POLITECNICO DI BARI

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

TESI DI LAUREA IN SISTEMI PER LA PROGETTAZIONE AUTOMATICA

COMPUTER VISION:

VERSO UN RICONOSCIMENTO AUTOMATICO DEI GESTI COMUNICATIVI

Relatore:
Chiar.mo Prof. MARIO REFICE

Laureando:
DONATO DI PIERRO

ANNO ACCADEMICO 2010-2011

Nota: Il presente documento è un breve estratto dello studio di tesi.

Gli obiettivi del presente studio di tesi sono:

- L'identificazione automatica dei movimenti;
- L'etichettatura automatica dei gesti;
- L'identificazione di particolari tipologie di gesti.

Definizione

Un gesto, secondo Adam Kendon, è una escursione di un arto da una posizione di riposo, e comprende il ritorno dell'arto nella posizione iniziale di riposo.

Esso consta di tre fasi:

1) **Preparation**: è la fase in cui l'arto si accinge a compiere il gesto;
2) **Stroke**: è la parte saliente del gesto;
3) **Retraction**: è la fase in cui l'arto, dopo lo stroke, ritorna verso la posizione iniziale di riposo.

Per poter condurre uno studio sul riconoscimento automatico dei gesti, è necessario in una prima fase, considerare una sola tipologia di gesti.

La tipologia di gesti presa in esame è quella dei gesti di tipo 'battito' (detti anche ***beats***). E' stata scelta questa tipologia per la possibilità che essa offre di identificare le fasi descritte da Kendon, e con lo scopo di studiare la correlazione tra i gesti ed il parlato:

- Confini delle unità prosodiche
- Prominenze accentuali.

Computer Vision: che cos'è

E' un insieme di tecniche di visione artificiale o computazionale volte a riprodurre le abilità della visione umana in dispositivi elettronici per l'acquisizione e la comprensione delle immagini.

Possibili campi applicativi della Computer Vision sono:

a) Controllo dei Processi
b) Rilevazione di eventi
c) Organizzazione delle informazioni
d) Modellazione di oggetti
e) Interazione Uomo-Macchina

Il Trattamento delle immagini digitali

Considerando la forma matriciale delle immagini digitali:

$$f(x,y)=\begin{bmatrix} f(0,\ 0) & f(0,\ 1) & \dots & f(0,\ N-1) \\ f(1,\ 0) & f(1,\ 1) & \dots & f(1,\ N-1) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ f(M-1,\ 0) & f(M-1,\ 1) & \dots & f(M-1,\ N-1) \end{bmatrix}$$

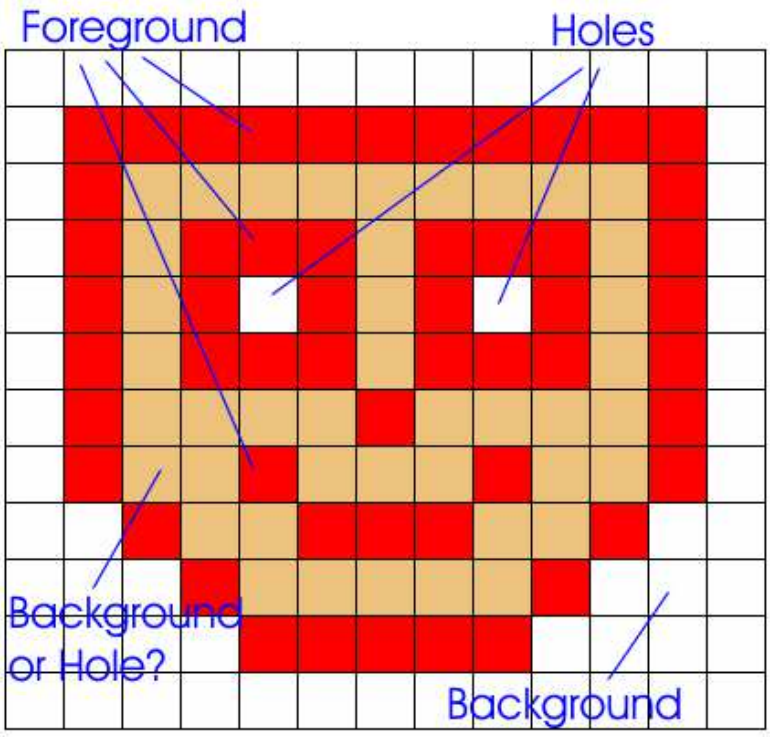
si ha che, che ognuna delle funzioni rappresentanti i valori di colore dei pixel è tale che:

$$f(x,y)\in\{0\}\cup\{1\}$$

allora l'immagine si dice binarizzata.

La binarizzazione di una immagine digitale, consente di facilitare la segmentazione dell'immagine in modo automatico.

