

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

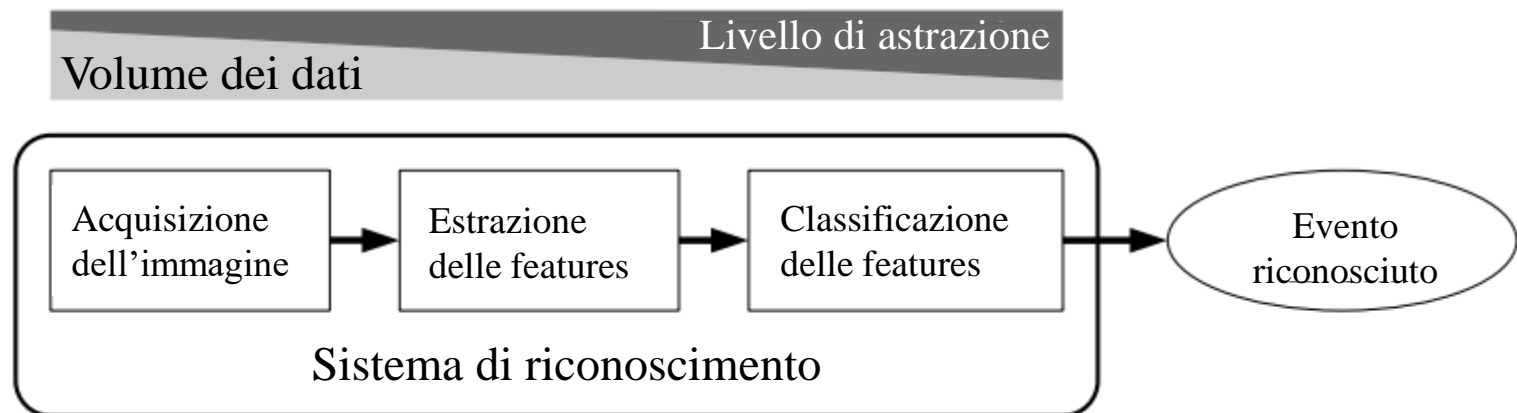
Si vogliono introdurre alcuni concetti di base riguardanti l'acquisizione non invasiva dell'azione umana mediante tecniche di visione.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Introduzione

L'uso dei gesti come mezzo per trasmettere informazioni è una parte importante della comunicazione umana. Poiché i gesti delle mani sono un mezzo comodo e naturale di interagire, le interfacce uomo-macchina possono trarre vantaggio dal loro utilizzo come modalità di input.

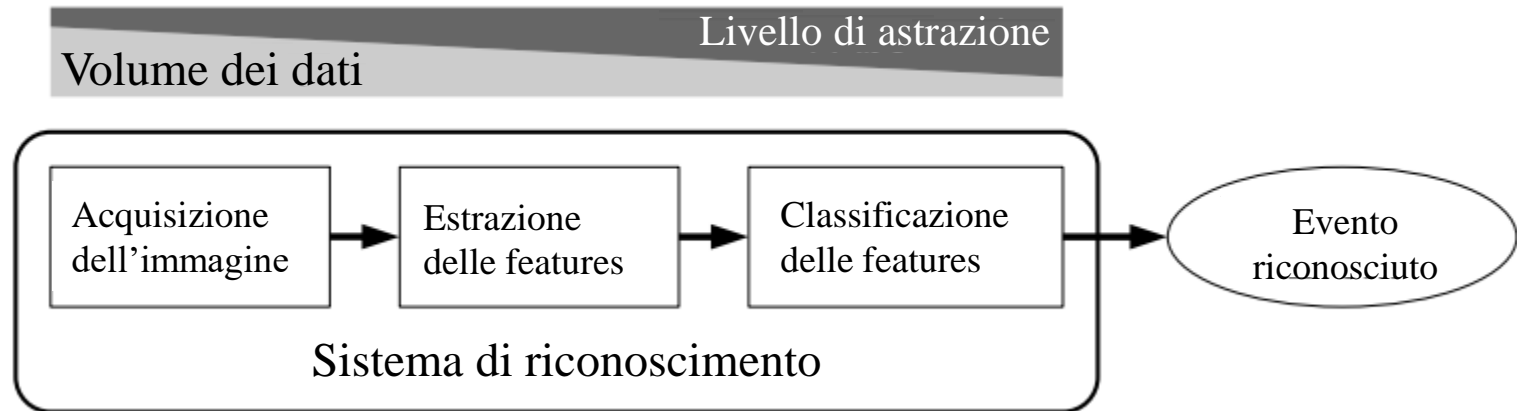
Il controllo dei gesti consente un'interazione silenziosa, remota e senza contatto, per la quale non è necessario individuare pulsanti o tasti.



Procedura di elaborazione presente in molti sistemi di riconoscimento di pattern

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Introduzione



Procedura di elaborazione presente in molti sistemi di riconoscimento di pattern

Molti sistemi di riconoscimento di pattern hanno una procedura di elaborazione comune che può essere divisa in tre fasi successive, ognuna delle quali inoltra il proprio output alla fase seguente.

In primo luogo, viene acquisito un singolo fotogramma (frame), ad esempio da una camera.

Nella fase di estrazione delle features, viene calcolato un numero prestabilito di parametri scalari. Ognuno di tali parametri descrive una singola caratteristica (feature) della scena osservata.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

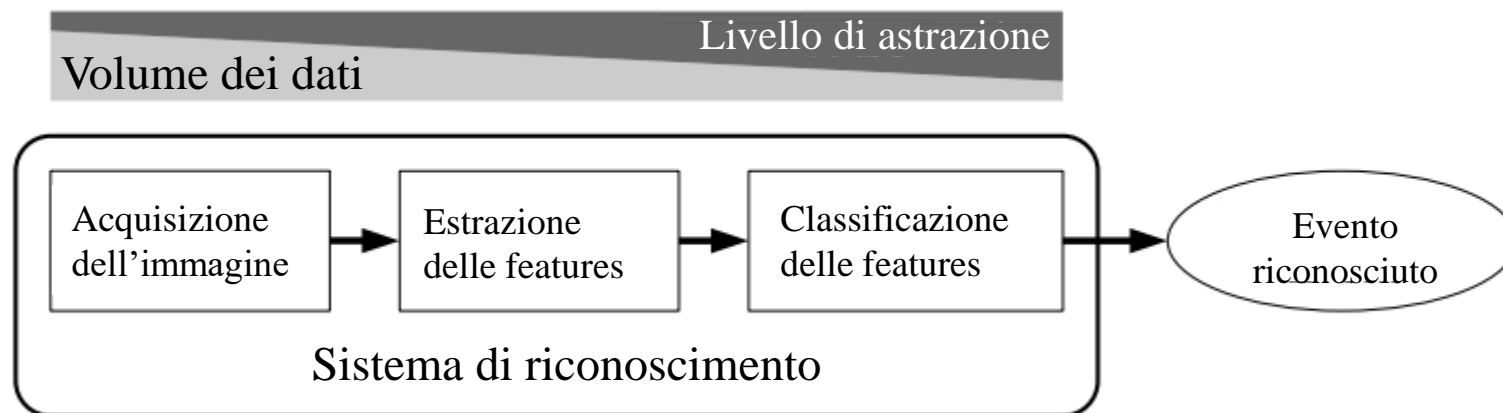
Gesti della mano – Introduzione

Esempi di parametri sono le coordinate di un oggetto (esempio: una mano), la sua dimensione o il suo colore. L'immagine viene quindi definita mediante una rappresentazione numerica, denominata *feature vector*. Questa fase è spesso la più complessa e computazionalmente costosa della procedura di elaborazione. L'accuratezza e l'affidabilità del processo di estrazione delle features possono influire in modo significativo sulle prestazioni complessive del sistema.

Infine, le caratteristiche estratte per un singolo fotogramma (gesto statico) o per una sequenza di fotogrammi (gesto dinamico) vengono classificate ad esempio mediante un insieme di regole o sulla base di una fase di training (addestramento) che precede la fase di applicazione (test). Durante la fase di training, la classificazione corretta delle features è nota; le features vengono memorizzate o apprese in modo da consentire successivamente di classificare le features calcolate nella fase di applicazione (test), anche in presenza di rumore e piccole variazioni che inevitabilmente si verificano nel mondo reale.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Introduzione



Procedura di elaborazione presente in molti sistemi di riconoscimento di pattern

Analizzando la procedura di elaborazione, il volume di dati da elaborare diminuisce continuamente (da un'immagine di diverse centinaia di kB a un vettore di features di alcuni byte a una descrizione di un singolo evento) mentre il livello di astrazione aumenta di conseguenza (da un insieme di pixels* a un insieme di features a un'interpretazione semantica della scena osservata).

*Un pixel (abbreviazione di "picture element") è il più piccolo elemento (distinto per colore, intensità,...) da cui è composta un'immagine.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Acquisizione dell'immagine e dati di input

Per quanto riguarda l'acquisizione dell'immagine e i dati di input, definiamo tre importanti proprietà:

- Il vocabolario, cioè l'insieme dei gesti che devono essere riconosciuti. La dimensione del vocabolario è di solito identica al numero di risposte possibili del sistema (a meno che più gesti non attivino la stessa risposta).
- Le condizioni di acquisizione dell'immagine in cui opera il sistema. Per semplificare la progettazione del sistema, possono essere imposte restrizioni sullo sfondo dell'immagine, sull'illuminazione, sulla distanza minima e massima del target dalla camera, ...
- Le prestazioni del sistema in termini di latenza (il tempo di elaborazione tra la fine del gesto e la disponibilità del risultato del riconoscimento) e di accuratezza del riconoscimento (la percentuale o la probabilità di risultati di riconoscimento corretti).

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Acquisizione dell'immagine e dati di input

Per quanto riguarda il vocabolario, la sua struttura e le sue proprietà hanno un impatto significativo sull'intera procedura di elaborazione.

Se il vocabolario è noto prima dell'inizio dello sviluppo del sistema di riconoscimento, la progettazione e l'implementazione del sistema sono in genere ottimizzate di conseguenza. Ciò può comportare problemi nel caso in cui il vocabolario debba essere modificato o ampliato.

Nel caso in cui il sistema sia completamente operativo prima che venga decisa la struttura del vocabolario, l'accuratezza del riconoscimento può essere ottimizzata semplicemente evitando i gesti che vengono frequentemente classificati in modo errato.

In pratica, spesso si applica un connubio di entrambi gli approcci.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Acquisizione dell'immagine e dati di input

Il vocabolario influisce sulla complessità del compito di riconoscimento in due modi. Primo, con l'aumentare delle dimensioni del vocabolario aumenta il numero di possibili errori di classificazione, rendendo il riconoscimento sempre più difficile. In secondo luogo, è necessario considerare il grado di somiglianza tra gli elementi del vocabolario. Questo è complesso da quantificare, e anche una stima qualitativa basata sulla percezione umana non può essere facilmente trasferita a un sistema automatico.

Pertanto, per valutare le prestazioni di un sistema di riconoscimento, la conoscenza della dimensione del vocabolario e dell'accuratezza del riconoscimento potrebbero essere insufficienti. Anche se spesso viene omesso, è necessario conoscere anche l'esatta struttura del vocabolario (o almeno il grado di somiglianza tra i suoi elementi).

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Acquisizione dell'immagine e dati di input

Le condizioni di acquisizione delle immagini in input sono cruciali per la progettazione di qualsiasi sistema di riconoscimento. È quindi importante specificarle il più precisamente ed esaustivamente possibile prima di iniziare l'implementazione.

Tuttavia, a meno che non vengano utilizzati algoritmi e dispositivi hardware già noti e documentati, non è sempre possibile predire come i passi successivi della procedura di elaborazione risponderanno a certe proprietà dei dati di input. Ad esempio, un piccolo cambiamento nella direzione dell'illuminazione può causare un cambiamento nell'aspetto di un oggetto che influisce notevolmente su una determinata feature (anche se questo cambiamento non è immediatamente percepibile dall'occhio umano), portando a una scarsa accuratezza nel riconoscimento. In questo caso, o si devono revisionare le specifiche dei dati di input (non ammettendo cambiamenti nell'illuminazione), oppure è necessario impiegare algoritmi più robusti per l'estrazione e/o la classificazione delle features (in modo da gestire l'effetto di tali variazioni).

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Acquisizione dell'immagine e dati di input

Le problematiche precedentemente spiegate riguardo all'acquisizione delle immagini possono essere evitate cercando di mantenere al minimo qualsiasi tipo di variazione o rumore sulle immagini da acquisire. Ciò porta alla definizione di quelle che sono comunemente denominate «condizioni di acquisizione in laboratorio» («laboratory recording conditions»).

Sebbene queste condizioni possano essere soddisfatte nella fase di training (se la fase di classificazione richiede un training), esse rendono molti sistemi inadeguati per qualsiasi tipo di applicazione pratica.

In contrasto con le «condizioni di acquisizione in laboratorio» («laboratory recording conditions»), una o più «condizioni di acquisizione nel mondo reale» («real world recording conditions») spesso si verificano quando un sistema deve essere messo in uso.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Acquisizione dell'immagine e dati di input

Dominio	Condizione/i
Contenuto dell'immagine	L'immagine contiene solo l'oggetto target, di solito davanti a uno sfondo monocromatico e senza texture. Non sono visibili altri oggetti. L'utente indossa abiti a maniche lunghe e non colorati, in modo da poter separare la mano dal braccio in base al colore.
Illuminazione	Una forte illuminazione diffusa fornisce un'illuminazione uniforme del target, senza ombre o riflessi.
Setup	Il posizionamento della camera, dello sfondo e del target rimane invariato. La distanza del target dalla camera e la sua posizione nell'immagine non cambiano (se il target è in movimento, ciò vale solo per il punto di partenza del movimento).
Camera	L'hardware e i parametri della camera (esempi: risoluzione e dimensione dell'immagine) non vengono mai modificati. Vengono inoltre impostati in modo ottimale per evitare la sovraesposizione e l'alterazione del colore. Nel caso di un target in movimento, la velocità dell'otturatore e la frequenza dei fotogrammi sono sufficientemente elevate per evitare sfocature e discontinuità del movimento. Si utilizza una camera professionale ad alta risoluzione.

Esempio di «laboratory recording conditions» per l'acquisizione dell'immagine (mano)

Dominio	Condizione/i
Contenuto dell'immagine	L'immagine può contenere altri oggetti (distrattori) oltre, o addirittura al posto, dell'oggetto target. Lo sfondo è sconosciuto e può anche essere in movimento (ad esempio: alberi, nuvole). Un target può essere parzialmente o completamente nascosto da distrattori o altri target. L'utente può indossare abiti a maniche corte o color pelle che impediscono la separazione del braccio e della mano in base al colore.
Illuminazione	L'illuminazione potrebbe non essere uniforme, con conseguente presenza di ombre e di un'illuminazione non omogenea all'interno dell'area dell'immagine. Può anche variare nel tempo.
Setup	La posizione del target rispetto alla camera non è fissa. Il target può trovarsi in un punto qualsiasi dell'immagine e la sua dimensione può variare in base alla sua distanza dalla camera.
Camera	L'hardware e/o i parametri della camera possono cambiare da una acquisizione all'altra. È possibile che si verifichino la sovraesposizione e l'alterazione del colore. In caso di target in movimento, basse velocità dell'otturatore possono causare sfocature del movimento; inoltre, una bassa frequenza dei fotogrammi (o un'elevata velocità del target) può causare discontinuità tra fotogrammi successivi. Viene utilizzata una camera economica.

