

# Head and Shoulder Profile Detector

Ilario Pierbattista

24 agosto 2015

## 1 Panoramica generale

5 Si vuole sviluppare un software in grado di riconoscere in modo efficiente e quanto più accurato il profilo umano ripreso dall'alto dal sensore di profondità di un dispositivo Kinect v2.

### 1.1 Hardware e setup

10 Il sensore di profondità del Kinect ha una risoluzione di 16bit ed una precisione intorno ai 3mm (che diminuisce con l'aumentare della distanza). È montato al soffitto di una stanza a circa tre metri da terra, ha un campo visivo di  $70^\circ \times 60^\circ$ , i quale, in relazione alla distanza dal pavimento, permette di catturare un'area di circa  $4m \times 5m$ .

Il Kinect cattura 30 frame ogni secondo. La dimensione di ogni frame è pari a  $512 \times 424$  pixel.

### 15 1.2 Struttura software

Il software è diviso in due moduli: il modulo per il riconoscimento e il modulo di allenamento.

20 Nella prima fase, quella di allenamento, il relativo modulo software, a partire da un insieme di campioni classificati, genererà un insieme di regole per classificare campioni con le medesime caratteristiche. Implementa l'algoritmo Adaboost.

La seconda fase, quella di riconoscimento, lavora su dati reali: ad ogni frame acquisito tramite il sensore di profondità del Kinect si applicano le regole di classificazione generate nella fase precedente, stabilendo se nel frame è stata ripresa una persona e - in caso affermativo - la posizione di quest'ultima.

## 25 2 Adaboost

Adaboost è un *meta algoritmo* di *machine learning* che viene utilizzato per aumentare le prestazioni di altri algoritmi di apprendimento. Ciò avviene estraendo un *classificatore forte* a partire da un insieme di *classificatori deboli*, dove per classificatore si intende una funzione  $f(x)$  che identifichi la *classe* di appartenenza dell'oggetto  $x$  in input.

30 La realtà d'interesse del problema prevede l'esistenza di due classi di oggetti: *umano* e *non umano*.

I dati in input per la fase di allenamento sono costituiti dal *dataset di allenamento*, ovvero una raccolta di oggetti preclassificati attraverso i quali Adaboost riuscirà a selezionare una combinazione di classificatori deboli che meglio approssimano la classificazione degli elementi del dataset.

## 2.1 Il dataset dall'allenamento

La finestra di visualizzazione, come accennato in precedenza è di  $512 \times 424$  pixel. In una registrazione che riprende dall'alto una persona che attraversa la stanza camminando, la figura della persona occupa una porzione della finestra di circa  $160 \times 100$  pixel.

40 Un dataset di allenamento non è altro che un insieme di porzioni di frame di una certa dimensione (la stessa per tutti) che vegono classificati manualmente dal creatore<sup>1</sup> del dataset.

La creazione del dataset è un'attività delicata. I criteri di scelta vengono trattati nella sezione 3.

## 45 2.2 Preprocessing

Gli elementi del dataset verranno preprocessati prima di poter essere dei dati di input adeguati per l'algoritmo.

Da ricordare che stiamo trattando di matrici  $160 \times 100$  (porzione del frame principale) i cui valori corrispondono al risultato della misurazione del sensore di profondità del Kinect ed è proporzionale alla distanza rilevata.

### 2.2.1 Conversione delle distanze

La prima operazione di preprocessing consiste nel trasformare tali distanze (sensore-superficie) in distanze dal pavimento della stanza. Come è facilmente intuibile

$$d' = d_{\text{pavimento}} - d$$

dove la distanza del pavimento dal sensore ( $d_{\text{pavimento}}$ ). La distanza del pavimento può essere estratta dall'immagine stessa

Questa prima operazione di preprocessing ha complessità computazionale  $\Theta(n \cdot m)$  con  $n$  ed  $m$  dimensioni della porzione del frame.

### 2.2.2 Immagine integrale

Costruire una funzione riesca a classificare un'immagine (la matrice dei valori del sensore di profondità è a tutti gli effetti considerabile come un'immagine) prevede l'utilizzo delle *Haar-like features*.

