Mood Lamp - Un sistema di interazione naturale basato sul riconoscimento dell'espressione facciale

Vittorio Cuculo
vittorio.cuculo@studenti.unimi.it

24 marzo 2013

Sommario

Il progetto Mood Lamp è un sistema che si avvale dell'espressione facciale per comunicare ad una lampada a colori RGB il proprio stato emotivo, e ricevere da questa una risposta coerente con lo stato emotivo inferito tramite la variazione di tonalità della luce emessa.

1 Introduzione

L'obiettivo dei sistemi che simulano interazioni naturali è quello di rimuovere la mediazione tra macchina e umano tipica delle interfacce classiche. Tra le principali modalità di interazione naturale ci sono il parlato, i gesti, il tracciamento dello sguardo e le espressioni facciali. Quest'ultima modalità è particolarmente rilevante perché ricopre un ruolo fondamentale anche nella comunicazione non verbale tra gli esseri umani.

Nell'ambito dell'interazione Uomo-Macchina, la capacità di riconoscere e sintetizzare espressioni facciali permette alla macchina di acquisire maggiori capacità comunicative, interpretando da un lato le emozioni presenti sul volto di un soggetto e dall'altro traducendo il proprio intento comunicativo per mezzo di un output, quale un movimento, una risposta sonora o una variazione di colore.

Il progetto Mood Lamp si pone l'obiettivo di realizzare un sistema che si avvale dell'interazione naturale per comunicare ad un oggetto fisico il proprio stato emotivo e ricevere da questa una risposta coerente. In particolare, per mezzo della propria espressione facciale si comunica ad una lampada a colori RGB il proprio stato emotivo. La lampada, a questo punto, risponderà all'interazione cambiando il colore della luce emessa coerentemente con lo stato emotivo inferito.

2 Analisi dello stato dell'arte

Come anticipato nell'introduzione, le espressioni facciali giocano un ruolo molto importante nell'interazione naturale e la loro analisi ha rappresentato da sempre

una sfida complessa.

Esse sono state oggetto di studio fin dal 1872, anno in cui Charles Darwin pubblica il libro *The Expression of the Emotions in Man and Animals*[1], nel quale esprime la sua tesi sull'universalità delle emozioni come risultato del processo evolutivo, e considerando le espressioni facciali come residuo di un comportamento, secondo il principio dei "serviceable habits".

Questi studi sono stati ripresi, circa un secolo dopo, da Paul Ekman e Wallace Friesen[2] i quali hanno dimostrato l'esistenza di un insieme di manifestazioni emotive di base, trasversali all'interno dell'umanità ed indipendenti dal contesto socio-culturale di appartenenza.

Secondo il sistema Facial Action Coding System (FACS)[3], sviluppato da Ekman, le manifestazioni emotive avvengono tramite l'attivazione di una serie di muscoli facciali che sono descritti dalle action units (AU). Questo sistema di codifica permette di descrivere circa 7000 espressioni riscontrabili su un volto umano tramite la combinazione di più AU.

Un approccio simile, ma con diversa motivazione, è stato realizzato da MPEG-4, che identifica un insieme di Face Animation Parameters (FAP)[4], ognuno dei quali corrisponde allo spostamento di un sottoinsieme degli 84 Feature Points (FP) presenti sul volto. Questi spostamenti sono misurati in FAP Units, ovvero la distanza tra i punti fiduciali del volto.

Per quanto riguarda la modellazione delle espressioni facciali generate da emozioni, Ekman individua sei emozioni di base come categorie discrete, quali: rabbia, disgusto, paura, felicità, tristezza e sorpresa.

Negli stessi anni vengono sviluppate ulteriori ipotesi in contrapposizione al modello discreto proposto da Ekman, tra questi c'è quello dimensionale di James Russell, che presenta il *circumplex model* [5], ovvero uno spazio circolare bidimensionale sul quale individuare gli stati emotivi di una persona. Lungo l'asse verticale di questa circonferenza sono indicati i valori di arousal, mentre su quella orizzontale i valori di valence.