Esempio di «real world recording conditions» per l'acquisizione dell'immagine (mano)

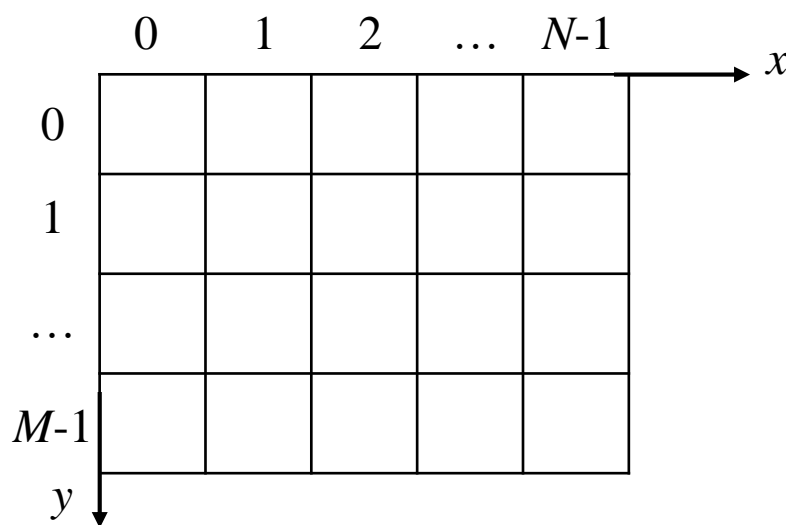
ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Acquisizione dell'immagine e dati di input

La descrizione e l'implementazione di algoritmi di elaborazione delle immagini richiede una opportuna rappresentazione matematica dei vari tipi di immagine. Una rappresentazione comunemente utilizzata per immagini rettangolari composte da M righe e N colonne è una funzione di variabili discrete

$$\mathbf{I}(x, y) \quad \text{con} \quad \begin{aligned} x &\in \{0, 1, \dots, N-1\} \\ y &\in \{0, 1, \dots, M-1\} \end{aligned}$$

Una 2-tupla (x, y) indica le coordinate dei pixels con l'origine $(0, 0)$ nell'angolo in alto a sinistra. Il valore di $\mathbf{I}(x, y)$ descrive una proprietà del pixel corrispondente. Questa proprietà può essere ad esempio il colore del pixel, la sua luminosità o la probabilità che esso rappresenti pelle umana.



In tale rappresentazione, un pixel corrisponde a una cella della griglia

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Acquisizione dell'immagine e dati di input

Il colore può essere rappresentato come una n-tupla di valori scalari, in base al modello scelto. Un modello comunemente utilizzato è il modello RGB (Red, Green, Blue). Il modello RGB utilizza una 3-tupla (r, g, b) specificando le componenti rosso, verde e blu di un colore:

$$\mathbf{I}(x, y) = \begin{pmatrix} r(x, y) \\ g(x, y) \\ b(x, y) \end{pmatrix}$$

Per immagini caratterizzate da colore, $\mathbf{I}(x, y)$ è una funzione vettoriale.

Nelle applicazioni, i colori r , g e b hanno di solito una risoluzione di 8 bit ciascuno:

$$r, g, b \in \{0, 1, \dots, 255\}$$

Utilizzando 8 bit infatti si hanno $2^8=256$ livelli per ognuno dei colori r , g e b .

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Acquisizione dell'immagine e dati di input

Per immagini caratterizzate da colore, $I(x,y)$ è una funzione vettoriale.

Per proprietà differenti dal colore, come ad esempio la luminosità e proprietà associate ad una probabilità, $I(x,y)$ è una funzione scalare e può essere visualizzata come un'immagine in scala di grigi (detta anche immagine dell'intensità).

Utilizzando 8 bit, l'intensità di ogni pixel di un'immagine in scala di grigi è rappresentata da un valore intero compreso tra 0 (nero) e 255 (bianco).



Esempio di immagine in scala di grigi

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Acquisizione dell'immagine e dati di input

Per il riconoscimento dei gesti, una classificazione importante è la distinzione tra primo piano (target) e sfondo. In tal modo si ottiene un'immagine binaria, comunemente chiamata maschera:

$$\mathbf{I}_{\text{mask}}(x, y) \in \{0, 1\}$$

I valori binari vengono di solito visualizzati come bianco e nero.

Una sequenza (esempio: videoclip) di T frame (fotogrammi) può essere descritta utilizzando una variabile tempo t :

$$\mathbf{I}(x, y, t), \quad t = 1, 2, \dots, T$$

Il tempo reale trascorso tra due fotogrammi successivi è l'inverso della frequenza utilizzata per l'acquisizione delle immagini.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Acquisizione dell'immagine e dati di input

Ora specificheremo esempi di dati di input (gesti statici, gesti dinamici) che possono essere elaborati da applicazioni di riconoscimento e le condizioni di acquisizione dei gesti statici e dinamici.

Gesti statici – Il sistema deve riconoscere tre gesti differenti («sinistra», «destra» e «stop»), come mostrato in figura. Inoltre, deve essere rilevato il caso in cui la mano non sia visibile nell'immagine o si trovi in una posizione inattiva e non esegua nessuno di questi gesti. Dato che questo vocabolario è ovviamente molto piccolo e i suoi elementi sono sufficientemente dissimili, ci si può aspettare un'elevata precisione di riconoscimento.



«sinistra»



«destra»



«stop»



posizione di riposo

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Acquisizione dell'immagine e dati di input

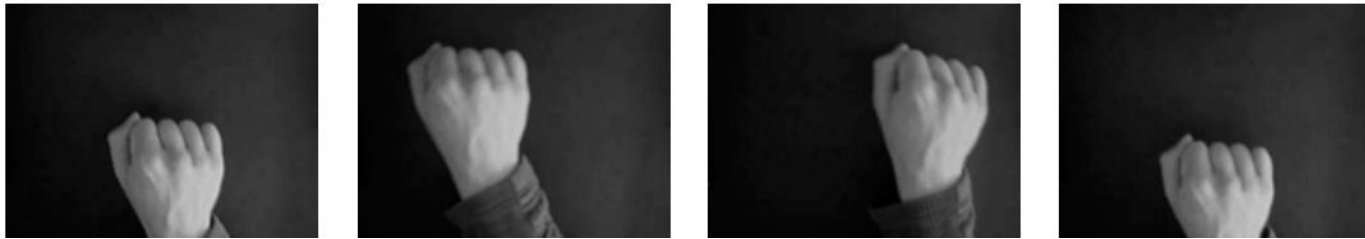
Gesti dinamici – Il vocabolario associato ai gesti dinamici è composto da sei gesti, di cui tre sono esecuzioni a ritroso degli altri tre. Poiché non si considerano le texture, non importa se è il palmo o il dorso della mano ad essere rivolto verso la camera.

Gesto	Descrizione
«senso orario»	Movimento circolare in senso orario della mano chiusa, che inizia e termina alla base di un cerchio immaginario.
«senso antiorario»	Esecuzione a ritroso di «senso orario».
«aperto»	Movimento di apertura della mano piatta mediante rotazione intorno all'asse del braccio di 90° (dita estese e a contatto).
«chiuso»	Esecuzione a ritroso di «aperto».
«presa»	Partendo da tutte le dita estese e non a contatto, si stringe il pugno
«rilascio»	Esecuzione a ritroso di «presa».

Descrizione dei sei gesti dinamici da riconoscere

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

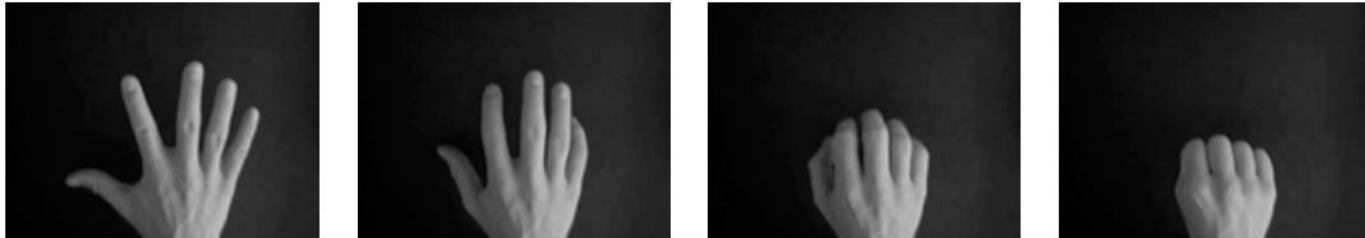
Gesti della mano – Acquisizione dell'immagine e dati di input



(a) «senso orario» (da sinistra a destra) e «senso antiorario» (da destra a sinistra)



(b) «aperto» (da sinistra a destra) e «chiuso» (da destra a sinistra)



(c) «presa» (da sinistra a destra) e «rilascio» (da destra a sinistra)

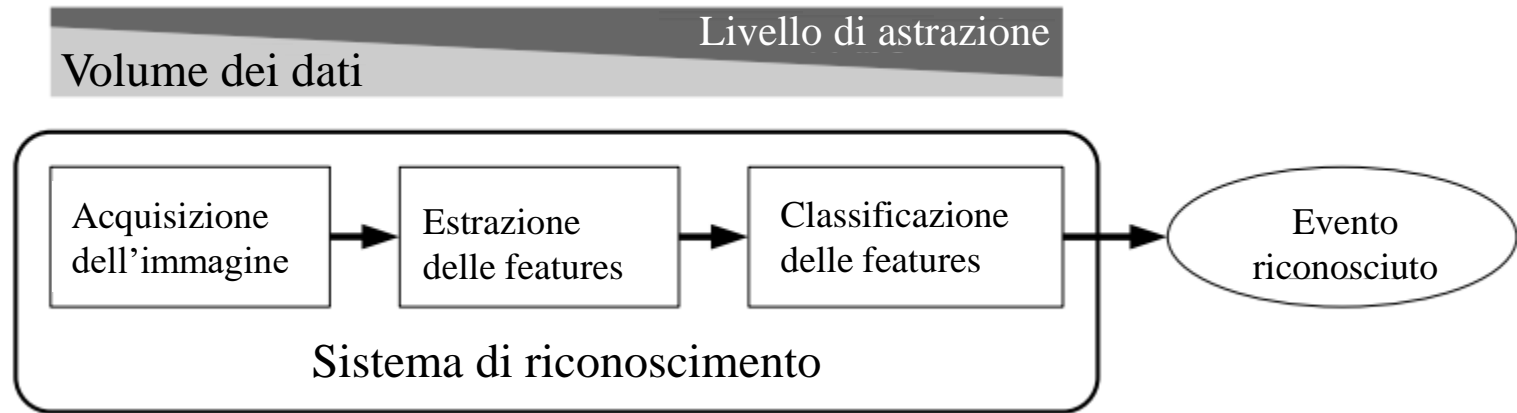
I sei gesti dinamici da riconoscere. In ciascuna delle tre sequenze di immagini, il secondo gesto è un'esecuzione a ritroso del primo, cioè corrisponde alla lettura della sequenza di immagini da destra a sinistra.

Dominio	Condizione/i
Contenuto dell'immagine	Idealmente, l'unico oggetto color pelle nell'immagine è la mano dell'utente. Altri oggetti color pelle possono essere visibili, ma devono essere piccoli rispetto alla mano. Questo vale anche per il braccio, che deve essere coperto da una camicia a maniche lunghe se è visibile. Nei gesti dinamici la mano deve essere interamente presente nell'immagine prima dell'inizio dell'acquisizione e non deve uscirne prima della fine della stessa, in modo che sia interamente visibile in ogni fotogramma acquisito.
Illuminazione	L'illuminazione è sufficientemente diffusa in modo che non siano visibili ombre significative sulla mano. Ombre leggere sono accettabili.
Setup	La distanza tra la mano e la camera è scelta in modo che la mano occupi circa il 10-25% dell'immagine. La posizione esatta della mano nell'immagine è arbitraria, ma nessuna parte della mano deve essere ritagliata. La camera non è ruotata, cioè il suo asse x è orizzontale.
Camera	La risoluzione può variare, ma l'immagine deve essere almeno 320×240 . L' <i>aspect ratio</i> rimane costante. La camera viene regolata in modo che non si verifichino sovraesposizioni e solo una lieve alterazione del colore. Per i gesti dinamici, è necessario utilizzare una frequenza di 25 fotogrammi al secondo e la velocità dell'otturatore deve essere sufficientemente alta da evitare la sfocatura del movimento. Una camera economica è sufficiente.

«Recording conditions» per le acquisizioni delle immagini negli esempi (gesti statici, gesti dinamici)

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features



Procedura di elaborazione presente in molti sistemi di riconoscimento di pattern

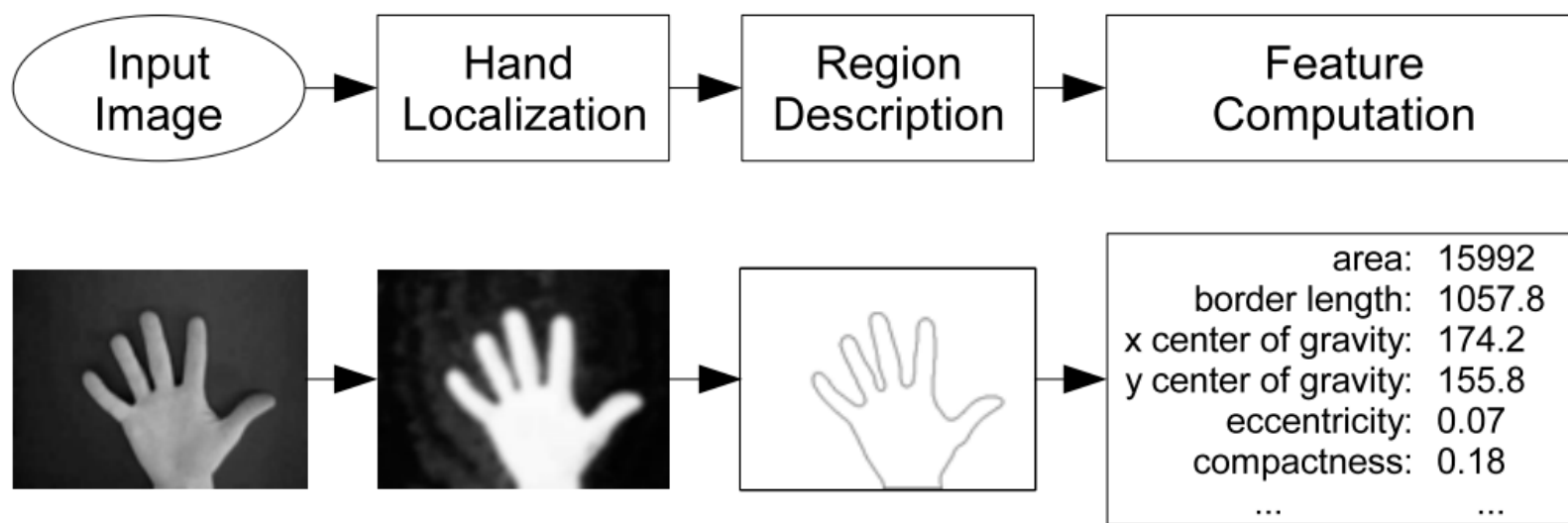
Il passaggio da dati di basso livello (immagine) a una loro descrizione di livello superiore, rappresentata come un vettore di valori scalari, si chiama estrazione delle features.

In questo processo, le informazioni irrilevanti (sfondo) vengono scartate, mentre quelle rilevanti (primo piano o target) vengono isolate. La maggior parte dei sistemi di riconoscimento dei pattern esegue questa fase, perché l'elaborazione dell'immagine completa è troppo onerosa dal punto di vista computazionale e introduce una quantità inaccettabile di rumore (le immagini spesso contengono più pixel di sfondo che pixel in primo piano).

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

Ricordando i due esempi (gesti statici, gesti dinamici), si vogliono presentare alcuni metodi utili per rilevare la mano dell'utente nell'immagine, descrivere la sua forma e calcolare diverse features geometriche che permettano la distinzione di diverse configurazioni statiche della mano e di diversi gesti dinamici. Questo processo viene svolto all'interno del modulo di estrazione delle features.



Visualizzazione della fase di estrazione delle features. La riga superiore indica le fasi di elaborazione, la riga inferiore mostra esempi per i dati corrispondenti.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

A partire dall'immagine acquisita, occorre effettuare la localizzazione del target, cioè la mano negli esempi proposti («Hand Localization»). L'identificazione delle regioni in primo piano o associate al target costituisce un'interpretazione dell'immagine basata su una conoscenza che di solito è specifica dello scenario applicativo. Questa conoscenza può essere codificata in modo esplicito (ad esempio come insieme di regole) o implicito (ad esempio in un istogramma o in una rete neurale). Le proprietà note dell'oggetto target, come la forma, le dimensioni o il colore, possono essere sfruttate. Nel riconoscimento dei gesti, il colore è la feature più frequentemente utilizzata per la localizzazione della mano, poiché la forma e le dimensioni della proiezione della mano nel piano bidimensionale dell'immagine variano notevolmente. Il colore è anche l'unica feature esplicitamente memorizzata nell'immagine.