---

<sup>1</sup>Il sistema di apprendimento che si sta trattando ricade nella categoria dei sistemi di *apprendimento supervisionato*.

L'obiettivo di una feature di Haar è quello di evidenziare le *differenze d'intensità* tra due regioni rettangolari adiacenti di un'immagine. Il valore di una semplice feature a due rettangoli corrisponde alla differenza tra la somma delle intensità di tutti i pixel della prima area e la somma delle intensità di tutti i pixel della seconda. Per intensità del pixel si intende il valore che il pixel assume. Inoltre si possono considerare aree partizionate in più di due rettangoli con una semplice generalizzazione del calcolo.

$$feature = \sum Area_{White} - \sum Area_{Black}$$

60 Saranno utilizzate solamente features a due e a tre rettangoli (bandate).

Il calcolo di una feature è un'operazione computazionalmente costosa. Si introduce quindi il concetto di *immagine integrale* (altrimenti nota come *summed area table*): ogni pixel dell'immagine integrale corrisponde alla somma dei valori di tutti i pixel della regione in alto a sinistra dell'immagine. Rigorosamente, sia  $II(x, y)$  l'intensità del pixel alla posizione  $(x, y)$  dell'immagine integrale:

$$II(x, y) = \sum_{i=0}^x \sum_{j=0}^y I(i, j)$$

Calcolare la somma delle intensità di tutti i pixel in un'area rettangolare attraverso l'immagine integrale diventa un'operazione molto veloce. Infatti:

$$\sum_{i=x_1}^{x_2} \sum_{j=y_1}^{y_2} I(i, j) = II(x_2, y_2) + II(x_1, y_1) - II(x_1, y_2) - II(x_2, y_1)$$

Grazie all'immagine integrale, una feature di qualsiasi dimensione può essere calcolata in un tempo costante.

### 2.3 Estrazione del classificatore forte

65 Sia  $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$  un insieme di  $n$  coppie costituite da un'immagine  $(x_i)$  e la relativa classificazione  $(y_i \in \{0, 1\})$ . Se  $y_i = 1$ , allora  $x_i$  appartiene alla classe *umano* (la coppia  $(x_i, y_i)$  prende il nome di *esempio positivo*), altrimenti appartiene alla classe *non umano* (la coppia  $(x_i, y_i)$  prende il nome di *esempio negativo*).

L'insieme  $D$  può essere partizionato come segue:

$$P = \{(x, y) \in D | y = 1\} \text{ e } N = \{(x, y) \in D | y = 0\}$$

Si tenga a mente che, essendo  $P$  ed  $N$  partizioni di  $D$ , valgono le seguenti<sup>2</sup>:

$$D = P \cup N \tag{1}$$

$$P \cap N = \emptyset \tag{2}$$

$$\#(D) = \#(P \cup N) = \#(P) + \#(N) \tag{3}$$

70

---

<sup>2</sup>La scrittura  $\#(D)$  denota il numero di elementi dell'insieme  $D$ .

Si introduce anche il concetto di *classificatore debole*. Si tratta di una funzione che per una data immagine  $x$  in ingresso, assume il valore che simboleggia la presunta classe di appartenenza di quest'ultima. Nel dettaglio:

$$h(x) = \begin{cases} 1 & \text{se } pf(x) < p\theta \\ 0 & \text{altrimenti} \end{cases} \quad (4)$$

dove  $f(x)$  è il valore di una feature di Haar applicata all'immagine  $x$ ,  $p \in \{-1, 1\}$  è detta *polarità* e  $\theta$  è la *soglia* (*threshold*). Tutti i classificatori deboli sono costruiti usando un'unica feature.

L'obiettivo di Adaboost è quello di formare un *classificatore forte* come combinazione lineare dei migliori classificatori deboli estraibili dal set di allenamento, dove il fattore moltiplicativo di ogni classificatore nella combinazione è inversamente proporzionale agli errori di classificazione compiuti da quest'ultimo in fase di allenamento.

