RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Introduzione

Il linguaggio dei segni viene trasmesso visivamente, utilizzando una combinazione di mezzi espressivi basati e non basati sull'utilizzo delle mani (forma della mano, postura della mano, posizione della mano, movimento della mano; posizione della testa e postura del corpo, sguardo, mimica facciale).

Alcuni segni possono essere riconosciuti utilizzando solo parametri basati sull'utilizzo delle mani, mentre altri segni possono rimanere ambigui a meno che non vengano rese disponibili informazioni addizionali non basate sull'utilizzo delle mani.

I segni vengono effettuati in uno spazio 3D vicino al busto e alla testa; tale spazio viene chiamato *spazio dei segni*.

La grammatica del linguaggio dei segni è molto differente da quella del linguaggio parlato. La struttura di una frase nel linguaggio parlato è lineare, nel senso che una parola è seguita da un'altra parola. Nel linguaggio dei segni, esiste una struttura con configurazione temporale e spaziale. La configurazione di una frase del linguaggio dei segni è caratterizzata da una notevole quantità di informazione sul tempo, sui luoghi, sulle persone.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Introduzione

Il metodo di acquisizione dei dati riguardanti il linguaggio dei segni definisce la qualità dell'interfaccia utente e rappresenta la caratteristica primaria per la classificazione. Esistono tecniche invasive e non invasive. Le tecniche invasive sono affidabili e robuste: i data gloves (guanto-sensore) misurano la flessione delle articolazioni delle dita, i markers magnetici o ottici posizionati sul volto e sulle mani facilitano la determinazione delle configurazioni relative alle mani e alle espressioni facciali. In alcuni ambiti, il costo di tali attrezzature potrebbe però rappresentare un problema.

L'usabilità di un sistema di riconoscimento non invasivo viene influenzata in maniera significativa dalla robustezza della fase di elaborazione delle immagini, cioè dalla sua abilità nella gestione di sfondi non omogenei, dinamici e generalmente non controllati e di condizioni di illuminazione non ottimali.

Un problema importante di molti sistemi di riconoscimento del linguaggio dei segni è la loro customizzazione per la singola persona, cioè ogni utente deve «allenare» (training) il sistema su se stesso prima di poterlo effettivamente utilizzare. Un sistema non dipendente dalla singola persona richiede una adeguata normalizzazione delle features nelle prime fasi della procedura di elaborazione in modo da

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Introduzione

eliminare la dipendenza delle features dalla posizione del segnante nell'immagine, dalla distanza del segnante dalla camera, e dalla risoluzione della camera.

Author, Year	Features	Interface	Vocabulary	Language Level	Recognition Rate in %
Vamplew 1996 [24]	manual	data glove	52	word	94
Holden 2001 [10]	manual	optical markers	22	word	95,5
Yang 2002 [28]	manual	video	40	word	98,1
Murakami 1991 [15]	manual	data glove	10	sentence	96
Liang 1997 [14]	manual	data glove	250	sentence	89,4
Fang 2002 [8]	manual	data glove	203	sentence	92,1
Starner 1998 [20]	manual	video	40	sentence	97,8
Vogler 1999 [25]	manual	video	22	sentence	91,8
Parashar 2003 [17]	manual and facial	video	39	sentence	92

Alcuni esempi di caratteristiche di classificatori utilizzati per il riconoscimento del linguaggio dei segni (customizzazione per la singola persona).

Osservando la tabella riportata, notiamo che negli esempi che utilizzano i data gloves c'è un vocabolario ampio. È comunque difficile confrontare i differenti esempi poiché non si riferiscono ad un unico caso di studio.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Introduzione

Author, Year	Features	Interface	Vocabulary	Language Level	Recognition Rate in %
Vamplew 1996 [24]	manual	data glove	52	word	94
Holden 2001 [10]	manual	optical markers	22	word	95,5
Yang 2002 [28]	manual	video	40	word	98,1
Murakami 1991 [15]	manual	data glove	10	sentence	96
Liang 1997 [14]	manual	data glove	250	sentence	89,4
Fang 2002 [8]	manual	data glove	203	sentence	92,1
Starner 1998 [20]	manual	video	40	sentence	97,8
Vogler 1999 [25]	manual	video	22	sentence	91,8
Parashar 2003 [17]	manual and facial	video	39	sentence	92

Alcuni esempi di caratteristiche di classificatori utilizzati per il riconoscimento del linguaggio dei segni (customizzazione per la singola persona).

Inoltre il rate riportato nell'ultima colonna si riferisce allo scenario utilizzato; non vengono riportate informazioni riguardanti la robustezza in applicazioni caratterizzate da scenari del mondo reale.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Introduzione

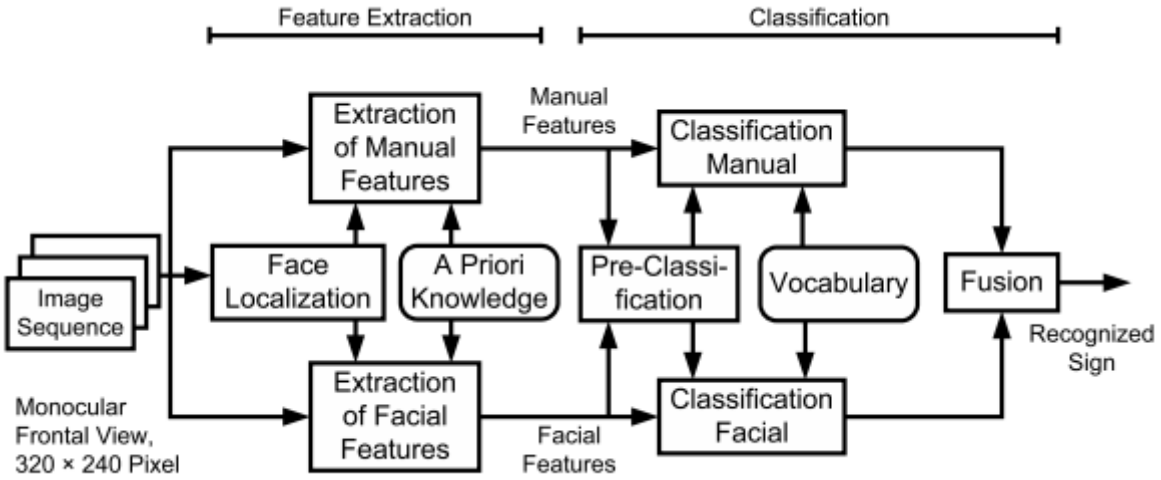
Al fine di risolvere il problema del riconoscimento del linguaggio dei segni, può essere sviluppato ad esempio un sistema mobile basato su un dispositivo portatile dotato di webcam; tale dispositivo legge mediante visione la mimica facciale e i gesti di una persona, fornendo una traduzione in linguaggio parlato. Tale sistema può essere inteso come un interprete.

Sistema di riconoscimento del linguaggio dei segni basato su PC portatile e webcam.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento non invasivo basato su tecniche di visione

Al fine di risolvere il problema del riconoscimento non invasivo del linguaggio dei segni, possono essere sviluppati classificatori basati su video.