Nel modello di colore RGB (e nella maggior parte degli altri), anche gli oggetti che si potrebbero definire come caratterizzati da un unico colore, di solito sono rappresentati da un range di valori numerici. Questo range può essere descritto dal punto di vista statistico utilizzando un istogramma discreto tridimensionale, con le dimensioni corrispondenti alle componenti di rosso, verde e blu:

$$h_{\text{object}}(r, g, b)$$

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

$h_{\text{object}}(r, g, b)$ viene calcolato a partire da un numero sufficientemente elevato di pixels. Tali pixels vengono ricavati ad esempio mediante marcatura manuale in un insieme di immagini sorgente (di partenza) che coprono tutte le configurazioni in cui il sistema è destinato a essere utilizzato (ad esempio, più utenti, condizioni di illuminazione variabile, ...). Il valore di h_{object} in corrispondenza di (r, g, b) indica il numero di pixels con il colore corrispondente. La somma totale degli h_{object} considerando tutti i colori è quindi uguale al numero di pixels n_{object} dell'oggetto considerato:

$$\sum_r \sum_g \sum_b h_{\text{object}}(r, g, b) = n_{\text{object}}$$

Di seguito viene riportato un esempio riguardo alla creazione di $h_{\text{object}}(r, g, b)$ considerando un singolo frame.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

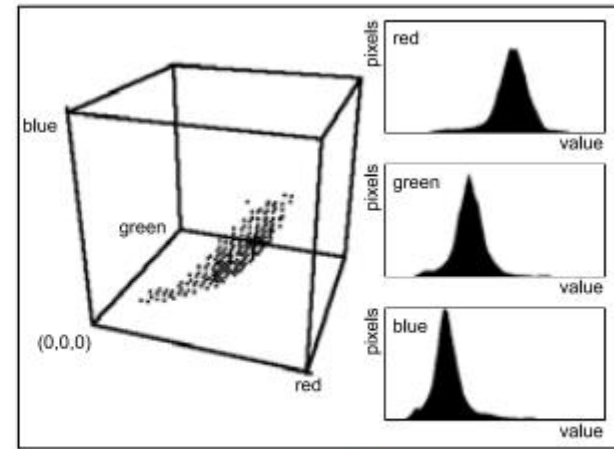
Gesti della mano – Estrazione delle features



Immagine sorgente



Maschera dell'oggetto



Istogramma del colore dell'oggetto

Calcolo dell'istogramma del colore dell'oggetto a partire dall'immagine sorgente e dalla corrispondente maschera generata manualmente.

La figura mostra l'immagine sorgente di una mano e una corrispondente maschera binaria generata manualmente che indica i pixel dell'oggetto (bianco) e i pixel dello sfondo (nero). La visualizzazione tridimensionale utilizza punti per indicare colori che si sono presentati con una certa frequenza minima. I tre grafici monodimensionali a destra mostrano la proiezione sull'asse del rosso, del verde e del blu.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

Sulla base di h_{object} , è possibile eseguire la rilevazione dell'oggetto basata sul colore in nuove immagini relative all'oggetto stesso. L'obiettivo è calcolare, a partire dal colore di un pixel, un valore di probabilità che indichi la possibilità che esso rappresenti una parte dell'oggetto target. Questo valore si ottiene per ogni pixel (x,y) e viene memorizzato in una immagine probabilità con le stesse dimensioni dell'immagine analizzata.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

Dato un pixel relativo all'oggetto, la probabilità di avere un certo colore (r,g,b) può essere calcolata da h_{object} come

$$P(r, g, b|\text{object}) = \frac{h_{\text{object}}(r, g, b)}{n_{\text{object}}}$$

Si può creare un istogramma complemento h_{bg} associato ai colori dello sfondo (background) considerando un numero totale n_{bg} di pixels associati allo sfondo. La probabilità di avere un pixel di sfondo è:

$$P(r, g, b|\text{bg}) = \frac{h_{\text{bg}}(r, g, b)}{n_{\text{bg}}}$$

Applicando la formula di Bayes, la probabilità che un pixel rappresenti una parte dell'oggetto può essere calcolata dal suo colore (r,g,b) utilizzando le due equazioni precedenti:

$$P(\text{object}|r, g, b) = \frac{P(r, g, b|\text{object}) \cdot P(\text{object})}{P(r, g, b|\text{object}) \cdot P(\text{object}) + P(r, g, b|\text{bg}) \cdot P(\text{bg})}$$

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

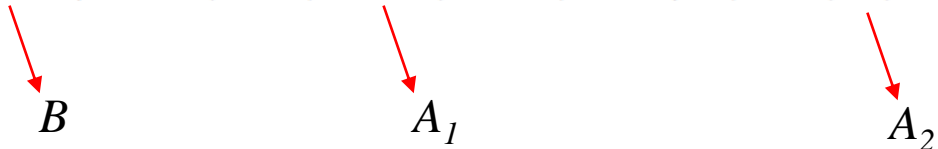
Gesti della mano – Estrazione delle features

Formula di Bayes

$$P(A_i | B) = \frac{P(A_i) P(B | A_i)}{P(B)} = \frac{P(A_i) P(B | A_i)}{\sum_{k=1}^n P(A_k) P(B | A_k)}$$

Nel caso considerato $n = 2$ e $i = 1$.

$$P(\text{object} | r, g, b) = \frac{P(r, g, b | \text{object}) \cdot P(\text{object})}{P(r, g, b | \text{object}) \cdot P(\text{object}) + P(r, g, b | \text{bg}) \cdot P(\text{bg})}$$



$B \qquad A_1 \qquad A_2$

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

$P(\text{object})$ e $P(\text{bg})$ rappresentano le probabilità (a priori) relative all'oggetto e allo sfondo:

$$P(\text{object}) + P(\text{bg}) = 1$$

Utilizzando l'equazione illustrata in precedenza

$$P(\text{object}|r, g, b) = \frac{P(r, g, b|\text{object}) \cdot P(\text{object})}{P(r, g, b|\text{object}) \cdot P(\text{object}) + P(r, g, b|\text{bg}) \cdot P(\text{bg})}$$

si ottiene, a partire dall'immagine \mathbf{I} , l'immagine probabilità associata all'oggetto

$$\mathbf{I}_{\text{obj,prob}}(x, y) = P(\text{object}|\mathbf{I}(x, y))$$

Al fine di classificare ogni pixel come sfondo o target, occorre definire una soglia Θ associata alla probabilità dell'oggetto. Probabilità maggiori o uguali a tale soglia verranno considerate associate al target, mentre le altre allo sfondo. Una struttura dati utile alla rappresentazione di questa classificazione è una maschera binaria

$$\mathbf{I}_{\text{obj,mask}}(x, y) = \begin{cases} 1 & \text{if } \mathbf{I}_{\text{obj,prob}}(x, y) \geq \Theta \quad (\text{target}) \\ 0 & \text{otherwise} \quad (\text{background}) \end{cases}$$

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

$$\mathbf{I}_{\text{obj,mask}}(x, y) = \begin{cases} 1 & \text{if } \mathbf{I}_{\text{obj,prob}}(x, y) \geq \Theta \quad (\text{target}) \\ 0 & \text{otherwise} \quad (\text{background}) \end{cases}$$

La maschera binaria rappresenta una segmentazione a soglia dell'immagine sorgente poiché essa partiziona l'immagine sorgente \mathbf{I} in regioni associate al target e allo sfondo.

E' possibile dimostrare che i valori di $P(\text{object})$ e $P(\text{bg})$ in

$$P(\text{object}|r, g, b) = \frac{P(r, g, b|\text{object}) \cdot P(\text{object})}{P(r, g, b|\text{object}) \cdot P(\text{object}) + P(r, g, b|\text{bg}) \cdot P(\text{bg})}$$

non influenzano la maschera binaria se la soglia Θ viene opportunamente manipolata. Per tale motivo, da un punto di vista pratico, si possono scegliere valori arbitrari per $P(\text{object})$ e $P(\text{bg})$, ad esempio

$$P(\text{object}) = P(\text{bg}) = 0.5$$

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

$$\mathbf{I}_{\text{obj,mask}}(x, y) = \begin{cases} 1 & \text{if } \mathbf{I}_{\text{obj,prob}}(x, y) \geq \Theta \quad (\text{target}) \\ 0 & \text{otherwise} \quad (\text{background}) \end{cases}$$

La scelta della soglia Θ rappresenta un passaggio cruciale per la corretta distinzione tra sfondo e target.

Nel caso in cui le condizioni di acquisizione rimangano costanti e siano note a priori, la soglia può essere scelta manualmente, ma quando tali condizioni non sono verificate, si desidera un calcolo automatico di tale soglia. Se non è disponibile alcuna informazione riguardo alla forma o alla posizione dell'oggetto, si possono utilizzare algoritmi iterativi. Nel caso in cui siano note informazioni approssimative riguardo alla forma e alla posizione, si può definire una soglia adeguata osservando la maschera associata all'oggetto. Tale approccio può essere implementato definendo una forma attesa del target, creando maschere differenti utilizzando differenti soglie e scegliendo la soglia che restituisce la forma più simile a quella attesa. Tale approccio richiede un'estrazione delle features e una metrica (misura) per quantificare la distanza di una determinata forma dalla forma attesa.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features



Immagine sorgente



Immagine probabilità (pelle)

$$I_{\text{obj,prob}}(x, y)$$

(visualizzata come
un'immagine
in scala di grigi)

$$I_{\text{obj,mask}}(x, y)$$



Classificazione
pelle/sfondo θ_1



Classificazione
pelle/sfondo θ_2



Classificazione
pelle/sfondo θ_3

$$\theta_1 < \theta_2 < \theta_3$$

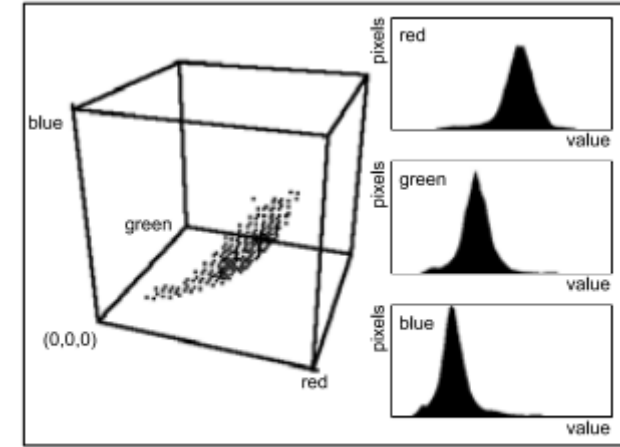
ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

La slide precedente mostra un esempio di «Hand Localization». È stato utilizzato l'istogramma mostrato in precedenza (pelle).

Inoltre è stato utilizzato un istogramma generico per lo sfondo (h_{bg}) e

$$P(\text{object}) = P(\text{bg}) = 0.5.$$



Istogramma del colore dell'oggetto

Vengono utilizzate tre soglie differenti, le quali producono tre maschere binarie. Nessuna delle tre maschere è del tutto corretta:

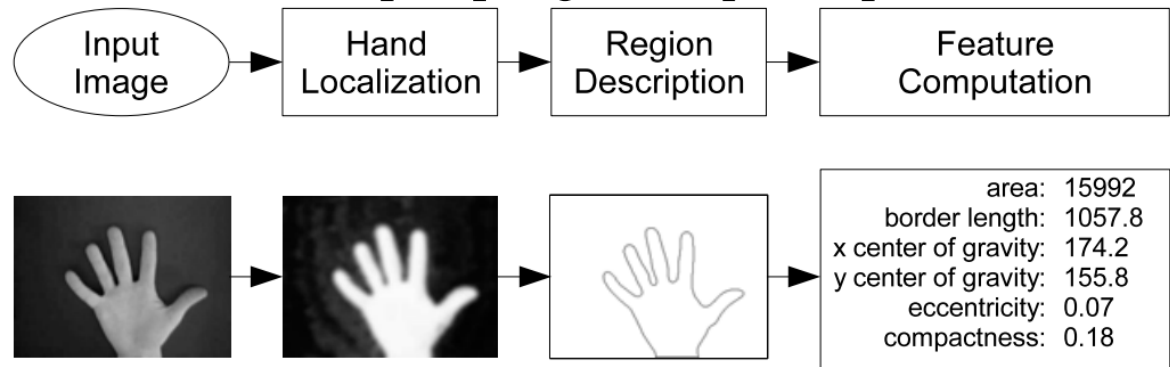
- Θ_1 : numerose regioni di sfondo (come il tappo della bottiglia nell'angolo in basso a destra) sono classificate come primo piano.
- Θ_3 : si notano buchi nel target, soprattutto sui bordi.
- Θ_2 : compromesso che può essere considerato come la soluzione migliore (soglia calcolata automaticamente con un algoritmo iterativo riportato in [1]).

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

Mentre la percezione umana del colore di un oggetto è in gran parte indipendente dall'illuminazione corrente (un effetto chiamato *color constancy*), i colori registrati da una camera sono fortemente influenzati dall'illuminazione e dalle caratteristiche dell'hardware. Questo limita l'uso degli istogrammi alle condizioni di acquisizione in cui sono stati creati i dati di origine, oppure richiede l'applicazione di algoritmi di costanza del colore.

In generale (come mostra l'esempio della prossima slide), non esiste una soglia in grado di poter garantire un risultato corretto. Ci può essere il rischio che un oggetto non desiderato venga riconosciuto come target (*false alarm*). A meno che non sia possibile ricavare istogrammi in grado di ridurre il numero di falsi allarmi, le fasi di elaborazione successive devono essere concepite per gestire questo problema.



Visualizzazione della fase di estrazione delle features. La riga superiore indica le fasi di elaborazione, la riga inferiore mostra esempi per i dati corrispondenti.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features



Immagine sorgente



Immagine probabilità (pelle)

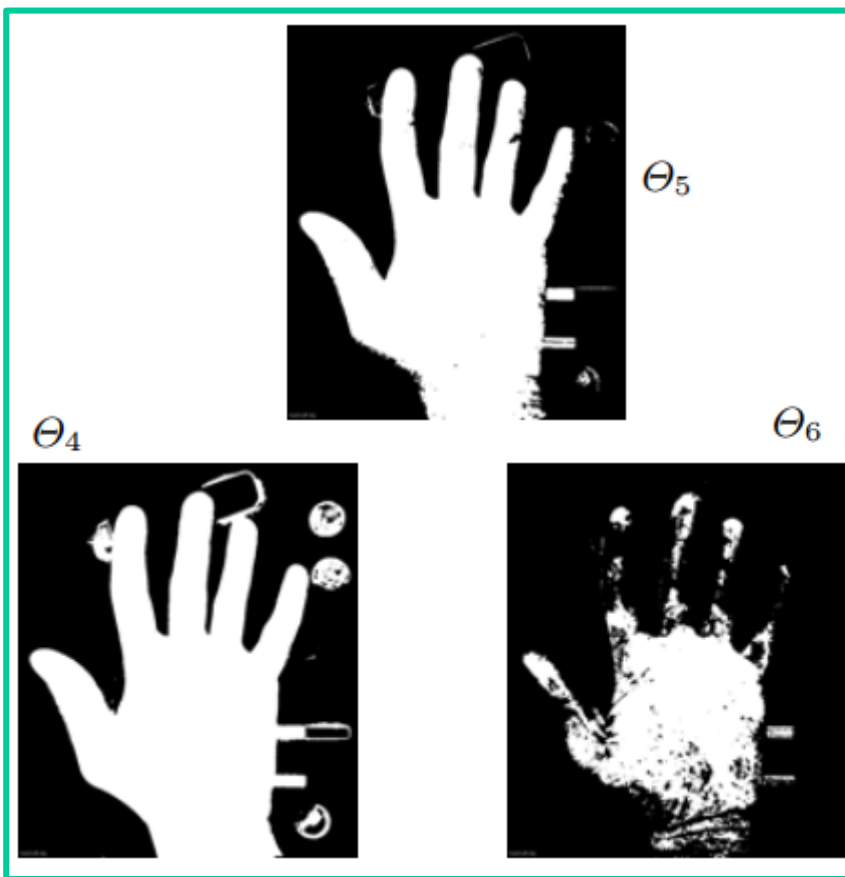
$$I_{\text{obj,prob}}(x, y)$$

Altro esempio di
«Hand Localization». È stato
utilizzato un altro istogramma
rispetto all'esempio delle slides
precedenti con

Classificazione
pelle/sfondo

$$I_{\text{obj,mask}}(x, y)$$

$$P(\text{object}) = P(\text{bg}) = 0.5$$



$$\Theta_4 < \Theta_5 < \Theta_6$$

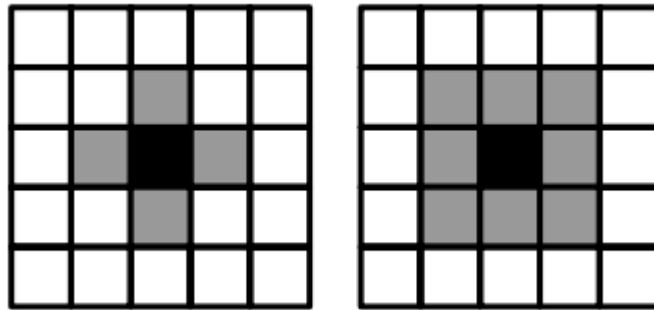
Vengono utilizzate tre soglie differenti, le quali
producono tre maschere binarie.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

$$\mathbf{I}_{\text{obj,mask}}(x, y) = \begin{cases} 1 & \text{if } \mathbf{I}_{\text{obj,prob}}(x, y) \geq \Theta \quad (\text{target}) \\ 0 & \text{otherwise} \quad (\text{background}) \end{cases}$$

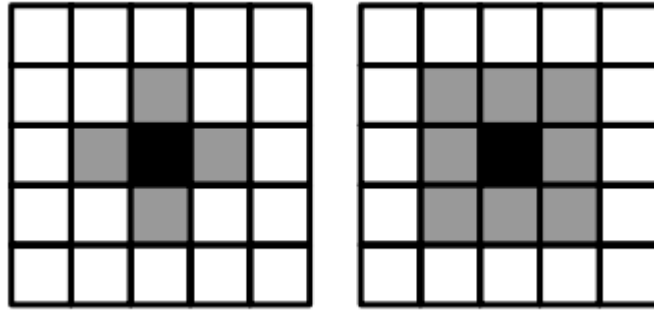
La maschera binaria definita in precedenza può essere utilizzata per effettuare la descrizione della regione («Region Description»). Mediante tale maschera, l'immagine sorgente \mathbf{I} può essere partizionata in regioni. Una regione R è un insieme contiguo di pixels p per i quali la maschera binaria $\mathbf{I}_{\text{obj,mask}}$ ha lo stesso valore. Il concetto di contiguità richiede la definizione di adiacenza tra pixels.