Il seguente algoritmo descrive la procedura di estrazione e combinazione di  $T$  classificatori deboli.

1. Si associa ad ogni elemento  $(x_i, y_i) \in D$  un peso  $w_i$  tale che  $w_i = \frac{1}{2l}$  se  $(x_i, y_i) \in P$  oppure  $w_i = \frac{1}{2m}$  se  $(x_i, y_i) \in N$ , dove  $l = \#(P)$  ed  $m = \#(N)$  (numero degli esempi positivi e numero degli esempi negativi).

Sia inoltre  $n := \#(D) = \#(P) + \#(N) = l + m$ .

2. For  $t = [1 : T]$

- (a) Si normalizzano i pesi, in modo che la loro somma sia pari ad 1:

$$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^n w_{t,j}}$$

- (b) Si estrae il miglior classificatore debole. La procedura viene esposta nel dettaglio nella sezione 2.4, ma si tenga presente che il miglior classificatore è quello il cui *errore pesato* è minimo per la corrente iterazione.

$$\epsilon_t = \min_{f,p,\theta} \left\{ \sum_{i=1}^n w_{t,i} \cdot |h(x_i, f, p, \theta) - y_i| \right\}$$

Siano inoltre  $f_t, p_t, \theta_t$  i parametri del classificatore debole che ne minimizzano l'errore pesato:

$$h_t(x) := h(x, f_t, p_t, \theta_t)$$

- (c)  $\beta_t \leftarrow \frac{\epsilon_t}{1-\epsilon_t}$

- (d) Si aggiornano i pesi

$$w_{t+1,i} \leftarrow w_{t,i} \cdot \beta_t^{e_i}$$

dove  $e_i = 1$  se  $(x_i, h_t(x_i)) \in D$  (ovvero se  $x_i$  è classificata correttamente),  $e_i = 0$  altrimenti.

(e)  $\alpha_t \leftarrow \log(\frac{1}{\beta_t})$

3. Il classificatore forte è dato da:

$$F(x) = \begin{cases} 1 & \text{se } \sum_{t=1}^T \alpha_t h_t(x) > \theta \sum_{t=1}^T \alpha_t \\ 0 & \text{altrimenti} \end{cases} \quad (5)$$

dove  $\theta \in [0, 1]$  è la soglia.

Si noti che, nell'operazione 2b, l'errore pesato non è altro che la somma dei pesi degli esempi non classificati correttamente. Infatti:

$$h(x_i, f, p, \theta) = y_i \Rightarrow |h(x_i, f, p, \theta) - y_i| = 0$$

$$h(x_i, f, p, \theta) \neq y_i \Rightarrow |h(x_i, f, p, \theta) - y_i| = 1$$

Al punto 2c, il valore di  $\beta_t$  non è altro che il rapporto tra l'errore pesato del classificatore debole e la somma dei pesi delle immagini classificate correttamente. Tale valore è chiaramente  $0 < \beta_t < 1$ .

In fase di aggiornamento dei pesi (punto 2d), i pesi relativi ad esempi classificati correttamente vengono moltiplicati per  $\beta_t$  ( $\beta_t^1 = \beta_t < 1$ ) e quindi decrementati, mentre gli altri vengono lasciati inalterati ( $\beta_t^0 = 1$ ). Fare in modo che gli esempi non classificati correttamente abbiano un peso maggiore di quelli classificati correttamente è il modo per influenzare la scelta del classificatore debole successivo che andrà a colmare le lacune del suo predecessore.

La scelta della soglia per il classificatore forte (punto 3) deve minimizzare il numero di esempi classificati in modo errato.

## 2.4 Il miglior classificatore debole

Si è detto che un classificatore debole è costruito a partire da una feature di Haar. La scelta del migliore, quindi, mira ad identificare la feature, la polarità e la soglia che minimizzano l'errore pesato di classificazione.