Considerando una immagine RGB, essa è formata da pixel, ognuno dei quali è rappresentato con una intensità di colore data dalla presenza delle tre componenti di colore (rosso, verde e blu).

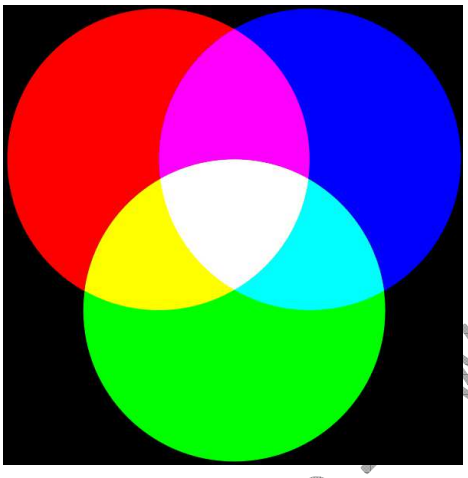


Come schematizzato nell'immagine sopra, è possibile distinguere i pixel in primo piano (indicati con Foreground), i pixel di sfondo (Background) e le lacune (Holes). In questo esempio, il colore della pelle non è detto che appartenga al primo piano, ma sorge per il sistema automatico un dubbio: si tratta di pixel di sfondo o lacune?

Gli spazi di colore

Per poter effettuare l'analisi automatica delle immagini, è necessario comprendere il dominio dei colori nella loro rappresentazione in formato analogico oltre che la loro conversione in formato digitale.

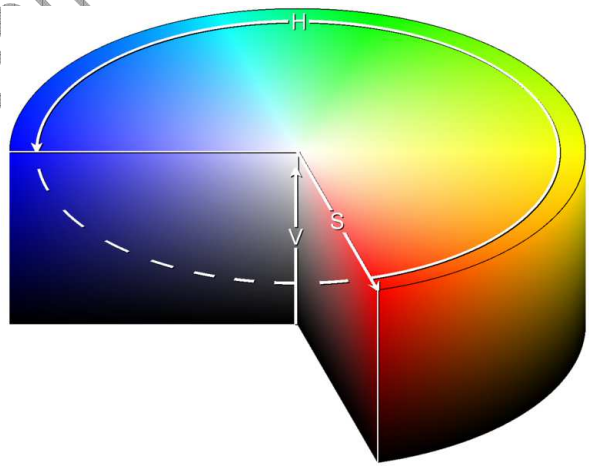
Lo spazio di colore più utilizzato è lo spazio RGB, secondo il quale, attraverso la miscelazione additiva dei colori primari Rosso, Verde e Blu, si ottengono tutti gli altri colori. In particolare, la somma delle tre componenti, porta al colore bianco.



Purtroppo, però, lo spazio di colore RGB presenta delle caratteristiche che poco si prestano all'identificazione automatica delle immagini, e cioè:

- Non è uno spazio assoluto;
- Il fattore della luminanza dipende dalle componenti R, G, B.

Uno spazio di colore molto utilizzato nella Computer Vision è lo spazio HSV, in cui il fattore della luminanza è indipendente dalla crominanza, ed in più è uno spazio assoluto (quindi ogni colore ha la medesima rappresentazione su diversi dispositivi se questi utilizzano lo spazio di colore HSV).



Dunque per poter applicare alle immagini gli algoritmi tipici della Computer Vision, si rende necessaria una conversione tra gli spazi di colore, in particolare si è proceduto convertendo le immagini da RGB ad HSV.