Sempre nel 1980 Albert Mehrabian aggiunge una ulteriore dimensione al modello di Russel e realizza lo spazio Pleasure-Arousal-Dominance (PAD)[6].

Ulteriori studi sono stati effettuati da Mehrabian e Valdez[7] circa gli effetti del colore sulle emozioni, utilizzando il modello a tre dimensioni PAD.

3 Modello teorico

La parte più impegnativa nella realizzazione del progetto Mood Lamp consiste nell'implementazione del sistema di riconoscimento delle espressioni facciali. Per fare questo è necessario dividere il problema in un numero finito di sottoproblemi, e per ognuno di essi realizzare un sottosistema che opera in maniera sequenziale con gli altri.

L'approccio comunemente usato è quello di partire da un processo di face detection per passare ad uno di feature extraction ed infine ad una fase di recognition.

Affinché l'interazione avvenga nel modo più naturale possibile, il sistema deve necessariamente lavorare in tempo reale ed è dunque impossibile utilizzare tecniche computazionalmente costose per analizzare l'espressione di un volto, ciò renderebbe il sistema lento causando una perdita di qualità nell'interazione.

A tale scopo, per realizzare la prima fase, quella di face detection, è stato utilizzato l'algoritmo di Viola e Jones, approccio consolidato e presente nelle principali librerie di computer vision come OpenCV.

Una volta individuato il volto all'interno dell'immagine, si passa all'estrazione dei punti fiduciali del volto tramite l'utilizzo di una Active Shape Model.

La fase di riconoscimento dell'espressione, infine, è affidata ad un semplice processo di regressione lineare, che a partire dalle distanze tra i punti fiduciali restituisce in output una tripla di valori di pleasure, arousal e valence.

3.1 Algoritmo di Viola e Jones

L'algoritmo di face detection è basato sul lavoro di Viola e Jones [8] nel quale le immagini sono identificate da un insieme di 3 tipi feature di Haar, mostrate in Figura 1.

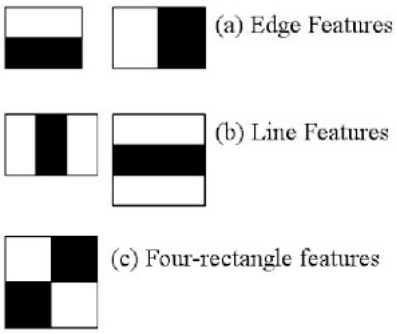

Figura 1: Feature di Haar.

Il valore della feature su una immagine è data dalla differenza tra la somma dei pixel nel rettangolo bianco e la somma dei pixel nel rettangolo nero. Per ridurre il costo computazionale di questa operazione, si utilizzano le immagini integrali, nelle quali il valore di ogni pixel è dato dalla somma di tutti i pixel precedenti, formalmente si ha che

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y') \tag{1}$$

In questo modo nelle operazioni di sottrazione è sufficiente considerare solo un pixel per ogni rettangolo.

Una volta ridotta l'immagine ad un insieme di feature, viene utilizzata una versione modificata dell'algoritmo di Adaboost che si occupa di scegliere le fea-

ture maggiormente discriminanti che separano le immagini di volti da quelle di non-volti.

L'algoritmo di apprendimento è costituito da una cascata di classificatori dove ad ogni passo aumenta il numero di feature analizzate e si procede al passo successivo solo se si ottiene una risposta positiva dal classificatore corrente.

3.2 Active Shape Model

Gli Active Shape Model (ASM)[9] rappresentano modelli parametrici deformabili realizzati mediante l'utilizzo di dataset contenenti immagini etichettate in una serie di punti di controllo, chiamati landmarks, sui quali vengono definiti dei vincoli riguardo le deformazioni ammissibili.