Pixels adiacenti (grigio) nel 4-intorno (immagine a sinistra) e nel 8-intorno (immagine a destra) di un pixel di riferimento (nero).

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features



Pixels adiacenti (grigio) nel 4-intorno (immagine a sinistra) e nel 8-intorno (immagine a destra) di un pixel di riferimento (nero).

Come mostrato nella figura precedente, l'adiacenza può essere basata sul concetto di *4-intorno* o sul concetto di *8-intorno*. In generale, le regioni possono contenere altre regioni e/o buchi, ma questo aspetto non verrà considerato perché è di scarsa importanza in molte applicazioni di riconoscimento dei gesti.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

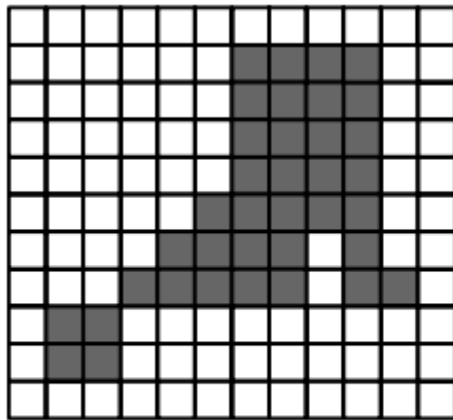
Ipotizzando «condizioni di acquisizione in laboratorio» («laboratory recording conditions»), la maschera binaria $I_{obj,mask}$ conterrà un'unica regione associata al target. In realtà, in molte applicazioni reali, molti oggetti potrebbero essere confusi con il target (come mostrato nell'ultimo esempio nelle slides precedenti). Se non vengono attuate ulteriori strategie, la fase di estrazione delle features è responsabile dell'identificazione della regione che rappresenta il target (la mano nell'esempio), dovendo eventualmente scegliere in un insieme di regioni candidate. Ciò viene fatto considerando features geometriche associate alle regioni.

Il primo passo da eseguire per il calcolo di tali features è la creazione di una descrizione esplicita di ogni regione contenuta nella maschera binaria. Esistono differenti algoritmi in grado di generare, per ogni regione, una lista dei pixels corrispondenti. Una rappresentazione più compatta e più efficiente dal punto di vista computazionale di una regione può essere ottenuta memorizzando solo i suoi punti di contorno (border points); questo poiché i punti di contorno sono generalmente in numero inferiore rispetto ai pixels della regione.

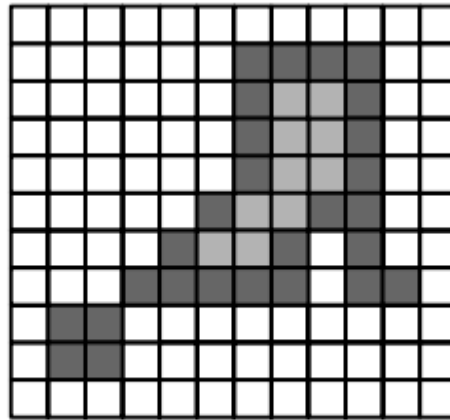
ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

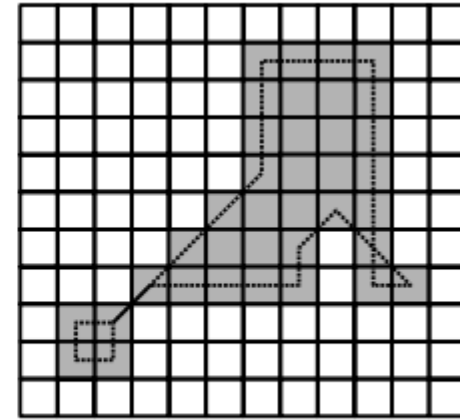
Un punto di contorno è definito come un pixel $p \in R$ che ha almeno un pixel $q \notin R$ all'interno del suo 4-intorno.



Pixels della regione
(grigio scuro)



Punti di contorno
(grigio scuro)



Poligono

$$B_R = \{(x_0, y_0), (x_1, y_1), \dots, (x_{n-1}, y_{n-1})\}$$

Una struttura dati efficiente per la rappresentazione di una regione è una lista ordinata degli n punti di contorno. Essa può essere interpretata come un poligono chiuso i cui vertici sono i centri dei punti di contorno. Si può quindi definire il contorno di un oggetto come questo poligono.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features



Immagine sorgente



Immagine probabilità (pelle)

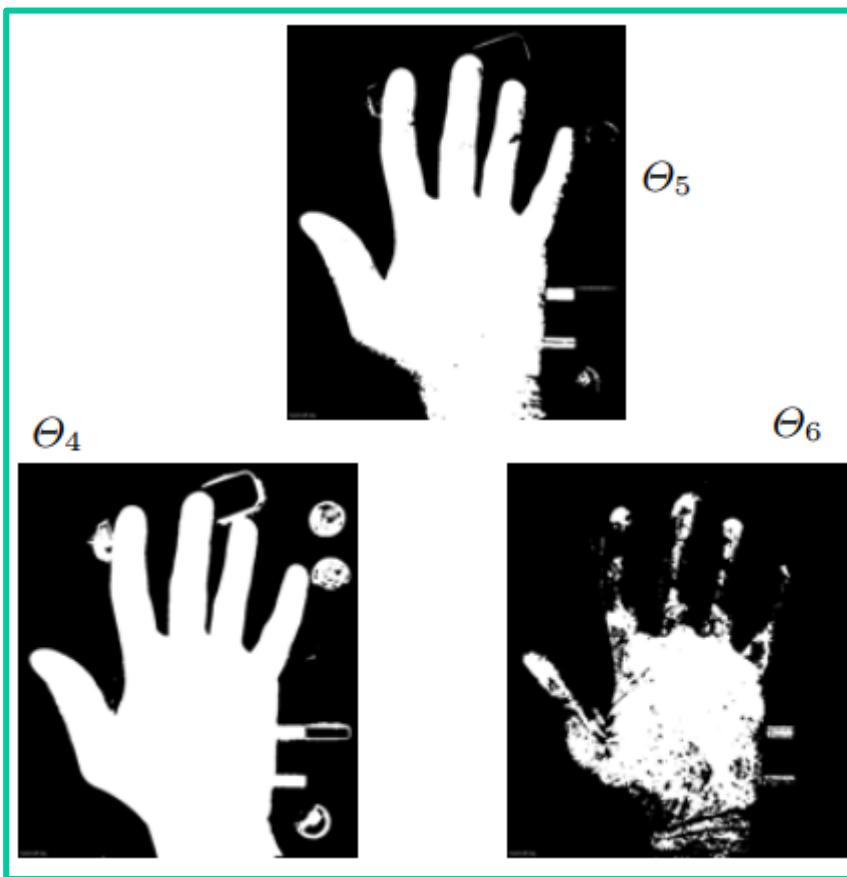
$$I_{\text{obj,prob}}(x, y)$$

Altro esempio di
«Hand Localization». È stato
utilizzato un altro istogramma
rispetto all'esempio delle slides
precedenti con

Classificazione
pelle/sfondo

$$I_{\text{obj,mask}}(x, y)$$

$$P(\text{object}) = P(\text{bg}) = 0.5$$



$$\Theta_4 < \Theta_5 < \Theta_6$$

Vengono utilizzate tre soglie differenti, le quali
producono tre maschere binarie.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

θ_4



θ_5



θ_6



Classificazione
pelle/sfondo

$$I_{\text{obj,mask}}(x, y)$$

$$\theta_4 < \theta_5 < \theta_6$$



(a)



(b)



(c)

Punti di contorno
(bordi) delle regioni
mostrate nelle
maschere ottenute in
precedenza
(è stato utilizzato un
algoritmo riportato
in [1])

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

Osservando i punti di contorno (bordi) delle regioni mostrate nelle maschere ottenute in precedenza, si può notare che nelle aree dove la probabilità associata alla pelle umana è vicina alla soglia Θ utilizzata, i bordi diventano frastagliati (specialmente nei casi (b) e (c)).

Questo effetto provoca un aumento casuale della lunghezza dei bordi; tale aumento è indesiderato perché riduce il contenuto informativo del valore calcolato associato alla lunghezza del bordo (forme simili potrebbero avere lunghezze del bordo significativamente differenti, rendendo tale feature poco utile per il riconoscimento).

Un miglioramento può essere ottenuto effettuando, prima della segmentazione

$$\mathbf{I}_{\text{obj,mask}}(x, y) = \begin{cases} 1 & \text{if } \mathbf{I}_{\text{obj,prob}}(x, y) \geq \Theta \quad (\text{target}) \\ 0 & \text{otherwise} \quad (\text{background}) \end{cases}$$

una convoluzione dell'immagine probabilità $\mathbf{I}_{\text{obj,prob}}$ con un kernel Gaussiano (si vedano le figure (d), (e), (f) nella slide successiva).

Convoluzione: la convoluzione è il processamento di una matrice attraverso un'altra che viene chiamata «kernel». Il tipo di kernel che viene utilizzato dipende dall'effetto che si vuole ottenere.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features



(a)



(b)



(c)



(d)



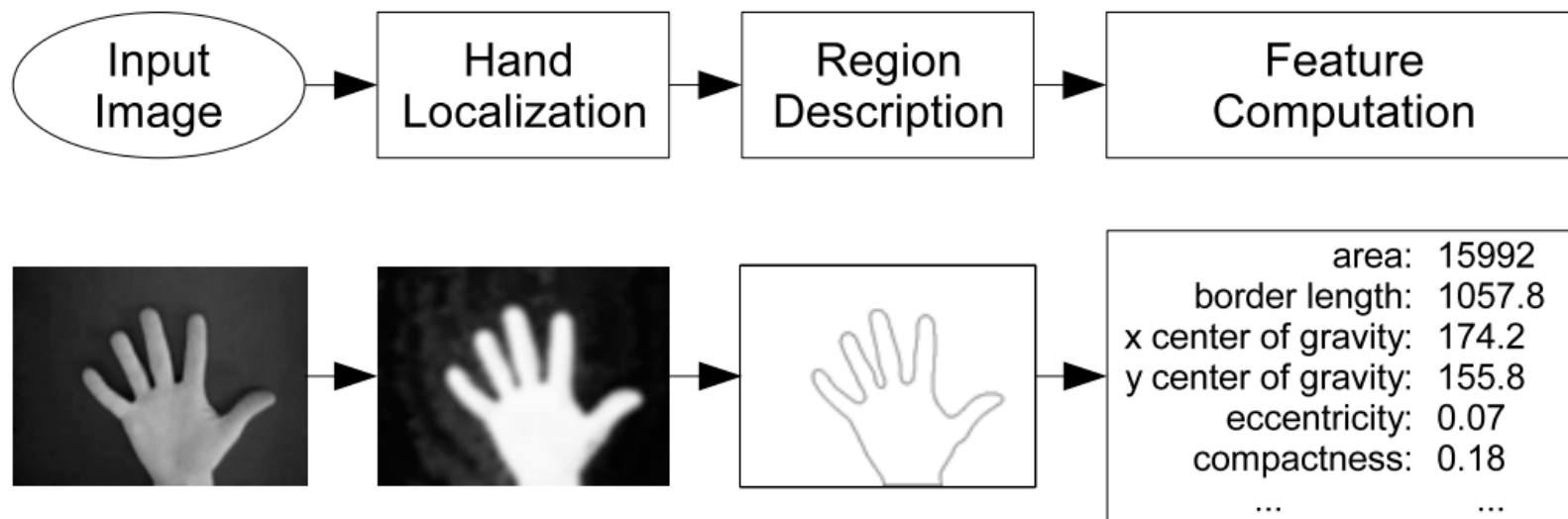
(e)



(f)

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features



Visualizzazione della fase di estrazione delle features. La riga superiore indica le fasi di elaborazione, la riga inferiore mostra esempi per i dati corrispondenti.

Si procede quindi con il calcolo delle features (vengono considerate features geometriche). Possono essere calcolate tantissime features per un poligono chiuso, ma solo un sottoinsieme di esse è adatto per il riconoscimento. Una feature è adatta se essa è caratterizzata da un'alta *inter-gesture variance* (varia significativamente tra gesti differenti) e una bassa *intra-gesture variance* (varia poco tra più riproduzioni dello stesso gesto).

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

La prima proprietà (*inter-gesture variance*) indica che la feature contiene molte informazioni, mentre la seconda (*intra-gesture variance*) indica che la feature non è affetta significativamente da rumore o variazioni non intenzionali che inevitabilmente si verificano.

Ogni feature deve essere stabile, cioè piccoli cambiamenti nei dati di input non devono risultare in grandi cambiamenti nella feature.

Infine, la feature deve poter essere calcolata con sufficiente accuratezza e velocità.

Il fatto che una determinata feature sia adatta dipende quindi dalla struttura del vocabolario (poiché questa influenza le proprietà *inter-gesture variance* e *intra-gesture variance*) e dallo scenario applicativo in termini di «condizioni di acquisizione» («recording conditions»), hardware, ... Quindi si può pensare di calcolare quante più features possibile e poi esaminare ognuna di esse in relazione alla sua idoneità per lo specifico compito di riconoscimento.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

La scelta delle features è di grande importanza, in quanto influenza la progettazione del sistema in tutta la procedura di elaborazione. Le prestazioni del sistema dipendono in modo significativo dalle features scelte e dalle loro proprietà.

Quando si definisce una feature, si considera la sua equazione e il modo in cui ciascuna feature viene influenzata dalla prospettiva della camera, dalla risoluzione e dalla distanza dall'oggetto.

Nel dominio dell'immagine, un cambiamento nella prospettiva della camera risulta in una rotazione o traslazione dell'oggetto, mentre la risoluzione e la distanza dall'oggetto influenzano la scala dell'oggetto (in termini di pixels).

In un'immagine discretizzata, features dichiarate «invarianti» potrebbero essere affette da piccole variazioni (ciò può essere ignorato a meno che la dimensione della forma è dell'ordine di alcuni pixels).

