Università degli Studi di Padova
Corso di Laurea in Scienze Statistiche

TESI DI LAUREA

UN APPROCCIO BAYESIANO PER IL RICONOSCIMENTO DI ESPRESSIONI FACCIALI

Relatore: Prof. NICOLA SARTORI
Dipartimento di Scienze Statistiche

Co-relatore: Prof. BRUNO SCARPA
Dipartimento di Scienze Statistiche

Laureando: EMANUELE NARDO

Anno Accademico 2011/2012

Indice

Introduzione

Le espressioni facciali sono il più potente e naturale mezzo di comunicazione tra gli esseri umani. Esse rappresentano oggetto di studio sin dal XIX secolo, quando Darwin con il saggio *The Expression of Emotions in Man and Animals* dimostrò l'universalità delle espressioni facciali e la continuità di rappresentazione delle stesse fra uomo e animali, rivendicando la presenza di specifiche emozioni innate come prodotto di una funzione biologica adattiva.

Nel 1971 Ekman e Friesen, recuperando le tesi darwiniane, svolsero una ricerca interculturale attraverso le più diverse forme di civiltà esistente per provare l'esistenza di alcune manifestazioni emotive originali per la specie, chiamate emozioni primarie o di base, trasversali all'interno di tutta l'umanità e indipendenti dal contesto socio-culturale di provenienza.

L'espressione delle emozioni avviene tramite l'attivazione di determinati muscoli, negli animali come nell'uomo. Quest'ultimo però possiede una maggiore abilità nell'articolazione delle espressioni facciali, tramite 46 muscoli che risultano il principale vettore di comunicazione emozionale: ognuno di questi è rappresentato da unità d'azione (AU) nel sistema sviluppato Ekman e Friesen, il *Facial Action Coding System* (FACS). La contrazione di questi muscoli genera oltre 7000 combinazioni diverse di espressioni riscontrabili su un volto umano, sebbene la maggior parte di esse si differenzino per piccoli cambiamenti di alcune caratteristiche del viso.

È nato così lo studio delle *microespressioni*, ossia espressioni del viso che appaiono in un venticinquesimo di secondo per poi svanire e che solitamente sono inconsapevoli.

Il sistema FACS si è dimostrato essere un potente mezzo per misurare e classificare l'attività facciale. Tuttavia, la formazione di personale esperto e la registrazione manuale delle AU sono attività dispendiose, inoltre l'affidabilità della codifica manuale non può essere in alcun modo garantita. Pertanto, un sistema che possa riconoscere le AU in tempo reale senza intervento umano è desiderabile in vari campi di applicazione, come per la ricerca comportamentale, la creazione di interfacce uomo-macchina e la sicurezza.

In generale, un sistema di riconoscimento automatico di azioni facciali è costituito da tre fasi peculiari, come illustrato nella Figura 1: l'acquisizione del volto con il riconoscimento della struttura facciale, l'estrazione delle caratteristiche del viso e la classificazione di quest'ultime in AU o in emozioni.

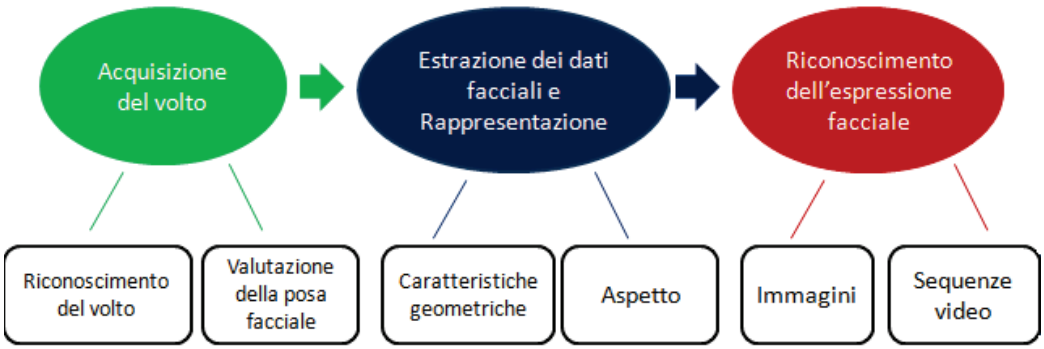

Figura 1: Struttura di base di un sistema di analisi delle espressioni facciali, rappresentazione tratta da Kanade e Tian, 2005, Capitolo 11.

Decenni di studi sono stati spesi per lo sviluppo di questi sistemi. Nonostante i risultati raggiunti dalla comunità scientifica, è tuttora complesso sviluppare procedure automatiche che si dimostrino flessibili e precise al tempo stesso. Inoltre, la vasta quantità di dati relativi alla configurazione facciale prodotta dagli strumenti di rilevazione, spesso non si converte in una capacità di riconoscimento sufficiente per affidarsi a dispositivi completamente indipendenti dalla supervisione umana.

L'obiettivo principale della tesi è valutare alcuni metodi statistici per la classificazione delle caratteristiche facciali, raccolte da fotogrammi di sequenze video. Il database di riferimento è il *Cohn-Kanade AU-Coded Facial Expression Database* (CK+), rilasciato nel 2000 dalla *Carnegie Mellon University* e successivamente completato nella parte finale del 2010. L'archivio contiene sequenze di fotogrammi relativi a pose facciali spontanee e non spontanee di alcuni studenti universitari in prospettiva frontale, ciascuna rappresentante uno stato emotivo specifico. Ogni sequenza è composta da fotogrammi con intensità crescente di espressione, a cui vengono associati 68 punti geometrici di riferimento estratti dall'immagine, le AU coinvolte e l'emozione associata secondo la codifica FACS, qualora fosse univocamente identificabile.

Il lavoro ha lo scopo di costruire e valutare modelli che associno ciascuna mappatura di punti estratti dall'immagine (i *landmarks*) alla relativa emo-

zione presente sul volto dell'individuo, senza una codifica intermedia in AU. L'aspetto del riconoscimento diretto delle emozioni verrà affrontato sia con procedure classiche di analisi dei dati che con un approccio nuovo di stampo bayesiano.

Nel Capitolo 1 vengonoi introdotti gli strumenti adottati all'interno dello studio, con particolare attenzione alle microespressioni, descrivendo il sistema di codifica FACS per l'analisi delle espressioni facciali. Si richiamano alcune nozioni introduttive relative allo studio delle espressioni, soffermandosi sulle applicazioni per le quali questo tipo di ricerca assume un'importanza rilevante nella realtà di tutti i giorni. Infine si presenta la fonte utilizzata per reperire i dati, l'*Extended Cohn-Kanade Database*, pubblicato nel 2010 allo scopo di favorire lo sviluppo di soluzioni e metodi per l'analisi delle espressioni.

Il Capitolo 2 descrive i recenti sviluppi nel campo dell'analisi delle espressioni, analizzando il processo in tutte le sue parti, partendo dall'acquisizione delle immagini, l'identificazione del profilo facciale, l'estrazione delle caratteristiche del volto, fino alla vera e propria classificazione in AU e il conseguente riconoscimento del comportamento emotivo. Gli attributi presi in considerazione possono essere sia punti geometrici che peculiarità estetiche, mentre la classificazione può articolarsi in modo diverso a seconda se viene effettuata su immagini o su sequenze video. Il punto d'arrivo della rassegna è costituito dalla recente analisi realizzata da Lucey *et al.* (2010) sullo stesso campione di immagini utilizzato nel presente lavoro.

Nel Capitolo 3 si adattano alle immagini del CK+ alcune metodologie di data mining per la classificazione in presenza di variabili risposta multiclasse. Congiuntamente alla classificazione in emozioni, si affronta il problema della selezione delle caratteristiche del volto più rilevanti. Dopo una valutazione con campioni di stima e di verifica indipendenti, si effettua un confronto in termini di risultati con Lucey *et al.* (2010).

Il Capitolo 4 descrive lo sviluppo di un approccio bayesiano completamente differente, identificando un modello probabilistico completo per l'algoritmo dei *K-Nearest Neighbors* e applicandolo all'analisi delle componenti facciali. L'analisi bayesiana per il riconoscimento delle emozioni e la selezione del sottoinsieme di variabili più rilevante è stato realizzata tramite l'uso di metodi Monte Carlo basati su catene di Markov.

La trattazione si conclude con una valutazione complessiva dei risultati ottenuti e con alcune considerazioni finali sul lavoro svolto, delineando le prospettive di ricerca future.

Capitolo 1

L'analisi delle espressioni

Tra le varie modalità di comunicazione non verbale dello stato emotivo, quali il sistema vocale, i gesti, la postura, non si può non tenere in considerazione lo sguardo ed il volto, ovvero la regione principale per attirare l'attenzione e l'interesse degli interlocutori: un sistema privilegiato di comunicazione e di trasmissione dei significati. Il presente capitolo introduce l'analisi delle espressioni facciali.

I Paragrafi 1.1 e 1.2 si concentrano sulla natura delle espressioni e la loro descrizione analitica. Nei Paragrafi 1.3 e 1.4 viene affrontata la tematica dello studio delle microespressioni, presentando il sistema di misurazione dei movimenti facciali *Facial Action Coding System* (FACS). Segue una rassegna su alcune applicazioni concrete che svolgono funzioni di utilità nella vita reale. Infine nel Paragrafo 1.6 si descrive la base di lavoro adottata, il *Cohn-Kanade Database*, articolato nella sua prima edizione del 2000 fino all'attuale versione estesa e arricchita da metadati.

1.1 Le emozioni e le espressioni facciali: breve panoramica dei principali orientamenti

Osservazioni sull'apparire di emozioni sul volto si possono trovare già in vari scrittori antichi e medievali, sia d'Oriente sia d'Occidente, ma lo studioso al cui lavoro fanno ancora riferimento gli studi moderni è Charles Darwin (1872), il quale formulò il primo trattato sulle espressioni facciali. L'obiettivo di Darwin era quello di mettere in evidenza il termine "espressione", indicante un'azione accompagnatoria di uno stato mentale, che non comprendeva però soltanto le emozioni, ma anche sensazioni, comportamenti e tratti della personalità. La rilevanza maggiore del suo lavoro, in questo ambito, è sta-

ta quella di considerare le espressioni facciali entro "gruppi" tra i quali c'è diversità e non come categorie fisse, immutabili.

L'influenza di Darwin ha dato vita a due corsi distinti, uno etologico (da intendere come studio del comportamento animale nel suo ambiente naturale) ed uno psicologico. Gli studi prodotti dagli etologi si sono concentrati prevalentemente sulle esibizioni facciali sul piano dell'interazione, mentre gli psicologi si sono avvicinati alle teorie darwiniane, in epoca moderna sotto la guida di Ekman negli anni '70. In quegli anni si distinsero tre scuole di pensiero.

La prima è quella teorizzata da Woodworth (1938) e successivamente sviluppata da suoi allievi i quali ipotizzarono che le espressioni facciali comunichino famiglie di emozioni, accomunate secondo basi quali la piacevolezza o spiacevolezza, l'attivazione autonoma o il rilassamento, l'attenzione o il rigetto.

La seconda scuola di pensiero fu sviluppata da Osgood *et al.* (1966) il quale diede rilievo alla esibizione dell'espressione facciale ed alla risposta dell'osservatore ad essa. Fornì anche prove della generalità transculturale del significato del volto.

La terza è rappresentata da Frijda *et al.* (1989), i quali proposero un modello di percezione dell'emozione del volto basato sull'elaborazione delle informazioni. Chi valutava le espressioni doveva descrivere non solo la singola emozione che riconosceva nel volto, ma anche immaginare gli stati interni della persona che esibiva l'espressione oltre agli stati interni provocati nell'osservatore stesso. Un ulteriore aspetto riguardo le espressioni facciali riguarda i meccanismi alla base della loro produzione. A questo proposito si sono distinte due opposte teorie, l'ipotesi globale e l'ipotesi dinamica.

Secondo l'ipotesi globale, sottolineata da Fridlund *et al.* (1987) ed altri ricercatori quali Izard (1977), le configurazioni espressive del volto sono unitarie, chiuse, universalmente condivise, sostanzialmente fisse e specifiche per ogni emozione.

In alternativa, la teoria dinamica prevede un processo sequenziale e cumulativo in ogni espressione facciale, che sarebbe, quindi, il risultato della progressiva accumulazione e dell'integrazione dinamica delle singole fasi. Secondo questo approccio, le espressioni facciali sono configurazioni motorie momentanee, flessibili e variabili ed in grado di adattarsi alle varie situazioni.

Un'ulteriore focus riguarda la trasmissione di significato delle espressioni facciali distinta tra prospettiva emotiva e prospettiva comunicativa. Per Ekman ed Izard, sostenitori dell' approccio emotivo, le microespressioni, ossia espressioni facciali involontarie della durata di 1/25 di secondo, hanno prevalentemente un valore emotivo in quanto risposta immediata, spontanea ed involontaria delle emozioni. Guardando le microespressioni facciali

di un soggetto si possono "leggere" le emozioni che egli sta provando, intese come categorie discrete. Da questa concezione deriva l'invariabilità culturale che Ekman *et al.* (1971) hanno verificato con soggetti appartenenti a diverse culture rispetto ad espressioni corrispondenti a sette emozioni di base. Tali studi non sono stati però privi di critiche, che facevano riferimento soprattutto ai metodi utilizzati e quindi ai dati ricavati. Tuttavia, restano comunque delle corrispondenze che hanno dato vita alla teoria dell'universalità minima, secondo la quale esiste un certo grado di somiglianza fra le culture nell'interpretazione delle microespressioni facciali.

Atri studiosi, come Fridlund (1994), hanno sostenuto la prospettiva comunicativa. Secondo tale approccio, le espressioni facciali hanno un valore prevalentemente comunicativo poiché manifestano agli altri le intenzioni del parlante. Le microespressioni, perciò, variano in funzione del contesto ed hanno un valore sociale: trasmettono agli altri obiettivi comunicativi. Il concetto di semplicità implicita spiegherebbe la produzione di microespressioni facciali anche in assenza di interlocutori: si è in presenza di un uditorio implicito anche quando si è soli.

Al di là degli studi effettuati e delle teorie che si sono sviluppate attorno alle espressioni facciali, tutti sono concordi nel sostenere l'universalità della comunicazione espressiva delle emozioni: "**molte microespressioni facciali sono presenti in tutto il mondo, in ogni razza e cultura umana**" (Frijda, 1989), sono queste conclusioni alle quali era giunto lo stesso Darwin e, in tempi molto più recenti Fridlund *et al.* (1987), secondo i quali le espressioni di **rabbia, disgusto, felicità, tristezza, sorpresa, disprezzo e paura** sono universali.

Ognuna di queste emozioni può essere riconosciuta nel proprio viso ed in quello di tutte le persone, a prescindere da etnia, cultura, genere, religione e, con un opportuno allenamento, è possibile riconoscerle e capire la reale emozione che una persona prova, persino se quella persona siamo noi stessi. Di seguito presentiamo una descrizione delle principali caratteristiche di ciascuna delle sette emozioni primarie.

1.1.1 La rabbia

Molte sono le cause che possono far provare un senso di rabbia. I motivi per cui ciò può accadere sono diversi, ad esempio quando un imprevisto reca un danno immediato, ma anche un senso di ingiustizia può sortire lo stesso effetto, o come conseguenza di un'aggressione fisica o verbale. Naturalmente viene espressa con vari livelli. Differentemente da altre emozioni, come paura e sorpresa, non si manifesta immediatamente nella sua interezza, ma tende a raggiungere il suo apice per gradi. Può iniziare come senso di fastidio e

andarsene immediatamente, percepita solo come un accenno, oppure se non interrotta autoalimentarsi rendendosi così sempre più difficile da controllare. Ovviamente dipende dal motivo per il quale si prova rabbia, perché maggiore è il "danno" subìto, maggiore sarà la velocità e la probabilità che si autoalimenti. In Figura 1.1 sono evidenziate le peculiarità dell'espressione di rabbia. Quando si manifesta, compare su tutte e tre le parti del viso: fronte, occhi, bocca.

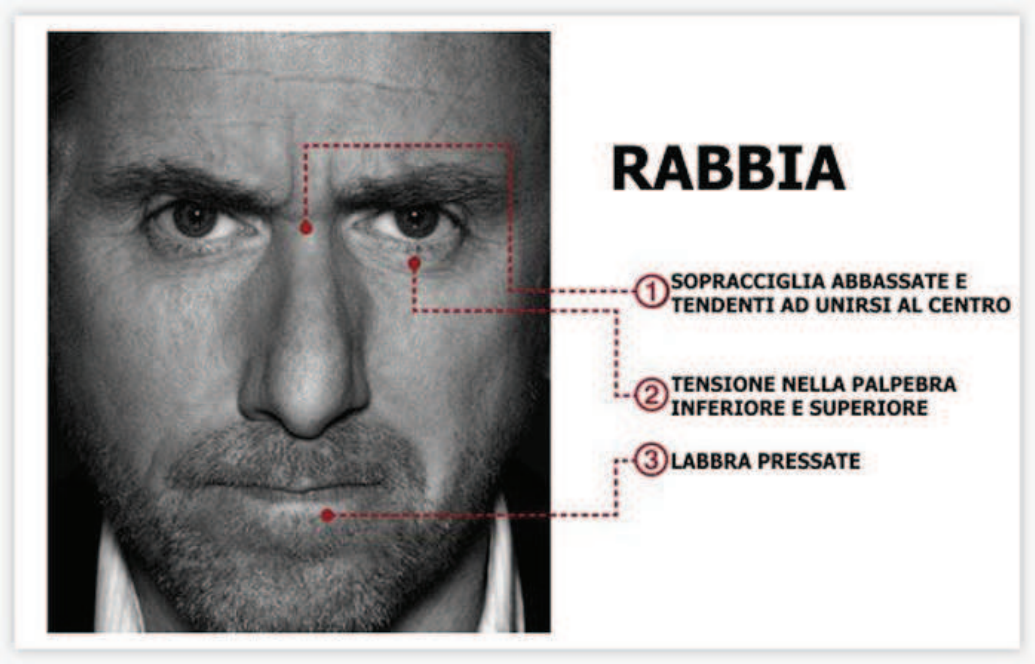

Figura 1.1: Rabbia. Immagine raffigurante Tim Roth, attore e regista inglese. Fonte *www.movieplayer.it*.

Come è possibile notare nella fronte, le sopracciglia si abbassano, gli angoli interni sempre delle sopracciglia si abbassano anch'essi muovendosi verso il centro formando due linee verticali nel mezzo. La palpebra inferiore e quella superiore risultano essere in tensione e le labbra sono pressate con forza. Restando invariata la parte della fronte e quella relativa agli occhi, può variare quella della bocca. Le labbra vengono pressate, a meno che non si stia parlando o urlando mentre si sfoga la propria rabbia, in questo caso la bocca risulterà aperta ma in tensione.

1.1.2 Il disgusto

Si prova una sensazione di disgusto quando si avverte repulsione per qualcosa. Può essere un oggetto, un odore, un sapore o un pensiero, a volte solo il ricordo di ciò per cui si è provato disgusto può far provare questa sensazione. In tutte queste occasioni la prima reazione è quella di allontanarsi o liberarsi da ciò che provoca disgusto.

Figura 1.2: Disgusto. Per i riferimenti all'immagine si veda la Figura 1.1.

Come presentato in Figura 1.2, un'espressione di disgusto si manifesta con il sollevamento del labbro superiore della bocca, le guance si alzano provocando l'innalzamento delle palpebre inferiori, le sopracciglia scendono abbassando a loro volta le palpebre superiori, mentre il naso si arriccia.

1.1.3 La felicità

La felicità è l'emozione che tutti vogliono provare e sentire il più spesso possibile. E' l'emozione più piacevole perchè quando si prova si sta bene, tanto che si preferisce frequentare persone che ridono e sono felici piuttosto di altre che non lo sono. A differenza della paura, le situazione dove si prova felicità si memorizzano per cercare di ripeterle o fare in modo che riaccadano.

Figura 1.3: Felicità. Per i riferimenti all'immagine si veda la Figura 1.1.

Un volto felice come quello presentato nella Figura 1.3 è contraddistinto da un'espressione che comporta il movimento della bocca verso l'esterno e verso l'alto, le guance si alzano provocando a loro volta l'innalzamento delle palpebre inferiori e delle rughe si formano attorno agli occhi. Inoltre una linea si forma partendo dal naso fino agli angoli della bocca.

1.1.4 La tristezza

Quando si subisce una perdita si prova tristezza. Si può trattare dell'allontanamento di una persona o la perdita di un oggetto a cui si era particolarmente affezionati. Può accadere dopo la separazione da un amico o da una persona amata, per la perdita del lavoro o della propria salute. La tristezza è un'emozione che permette di superare queste perdite e di comunicare agli altri il nostro bisogno di essere confortati. Il volto di una persona triste (Figura 1.4) apparirà in questo modo.

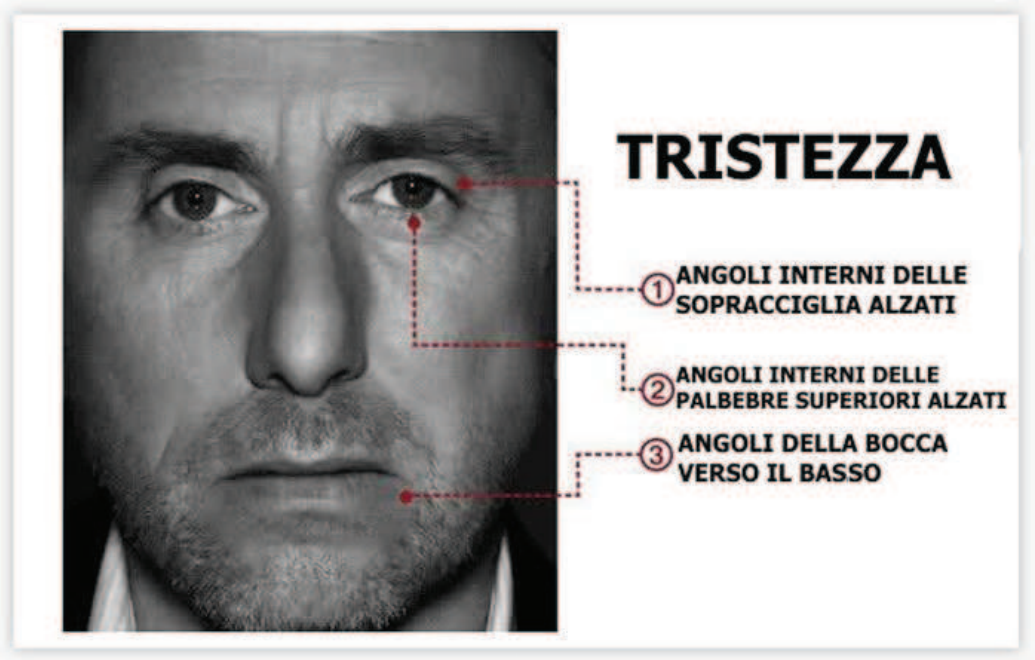

Figura 1.4: Tristezza. Per i riferimenti all'immagine si veda la Figura 1.1.

Nella parte relativa alla fronte si nota che gli angoli interni delle sopracciglia si alzano verso il centro. Le palpebre superiori tendono ad alzarsi negli angoli interni. Gli angoli della bocca risultano abbassati e con l'intensificarsi della tristezza le labbra risulteranno tremanti e il labbro inferiore esteso in avanti.

