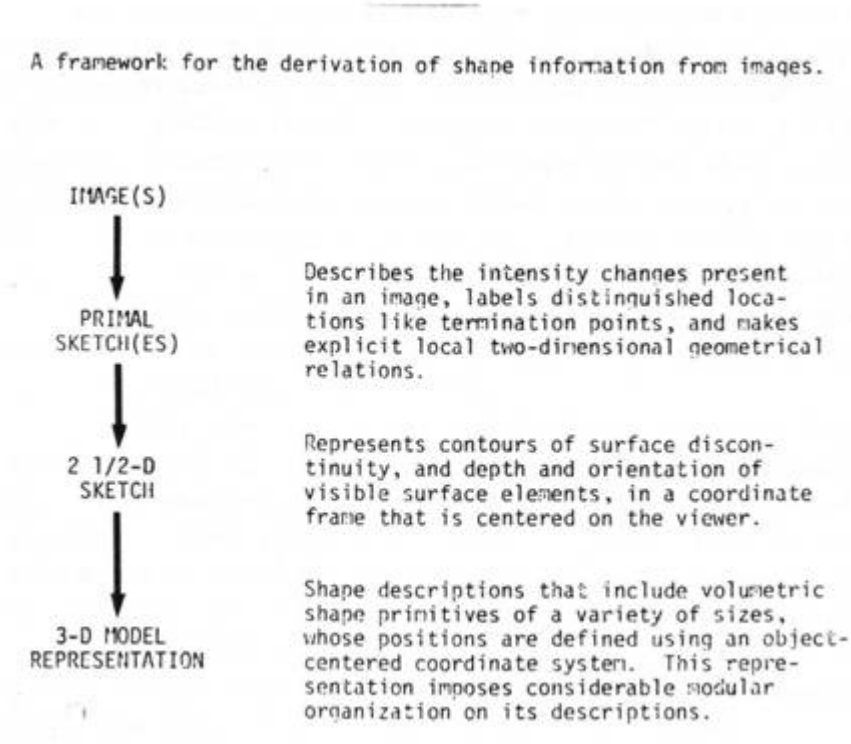

Figura 3

Dagli anni Ottanta ad oggi il metodo di Marr, con migliorie e aggiunte, è quello che ha più credito e seguito fra i cultori della visione delle macchine.

Personalmente due considerazioni mi tengono lontano dal modo di pensare di Marr:

1- la segmentazione, anzi addirittura la costruzione dell'immagine tridimensionale (intesa come precisa forma geometrica di un oggetto), hanno un ruolo fondamentale nel suo metodo. Posto che ci si riesca a ricavare quest'immagine tridimensionale, ai fini del riconoscimento, secondo me si è lavorato inutilmente: noi riconosciamo le figure disegnate senza le ombreggiature e senza quegli elementi che permettono (a volte!) la costruzione tridimensionale. Inoltre il bambino piccolo sorride di fronte a stimoli che solo 24 settimane possono essere considerati un volto completo, si veda la fig. 4. Da Bower, A primer of infant developement, Freeman, San Francisco 1977.

Figura 4

2- l'incastro di cilindri o volumi "ad hoc", entro l'immagine 3D, mi sembra porti ancor più molto lontano dall'operare cerebrale e

[Decomposizione in parti]

in sostanza non mi è chiaro il vantaggio di una rappresentazione tridimensionale, su cui applicare il vecchio metodo della pattern recognition detto "decomposizione in parti", che non funziona nella visione bidimensionale, anche con immagini perfettamente segmentate. Che Nostro centri il sistema di riferimento nell'oggetto, può portare a semplificazioni, che usi coordinate opportune (ad esempio quelle cilindriche) può agevolare i calcoli. Tuttavia, chiunque abbia un po' di pratica di matematiche sa che questi procedimenti non mutano la sostanza del problema.

Con Marr concordo con il metodo multiscala, e con la gerarchia secondo cui organizzare le caratteristiche primitive (i volumi dell'esempio) e con l'attenzione che pone alla neurofisiologia e alla psicofisica. Negli ultimi anni Settanta e nei primi anni Ottanta non era facile ragionare così.