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

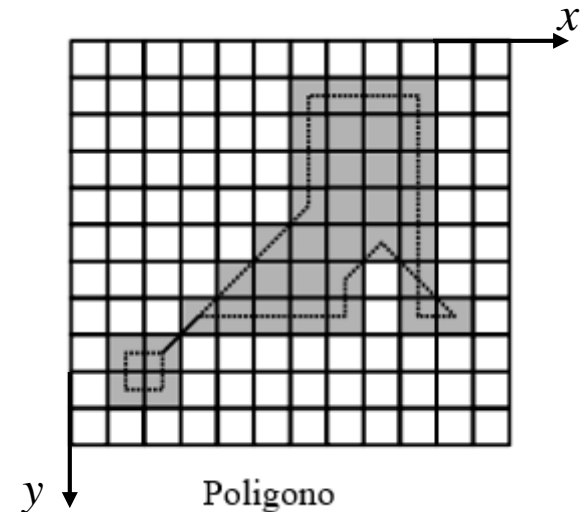
Gesti della mano – Estrazione delle features

Di seguito vengono presentate alcune delle features geometriche più utilizzate. Si ricorda che le features vengono calcolate sul generico poligono chiuso rappresentato da

$$B_R = \{(x_0, y_0), (x_1, y_1), \dots, (x_{n-1}, y_{n-1})\}$$

Lunghezza del bordo (border length) l

Tale feature può essere calcolata dal bordo della regione, tenendo conto che la distanza tra due punti di contorno successivi è 1 se le loro coordinate x o y sono uguali, $\sqrt{2}$ altrimenti. l dipende dalla scala/risoluzione ed è invariante alla traslazione e alla rotazione.



ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

Area, Centro di gravità (center of gravity, COG) e Momenti del Secondo Ordine

Tali features rientrano nella generica definizione di momenti arbitrari $\nu_{p,q}$ dei poligoni. Nelle espressioni che seguono, x_i e y_i ($i = 0, 1, \dots, n-1$) sono gli elementi di B_R . Poiché si richiede che il poligono sia chiuso, per $i = n$ si ha che

$$x_n = x_0 \quad y_n = y_0$$

$$\nu_{0,0} = a = \frac{1}{2} \sum_{i=1}^n x_{i-1} y_i - x_i y_{i-1} \quad \text{Area } a$$

$$\alpha_{1,0} = x_{\text{cog}} = \frac{1}{6a} \sum_{i=1}^n (x_{i-1} y_i - x_i y_{i-1}) (x_{i-1} + x_i)$$

Centro di gravità (center of gravity, COG) $(x_{\text{cog}}, y_{\text{cog}})$

$$\alpha_{0,1} = y_{\text{cog}} = \frac{1}{6a} \sum_{i=1}^n (x_{i-1} y_i - x_i y_{i-1}) (y_{i-1} + y_i)$$

x center of gravity

y center of gravity

$$\alpha_{2,0} = \frac{1}{12a} \sum_{i=1}^n (x_{i-1} y_i - x_i y_{i-1}) (x_{i-1}^2 + x_{i-1} x_i + x_i^2)$$

$\alpha_{p,q}$ Momenti normalizzati

$\mu_{p,q}$ Momenti centrali

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

$$\alpha_{1,1} = \frac{1}{24a} \sum_{i=1}^n (x_{i-1}y_i - x_iy_{i-1})(2x_{i-1}y_{i-1} + x_{i-1}y_i + x_iy_{i-1} + 2x_iy_i)$$

$$\alpha_{2,0} = \frac{1}{12a} \sum_{i=1}^n (x_{i-1}y_i - x_iy_{i-1})(y_{i-1}^2 + y_{i-1}y_i + y_i^2)$$

$$\mu_{2,0} = \alpha_{2,0} - \alpha_{1,0}^2$$

$$\mu_{1,1} = \alpha_{1,1} - \alpha_{1,0}\alpha_{0,1}$$

$$\mu_{0,2} = \alpha_{0,2} - \alpha_{0,1}^2$$

$\alpha_{p,q}$ Momenti normalizzati

$\mu_{p,q}$ Momenti centrali

L'area a dipende dalla scala/risoluzione, ed essa è indipendente da traslazione e rotazione. Il centro di gravità $(x_{\text{cog}}, y_{\text{cog}})$ varia con la traslazione e dipende dalla risoluzione.

I momenti del secondo ordine ($p+q=2$) vengono utilizzati per calcolare altri descrittori della forma.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

Eccentricità (eccentricity) e

$$e = \frac{(\mu_{2,0} - \mu_{0,2})^2 + 4\mu_{1,1}}{a}$$

L'eccentricità definita mediante tale formula è 0 per forme circolari e cresce per forme allungate. È invariante alla traslazione e alla rotazione, ma varia con la scala/risoluzione.

Orientamento (orientation) α

Ciò che è intuitivamente chiamato «orientamento dell'oggetto» viene definito da

$$\alpha = \frac{1}{2} \arctan \left(\frac{2\mu_{1,1}}{\mu_{2,0} - \mu_{0,2}} \right)$$

L'orientamento è invariante alla traslazione e alla scala/risoluzione.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

Compattezza (compactness) c

La compattezza di una forma $c \in [0, 1]$ è definita come

$$c = \frac{4\pi a}{l^2}$$

Forme compatte ($c \rightarrow 1$) hanno bordi corti l che contengono una grande area a . La forma più compatta è il cerchio ($c = 1$), mentre per le forme allungate si ha $c \rightarrow 0$. La compattezza è invariante alla scala/risoluzione, alla traslazione e alla rotazione.

Features relative al bordo

Oltre alla lunghezza del bordo l , si possono calcolare

- x_{\min} , x_{\max} , y_{\min} e y_{\max} (coordinata minima e coordinata massima)
- r_{\min} e r_{\max} (distanza minima e distanza massima dal centro di gravità al bordo)

La coordinata minima e la coordinata massima non sono invarianti ad alcuna trasformazione. La distanza minima e la distanza massima dal centro di gravità al bordo sono invarianti alla traslazione e alla rotazione, ma variano con la scala/risoluzione.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

Alcune delle features elencate possono essere utilizzate in un'applicazione reale solo se si esegue un'adeguata *normalizzazione* per annullare la loro variazione rispetto alla traslazione e alla scala. Ad esempio, l'utente potrebbe eseguire gesti in posizioni diverse dell'immagine e a diverse distanze dalla camera.

La miglior strategia di normalizzazione dipende dall'applicazione e dalle condizioni di acquisizione («recording conditions»); essa viene di solito individuata empiricamente. Nel seguito, il risultato della normalizzazione di una feature f verrà indicato con f' .

L'invarianza può essere ottenuta anche calcolando nuove features da due o più features non normalizzate. Ad esempio, una feature invariante rispetto alla scala e alla risoluzione è

$$x_p = \frac{x_{\max} - x_{cog}}{x_{cog} - x_{\min}}$$

Tale feature, detta *protrusion ratio* (*rapporto di sporgenza*), specifica il rapporto della protrusione orizzontale più lunga, misurata dal centro di gravità, verso destra e verso sinistra.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

Riprendiamo ora gli esempi riguardanti i *gesti statici* e i *gesti dinamici* e mostriamo come si possono applicare le tecniche appena descritte.

Gesti statici – I gesti statici



«sinistra»



«destra»

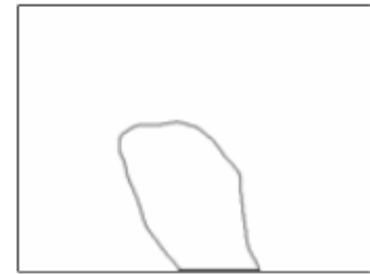
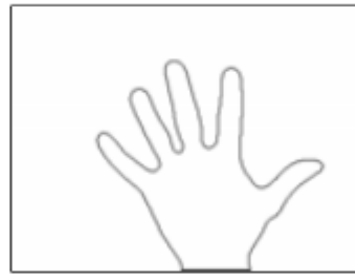
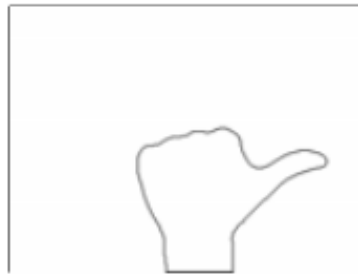


«stop»



posizione di riposo

vengono segmentati utilizzando un algoritmo iterativo per il calcolo automatico della soglia Θ e un algoritmo per il calcolo dei punti di contorno riportati in [1], ottenendo i contorni seguenti



ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

e le features (calcolate dai contorni riportati nella slide precedente)

Feature	Symbol	Gesture			
		“left”	“right”	“stop”	none
Normalized Border Length	l'	1.958	1.705	3.306	1.405
Normalized Area	a'	0.138	0.115	0.156	0.114
Normalized Center of Gravity	x'_{cog}	0.491	0.560	0.544	0.479
	y'_{cog}	0.486	0.527	0.487	0.537
Eccentricity	e	1.758	1.434	1.722	2.908
Orientation	α	57.4°	147.4°	61.7°	58.7°
Compactness	c	0.451	0.498	0.180	0.724
Normalized Min./Max. Coordinates	x'_{min}	0.128	0.359	0.241	0.284
	x'_{max}	0.691	0.894	0.881	0.681
	y'_{min}	0.256	0.341	0.153	0.325
	y'_{max}	0.747	0.747	0.747	0.747
Protrusion Ratio	x_p	0.550	1.664	1.110	1.036

posizione di riposo

$$x_p = \frac{x_{max} - x_{cog}}{x_{cog} - x_{min}}$$

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

Feature	Symbol	Gesture			
		“left”	“right”	“stop”	none
Normalized Border Length	l'	1.958	1.705	3.306	1.405
Normalized Area	a'	0.138	0.115	0.156	0.114
Normalized Center of Gravity	x'_{cog}	0.491	0.560	0.544	0.479
	y'_{cog}	0.486	0.527	0.487	0.537
Eccentricity	e	1.758	1.434	1.722	2.908
Orientation	α	57.4°	147.4°	61.7°	58.7°
Compactness	c	0.451	0.498	0.180	0.724
Normalized Min./Max. Coordinates	x'_{min}	0.128	0.359	0.241	0.284
	x'_{max}	0.691	0.894	0.881	0.681
	y'_{min}	0.256	0.341	0.153	0.325
	y'_{max}	0.747	0.747	0.747	0.747
Protrusion Ratio	x_p	0.550	1.664	1.110	1.036

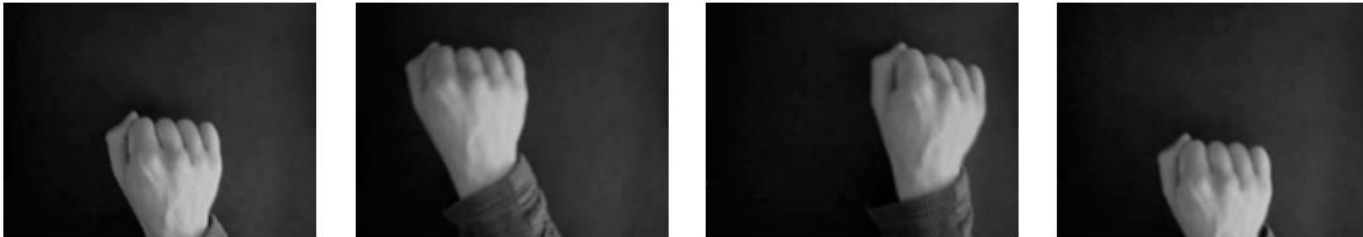
posizione di riposo

Per ottenere invarianza rispetto alla risoluzione, le lunghezze e le coordinate sono state normalizzate tenendo conto della larghezza dell'immagine N , mentre l'area è stata normalizzata con un fattore N^2 . α specifica l'angolo di cui l'oggetto dovrebbe ruotare per allinearsi all'asse x .

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

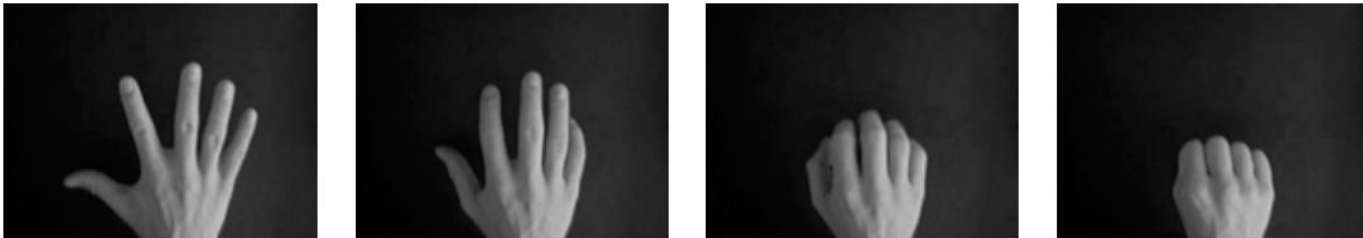
Gesti dinamici – Per quanto riguarda i gesti dinamici



(a) «senso orario» (da sinistra a destra) e «senso antiorario» (da destra a sinistra)



(b) «aperto» (da sinistra a destra) e «chiuso» (da destra a sinistra)

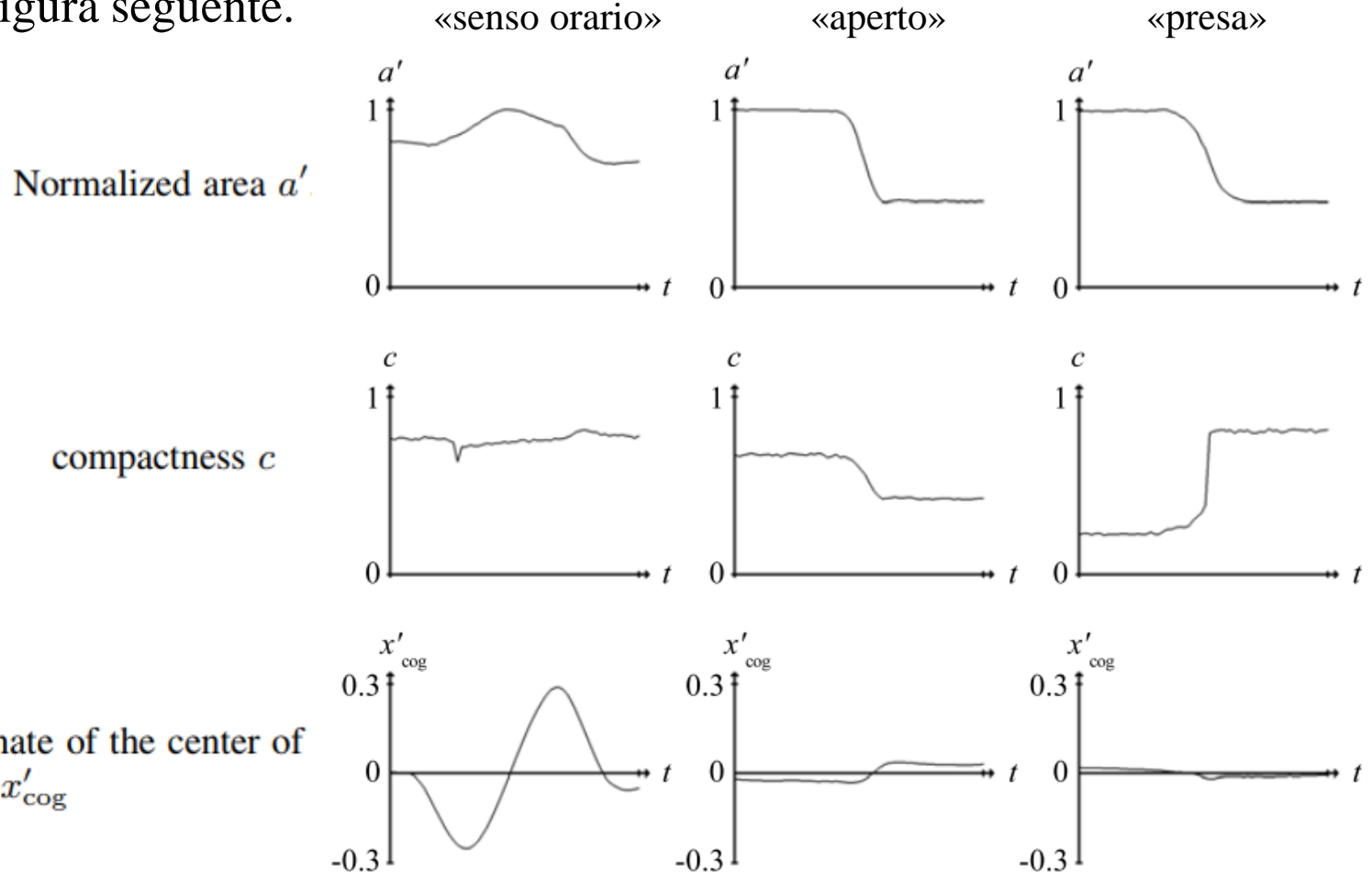


(c) «presa» (da sinistra a destra) e «rilascio» (da destra a sinistra)

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

considerando «senso orario», «aperto» e «presa», vengono calcolate tre features, come mostrato nella figura seguente.