1.1.5 La sorpresa

Quando ci si trova di fronte ad un fatto inaspettato, la prima sensazione che si prova è sorpresa. Può accadere quando si incontra qualcuno che non si vedeva da molto tempo, oppure quando viene data un'informazione che non ci si aspettava o in mille altre circostanze che in comune hanno il fatto di accadere senza alcun preavviso. Due sono le particolarità di questa emozione. È istantanea, infatti ha una durata brevissima, anche una sola frazione di secondo, il tempo necessario al cervello di elaborare questa nuova informazione. Una volta elaborata subentrerà l'emozione successiva che dipende dal motivo di tale sorpresa. Per esempio può essere di gioia se si apprende di avere vinto al superenalotto o di paura se una macchina esce all'improvviso tagliando la strada. Secondariamente la sorpresa non è un'emozione né positiva né negativa. Non dà né gioia né turbamento. Come detto serve per apprendere ed

elaborare una nuova informazione, sarà poi l'emozione successiva ad essere positiva o negativa.

Figura 1.5: Sorpresa. Per i riferimenti all'immagine si veda la Figura 1.1.

Come è possibile osservare in Figura 1.5, nella parte relativa alla fronte, le sopracciglia sono alzate e formano una curvatura verso l'alto. Spesso come conseguenza di questo movimento si formano linee orizzontali che attraversano tutta la fronte. Gli occhi risultano spalancati, permettendo così al cervello di acquisire il maggior numero di dati in merito a questa nuova informazione che crea sorpresa. Nessuna tensione si crea sulla palpebra superiore o inferiore. Come conseguenza di una sorpresa si apre la bocca con un movimento della mascella verso il basso che separa le arcate dentali.

1.1.6 Il disprezzo

Il disprezzo è un'emozione rivolta esclusivamente verso le persone. Non si prova disprezzo per oggetti od odori, piuttosto disgusto. Si manifesta quando ci si trova di fronte a situazioni ritenute immorali e si prova un senso di superiorità rispetto a chi ha compiuto l'azione. Può accadere quando si vede un uomo maltrattare una donna o un bambino, oppure se si nota una persona che maltratta un animale. Si prova disprezzo anche quando il proprio

consiglio o la propria opinione non viene ascoltata o contraddetta da qualcuno che occupa un grado superiore (capoufficio per esempio) ma ritenuto moralmente inferiore. In Figura 1.6 una tipica espressione di disprezzo.

Figura 1.6: Disprezzo. Per i riferimenti all'immagine si veda la Figura 1.1.

Tra le emozioni di base, il disprezzo è l'unica che si presenta in modo asimmetrico, ovvero su un solo lato del volto. Compare solo nella parte inferiore del viso dove la bocca da un lato si tende verso l'esterno con l'angolo rivolto all'insù.

1.1.7 La paura

La paura è un'emozione a cui qualsiasi essere vivente rinuncerebbe, non tanto per le sensazioni che comporta, ma quanto per le cause che la scatenano. La priorità del cervello è quella di sopravvivere e per farlo ha bisogno di allontanarsi da tutte le situazione che lo mettano in pericolo, da tutto ciò che può recare danno fisico, morale o psicologico. Per raggiungere questo scopo viene "attivata" l'emozione della paura, una sensazione negativa e sgradevole che provoca una reazione fisica immediata di fuga per creare distanza dal pericolo, inoltre permette al cervello di memorizzare quel pericolo associandolo alla sensazione di paura in modo tale da evitarlo in futuro.

Figura 1.7: Paura. Per i riferimenti all'immagine si veda la Figura 1.1.

Nella Figura 1.7, le sopracciglia sono alzate e posizionate quasi in linea orizzontale (a differenza della sorpresa dove sono piegate verso l'alto), gli angoli interni vanno verso il centro e sulla fronte si formano linee orizzontali solo in centro (a differenza della sorpresa dove le linee orizzontali attraversano tutta la fronte). Gli occhi si aprono perché sia le sopracciglia che le palpebre superiori si alzano, ma non risulta lo stesso effetto 'spalancato' come nella sorpresa in quanto sono in tensione e tendono ad alzarsi le palpebre inferiori. La bocca può aprirsi come nella sorpresa, ma anziché essere rilassata, presenta tensione nelle labbra che si tendono verso l'esterno.

1.2 Lo studio delle microespressioni

Quando si osserva un volto che esprime un'emozione, la decodifica della sua espressività sembra non essere un problema significativo per la maggior parte delle persone. L'informazione contenuta in esse sembra essere una aspetto naturale delle interazioni e delle relazioni sociali e la sua elaborazione procede solitamente senza difficoltà o problemi nella comunicazione. A dispetto di questa facilità di gestione delle espressioni facciali nell'interazione quotidiana, i ricercatori si trovano da sempre discordi sull'interpretazione delle informazioni veicolate dal volto e sulla metodologia più adeguata per

studiarle. Il problema principale consiste nel riconoscimento o meno, dell'esistenza o meno e della natura di un'intenzione comportamentale o di uno stato di sentimento emozionale all'interno delle espressioni del volto. Per rispondere a svariati quesiti riguardanti i legami esistenti tra le microespressioni facciali e le caratteristiche di personalità, l'esperienza emotiva ed i processi comunicativi, sono state messe a punto, a partire dagli anni '60, numerose tecniche di rilevazione e di analisi delle espressioni facciali. Le tecniche che si concentrano sul volto, quale elemento predominante delle espressioni delle emozioni, si suddividono in studi di giudizio ed in studi di misurazione.

Gli studi di giudizio riguardano le informazioni veicolate dal comportamento facciale, le emozioni che ne possono derivare, nonché quali tratti della personalità vengono espressi dalla mimica facciale. Complessivamente si sono diretti verso due orientamenti non esclusivi: la formulazione di giudizi qualitativi, rappresentato, per esempio, dalla scelta forzata del giudizio entro un numero limitato di possibilità, o il metodo di valutazione, che permette di valutare in quale misura compaiono una serie di proprietà nei volti presi in esame e la scelta è sempre effettuata dal ricercatore.

Gli studi di misurazione si concentrano in prevalenza sulla rilevazione dei movimenti, ossia sugli aspetti che non prendono in considerazione ciò che vuole essere comunicato con il comportamento facciale. Si basano sulla identificazione di unità di comportamento facciali visibili. Si distinguono dai metodi di giudizio poiché, nonostante implichino il giudizio di osservatori, le valutazioni sono puramente descrittive e non interpretative. Gli indirizzi di tali studi si suddividono a loro volta in due gruppi: quelli bastati sulla riflessione teorica e quelli fondati sull'anatomia muscolare del volto.

I primi puntano all'identificazione delle combinazioni di movimento facciale, presumibilmente associate ad emozioni particolari, quelle "universali", ma non permettono di misurare l'intensità del comportamento e non danno giustificazione ad azioni diverse da quelle precostituite, quali prototipo di determinati stati emotivi. A questa categoria di sistemi di valutazione fanno parte il *Facial Affect Scoring Technique* di Ekman *et al.* (1971), il *Maximally Discriminative Facial Movement Coding System* di Izard (1979) ed il *System for Identifying Affect Expression by Holistic Judgment* di Izard e Dougherty (1980).

Per quanto riguarda gli studi sull'anatomia muscolare, una prima trattazione è stata presentata da Hjortsjö (1969). Imparando a muovere volontariamente i muscoli facciali, Hjortsjö ha sviluppato una descrizione precisa dei cambiamenti risultanti da ciascun movimento muscolare facciale ed ha fornito un sistema di codifica numerico. Riprendendo questa idea, Ekman e Friesen (1978) hanno elaborato un sistema maggiormente sofisticato, il *Facial Action Coding System (FACS)*. In questo sistema ogni movimento singolarmente ri-

levabile è stato indicato come unità d'azione, a cui può essere attribuito un punteggio in base all'intensità. È un sistema piuttosto elaborato e non tutti i movimenti rilevati possono essere attribuiti ad una specifica emozione. Per ovviare alle difficoltà per i ricercatori, gli stessi autori hanno elaborato l'EMFACS (1982) che assegna un punteggio soltanto alle unità o combinazioni di unità d'azione che la teoria considera segnali emozionali. Uno schema esplicativo sulle tipologie di studio delineate è fornito in Figura 1.8.

Figura 1.8: Metodi per lo studio delle espressioni

1.3 Il metodo FACS: Facial Action Coding System

Ekman e Friesen hanno elaborato il Facial Action Coding System (1978) come sistema di osservazione e classificazione di tutti i movimenti facciali visibili, anche quelli minimi, in riferimento alle loro componenti anatomo-fisiologiche che sono gli elementi costitutivi delle esibizioni facciali. Tra i vari sistemi di analisi e misurazione delle espressioni facciali elaborate fino ad ora il FACS è il più comprensibile, completo e versatile. In esso il volto è considerato come un sistema di risposta multidimensionale, capace di flessibilità e specificità. Il volto veicola informazioni attraverso quattro classi di segnali:

- segnali statici, che rappresentano tratti relativamente permanenti del volto, come la struttura delle ossa e le masse di tessuto sottocutaneo che contribuiscono a dare forma al volto;

- segnali lenti, che sono costituiti dai cambiamenti che avvengono sul volto nel corso del tempo e che segnano il viso con le rughe, con il cambiamento nella grana della pelle;

- segnali artificiali, rappresentati dai tratti del volto determinati da elementi artificiosi, come gli occhiali, la cosmesi o interventi di chirurgia ricostruttiva;

- segnali rapidi, variazioni che si configurano sul volto dovute all'attività neuromuscolare e che determinano le vere e proprie espressioni facciali.

La somma delle quattro classi di segnali determinano la fisionomia di un volto, e singolarmente o cumulativamente comunicano molti messaggi. Ai fini dello studio sulle emozioni si analizzano i segnali rapidi che attuano variazioni della forma degli occhi, delle sopracciglia, della bocca e delle labbra.

Al fine di fornire un indice specifico per ogni tipo di movimento e di espressione, il FACS prende in considerazione 44 unità fondamentali denominate da Ekman e Friesen "Unità d'Azione" (Action Units, nel seguito AU) che possono dare luogo a più di 7000 combinazioni possibili. In totale sono classificati 58 diversi movimenti o caratteristiche, alcuni dei quali sono associati tipicamente a un'emozione specifica, mentre altri non sono associati ad alcuna emozione in particolare.

Per ciascuna AU sono descritte nel manuale del FACS, nella sezione A) i cambiamenti osservabili sul volto, provocati da quel movimento; nella sezione B), indicazioni su come effettuare quel specifico movimento, con la descrizione di eventuali difficoltà e soluzioni utili a superarle; nella sezione C), come registrare l'intensità di quella singola AU.

Il FACS è un sistema di osservazione puramente descrittivo, e in quanto tale, non soggetto alla tendenza di attribuire immediatamente un significato interpretativo alla mimica facciale. Per raggiungere l'obbiettivo, gli autori hanno fatto riferimento esclusivamente all'analisi del fondamento anatomico dei movimenti del volto umano. Poiché ognuno di questi movimenti è il risultato dell'azione singola o sinergica dei muscoli facciali, il FACS prende in considerazione il modo in cui ogni muscolo facciale agisce nel modificare visibilmente la configurazione del volto stesso. Inoltre, Ekman e Friesen hanno fatto ricorso anche alla tecnica di Duchenne che consiste nello stimolare diversi fasci di fibre muscolari con elettrodi, così da determinare, in modo riflesso, gli effetti cinesici esteriormente osservabili.

Una limitazione del FACS, consiste tuttavia nel prendere in considerazione solo i mutamenti visibili del volto umano, non quelli non osservabili. Si tratta di una esplicita scelta di Ekman al fine di una concentrazione su ciò che può essere effettivamente impiegato nella comunicazione non verbale. Nel FACS infatti, vengono trascurate le caratteristiche statiche del volto quali il trucco o l'acconciatura ed espressioni facciali quali il rossore e la sudorazione. Per gli autori andrebbero in questi casi applicate altre metodologie di ricerca. Per descrivere in modo preciso e attendibile ogni AU, Ekman e Friesen hanno preso in considerazione il linguaggio del corpo nella sua processualità, non soltanto la descrizione delle configurazioni statiche finali.

Un aspetto rilevante della metodologia è la possibilità di rilevare livelli di intensità di ogni attività muscolare. Gli autori hanno creato uno strumento oggettivo di categorizzazione dei movimenti facciali, che può essere rielaborato in relazione agli studi che si desidera condurre, e che ne prevedano l'applicazione.

I livelli di intensità sono cinque e così suddivisi: livello A, in presenza di tracce deboli dell'azione; livello B, con evidenza leggera; livello C, dinanzi a segni marcati o pronunciati del movimento; livello D, riscontrabile con segni intensi o estremi dell'azione; livello E, per intensità massima del movimento.

Questa scala d'intensità non è una scala ad intervalli regolari, i livelli A, B e E hanno un'ampiezza abbastanza ristretta, mentre i livelli C e D comprendono la maggior parte dei movimenti facciali. In un'analisi di un movimento facciale la lettera relativa ad un dato livello di intensità individuata viene posta a fianco del numero della relativa AU. Ad esempio, è possibile rilevare le differenze di intensità per l'AU che prevede la contrazione della palpebra inferiore per azione del muscolo orbicolare dell'occhio, in due immagini che sono state valutate dagli stessi autori rispettivamente di intensità B ed E.

Congiuntamente alla creazione di tale codifica, gli autori hanno provveduto anche a creare un manuale di insegnamento pubblicato nel 1992 per quei ricercatori che intendessero applicare la ricerca sull'analisi dei singoli movimenti facciali. Un analista FACS, infatti, oltre a conoscere le singole AU, deve anche essere in grado di selezionare un'espressione osservata, decomponendola nelle singole unità che sottostanno al movimento facciale. Deve saper attribuire un punteggio all'intensità, valutare l'apparizione e la scomparsa di ogni azione muscolare oltre alla eventuale presenza di asimmetrie bilaterali, caratterizzate dalla lettera L o R, che contraddistinguono rispettivamente il lato sinistro ed il lato destro del volto.

Le unità di punteggio, ovvero una lista di AU coinvolte in un'espressione facciale con i rispettivi livelli di intensità, sono puramente descrittive: non interferiscono cioè con l'interpretazione delle emozioni e possono essere con-

vertite da un software usando un dizionario di interpretazione e predizione delle emozioni, come quello presentato nella Tabella 1.1.

Emozione	**AU**
Felicità	6+12
Tristezza	1+4+15
Sorpresa	1+2+5B+26
Paura	1+2+4+5+20+26
Rabbia	4+5+7+23
Disgusto	9+15+16
Disprezzo	6 R12A+R14A

Tabella 1.1: Tabella di interpretazione delle emozioni

I vantaggi offerti dall'uso del FACS come sistema di analisi delle espressioni facciali, quindi, sono innanzitutto di ordine metodologico: è un sistema basato sull'anatomia dei muscoli mimici ed in quanto tale, maggiormente preciso e permette, inoltre, di identificare e rilevare ogni movimento possibile con il numero esatto di muscoli che si sono contratti o rilassati per eseguirlo.

L'attribuzione dei movimenti ad unità d'azione, piuttosto che all'azione dei singoli muscoli, rende più chiara la codifica, dal momento che ogni AU identifica un singolo movimento facciale, mentre un muscolo può sottostare a più movimenti ovvero un'azione facciale può essere la risultante dell'azione di più di un muscolo. Per contro, svantaggi derivanti dall'uso di questo sistema riguardano la difficoltà di apprendimento delle tecniche ed i tempi necessari per il suo utilizzo.

1.4 Rassegna di applicazioni

La codifica delle espressioni facciali trova spazio in un insieme molto esteso di applicazioni concrete: nella ricerca psicologica, per realizzare sistemi di interpretazione emozionale e la simulazione della mimica facciale, come nell'interazione uomo-macchina, spaziando in qualsiasi ambito che coinvolga dinamiche sociali e umane. Di seguito vengono presentate alcune esemplificazioni di come questo approccio multidisciplinare trovi spazio nella vita di tutti i giorni. La metrica di riferimento per tutti i lavori menzionati è in AU FACS.

In uno studio della Stanford University (Ahn *et al.* 2008) è stata utilizzata l'analisi dell'espressione facciale per valutare lo stato emotivo dei consumatori allo scopo di identificare i reali acquirenti, sia che si rechino al punto

vendita o compiano le loro scelte via web. L'obiettivo è risparmiare tempo e dirigere il personale ai punti vendita dove è più probabile ricevere acquirenti interessati, mentre nello shopping online è offrire una maggiore libertà di navigazione e personalizzazione del servizio agli utenti aumentando così il potenziale per le vendite future. Le applicazioni di questi studi permetteranno ai computer di 'leggere' i volti degli utenti, prevedere la sequenza successiva di comportamenti e di agire preventivamente. Questo tipo di progressi nei nuovi media permette di interagire con la tecnologia, invece di reagire passivamente al comportamento degli utenti.

Una ricerca pubblicata nel luglio 2007 affronta la spinosa questione della percezione del dolore (Feldt *et al.*, 2007) nei soggetti con deterioramento cognitivo. Lo studio valuta le espressioni facciali come mezzo per identificare il dolore orofacciale in soggetti anziani sia con capacità cognitive intatte che deficitarie, confrontando i risultati rispetto ad altri strumenti disponibili per la valutazione del dolore.

L'università di Ottawa ha affrontato il problema di rappresentare espressioni facciali di emozioni combinate in modo percettivamente valido (Arya *et al.*, 2009). La ricerca è stata effettuata nel contesto di sanità e istruzione al fine di migliorare la competenza sociale e la capacità di riconoscimento delle espressioni facciali in bambini autistici, nonché l'apprendimento delle lingue native, ma i risultati possono essere estesi a molte altre applicazioni come giochi con necessità di espressioni facciali dinamiche o strumenti per automatizzare la creazione di animazioni del volto umano.

Un'altra applicazione molto attuale si può riscontrare nell'ambito della sicurezza sociale. Per esempio, i tirocinanti presso l'Accademia Nazionale di FBI vengono addestrati in modo da aumentare la loro capacità di riconoscere microespressioni, utilizzando software che permettono un apprendimento in alcuni casi superiore al 90%; strumenti di questo tipo supportano anche la formazione di alti ufficiali inquirenti della US Coast Guard, i quali hanno dimostrato una capacità post-apprendimento oltre l'80%. Nei più importanti aeroporti statunitensi e tedeschi, sono attivi sistemi automatici di riconoscimento facciale, che offrano un supporto alla sicurezza e alla lotta al terrorismo.

Il Dipartimento di Informatica dell'Università di San Diego (California), ha presentato nel 2006 i risultati di alcune ricerche sulle espressioni facciali, in cui sono stati raggiunti risultati concreti in settori cruciali quali educazione, medicina e psicologia del comportamento umano, utilizzando uno strumento automatico di riconoscimento delle espressioni facciali in tempo reale, presentato come *Computer Expression Recognition Toolbox* (CERT). Nella sua ultima versione del 2011, può rilevare automaticamente l'intensità di 19 diverse azioni facciali del FACS e 6 differenti espressioni facciali prototipiche

(Bartlett *et al.*, 2006). Calcola inoltre le localizzazioni di 10 caratteristiche facciali, nonché l'orientamento 3-D (movimento laterale, sollevamento, rotazione) della testa.

Un utilizzo del CERT è stato effettuato per discriminare le espressioni di dolore genuine da quelle simulate (Bartlett *et al.*, 2009). L'obiettivo finale dello studio non riguardava l'individuazione di espressioni non-autentiche in sé, ma piuttosto dimostrare la capacità di un sistema automatizzato di rilevare un comportamento facciale che l'occhio inesperto potrebbe non riuscire ad interpretare, e di approfondire le conoscenze sul controllo nervoso del viso.

Il Ministero dei Trasporti americano nel 2001 ha stimato che la sonnolenza del conducente provoca più incidenti mortali che la guida in stato di ebrezza. Quindi un sistema automatizzato in grado di rilevare la sonnolenza e avvertire il conducente o gli altri passeggeri potrebbe salvare molte vite. Mentre vi è una notevole evidenza empirica che la frequenza di ammiccamento può essere relazionata all'insorgere della sonnolenza, non è noto se ci siano altri comportamenti facciali in grado di predire tali occorrenze. Tramite uno studio apposito si è cercato di scoprire quali configurazioni facciali siano sintomatiche della stanchezza, ottenendo risultati sorprendenti. Il movimento facciale è stato analizzato automaticamente utilizzando il CERT, supportando la raccolta dati mediante un accelerometro posto sulla testa del soggetto e un rilevatore di oscillazioni sul volante.

Con un progetto simile, Dinges *et al.* (2005) hanno lavorato a un sistema automatizzato in grado di discriminare fra bassi e alti livelli di stress espressi da un volto umano. Questa particolare applicazione fu sviluppata a supporto delle missioni aerospaziali con lo scopo di quantificare lo stress manifestato dagli astronauti inviati nello spazio.

Whitehill *et al.* (2008) hanno studiato l'utilità di integrare il riconoscimento delle espressioni in un sistema di insegnamento automatizzato. Questo lavoro utilizza l'espressione per stimare la velocità di visualizzazione dei video preferita dallo studente e il livello di difficoltà della lezione come viene percepito dal singolo studente, in ogni istante di tempo. Questo studio pilota ha compiuto i primi passi verso lo sviluppo di metodi di insegnamento con politiche a ciclo chiuso, come sistemi che danno accesso a stime in tempo reale degli stati cognitivi ed emotivi degli studenti e agiscono di conseguenza.

Applicazioni di successo sono state raggiunte anche nella studio di disturbi psichiatrici (Wang *et al.*, 2008), rivolgendosi a due casistiche di pazienti: quelli affetti da schizofrenia e quelli con la sindrome di Asperger. Mentre è ben noto che i pazienti con problemi di schizofrenia presentano nell'espressione facciale appiattimenti e anormalità, ci sono pochi dati oggettivi sul comportamento facciale a causa del tempo richiesto per la codifica manuale.

Allo stesso modo, esistono poche prove scientifiche relative alla produzione di espressioni facciali in soggetti affetti da disturbi dello spettro autistico.

Cassel *et al.* (2008) hanno prodotto la prima applicazione per la misurazione automatica dell'associazione naturale che si instaura in termini di espressioni facciali nelle interazioni fra bambino e genitore. Hanno studiato il comportamento facciale di due coppie di madri e bambini durante le sessioni di gioco, riscontrando sincronie nella produzione dei sorrisi (AU 12), contrazione degli occhi (AU 6) e apertura della bocca (AU 25, 26).

Le tecnologia per l'implementazione di questo tipo di sistemi è progredita al punto da poter applicare le analisi a espressioni facciali spontanee con un discreto successo. L'accuratezza spesso non è tale da renderli uno strumento indipendente dalla supervisione umana, però presentano punti di forza quali capacità di astrazione delle problematiche, identificazione di strutture latenti, rappresentazione dinamica tridimensionale e applicazione estensiva che rivelano come questo campo di ricerca possa risultare accattivante e proficuo negli anni a venire.

1.5 Il Cohn-Kanade AU-Coded Facial Expression Database

Per suffragare adeguatamente lo sviluppo di soluzioni per l'analisi delle espressioni facciali è indispensabile disporre di una base di lavoro comune a tutta la comunità scientifica. Una piattaforma universalmente condivisa attualmente non è ancora stata predisposta, ma singoli gruppi di ricerca lavorano a progetti indipendenti, utilizzando basi di dati create appositamente per soddisfare specifiche esigenze sperimentali.