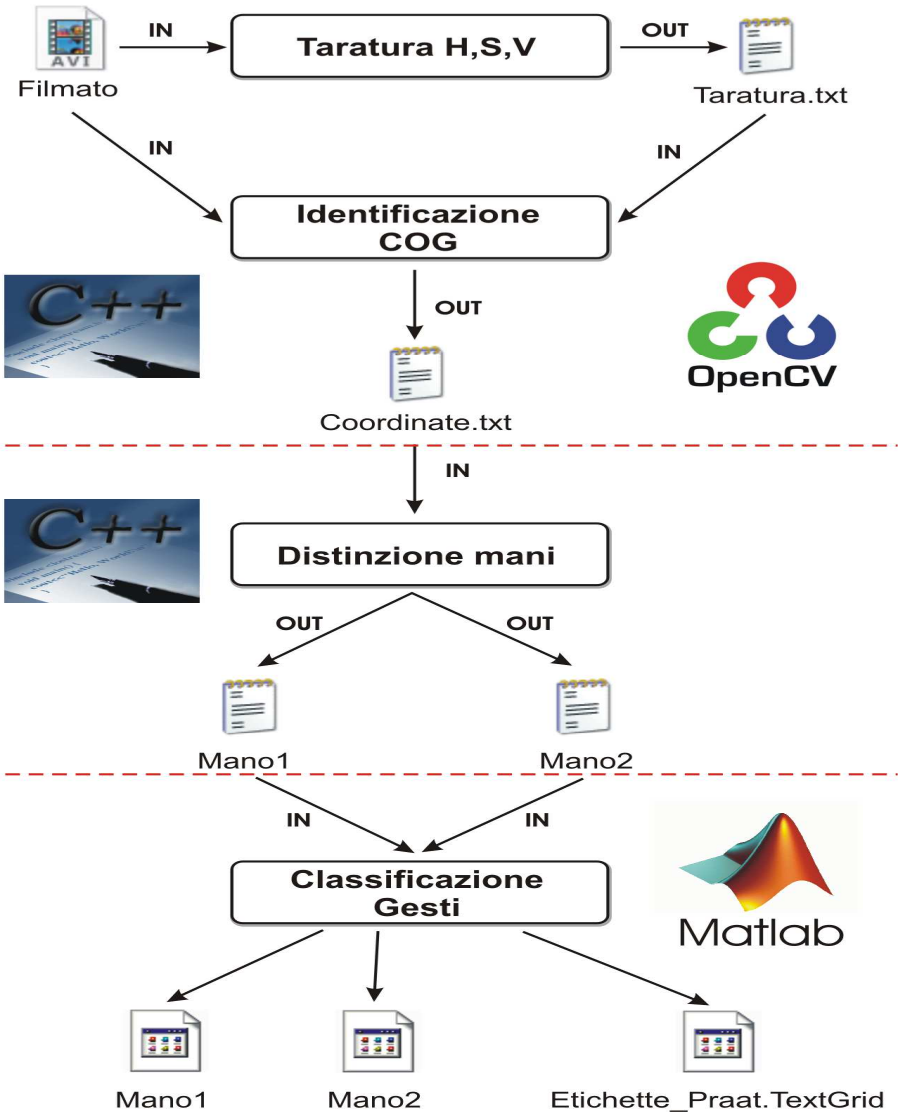
La Segmentazione delle immagini

Si tratta di una tecnica che consente di suddividere un'immagine nelle regioni o negli oggetti che la compongono (definizione di Gonzalez – Woods).

Le proprietà che contribuiscono all'operazione di segmentazione sono:

1) Discontinuità (bruschi cambiamenti di intensità da un pixel all'altro;
2) Similarità (confronto con pattern, ovvero immagini predefinite).

Visione d'insieme del sistema



Il sistema software sviluppato, è stato organizzato secondo tre moduli:

Il primo modulo, realizzato in C++ con l'ausilio delle librerie OpenCV, permette di effettuare una taratura manuale dei parametri HSV aprendo il file video da analizzare, dopodichè scorre in

automatico il filmato, identificando il centro di massa della mano destra e quello della mano sinistra del soggetto presente nel video. Durante l'identificazione, vengono registrate le coordinate della mano destra e quelle della mano sinistra, in un file di testo: coordinate.txt.

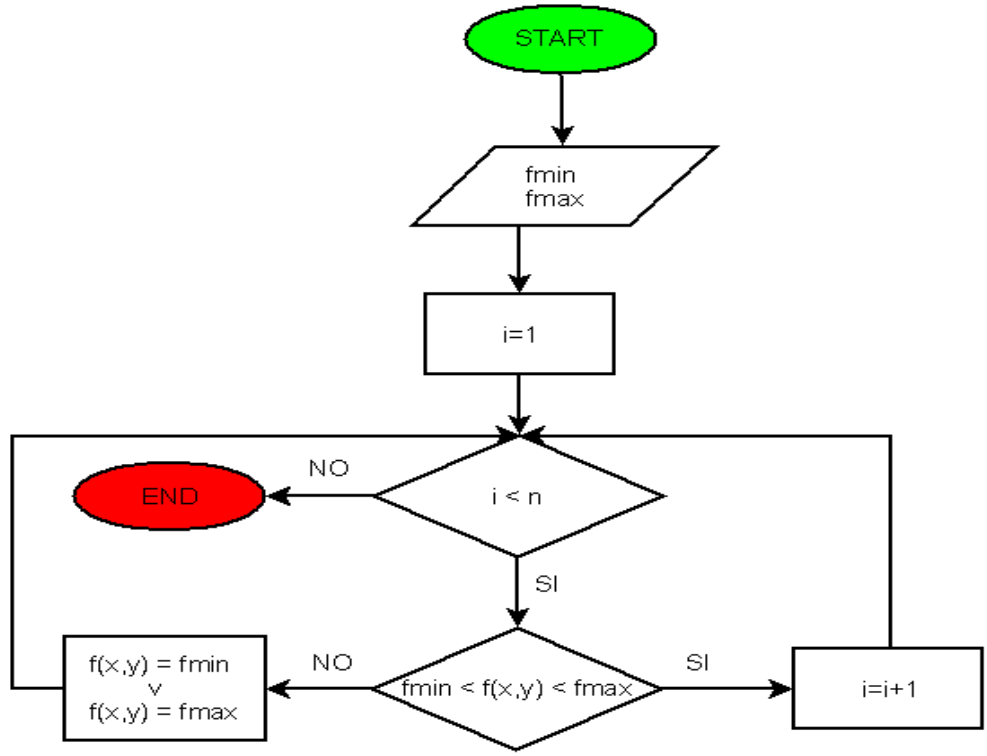
Il secondo modulo, sviluppato secondo lo standard ANSI C++, discrimina le coordinate dal file di testo iniziale, generando due nuovi file di testo, uno contenente le coordinate della mano destra e l'altro contenente le coordinate della mano sinistra.