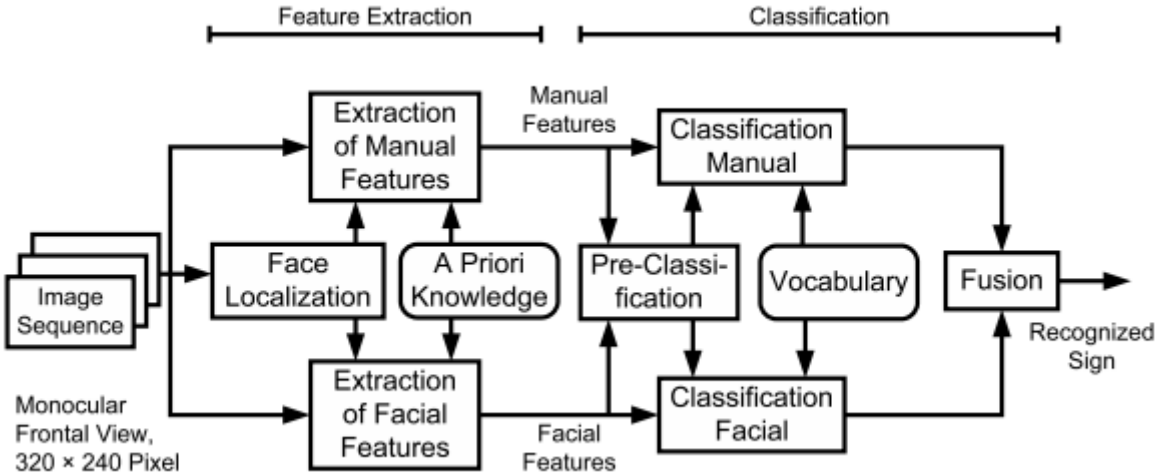
Schema di un sistema per il riconoscimento del linguaggio dei segni.

Nello schema riportato, è evidente la presenza della fase di estrazione delle features e della fase di classificazione.

La sequenza delle immagini di input viene inviata a due fasi di elaborazione parallele le quali estraggono le features relative alle mani e quelle relative al volto. Tali procedure di estrazione utilizzano conoscenza a priori sul processo dei

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento non invasivo basato su tecniche di visione



Schema di un sistema per il riconoscimento del linguaggio dei segni.

segni. Prima della fase finale di classificazione, un opportuno modulo svolge una fase di pre-classificazione al fine di restringere il vocabolario da considerare. Tale fase è utile per ridurre il tempo di esecuzione.

Le features relative alle mani e quelle relative al volto vengono poi classificate separatamente; i risultati ottenuti vengono poi uniti per ottenere un unico risultato finale di classificazione.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

Il riconoscimento del linguaggio dei segni in scenari reali ha esigenze più elevate riguardo all'estrazione delle features e agli algoritmi rispetto al riconoscimento dei gesti della mano in ambienti controllati (problematica analizzata precedentemente).

In primo luogo, il linguaggio dei segni fornisce un framework che restringe la libertà dello sviluppatore nella scelta del vocabolario. Rispetto ad una scelta arbitraria dei gesti, problematiche comuni nel riconoscimento dei segni sono la sovrapposizione di entrambe le mani e della faccia (immagine *a* riportata nella figura della slide seguente), le ambiguità (immagine *b* riportata nella figura della slide seguente), e le somiglianze tra segni differenti (immagini *c* e *d* riportate nella figura della slide seguente).

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

(a) “digital”

(b) “conference”

(c) “food”

(d) “date”

Difficoltà nel riconoscimento del linguaggio dei segni (British Sign Language).

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

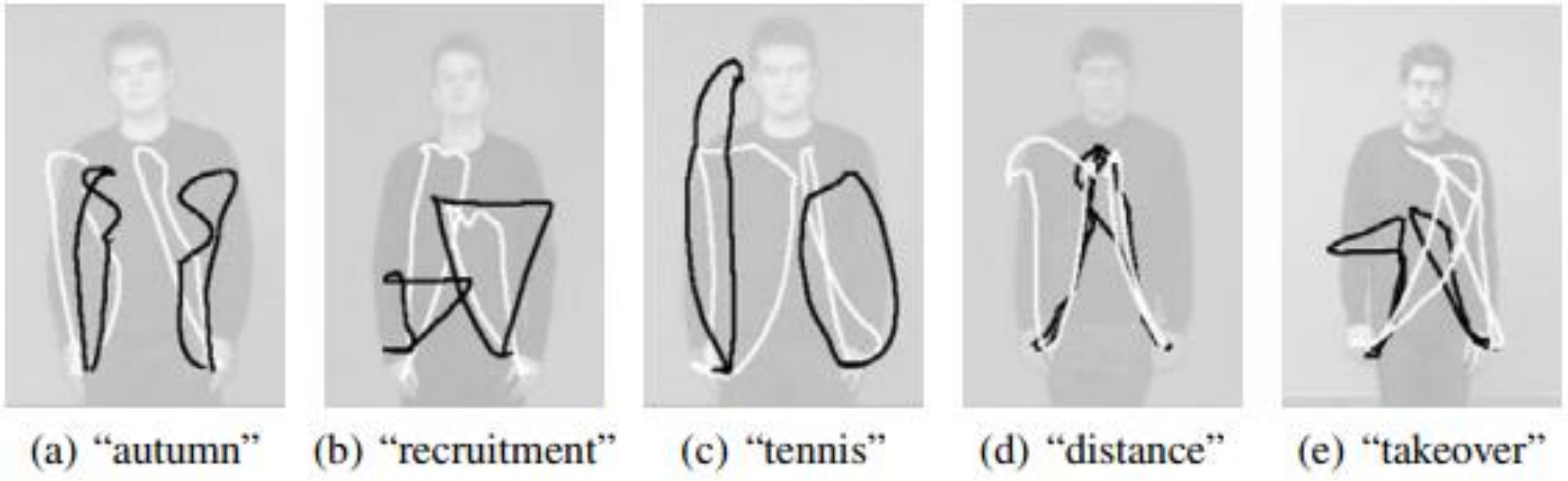
Come nel linguaggio parlato, anche nel linguaggio dei segni esistono i dialetti. Comunque, rispetto alla pronuncia delle parole, non ci sono standard per i segni: le persone potrebbero usare un segno differente per la stessa parola.

Due segni differenti per «actor»: movimento circolare delle mani davanti alla parte superiore del corpo (a) e mano dominante (mano destra in questo esempio) posizionata sulla mano non dominante (b).

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

Inoltre, le differenze nel modo in cui persone differenti eseguono lo stesso segno possono essere notevoli.