Per la valutazione di modelli statistici è consigliabile l'utilizzo di un insieme di dati omogeneo e completo che permetta un confronto diretto con i risultati ottenuti in altri studi. La scelta di una banca dati appropriata dovrebbe essere effettuata in base agli obiettivi prefissati e alle proprietà che si intende studiare (Gross *et al.*, 2005). In funzione di tale criterio, abbiamo fatto uso del dabase raccolto, strutturato e messo a disposizione da Cohn, Kanade e Tian, la cui prima pubblicazione risale all'anno 2000.

Il gruppo di ricerca interdisciplinare formato da psicologi e informatici dell'Università di Pittsburgh ha sviluppato quest'ampia collezione di espressioni facciali per lo sviluppo e il controllo di algoritmi per l'analisi delle espressioni, denominato *CMU-Pittsburgh AU-Coded Face Expression Image Database*. La collezione nella sua distribuzione attuale è composta da 488 sequenze di immagini relative a 97 persone diverse. Ogni sequenza di imma-

gini contiene 18 immagini in media, da un minimo di 4 a un massimo di 66. Ogni sequenza mostra una faccia neutra all'inizio per poi svilupparsi in una delle sette espressioni facciali universali definite da Ekman e Friesen. Inoltre vengono fornite l'insieme delle AU coinvolte in ciascuna sequenza, rilevate da esperti FACS. Si noti che il *Cohn-Kanade AU-Coded Facial Expression Database* non contiene espressioni naturali, bensì ha chiesto ai soggetti volontari di rappresentarle. Le sequenze di immagini, derivanti da riprese video, sono state registrate in un ambiente di laboratorio con condizioni di illuminazione predefinite, fondo omogeneo e prospettiva frontale. I risultati ottenuti dai modelli con queste sequenze di immagini non possono dunque essere implicitamente estesi a un contesto di vita reale, ma forniscono una base sperimentale per sviluppi più complessi.

1.5.1 Le specifiche del database

I comportamenti facciali sono stati registrati su un campione originario di 210 adulti con età compresa fra i 18 e i 50 anni. La composizione è volutamente eterogenea: 69% donne, 31% uomini; 81% Americani di origine europea, 13% Americani di origine africana, 6% di altri gruppi etnici. Le riprese sono avvenute in una sala d'osservazione contenente una sedia dove sono stati fatti sedere i soggetti e 2 fotocamere, ciascuna collegata a un videoregistratore fornito di un time-generator. Una delle fotocamere è stata posizionata frontalmente al soggetto, l'altra con un angolo di 30 gradi a destra. Con circa un terzo dei volontari la luminosità della sala è stata aumentata con una lampada ad alta intensità, mentre il resto degli individui ha usufruito di un sistema di illuminazione uniforme apportata da due differenti fonti di luce. I soggetti sono stati istruiti da uno sperimentatore a eseguire una serie di 23 pose facciali; tra queste figurano sia singole unità d'azione che combinazioni di AU. Ulteriori dettagli sulla composizione del database sono reperibili nel documento associato alla pubblicazione (Kanade *et al.*, 2000). La Tabella 1.2 riassume le principali caratteristiche tecniche.

1.5.2 The Extended Cohn-Kanade Database (CK+)

Il *Cohn-Kanade (CK) Database* è stato distribuito allo scopo di promuovere la ricerca nell'individuazione automatica delle espressioni facciali individuali. Da allora, il CK database è diventato uno dei più diffusi banchi di prova per lo sviluppo e la valutazione di sistemi di classificazione. Durante questo periodo, tre limitazioni sono apparse evidenti: mentre i codici AU vengono correttamente identificati, per le etichette delle emozioni non si può dire altrettanto, riferendosi a ciò per cui erano state richieste piuttosto che a

Soggetti	
Numero di soggetti	210
Età	18-50 anni
Donne	69%
Uomini	31%
Euro-Americani	81%
Afro-Americani	13%
Altri	6 %
Sequenze digitalizzate	
Numero di soggetti	182
Risoluzione	640x490 in scala di grigi
	640x480 per colori 24 bit
Visuale frontale	12105
Visuale a 30 gradi	Solo in ripresa video
Durata della sequenza	9-60 frames/sequenza

Tabella 1.2: CMU-Pittsburh AU-Coded Facial Expression Image Database. La quantità di sequenze realizzate (con 210 soggetti coinvolti) è superiore a quelle successivamente digitalizzate (182 soggetti); da queste sono state poi selezionate le 488 sequenze finali pubblicate nella distribuzione ufficiale (97 individui).

ciò che viene riprodotto in realtà; la mancanza di una metrica comune delle prestazioni con cui valutare i nuovi modelli; non esiste un protocollo per il confronto dei risultati fra differenti basi di dati. Di conseguenza, il database CK è stato utilizzato sia per il rilevamento di AU che delle emozioni, sebbene le etichette per quest'ultime non siano state in alcun modo verificate; manca inoltre un confronto con i modelli statistici di riferimento, e l'uso di sottoinsiemi casuali del database originale rende la meta-analisi difficile. Per affrontare queste e altre questioni, è stato sviluppato il *Cohn-Kanade Extended (CK+) Database*. Il numero di sequenze è aumentato del 22% e il numero di soggetti del 27%, giungendo a un totale di 593 sequenze e 123 soggetti coinvolti. L'espressione obiettivo per ogni sequenza è completamente codificato con FACS e le etichette delle emozioni sono state riviste e convalidate secondo la codifica EMFACS. In aggiunta a questo, sono state aggiunte sequenze non in posa per i diversi tipi di sorrisi con associate i rispettivi metadati. Allo stato attuale, nella base di dati vengono rappresentate 30 tipi di AU e le 7 emozioni di base. Alcuni esempi dal CK+ sono riportati in Figura 1.9.

Figura 1.9: Esempi dal CK+ Database. Le immagini in alto provengono dal CK Database originale, mentre quelle in basso dalla sua estensione del 2010. Esempi di emozioni con le relative AU associate: (a) Disgusto - AU 1+4+15+17, (b) Felicità - AU 6+12+25, (c) Sorpresa - AU 1+2+5+25+27, (d) Paura - AU 1+4+7+20, (e) Rabbia - AU 4+5+15+17, (f) Disprezzo - AU 14, (g) Tristezza - AU 1+2+4+15+17, (h) Neutrale - AU0.

Un elemento fondamentale della versione aggiornata del CK Database è l'inclusione delle caratteristiche facciali globali riportate per ogni fotogramma, argomento che verrà affrontato nel Capitolo 2.

Capitolo 2

Recenti progressi nell'analisi delle espressioni facciali

2.1 Introduzione

Dal 1978, quando venne tracciato per la prima volta il movimento di 20 punti individuati su una sequenza di immagini, e più concretamente dal 1990, con l'introduzione di strumenti tecnologici e informatici avanzati, sono stati sviluppati un certo numero di approcci in ambito di visione artificiale per riconoscere AU e le loro combinazioni.

Generalmente, si adotta una procedura in tre fasi: dapprima, è necessario acquisire le immagini e riconoscere il volto contenuto in esse; a questo punto avviene l'estrazione delle caratteristiche facciali relative al volto rappresentato, con l'eventuale costruzione di trasformazioni geometriche delle stesse; infine quest'ultime vengono classificate in AU da appositi algoritmi di riconoscimento. Una trattazione completa su sistemi di questo tipo, è presentata da Li e Jain (2005).

In questo capitolo, vengono illustrati brevemente i recenti progressi nell'analisi delle espressioni facciali, presentando le varie fasi del processo in questa triplice ottica (acquisizione, estrazione e rappresentazione, riconoscimento).

I successi più rilevanti in questa tematica sono stati raggiunti con i sistemi sviluppati dalla *Carnegie Mellon University* (in seguito CMU: Lien *et al.*, 1998), *Univesity of California, San Diego* (UCSD: Bartlett *et al.*, 2001), *University of Illinois at Urbana-Champaign* (UIUC: Cohen *et al.*, 2003), *Massachusetts Institute of Technology* (MIT: Essa e Pentland, 1997), *University of Maryland* (UMD: Yacoob e Davis, 1996), *Delft University of Technology* (TUDELFT: Pantic e Rothkrantz, 2000), *Dalle Molle Institute for Perceptual Artificial Intelligence* (IDIAP: Fasel e Luttin, 2003) e altri, principalmente

della *Kyoto University* (Lyons *et al.*, 1999).

Dal momento che l'essenza del lavoro di MIT, UMD, TUDELFT, IDIAP e altri è ben sintetizzato in Fasel e Luttin (2003) e Pantic e Rothkrantz (2000), qui verranno descritti gli studi di CMU, UCSD e UIUC, sviluppati dopo il 2000.

In particolare le recenti ricerche per la realizzazione di sistemi automatici di analisi delle espressioni facciali sono orientate nelle seguenti direzioni:

- costruzione di più robusti sistemi capaci di gestire efficacemente movimenti della testa, chiusura degli occhi, cambiamenti di luce e basse intensità di espressione;
- utilizzo di un numero maggiore di caratteristiche del volto allo scopo di riconoscere più espressioni e raggiungere livellli di precisione superiori;
- riconoscimento delle AU e le loro combinazioni piuttosto di specifiche emozioni;
- realizzazione di sistemi di analisi completamente automatici.

Nella tabella 2.1 sono schematizzate le proprietà essenziali di sei sistemi provenienti dalle università sopraccitate: con l'acronimo CMU1 ci si riferisce a Kanade *et al.* del 2000; con CMU2 si indica la produzione di Cohn *et al.* del 2001; con UCSD1 si fa riferimento a Ford (2002); UCSD2 per Bartlett *et al.* del 2001; UIUC1 è descritto da Cohen *et al.* (2003); UIUC2 da Wen e Huang (2003). Una descrizione a parte è stata fatta per le metodologie di lavoro articolate da Lucey *et al.* (CMU) che nel 2010 con il CK+, realizzarono una classificazione combinando *Active Appearance Models* (AAM) e *Support Vector Machine* (SVM), come naturale sviluppo del lavoro presentato nel CMU1 e nel CMU2.

Per ogni sistema vengono definiti i metodi per ciascuna delle tre fasi di cui il sistema è composto.

	Metodo	Proprietà		
		Tempo Reale	Auto	Altro
CMU1	Individuazione volto alla prima immagine	Sì	Sì	Non gestisce movimenti della testa; volti con occhiali e capelli; più di 30 AU e combinazioni; prima immagine frontale e inespressiva
	Estrazione caratteristiche alla prima immagine	No	No	
	Rilevamento proprietà geometriche	Sì	Sì	
	Estrazione proprietà estetiche	Sì	Sì	
	Classificazione via reti neurali	Sì	Sì	
CMU2	Individuazione volto alla prima immagine	Sì	Sì	Gestisce movimenti della testa; resistente a cambiamenti di luminosità e chiusura occhi; riconosce espressioni spontanee; prima immagine frontale e inespressiva
	Modello facciale alla prima immagine	Sì	Sì	
	Rilevamento 3D della testa	Sì	Sì	
	Stabilizzazione della regione facciale	Sì	Sì	
	Estrazione proprietà geometriche	Sì	Sì	
	Classificazione basata su regole	Sì	Sì	
UCSD1	Individuazione volto ad ogni immagine	Sì	Sì	Prospettiva frontale; riconoscimento di espressioni elementari; necessità di volto neutro come base
	Ridimensionamento del volto	Sì	Sì	
	Estrazione proprietà estetiche	Sì	Sì	
	Classificazione via SVM	Sì	Sì	
UCSD2	Modellazione del volto in 3D	No	No	Gestisce movimenti della testa; necessità di base a volto neutro; riconoscimento di singole AU; valido per espressioni spontanee
	Stabilizzazione della regione facciale	Sì	Sì	
	Estrazione proprietà estetiche	Sì	Sì	
	Classificazione con HMM-SVM	Sì	Sì	
UIUC1	Definizione del modello facciale	No	No	Gestisce limitati movimenti della testa; riconoscimento di espressioni elementari; prima immagine frontale e inespressiva
	Rilevamento del volto in 3D	Sì	Sì	
	Estrazione proprietà geometriche	Sì	Sì	
	Classificazione con HMM	Sì	Sì	
UIUC2	Definizione del modello facciale	No	No	Gestisce limitati movimenti della testa; resistente a cambiamenti di luminosità; riconoscimento di espressioni elementari; prima immagine frontale e inespressiva
	Rilevamento del volto in 3D	Sì	Sì	
	Estrazione proprietà geometriche	Sì	Sì	
	Estrazione proprietà estetiche	No	Sì	
	Classificazione con NN-HMM	Sì	Sì	

Tabella 2.1: Recenti approcci nell'analisi delle espressioni facciali. Si veda la pagina precedente per una descrizione delle sigle riportate

2.2 Riconoscimento facciale

2.2.1 Rilevamento del volto

Numerosi metodi di acquisizione del volto sono stati sviluppati per identificare la presenza di volti in un contesto qualsiasi. La maggior parte di questi è in grado di rilevare pose frontali o quasi-frontali. Heisele *et al.* (2001) hanno realizzato un sistema addestrabile basato su componenti per la rilevazione di facce in posizione frontale e quasi frontale, utilizzando immagini grigie a inquadratura fissa. Rowley *et al.* (1998) hanno sviluppato un sistema basato su reti neurali per pose facciali frontali. Viola e Jones (2001) hanno progettato un rilevatore di volti in tempo reale basato su una serie di caratteristiche rettangolari. Per gestire il movimento della testa, alcuni ricercatori hanno sviluppato sensori per identificare il viso da diversi punti di vista. Pentland *et al.* (1994) sono ricorsi all'uso in tempo reale di un sistema di autospazi, in grado di gestire diverse posizioni della testa. Schneiderman e Kanade (2000) hanno proposto un metodo statistico per il rilevamento di oggetti 3D che può identificare in modo affidabile facce umane con rotazione fuori piano. Tale approccio è costituito da statistiche di aspetto dell'oggetto e di strutture fuori-oggetto tramite un prodotto di istogrammi. Li *et al.* nel 2001 hanno lavorato a un approccio di tipo *AdaBoost* per rilevare i visi da differenti prospettive. Una rassegna dettagliata sul rilevamento facciale è presentata da Ahuja *et al.* (2002).

La UCSD1 ha utilizzato il rilevatore sviluppato da Jones e Viola per riconoscere i volti per ogni fotogramma. I sistemi CMU assumono che il primo fotogramma della sequenza sia frontale e senza espressione. Essi individuano i volti solo nel primo fotogramma grazie a un identificatore facciale basato su una rete neurale.

2.2.2 Stima della posizione della testa

In un ambiente reale, i movimenti della testa fuori piano sono un elemento ricorrente nell'analisi dell'espressione facciale. Per gestire questa problematica, si cerca di calcolare la posizione della testa. Gli approcci esistenti possono essere raggruppati in metodi basati su modelli 3D e metodi basati su immagini 2D.

I sistemi CMU2, UCSD2, UIUC1 e UIUC2 hanno utilizzato un metodo basato su un modello 3D per stimare la posizione della testa.

UCSD2 ha costruito un modello facciale a maglia metallica per stimare la geometria e la posa in 3D di un insieme di punti facciali segnati a mano. I sistemi UIUC si sono serviti di un modello 3D a rappresentazione grafica

per tracciare le caratteristiche geometriche del viso definite nel modello. Il modello 3D è montato sul primo fotogramma della sequenza selezionando manualmente dei punti di riferimento, come gli angoli degli occhi e della bocca. Il modello facciale generico, che consiste di 16 macchie superficiali, è deformato per adattarsi alle caratteristiche facciali selezionate. Per stimare il movimento della testa e le deformazioni delle caratteristiche facciali, si ricorre a un processo in due fasi: il movimento viene monitorato utilizzando una sagoma che rileva corrispondenze tra fotogrammi a diverse risoluzioni. Dai movimenti in 2D di molti punti sul modello facciale, viene stimato il movimento in 3D della testa poi è stimato risolvendo un sistema di equazioni dei moti proiettivi con un algoritmo simile ai minimi quadrati.

Nel CMU2, si è fatto uso di un modello facciale cilindrico per stimare automaticamente i 6 gradi di libertà di movimento della testa in tempo reale. Un *Active Appearance Model* (di seguito AAM, da Cootes *et al.*, 1998), che verrà descritto nel Paragrafo 2.3.2, viene utilizzato per associare automaticamente il modello alla regione facciale. Per un dato fotogramma, il modello è l'immagine del volto nell'istantanea precedente che viene proiettato sul modello cilindrico. Quindi quest'ultimo viene registrato con l'aspetto del fotogramma corrente per ricostruire il movimento completo della testa. Dal momento che la struttura della testa viene recuperata utilizzando degli stampi costantemente aggiornati e la posizione stimata per il frame corrente viene utilizzata per stimare quella del frame successivo, se non vengono gestiti preventivamente si potrebbero accumulare degli errori. Per risolvere questo problema, il sistema seleziona automaticamente e memorizza uno o più frame e assume come riferimento le rispettive pose facciali.

Nell'ambito dei metodi basati su immagini, Tian *et al.* (2003), per gestire l'intera gamma di movimenti del capo, hanno rilevato la testa invece del volto. L'acquisizione è realizzata tramite una sagoma levigata dell'oggetto in primo piano con sottrazione del fondo, calcolando la curvatura negativa minima dei punti della sagoma. Dopo che la testa è stata rilevata, l'immagine viene convertita in scala di grigi, ripulita e ridimensionata alla risoluzione stimata. Successivamente una rete neurale (NN) a tre strati latenti effettua una stima della posizione della testa. La variabile d'ingresso per la rete è l'immagine elaborata. Le uscite sono tre pose facciali: (1) vista frontale o quasi-frontale, (2) vista laterale o di profilo, (3) altri, come parte posteriore della testa o volto coperto. Nella prospettiva frontale, gli angoli di occhi e labbra sono visibili. Nella laterale, almeno un occhio o un angolo della bocca non è visibile a causa della testa. Il processo di analisi di espressione viene applicata solo sulle facce vista frontale e quasi-frontale. Il loro sistema funziona bene anche con risoluzione delle immagini molto bassa.

2.3 Estrazione e rappresentazione delle caratteristiche facciali

Dopo aver rilevato il volto, il passo successivo è l'estrazione delle caratteristiche facciali. Due tipi di attributi possono essere selezionati: le caratteristiche geometriche e le caratteristiche estetiche. Le caratteristiche geometriche presentano la forma e la posizione delle componenti del viso (tra cui bocca, occhi, sopracciglia, naso). Le componenti facciali vengono estratte per formare un vettore di caratteristiche che rappresenta la geometria facciale. Le peculiarità estetiche rappresentano i cambiamenti nell'aspetto del viso (struttura della pelle), come rughe e solchi: possono essere estratti sia sull'intero volto o su regioni specifiche. Per riconoscere le espressioni facciali, un sistema automatizzato può utilizzare solo proprietà geometriche (Cohen *et al.*, 2003; Pantic e Rothkrantz, 2000), solo proprietà estetiche (Bartlett *et al.*, 2001; Ford, 2002; Zhang *et al.*, 1998), o caratteristiche ibride, sia geometriche che estetiche (Bartlett *et al.* 1999; Tian *et al.*, 2001 e 2002; Wen e Huang, 2003). La ricerca mostra che utilizzano caratteristiche ibride si possono ottenere migliori risultati per alcune espressioni. Per rimuovere gli effetti della variazioni di scala, movimento, illuminazione, e di altri fattori, si può prima allineare e normalizzare la faccia, manualmente o automaticamente, a una sagoma standard (2D o 3D), e quindi ottenere le misure normalizzate utilizzando un'immagine di riferimento (espressione neutra).

2.3.1 Estrazione delle caratteristiche geometriche

Nel CMU1, per rilevare e tener traccia dei cambiamenti nelle componenti facciali in pose frontali, si sono sviluppati modelli multistato per estrarre le peculiarità geometriche del volto. Il profilo delle labbra è stato riprodotto in 3 momenti: aperto, chiuso e serrato. Per ciascuno degli occhi viene associata una struttura a 2 stati, mentre per fronte e guance una monostato. Alcune caratteristiche estetiche come il solco naso-labiale e le rughe del contorno occhi sono rappresentate in modo esplicito con due stati esclusivi (presente/assente). Data una sequenza di immagini, la regione facciale e la posizione approssimativa degli attributi individuali del viso vengono rilevate automaticamente nella struttura iniziale. I contorni e le componenti aggiuntive vengono regolati manualmente su di essa. Dopo questa fase di inizializzazione, tutte i cambiamenti espressivi vengono rilevati e registrati nella sequenza di immagini. Questo sistema comprende 15 parametri per la parte superiore del viso e 9 per quella inferiore, che descrivono forma, movimenti e stato delle componenti del viso e delle rughe. Tutti i valori so-

no standardizzati rispetto al valore di riferimento della struttura anatomica iniziale.

Nel CMU2, il rilevamento facciale in 3D è stato applicato per gestire grandi movimenti della testa fuori piano e riconoscere componenti non rigide. La pre-elaborazione manuale della funzione di inizializzazione geometrica nel CMU1 venne ridotta utilizzando una mappatura tramite AAM. Una volta che la posa della testa è stata recuperata, la regione facciale è stabilizzata trasformando l'immagine in un orientamento opportuno per il riconoscimento dell'espressione.

I sistemi UIUC si sono serviti di un modello 3D a griglia per tracciare i punti geometrici del viso. Il modello 3D è calcolato sul primo fotogramma della sequenza selezionando manualmente i punti facciali di riferimento, come gli angoli degli occhi e della bocca. Il modello facciale generico, che consiste di 16 aree superficiali, viene deformato per adattarsi alle caratteristiche selezionate.

2.3.2 Gli Active Appearance Models

L'analisi delle espressioni realizzata dalla CMU nel 2010 sul CK+, è stata effettuata a partire da caratteristiche geometriche (Landmarks), estratte e modellate attraverso gli AAM.

Formalizzati per la prima volta da Cootes *et al.* (1998), rappresentano un buon metodo per associare un modello lineare per immagini, in grado di gestire anche variazioni lineari di forma, ad un'immagine contenente l'oggetto di interesse. In generale, l'AAM modella le componenti estetiche e di forma attraverso una ricerca a discesa del gradiente. La forma s di un AAM è descritta da una maglia di punti divisi in triangoli, viene presentata in Figura 2.1. In particolare, le coordinate dei vertici definiscono la forma $s = [x_1, y_1, x_2, y_2, ..., x_n, y_n]$, dove n è il numero dei vertici. La posizione di questi vertici funge da riferimento per l'aspetto da rappresentare, dai quali la forma viene delineata. Dal momento che gli AAM permettono variazioni lineari di forma, il vettore di punti s può essere espresso come una forma di base s_0 più una combinazione lineare di m vettori di forma s_i:

$$s = s_0 + \sum_{i=1}^{m} p_i s_i, \tag{2.1}$$

dove i coefficienti $p = (p_1, ..., p_m)^T$ rappresentano i parametri di forma. Questi parametri possono essere divisi in parametri di somiglianza rigida p_s e parametri di deformazione non rigida p_0, tali da esser rappresentabili nella forma $p^T = [p_s^T, p_0^T]$. I parametri di similarità sono associati a trasformazioni

geometriche quali traslazione, rotazione e riduzione di scala. Le misure specifiche dell'oggetto in questione risultano essere i parametri rappresentanti variazioni geometriche non rigide che determinano la forma dell'oggetto stesso (i.e. apertura della bocca, chiusura degli occhi etc.). Per stimare la forma di base s_0 si utilizza un'analisi di struttura chiamata allineamento Procuste. I fotogrammi chiave all'interno di ogni ripresa sono stati manualmente etichettati, mentre i rimanenti sono stati modellati automaticamente utilizzando un algoritmo di stima a discesa del gradiente (Matthews e Baker, 2004).