La fase di valutazione viene eseguita calcolando in modo automatico le posizioni iniziali dei landmarks nell'immagine, per poi modificarle iterativamente variando i parametri entro i vincoli imposti dallo shape, utilizzando per ogni landmark il valore che migliora il matching globale tra il modello e l'immagine.

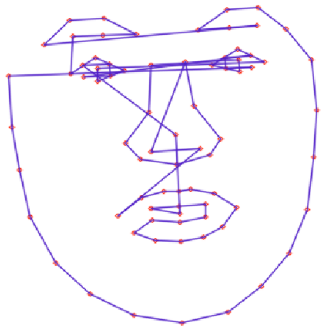

Figura 2: Esempio di modello di un volto con 76 landmarks.

3.3 Regressione lineare

La regressione è un metodo di apprendimento supervisionato con output continuo nello spazio R. Questo metodo viene utilizzato per individuare una relazione funzionale $y = f(x)$ tra due valori x e y sulla base della conoscenza di esse in un numero finito di casi, tale che y non sia troppo diverso dal valore di output reale t associato a x.

Non conoscendo la funzione f, si assume essere di tipo lineare, ovvero che i dati siano modellabili con la retta

$$f(x; w_0, w_1) = w_1 x + w_0 \tag{2}$$

dove x è la variabile indipendente (o regressore), w_0 è l'intercetta della retta di regressione e w_1 il suo coefficiente angolare.

L'obiettivo è quello di minimizzare una funzione di costo scelta al variare dei parametri w_0 e w_1. Dunque, cercando gli zeri delle derivate della funzione costo rispetto ai due coefficienti, si ottiene

$$\begin{aligned} \hat{w_0} &= \overline{t_n} - w_1 \overline{x} \\ \hat{w_1} &= \frac{\overline{x_n t_n} - \overline{t}\,\overline{x}}{\overline{x^2} - \overline{x}^2} \end{aligned} \tag{3}$$

dove $\overline{t_n}$ e $\overline{x}$ sono rispettivamente le medie aritmetiche degli t_i e degli x_i di training.

Con questi valori si w_0 e w_1 è possibile effettuare la predizione sugli x di test effettuando il fitting della retta di Eq.(2).

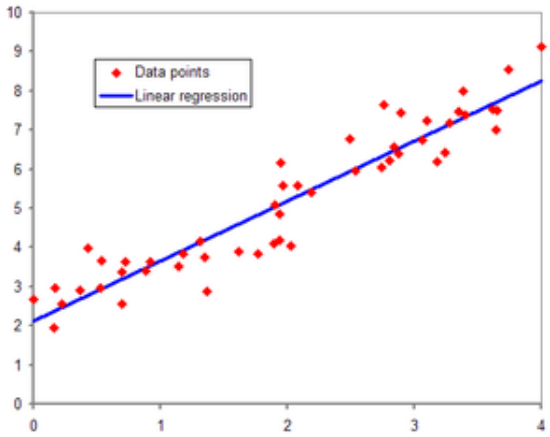

Figura 3: Esempio di retta di regressione.

Nei casi in cui la variabile t dipende da più di una variabile x, il modello si dice di *regressione lineare multipla* e si esprime in notazione matriciale come

$$t = Xw$$

dove, moltiplicando per la matrice X trasposta entrambi i termini, si ottiene

$$X^T t = (X^T X)w$$

ovvero

$$w = (X^T X)^{-1} X^T t$$

che corrisponde al prodotto tra la pseudoinversa di X e il vettore t.

4 Simulazione e esperimenti

4.1 Dataset

MUCT http://www.milbo.org/muct/

Il dataset MUCT[10] consiste in 3755 immagini di volti annotati manualmente con 76 landmarks corrispondenti ai punti fiduciali collocati nei contorni del volto, della bocca, del naso, degli occhi e delle sopracciglia, come mostrato in Figura 4. Questo dataset è stato utilizzato nel progetto per la costruzione dell'ASM.