Si ricordi che le feature di Haar sono degli indicatori di quanto le intensità dei pixel variano da una regione della feature ad un'altra. Il classificatore debole, quindi, classificherà l'immagine a seconda che tale indice sia maggiore o minore di una certa soglia. Il compito della polarità è quello di stabilire il verso della disequaglianza.

Il pool di feature da testare è costituito - teoricamente - da tutte quelle individuabili nell'immagini di allenamento. Nell'opera di Viola e Jones vengono utilizzate immagini di allenamento di  $24 \times 24$  pixel e 5 tipologie di feature ([2], sezione 2.2): il numero di possibili features in tale area è maggiore di 160000. In questa applicazione si utilizzano immagini di  $160 \times 100$  pixel e 4 tipologie di feature: il pool è costituito da un numero di elementi maggiore di almeno 4 ordini di grandezza.

È da tener presente che moltissime di queste feature sono poco significative in questa situazione: non ha senso calcolare la variazione di intensità di due aree molto piccole con delle immagini che hanno una risoluzione tanto alta. Effettuando una prima scrematura,

si cercherà di avere un pool di feature selezionabili la cui dimensione non superi quella del pool di Viola-Jones per più di un ordine di grandezza.

125 Sia  $\{f_1, \dots, f_k\}$  l'insieme di tutte le feature selezionabili,  $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$  l'insieme degli esempi di allenamento e  $\{w_1, \dots, w_n\}$  l'insieme dei relativi pesi. La scelta del classificatore debole avviene come descritto dal seguente algoritmo in pseudocodice:

1. Si calcolano  $T^+$  e  $T^-$ , rispettivamente, somma dei pesi degli esempi negativi e di quelli negativi:

$$T^+ \leftarrow \sum_{i=1}^n (w_i y_i), \quad T^- \leftarrow \sum_{i=1}^n [w_i (1 - y_i)]$$

2. *For*  $f = [f_1, \dots, f_k]$

- (a) Si inizializza una lista di  $n$  elementi per memorizzare i valori della feature  $i$ -esima applicata ad ogni immagine di allenamento:

$$values[i] \leftarrow f(x_i) \quad \forall x_i \in D$$

- (b) Si ordinano gli elementi della lista in ordine crescente. Si tenga in conto che, dopo tale operazione, all' $i$ -esima posizione della lista non corrisponderà più il valore della feature applicata all' $i$ -esima immagine di allenamento.

- (c) Si inizializzano  $S^+$  ed  $S^-$ , con le quali, scorrendo gli elementi della lista con un cursore, indicheremo rispettivamente la somma dei pesi degli esempi positivi e di quelli negativi:  $S^+ \leftarrow 0, S^- \leftarrow 0$

- (d) *For*  $i = [1 : n]$

- i. A causa del rilocamento degli indici, alla posizione  $i$ -esima della lista corrisponderà il valore della feature dell'elemento  $x_j$  con classificazione  $y_j$  e peso  $w_j$  tale che  $(x_j, y_j) \in D$ :

$$x_j, y_j, w_j \leftarrow values[i]$$

- 135 ii. *If*  $y_i = 1$  *Then*

$$A. \quad S^+ \leftarrow S^+ + w_j$$

- iii. *Else*

$$A. \quad S^+ \leftarrow S^- + w_j$$

- iv. Si calcola l'errore pesato di classificazione:

$$e_i = \min\{S^+ + (T^- - S^-), S^- + (T^+ - S^+)\}$$

- (e) Si determinano la polarità ( $p_f$ ) e la soglia ( $\theta_f$ ) per cui l'errore pesato ( $\epsilon_f$ ) di classificazione per un classificatore che utilizza la feature  $f$  è minimo:

$$p_f, \theta_f | \epsilon_f = \min\{e_1, \dots, e_n\}$$

3. Si scelgono polarità ( $p$ ) e soglia ( $\theta$ ) finali, ovvero quelle del classificatore debole con errore pesato  $\epsilon$  minore tra tutti i classificatori possibili:

$$p, \theta | \epsilon = \min\{\epsilon_1, \dots, \epsilon_k\}$$

Il miglior classificatore debole è quindi:

$$h(x) := \begin{cases} 1 & \text{se } pf(x) < p\theta \\ 0 & \text{altrimenti} \end{cases} \quad (6)$$