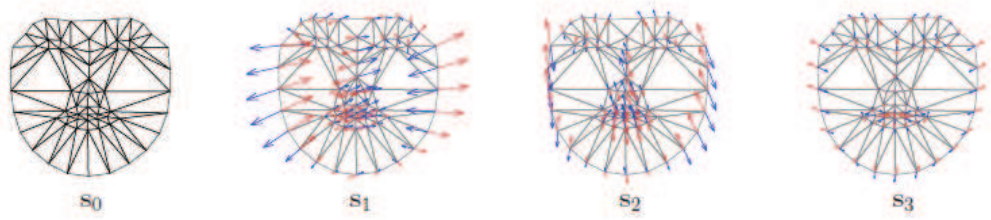

Figura 2.1: Struttura fisica di un AAM. La mappatura a triangoli s_0 e 3 possibili variazioni di forma applicabili

La struttura di similarità normalizzata s_n, è costituita da 68 punti in coordinata doppia $x-$ e $y-$ che danno origine a un vettore di 136 caratteristiche facciali per ogni immagine analizzata. Ulteriori informazioni a riguardo sono contenute nel documento di accompagnamento del database redatto da Lucey *et al.* nell'anno 2010.

2.3.3 Estrazione delle caratteristiche estetiche

Per quanto concerne l'estrazione di caratteristiche fisionomiche del volto, le *Gabor wavelets* (Daugman, 1985) sono ampiamente utilizzate in questo senso, fornendo una serie di coefficienti di scala e orientamento. Il filtro di Gabor (un filtro lineare creato per infinite combinazioni di rotazioni e dilatazioni, utilizzato diffusamente nell'elaborazione delle immagini) può essere applicato a specifiche porzioni del volto o alla sagoma intera. Zhang *et al.* (1998) sono stati i primi a confrontare due tipi di caratteristiche per il riconoscimento di espressioni: le misure geometriche di 34 riferimenti e i 612 coefficienti *wavelets* di Gabor estratti dalla stessa immagine in corrispondenza dei 34 punti. I tassi di riconoscimento per le 6 emozioni specificate erano significativamente più alti per i coefficienti *wavelets*.

Bartlett *et al.* (1999) hanno confrontato diverse tecniche per il riconoscimento di 6 AU della parte superiore e 6 della parte inferiore del viso. Queste tecniche comprendono *optical flow* (Barron e Beauchemin, 1995), analisi

delle componenti principali, analisi delle componenti indipendenti, l'analisi di caratteristiche locali, e la rappresentazione tramite Gabor *wavelets*. Le prestazioni migliori sono state ottenute da quest'ultima e dalle componenti indipendenti. Tutti questi sistemi sono stati sottoposti a un passaggio manuale per allineare ogni immagine in ingresso con una immagine facciale di riferimento utilizzando il centro degli occhi e la bocca.

Nel CMU1, Tian *et al.* (2002) hanno studiato le caratteristiche geometriche e i coefficienti di Gabor per riconoscere singoli AU e loro combinazioni, mentre Lyons *et al.* (1998), hanno utilizzato 480 coefficienti di Gabor nella parte superiore per 20 posizioni e 432 coefficienti di Gabor nella parte inferiore per altre 18 posizioni. In tale circostanza si è realizzato come i Gabor *wavelets* funzionino bene per il riconoscimento di singole in soggetti omogenei, senza movimenti della testa. Tuttavia, per il riconoscimento di combinazioni di AU nel caso di sequenze di immagini con soggetti non omogenei e piccoli movimenti del capo, i risultati usando solo le caratteristiche Gabor sono relativamente poveri. Diversi fattori possono giustificare tale differenza, in primis l'uso di soggetti sempre uguali, negli studi precedenti e il riconoscimento di specifiche emozioni o solo singole AU; inoltre, le immagini venivano ritagliate e allineate manualmente. In questo sistema viene omesso questo passaggio di pre-elaborazione e combinando coefficienti di Gabor *wavelets* con caratteristiche geometriche, le prestazioni di riconoscimento risultano maggiori per tutte le AU.

Nel UCSD1, l'immagine 2D in vista frontale del volto venne rilevata per ciascun fotogramma. Per UCSD2, la posizione in 3D e la geometria facciale vengono stimate a partire da punti etichettati a mano utilizzando un modello di rete facciale. Una volta stimata la posa in 3D, le facce vengono ruotate in posizione frontale e deformate in una geometria facciale canonica. Quindi, le immagini facciali vengono automaticamente ridimensionate e ritagliate sulla base di un volto standard con una distanza fissa tra i due occhi sia in UCSD1 che UCSD2; viene poi sottratta una faccia a espressione neutra. Entrambi gli approcci impiegano una famiglia di Gabor *wavelets* a cinque frequenze spaziali e otto orientamenti per ciascuna immagine. Invece che in determinate posizioni su di un volto, il filtro di Gabor viene applicato all'intera immagine. Per dare maggiore solidità alle condizioni di illuminazione e ai movimenti dell'immagine hanno impiegato una rappresentazione in cui i risultati di due filtri di Gabor in quadratura sono elevati al quadrato e poi sommati. Questa rappresentazione è nota come filtri di energia di Gabor ed i relativi modelli complessi come cellule della corteccia visiva primaria. Questo tipo di estrazione dati è illustrato in Bartlett *et al.* (2001).

Nel UIUC2, Wen e Huang (2003) hanno impostato un metodo basato sul rapporto d'immagine per estrarre le caratteristiche estetiche, che è indipen-

dente dalla quantità di luce riflessa dal volto di una persona. Per limitare gli effetti del rumore durante il monitoraggio e la variabilità individuale, hanno estratto le caratteristiche estetiche delle regioni del viso invece dei singoli punti, registrando poi una media ponderata come caratteristica finale per ogni regione. Per monitorare le variazioni di aspetto del viso, è stata creata una struttura a superficie piana per immagini utilizzando una mistura di modelli gaussiani. Infine si è ricorso a un algoritmo di adattamento in tempo reale per adeguare progressivamente il modello a nuove condizioni ambientali, quali cambiamenti di illuminazione o inserimento di nuovi individui.

2.4 Classificazione dell'espressione facciale

L'ultima fase di un sistema automatico è il riconoscimento delle espressioni facciali sulla base delle caratteristiche estratte. Molti classificatori sono stati applicati a tale scopo come Reti Neurali (NN), *Support Vector Machines* (SVM), Analisi Discriminante Lineare (LDA), *K-Nearest Neighbors* (K-NN), *Multinomial Logistic Ridge Regression* (MLR), *Hidden Markov Models* (HMM), *Tree Augmented Naive Bayes* (TAN) e molti altri. Alcuni sistemi utilizzano solo una classificazione basata su regole definite e automatizzate. Di seguito riassumiamo i metodi più rappresentativi, sia per il riconoscimento da immagine singola che da sequenza video. Le metodologie per immagini statiche lavorano soltanto sul fotogramma corrente con o senza un'immagine di riferimento (tipicamente un'espressione neutrale) su cui rilevare i cambiamenti. I metodi per sequenze video si servono invece anche dell'informazione temporale per riconoscere le espressioni su uno o più fotogrammi in successione. La Tabella 2.2 riassume i metodi di classificazione, il tasso di riconoscimento e i database utilizzati per valutare le procedure.

2.4.1 Riconoscimento di espressioni da immagini statiche

Il riconoscimento di espressioni da fotogrammi non considera alcuna informazione temporale associata alle immagini in ingresso. Queste possono essere immagini statiche o una istantanea di una sequenza che viene trattata in modo indipendente.

CMU1 si è servito di una rete neurale in grado di riconoscere le AU del FACS. In particolare sono state addestrate reti neurali con uno strato latente in grado di riconoscere le AU attraverso un classico metodo di *backpropagation*. Si sono utilizzate reti separate per la parte superiore e inferiore del volto. Le variabili in ingresso possono essere le caratteristiche geometriche

Sistemi	Metodo	Precisione	Database
CMU1	Rete neurale (immagini)	95.5%	Ekman-Hager (1969) Cohn-Kanade (2000)
CMU2	Basato su regole (sequenze)	100%	Frank-Ekman (1997)
UCSD1	SVM + MLR (immagini)	91.5%	Cohn-Kanade (2000)
UCSD2	SVM + HMM (sequenze)	98%	Frank-Ekman (1997)
UIUC1	BN + HMM (immagini e sequenze)	73.22% 66.53%	Cohn-Kanade (2000) UIUC-Chen (2000)
UIUC2	NN + GMM (immagini)	71%	Cohn-Kanade (2000)

Tabella 2.2: Metodi per la classificazione delle emozioni espresse tramite espressioni facciali. SVM, Support Vector Machines; MLR, Multinomial Logistic Ridge Regression; HMM, Hidden Markov Models; BN Bayes Network; GMM, Gaussian Mixture Model.

normalizzate, le caratteristiche estetiche, o entrambe, mentre le variabili risposta sono le AU riconosciute. La rete è addestrata per rispondere alle AU designate sia se si verificano singolarmente o in combinazione. Nel secondo caso, vengono generati nodi di uscita multipli. Il CMU1 fu il primo sistema in grado di gestire combinazioni di AU. Complessivamente sono state osservate più di 7.000 diverse combinazioni differenti di AU, con un tasso di riconoscimento globale del 95.5% per l'espressione neutra e 16 AU, verificatesi singolarmente o in combinazione.

Il UCSD1 ha fatto uso di un classificatore in due fasi per riconoscere un'espressione neutra e sei espressioni associate a emozioni specifiche. In primo luogo, SVM sono state utilizzate per i classificatori a coppie, vale a dire che la procedura era chiamata a distinguere tra due emozioni. In seguito hanno testato diversi approcci, come ad esempio *Nearest neighbor*, un semplice schema di voto e *Multinomial logistic ridge regression* (MLR) per convertire la rappresentazione prodotta nel primo stadio in una distribuzione di probabilità su sei determinate espressioni, più quella neutra. Le migliori prestazioni raggiungono un 91.5% di precisione.

Anche UIUC2 è ricorso a un classificatore in due fasi per riconoscere un'espressione neutra e sei espressioni relative a emozioni specifiche. Il primo passo è stato utilizzare una rete neurale per classificare tra espressione

neutrale e non neutrale. Quindi hanno costruito dei *Gaussian mixture models* (GMM) per le espressioni rimanenti. Il tasso complessivo di riconoscimento medio è stato del 71% con un campione di prova indipendente.

2.4.2 Riconoscimento di espressioni da sequenze video

La classificazione di espressioni basata su fotogrammi di sequenze video fa ampio uso delle informazioni temporali come supporto all'analisi. Le tecniche più diffuse sono state storicamente HMM, reti neurali ricorsive e classificatori pre-addestrati. I sistemi CMU2, UCSD2 e UIUC1 fanno uso di classificatori per sequenze.

Si noti che i sistemi CMU2 e UCSD2 sono studi comparativi per il riconoscimento di AU spontanee, utilizzando lo stesso database. In questi lavori, i soggetti erano etnicamente eterogenei e le AU si sono manifestate durante il discorso, con ricorrenti passaggi fuori piano e occlusioni dovute al movimento della testa e alla presenza di occhiali.

CMU2 si è servito di un classificatore basato su regole per riconoscere AU dell'occhio e della fronte che si verificavano spontaneamente, utilizzando un certo numero di fotogrammi in sequenza. L'algoritmo ha ottenuto un'accuratezza complessiva del 98% per tre comportamenti degli occhi: ammiccante (AU 45), chiuso a intermittenza, e aperto (AU 0). Una precisione del 100% è stata raggiunta nel discernere tra aperti e chiusi. L'accuratezza nelle tre categorie della regione fronte (sopracciglia alzate, abbassate e rilassate) è stata del 57%.

UCSD2 inizialmente ha utilizzato le SVM per riconoscere AU, servendosi delle rappresentazioni di Gabor; in un secondo momento hanno usato gli HMM per rilevare AU dinamiche. HMM sono state applicate in due modi: (1) tenendo le rappresentazioni di Gabor come input, e (2) prendendo i risultati delle SVM come ingresso. Quando si utilizzano le rappresentazioni di Gabor come input per addestrare gli HMM, i coefficienti di Gabor sono stati ridotti a dimensione 100 per immagine attraverso l'analisi delle componenti principali (PCA). Complessivamente sono stati improntati due HMM, uno per gestire l'ammiccamento degli occhi e uno per le situazioni restanti, utilizzando una convalida incrociata *leave-one-out*. Una delle migliori prestazioni, con un tasso di riconoscimento del 95.7%, è stato ottenuto utilizzando cinque stati e tre funzioni gaussiane. Utilizzando in ingresso i risultati delle SVM si è raggiunto un tasso di riconoscimento del 98.1%, con HMM formato da cinque stati e tre funzioni gaussiane. La precisione nelle tre categorie della regione fronte è stato pari rispettivamente al 70.1% (HMM addestrato con Gabor-PCA) e 66.9% (HMM costituito su risultati SVM). Tralasciando lo

stato con sopracciglia abbassate, la precisione aumenta al 90.9% e 89.5% per i due approcci.

Nel UIUC1, Cohen *et al.* (2003) dapprima hanno costruito reti bayesiane quali *Gaussian naive Bayes* (NB-Gaussian), *Cauchy naive Bayes* (NB-Cauchy), e *Tree-augmented-naive Bayes* (TAN) per la valutazione di immagini statiche, concentrandosi sui cambiamenti nelle ipotesi di distribuzione e sulle strutture di dipendenza. In un secondo tempo hanno proposto una nuova architettura di HMM per classificare 7 espressioni specifiche (inclusa quella neutra) da sequenze video. Con un insieme di verifica indipendente selezionato dal *Cohn-Kanade Database*, la migliore prestazione, con un tasso di riconoscimento del 73.2%, è stata raggiunta dal classificatore TAN.

Un riconoscimento tramite SVM è stato adottato dal gruppo di lavoro della CMU che ha pubblicato la versione estesa del *Cohn-Kanade Database* nel 2010, raggiungendo un risultato discreto per tutte le emozioni coinvolte, come presentato in Tabella 2.3. In questo caso sono state utilizzate le caratteristiche di forma estratte tramite AAM, per costruire un sistema basato su SVM per identificare le 7 emozioni di base teorizzate da Ekman e Friesen.

	Ang	*Con*	*Dis*	*Fea*	*Hap*	*Sad*	*Sur*
Ang	**75.0**	5.0	7.5	5.0	0.0	5.0	2.5
Con	3.1	**84.4**	3.1	0.0	6.3	3.1	0.0
Dis	5.3	0.0	**94.7**	0.0	0.0	0.0	0.0
Fear	4.4	8.7	0.0	**65.2**	8.7	0.0	13.0
Hap	0.0	0.0	0.0	0.0	**100.0**	0.0	0.0
Sad	12.0	8.0	4.0	4.0	0.0	**68.0**	4.0
Sur	0.0	0.0	0.0	0.0	0.0	4.0	**96.0**

Tabella 2.3: Matrice di confusione per il riconoscimento delle 7 emozioni di base, ottenuta da Lucey *et al.* (2010). Per ogni emozione, è stata evidenziata la classificazione con percentuale maggiore. Le etichette rappresentano: Ang = Rabbia, Con = Disrezzo, Dis = Disgusto, Fea = Paura, Hap = Felicità, Sad = Tristezza, Sur = Sorpresa. Il metodo utilizzato è un SVM multiclasse.

Capitolo 3

L'analisi dei dati

3.1 Introduzione

Dopo la disamina sulla situazione attuale nella progettazione di sistemi per l'analisi delle espressioni facciali, in questo capitolo si affronta il problema da un'ottica puramente statistica, concentrandosi sull'ultimo aspetto del processo, quello del riconoscimento delle espressioni facciali a partire da un insieme di dati in ingresso, tipicamente le caratteristiche geometriche precedentemente estratte dai fotogrammi in esame. L'attività di classificazione, come evidenziato nel Paragrafo 2.4, avviene solitamente identificando le AU coinvolte nell'espressione. Attraverso un approccio personalizzato, basato sull'analisi diretta delle relazioni esistenti fra punti geometrici del volto e caratterizzazione emozionale, si cerca di superare questa metrica canonica e di riconoscere direttamente un insieme limitato di emozioni, precisamente le 7 emozioni universali a cui si è fatto menzione nel Capitolo 1 e l'espressione neutra, senza condizionare il risultato al rilevamento delle singole AU. Con questo criterio operativo si riduce il numero di variabili in uscita del sistema (da 44 AU combinabili a 8 stati emotivi) con l'auspicio di aumentare la capacità predittiva dei modelli, dal momento che la fase di identificazione delle emozioni viene inclusa nel processo di stima e non assegnato posteriormente attraverso un sistema di codifica come quello presentato in Tabella 1.1. Di seguito si valutano alcuni tra i più diffusi metodi di classificazione per variabili risposta multinomiali, utilizzando come campione di riferimento un sottoinsieme di immagini contenute nell'*Extended Cohn-Kanade Dataset* descritto nel Paragrafo 1.5.1. Per dettagli sulla struttura del campione, riferirsi al Paragrafo 3.2.

All'interno dello studio si considerano le 7 emozioni universali che danno vita alle microespressioni facciali, secondo la teoria sviluppata da Ekman

e Friesen nel FACS, a cui viene aggiunta l'espressione neutra, che si unisce alle altre classi e in alcuni metodi funge da riferimento per l'analisi. Il taglio complessivo del lavoro è associabile al filone che si occupa del riconoscimento di espressioni facciali da immagini statiche. Si è tenuto conto della sequenza cronologica originaria esclusivamente per la scelta dei fotogrammi da includere nella campione, dal momento che le riprese rappresentano un'intensità espressiva crescente. Come verrà descritto nel prossimo paragrafo, sono state prese in esame solo immagini in cui l'emozione espressa sia definita integralmente.

Il proposito di questa fase dello studio è la valutazione delle capacità di riconoscimento dei classificatori considerati, operando direttamente sui *landmarks* estratti tramite AAM, ed identificando tramite essi le classi di emozioni rappresentate nella Figura 1.9. Questo aspetto rappresenta un elemento di assoluta novità rispetto a tutti gli studi precedenti, che si sono dedicati esclusivamente alla classificazione delle AU.

L'approccio alla questione è affine a quello presentato da Lucey *et al.* sulle espressioni facciali (2010). Tali autori sono sviluppatori di sistemi per l'analisi automatica di espressioni facciali e coordinatori del progetto che ruota attorno al database CK+. La differenza sostanziale si articola nell'utilizzo diretto dei Landmarks, nella modalità di catalogazione (in emozioni) e nell'uso estensivo e articolato di un numero consistente di modelli statistici per ottimizzare la capacità di riconoscimento, e quindi la funzionalità stessa della procedura di analisi. Per approfondimenti sui recenti sviluppi e le metodologie in uso, si faccia riferimento al Capitolo 2 e nello specifico per il database CK+ a Lucey *et al.* (2010).

Al di là del giudizio complessivo sulle prestazioni offerte dai modelli, è anche interessante individuare quali metodi offrono maggiori garanzie dinanzi a un problema di classificazione attraverso coordinate geometriche, selezionando eventualmente alcuni punti più rilevanti di altri.

Il capitolo comprende una prima parte dedicata all'analisi esplorativa (Paragrafo 3.1). Nella seconda parte (Paragrafo 3.2) si articola il processo di analisi e classificazione delle espressioni: dapprima si affronta il problema della dimensionalità elevata della matrice dei dati, effettuando un tentativo di riduzione del numero di covariate coinvolte tramite un'analisi delle componenti principali (PCA); quindi si sviluppa la modellazione del problema, presentando gli approcci di cui si è fatto uso e descrivendo il loro comportamento come classificatori di emozioni; quindi si discute il tema della selezione delle variabili, raccogliendo e confrontando le indicazioni fornite dai sistemi che hanno dimostrato maggior precisione sul piano della classificazione; infine si convalidano i modelli stimati su un'insieme di verifica indipendente, ovvero composto da immagini realizzate su soggetti diversi rispetto alla fase

di stima. Nel Paragrafo 3.3., si attua un confronto sul piano dei risultati con il precedente lavori sul CK+ di Lucey *et al. (2010)*. Le analisi statistiche sono state interamente sviluppate in *R* (R Development Core Team, 2012). Il codice relativo alle analisi viene allegato in Appendice.

3.2 Analisi esplorativa e descrizione del campione

Per le analisi sono stati selezionati fotogrammi da ciascuna sequenza realizzata e inclusa nel CK+; per ognuna sono state considerate le pose relative agli ultimi 4 istanti in cui l'espressione ha un'intensità massima, di livello D-E (descritti nel Paragrafo 1.3). Considerando i 4 istanti come realizzazioni di una singola espressione prodotta da un soggetto, si ottiene un considerevole numero di pose a disposizione per l'analisi.

In secondo luogo sono stati aggiunti 593 fotogrammi relativi alle espressioni neutre iniziali di ogni ripresa video, per un totale di 2965 unità campionarie. Ciascuna operazione di estrazione dati dal database originale è stata realizzata con l'ausilio di un programma realizzato in linguaggio Java, il cui codice principale è riportato in Appendice.

Dall'insieme sono state successivamente escluse tutte quelle pose dove l'espressione facciale risultante non fosse codificabile univocamente con una sola etichetta emozionale in base al sistema FACS, i.e. espressioni miste. Si è operata dunque la scelta di campo di valutare solo stati emotivi chiaramente definiti e con un livello di espressione non equivocabile. Il campione finale di 1881 osservazioni è stato quindi suddiviso in 3 sottoinsiemi: un primo raggruppamento di stima, corrispondente a circa l'80% del totale, per l'addestramento dei modelli, e due campioni pari al 10% ciascuno: il primo svolge una funzione principale, ovvero valutare la capacità di riconoscimento dei classificatori; il secondo viene utilizzato per eventuali stima iterative dei parametri di controllo. La suddivisione delle immagini nei 3 gruppi è avvenuta in modo completamente casuale. La presenza di immagini raffiguranti gli stessi soggetti in gruppi diversi può portare a una distosione delle stime, tale questione verrà discussa nel Paragrafo 3.3.4. Confrontando le coordinate di punti con la rispettiva immagine, si può osservare (Figura 3.1) come si ottenga una rappresentazione spaziale dell'espressione, sostanzialmente non vincolata alle specifiche caratteristiche somatiche del soggetto.