Features calcolate per i gesti dinamici (tempo $t=1, \dots, 60$)

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Estrazione delle features

Poiché i rimanenti tre gesti sono esecuzioni a ritroso degli altri tre, i grafici associati alle loro features sono semplicemente versioni temporalmente speculari dei grafici mostrati nella slide precedente.


Poiché la feature a dipende dall'anatomia della mano e dalla sua distanza dalla camera, essa è stata normalizzata tenendo conto del suo valore massimo per eliminare tale dipendenza:

$$f'(t) = \frac{f(t)}{\max |f(t)|}$$

La feature x_{cog} è stata divisa per N in modo da eliminare la dipendenza dalla risoluzione e inoltre è stata applicata la seguente formula in modo da avere valore medio nullo:

$$f'(t) = f(t) - \overline{f}$$

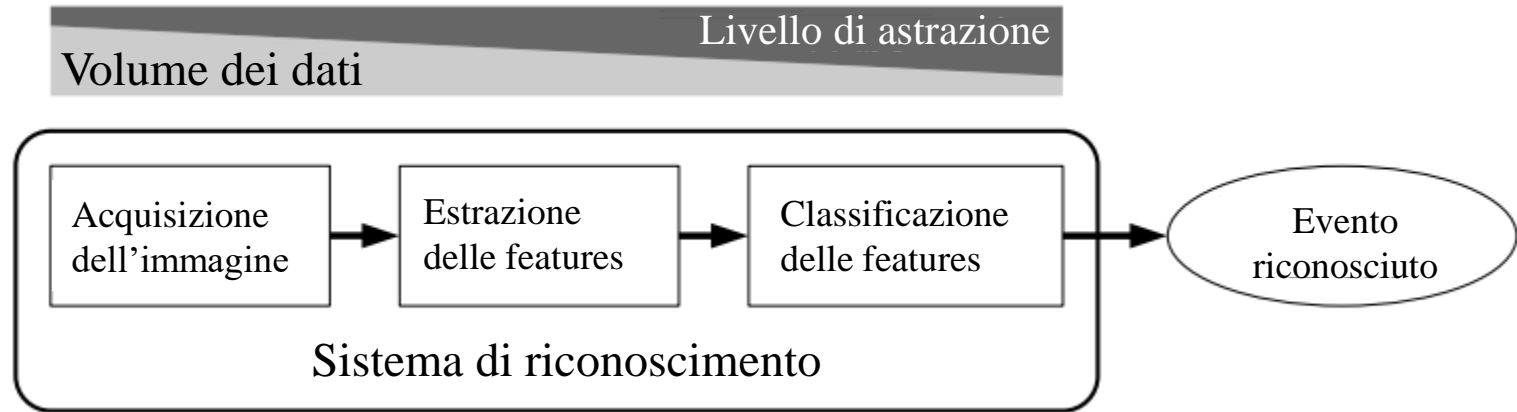
Media
aritmetica



Tale accorgimento permette di eseguire i gesti in qualsiasi punto dell'immagine.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features



Procedura di elaborazione presente in molti sistemi di riconoscimento di pattern

La classificazione delle features viene eseguita in tutti i sistemi di riconoscimento di pattern. È disponibile un elevato numero di algoritmi per costruire classificatori che possono essere utilizzati per diversi scopi.

Alcune categorie di classificatori operano in due fasi: la fase di training («training phase») e la fase di classificazione («classification phase»). Nella fase di training, il classificatore «impara» il vocabolario da un numero sufficientemente grande di esempi rappresentativi (denominati *training samples*). Questa «conoscenza» viene poi applicata nella fase di classificazione.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Dal punto di vista matematico, il compito della classificazione delle features è quello di identificare un evento non noto ω dato un insieme finito Ω di n possibili eventi mutuamente esclusivi.

$$\omega \in \Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$$

Gli elementi di Ω vengono chiamati classi. Nel riconoscimento dei gesti, Ω è il vocabolario, e ogni classe ω_i ($i = 1, 2, \dots, n$) rappresenta un gesto.

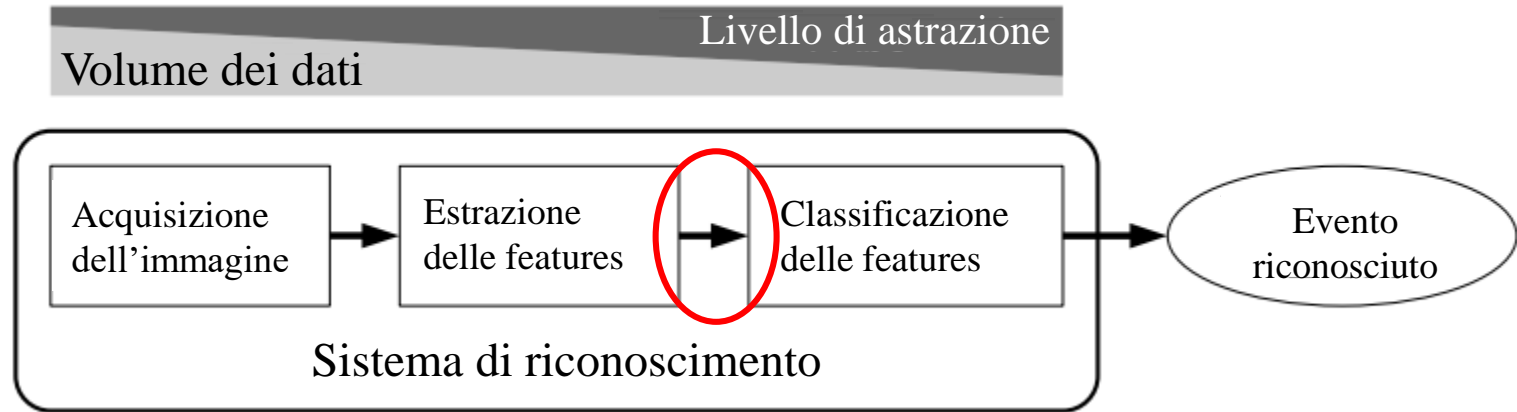
Si noti che

$$\omega \in \Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$$

restringe gli inputs agli elementi di Ω . Ciò significa che al sistema non viene mai presentato un gesto sconosciuto, il che influenza la progettazione e gli algoritmi del classificatore.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features



Procedura di elaborazione presente in molti sistemi di riconoscimento di pattern

Dalla fase di estrazione delle features, il classificatore riceve un'osservazione \mathbf{O} (si veda il cerchio rosso nella figura). Tale osservazione può essere un singolo *feature vector* (nel caso di gesti statici) oppure una sequenza di *feature vector* (nel caso di gesti dinamici).

l'osservazione \mathbf{O} indica una sequenza di T osservazioni singole \mathbf{o}_t

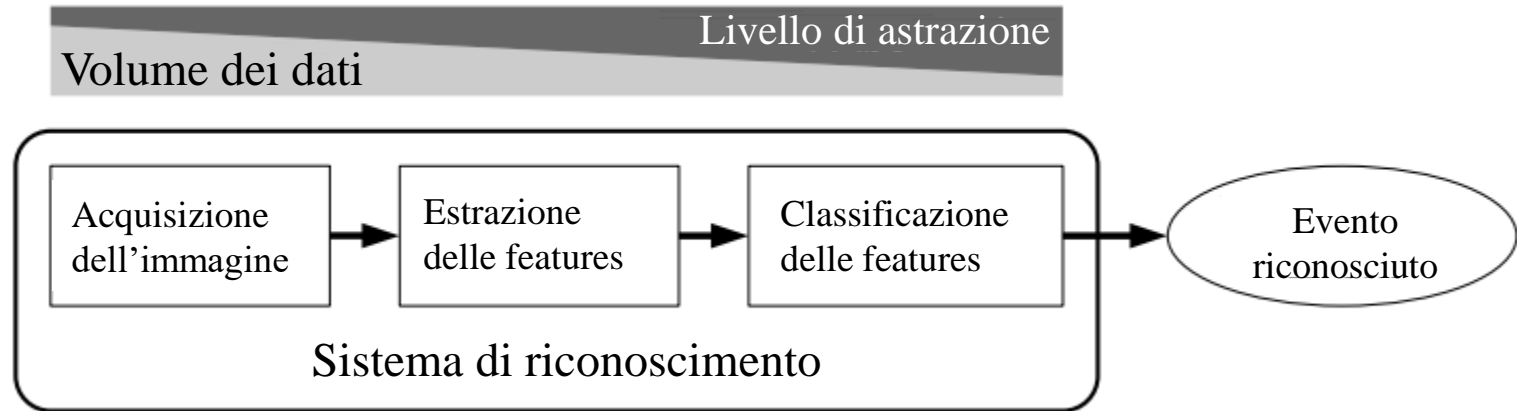
$$\mathbf{O} = (\mathbf{o}_1 \ \mathbf{o}_2 \ \cdots \ \mathbf{o}_T) \quad \mathbf{o}_t = \begin{pmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_K(t) \end{pmatrix}$$

l'osservazione \mathbf{O} è un vettore di K features scalari f_i :

$$\mathbf{O} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_K \end{pmatrix}$$

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features



Procedura di elaborazione presente in molti sistemi di riconoscimento di pattern

Sulla base dell'osservazione ricevuta, il classificatore restituisce in uscita un risultato (si veda «Evento riconosciuto» nella figura)

$$\hat{\omega} \in \Omega \quad \text{dove} \quad \hat{\omega} = \omega_k, \quad k \in \{1, 2, \dots, n\}$$

k indica l'indice della classe dell'evento ω che si ipotizza essere la sorgente dell'osservazione \mathbf{O} . Se $\hat{\omega} = \omega$ il risultato della classificazione è corretto, altrimenti è errato.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Per considerare il caso in cui \mathbf{O} non presenti una sufficiente somiglianza con alcun elemento di Ω , si può decidere di permettere un rifiuto di \mathbf{O} . Questo si ottiene introducendo una pseudo-classe ω_0 e definendo un insieme di uscite del classificatore $\hat{\Omega}$ come

$$\hat{\Omega} = \Omega \cup \{\omega_0\}$$

Quindi la

$$\hat{\omega} \in \Omega \quad \text{dove} \quad \hat{\omega} = \omega_k, \quad k \in \{1, 2, \dots, n\}$$

diventa

$$\hat{\omega} \in \hat{\Omega} \quad \text{where} \quad \hat{\omega} = \omega_k, \quad k \in \{0, 1, \dots, n\}$$

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Un'alternativa al rifiuto di **O** è costituita dalla possibilità di includere esplicitamente in Ω classi che rappresentano eventi per i quali non si desidera una reazione del sistema (dette *garbage classes*). Il classificatore tratta tali classi come classi regolari, ma i passi successivi non eseguono alcuna azione quando il risultato del classificatore è una di queste *garbage classes*.

Ad esempio, un sistema di riconoscimento dei gesti può osservare costantemente la mano dell'utente che tiene il volante, ma reagire solo a gesti specifici diversi dai movimenti dello sterzo. Per questa applicazione, le *garbage classes* conterrebbero i movimenti della mano lungo la forma circolare del volante.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Alcuni tipi di classificatori possono essere progettati per considerare un valore di costo (cost) o di perdita (loss) per errori di classificazione, includendo la reiezione. In generale, tale valore dipende dalla classe di input ω e dall'uscita (risultato) $\hat{\omega}$ del classificatore:

$$L(\omega, \hat{\omega}) = \text{costo per la classificazione dell'evento della classe } \omega \text{ come } \hat{\omega}$$

Il costo di un risultato corretto viene settato a 0:

$$L(\omega, \omega) = 0$$

Tale accorgimento permette di prendere in considerazione applicazioni nelle quali alcuni errori di classificazione sono più gravi di altri.

Ad esempio, consideriamo un'applicazione di riconoscimento dei gesti che esegue la navigazione della struttura di un menù. L'errata classificazione del gesto di «ritorno al menù principale» come «spostamento del cursore verso il basso» richiede che l'utente esegua nuovamente «ritorno al menù principale».

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

L'errata classificazione di «sposta il cursore verso il basso» come «torna al menù principale», tuttavia, elimina la voce del menù attualmente selezionata, il che è un errore più grave perché l'utente dovrà navigare di nuovo verso questa voce del menù (supponendo che non ci sia una funzionalità di «annullamento» o «indietro»).

Nel progetto di un classificatore, si può decidere di non considerare la possibilità di rifiuto, quindi

$$\hat{\omega} \in \Omega \quad \text{dove} \quad \hat{\omega} = \omega_k, \quad k \in \{1, 2, \dots, n\}$$

Inoltre, si può decidere che il valore di perdita provocato da un errore di classificazione sia un valore costante L :

$$L(\omega, \hat{\omega}) = L \quad \text{for} \quad \omega \neq \hat{\omega}$$

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Per quanto riguarda le procedure di classificazione, occorre fare una distinzione tra classificazione supervisionata (supervised) e classificazione non supervisionata (unsupervised).

Nella classificazione supervisionata i campioni che vengono utilizzati nella fase di training sono etichettati (labeled), cioè è nota la classe di ogni campione di training. La classificazione di un nuovo campione viene eseguita attraverso un confronto con tali campioni di training (o con modelli creati a partire da essi) e restituisce la classe che meglio corrisponde all'osservazione, secondo un criterio di corrispondenza.

Nella classificazione non supervisionata i campioni che vengono utilizzati nella fase di training non sono etichettati. Vengono utilizzati algoritmi di clustering al fine di raggruppare campioni simili prima di effettuare la classificazione. Quindi nella classificazione non supervisionata il compito di etichettare i dati è affidato al classificatore, il quale può introdurre errori. Il risultato della classificazione non supervisionata è un indice di clustering e non un'etichetta.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Per il riconoscimento dei gesti, la procedura di classificazione più comunemente utilizzata è la classificazione supervisionata poiché nel processo di acquisizione dei campioni di input (immagini) è agevole assegnare le etichette.

Per molti algoritmi di classificazione, l'ottimizzazione delle prestazioni sui campioni di training rischia di peggiorare le prestazioni riguardanti altri campioni generici. Tale fenomeno è chiamato *overfitting* e si presenta soprattutto per piccoli insiemi di campioni di training. Un classificatore con un insieme di parametri p si adatta in modo eccessivo (cioè si ha l'*overfitting*) ai campioni di training T quando esiste un altro insieme di parametri m che restituisce prestazioni peggiori su T ma prestazioni migliori nelle applicazioni nel «mondo reale» (cioè su campioni diversi da quelli utilizzati nella fase di training).

Una strategia per evitare l'*overfitting* consiste nell'utilizzare insiemi di campioni disgiunti nelle fasi di training e di test. Ciò misura esplicitamente la capacità di generalizzazione del classificatore. I campioni di test devono essere sufficientemente distinti da quelli di training perché questo approccio sia efficace.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Alcune delle strategie di classificazione più comunemente utilizzate sono:

- Approcci basati su regole (*rule-based*), adatti quando si hanno vocabolari di gesti statici di piccole dimensioni.
- Concetto di *massima verosimiglianza* (*maximum likelihood*), il quale può essere applicato a diversi problemi.
- *Hidden Markov Models* (*HMMs*), i quali vengono utilizzati per gesti dinamici. Inoltre gli *HMMs* vengono frequentemente utilizzati per la classificazione di vari processi dinamici, come ad esempio il parlato e il linguaggio dei segni.
- Algoritmi basati su *Artificial Neural Networks* (*ANNs*). Il loro nome e la loro struttura sono ispirati al cervello umano, imitando il modo in cui i neuroni biologici si inviano segnali. Le *ANNs* sono una categoria di modelli utilizzati nel machine learning.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Come menzionato in precedenza, la fase di classificazione delle features riceve, per ogni campione di dati di input, un *feature vector* dalla fase di estrazione delle features. Per gesti statici si ha un unico *feature vector*, mentre per gesti dinamici si ha una sequenza di *feature vector*. La scelta degli elementi del *feature vector* (che rimangono fissi nelle fasi di training e di classificazione) è un'importante decisione di progetto che può significativamente influenzare le prestazioni del sistema di riconoscimento.