Tuttavia, poiché la segmentazione è una necessità sottostante a tutto il lavoro di Marr non posso fare a meno di rilevare che il mio approccio al problema del riconoscimento delle forme è stato ed è diversissimo ed in esso la segmentazione ha un rilievo minimo, inessenziale e segue il riconoscimento. Spero in questo di aver imitato il comportamento del cervello.

La segmentazione e il riconoscimento nel sistema visivo umano.

[Gradienti e formazione dell'oggetto]

Le bande di Mach, permettono di concludere che nel sistema visivo umano i contorni si formano con il gradiente, anche se non mi è stato possibile appurare quale tipo di gradiente. Io uso un mio sistema (che non è il laplaciano che usa Marr) ma può andar bene qualunque tipo gradiente: tanto secondo il mio metodo di riconoscimento non importa avere contorni precisi da procedure a basso livello. Se si ricavano i gradienti da una figura (con lo stesso operatore!) essi risultano più o meno intensi. I contorni, con il mio metodo, formano attraverso un processo di memorizzazione, a partire dai più intensi. Quando io ho visto la figura 4 ho compreso che ero sulla strada giusta: il mio metodo rifletteva il funzionamento del sistema visivo umano. Si infatti osservi la fig. 5 e la si confronti con la fig. 4, che ho tratto da una rivista della quale mi sfugge il nome, i gradienti più intensi sono quelli memorizzati prima dal bambino.

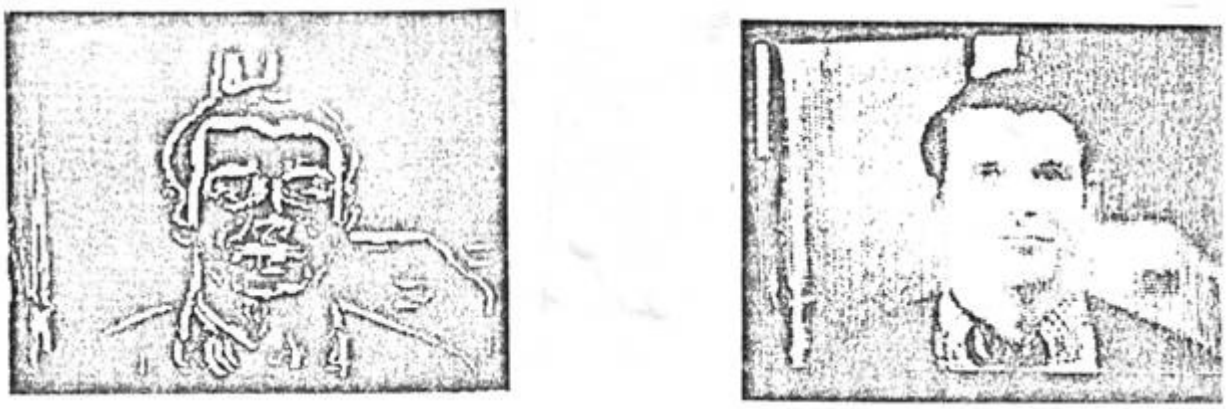

Figura 5

[Inessezialità della segmentazione ai fini del riconoscimento nel bambino]

Sempre osservando la fig. 4 si capisce che la segmentazione segue di molto il riconoscimento.La segmentazione, per la quale è importante estrarre i contorni esterni di una figura è una meta secondaria dell'apparato visivo umano, come del mio metodo.