Un'analisi preliminare consiste nel valutare la distribuzione di frequenza del campione selezionato (Figura 3.2), sulla base della codifica emozionale: alcuni stati sono più rappresentati di altri, manifestando la necessità di for-

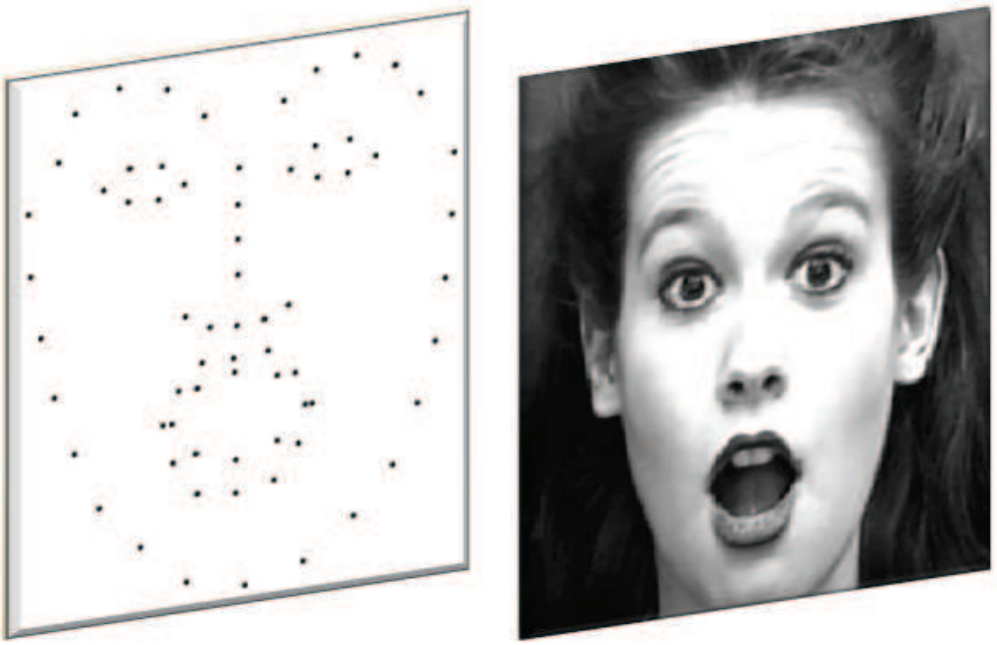

Figura 3.1: Espressione di sorpresa, stupore. Confronto fra punti estratti da un'immagine con il fotogramma originale.

nire un numero sufficiente di osservazioni per ogni classe, alle procedure di apprendimento. L'espressione neutra è invece presente per tutte le sequenze, rivestendo una particolare importanza dal momento che alcuni sistemi utilizzano una modalità della variabile categoriale come riferimento per la stime: a tale scopo una posa neutra del soggetto funge da stampo per la modellazione e allo stesso tempo da alternativa fra le etichette emozionali selezionabili dal classificatore. La codifica adottata è la stessa del CK+, ossia: 0=neutrale, 1=rabbia, 2=disprezzo, 3=disgusto, 4=rabbia, 5=felicità, 6=tristezza, 7=sorpresa.

Ad ogni osservazione sono associate 68 coppie di coordinate (x, y), come quelle raffigurate in Figura 3.1, relative a ciascun punto estratto dal volto umano presente nella foto. La posizione di ciascun punto è rappresentata in Figura 3.3. Nell'immagine si può notare come venga ricostruita la sagoma del volto umano in tutte le sue componenti mobili: tramite lo spostamento di uno o più di questi punti dalla posizione iniziale è possibile rilevare l'attività muscolare tipica delle espressioni facciali umane. I *landmarks* associati ai bordi della bocca (49, 50, 54, 55, 56, 60) e alle sopracciglia (19, 20, 21, 24, 25, 26) sono soggetti a una variazione posizionale più marcata rispetto ad altri punti del viso.

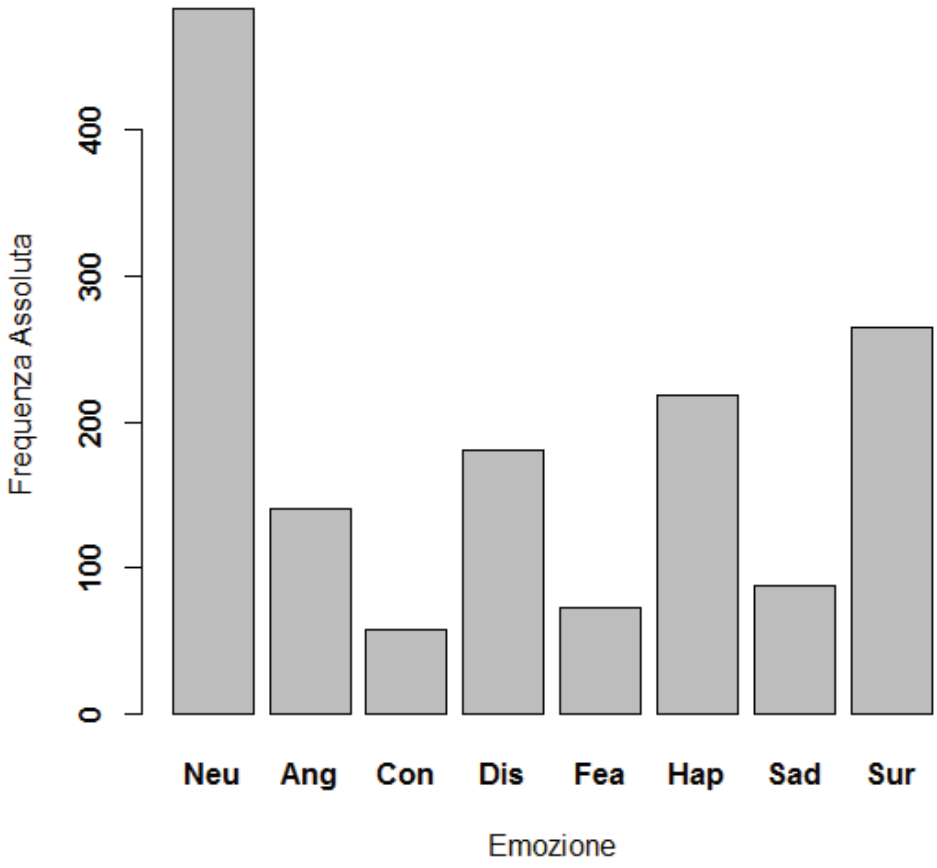

Figura 3.2: Distribuzione di frequenza per emozione dell'insieme di stima.

La variabilità delle coordinate dei punti è piuttosto eterogenea, soprattutto confrontando le misure in ascissa rispetto a quelle in ordinata, come si può vedere in Figura 3.4: quasi per ogni *landmark*, indicizzato da $i = 1, \ldots, 136$, il valore dell'ascissa x_i ha una deviazione standard superiore alla corrispondente ordinata y_i, di conseguenza ci si aspetta che i classificatori manifestino una sensibilità differente nel discriminare rispetto a spostamenti dei punti in direzione orizzontale e verticale. Fanno eccezione i punti 8, 9 e 10 (mento), 57, 58, 59 e 67 (parte inferiore della bocca).

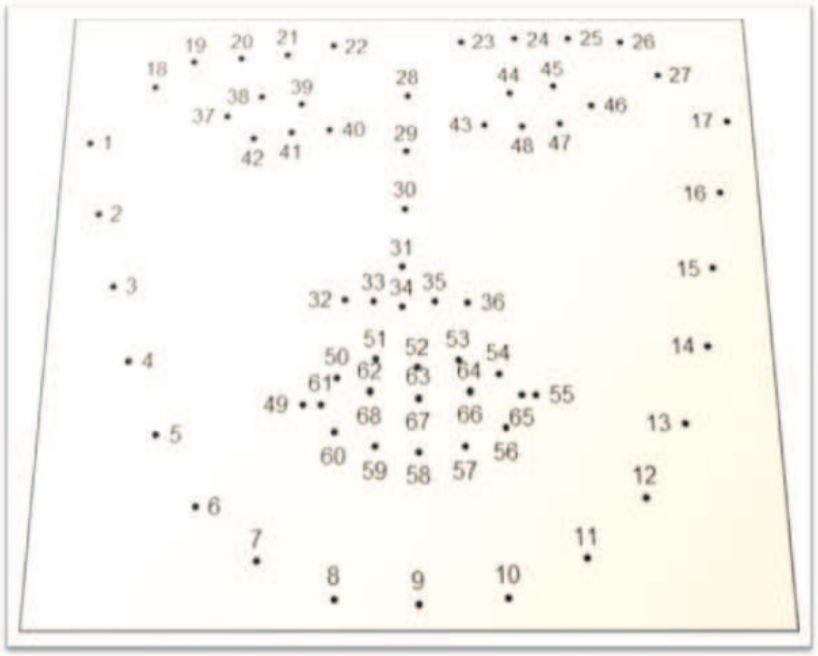

Figura 3.3: Rappresentazione dei 68 punti facciali estratti da un'immagine, raffigurante l'espressione neutra. I numeri identificano la posizione di ogni *landmarks* sul volto. Si noti come nell'immagine neutra i punti 62 e 68, 63 e 67, 64 e 66 siano a 2 a 2 sovrapposti.

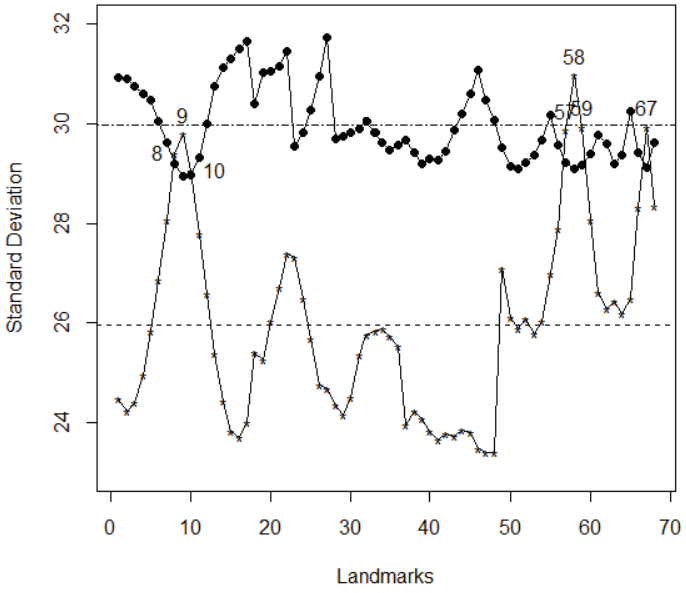

Figura 3.4: Deviazione standard dei valori per coordinata. Le corrispondenze in ascissa sono rappresentate da punti, quelle in ordinata da stelle.

La struttura finale della matrice dei dati è costituita da 1881 osservazioni, ciascuna corredata da 68 coppie di coordinate di punti definiti *landmarks* e dall'etichetta che definisce l'emozione presente, per un totale di 137 variabili, come è presentato in Figura 3.5.

	Emotion	X.Landmark 1	Y.Landmark 1	...	X.Landmark n	Y.Landmark n
1	Dis	217,79	217,08	...	352,81	330,70
2	Sur	263,46	230,32	...	360,81	390,53
3	Ang	262,26	225,81	...	347,09	347,09
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
n-1	Hap	255,62	235,24	...	264,73	364,73
n	Sur	260,66	239,09	...	351,89	351,89

Figura 3.5: Struttura dei dati ottenuta dopo la fase di estrazione dei valori relativi ai fotogrammi considerati.

3.3 Valutazione dei modelli statistici

3.3.1 Riduzione della dimensionalità dei dati

Una delle questioni più spinose di cui è necessario occuparsi prima di avventurarsi nella modellazione delle caratteristiche facciali, è senza dubbio la gestione della dimensionalità del campione. Il numero di osservazioni n è di 1881 (immagini del CK+), mentre l'insieme delle covariate p della matrice dei dati è costituito da 136 variabili quantitative contenenti coordinate normalizzate di punti, a cui si aggiunge la variabile indicatrice dell'emozione raffigurata, per una struttura dati complessiva di dimensione 1881×137. Se si considera il numero di classi (8), modelli specifici per variabili multiclasse devono affrontare operazioni di stima particolarmente onerose . E' quindi opportuno cercare una semplificazione per rendere computazionalmente più gestibile l'esecuzione degli algoritmi. Una possibilità per cercare di sintetizzare il contenuto informativo in un numero limitato di variabili è l'analisi delle componenti principali (PCA), descritta su Mardia *et al.*, 1979). Lo scopo primario di questa tecnica è la riduzione di p variabili in una quantità $k < p$ di variabili latenti: ciò avviene tramite una trasformazione lineare delle variabili che proietta quelle originarie in un nuovo sistema cartesiano nel quale la nuova variabile con la maggiore varianza viene proiettata sul primo asse,

la variabile nuova, seconda per dimensione della varianza, sul secondo asse e così di seguito fino alla p-esima. La riduzione della complessità avviene limitandosi ad analizzare le componenti principali, ovvero quelle che descrivono la maggior porzione di varianza. Dopo la trasformazione le nuove coordinate dei vettori non sono facilmente interpretabili, gli unici elementi che forniscono informazioni in questo senso sono i *loadings*, ovvero i coefficienti delle combinazioni lineari che definiscono i nuovi *scores*, i quali forniscono una misura del contributo di ogni variabile alle componenti principali.

Con i dati a disposizione, è stata realizzata una PCA con i *landmarks* estratti dalle immagini del campione di stima, al fine di valutare la possibilità di ridurre il numero di covariate, senza una considerevole perdita di informazione. Il calcolo delle componenti è stato effettuato utilizzando la matrice di correlazione. I grafici relativi alla porzione di varianza spiegata sono presentati in Figura 3.6 e Figura 3.7. Solitamente è conveniente selezionare un numero di componenti tale da raggiungere una percentuale predeterminata, o escludere quelle in cui l'apporto in termini di varianza è poco considerevole. Tale soglia è identificabile osservando l'altezza delle colonne nell'istogramma e il cosiddetto "punto di gomito" nella rappresentazione a punti congiunti, che nel nostro caso è osservabile tra la terza e la quarta componente.

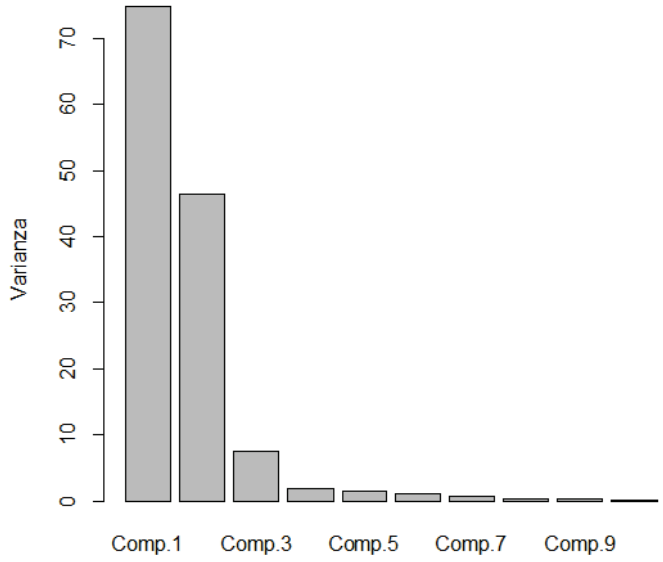

Figura 3.6: Componenti principali per il campione di stima. Porzione di varianza spiegata. Rappresentazione in forma di istogramma.

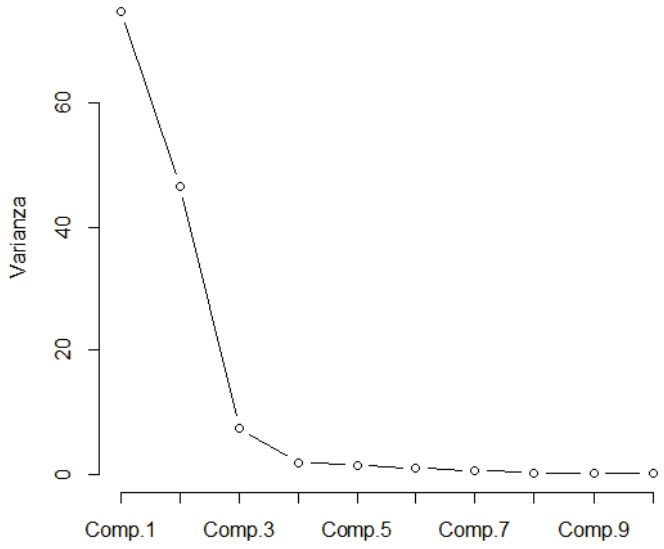

Figura 3.7: Componenti principali per il campione di stima. Porzione di varianza spiegata. Rappresentazione in punti congiunti.

Nelle prime 3 componenti si riassume il 94.7% della varianza totale presente nelle 136 coordinate di punti. La riduzione di dimensionalità è considerevole, e se sotto l'aspetto computazionale risulta conveniente, sul piano previsivo potrebbe causare una classificazione delle classi non sufficientemente precisa. In ogni caso l'inclusione di un numero più elevato di componenti non risolverebbe del tutto la questione, ragion per cui si è deciso di procedere costruendo alcuni modelli di prova con le variabili esplicative trasformate in componenti principali, e valutare come procedere in funzione dei risultati ottenuti. I pesi delle variabili all'interno delle 3 componenti selezionate sono rappresentati in Figura 3.8. In ascissa le 136 caratteristiche facciali, in ordinata il valore osservato e in legenda il tipo di linea associato ad ogni componente.

Con la matrice dei dati trasformata con la PCA, sono stati realizzati un modello LDA (Analisi Discriminante Lineare) e un modello GLM *multilogit*. Entrambi hanno registrato capacità predittive molto basse, inferiori al 50%. Si è deciso quindi di non effettuare alcuna semplificazione tramite PCA, ma di fornire ai vari modelli tutte le variabili originarie ed effettuare una selezione

di quelle significative a posteriori. Le operazioni di stima richiederanno un costo considerevole in termini di tempo, ma data la natura del problema è opportuno privilegiare una maggiore capacità di classificazione. Un ulteriore approccio che realizza una automatica selezione delle variabili è il GLM con regolazione LASSO, di cui si fa menzione nel prossimo paragrafo.

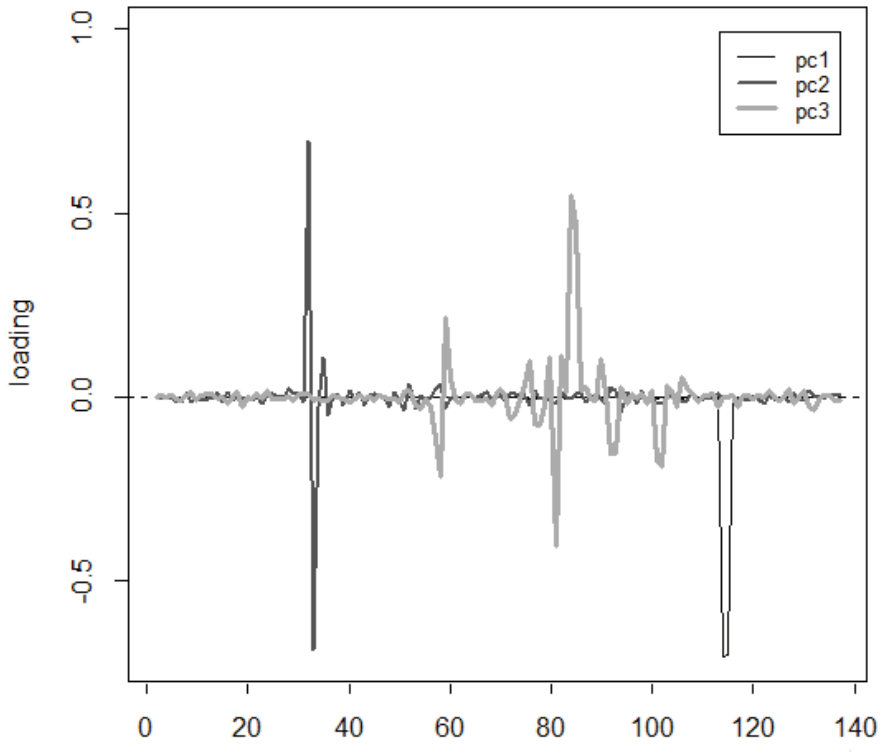

Figura 3.8: Analisi delle componenti principali. Valori dei *loadings* delle prime 3 componenti.

3.3.2 Tecniche statistiche di classificazione

In questa sezione viene affrontato dunque il problema di classificazione rispetto a una variabile risposta categoriale, composta da 8 modalità di emozione. Gli approcci metodologici predisposti coinvolgono tecniche statistiche specifiche per variabili risposta multiclasse. In alcuni casi la previsione sull'in-

sieme di verifica ha restituito le etichette delle classi, in altri una probabilità a posteriori che è stata discretizzata attraverso un voto di maggioranza. Tale operazione ha influenzato lievemente le percentuali di precisione dei modelli proposti. In Tabella 3.1 presentiamo i risultati ottenuti. Gli approcci considerati fanno riferimento a Hastie *et al.* (2009) e a Azzalini e Scarpa (2004), in particolare sono state improntate procedure riconducibili ai modelli lineari generalizzati, con o senza tecniche di regolazione dei coefficienti; metodi specifici per la classificazione, come LDA, SVM e K-NN; alberi di classificazione e classificatori combinati; strutture non parametriche come modelli additivi (GAM) e funzioni polinomiali a tratti (MARS).

Metodo	**Neu**	**Ang**	**Con**	**Dis**	**Fea**	**Hap**	**Sad**	**Sur**	**Totale**
Multilogit	88%	100%	100%	93%	100%	100%	100%	94%	**95%**
Log Linear Model	80%	100%	90%	93%	100%	100%	100%	97%	**93%**
LDA	92%	100%	90%	100%	100%	100%	89%	94%	**96%**
Tree	84%	80%	90%	77%	53%	100%	100%	94%	**86%**
Neural Networks	98%	100%	100%	96%	100%	100%	100%	97%	**98%**
GAM	84%	100%	100%	96%	100%	100%	100%	97%	**95%**
LASSO	96%	80%	100%	96%	100%	100%	89%	94%	**95%**
MARS	96%	100%	70%	96%	86%	100%	89%	94%	**94%**
SVM	98%	93%	90%	88%	73%	100%	78%	74%	**89%**
K-NN	73%	100%	90%	100%	93%	100%	100%	100%	**92%**
Boosting	96%	93%	100%	96%	93%	96%	100%	91%	**94%**
Bagging	92%	93%	100%	100%	80%	100%	89%	91%	**94%**
Random Forest	92%	93%	100%	100%	100%	100%	100%	94%	**96%**

Tabella 3.1: Prestazioni dei modelli per la classificazione delle emozioni.

Il primo approccio utilizzato è regressione logistica multinomiale, con funzione legame di tipo *multilogit*. La modalità 0 relativa all'espressione neutra funge da riferimento per la stima dei coefficienti.

$$\begin{aligned} \ln \frac{Pr(Y_i = 1)}{Pr(Y_i = K)} &= \beta_1 \cdot X_i \\ \ln \frac{Pr(Y_i = 2)}{Pr(Y_i = K)} &= \beta_2 \cdot X_i \\ &\vdots \\ \ln \frac{Pr(Y_i = K-1)}{Pr(Y_i = K)} &= \beta_{K-1} \cdot X_i \end{aligned} \tag{3.1}$$

Come descritto in Yee e Wild (1996), un modo per realizzare una regressione multinomiale è quello di immaginare, per K risultati possibili, l'esecuzione di $K-1$ modelli binari indipendenti di regressione logistica, in cui un

risultato viene scelto come base, dalla quale le $K-1$ probabilità rimanenti vengono calcolate, seguendo la (3.1). Le stime dei parametri ignoti β_i si ottengono tramite massima verosimiglianza. Per ogni osservazione, la classe assegnata è quella a cui viene associata la probabilità maggiore dal modello. Una implementazione dell'algoritmo in *R* è disponibile con il comando *vglm* contenuto del pacchetto *VGAM*, che è stato realizzato dallo stesso Yee. Sono stati inseriti tutti i 136 *landmarks* come regressori del modello, senza interazioni. Il classificatore GLM ha incontrato alcune difficoltà per riconoscere correttamente l'espressione neutra, registrando un tasso di riconoscimento inferiore al 90%; in tutti gli altri casi ha fornito prestazioni ottime, a esclusione di alcune errate classificazioni per il disgusto, assegnate a tristezza e sorpresa.