Tracciamento di (x_{cog} , y_{cog}) per entrambe le mani considerando due persone differenti (bianco e nero) che eseguono lo stesso segno. Lo sfondo mostra una delle due persone come riferimento.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

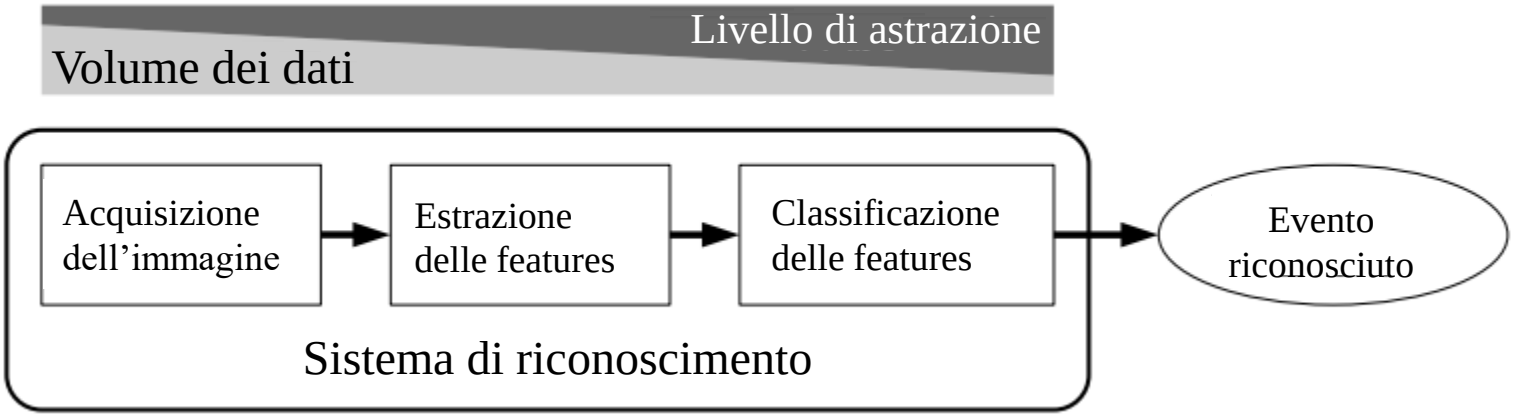
Riconoscimento di segni isolati in scenari reali

Infine, quando si utilizza il colore e/o il movimento per rilevare la mano e la faccia del segnante, sfondi non controllati possono causare numerosi *false alarms* che richiedono un certo tipo di ragionamento per distinguere i target dai distrattori. Tale processo è oneroso dal punto di vista computazionale ed è molto soggetto ad errori perché la conoscenza e l'esperienza richieste, che sono naturali per gli esseri umani, sono difficili da codificare in forma comprensibile per le macchine.

Esempi di sfondi non controllati che possono interferire con il tracciamento di mani e faccia.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

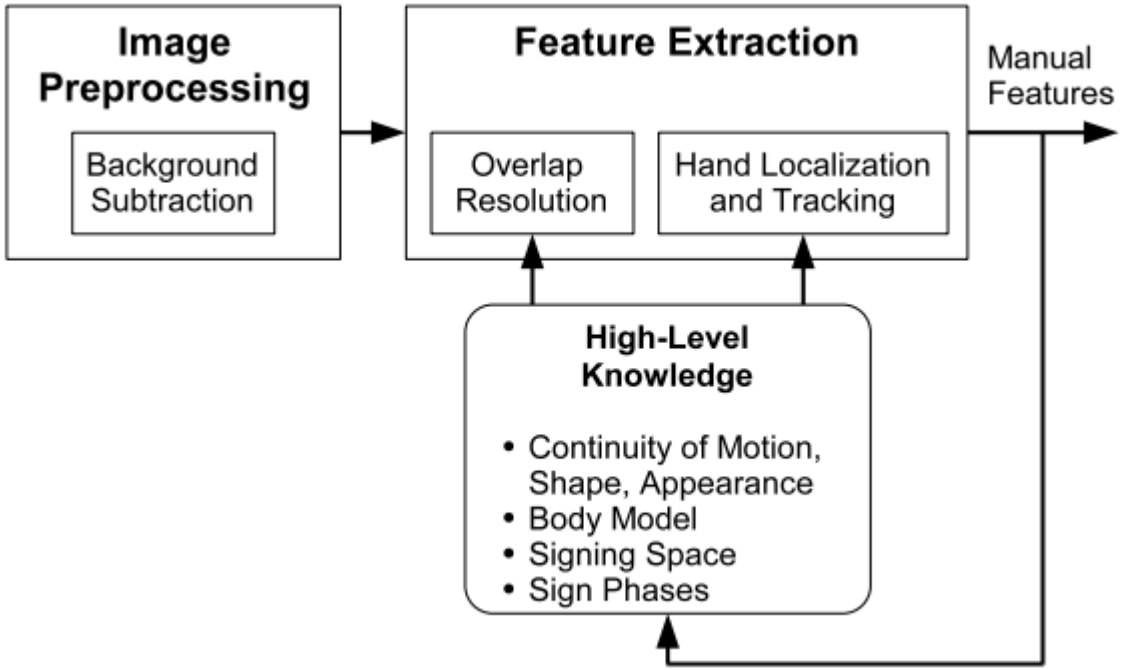


Procedura di elaborazione presente in molti sistemi di riconoscimento di pattern

La procedura di elaborazione per il riconoscimento di segni isolati in scenari reali si basa sullo schema riportato (già analizzato in precedenza). Nel caso del riconoscimento di segni, la fase di estrazione delle features viene estesa, come mostrato nella figura riportata nella slide seguente. Viene aggiunta una fase di preelaborazione delle immagini, nella quale vengono applicati algoritmi per il miglioramento dell'immagine, come ad esempio la modellizzazione dello sfondo. Viene applicata inoltre la conoscenza ad alto livello al fine di risolvere la problematica della sovrapposizione e per supportare il tracciamento e la localizzazione della mano.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

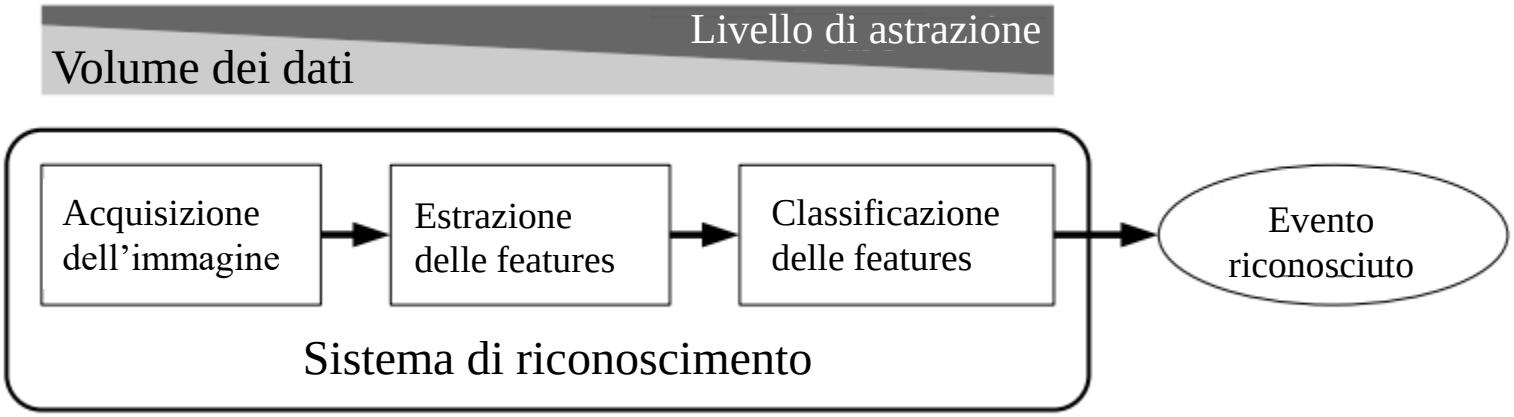
Riconoscimento di segni isolati in scenari reali



Estensione della fase di estrazione delle features con una fase di preelaborazione delle immagini. In tale fase di preelaborazione si applica la conoscenza ad alto livello del processo dei segni al fine di risolvere la problematica della sovrapposizione e per supportare il tracciamento e la localizzazione della mano.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali



Procedura di elaborazione presente in molti sistemi di riconoscimento di pattern

Per quanto riguarda l'acquisizione delle immagini, il passaggio dai gesti al linguaggio dei segni propone le seguenti sfide addizionali:

1) Mentre molti gesti si basano sull'utilizzo di una sola mano, i segni possono essere caratterizzati dall'utilizzo di una o due mani. Quindi il sistema non solo deve essere capace di gestire due oggetti in movimento, ma anche di rilevare se il segno è eseguito con una mano o con due mani (cioè se una mano è inattiva).