[Movimento e grossolana segmentazione]

In effetti considerando come è costituita la retina dell'occhio e ricordato che l'occhio esplora la scena con movimenti detti saccadi si capisce che il problema dell'estrazione dei contorni e della segmentazione è mal posto. Circa la retina, verso la sua periferia ci sono molti bastoncelli che si uniscono in fascetti e convergono su una sola fibra del nervo ottico; verso il centro vi sono molti coni ognuno dei quali, a volte, converge su una fibra del nervo ottico. Ne risulta che nell'area visiva le figure che si proiettano sul centro della retina sono ampliate (e di molto) rispetto a quelle che si proiettano alla periferia della retina. Inoltre relative ai bordi della retina vi sono molte cellule fasiche, che rispondono al movimento, mentre per la visone centro tali cellule sono quasi assenti: qui predominano le cellule toniche, che rispondono alla variazione dell'intensità luminosa. Si può pensare che le cellule fasiche individuino la proiezione sulla retina l'oggetto in moto e la portino al centro occupando così l'intero campo visivo. Tale processo è facilmente meccanizzabile. Rozzamente parlando, pensato l'ambiente immoto, basta la differenza fra due matrici di pixels ottenute in successione temporale per rilevare il corpo in movimento. Quando l'immagine è al centro della retina e viene seguita dall'occhio, se essa ancor si muove rispetto all'ambiente, i suoi contorni non sono più ben marcati da una differenza fra matrici di pixels a meno che la figura non sia rigida, non ruoti ed evito di parlare di altre condizioni. Quindi, in generale, dal movimento si ottiene una grossolana segmentazione, che non definisce i contorni della figura. Si limita ad ampliare l'immagine dell'oggetto nell'area visiva, facendolo diventare di gran lunga più importante di ogni altra parte ambientale. La visione centrale può quindi, con processo facilmente meccanizzabile, inibire quella periferica, fenomeno che pare accadere nel sistema visivo umano. In effetti, questo comportamento è evidente anche in alcuni animali superiori. Per esempio, il rinoceronte vede solo gli

oggetti in moto e li posiziona ma li riconosce soprattutto attraverso l'odore.

Nell'uomo, specie in quello civilizzato, posizionamento e riconoscimento sono entrambe attività del sistema visivo ma sono attività differenti e cominciano ad apparire conferme neurofisiologiche di questa mia conclusione. Siccome nell'uomo il riconoscimento degli oggetti attraverso la vista, ha un'importanza di gran lunga superiore di quanto ne abbia negli animali, questo provoca confusione fra le due attività del sistema visivo.

[L'immagine e il suo riconoscimento]

Perché vediamo il mondo intorno a noi, (ed anche udiamo percepiamo gli odori,….) mentre invece non abbiamo sensazione del funzionamento del funzionamento del nostro fegato? Sotto il profilo della nostra sopravvivenza entrambe le attività sono importanti.

La mia risposta, che ho spiegato in altra parte della pubblicazione, è perché il fegato si trova a svolgere sempre lo stesso compito mentre l' essere umano si trova in un ambiente vario e mutevole ed ha esigenze varie e mutevoli. Le opportunità che fornisce l'ambiente vanno colte, i suoi pericoli vanno evitati e tutto questo va posto in relazione ai bisogni del vivente. Quest'attività cerebrale di confronto continuo fra bisogni del vivente e stati ambientali può venire descritta da parole come intenzionalità, consapevolezza, coscienza,…. e produce, parlando lo stesso linguaggio, l'immagine del mondo, la visione, ecc… Tuttavia, quello che noi vediamo non è l'"essere" e tanto meno la sua "essenza" è l'informazione che giunge dal nostro occhio, risultato dell'evoluzione. In altre parole è quanto ci serve per soddisfare i nostri bisogni primari in un certo ambiente. La mosca, la rana, stando alla struttura del loro sistema visivo, hanno una visione del mondo diversa dalla nostra e questo per loro quello è il mondo. Se noi pensassimo che quanto vediamo sia il mondo cadremmo in un antropocentrismo tanto improbabile quanto il geocentrismo. Per me, che sono un discepolo di David Hume, è facile accettare questi ragionamenti. Concordo con lui nell'affermare che è strambo parlare di "realtà fisica": le uniche cose di cui ha senso parlare sono le percezioni e le relazioni fra esse. Tuttavia, essendo nato due secoli e mezzo dopo di lui, filtro il suo pensiero con quello di Darwin e aggiungo che la percezione è legata all'evoluzione e alla sua economia, per cui noi percepiamo quanto ci serve per sopravvivere.