140 Il metodo di identificazione della polarità e della soglia non viene riportato nell'algoritmo, essendo un passaggio che merita una trattazione a parte. Selezionare un valore di soglia vuol dire trovare il *punto che partiziona al meglio la lista dei valori della feature calcolata sulle immagini di allenamento, in modo tale da minimizzare gli errori di classificazione*. La miglior soglia di una buona feature fa in modo che la maggior parte delle  
145 immagini appartenenti alla stessa classe assumano un valore minore (o maggiore). Si può notare, come diretta conseguenza, che con una feature pessima per la classificazione non sarà possibile trovare un valore di soglia che soddisfi tale criterio.

Al punto 2(d)iv viene calcolato l'errore pesato di classificazione per una feature  $f$  con soglia  $values[i]$ <sup>3</sup>. Per essere più espliciti, bisogna effettuare una serie di osservazioni:

- 150 1.  $T^+$  ( $T^-$ ) corrisponde alla somma dei pesi degli esempi positivi (negativi);
2.  $S^+$  ( $S^-$ ) corrisponde alla somma dei pesi degli esempi positivi (negativi) dalla prima posizione fino all' $i$ -esima della lista (quella su cui è posizionato il cursore);
3.  $T^+ - S^+$  ( $T^- - S^-$ ) corrisponde alla somma dei pesi degli esempi positivi (negativi) dalla posizione  $i + 1$  della lista fino alla fine;
- 155 4. Per un classificatore della forma (6) con  $p = 1$ ,  $S^+$  corrisponde alla *somma dei pesi degli esempi classificati correttamente*, mentre  $S^- + (T^+ - S^+)$  corrisponde alla somma dei pesi degli esempi classificati in modo scorretto;
5. Per l'osservazione 4 un classificatore (6) con  $p = 1$ , la quantità  $S^- + (T^+ - S^+)$  è *l'errore pesato*;
- 160 6. Analogamente alle osservazioni 4 e 5, un classificatore (6) con  $p = -1$  commetterà un errore pesato pari alla quantità  $S^+ + (T^- - S^-)$ .

Grazie alle osservazioni 5 e 6, dal semplice calcolo dell'errore di classificazione pesato, si ottiene anche il valore della polarità:

$$p = \begin{cases} 1 & \text{se } S^- + (T^+ - S^+) < S^+ + (T^- - S^-) \\ -1 & \text{altrimenti} \end{cases} \quad (7)$$

---

<sup>3</sup>I possibili valori delle soglie corrispondono esattamente ai valori della feature calcolata sugli esempi di allenamento

In definitiva, con uno scorrimento della lista ordinata si ottengo i parametri per la  
165 costruzione del classificatore debole. La complessità di tale operazione è fortemente  
legata all'algoritmo di ordinamento della lista, la quale ha  $\Theta(n \log n)$  come limite teorico  
inferiore ([3], p. 167). Nell'implementazione è stato scelto proprio un algoritmo che  
avesse complessità  $\Theta(n \log n)$  nel caso peggiore. Ripetendo queste operazioni per ognuna  
delle feature selezionabili, si ottiene che l'algoritmo di selezione del miglior classificatore  
170 debole ha complessità  $\Theta(kn \log n)$ .

### 3 Considerazioni sulla costruzione dei dataset

La descrizione dell'algoritmo di allenamento dovrebbe far intuire quanto la qualità del  
dataset di allenamento sia importante per ottenere dei buoni risultati. Quest'ultimo,  
infatti, costituisce un'universo di campioni dal quale è possibile estrarre un classificatore  
175 forte.

Un campione dovrebbe avere le seguenti caratteristiche:

1. dall'immagine del profilo ci dovrebbe essere uno spazio vuoto di fronte e dietro di  
esso;
2. al lato di ciascuna spalla ci dovrebbe essere uno spazio vuoto;
- 180 3. tra la testa e le spalle ci deve essere una differenza d'altezza.