Sotto un'altra prospettiva, si possono considerare dei modelli lineari generalizzati che trattano le $n \times p$ celle di una tabella di contingenza che coinvolge variabili esplicative e modalità della variabile categoriale come osservazioni di una distribuzione casuale di Poisson. La tabella ha come colonne le modalità della variabile risposta, i cui valori di frequenza condizionati alle covariate X_i sono interpretati come conteggi. Utilizzando la funzione *multinom* del pacchetto *nnet*, le stime del modello loglineare vengono calcolate tramite delle reti neurali: alla prima classe vengono assegnati coefficienti nulli, mentre per le altre vengono stimati tanti coefficienti quanto il numero di regressori considerati (136). Il tasso di riconoscimento sul campione di verifica è molto alto per tutte le emozioni, ad eccezione dell'espressione neutra dove si incorre in un errata classificazione nel 20% dei casi.

Una formulazione particolare dei modelli di regressione logistica è stata adottata tramite massima verosimiglianza penalizzata con regolarizzazione LASSO (*Least Absolute Shrinkage and Selection Operator*). Tale procedura impone che la somma dei valori assoluti dei coefficenti calcolati con i minimi quadrati sia inferiore a un valore di soglia λ compreso tra 0 e 1. Questa penalizzazione fa sì che al descrescere del valore λ alcuni coefficienti vengano posti uguali a 0, effettuando di fatto una selezione implicita delle variabili sulla base della loro importanza all'interno del modello. Per ogni valore di λ vengono ricalcolato i valori di ciascun β. Una giustificazione di questa metodologia è contenuta in Friedman *et al.*(2008). Utilizzando l'apposita funzione contenuta nel pacchetto *glmnet* di *R*, è stato costruito un modello di regressione logistica multinomiale con penalizzazione LASSO per il nostro campione di stima, con numero massimo di iterazioni sui dati per tutti i λ pari a 200000. Il valore di λ è stato calcolato attraverso una convalida incrociata che minimizzasse l'errore di previsione, selezionando 1.41×10^{-4}. La previsione finale ha prodotto un tasso di riconoscimento del 95%, percentuali elevate per tutte le classi. Figura 3.9 presenta graficamente l'evoluzione dei

coefficienti all'interno del modello per la classe sorpresa, in funzione della soglia λ.

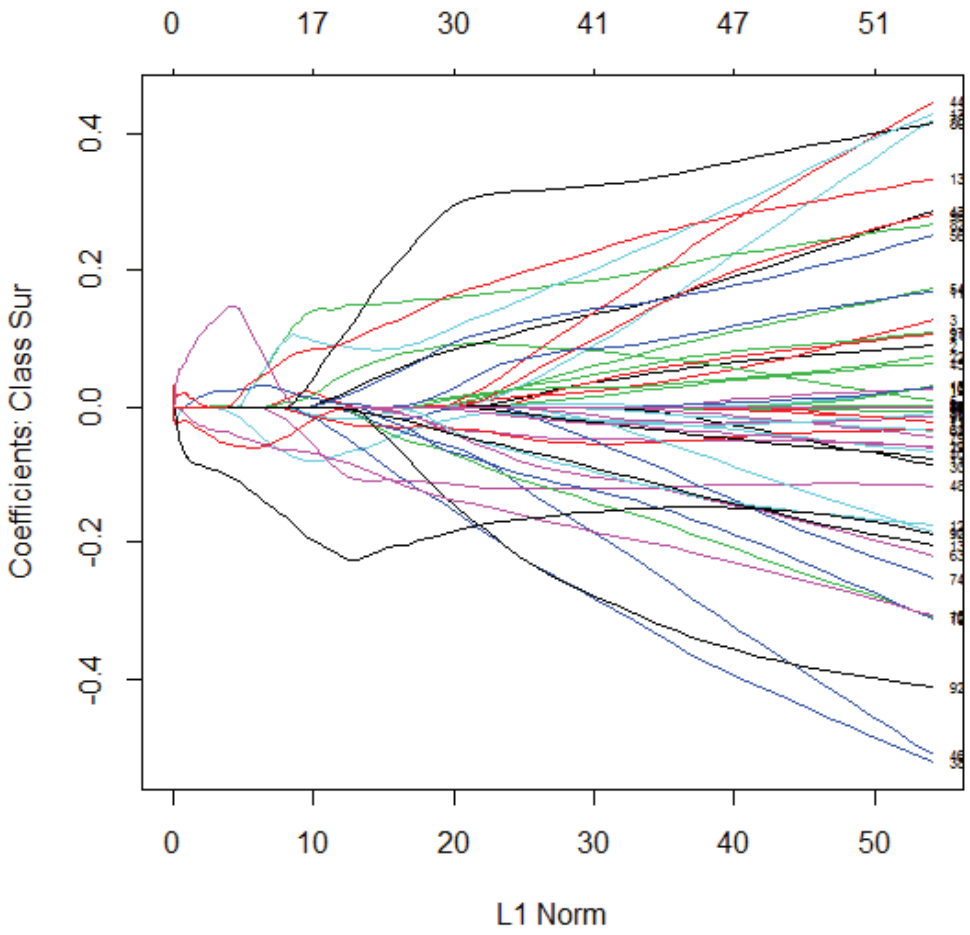

Figura 3.9: Evoluzione dei coefficienti delle variabili nel modello LASSO per la sorpresa. Si osservi come al diminuire della soglia di compressione (asse orizzontale), numerosi coefficienti collassino a 0 (asse verticale). Sulla destra le etichette delle coordinate geometriche, ordinate da 1 a 136.

L'analisi discriminante lineare è stata realizzata in *R* tramite la funzione *lda* del pacchetto *MASS*. Come probabilità a priori sono state utilizzate le proporzioni campionarie ed è stato inserito nel modello l'insieme completo delle variabili esplicative, producendo stime per i coefficienti di 7 funzioni discriminanti che associano ad ogni osservazione una probabilità a posteriori

di appartenere alle classi. La previsione sull'insieme di verifica ha prodotto un tasso di riconoscimento complessivo del 96%.

Uno soluzione efficace per problemi con variabile risposta multiclasse è costituita dagli alberi di classificazione (Breiman *et al.*, 1984) e dalle procedure sviluppate con loro combinazioni realizzate attraverso tecniche di *boosting* *bagging* e foreste casuali. Gli alberi singoli sono molto sensibili all'insieme di dati fornito in fase di addestramento e alla scelta dei parametri di modellazione, mentre i classificatori combinati forniscono risultati più robusti e spesso riescono a cogliere particolarità che altrimenti potrebbero non esser rilevate.

Il passo successivo è stato costruire degli alberi di classificazione. Per l'albero semplice, la fase di crescita è stata realizzata utilizzando come regressori tutte le coordinate geometriche disponibili e impostando i parametri di regolazione nel seguente modo: numero di osservazioni pari all'intero insieme di stima (1505), dimensione minima dei nodi pari a 2, devianza minima consentita dentro ad ogni nodo pari a 1×10^{-4}. Come misura di impurità per la variabile risposta si considera la devianza. Successivamente l'albero è stato potato tramite minimizzazione della devianza calcolata sul secondo campione di verifica, ottenendo una struttura finale di 145 nodi terminali, presentata in Figura 3.10. In termini previsivi il classificatore si dimostra impreciso nel catalogare espressioni di rabbia, disprezzo e paura, con tassi di corretta identificazione che non superano l' 80%. Tutte le procedure utilizzate sono contenute nella libreria *tree* di *R*.

La prima combinazione di alberi binari è stata effettuata con una tecnica *bagging* (Breiman, 1996), contenuta nella libreria *ipred*. È stato costruito dapprima un albero di classificazione, con almeno 20 unità per ogni nodo e profondità massima pari a 30 ramificazioni successive, che dipendono da un parametro di complessità di 0.01. La crescita dell'albero è stata realizzata tramite convalida incrociata sul campione originale suddiviso in 10 partizioni. Successivamente 50 versioni *bootstrap* del dataset di stima sono state utilizzate per crescere l'albero, utilizzando l'insieme originale per la potatura. La classificazione finale è stata effettuata per "voto di maggioranza" sui 50 alberi cresciuti con le versioni *bootstrap*. Il risultato ottenuto è un tasso complessivo di corretta classificazione pari al 94%.

Un'altra possibilità per ottenere combinazioni di modelli ad albero consiste nel considerare diversi sottoinsiemi delle variabili esplicative selezionate casualmente, ottenendo così delle stime da combinare: tale procedura prende il nome di foresta casuale (Breiman, 2001). Un'implementazione di questo algoritmo è presente in *R* nella libreria *randomForest*. Sono stati calcolati 500 alberi, con 12 variabili selezionate casualmente per effettuare la divisione ad ogni nodo, campionando dall'insieme completo delle esplicative. Con questa

Figura 3.10: Albero di classificazione singolo.

tecnica è stato riconosciuto correttamente il 96% delle espressioni contenute nel campione di verifica, con precisione superiore al 90% per tutte le classi.

Un'evoluzione del *bagging* è stata fornita con l'algoritmo *adaboost*, nella versione per variabili multiclasse (Freund e Schapire, 1996). In questo caso si utilizzano diversi insiemi di dati selezionando, per ciascuna unità, probabilità di entrata nel campione di stima non uguali: assegnando maggior peso alle osservazioni che sono state precedentemente classificate erroneamente, il classificatore viene "costretto" a lavorare maggiormente su questi punti. Inizialmente ogni osservazione ha peso $p_i = 1/N$, con N numerosità del campione. Alla iterazione t-esima si disegna un campione selezionato con le probabilità correnti $p_1, p_2, ..., p_N$, adattandolo al classificatore e calcolando tasso di errore e_t sul campione di verifica. Dato un parametro

$$\beta_t = \frac{e_t}{1 - e_t} \tag{3.2}$$

per i punti classificati correttamente si abbassano le probabilità di entrare nel campione ponendo $p_i = p_i \cdot \beta_t$ e li si rinormalizza. Questa operazione, ripetuta per molte iterazioni, stabilizza i risultati. Infine si sceglie come classificatore finale il voto ponderato con peso $\log(1/\beta_t)$ delle classificazioni ottenute. Zhu *et al.* (2009), hanno introdotto l'algoritmo SAMME, una

generalizzazione dell'*Adaboost* per variabili multinomiali, dove il peso delle osservazioni dipende anche dal numero di classi della variabile risposta. Un'implementazione di questa versione è contenuta nella libreria *adabag* di *R*, grazie alla quale abbiamo calcolato 100 iterazioni, ottenendo una precisione media sul primo campione di verifica del 94%, evidenziando una capacità di riconoscimento eccellente per tutte le tipologie di espressioni.

Un tipo di approccio completamente differente è quello delle reti neurali (libreria *nnet*), ovvero strutture non lineari calcolate per modellare relazioni complesse tra un insieme di variabili di ingresso e i valori assunti in uscita dalle osservazioni corrispondenti, secondo uno schema che simula i reali circuiti neuronali degli esseri viventi. Tale sistema si compone di 3 strati interconnessi tramite funzioni non lineari o parzialmente non lineari, costituiti da variabili di ingresso, variabili latenti e variabili d'uscita. Dettagli tecnici sulla metodologia si possono trovare in Ripley (1996, Capitolo 5). Per quanto concerne il modello realizzato per le espressioni facciali, i parametri di controllo sono stati stimati iterativamente al fine di ottenere la combinazione che riducesse al minimo l'errore di classificazione sull'insieme di verifica secondario: *weight decay* pari 0.107, 8 unità nello strato latente, 1168 pesi per le connessioni tra i nodi e un massimo di 2000 iterazioni per la convergenza dell'algoritmo, che è stata effettivamente raggiunta. Il tasso di errore sul campione di verifica principale è pari all'1.6%, con precisione massima per rabbia, disprezzo, paura, gioia e tristezza.

Proseguendo con approcci non parametrici, è stato costruito un modello additivo GAM per variabili multinomiale, seguendo la proposta di Yee e Wild (1996). In questa modellazione dei dati, il legame tra variabile dipendente e ciascun predittore è non lineare. Si ipotizza che il valore atteso condizionato di Y data la matrice X di esplicative è uguale alla somma di funzioni incognite delle variabili esplicative più un termine costante; la stima delle funzioni avviene tramite una procedura iterativa conosciuta come algoritmo di *backfitting*, sostanzialmente una variante dell'algoritmo di Gauss-Seidel, mentre la variabile risposta viene trasformata con una funzione legame di tipo *logit* come nel caso della regressione multinomiale visto in precedenza. Ricorrendo all'implementazione contenuta nella libreria *vgam*, sono state inserite nel vettore di variabili da lisciare tutte le caratteristiche geometriche a disposizione, come classe di riferimento è stata scelta l'espressione neutra. Per limitare il carico computazionale è stato fissato a 30 sia il numero massimo di iterazioni compiute dall'algoritmo di *backfitting*, sia per le iterazioni di *Newton-Raphson*, calcolate per risolvere localmente le equazioni delle funzioni che caratterizzano il modello. Una volta categorizzate le classi predette sull'insieme di verifica (per voto di maggioranza), si è ottenuto un tasso di errore del 5%, quasi totalmente relativo a errate classificazioni di immagini

con espressione neutra.

Un altra struttura per modellare problemi con molte variabili esplicative è il MARS (*Multivariate adaptive regression splines*), una specificazione iterativa delle *spline* di regressione. Per ogni variabile esplicativa si determina una coppia di basi che operano localmente; il loro prodotto tensoriale, combinato a coefficienti β stimati minimizzando la somma dei quadrati dei residui, dà luogo a una particolare forma di modello lineare che automaticamente gestisce non linearità e interazioni tra le variabili. In una prima fase di analisi si sono adattati i dati a una versione di MARS sviluppata da Stone *et al.* (1997), specifica per problemi di classificazione multiclasse. Tale formulazione, che prende il nome di *PolyMARS*, avente una struttura logistica multipla, costruisce il modello con una procedura *forward stagewise* come nei MARS classici, dove però ad ogni ciclo la funzione di verosimiglianza viene sostituita da una approssimazione quadratica per cercare la coppia di basi successive. Una volta aggregata la nuova coppia al modello, si ristimano tutti i coefficienti via massima verosimiglianza e si passa al ciclo seguente. L'applicazione della tecnica sul nostro insieme di dati (libreria *polspline* di *R*) ha fornito un tasso di errore del 18%, non equiparabile ai risultati ottenuti attraverso gli altri classificatori. Si è scelto dunque di costruire un modello MARS più tradizionale per variabili continue. Rriferendosi a quello introdotto da Friedman (1991), si è fatto ricorso all'implementazione *earth* contenuta nell'omonimo pacchetto *R*. La procedura MARS attua una selezione automatica delle variabili di ingresso, includendo nel modello finale 80 predittori su 136 e 129 funzioni di base. La valutazione sul campione di verifica produce un tasso di riconoscimento del 94%, con percentuali sub-ottimali per quasi tutte le classi.

Le SVM sono state utilizzate da Lucey *et al.* (2010) per valutare le capacità di riconoscimento espressivo sul CK+. La formulazione è opera di Cortes e Vapnik (1995): l'algoritmo di classificazione opera individuando dei margini che massimizzano la distanza tra gli elementi delle diverse modalità, ammettendo un numero limitato di eccezioni. L'insieme di punti che determinano la posizione dei margini è detta *support vector*; la funzione obiettivo considera trasformate delle variabili originarie, che entrano nel modello tramite prodotti interni che seguono la struttura delle funzioni nucleo più diffuse. Nel caso delle espressioni, è stata utilizzata una funzione radiale di parametro γ pari a 7.35×10^{-3} e un costante di penalizzazione per i punti eccedenti di valore 2, ottenuta tramite una convalida incrociata costituita da 20 partizioni dell'insieme di stima. Il numero totale di vettori di supporto calcolati (funzione *svm* della libreria *e1071* di *R*) per le 8 modalità è 1141, per un errore di classificazione dell'11% sul campione di verifica; le principali difficoltà sono state riscontrate per le classi disgusto, paura, tristezza e sorpresa.

Come ultimo approccio all'analisi delle espressioni facciali è stato scelto l'algoritmo più semplice fra quelli utilizzati nell'apprendimento automatico: i *K-nearest neighbors*. Un oggetto è classificato in base alla maggioranza dei voti dei suoi k vicini. Se $k = 1$ allora l'oggetto viene assegnato alla classe del suo vicino. Secondo la teoria di Dasarathy (1991), ai fini del calcolo della distanza gli oggetti sono rappresentati attraverso vettori di posizione in uno spazio multidimensionale. Di solito viene usata la distanza euclidea, ma anche altri tipi di distanza sono ugualmente utilizzabili, ad esempio la distanza *Manhattan*. Il classificatore assegna un punto alla classe C se questa è la più frequente fra i k esempi più vicini all'oggetto sotto esame, dato un insieme di oggetti per cui è nota la classificazione corretta. Il valore del parametro k ottimale nel caso corrente è stato stimato tramite convalida incrociata, utilizzando la procedura apposita contenuta nella libreria *ipred*. Costruito un modello con $k = 2$ e distanza euclidea, si è ottenuta una percentuale di corretta classificazione nel 92% dei casi, con difficoltà predittive solo per la classe neutra. Nei casi in cui i due punti più vicini appartengano a classi diverse, l'osservazione viene assegnata casualmente a una delle due emozioni.

Ultimata la fase di costruzione dei classificatori, si può effettuare una valutazione complessiva dei risultati ottenuti. Se si prende in considerazione il comportamento generale nelle 8 classi, l'apprendimento per l'espressione felice è stato ottimale, mentre le immagini neutre hanno creato difficoltà a tutti i classificatori. Nella maggior parte dei casi, l'errata classificazione è stata provocata da espressioni neutre catalogate come emozioni distinte, ovvero da una minore sensibilità nel discernere un volto espressivo da uno emozionalmente indifferente. Questo tipo di carenza predittiva si può osservare nella matrice di dispersione del modello logistico, presentata in Tabella 3.2. Questo può esser dovuto a due ragioni: in primis, sposando la teoria di Ekman e Friesen (1971), la struttura espressiva del volto umano è determinata dall'attivazione di muscoli facciali che causano una variazione rispetto alla morfologia naturale del soggetto, ragion per cui i punti geometrici di riferimento modificano le proprie coordinate in relazione a un posizionamento di base originario (quello definito dall'espressione neutra). Laddove dunque la traslazione nello spazio non risulti particolarmente accentuata, le procedure faticano a distinguere e catalogare correttamente. Una seconda ragione può ricercarsi nella non indipendenza complessiva delle osservazioni selezionate, dovuta alla ripetizione dei soggetti coinvolti. Mentre per il primo assunto non si opererà nessun approfondimento, considerandolo parte integrante della natura stessa del problema, per la seconda questione verrà accertato nel Paragrafo 3.3.4 se l'uso di campioni di stima e verifica indipendenti conduce a risultati sensibilmente diversi.

La scelta del modello migliore non è però così banale, dal momento che

	Neu	*Ang*	*Con*	*Dis*	*Fea*	*Hap*	*Sad*	*Sur*
Neu	**45**	0	0	0	0	0	0	1
Ang	1	**15**	0	0	0	0	0	1
Con	1	0	**10**	0	0	0	0	0
Dis	1	0	0	**25**	0	0	0	0
Fea	1	0	0	0	**15**	0	0	0
Hap	0	0	0	0	0	**28**	0	0
Sad	1	0	0	1	0	0	**9**	0
Sur	1	0	0	1	0	0	0	**32**

Errore Totale: 5.3%

Tabella 3.2: Classificazione sull'insieme di verifica ottenuta con il modello di regressione logistica multinomiale. Matrice di confusione per il riconoscimento delle 8 emozioni considerate. Le colonne corrispondono ai valori effettivi, le righe a quelli predetti. La maggior parte delle errate classificazioni coinvolge immagini raffiguranti l'espressione neutra.

più di un approccio ha fornito risultati 'ottimali'. In alcuni casi, come le reti neurali, la struttura di classificazione è difficilmente interpretabile ed è sovente soggetta a sovradattamento dei dati. Uno strumento relativamente obiettivo per misurare le prestazioni complessive dei classificatori binari è il valore AUC (*Area Under the ROC Curve*), ovvero una stima della percentuale di area sottesa dalla curva ROC (*Receiver Operating Characteristic*), che è una rappresentazione della proporzione di veri positivi (sensibilità) rispetto al numero di positivi effettivi, associata alla proporzione di falsi positivi rispetto al totale dei negativi effettivi (1-specificità), al variare del valore di soglia che determina le classi. Questa struttura, usuale strumento di studio per classificazioni binarie, è stato riadattata al caso multiclasse da Hand and Till (2001) e implementata nella libreria *pROC*. In questa versione, si considerano i valori di sensibilità e specificità per tutti i confronti binari tra le classi ed una volta calcolati i singoli AUC, si tiene conto della media come misura complessiva per tutte le categorie. I valori relativi ai modelli che hanno fornito previsioni più precise sono riportati in Tabella 3.3.

Valutando stabilità dei risultati, interpretabilità, prestazioni previsive e valore AUC, fra tutti i modelli considerati selezioniamo il classificatore LDA, la regressione logistica LASSO ed i classificatori combinati tramite *boosting* e foreste casuali come i più adeguati per l'analisi effettuata sulle espressioni facciali della base di dati CK+. Per queste 4 modellazioni si porterà avanti l'analisi andando a indagare quali variabili risultano maggiormente

Modello	*AUC*	*Tasso di riconoscimento*
Multilogit	93.20%	94.7%
LDA	97.36%	95.8%
Neural Networks	95.53%	98.4%
LASSO	98.28%	95.2%
Boosting	95.26%	93.7%
Random Forest	94.52%	96.7%

Tabella 3.3: Valore AUC multiclasse e tasso di riconoscimento per i modelli selezionati.

informative.

3.3.3 Importanza e selezione delle variabili

Un aspetto fondamentale quanto il raggiungimento di una considerevole abilità previsiva, è l'identificazione delle variabili sulla base della loro rilevanza statistica all'interno del modello, che dal punto di vista interpretativo significa stabilire un ordinamento delle caratteristiche geometriche sulla base del loro utilità per il riconoscimento dell'espressione. Utilizzando degli indici per quantificare l'importanza delle covariate, vengono presentati alcuni criteri di selezione e ordinamento per i modelli che hanno ottenuto le prestazioni di classificazione migliori. In particolare gli approcci a cui si fa riferimento in questo paragrafo non devono aver riscontrato difficoltà in nessuna delle classi.

Per LDA, è stata adottata una procedura di selezione (*stepwise*) che opera in entrambe le direzioni (*backward-forward*). Il modello iniziale è definito dalle variabili di partenza; ad ogni passo vengono generati nuovi modelli includendo ogni singola variabile che non è presente, ed escludendo ogni singola variabile che è già stata inserita. Le prestazioni per i sottomodelli sono valutate tramite a convalida incrociata, e se il massimo miglioramento ottenuto è superiore a un valore di soglia, la variabile corrispondente è inclusa o esclusa. La procedura si interrompe, se il nuovo valore massimo di miglioramento non è sufficiente, o se viene raggiunto il numero massimo di variabili. Un'implementazione di questa procedura si può effettuare con la funzione *stepclass* della libreria *klaR*, e nel caso in esame è stata utilizzata una soglia di miglioramento di 0.01 per ogni iterazione dell'algoritmo. La soglia è stata scelta in modo che a ciascun cambiamento di stato delle variabili corrispondesse un miglioramento sensibile nella qualità del modello. I risultati sono riportati nella Tabella 3.4. Il numero di variabili nel modello finale è di 12 volte in-

feriore rispetto all'insieme completo e racchiude solo coordinate verticali al suo interno. Il tasso di corretta classificazione sull'insieme di verifica (85%) e l'AUC complessivo (90%) con questa configurazione delle variabili è peggiore rispetto al modello saturo, però i tempi di calcolo sono di 93 volte inferiori, rendendo la fase di adattamento dei dati al modello decisamente più veloce.