In altre parole, proprio per l'economia del processo evolutivo solo il posizionamento degli oggetti deve essere un'attività cosciente, nel senso che implica un rapporto incessante fra il cervello e il mutevole mondo esterno ad esso. Limitandoci alla vista e omettendo di parlare degli altri sensi, questo funzionamento cerebrale ci appare come la visione del mondo: uno spazio tridimensionale in cui gli oggetti si dispongono e si muovono. Il riconoscimento visivo dell'oggetto è invece una funzione inconscia, automatica, perché coinvolge il solo cervello. E' come il funzionamento del fegato o la respirazione. Siccome ***lo spazio ridimensionale, come lo percepiamo, è un modello del mondo atto a localizzare*** *(grosso modo)* ***e schivare gli oggetti***, ***ritengo sia inutile al riconoscimento visivo dei corpi. Marr invece usa questo modello***, lo vuole precisissimo e come se non bastasse lo scompone secondo le figure della geometria euclidea, che sono tratte da esso. Per questo il mio pensiero è lontanissimo da quello Marr io non sono interessato ad ottenere la perfetta forma tridimensionale di un oggetto per poterlo riconoscere e neanche per poterlo localizzare in modo precisissimo. Infatti operando "human like", per la localizzazione infatti mi basta definire un'area in cui l'oggetto può essere contento.

[Gli elementi utili riconoscimento]

Ripeto, riassumendo, quanto ho già scritto in altra parte di questa pubblicazione. Al fine del riconoscimento delle immagini il mio metodo si incardina due evidenze della neurofisiologia:

1- l'inibizione laterale che mostra come il sistema visivo rilevi i gradienti nelle immagini e

2- la scoperta di Hubel e Wiesel, che notarono come il funzionamento dell'area 17 della corteccia visiva, sia volto a rilevare gli angoli delle figure.

Sono stato il più vicino possibile al funzionamento del sistema visivo ma la neurofisiologia fornisce

poche risposte, per cui ho operato tagli e interpretazioni opinabili, di cui mi assumo la responsabilità. Ad esempio, ho trascurato la visione mesencefalica, importantissima per gli animali inferiori, perché la ritengo una strada che l'evoluzione ha abbandonato. Invece mi sono stato molto utili le illusioni ottiche e mi ha confortato che il mio apparato della visione artificiale cada nelle stesse illusioni ottiche in cui cade l'uomo. Accanto allo schema del riconoscimento secondo Marr, vorrei aggiungere anche il mio, in fig 6,rimandando il lettore alle restanti pubblicazioni su questo sito per la spiegazione dettagliata.