Poichè la forma del profilo umano in una registrazione dall'alto è mutevole a seconda  
della *direzione*, della *corporatura* e della *postura* dell'individuo, una tale varietà di forme  
potrebbe portare a dei classificatori che commetterebbero troppi errori in un caso di  
riconoscimento reale.

185 **Postura** Per valutare un classificatore forte bisogna sempre tenere conto del contesto  
applicativo: un dataset di allenamento per un'applicazione di riconoscimento in  
un ambiente assistivo, ritrarrà individui con una postura diversa da altri ambienti.  
La postura è quindi un parametro legato al contesto applicativo e tutti i campioni  
del dataset dovrebbero essere simili tra loro sotto tale aspetto.

190 **Corporatura** A differenza dalla postura, la corporatura è un parametro che sfugge alla  
classificazione in un singolo contesto. I campioni del dataset dovrebbero ritrarre  
individui di corporatura diversa.

**Direzione** Anche la direzione che l'individuo ha rispetto alla stanza è un parametro  
che non può essere legato al contesto, d'altra parte un dataset in cui i campioni  
195 abbiano tutti direzioni diverse non può fornire un classificatore affidabile.

La soluzione al problema della direzione dei campioni consiste nell'utilizzare più  
classificatori forti per il riconoscimento. In prima approssimazione si possono individuare  
2 direzioni, quella perpendicolare alla larghezza della finestra di registrazione e quella



pendicolare all'altezza. Nel primo caso il profilo umano apparirà orizzontale, nel  
 200 secondo verticale.

Costruendo un dataset di campioni in posizione un posizione orizzontale ed uno  
 un posizione verticale, si possono estrarre i classificatori  $F_{hor}(x)$  ed  $F_{ver}(x)$ . A questo  
 punto, si possono combinare i due in un classificatore unico che riesca a riconoscere  
 correttamente individui che, in casi reali, possono assumere entrambe le direzioni. Con  
 un piccolo abuso di notazione:

$$F(x) := F_{hor}(x) \vee F_{ver}(x)$$

Allo stesso modo possono essere costruiti classificatori ancora più stringenti considerando  
 altre direzioni intermedie, quindi combinandoli in parallelo.

Un ulteriore problema che merita di essere affrontato è quello della *distorsione pro-*  
*spettica* del sensore, che muta la forma del profilo umano a seconda della sua posizione  
 205 nella stanza. Purtroppo questa è una limitazione hardware del sensore che non può  
 essere eliminata a meno operazioni di preprocessing più potenti come l'estrazione del  
*point cloud* dal frame di profondità (e quindi formulando anche un nuovo algoritmo di  
 allenamento).

Sarebbe possibile invece, sulla falsa riga di quanto concluso per il problema della  
 210 direzione, creare più classificatori, ognuno dei quali opera in una regione della finestra  
 ben specifica, una regione che è caratterizzata da una particolare distorsione.

$$\begin{cases} F_1(x) = F_{1,hor}(x) \vee F_{1,ver}(x) \\ \dots \\ F_p(x) = F_{p,hor}(x) \vee F_{p,ver}(x) \end{cases} \quad (8)$$

## Riferimenti bibliografici

- [1] Lei Zhu, Kin-Hong Wong. *Human Tracking and Counting Using the KINECT Range  
 Sensor Based on Adaboost and Kalman Filter*. Advances in Visual Computing, 9th  
 215 International Symposium, ISVC 2013, Rethymnon, Crete, Greece, July 29-31, 2013.  
 Proceedings, Part II, pp 582-591.
- [2] Paul Viola, Michael J. Jones. *Robust Real-Time Face Detection*. International Journal  
 of Computer Vision 57(2), pp 137-154, 2004.
- [3] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein. *Introduction to Algorithms*. The  
 220 MIT Press, second edition, 2001.
- [4] Yoav Freund, Robert E. Schapire. *A Decision-Theoretic Generalization of On-Line  
 Learning and an Application to Boosting*. Journal of Computer and System Science  
 55, pp 119-139, 1997.