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

2) Poiché la posizione delle mani rispetto al volto rappresenta un'informazione rilevante nel linguaggio dei segni, la faccia deve essere inclusa nell'immagine come punto di riferimento. Ciò riduce la risoluzione dell'immagine disponibile per le mani e introduce un ulteriore compito di localizzazione.

3) Le mani possono occludersi l'una con l'altra e/o con la faccia. Per oggetti occlusi, l'estrazione delle features mediante segmentazione a soglia non è possibile; ciò causa una perdita di informazione.

4) I segni possono essere estremamente simili (o anche identici) nelle loro features associate alle mani e possono differire principalmente (o esclusivamente) nelle features non associate alle mani. Ciò rende difficile il riconoscimento automatico basato su features associate alle mani o anche impossibile nel caso di segni caratterizzati da features (associate alle mani) identiche.

5) A causa dei dialetti e della mancanza di standardizzazione, potrebbe non essere possibile mappare una determinata parola in un singolo segno. Come conseguenza, o il vocabolario deve essere ampliato per includere più varianti dei segni interessati, oppure il segnante deve conoscere i segni del vocabolario. Quest'ultimo aspetto è problematico perché i madrelingua possono trovare innaturale il fatto di essere costretti a usare un certo segno quando normalmente ne userebbero uno diverso.

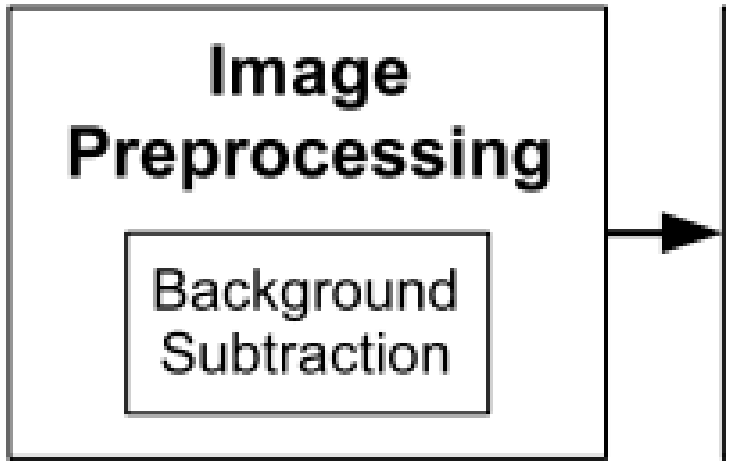
RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

La fase di preelaborazione migliora la qualità dei dati di input al fine di ottenere risultati migliori (in termini di velocità di elaborazione e/o accuratezza) nelle fasi successive.

Potrebbero essere sfruttate informazioni ad alto livello ottenute nelle fasi successive ma ciò comporta i rischi usuali di un feedback loop (anello di retroazione), come ad esempio l'instabilità o il rinforzo degli errori.

Possono quindi essere utilizzati algoritmi a basso livello basati sul concetto di pixel, i quali non hanno tale problema e possono spesso essere utilizzati in molte applicazioni. Un esempio importante è rappresentato dalla modellizzazione dello sfondo (background) dell'immagine.



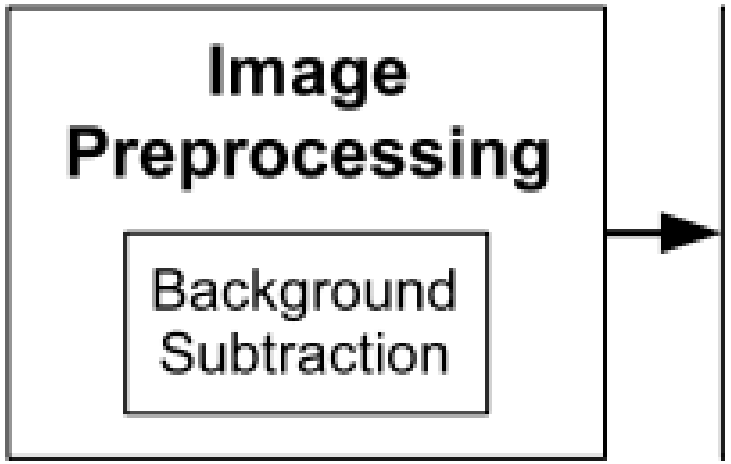
RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

Il rilevamento delle aree di sfondo (ad esempio statico) in scene dinamiche rappresenta una fase importante in molti problemi di riconoscimento di pattern. La sottrazione dello sfondo (background subtraction), cioè l'esclusione di tali aree dalle fasi successive di elaborazione, può ridurre in maniera significativa il disordine nelle immagini di input e diminuire la quantità di dati da processare nelle fasi successive.

Se è disponibile una vista dello sfondo senza oggetti in primo piano, può essere creato un modello statistico in una fase iniziale di calibrazione.

Nel seguito, si assume che tutti i dati di input contengano il segnante. Quindi, se ad esempio il processo dei segni dura 1-10 secondi, il modello dello sfondo deve essere creato senza una fase iniziale di calibrazione dai 25-250 frames disponibili (ipotizzando 25 fps), i quali contengono sia lo sfondo che il primo piano.



RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

Un possibile approccio per la modellizzazione di uno sfondo statico è basato sul concetto di mediana (median background model). Tale approccio permette di creare un modello dello sfondo nella forma di un'immagine RGB $\mathbf{I}_{\text{bg}}(x, y)$. Per ogni pixel *(x,y)*, viene calcolata la mediana del colore su tutti i frames:

$$\mathbf{I}_{\text{bg}}(x, y) = \text{median}\{\mathbf{I}(x, y, t) | 1 \leq t \leq T\}$$

dove la mediana di un insieme di vettori $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ $(n = T)$ è l'elemento per il quale la somma delle distanze Euclidee dagli altri elementi è minima:

trasposizione

$$\text{median}\, V = \underset{\mathbf{v} \in V}{\text{argmin}} \sum_{i=1}^{n} |\mathbf{v}_i - \mathbf{v}|$$

$$\mathbf{v}_i = (r_i \;\; g_i \;\; b_i)' \qquad \mathbf{v} = (r_{\text{bg}} \;\; g_{\text{bg}} \;\; b_{\text{bg}})'$$

$$|\mathbf{v}_i - \mathbf{v}| = \sqrt{(r_i - r_{\text{bg}})^2 + (g_i - g_{\text{bg}})^2 + (b_i - b_{\text{bg}})^2}$$

Spesso l'espressione precedente viene approssimata tramite la mediana divisa per canale (channel-wise median)

$$\mathbf{I}_{\text{bg}}(x, y) = \begin{pmatrix} \text{median}\{r(x, y, t) | 1 \leq t \leq T\} \\ \text{median}\{g(x, y, t) | 1 \leq t \leq T\} \\ \text{median}\{b(x, y, t) | 1 \leq t \leq T\} \end{pmatrix}$$

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

Il modello ricavato mediante l'approccio riportato nella slide precedente ha la proprietà di non richiedere alcun parametro; esso è quindi molto robusto. L'unico requisito è che lo sfondo dell'immagine debba essere visibile nel pixel considerato in più del 50% dei frames di input; tale ipotesi è comunque ragionevole in molti scenari. Un piccolo svantaggio è che tutti i frames di input devono essere memorizzati al fine di eseguire la procedura.