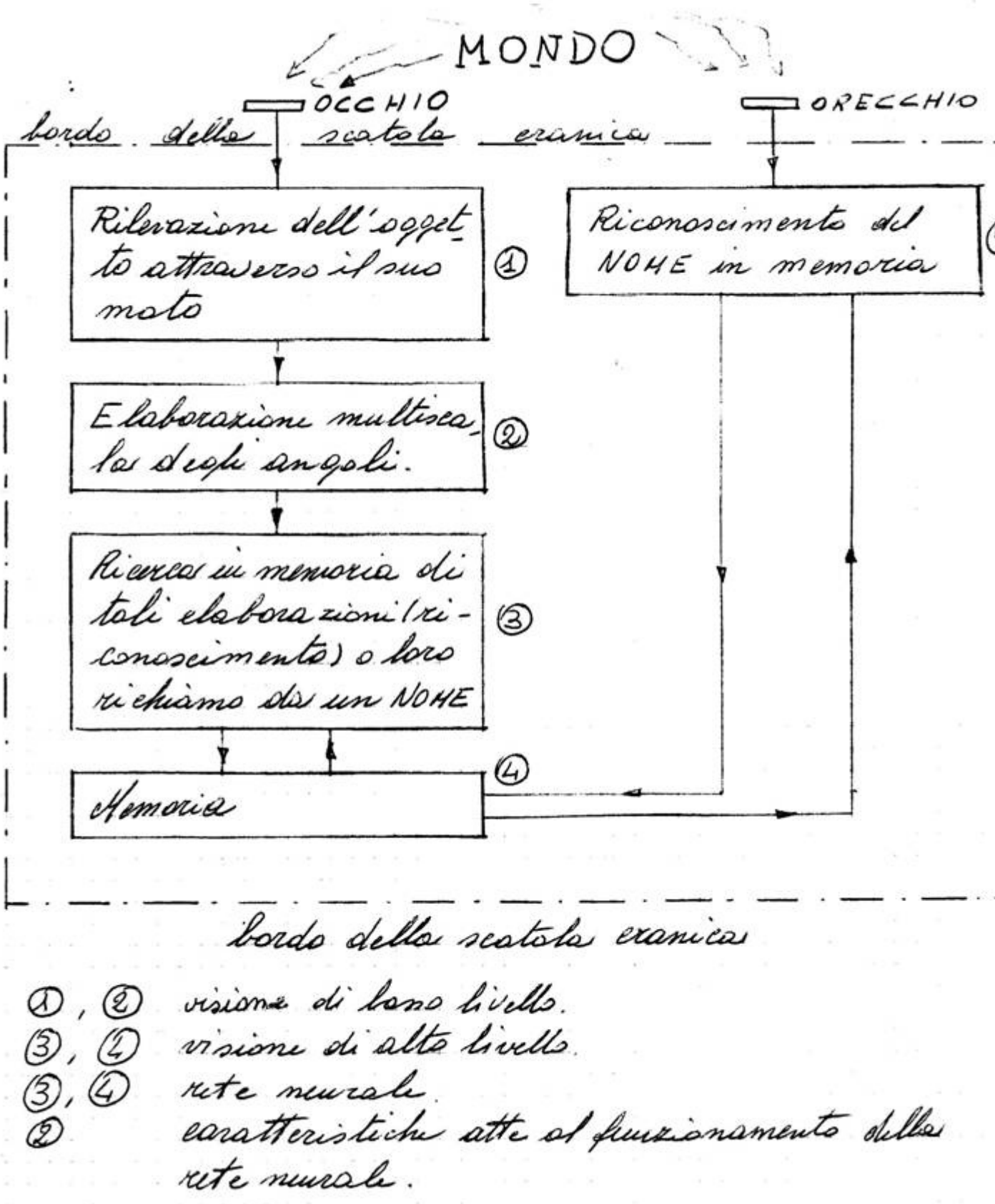

Figura 6

Precisazioni sullo schema in fig.6

Circa il modulo 1 non vi sono argomenti da approfondire.

Circa il modulo 2:

2a- il contorno non si forma sui massimi della superficie formata dai gradienti. Ogni gradiente forma un campo direzionale, estendentesi intorno ad esso, che inibisce il modulo dei gradienti paralleli. Il contorno si forma unendo i massimi dei gradienti così ricalcolati e allineandoli lungo la perpendicolare ad essi. La posizione del massimo del gradiente non viene trasmessa, si ottiene dall'angolazione. Tale processo favorisce la continuità del contorno. Esso è responsabile delle illusioni ottiche che riguardano la malpercezione degli angoli e la dislocazione spaziale delle linee;

2b- si consideri un angolo della geometria elementare: due semirette partenti da un vertice. Se un contorno ha tale forma, i gradienti sui suoi lati, lontano dal vertice, sono perpendicolari ai lati. Sul vertice tali gradienti si sommano e ne risulta (in generale) un gradiente coincidente con la bisettrice di modulo maggiore ai gradienti sui lati perché deriva dalla loro somma. Rilevo che per quanto detto nel punto 2a ognuno dei due lati subisce il campo direzionale dell'altro e quindi non viene percepito con la corretta angolazione, gli angoli acuti si ampliano e quelli ottusi si restringono. Inoltre la sensibilità al gradiente diventa massima lungo la bisettrice;