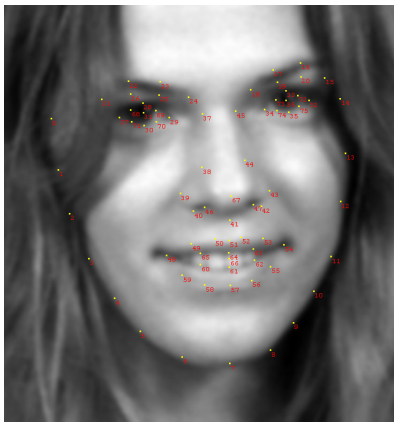

Figura 4: Localizzazione dei 76 landmarks del dataset MUCT.

JAFFE http://www.kasrl.org/jaffe.html

Il dataset JAFFE[11] consiste in 213 immagini di 10 donne giapponesi che esprimono 7 espressioni facciali (le 6 espressioni di base + 1 neutrale). Un sottoinsieme di questo dataset è stato utilizzato per la fase di training nella regressione lineare.

Figura 5: Esempio di immagini contenute nel dataset JAFFE.

4.2 Architettura del sistema

Il progetto Mood Lamp è riassumibile dal framework in Figura 6, esso è costituito principalmente da due componenti: un sottosistema di riconoscimento delle espressioni facciali (FER) ed un sottosistema di interfacciamento con l'oggetto fisico, ovvero la lampada.

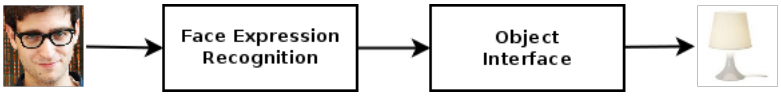

Figura 6: Framework generale del sistema.

Face Expression Recognition System Come anticipato nella Sezione 3, il sottosistema che si occupa del riconoscimento delle espressioni facciali è costituito da tre moduli che operano in tempo reale ed in maniera sequenziale, riassunti in Figura 7

Figura 7: Sistema di riconoscimento dell'espressione facciale.

Il primo modulo, quello di Face Detection, riceve in input un'immagine catturata dalla webcam e, se presente, ne estrae la sottoregione (ROI) in cui viene riconosciuto un volto umano.

La sottoimmagine viene passata al modulo che, adattando il modello dei punti fiduciali, si occupa di individuare le posizioni dei 76 landmarks.

A questo punto, il modulo che realizza la regressione costruisce la matrice delle feature calcolando le distanze tra i punti fiduciali ed effettua il fitting con la matrice dei parametri creata durante la fase di training, restituendo in output una tripla di valori $< p, a, d >$ che corrispondono rispettivamente ai valori di pleasure, arousal e dominance calcolati e che costituiranno l'input per il sottosistema di interfacciamento con la lampada.

Object Interface Il sottosistema che si occupa dell'interfacciamento con la lampada non presenta alcun problema in termini di efficienza. Questo, infatti, è costituito da due sottomoduli, il primo dei quali si occupa di convertire in tempo reale la tripla di valori $< p, a, d >$ ricevuta dal sottosistema precedente in una di valori $< r, g, b >$ corrispondenti ad altrettanti valori di rosso, verde e blu, espressi nell'intervallo [0,255].

Questa tripla di valori viene inviata tramite la porta USB al secondo modulo, realizzato da uno script per Arduino che si occupa di restare in attesa dei dati e, alla loro ricezione, di impostare in maniera adeguata i valori dei led contenuti nella lampada. Per completezza è mostrato in Figura 8 lo schema di questo sottosistema, che insieme al primo realizza il framework completo di Figura 6.

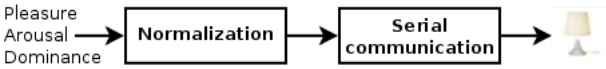

Figura 8: Sistema di interfacciamento con la lampada.