Al fine di applicare il modello dello sfondo, cioè per classificare se un dato pixel appartiene allo sfondo o al primo piano, è necessario definire una metrica opportuna Δ per quantificare la differenza tra un vettore $(r_{\text{bg}}\ g_{\text{bg}}\ b_{\text{bg}})^T$ (colore di sfondo) e un vettore associato a un dato colore $(r\ g\ b)^T$:

trasposizione

$$\Delta\left((r\ g\ b)^T, (r_{\text{bg}}\ g_{\text{bg}}\ b_{\text{bg}})^T\right) \geq 0$$

Calcolando Δ per ogni pixel di un'immagine di input $\mathbf{I}(x, y, t)$ e confrontandola con una soglia di sensitività del movimento Θ_{motion}, si ottiene una maschera $\mathbf{I}_{\text{fg,mask}}$ associata al primo piano (foreground):

$$\mathbf{I}_{\text{fg,mask}}(x, y, t) = \begin{cases} 1 & \text{if} \quad \Delta\left(\mathbf{I}(x, y, t), \mathbf{I}_{\text{bg}}(x, y)\right) \geq \Theta_{\text{motion}} \\ 0 & \text{otherwise} \end{cases}$$

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

$$\mathbf{I}_{\mathrm{fg,mask}}(x,y,t) = \begin{cases} 1 & \text{if} \quad \Delta\left(\mathbf{I}(x,y,t), \mathbf{I}_{\mathrm{bg}}(x,y)\right) \geq \Theta_{\mathrm{motion}} \\ 0 & \text{otherwise} \end{cases}$$

Θ_{motion} viene scelta sufficientemente grande in modo che il rumore della camera non venga classificato come primo piano.

Una implementazione immediata per la metrica Δ è la distanza Euclidea:

$$\Delta\left((r\ g\ b)^T, (r_{\mathrm{bg}}\ g_{\mathrm{bg}}\ b_{\mathrm{bg}})^T\right) = \sqrt{(r-r_{\mathrm{bg}})^2 + (g-g_{\mathrm{bg}})^2 + (b-b_{\mathrm{bg}})^2}$$

Esistono altri metodi per calcolare tale metrica, utili ad esempio al fine di ignorare le ombre.

Mentre il calcolo del median background model non richiede alcun parametro, la sua applicazione prevede l'introduzione del parametro Θ_{motion} e di eventuali altri parametri utilizzati per definizioni alternative della metrica Δ .

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

Quattro frames di esempio del segno «sbucciare». Il video di input è costituito da 54 frames (tutti con almeno una persona sullo sfondo).

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

(a) Background Model $\mathbf{I}_{\mathrm{bg}}$

(b) Example Frame

(c) Foreground Mask $\mathbf{I}_{\mathrm{fg,mask}}$

Modello dello sfondo (background model) calcolato per l'input mostrato nella slide precedente (a), frame di esempio (b) e maschera risultante associata al primo piano (c).

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

Molti approcci di modellizzazione dello sfondo (per ottenere background models), compreso quello appena presentato, ipotizzano uno sfondo statico. Comunque, gli sfondi possono essere anche dinamici, specialmente in scenari outdoor. Per esempio, un albero con foglie verdi che si muove di fronte a una parete bianca determina l'alternanza di verde e bianco in ogni pixel di sfondo.
Negli approcci che non implementano una modellizzazione di sfondi dinamici viene calcolato un singolo colore di sfondo, come precedentemente mostrato da

$$\mathbf{I}_{\mathrm{bg}}(x,y) = \mathrm{median}\{\mathbf{I}(x,y,t) | 1 \leq t \leq T\}$$

Negli approcci che implementano una modellizzazione di sfondi dinamici, lo sfondo e il primo piano sono modellizzati insieme per ogni pixel come una combinazione di distribuzioni Gaussiane, di solito nello spazio RGB. I parametri di tali distribuzioni vengono aggiornati online ad ogni nuovo frame. Vengono poi utilizzate regole per la classificazione di tali distribuzioni (sfondo o primo piano). Rispetto a metodi più semplici, tali approcci sono più onerosi dal punto di vista computazionale poiché richiedono molti frames di input per calcolare un modello accurato (dato l'alto numero dei parametri da stimare).

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

I sistemi che devono operare in scenari reali richiedono approcci sofisticati di estrazione delle features.

Piuttosto che localizzare una singola mano, occorre segmentare entrambe le mani e la faccia (che deve rappresentare un punto di riferimento). Poiché un segno può coinvolgere una o entrambe le mani, è impossibile conoscere in anticipo se la mano non dominante rimarrà inattiva o se essa si muoverà con la mano dominante.

Inoltre, la risoluzione disponibile per ogni oggetto target è significativamente più piccola rispetto alla maggior parte delle applicazioni di riconoscimento dei gesti. Infatti, nelle applicazioni di riconoscimento dei gesti, la mano tipicamente riempie l'immagine complessiva.

Se si applica solo la sottrazione dello sfondo, non ci si può aspettare di isolare oggetti in primo piano senza errori (primo piano classificato come sfondo) o *false alarms* (sfondo classificato come primo piano). Inoltre, la faccia (che è statica nella maggior parte dei casi) deve essere localizzata prima che venga applicata la sottrazione dello sfondo.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

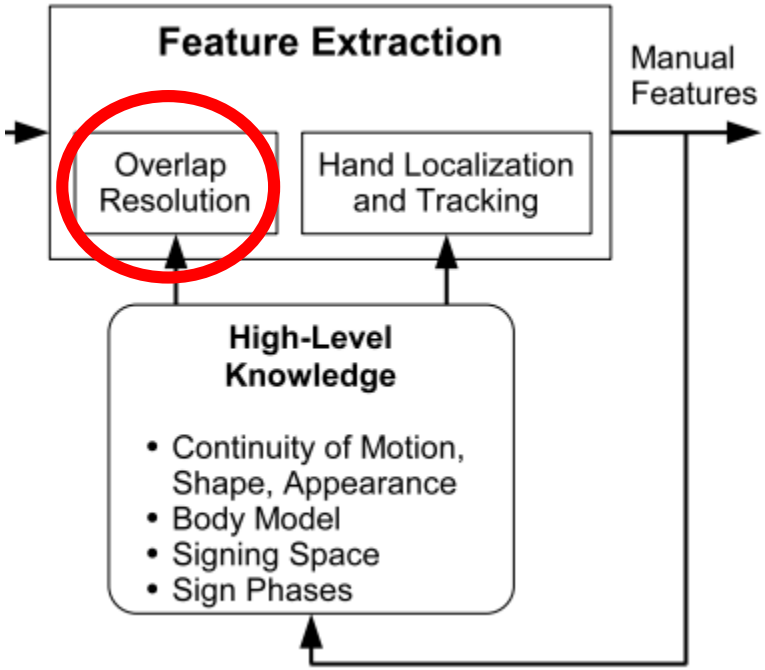
Riconoscimento di segni isolati in scenari reali

Poiché l'aspetto (appearance) estremamente variabile delle mani impedisce l'uso di indicazioni sulla forma e sulla superficie, è comune l'utilizzo del solo colore per la localizzazione della mano. Ciò può portare alla restrizione che il segnante debba indossare abiti a maniche lunghe e non colorati per facilitare la separazione della mano dal braccio per mezzo del colore.