Il terzo ed ultimo modulo, realizzato in Matlab, scorre i due file relativi alle coordinate della mano destra e della mano sinistra, effettuando il riconoscimento automatico dei gesti. Ogni gesto identificato viene registrato in un nuovo file, avente estensione TextGrid, con i rispettivi frame di inizio e di fine gesto.

In questo modo l'utente può utilizzare il file delle etichette generato, in un altro tool, Praat, in modo da poterne studiare la correlazione con il parlato.

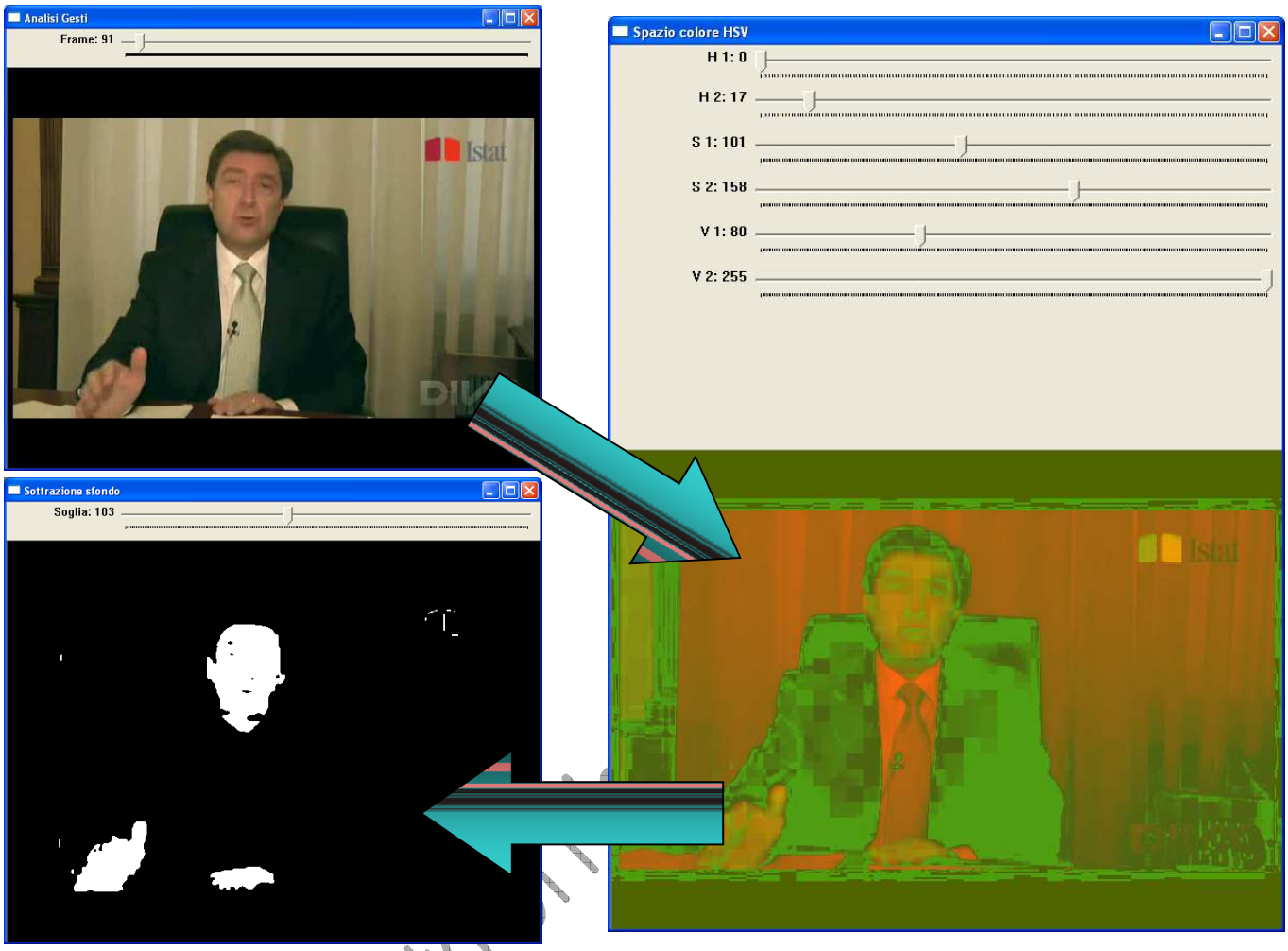
Il primo modulo

Il primo modulo effettua, tra le operazioni più importanti, quella di Hysteresis Thresholding, il cui algoritmo è riportato di seguito:



tale operazione permette di studiare qualunque tipo di filmato poiché si tratta di una operazione di taratura che viene eseguita dall'utente, fin quando non vengono eliminate dalle immagini

componenti il video da analizzare, tutte le parti non salienti (è dunque possibile effettuare una sottrazione dello sfondo, seppur con alcuni limiti).



Nell'immagine raffigurata sopra, è possibile osservare come, attraverso una opportuna taratura dei parametri HSV, sia possibile ottenere una sottrazione dello sfondo e binarizzando l'immagine, vengono evidenziate le sole aree di interesse.

Tuttavia, come si è accennato, questo è possibile solo se lo sfondo non presenta tonalità di colore simili alla pelle.

Il secondo modulo

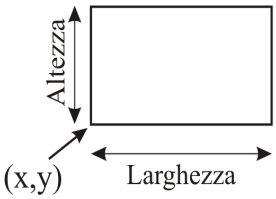
Il secondo modulo, partendo dall'immagine a colori, effettua fotogramma per fotogramma, le seguenti operazioni:

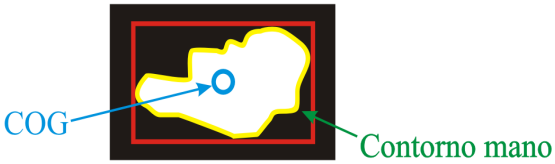
binarizza l'immagine

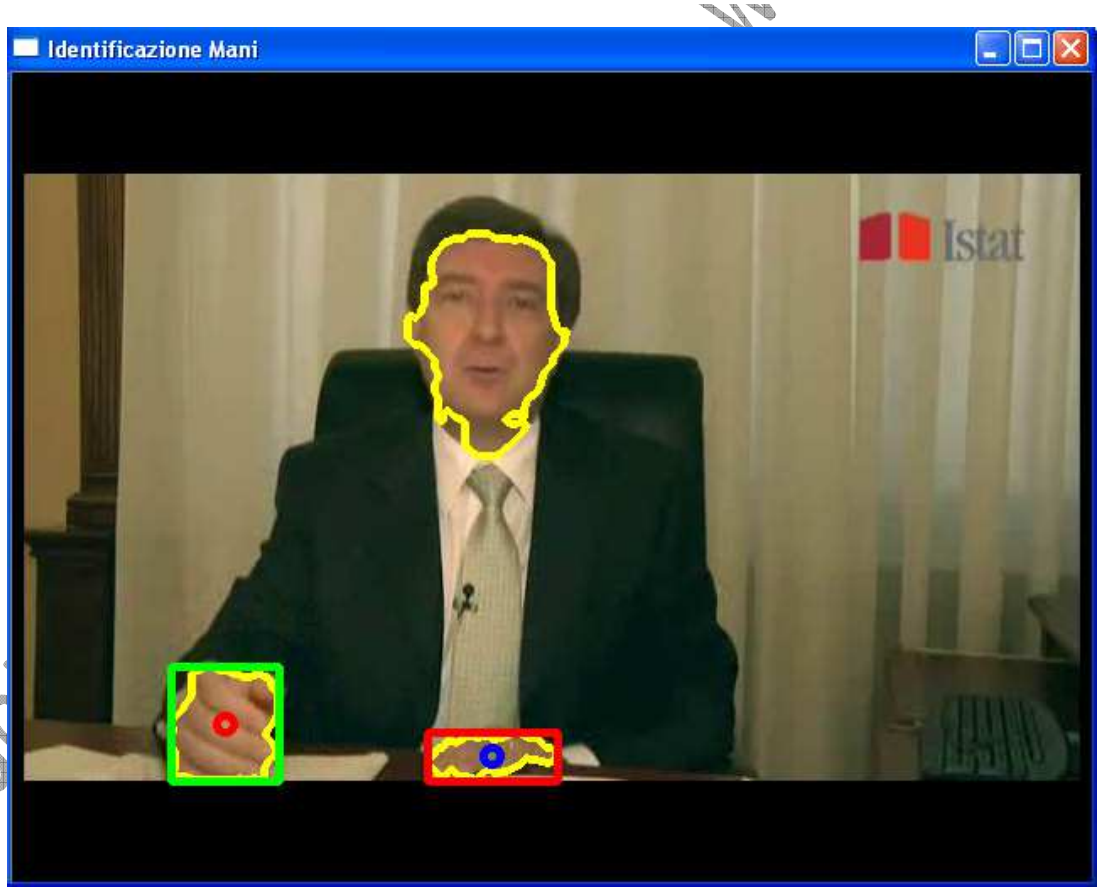
identifica le aree aventi la forma tipica di una mano

calcola il minimo rettangolo contenente tali aree

deriva il centro di gravità delle mani come il centro del rettangolo minimo contenente le aree identificate.