4.3 Dettagli implementativi

Librerie utilizzate Il framework appena descritto è stato implementato in linguaggio C++, mentre lo script per l'interfacciamento con la lampada è in linguaggio C ed utilizza le librerie di Arduino.

Le operazioni di face detection e di manipolazione delle immagini sono state effettuate con l'ausilio della libreria OpenCV (http://opencv.org/). Per la creazione e il fitting del modello dei punti fiduciali è stata utilizzata la libreria asmlib-opencv (https://code.google.com/p/asmlib-opencv/). Le operazioni matriciali nella regressione lineare, infine, sono realizzate utilizzando la libreria Eigen (http://eigen.tuxfamily.org/).

Sintassi del programma Il software realizzato prevede due operazioni principali: una di training ed una di prediction. Per l'operazione di training sono richiesti in input tre parametri:

- il file che contiene il riferimento al dataset di training labellato con i valori di pleasure, arousal e dominance di ogni immagine (descritto in Appendice A);
- il file di output nel quale salvare la matrice dei parametri, risultato della regressione;
- il file che rappresenta il modello da usare come ASM per l'individuazione dei punti fiduciali.

Per l'operazione di prediction, invece, oltre al parametro che indica il modello da usare come ASM, sono richiesti:

- il file che contiene la matrice dei parametri, output dell'operazione di training (descritto in Appendice B);
- l'indice della webcam da utilizzare per l'acquisizione delle immagini (default: 0);
- la porta USB alla quale è collegata la lampada (default: /dev/ttyUSB0).

Training Per effettuare il labelling delle immagini di training con i valori di pleasure, arousal e dominance si è ricorso ad uno strumento chiamato AffectButton (http://mmi.tudelft.nl/ joostb/AffectButton)[12]. Questa applicazione permette, per mezzo di una emoticon, di individuare i valori di PAD corrispondenti all'espressione generata. I valori restituiti dall'applicazione sono stati combinati con alcuni dei volti presenti nel dataset JAFFE sulla base della somiglianza dell'espressione facciale prodotta, come mostrato in Figura 9.

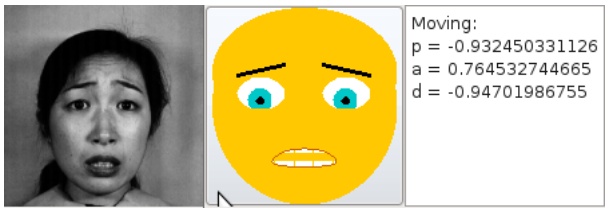

Figura 9: Labelling del dataset di training nello spazio PAD.

Prediction Sulla base degli studi effettuati da Mehrabian e Valdez[7] circa gli effetti del colore sulle emozioni, si è decisa una semplice corrispondenza tra i valori di Pleasure, Arousal e Dominance con quelli di Red, Green e Blue, secondo l'equazione

$$r = \frac{(p+1)*255}{2}, g = \frac{(d+1)*255}{2}, b = \frac{(a+1)*255}{2} \tag{4}$$

dove p, a e d sono valori reali nell'intervallo [-1,1] mentre r, g e b sono valori interi nell'intervallo [0,255].

5 Risultati ottenuti

Poiché il metodo di fitting dell'ASM si basa sul tracciamento di singoli punti, la posizione di questi punti può essere imprecisa se le immagini sono rumorose, causando valori errati nella costruzione della matrice di input.

Durante la fase di training si nota un fitting corretto del modello sulle immagini del dataset, come mostrato in Figura 10

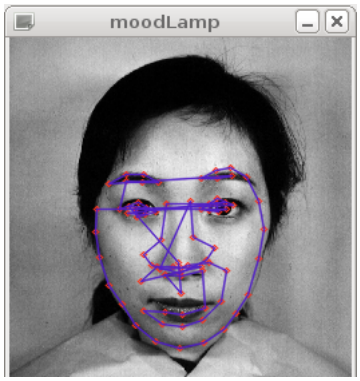

Figura 10: Fitting dell'ASM su una immagine di training.