Utilizzando generici modelli del colore, si può ottenere l'indipendenza dall'utente e dall'illuminazione al costo di avere un alto numero di *false alarms*. Questo metodo quindi richiede algoritmi di ragionamento ad alto livello al fine di gestire tali ambiguità. In alternativa, si può scegliere di calibrare esplicitamente o automaticamente il sistema in base all'utente e all'illuminazione corrente. Nel seguito verrà descritto il primo approccio.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

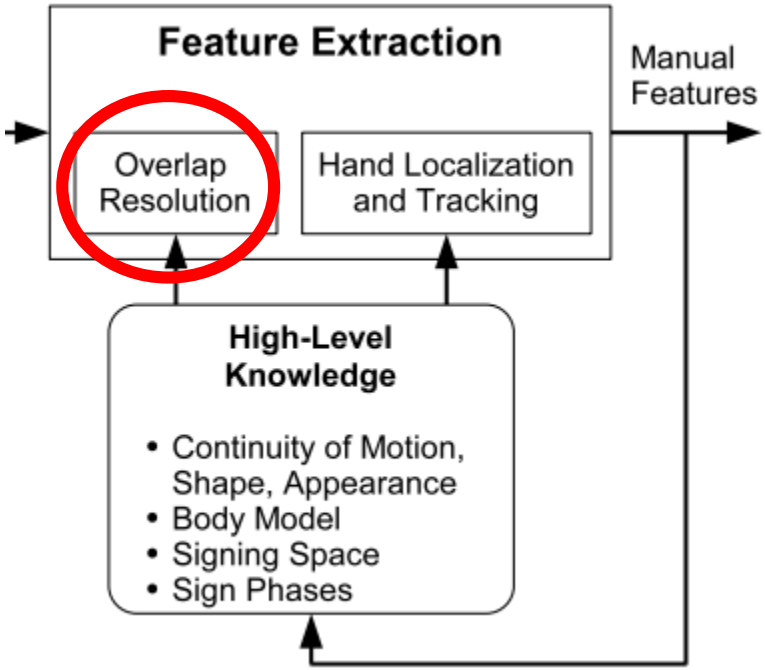


In numerosi segni, entrambe le mani si sovrappongono l'una all'altra e/o al viso. Quando due o più oggetti si sovrappongono nell'immagine, la segmentazione del colore (pelle) produce un unico blocco per tutti gli oggetti, rendendo impossibile l'estrazione diretta di features significative.

Il basso contrasto, la bassa risoluzione e l'aspetto variabile delle mani di solito non permettono una separazione degli oggetti sovrapposti mediante segmentazione basata sui bordi. La maggior parte delle features geometriche disponibili per oggetti non sovrapposti non può quindi essere calcolata per oggetti sovrapposti. Comunque, se l'aspetto di una mano è sufficientemente costante su diversi frames, può essere applicato il metodo del template matching.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

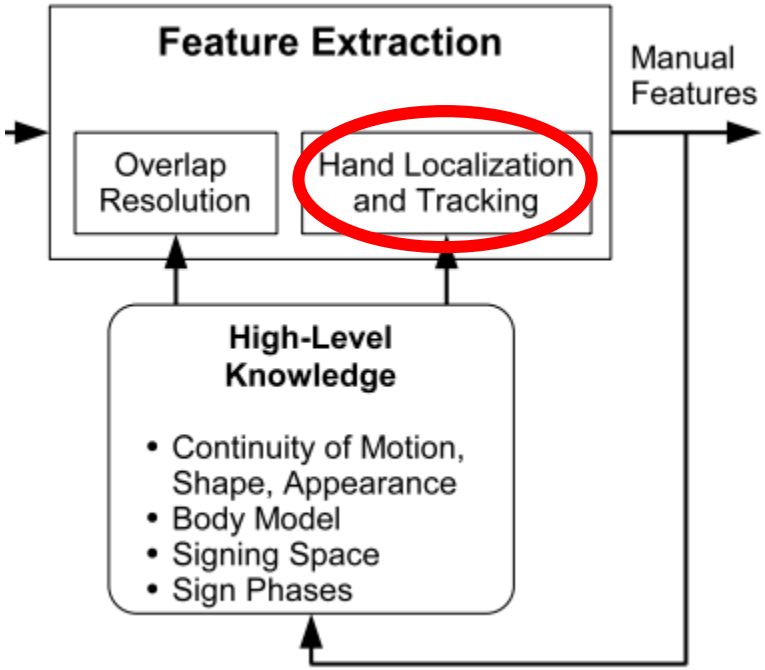
Riconoscimento di segni isolati in scenari reali



Nel template matching, utilizzando come template l'ultima vista non sovrapposta di ogni oggetto sovrapposto, durante la sovrapposizione possono essere calcolate almeno le features associate alla posizione. Tali features contengono una notevole quantità di informazioni. L'accuratezza di tale metodo decresce con la crescita dell'età del template. Fortunatamente, la stessa tecnica può essere utilizzata due volte per un singolo periodo di sovrapposizione. La seconda volta viene utilizzata iniziando con la prima vista non sovrapposta dopo la cessazione della sovrapposizione e proseguendo temporalmente all'indietro. Tale accorgimento riduce la massima età del template e aumenta considerevolmente la precisione.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali



L'esecuzione della localizzazione della mano in ogni frame non è sufficientemente affidabile in scenari complessi. Deve essere considerata anche la relazione tra frames temporalmente adiacenti al fine di migliorare le prestazioni del sistema di riconoscimento. La localizzazione è quindi sostituita dal tracciamento (tracking), il quale utilizza informazioni associate ai frames precedenti come base per trovare la mano nel frame corrente.

Un problema comune nel tracking è la gestione di situazioni ambigue dove è plausibile più di un'interpretazione (o ipotesi) a partire dalla segmentazione ottenuta (si veda l'esempio riportato nelle due slide seguenti).