	Step	*Variabile*	*% Riconoscimento*
0	start	–	0.00
1	in	Y. Landmark58	0.38
2	in	Y. Landmark50	0.55
3	in	Y. Landmark34	0.66
4	in	Y. Landmark49	0.70
5	in	Y. Landmark6	0.74
6	in	Y. Landmark68	0.76
7	in	Y. Landmark62	0.78
8	in	Y. Landmark31	0.80
9	in	Y. Landmark22	0.81
10	in	Y. Landmark41	0.83
11	in	Y. Landmark36	0.85

Tabella 3.4: Procedura *stepwise* di selezione delle variabili per LDA. Nella prima colonna il numero dell'iterazione, nella seconda il tipo di *step* affrontato (inserimento o rimozione), nella terza il nome della variabile e nella quarta il tasso di riconoscimento.

Il GLM con penalizzazione LASSO effettua una selezione implicita delle variabili, dal momento che la riduzione dello spazio dei parametri induce alcuni coefficienti, ovvero quelli meno rilevanti per una struttura funzionale del modello, ad assumere valore nullo. Fissato quindi il valore di soglia λ, si ottiene una combinazione di coefficienti diversi da 0 per ciascuna modalità della variabile risposta. L'applicazione al nostro insieme di dati ha fornito il risultato in Tabella 3.5. Le covariate senza coefficienti di valore assoluto almeno pari a 0.2 sono state escluse dalla selezione e considerate di peso trascurabile.

Var	*Neu*	*Ang*	*Con*	*Dis*	*Fea*	*Hap*	*Sad*	*Sor*
X.Landmark7		0.2						
Y.Landmark8							-0.2	
Y.Landmark9	-0.2	0.1		0.1	-0.3			
X.Landmark18					0.3			
Y.Landmark19		0.1						-0.2
Y.Landmark22			-0.1		-0.1	0.1	-0.2	0.1
Y.Landmark23				0.2			-0.2	-0.1
Y.Landmark24	0.2			0.3				-0.1
Y.Landmark31		-0.1					0.2	
X.Landmark32		-0.4					-0.1	
Y.Landmark32			0.1	-0.3				
Y.Landmark34		-0.3		0.5				
Y.Landmark36				-0.4			0.2	-0.1
X.Landmark39		0.1			0.2			
Y.Landmark40	-0.1							0.3
Y.Landmark41		-0.3		-0.1				0.2
Y.Landmark42			0.3	-0.2				
X.Landmark43	0.2			0.1				
Y.Landmark44					-0.2			
X.Landmark45							-0.3	
X.Landmark46	-0.1	0.2						
Y.Landmark46			0.1					-0.2
X.Landmark49				0.2	-0.2			
Y.Landmark49			-0.3			-0.3	0.7	
Y.Landmark50	0.1			-0.3				-0.1
Y.Landmark51	-0.1	0.6	0.5					
Y.Landmark53		0.2						
Y.Landmark54				-0.2				
X.Landmark55		-0.1	0.1					-0.2
Y.Landmark55						-0.3	0.4	
X.Landmark56		-0.3						
X.Landmark60					-0.2			0.1
Y.Landmark60		-0.3			0.3			-0.1
Y.Landmark62	0.3						-0.1	
Y.Landmark63				0.1				-0.2
X.Landmark65				-0.1	0.2	0.1		
Y.Landmark66								0.2
Y.Landmark67						0.3		
X.Landmark68			-0.2					-0.1
Y.Landmark68					0.1		-0.6	0.2

Tabella 3.5: Selezione delle variabili per il GLM con regolazione LASSO. I coefficienti stimati rappresentano i pesi con cui le variabili influenzano la variabile risposta, per ciascuna emozione.

Il numero di variabili esplicative si riduce sensibilmente a 40 coordinate di punti, 13 ascisse e 27 ordinate. Alcune entrano nel calcolo di diverse emozioni, altre assumono valori non nulli per poche modalità. I *landmarks* Y.49 e Y.68 hanno una forte influenza per la classificazione della tristezza, il Y.51 per rabbia e disprezzo.

La modellazione combinata degli alberi di classificazione con tecniche di boosting ha un criterio di selezione delle variabili che associa a ciascuna un punteggio in funzione dell'importanza nel processo di classificazione. In particolare questa misura considera l'indice di Gini registrato da una variabile all'interno di un albero e il peso di tale albero nella struttura di classificazione. I risultati ottenuti dal modello per le espressioni facciali sono riportati in Tabella 3.6.

	Variabile	*Importanza*
1	Y. Landmark19	3.51
2	Y. Landmark41	2.89
3	Y. Landmark22	2.36
4	Y. Landmark42	2.14
5	Y. Landmark29	1.88
6	Y. Landmark43	1.85
7	Y. Landmark8	1.76
8	Y. Landmark5	1.73
9	Y. Landmark12	1.56
10	Y. Landmark4	1.55
11	X. Landmark65	1.52
12	Y. Landmark38	1.50
13	Y. Landmark62	1.39
14	Y. Landmark6	1.38
15	Y. Landmark54	1.26
16	Y. Landmark35	1.21
17	Y. Landmark13	1.17
18	Y. Landmark37	1.09
19	Y. Landmark36	1.03

Tabella 3.6: Selezione delle variabili per classificatori combinati via *boosting*. Le variabili con punteggi maggiori sono le più rilevanti per il riconoscimento delle espressioni. Vengono riportate le variabili con valori superiori a 1.

Anche questa procedura ha posto più in alto nella gerarchia delle variabili principalmente coordinate Y dei punti facciali: unica eccezione l'ascissa del

landmark65.

L'ultima verifica per l'identificazione delle esplicative che svolgono un ruolo principale per la classficazione in categorie emozionali è svolta sul modello combinato con la tecnica delle foreste casuali. In questo caso si misura l'importanza di una variabile valutando la perdita di precisione causata dall'esclusione della stessa dal modello. Anche in quest'ottica, i punteggi più alti (Tabella 3.7) vengono associati a un sottoinsieme di ordinate.

	Variabile	*Mean Decrease Accuracy*	*Mean Decrease Gini*
1	Y. Landmark67	0.70	37.89
2	Y. Landmark58	0.66	31.15
3	Y. Landmark59	0.63	25.68
4	Y. Landmark68	0.62	25.61
3 5	Y. Landmark55	0.67	24.93
6	Y. Landmark49	0.68	24.76
7	Y. Landmark57	0.62	23.57
8	Y. Landmark66	0.57	23.08
9	Y. Landmark24	0.62	21.58
10	Y. Landmark21	0.59	20.95
11	Y. Landmark50	0.64	19.23
12	Y. Landmark23	0.58	18.98
13	Y. Landmark54	0.63	18.72
14	Y. Landmark22	0.57	18.27
15	Y. Landmark20	0.56	18.26
16	Y. Landmark25	0.55	17.40
17	Y. Landmark63	0.57	15.63
18	Y. Landmark65	0.56	15.27

Tabella 3.7: Selezione delle variabili per importanza. Foreste casuali, con ordinamento per diminuzione media della precisione e indice di Gini.

Una valutazione complessiva dei risultati evidenzia un aspetto centrale all'interno dell'intera analisi: la prevalenza di coordinate verticali nell'insieme delle variabili più rilevanti. Questo comportamento è indicativo di come da un punto di vista mimico sia più difficile realizzare movimenti orizzontali, rendendo le rispettive componenti meno informative per identificare il tipo di emozione rappresentata. Le zone più "espressive", sono l'arcata sopracciliare e la bocca, mentre i margini laterali non sono quasi mai considerati dai modelli. Una parziale interpretazione a questo riscontro empirico si può

ricercare nella maggiore mobilità di queste 2 componenti rispetto alle altre. Un altro elemento interessante riguarda l'importanza dell'asse centrale del volto, in corrispondenza della zona naso-labiale (punti 31-34, 63-67). Generalmente l'asimmetria (prevalenza di punti rilevanti sul lato destro o sinistro del volto) non si configura come una caratteristica peculiare nell'identificazione delle espressioni, ad eccezione che per l'emozione del disprezzo. Per un' analisi delle corrispondenze facciali, presentiamo in Figura 3.11 la rappresentazione su volto inespressivo delle variabili significative per il modello LDA. A causa di ovvie limitazioni grafiche, è stato evidenziato l'intero punto e non la singola coordinata indicata dalla procedura di selezione.

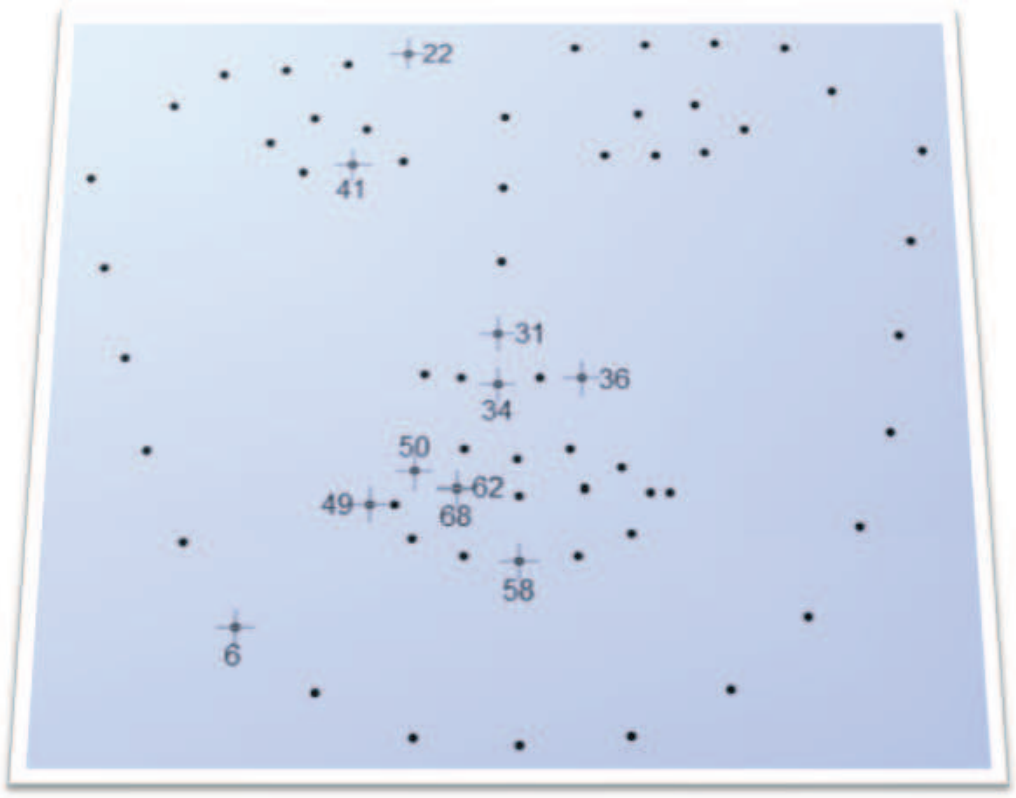

Figura 3.11: Rappresentazione dei 68 punti facciali estratti da un'immagine raffigurante l'espressione neutra. I numeri e le croci evidenziano la posizione dei *landmarks* selezionati dalla procedura *stepwise* per il modello LDA. Essendo un'immagine neutra, i punti 62 e 68 coincidono.

3.3.4 Valutazione con insieme di verifica indipendente

Come discusso in precedenza, una questione che mina l'attendibilità dei risultati contenuti in Tabella 3.1 riguarda la non indipendenza delle osservazioni, dovuta all'utilizzo di un numero limitato di soggetti per le pose. La variabilità delle caratteristiche di 20 misurazioni relative allo stesso soggetto è circa *10 volte inferiore* a quella di 20 osservazioni scelte casualmente, questo può portare a delle valutazioni non veritiere. Per contrastare questo effetto indesiderato, è consigliabile effettuare nuovamente le stime e valutare i modelli su un'insieme di verifica non più selezionato casualmente dall'intero campione, ma composto da sole unità relative a soggetti non inclusi nell'insieme di stima: in questo modo i raggruppamenti di apprendimento e di valutazione risultano indipendenti. È opportuno fornire una quantità di dati appropriata per ciascun algoritmo ed al contempo rendere comparabile il nuovo esperimento con il precedente, in termini di numerosità del campione. In quest'ottica le proporzioni definite precedentemente nel Paragrafo 3.2 per i due gruppi sono rimaste inalterate. Nella Tabella 3.8 sono riportati i risultati ottenuti dai classificatori con questa nuova configurazione.

Metodo	*N*eu	*A*ng	*C*on	*D*is	*F*ea	*H*ap	*S*ad	*S*ur	***Totale***
GLM Logit	93%	100%	100%	100%	100%	100%	100%	100%	**98%**
Log Linear Model	90%	100%	88%	93%	100%	100%	100%	97%	**96%**
LDA	87%	100%	94%	100%	100%	100%	100%	100%	**96%**
Tree	93%	100%	100%	100%	83%	100%	94%	100%	**95%**
Neural Networks	83%	100%	100%	100%	100%	100%	100%	100%	**96%**
Lasso	93%	100%	100%	100%	100%	100%	100%	100%	**98%**
MARS	90%	100%	93%	100%	100%	100%	100%	100%	**95%**
SVM	100%	100%	88%	100%	100%	100%	100%	100%	**98%**
K-NN	73%	100%	100%	100%	96%	100%	100%	100%	**93%**
Boosting	96%	100%	100%	100%	100%	100%	100%	100%	**98%**
Bagging	93%	100%	100%	100%	100%	100%	100%	100%	**98%**
Random Forest	87%	100%	100%	100%	100%	100%	100%	100%	**97%**

Tabella 3.8: Classificazione con insieme di verifica indipendente: i soggetti utilizzati in questa analisi sono nuovi rispetto alla fase di addestramento.

I modelli di classificazione dimostrano una capacità predittiva non inferiore a quella manifestata attraverso verifica con pose di soggetti casuali, registrando un tasso di errata classificazione non superiore al 7%. Gli approcci che hanno offerto una prestazione migliore sono stati GLM (sia classici che con regolazione LASSO), SVM e i classificatori combinati. La classe che ha creato maggiori difficoltà permane quella delle espressioni neutre, sicuramente la più impegnativa da un punto di vista strutturale per le ragioni descritte nel Paragrafo 3.3. Si considerano i risultati ottenuti complessivamente più soddisfacenti rispetto a quelli ottenuti nella precedente elaborazione.

La dipendenza delle immagini dalla fisionomia del soggetto non sembra dunque influire troppo sui risultati, probabilmente a causa della netta definizione dellle espressioni (i fotogrammi selezionati raffigurano stati emozionali a marcata intensità espressiva) e il numero ingente di variabili a disposizione per modellare il problema.

Da un punto di vista metodologico, i *landmarks* sono stati estratti da un formato foto standard, limitando largamente le incongruenze individuali nei contorni del volto. In virtù del riscontro empirico, si è deciso di non considerare la ripetizione dei soggetti raffigurati come un aspetto di rilevanza tale da giustificare modifiche sostanziali nella composizione del campione e nella valutazione complessiva dei risultati ottenuti.

3.4 Confronto con i risultati precedenti

Una classificazione molto simile avvenuta tramite SVM è stata effettuata dagli stessi creatori della collezione di sequenze, come riportato nel capitolo precedente in Tabella 2.3. Il riconoscimento delle emozioni di Lucey *et al.* (2010) è avvenuto utilizzando due tipi di trasformazioni dei *landmarks*, i quali sono stati adattati a un classificatore SVM addestrato a identificare le singole AU. Successivamente, utilizzando uno schema di conversione in emozioni per combinazioni di AU, è stata costruita la matrice di confusione finale per le emozioni. I risultati ottenuti sono stati complessivamente migliori rispetto allo studio del 2010, sia utilizzando la stessa metodologia SVM che improntando soluzioni differenti, ragion per cui si ritiene che per un'analisi mirata al riconoscimento diretto delle emozioni, sia più opportuno lavorare con gli strumenti proposti in questo elaborato. Confrontando l'attuale classificatore SVM con quello realizzato dal gruppo della *Carnegie Mellon University*, le variazioni percentuali ottenute per ciascuna emozione sono:

- Per la *rabbia*, +24% di precisione rispetto allo studio originario
- Per il *disprezzo*, +6.6%
- Per il *disgusto*, -7%
- Per la *paura*, +11.9%
- Per la *felicità*, 0%, dal momento che in entrambi gli studi si è raggiunto il 100% di corrette classificazioni
- Per la *tristezza*, +14.7%
- Per la *sorpresa*, -22.9% ,

manifestando un miglioramento complessivo di circa il 6%, a cui va aggiunta una capacità di riconoscimento del 98% per le espressioni neutre. Dal momento che, escludendo l'albero singolo, tutti gli altri approcci hanno ottenuto percentuali più elevate, i progressi rispetto al sistema di Cohn *et al.* sono piuttosto sensibili.

Le differenze sostanziali rispetto agli studi precedenti riguardano gli elementi di cui ci si è serviti per sviluppare l'analisi, ovvero le basi da cui ha avuto inizio la classificazione: i *landmarks* del viso. I punti di riferimento che sono stati estratti sono la forma più semplice per sintetizzare le sembianze di un volto umano, sono universali e rintracciabili in qualunque foto con soggetto in primo piano frontale; possono essere tracciati automaticamente da un computer o manualmente da un esperto. Confrontando i risultati con le SVM applicate ai dati del CK+, si evince come, per il riconoscimento delle emozioni, la classificazione diretta dei punti facciali abbia una resa superiore in termini previsivi, rispetto alla classificazione intermedia delle AU.

Un secondo elemento di distinzione riguarda la tipologia di studio delle relazione fra punti ed espressione facciale. Un requisito essenziale per una buona riuscita dell'analisi è l'utilizzo di espressioni definite integralmente e appartenenti a una delle sette categorie di base di cui si è parlato nel Capitolo 1. Da un punto di vista statistico è possibile estendere le classi, perdendo però il supporto bibliografico relazionato all'universalità delle emozioni primarie (Fridlund *et al.*, 1987). Un via più complessa riguarda lo studio di intensità basse (livelli A e B), utile per indagini e situazioni dove l'evidenza dell'emozione non è così forte e non può essere individuata facilmente attraverso le naturali capacità dell'occhio umano, o lo studio di combinazioni di emozioni primarie, molto attuale e apprezzato in ambito professionale e investigativo: in entrambi i casi è spesso necessario ricorrere al riconoscimento intermedio delle AU attivate e successivamente "interpretare" i risultati in termini di comportamento emozionale, tuttavia la capacità dei classificatori di assegnare una probabilità di appartenenza a ciascuna classe non esclude sviluppi in questo senso. L'uso di procedure statistiche avanzate e diversificate, ha permesso la costituzione di classificatori ad alto tasso di riconoscimento e al contempo versatili nel gestire la dimensionalità e la natura spaziale del problema.

Capitolo 4

Un approccio bayesiano: K-NN probabilistici

4.1 Introduzione

Questo capitolo descrive lo sviluppo e l'implementazione di un approccio nuovo per la classificazione delle espressioni facciali. Il metodo statistico di riferimento sono i K-NN presentati nel Capitolo 3, rivisitati però in ottica bayesiana. Una caratteristica fondamentale della nuova formulazione è la creazione di un modello probabilistico completo, in cui i problemi di trattabilità associati alla funzione di verosimiglianza vengono superati attraverso simulazioni Monte Carlo sulle distribuzioni condizionate delle variabili.

La prima parte del capitolo è dedicata alla discussione e all'implementazione del nuovo modello. Partendo da una conversione dell'usuale algoritmo dei K-NN in una struttura probabilistica i cui elementi godono della *proprietà di Markov*, si illustra un'implementazione tramite *Markov Chain Monte Carlo*, in seguito MCMC (si veda ad esempio Robert e Casella, 2010), che fornisce una stima a posteriori dei parametri del modello e permette di effettuare l'analisi delle emozioni rappresentate nel CK+; una volta calcolata la stima bayesiana della probabilità a posteriori per ciascuna nuova osservazione, si effettua la classificazione sull'insieme di dati e si calcola il tasso di riconoscimento ottenuto.

Nella seconda parte si affronta il problema della selezione delle variabili per il modello probabilistico, implementando un metodo MCMC che valuta il contributo di ciascuna variabile tramite una procedura *stepwise* simile a quella utilizzata nella regressione lineare. In questo caso la scelta di includere o escludere una variabile avviene sulla base delle probabilità a posteriori di modelli che includono ed escludono ciascuna variabile ad ogni passo dell'algo-

ritmo. Le tecniche statistiche applicate sono state implementate in linguaggio R e il relativo codice è riportato in Appendice.

4.2 Formulazione del modello

4.2.1 Una versione probabilistica dei K-NN

La procedura classica dei K-NN permette di realizzare una classificazione semplice e con costi computazionali proporzionali alla taglia dell'insieme di dati in esame. Riprendendo la formulazione di Dasarathy (1991), questa tecnica richiede unicamente il calcolo della matrice delle distanze e di classificare le nuove osservazioni tramite un voto di maggioranza sulle classi dei k punti più vicini, ma non costituisce una modellazione statistica in termini propri.

La difficoltà di strutturare i K-NN in forma probabilistica è essenzialmente dovuta alla mancanza di una relazione simmetrica tra osservazioni consecutive, ovvero di connessione tra i primi n elementi e l'$n+1$-esimo elemento inserito. Questa associazione tra le osservazioni rappresenta un requisito fondamentale per l'esistenza di una distribuzione di probabilità congiunta per il nuovo campione di dimensione $n+1$. In altre parole, qualunque insieme completo di distribuzioni condizionate basata sul classico sistema K-NN "asimmetrico" non è compatibile con una distribuzione congiunta.

Una soluzione per ottenere una interpretazione probabilistica di questo metodo, è modificare la struttura originaria dei K-NN rendendo simmetrica la relazione fra punti "vicini". Esemplificando, se x_i appartiene all'insieme dei k valori più prossimi a x_j e la relazione inversa non si verifica (ovvero x_j non appartiene all'insieme dei punti più prossimi a x_i), il punto x_j viene aggiunto ugualmente all'insieme dei K-vicini di x_i. L'insieme di punti vicini ottenuto è chiamato *symmetrized k-nearest neighbor system* (Marin e Robert, 2007, Capitolo 8). Come illustrato nella Figura 4.1, il punto centrale ha 3 vicini più prossimi tali che nessuno di questi lo contenga a sua volta nell'insieme dei propri 3 punti più prossimi. Nella nuova versione a relazione simmetrica, il punto centrale viene invece a sua volta incluso nell'insieme dei vicini più prossimi dei 3 punti.

Una volta aggiornata questa definizione di "vicino", si denota con $i \sim_k j$ l'appartenenza di i all'insieme dei k vicini più prossimi di j e la conseguente relazione inversa. A questo punto è possibile proporre una distribuzione di probabilità condizionata completa, come avviene nel *Gibbs sampling* (si veda ad esempio Robert e Casella, 2010, Capitolo 7), per regolare l'appartenenza delle osservazioni alle classi.