Molti sistemi di riconoscimento non utilizzano tutte le features che teoricamente potrebbero essere calcolate dai dati di input. Si può effettuare una selezione di features adeguate mediante determinati criteri e algoritmi. La riduzione del numero delle features riduce il costo computazionale (tempo di calcolo, requisiti di memoria) e la complessità del sistema di riconoscimento sia nella fase di estrazione delle features sia nella fase di classificazione delle features. In alcuni casi, tale riduzione può migliorare l'accuratezza del sistema di riconoscimento eliminando features non adeguate e quindi enfatizzando le features rimanenti.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Nel processo di selezione delle features, vengono selezionate una o più features che consentono una identificazione affidabile di ogni elemento del vocabolario. Tale risultato si può ottenere se le features selezionate sono caratterizzate da un'alta *inter-gesture variance* (variano significativamente tra gesti differenti) e una bassa *intra-gesture variance* (variano poco tra più riproduzioni dello stesso gesto).

Non è richiesto che gesti differenti differiscano in ogni elemento del *feature vector*.

Al fine di determinare la *inter-gesture variance* e la *intra-gesture variance* di una feature, la feature deve essere calcolata per numerose riproduzioni di ogni gesto presente nel vocabolario. Deve essere utilizzata una composizione rappresentativa delle condizioni di acquisizione («recording conditions») permesse in modo da verificare in che modo ogni feature dipende da differenti condizioni, ad esempio condizioni di illuminazione. Inoltre, se si vuole effettuare un riconoscimento dei gesti non associato ad una persona specifica, tale procedura deve essere applicata utilizzando gesti di persone differenti. Sulla base dei campioni a disposizione, la selezione delle features può essere effettuata manualmente (utilizzando le regole

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

menzionate in precedenza) o attraverso un algoritmo automatico in grado di trovare il *feature vector* che ottimizza le prestazioni di un classificatore selezionato sui dati disponibili.

È importante tenere presente che, sebbene la selezione delle features risultante dall'analisi manuale o automatica possa essere ottimale per i dati sui quali è basata, essa potrebbe non essere ottimale per essere applicata su altri insiemi di dati. Di solito, i campioni a disposizione non possono coprire tutte le condizioni di acquisizione («recording conditions») nelle quali il sistema di riconoscimento verrà utilizzato, quindi le features adeguate per i dati disponibili potrebbero non essere adeguate per condizioni di acquisizione differenti da quelle relative ai dati disponibili.

Potrebbe quindi essere necessario riconsiderare la selezione delle features se l'accuratezza effettiva del riconoscimento è inferiore alle aspettative.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Come menzionato in precedenza, tra gli algoritmi di classificazione utilizzati ci sono gli algoritmi basati su regole (*rule-based*).

Un semplice approccio basato su regole è un insieme di regole esplicite IF-THEN riferito alle features del target. Tali regole possono richiedere ad esempio che alcune features del target appartengano ad un range tipico di un gesto specifico.

Feature	Symbol	Gesture			
		“left”	“right”	“stop”	none
Compactness	c	0.451	0.498	0.180	0.724

posizione di riposo



«stop»

Riguardo all'esempio dei gesti statici, si supponga che, da differenti acquisizioni dei gesti da riconoscere relative a diverse persone, risulti

Valore
massimo

$$\max c \approx 0.23$$

$$\bar{c} \approx 0.18$$

“stop”

$$c \gg 0.23$$

Molto
maggiore

Altri gesti

Media
aritmetica



ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Quindi per identificare il gesto statico «stop» tra gli altri gesti, si può utilizzare una semplice regola basata sulla feature compattezza c :

IF $c < 0.25$ THEN the observed gesture is "stop".

Ovviamente, con l'aumentare della dimensione del vocabolario, il numero e la complessità delle regole cresce ed esse potrebbero velocemente diventare difficili da gestire. La creazione manuale di regole appena descritta è quindi adeguata solo per vocabolari di dimensione molto piccola (come ad esempio il vocabolario associato all'esempio dei gesti statici), dove si possono facilmente individuare opportune soglie.

Esistono comunque algoritmi per l'apprendimento automatico di regole.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Per vocabolari di dimensione maggiore, le regole vengono spesso utilizzate prima dell'utilizzo di un altro classificatore. Ad esempio, la regola

IF $a < \Theta_a N^2$ THEN the object is not the hand.

scarta rumore e distrattori sullo sfondo che sono troppo piccoli per rappresentare una mano.

Un altro esempio è quello di determinare se un oggetto in una sequenza di immagini è in movimento o è fermo. La regola seguente verifica la presenza di un movimento minimo dell'oggetto tra i frame i e j :

IF $\max_{i \neq j} \sqrt{(x_{\text{cog}}(i) - x_{\text{cog}}(j))^2 + (y_{\text{cog}}(i) - y_{\text{cog}}(j))^2} < \Theta_{\text{motion}} N$
THEN the hand is idle.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Come menzionato in precedenza, tra gli algoritmi di classificazione più comunemente utilizzati ci sono gli algoritmi basati sul concetto di *massima verosimiglianza* (*maximum likelihood*).

Si ipotizzi che l'osservazione \mathbf{O} sia un vettore di K features scalari f_i :

$$\mathbf{O} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_K \end{pmatrix}$$

Il classificatore di massima verosimiglianza identifica la classe $\omega \in \Omega$ che più verosimilmente ha causato l'osservazione \mathbf{O} . Il classificatore si basa sulla massimizzazione della probabilità condizionata $P(\omega|\mathbf{O})$.

$$\hat{\omega} = \operatorname{argmax}_{\omega \in \Omega} P(\omega|\mathbf{O})$$

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

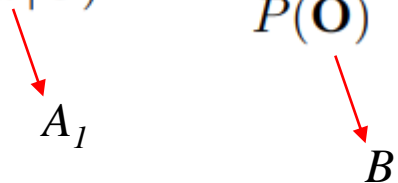
Gesti della mano – Classificazione delle features

Per risolvere tale problema di massimizzazione si può applicare la formula di Bayes.

Formula di Bayes

$$P(A_i | B) = \frac{P(A_i) P(B | A_i)}{P(B)}$$

Nel caso considerato $i = 1$.

$$P(\omega | \mathbf{O}) = \frac{P(\mathbf{O} | \omega) P(\omega)}{P(\mathbf{O})}$$


ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Per semplificare il problema, di solito si assume che le n classi sono equiprobabili:

$$P(\omega) = \frac{1}{n}$$

Quindi, poiché $P(\omega)$ e $P(\mathbf{O})$ non dipendono da ω , esse si possono non considerare quando si va a sostituire

$$P(\omega|\mathbf{O}) = \frac{P(\mathbf{O}|\omega)P(\omega)}{P(\mathbf{O})}$$

In

$$\hat{\omega} = \operatorname{argmax}_{\omega \in \Omega} P(\omega|\mathbf{O})$$

Si ottiene quindi

$$\hat{\omega} = \operatorname{argmax}_{\omega \in \Omega} P(\mathbf{O}|\omega)$$

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Mostriamo la classificazione delle features per gli esempi relativi ai gesti statici e ai gesti dinamici. Vogliamo identificare le features con un'alta *inter-gesture variance* (variano significativamente tra gesti differenti) considerando la seguente tabella per i gesti statici

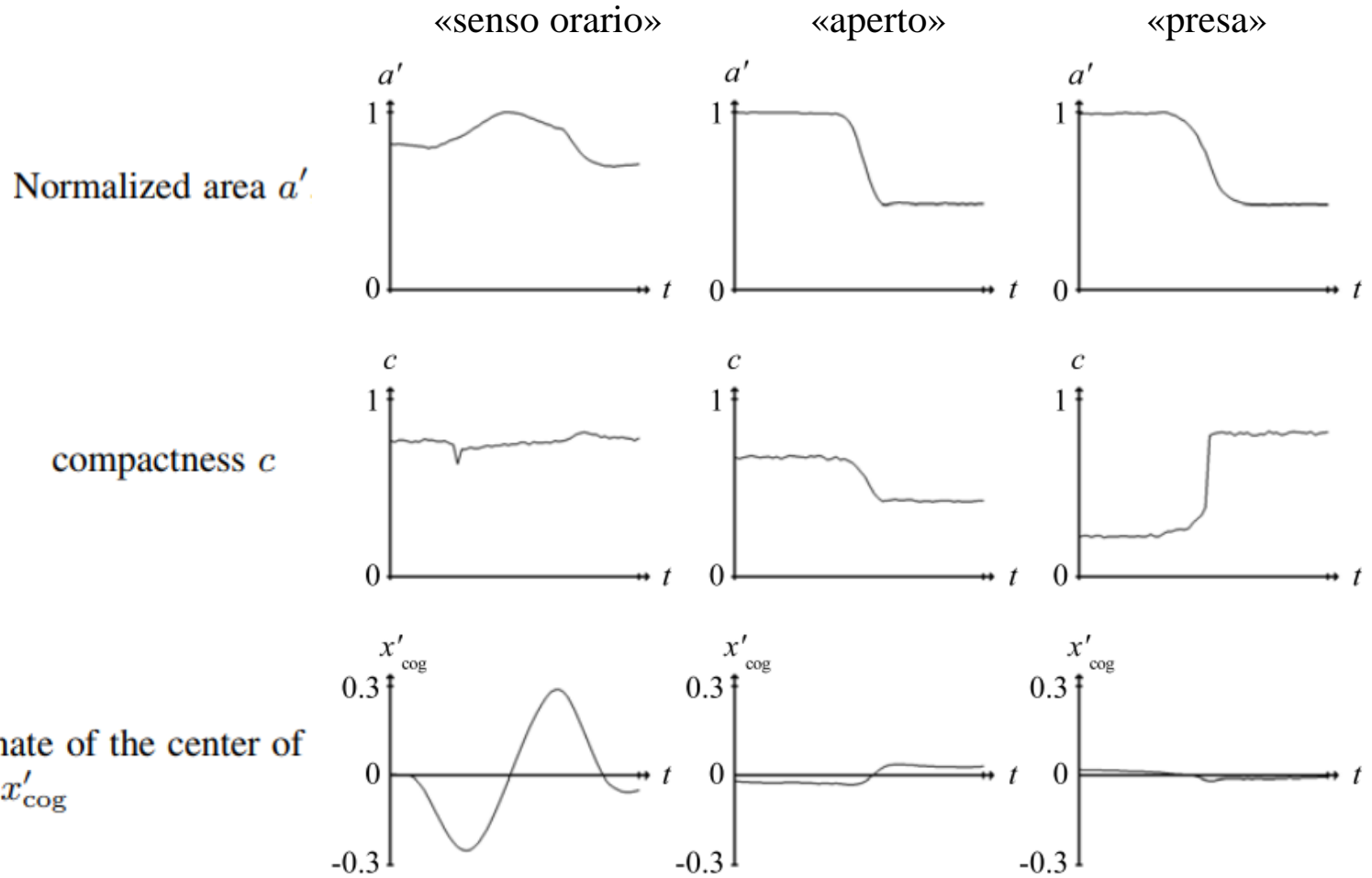
Feature	Symbol	Gesture			
		“left”	“right”	“stop”	none
Normalized Border Length	l'	1.958	1.705	3.306	1.405
Normalized Area	a'	0.138	0.115	0.156	0.114
Normalized Center of Gravity	x'_{cog}	0.491	0.560	0.544	0.479
	y'_{cog}	0.486	0.527	0.487	0.537
Eccentricity	e	1.758	1.434	1.722	2.908
Orientation	α	57.4°	147.4°	61.7°	58.7°
Compactness	c	0.451	0.498	0.180	0.724
Normalized Min./Max. Coordinates	x'_{min}	0.128	0.359	0.241	0.284
	x'_{max}	0.691	0.894	0.881	0.681
	y'_{min}	0.256	0.341	0.153	0.325
	y'_{max}	0.747	0.747	0.747	0.747
Protrusion Ratio	x_p	0.550	1.664	1.110	1.036

posizione di riposo

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

e i seguenti grafici per i gesti dinamici.



Features calcolate per i gesti dinamici (tempo $t=1, \dots, 60$)

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Poiché tutti i valori e i grafici descrivono solo un singolo gesto, devono prima essere calcolati per un certo numero di riproduzioni per avere un'idea della *intra-gesture variance* (tale processo non viene descritto esplicitamente).

Gesti statici

Feature	Symbol	Gesture			
		“left”	“right”	“stop”	none
Normalized Border Length	l'	1.958	1.705	3.306	1.405
Normalized Area	a'	0.138	0.115	0.156	0.114
Normalized Center of Gravity	x'_{cog}	0.491	0.560	0.544	0.479
	y'_{cog}	0.486	0.527	0.487	0.537
Eccentricity	e	1.758	1.434	1.722	2.908
Orientation	α	57.4°	147.4°	61.7°	58.7°
Compactness	c	0.451	0.498	0.180	0.724
Normalized Min./Max. Coordinates	x'_{min}	0.128	0.359	0.241	0.284
	x'_{max}	0.691	0.894	0.881	0.681
	y'_{min}	0.256	0.341	0.153	0.325
	y'_{max}	0.747	0.747	0.747	0.747
Protrusion Ratio	x_p	0.550	1.664	1.110	1.036

posizione di riposo

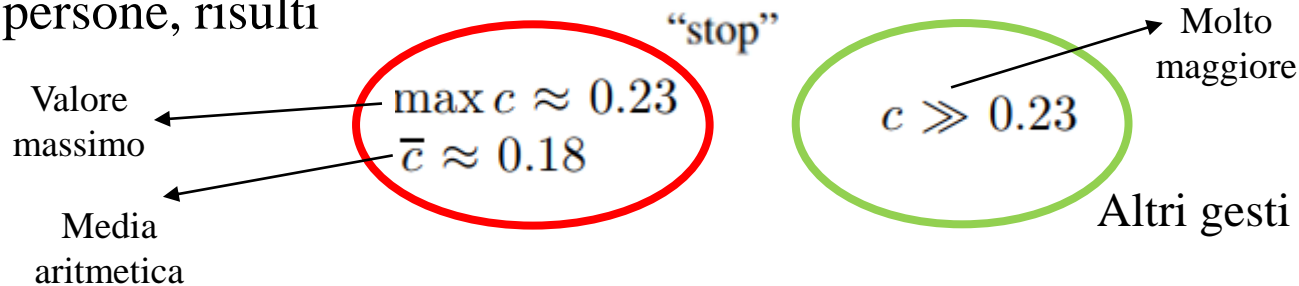
ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Feature	Symbol	Gesture			
		“left”	“right”	“stop”	none
Compactness	c	0.451	0.498	0.180	0.724

posizione di riposo

Riguardo all'esempio dei gesti statici, si supponga che, da differenti acquisizioni dei gesti da riconoscere relative a diverse persone, risulti



IF $c < 0.25$ THEN the observed gesture is “stop”.