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

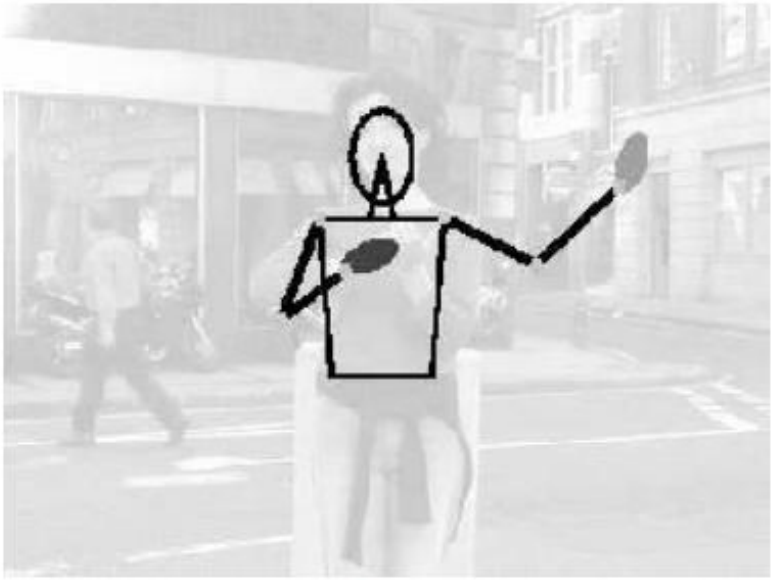
Riconoscimento di segni isolati in scenari reali

Example Frame

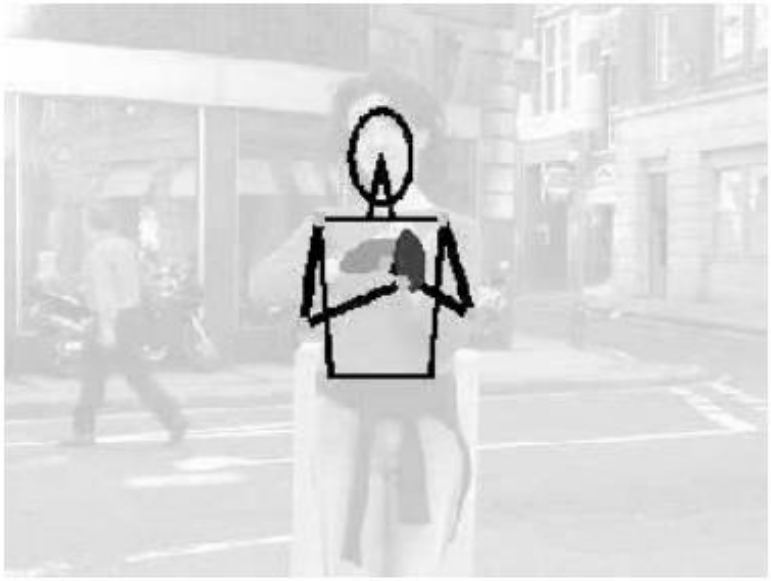
Segmentazione basata sul colore (pelle)

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

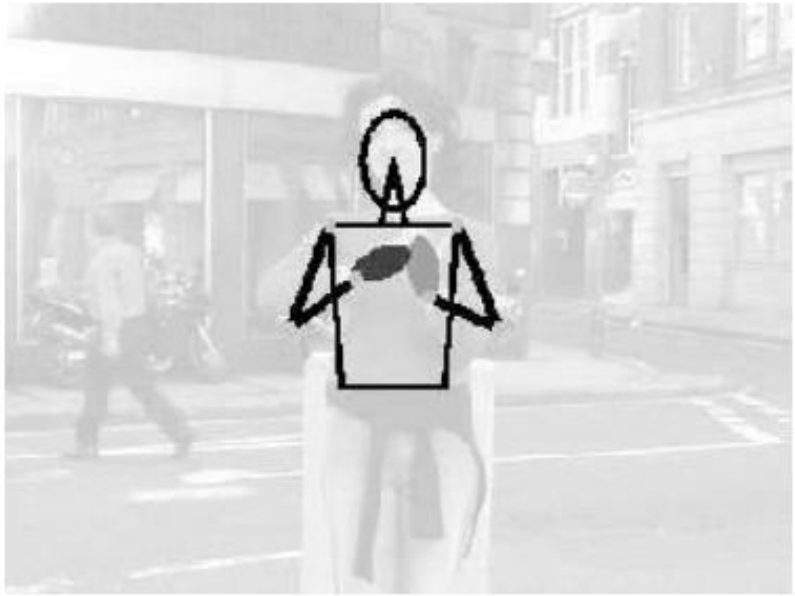
ni i



(b)



(c)

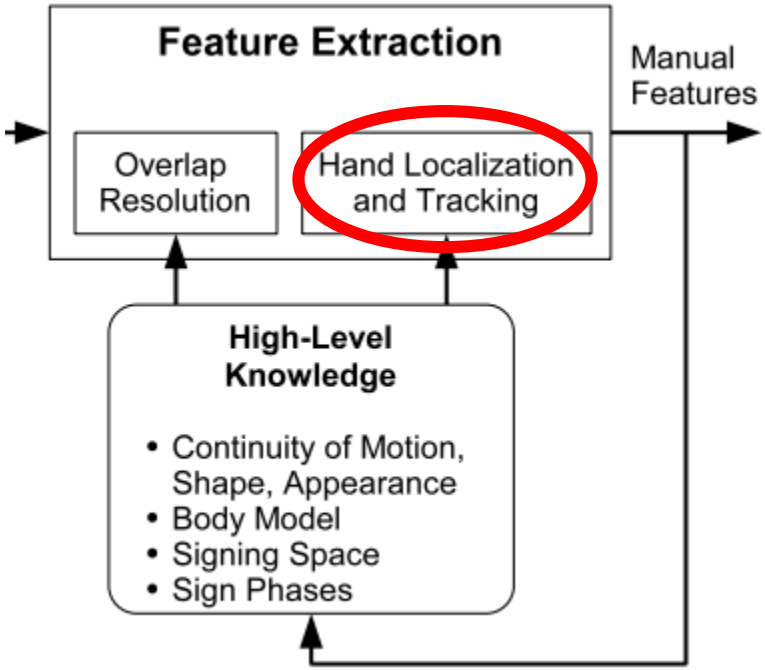


(d)

Sottoinsieme di possibili interpretazioni (b,c,d). La d è quella corretta.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali



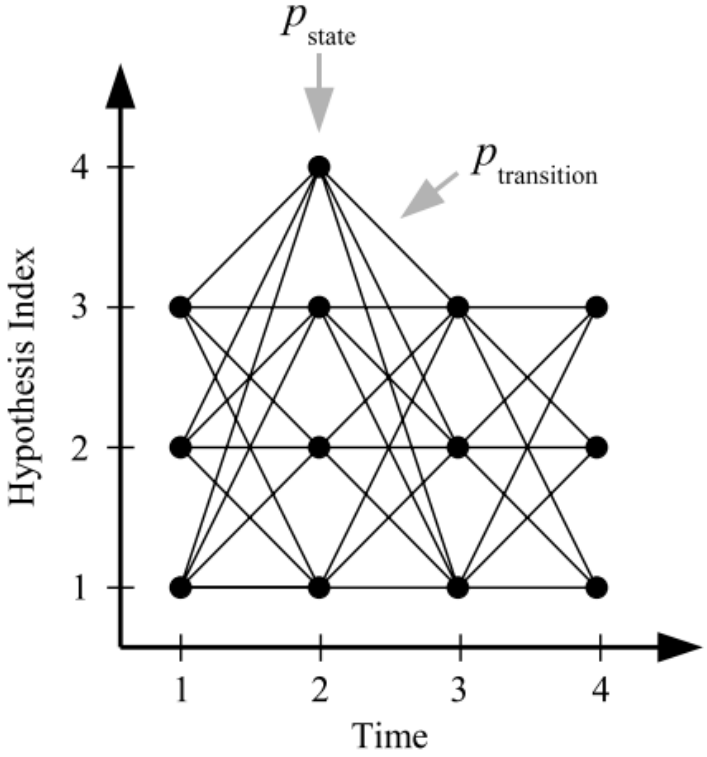
Approcci semplici pesano tutte le ipotesi e scelgono la più probabile in ogni frame sulla base delle informazioni raccolte nei frames precedenti, come, ad esempio, una predizione della posizione. Tutte le altre ipotesi vengono scartate.