Nella fase di testing, invece, sono presenti alcuni errori dovuti al fatto che la qualità della luce e della webcam utilizzata non erano ottimali. Tuttavia,

considerando solo i valori accettabili, si riescono ad ottenere buoni risultati, come mostrato in Figura 11.

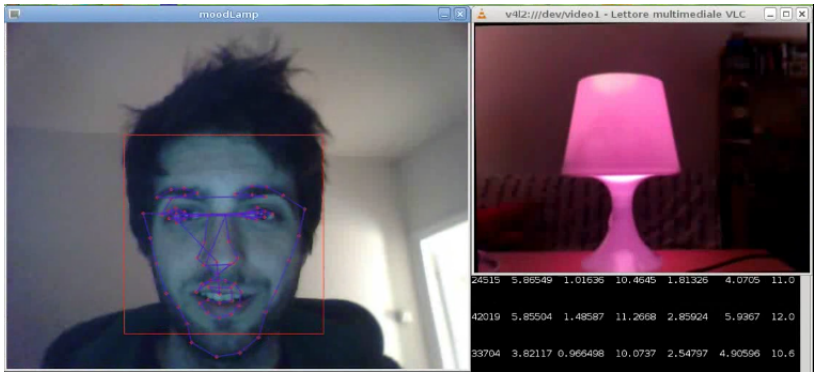

Figura 11: Esempio di risultato ottenuto con l'utilizzo della webcam.

A Formato del file di training

```
10              #dimensione del dataset
01.tiff         #nome del file
0.0145695364238 #valore di Pleasure
-1              #valore di Arousal
0.262251655629  #valore di Dominance
02.tiff
0.87417218543
0.635025754231
0.917880794702
...
```

B Formato del file di fitting

```
21        #numero di righe
3         #numero di colonne
0         #valore w[0,0]
0         #valore w[0,1]
0         #valore w[0,2]
1.19037   #valore w[1,0]
0.40213   #valore w[1,1]
1.49484   #valore w[1,2]
-0.476484
0.578047
...
```

Riferimenti bibliografici

[1] Darwin, C., 1872. The Expression of the Emotions in Man and Animals. Harper Collins, London, (Reprinted 1998).

[2] Ekman, P. e Friesen, W. V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology , 17, 124-129.

[3] Ekman, P. e Friesen, W. V. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement . Consulting Psychologists Press, Palo Alto.

[4] Pandzic, I. S. e Forchheimer, R. 2002. MPEG-4 Facial Animation: The Standard, Implementation and Applications. Wiley

[5] Russell, James (1980). A circumplex model of affect. Journal of Personality and Social Psychology 39: 1161-1178.

[6] Mehrabian, A. (1980). Basic Dimensions for a General Psychological Theory: OG&H Publishers.

[7] Valdez, P. e Mehrabian, A. (1994). Effects of color on emotions. Journal of Experimental Psychology: General 123, 4, 394-409.

[8] Viola, P. e Jones, M. J. (2004). Robust real-time object detection. International Journal of Computer Vision, Vol. 57, No. 2, 137-154.

[9] Cootes, T. F. e Taylor, C. J. (1992). Active shape models - smart snakes. 3rd Bitish Machine Vision Conference, 266-275.

[10] Milborrow, S., Morkel, J. e Nicolls, F. (2010). The MUCT Landmarked Face Database. PRASA.

[11] Lyons, M. J., Akamatsu, S., Kamachi, M. e Gyoba, J. (1998). Coding Facial Expressions with Gabor Wavelets. Third IEEE International Conference on Automatic Face and Gesture Recognition, 200-205

[12] Broekens, J. e W. P. Brinkman. (2009). AffectButton: Towards a Standard for Dynamic Affective User Feedback, In ACII 2009. IEEE.