«stop»



ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

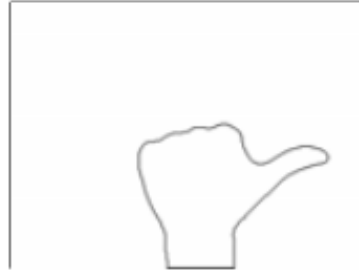
Gesti della mano – Classificazione delle features



«sinistra»



«destra»



I gesti statici di «pointing» (puntare, indicare con il dito) sono caratterizzati da una forma compatta e quasi circolare con una singola sporgenza (protrusion) riguardante il dito considerato. Per i gesti «sinistra» e «destra» tale sporgenza è lungo l'asse x . Ipotizzando che l'area di pointing non impatti troppo su x_{cog} , si può utilizzare il «protrusion ratio». Forme con $x_p \gg 1$ sono caratterizzate da pointing verso destra, mentre forme con $\frac{1}{x_p} \gg 1$ si riferiscono a pointing verso sinistra. Gesti con $x_p \approx 1$ non hanno pointing lungo l'asse x .

$$x_p = \frac{x_{\max} - x_{cog}}{x_{cog} - x_{\min}}$$

Feature	Symbol	Gesture			
		“left”	“right”	“stop”	none
Protrusion Ratio	x_p	0.550	1.664	1.110	1.036

$$x_p > 1.2 \text{ «destra»}$$

$$\frac{1}{x_p} > 1.2 \text{ «sinistra»}$$

Dominio	Condizione/i
Contenuto dell'immagine	<p>Idealmente, l'unico oggetto color pelle nell'immagine è la mano dell'utente. Altri oggetti color pelle possono essere visibili, ma devono essere piccoli rispetto alla mano. Questo vale anche per il braccio, che deve essere coperto da una camicia a maniche lunghe se è visibile. Nei gesti dinamici la mano deve essere interamente presente nell'immagine prima dell'inizio dell'acquisizione e non deve uscirne prima della fine della stessa, in modo che sia interamente visibile in ogni fotogramma acquisito.</p>
Illuminazione	<p>L'illuminazione è sufficientemente diffusa in modo che non siano visibili ombre significative sulla mano. Ombre leggere sono accettabili.</p>
Setup	<p>La distanza tra la mano e la camera è scelta in modo che la mano occupi circa il 10-25% dell'immagine. La posizione esatta della mano nell'immagine è arbitraria, ma nessuna parte della mano deve essere ritagliata. La camera non è ruotata, cioè il suo asse x è orizzontale.</p>
Camera	<p>La risoluzione può variare, ma l'immagine deve essere almeno 320×240. L'<i>aspect ratio</i> rimane costante. La camera viene regolata in modo che non si verifichino sovraesposizioni e solo una lieve alterazione del colore. Per i gesti dinamici, è necessario utilizzare una frequenza di 25 fotogrammi al secondo e la velocità dell'otturatore deve essere sufficientemente alta da evitare la sfocatura del movimento. Una camera economica è sufficiente.</p>

«Recording conditions» per le acquisizioni delle immagini negli esempi (gesti statici, gesti dinamici)

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Poiché nell'immagine ci potrebbero essere anche altri oggetti di colore simile a quello della mano, la segmentazione potrebbe restituire diverse regioni color pelle. Nell'applicazione proposta si ipotizza che la mano sia l'oggetto più grande. Le condizioni di acquisizione mostrate nella tabella della slide precedente richiedono una dimensione minima della forma uguale al 10% della dimensione dell'immagine:

$$a_{\min} = 0.1NM$$

Quindi, tutte le forme caratterizzate da $a < a_{\min}$ possono essere scartate.

Considerando le forme non scartate, può essere scelta quella con area maggiore.

Se la condizione mostrata in precedenza non è verificata per alcuna forma, l'immagine può essere considerata vuota.

Nell'esempio considerato, il *feature vector* per il riconoscimento di gesti statici è

$$\mathbf{O} = \begin{pmatrix} a \\ c \\ x_p \end{pmatrix}$$

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Le osservazioni riguardanti le tre features inserite nel *feature vector* suggeriscono di utilizzare un insieme di regole per la classificazione. Le soglie richieste possono essere facilmente identificate poiché gesti differenti (inclusa la «posizione di riposo») si sovrappongono poco nello spazio delle features.

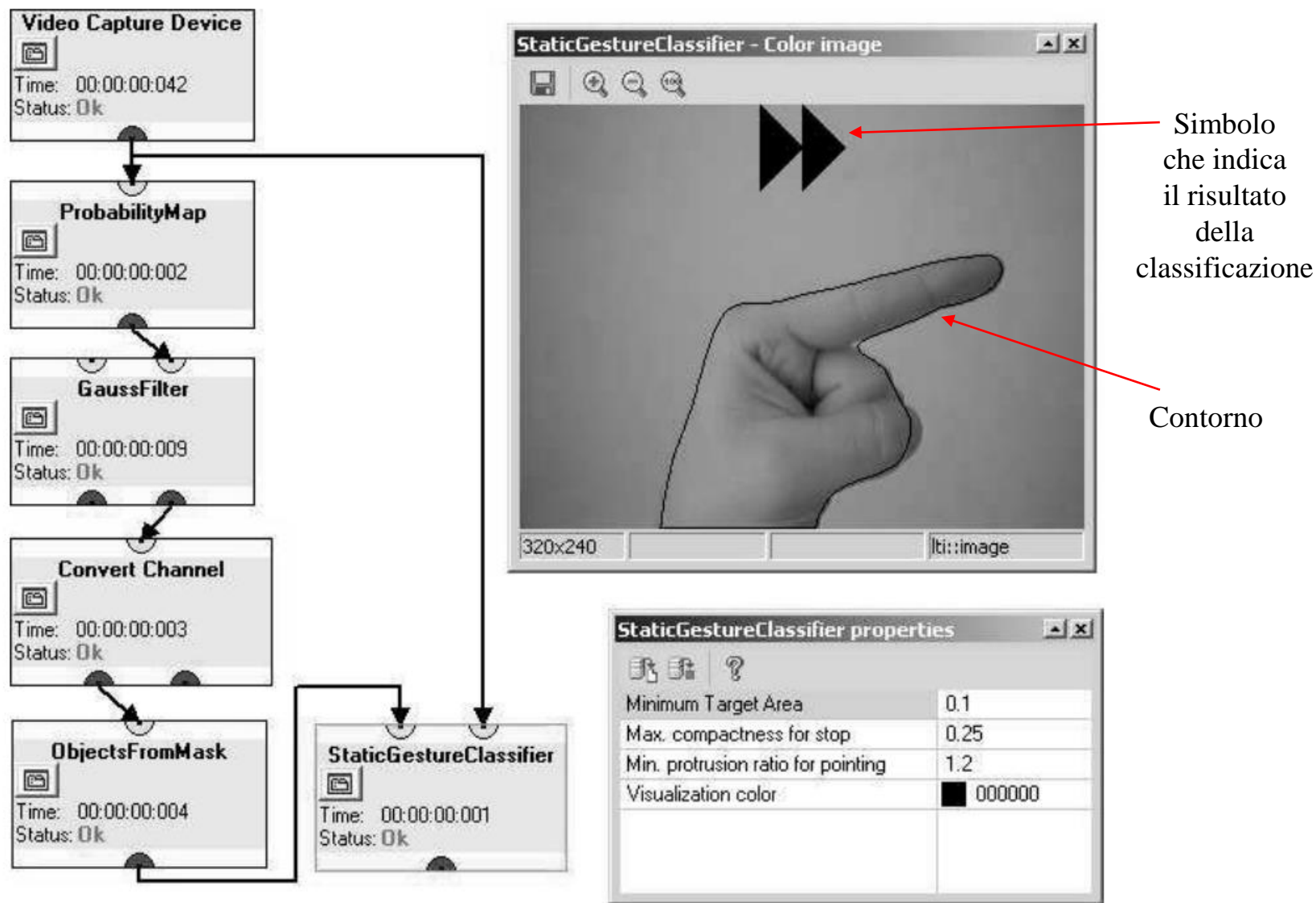
Riformulando le osservazioni precedenti, si ha

1. Discard all objects with $a < 0.1NM$.
2. If no objects remain then the image is empty.
3. Consider the object for which a is maximum.
4. If $c < 0.25$ then the observed gesture is “stop”.
5. If $x_p > 1.2$ then the observed gesture is “right”.
6. If $\frac{1}{x_p} > 1.2$ then the observed gesture is “left”.
7. Otherwise the hand is idle and does not perform any gesture.

Tali regole concludono la progettazione dell'applicazione di riconoscimento di gesti statici mostrata nell'esempio.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features



ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Video Capture Device : acquisizione di immagini da una camera. Ogni immagine viene processata individualmente e indipendentemente dalle altre.

ProbabilityMap : calcolo dell'immagine probabilità $\mathbf{I}_{\text{obj,prob}}(x, y)$.

GaussFilter : convoluzione dell'immagine probabilità $\mathbf{I}_{\text{obj,prob}}(x, y)$ con un kernel Gaussiano.

Convert Channel : l'output del passo precedente è un'immagine in scala di grigi floating point $\mathbf{I}(x, y) \in [0, 1]$. Viene quindi effettuata una conversione in un'immagine in scala di grigi fixed point $\mathbf{I}(x, y) \in \{0, 1, \dots, 255\}$.

ObjectsFromMask : estrazione di regioni contigue nell'immagine probabilità con la possibilità di utilizzare un algoritmo iterativo per il calcolo automatico della soglia Θ e utilizzando un algoritmo per il calcolo dei punti di contorno ([1]).

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

StaticGestureClassifier : tale blocco riceve la lista di regioni dal blocco precedente e scarta tutte le regioni eccetto la più grande. Viene quindi calcolato il *feature vector*

$$\mathbf{O} = \begin{pmatrix} a \\ c \\ x_p \end{pmatrix}$$

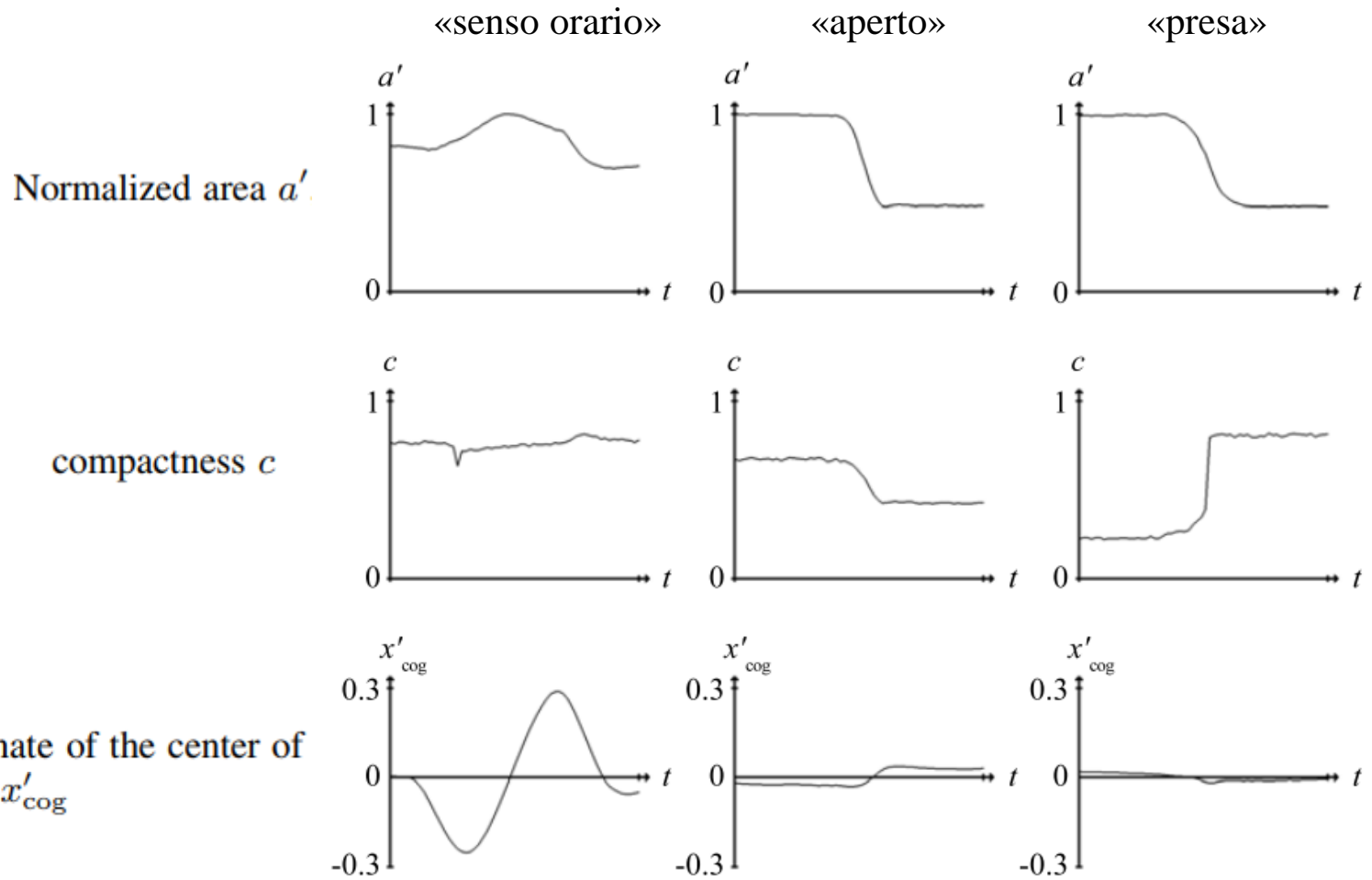
e viene effettuata la classificazione utilizzando l'insieme di regole

1. Discard all objects with $a < 0.1NM$.
2. If no objects remain then the image is empty.
3. Consider the object for which a is maximum.
4. If $c < 0.25$ then the observed gesture is “stop”.
5. If $x_p > 1.2$ then the observed gesture is “right”.
6. If $\frac{1}{x_p} > 1.2$ then the observed gesture is “left”.
7. Otherwise the hand is idle and does not perform any gesture.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Gesti dinamici



ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

Per i gesti dinamici, si può scegliere il seguente *feature vector*

$$\mathbf{o}_t = \begin{pmatrix} a' \\ c \\ x'_{\text{cog}} \end{pmatrix}$$

$$\mathbf{O} = (\mathbf{o}_1 \ \mathbf{o}_2 \ \cdots \ \mathbf{o}_T)$$

«senso orario» e «senso antiorario» possono essere identificati attraverso la forma sinusoidale della feature x'_{cog} .

«aperto» e «presa» mostrano una diminuzione della feature a' intorno all'istante $\frac{T}{2}$ ma possono essere distinti considerando la compattezza c . Infatti la compattezza diminuisce per «aperto» ma aumenta per «presa». «chiuso» e «rilascio» possono essere distinti in modo analogo.

Per processare la serie temporale di *feature vectors* ottenuta si può utilizzare un classificatore basato su *Hidden Markov Models* (HMMs).

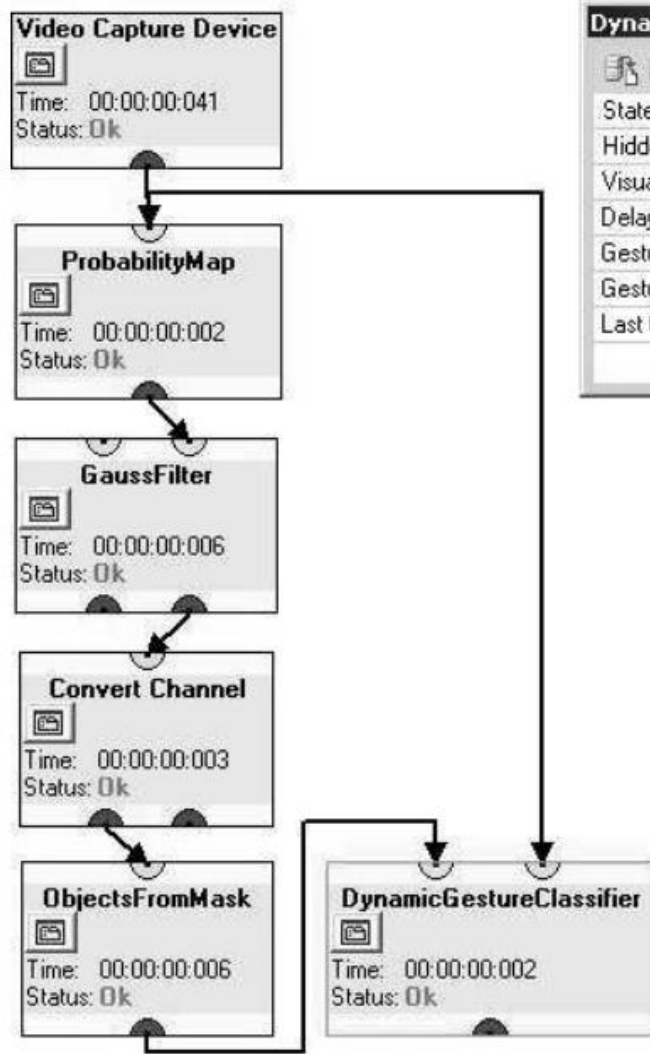
ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features

La classificazione di processi dipendenti dal tempo come i gesti, il linguaggio dei segni o il parlato suggerisce di modellizzare esplicitamente non solo la distribuzione dei campioni nello spazio delle features, ma anche la dinamica del processo da cui provengono i campioni. Questo requisito viene soddisfatto dagli *HMMs*. Gli *HMMs* derivano dalle catene di Markov.

ACQUISIZIONE NON INVASIVA DELL'AZIONE UMANA

Gesti della mano – Classificazione delle features



DynamicGestureClassifier properties

State	Idle
Hidden Markov Model	..\data\gestures.hmm
Visualization Color	000000
Delay Samples	50
Gesture Samples	60
Gesture Names	clockwise close counterclockwise drop grab open
Last Classified Gesture	ID: 3 (drop)



Riferimenti Bibliografici

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