Tale concetto è soggetto ad errori poiché situazioni ambigue che non possono essere interpretate in modo affidabile si presentano spesso nel linguaggio dei segni. La robustezza può essere aumentata significativamente tenendo in considerazione non solo i frames precedenti ma anche quelli successivi prima di scartare ogni ipotesi. Si vuole quindi ritardare la decisione finale sulla posizione delle mani in ogni frame fino a quando non siano stati analizzati tutti i dati disponibili e non sia disponibile la massima quantità di informazione.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

Tali considerazioni portano all'approccio multiple hypotheses tracking (MHT). In una prima elaborazione dei dati di input, per ogni frame vengono create tutte le ipotesi plausibili. Sono ammesse transizioni da ogni ipotesi associata al tempo *t* a tutte le ipotesi associate al tempo *t+1*, ottenendo uno spazio di stato (si veda la figura).



Spazio delle ipotesi e probabilità per stati e transizioni

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

Il numero totale di percorsi associati a tale spazio di stato è dato da

$$\prod_t H(t)$$

dove *H(t)* indica il numero di ipotesi al tempo *t*. Ammesso che la segmentazione basata sul colore (pelle) rilevi entrambe le mani e la faccia in ogni frame, uno di tali percorsi rappresenta il corretto risultato di tracking. Per trovare tale percorso (o il più vicino possibile ad esso), vengono calcolate le probabilità che indicano la plausibilità di ogni configurazione ipotizzata (p_{state}) e la plausibilità di ogni transizione ($p_{\text{transition}}$).

Il calcolo di p_{state} e $p_{\text{transition}}$ è basata su conoscenza ad alto livello, codificata mediante un insieme di regole o appresa durante una fase di training. Una base generale è fornita dai seguenti aspetti:

- La struttura fisica del corpo del segnante può essere dedotta dalla posizione e dalla dimensione della faccia e delle mani. Le configurazioni che sono anatomicamente improbabili o che non si verificano nel linguaggio dei segni riducono p_{state} .

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

- Le tre fasi di un segno (preparazione, esecuzione, ritorno), in connessione con la lateralità manuale del segnante (che deve essere nota al fine di interpretare correttamente il *feature vector* risultante), dovrebbero essere rispecchiate anche nel calcolo di p_{state} .
- Se si ipotizzano 25 fps, la forma della mano cambia poco tra frames successivi (anche in caso di movimenti rapidi). Se non si verificano sovrapposizioni, la forma al tempo *t* può essere quindi utilizzata come stima per il tempo *t+1*. Con l'aumentare della differenza della forma corrente dalla forma attesa, $p_{\text{transition}}$ viene ridotta. Improvvisi cambi di forma dovuti alla sovrapposizione richiedono una gestione differente.
- Similmente, la posizione della mano cambia poco da un frame all'altro (ipotizzando ad esempio 25 fps); quindi le coordinate al tempo *t* possono essere utilizzate come predizione al tempo *t+1*. Il filtro di Kalman può aumentare l'accuratezza della predizione mediante estrapolazione sulla base di tutte le misure passate.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

- Il tener traccia della media o della mediana del colore della mano può evitare di confondere la mano con distrattori vicini di dimensione simile ma colore differente. Tale criterio influenza $p_{\text{transition}}$.

Al fine di cercare lo spazio delle ipotesi, può essere applicato l'algoritmo di Viterbi in combinazione con l'eliminazione dei percorsi improbabili ad ogni passo.

L'approccio MHT assicura l'utilizzo di tutte le informazioni disponibili prima di determinare il risultato finale del tracking. La fase di tracking può quindi sfruttare, al tempo *t*, l'informazione che diventa disponibile solo al tempo $t_1 > t$. Gli errori vengono corretti in modo retroattivo non appena si manifestano.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

Dopo l'estrazione delle features, le features devono essere normalizzate. Ipotizzando che la dipendenza dalla risoluzione sia stata già eliminata utilizzando i metodi descritti per il riconoscimento dei gesti statici, l'area *a* ancora dipende dalla distanza del segnante dalla camera. Le coordinate della mano (x_{cog} , y_{cog}) dipendono anche dalla posizione del segnante nell'immagine. Utilizzando la posizione e la dimensione della faccia come un riferimento per la normalizzazione, entrambe le dipendenze possono essere eliminate.

Se per la faccia viene utilizzata una semplice segmentazione a soglia, il collo di solito viene incluso. Le differenti linee del collo possono quindi influenzare l'area e la posizione (COG) rilevate riguardo alla faccia. In questo caso, si può utilizzare il punto più alto della faccia e la sua larghezza per evitare tale problema. Inoltre, se le posizioni delle spalle possono essere rilevate o stimate, esse possono fornire un adeguato riferimento.

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

La classificazione del *feature vector* estratto può essere eseguita ad esempio utilizzando gli *Hidden Markov Models* (*HMMs*) (come per il riconoscimento di gesti dinamici). Inoltre, per la natura del linguaggio dei segni, è consigliabile eseguire alcune fasi di elaborazione aggiuntive. In primo luogo, i frames non significativi iniziali e/o finali dei dati di input possono essere rilevati mediante semplici regole. Si ricordi ad esempio la regola utilizzata in precedenza per determinare se un oggetto in una sequenza di immagini è in movimento o è fermo:

$$\text{IF} \quad \max_{i \neq j} \sqrt{(x_{\text{cog}}(i) - x_{\text{cog}}(j))^2 + (y_{\text{cog}}(i) - y_{\text{cog}}(j))^2)} < \Theta_{\text{motion}} N$$

$$\text{THEN} \quad \text{the hand is idle.}$$

Questa operazione viene eseguita singolarmente per ogni mano. Ritagliare i frames in cui entrambe le mani sono inattive accelera la classificazione e impedisce al classificatore di elaborare dati di input che non contengono informazioni. Ovviamente, questo processo deve essere applicato anche nella fase di training.

Se il segno è caratterizzato dall'utilizzo di una mano, cioè una mano rimane inattiva in tutti i frames, tutte le classi del vocabolario che rappresentano segni

RICONOSCIMENTO DEL LINGUAGGIO DEI SEGNI

Riconoscimento di segni isolati in scenari reali

caratterizzati dall'utilizzo di due mani possono essere disabilitate (o viceversa). Tale accorgimento riduce ulteriormente il costo computazionale nella fase di classificazione.

Riferimenti Bibliografici

[1] Kraiss, K. -F. (2006). Advanced Man-Machine Interaction: Fundamentals and Implementation. Springer-Verlag Berlin Heidelberg. ISBN-10: 3-540-30618-8