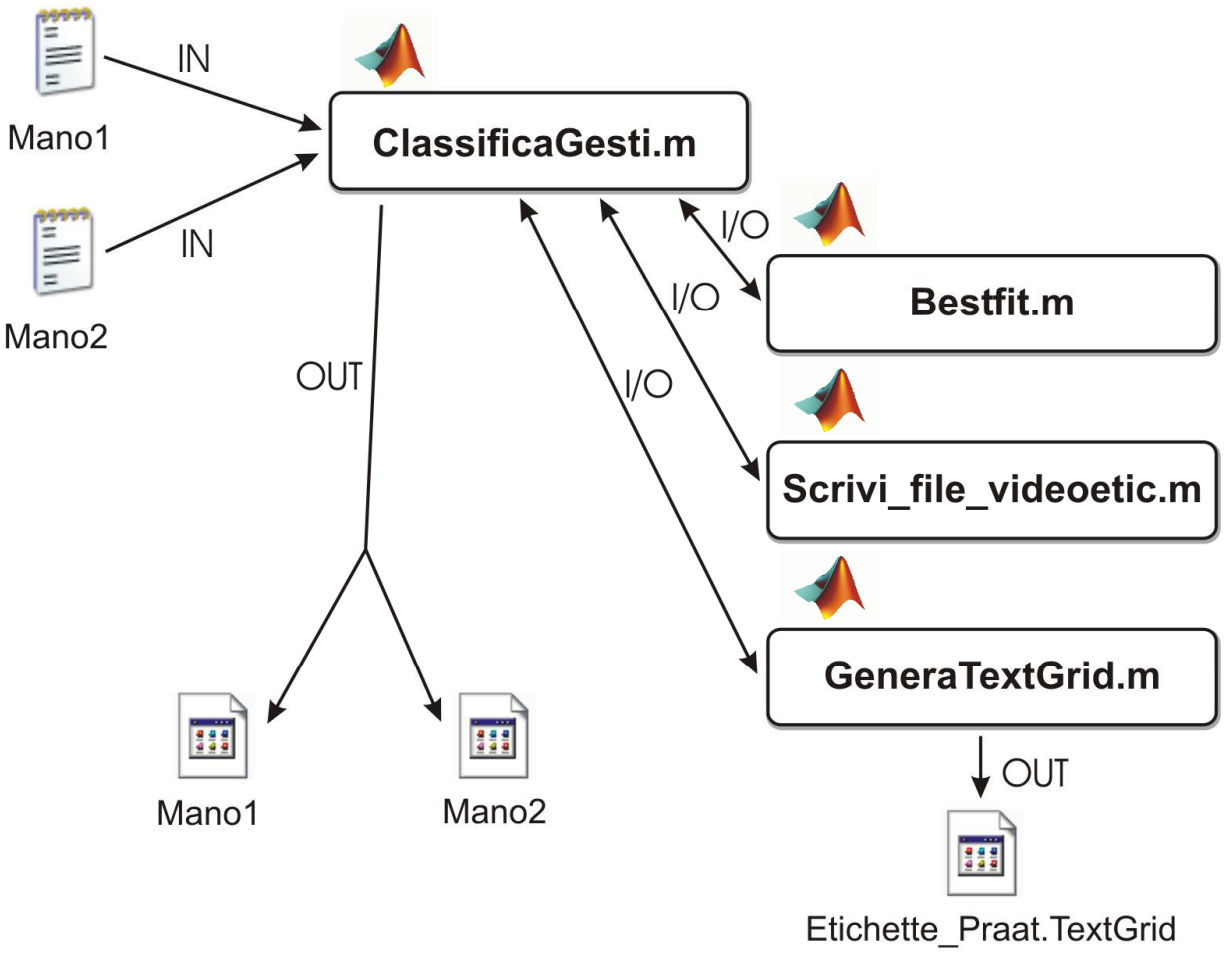






Il terzo modulo

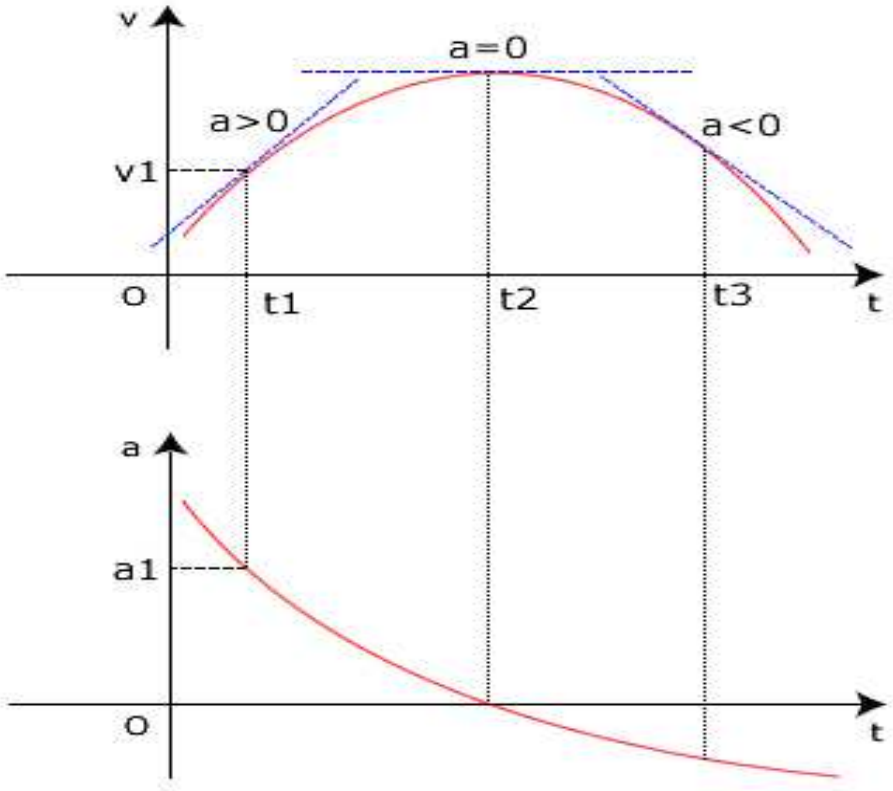
Il terzo modulo è organizzato in quattro script Matlab, uno dei quali funge da main script, come raffigurato nell'immagine seguente:



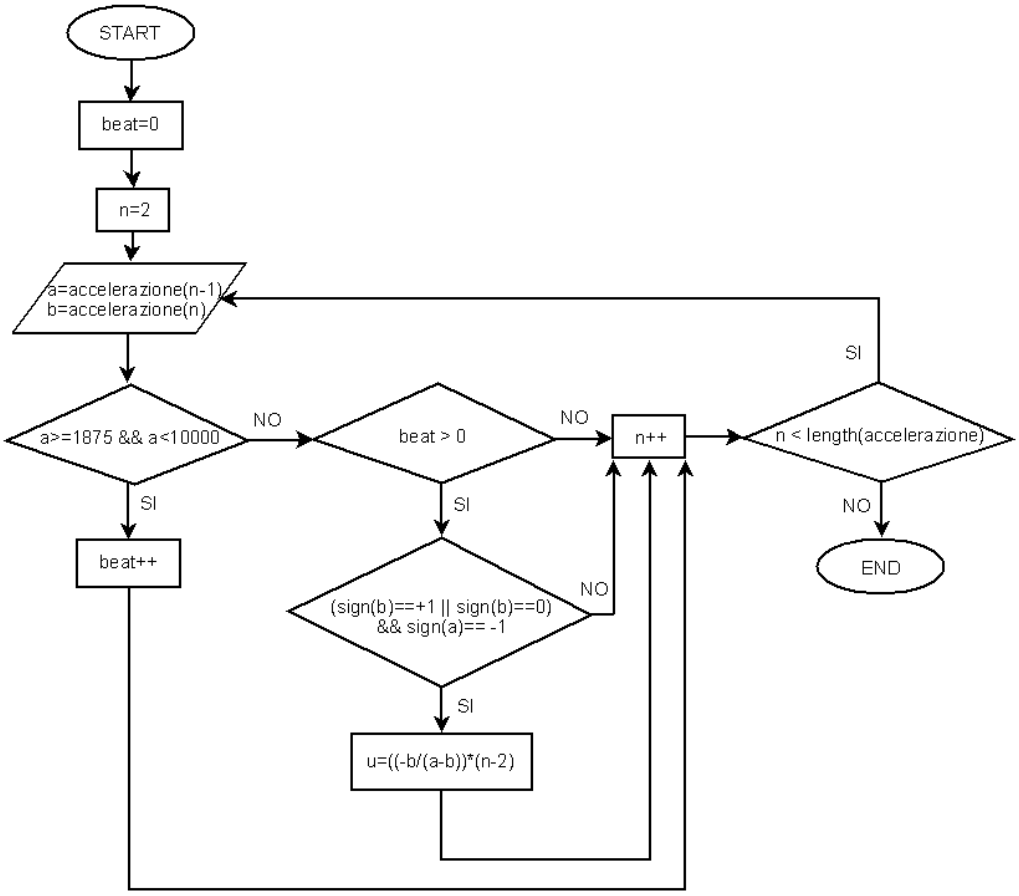
Tra i file di output, sono presenti anche due file Mano1 e Mano2, che possono essere letti con il software Videoetic, sviluppato nel laboratorio SIUM del Politecnico di Bari, e permettono all'operatore di verificare manualmente l'etichettatura automatica generata, con diretto riferimento alle porzioni di video.