2c- riflettendo sulla somma dei gradienti di cui ho parlato al punto 2c, si comprende che il gradiente sui vertici degli angoli è tanto maggiore quanto più l'angolo è acuto e richiede che ambo i lati siano lunghi oltre la misura in cui avviene la somma dei loro gradienti. Infatti la somma dei gradienti è inferiore se anche uno solo dei lati è minore di tale misura;

2d- sui gradienti si opera una sogliatura ripetuta (detta contrazione) che elimina gli angoli del contorno il cui valore è sotto la soglia via via crescente. Si ottiene così una spezzata, il contrario dello smoothing (vedi la fig delle R). Rozzamente parlando si può dire che vengono eliminati gli angoli con almeno un lato corto e quelli vicini all'angolo piatto.

I moduli 3 e 4 memorizzano il risultato di ogni sogliatura. Di ogni angolo si conserva in memoria:

3a – il numero cardinale della contrazione (detto potenza della contratta);

3b – se l'angolo è una sporgenza o una rientranza (per quanto questo sia un problema topologicamente irrisolvibile nell'ipotesi di un grossolano isolamento del corpo esso è facilmente meccanizzabile);

3c – l'angolazione della bisettrice;

3d – la sua distanza dall'inizio della contratta

I moduli 3 e 4 sono marcatamente una rete neurale che funziona sulle caratteristiche descritte nei punti 3a, 3b, 3c, 3d. Il confronto fra quanto in memoria e l'immagine presentata avviene attraverso il confronto approssimato delle caratteristiche 3a, 3c, 3d. La concavità o la convessità dell'angolo è invece una caratteristica su cui non è ammessa approssimazione. Questo perché la geometria del sistema visivo è la geometria proiettiva che accetta le sproporzioni e le variazioni degli angoli ma non che angolo diventi da concavo a convesso. E' inoltre essenziale l'ordine degli angoli lungo la linea. Riconosciuto l'oggetto si mediano le caratteristiche 3a, 3c, 3d e si sostituisce la memorizzazione. Ogni angolo ha un una sua potenza che viene aumentata o diminuita secondo lo schema classico della rete neurale. In tal modo i contorni si allungano e si affinano. Più è lungo il contorno più sono facili memorizzazione e riconoscimento. Per questo i pupazzi dei bambini hanno forti contrasti di colore. Le contratte di ogni figura sono memorizzate in una cella di memoria che ha un indirizzo.

Circa il modulo 5, non ho sviluppato studi di speech recognition e non conosco la neurofisiologia del sistema uditivo. Ho fatto l'ipotesi che il NOME sia un'inequivocabile serie di caratteri alfanumerici, che viene memorizzato in una cella di memoria, che ha un indirizzo.
La cella di memoria del nome può memorizzare l'indirizzo della cella di memoria delle contratte e viceversa, permettendo così il richiamo e il riconoscimento.

vorrei notare che le contratte permettono il riconoscimento della figura, non la riproduzione di essa, attraverso un disegno. Nella memorizzazione vengono, in larga misura, perse le proporzioni dei particolari di un singolo contorno. Per esempio in un profilo viene persa la proporzione l'ampiezza del mento, del naso e della fronte. Non viene invece perso il loro ordine, perché sono una stessa linea. Tuttavia se i contorni formano linee diverse, per esempio, nel profilo, la bocca e l'orecchio allora anche la loro posizione reciproca è perduta. Bocca e orecchio sono memorizzati come oggetti diversi della stessa percezione. I loro rapporti, anche qui molto mal definiti, formano un ulteriore capitolo della visione: la visione sintattica.

Vorrei rimarcare ***che in questo articolo è solo esposta una piccola parte del mio lavoro. Giusto degli addentellati per mostrare le differenze fra il metodo di Marr e il mio.***

Sito da cui è stato copiato il disegno dell'occhio a inizio pagina

---

Home