Il riconoscimento

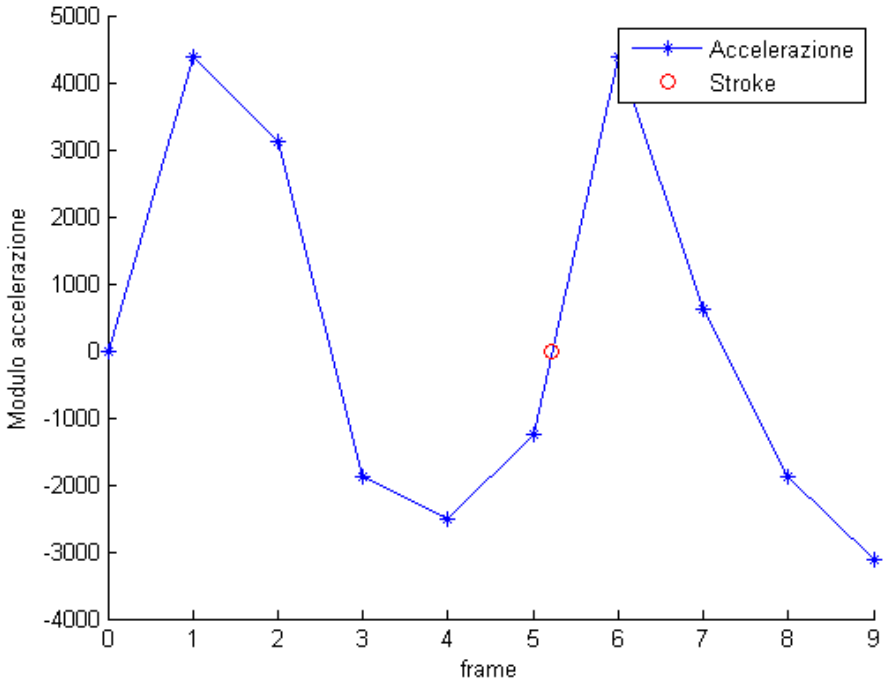
La procedura porta al riconoscimento automatico degli strokes, è basata sullo studio delle velocità e delle accelerazioni delle mani. Diagrammando le velocità in funzione delle accelerazioni, è facile osservare come lo stroke possa essere identificato con il momento in cui la mano ha accelerazione nulla.



Quindi sono stati identificati dei valori di soglia per le accelerazioni, ed è stato definito il seguente algoritmo:



Per il Teorema di Shannon sul campionamento, purtroppo la precisione di tale operazione è limitata proprio dal livello di campionamento del video digitale. Maggiore è il valore di fps del video, più accurata è la derivazione del momento in cui si verifica uno stroke.



Risultati ottenuti

Per poter avere una misura qualitativa dell'efficacia del riconoscimento automatico dei gesti, sono state valutati i gesti identificati dal sistema automatico rapportandoli ai gesti identificati correttamente.

Il grado di attendibilità è dunque stato definito come:

$$\text{Attendibilità (\%)} = \frac{\text{Gesti identificati correttamente}}{\text{Totale dei gesti identificati}} \times 100$$

Lo studio è stato effettuato valutando filmati aventi diverse caratteristiche (sfondo diverso, personaggi diversi). Ne è derivata la seguente tabella:

	Totale gesti identificati	Gesti identificati correttamente	Attendibilità del Sistema Automatico
Mano destra (video1)	**93**	**69**	**74,2%**
Mano sinistra (video1)	**151**	**136**	**90,1%**
Mano destra (video2)	**41**	**35**	**85,4%**
Mano sinistra (video2)	**16**	**10**	**62,5%**
Mano destra (video3)	**74**	**57**	**77,0%**
Mano sinistra (video3)	**29**	**13**	
Mano destra (video4)	**91**	**86**	**94,5%**
Mano sinistra (video4)	**0**	**--**	**--**

Possibili sviluppi futuri

- Riconoscimento di altre tipologie di gesti oltre i beats;
- Riconoscimento degli aspetti morfologici tipici delle mani in presenza di sfondi complessi;
- Riconoscimento dei gesti nelle tre dimensioni.