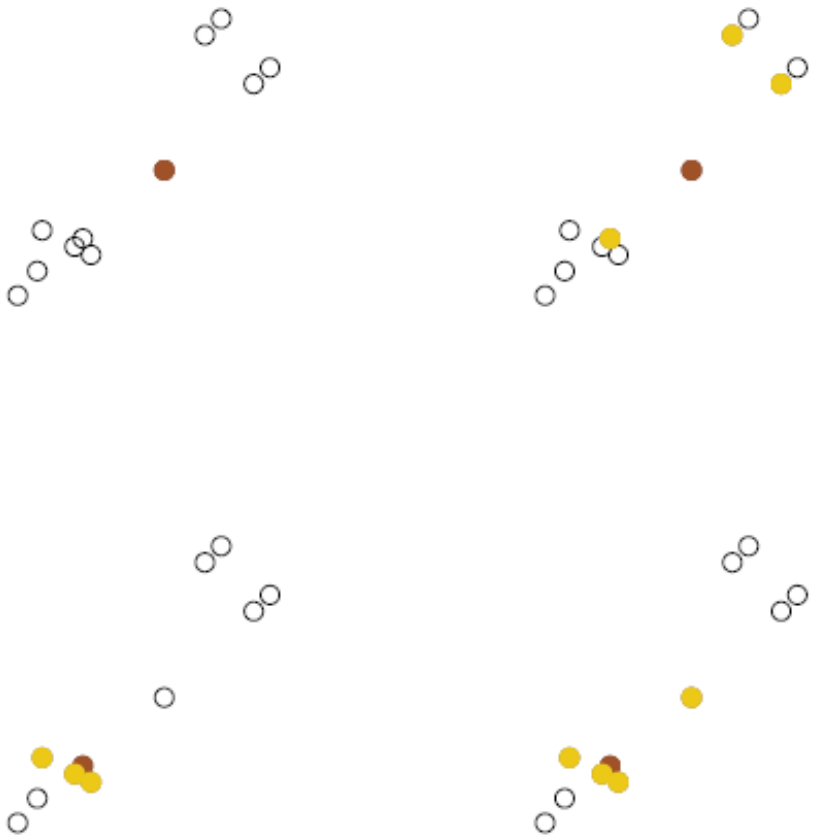

Figura 4.1: Rappresentazione del processo di simmetrizzazione, fonte Marin e Robert (2007). Nel riquadro in alto a sinistra, un campione di stima e un punto centrale x_1; in alto a destra, il punto x_1 e i suoi 3 vicini più prossimi, x_2, x_3 e x_4; in basso a sinistra, uno dei 3 punti (x_2) con i rispettivi 3 vicini più prossimi; in basso a destra, la relazione simmetrica include x_1 nell'insieme dei vicini più prossimi di x_2.

Adottando una approccio logistico simile al modello multinomiale logistico definito nel Paragrafo 3.3.2, si può definire una probabilità per ogni osservazione y_i nella forma

$$\mathbb{P}(Y_i = C_j|\mathbf{y}_{-i}, \mathbf{X}, \beta, k) = \frac{\exp\left(\beta \sum_{l\sim_{kj}} \mathbb{I}_{C_j}(y_l) \Big/ N_k\right)}{\sum_{g=1}^{G} \exp\left(\beta \sum_{l\sim_{kj}} \mathbb{I}_{C_g}(y_l) \Big/ N_k\right)} \tag{4.1}$$

dove N_k rappresenta la dimensione dei k vicini più prossimi considerati, $\beta > 0$, $y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ e $\mathbb{I}_{C_j}(y_l)$ indica l'appartenenza dell'osservazione y_l alla classe C_j, che dipende dalla distanza euclidea all'interno delle covariate $X = \{x_1, \ldots, x_n\}$.

A differenza del modello logistico tradizionale, quest'ultimo è parametrizzato da (β, k) ed è definito solo dalle sue distribuzioni condizionate, che vengono utilizzate anche per la previsione della classe di una nuova osservazione y_{n+1}, una volta noto il relativo vettore di covariate x_{n+1}. Il parametro β serve per misurare il grado di influenza della classe più rappresentata nei vicini, dal momento che $\beta = 0$ corrisponde a una distribuzione uniforme per tutte le classi, mentre $\beta = +\infty$ induce con probabilità massima alla classe prevalente.

Definendo la questione da un punto di vista bayesiano, data una distribuzione a priori $\pi(\beta, k)$ di supporto $[0, \beta_{max}] \times \{1, \ldots, K\}$, la distribuzione marginale di $y_n + 1$ risulta

$$\begin{aligned} &\mathbb{P}(Y_{n+1} = C_j|\mathbf{x}_{n+1}, \mathbf{y}, \mathbf{X}) = \\ &\sum_{k=1}^{K} \int_0^{\beta_{max}} \mathbb{P}(Y_{n+1} = C_j|\mathbf{x}_{n+1}, \mathbf{y}, \mathbf{X}, \beta, k)\pi\left(\beta, k|\mathbf{y}, \mathbf{X}\right) d\beta, \end{aligned} \tag{4.2}$$

dove $\pi\left(\beta, k|\mathbf{y}, \mathbf{X}\right)$ è la distribuzione a posteriori dato l'insieme $(\mathbf{y}, \mathbf{X})$. Questa soluzione formale ha però il grosso problema del calcolo della funzione di verosimiglianza $f\left(y|x, \beta, k\right)$: infatti, dal momento che la costante di normalizzazione è dipendente da β e k, non possiede una forma esplicita direttamente calcolabile.

Come approssimazione si utilizza la pseudo-verosimiglianza

$$\hat{f}\left(\mathbf{y}|\mathbf{X},\beta,k\right)=\prod_{g=1}^{G}\prod_{y_i=C_g}\mathbb{P}(y_i=C_g|\mathbf{y}_{-i},\mathbf{X},\beta,k), \tag{4.3}$$

formata dal prodotto delle probabilità condizionate di ciascuna osservazione di appartenere alla sua vera classe. Nella (4.3) è presente una approssimazione, dal momento che l'integrazione di $\hat{f}\left(\mathbf{y}|\mathbf{x},\beta,k\right)$ non vale 1 e la costante di normalizzazione non è conosciuta. Si può ovviare al problema utilizzando un algoritmo MCMC, che fornisce una stima della distribuzione a posteriori $\hat{\pi}\left(\beta,k|\mathbf{y},\mathbf{X}\right)\propto\hat{f}\left(\mathbf{y}|\mathbf{x},\beta,k\right)\pi\left(\beta,k\right)$ e la pseudo-distribuzione per le previsioni

$$\begin{aligned}&\hat{\mathbb{P}}(y_{n+1}=C_j|\mathbf{x}_{n+1},\mathbf{y},\mathbf{X})=\\&\sum_{k=1}^{K}\int_0^{\beta_{max}}\mathbb{P}(y_{n+1}=C_j|\mathbf{x}_{n+1},\mathbf{y},\mathbf{X},\beta,k)\hat{\pi}\left(\beta,k|\mathbf{y},\mathbf{X}\right)\,d\beta.\end{aligned} \tag{4.4}$$

4.2.2 K-NN Metropolis-Hastings Sampler

Il calcolo esplicito della (4.4) è difficilmente trattabile ed è richiesta una approssimazione MCMC, basata su una catena di Markov $\left((\beta^{(1)},k^{(k)},\right),\ldots,\left(\beta^{(M)},k^{(M)}\right))$, con distribuzione stazionaria $\hat{\pi}\left(\beta,k|y,x\right)$, da cui simulare un elevato numero di volte.

In un primo momento, potrebbe sembrare intuitivo utilizzare uno schema di *Gibbs sampling*. Non disponendo però delle distribuzioni condizionate di β e k, è ragionevole proporre un algoritmo *Metropolis-Hastings* (si veda ad esempio Robert e Casella, 2010, Capitolo 6). L'aggiornamento di β e k avviene utilizzando distribuzioni proposta con passeggiata casuale. Dal momento che $\beta\in[0,\beta_{max}]$, si può ricorrere a una trasformazione logistica del parametro θ, in cui β, invertendo la relazione, assume una struttura del tipo

$$\beta^{(t)}=\beta_{max}\exp\left(\theta^{(t)}\right)\Big/\left(\exp\left(\theta^{(t)}\right)+1\right), \tag{4.5}$$

simulando i valori di θ con passeggiata casuale da $N\left(\theta^{(t)},\tau^2\right)$, senza restrizioni al supporto. Per l'aggiornamento di k, si può utilizzare una proposta uniforme su r vicini di $k^{(t)}$, ovvero su uno spazio $\left\{k^{(t)}-r,\ldots,k^{(t)}+r\right\}\cap\{1,\ldots,K\}$. La distribuzione delle proposte per k è quindi $Q_r\left(k,\cdot\right)$ con funzione di probabilità $q_r\left(k,k^{'}\right)$ che dipende dal parametro di scala r. Il corrispondente algoritmo MCMC risultante per stimare i parametri del modello KNN è presentato in Figura 4.2:

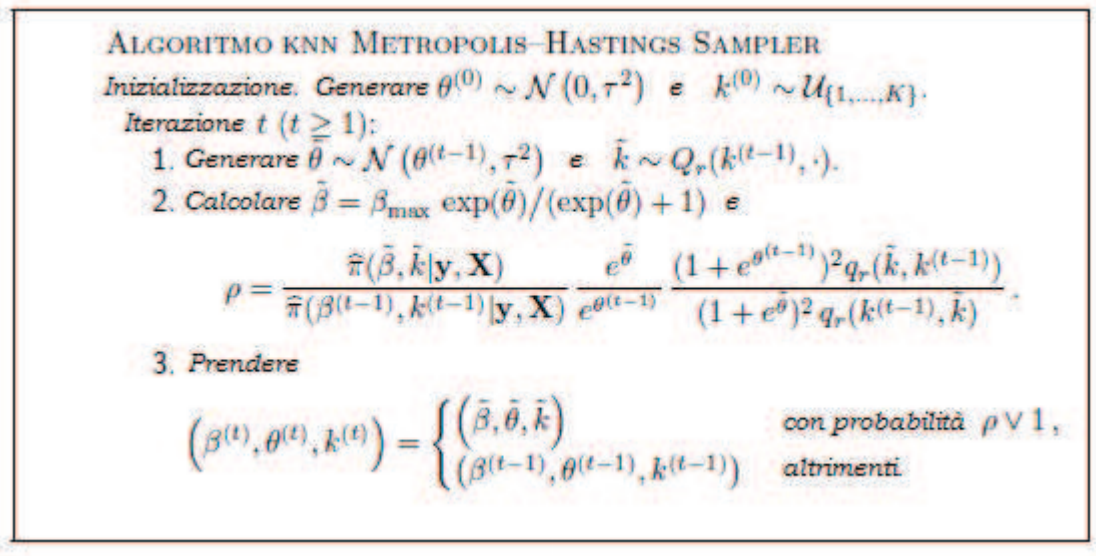

Figura 4.2: Algoritmo di campionamento *Metropolis-Hastings* per il modello K-NN.

Utilizzando questa procedura, è possibile ottenere la classe con maggiore probabilità a posteriori dato un vettore x_{n+1} di covariate, tramite

$$\frac{1}{M}\sum_{i=1}^{M}\hat{\mathbb{P}}\Big(y_{n+1} = l|x_{n+1}, y, x, (\beta^{(i)}, k^{(i)})\Big), \tag{4.6}$$

In cui M rappresenta il numero di simulazioni effettuate dopo la fase iniziale di calibrazione dell'algoritmo, definita comunemente *burn-in*. Il risultato è una approssimazione della stima MAP (*maximum* a posteriori) della probabilità di y_{n+1} di appartenere alla classe l.

4.2.3 Applicazione alle espressioni facciali

L'algoritmo MH illustrato al punto precedente è stato utilizzato per fornire una classificazione delle emozioni rappresentate nelle immagini del database CK+.

La funzione che lo implementa, riportata in Appendice, richiede in ingresso la matrice dei dati (variabile risposta ed esplicative), un valore di β_{max}, un numero massimo K di punti vicini considerati, il numero di simulazioni da calcolare e un valore per i parametri di scala τ e r delle proposte.

Nel caso delle espressioni, i parametri di scala sono scelti in modo da ottenere una coppia di valori accettabile per la dimensionalità del problema

($\tau = 0.5$, $r = 2$), ovvero tale che l'algoritmo avesse un tasso di accettazione adeguato. Come distribuzione a priori non informativa per β e k è stata scelta una uniforme definita sul loro dominio. Il numero di simulazioni calcolate è stato fissato a 5000, il valore massimo di β e K rispettivamente a 15 e 70: per il primo può essere scelto un valore positivo a piacere sufficientemente grande affinchè il supporto dei β venga esplorato senza limitazioni superiori, mentre il secondo viene fissato in funzione della numerosità del campione.

La simulazione ha prodotto una stima delle medie a posteriori per i parametri del modello pari a $(\hat{\beta}, \hat{k}) = (6.5, 3)$ I grafici nella Figura 4.3 e 4.4 presentano le distribuzioni a posteriori marginali per β e k. I valori di β presentano una distribuzione lievemente asimmetrica lungo il dominio $(0, \beta_{max})$ mentre il parametro k oscilla quasi essenzialmente su valori piccoli, privilegiando la scelta di un sottoinsieme ridotto di punti vicini. Effettuando la previsione delle classi sullo stesso campione utilizzato per le analisi del Capitolo 3, si è ottenuto un errore di classificazione per le nuove osservazioni pari al 7%, con alti tassi di riconoscimento per tutte le emozioni, ad esclusione dell'espressione neutra che risulta classificata erroneamente nel 22% dei casi. L'etichetta delle classi è stata assegnata selezionando per ogni osservazione la modalità della variabile risposta Y con probabilità a posteriori più elevata. In Tabella 4.1 viene presentata la matrice di confusione per il campione di punti facciali.

	Neu	*Ang*	*Con*	*Dis*	*Fea*	*Hap*	*Sad*	*Sur*
Neu	**77.7**	0.0	0.0	0.4	0.0	0.0	0.0	1.8
Ang	3.9	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0
Con	4.0	0.0	**100.0**	0.0	0.0	0.0	0.0	0.0
Dis	4.7	0.0	0.0	**99.6**	0.0	0.0	0.0	0.6
Fea	2.2	0.0	0.0	0.0	**100.0**	0.0	0.0	0.0
Hap	2.0	0.0	0.0	0.0	0.0	**100.0**	0.0	0.0
Sad	3.3	0.0	0.0	0.0	0.0	0.0	**100.0**	0.0
Sur	2.2	0.0	0.0	0.0	0.	0.0	0.0	**97.6**

Tabella 4.1: Classifiazione con K-NN probabilistico. Matrice di confusione per il riconoscimento delle 8 emozioni considerate. Le righe corrispondono ai valori predetti, le colonne ai valori osservati. Per ogni emozione, è stata evidenziata la classificazione con percentuale maggiore. Le etichette rappresentano: Neu = Neutrale, Ang = Rabbia, Con = Disrezzo, Dis = Disgusto, Fea = Paura, Hap = Felicità, Sad = Tristezza, Sur = Sorpresa.

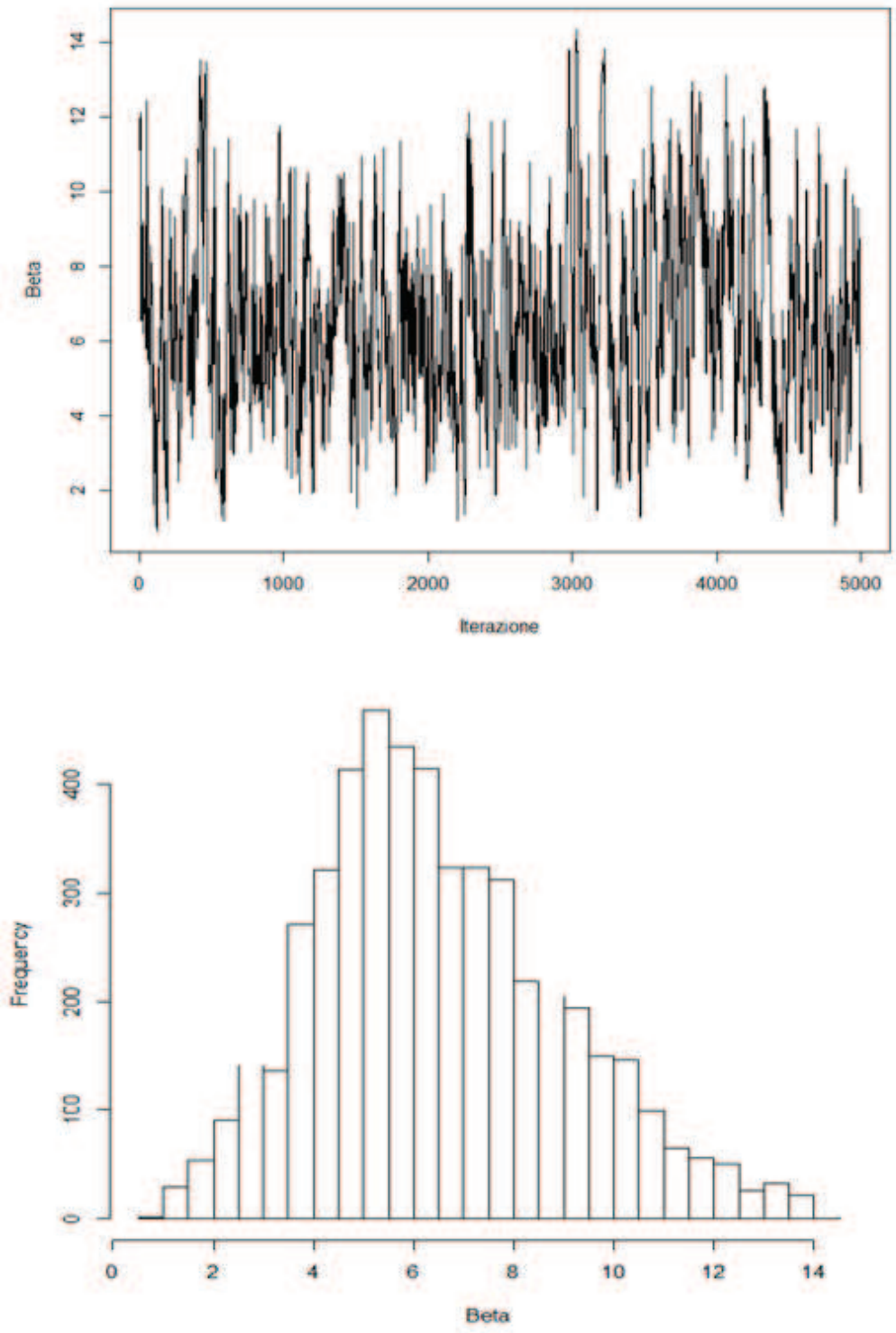

Figura 4.3: Sequenza di 5000 simulazioni con il K-NN Metropolis-Hastings. Distribuzione marginale a posteriori marginale di β.

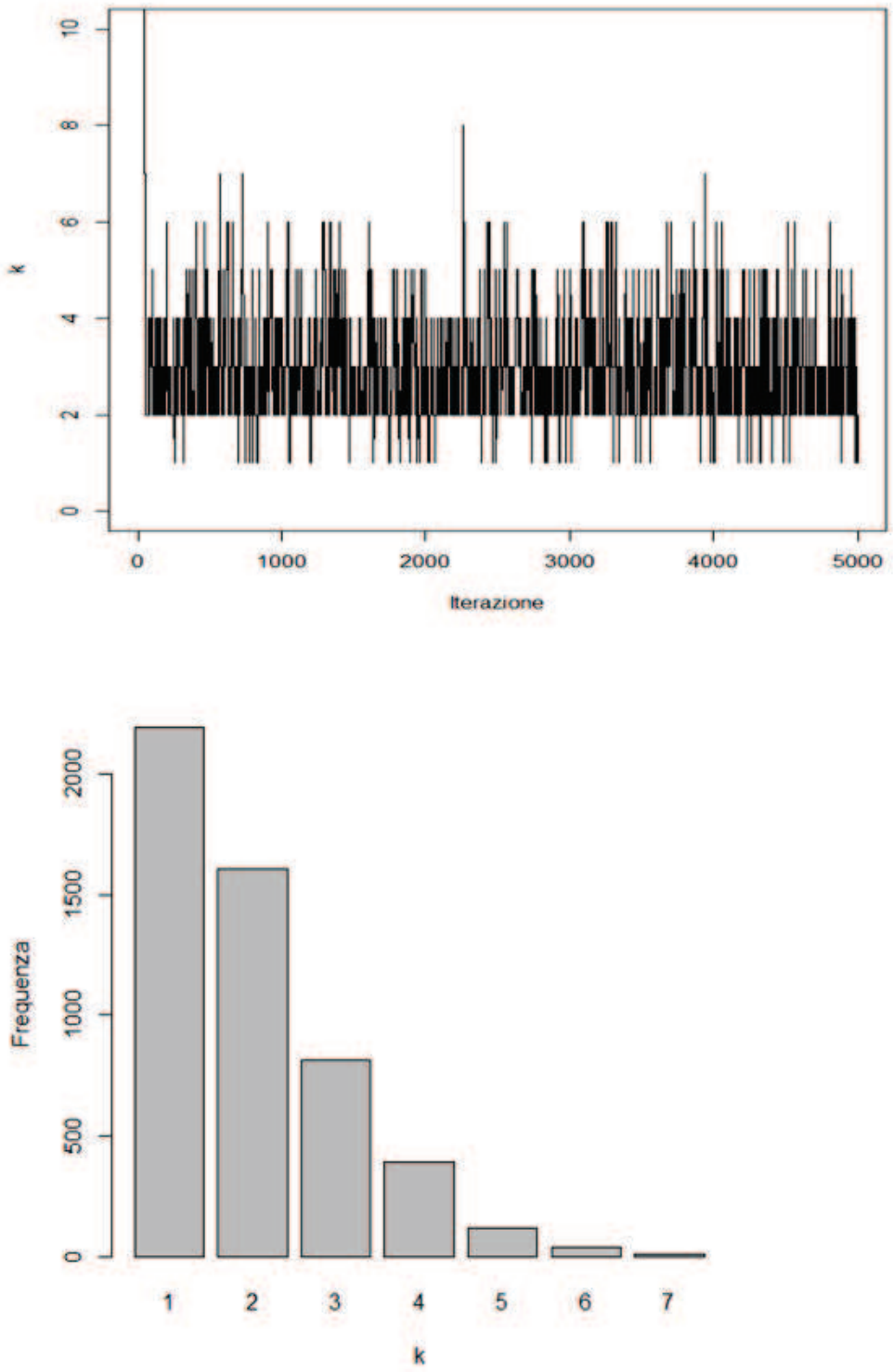

Figura 4.4: Sequenza di 5000 simulazioni con il K-NN Metropolis-Hastings. Distribuzione marginale a posteriori marginale di k.

4.3 Selezione delle variabili

Un ulteriore obiettivo dell'analisi è la selezione delle componenti più rilevanti nel vettore X costituito dalle 136 coordinate di punti, sulla base del loro contributo alla classificazione. Combinando criteri di parsimonia (l'eliminazione dei predittori superflui riduce i tempi di calcolo) ed efficienza (un numero eccessivo di variabili può rendere meno evidenti le differenze tra le classi), è desiderabile ottenere un sottoinsieme di variabili rilevanti che permettano di effettuare il riconoscimento delle emozioni con un margine di errore limitato.

In quest'ottica la coppia di parametri (β, k), di cui si vuole ottenere una stima della distribuzione a posteriori, viene condizionata a un'ulteriore classe di variabili indicatrici $\gamma_j \in \{0, 1\}$ con $(1 \leq j \leq p)$ che determina quali componenti di x sono attive nel modello. Si induce quindi l'algoritmo MCMC a eseguire simulazioni su sottomodelli di dimensione inferiore a p, riproducendo strutture di selezione gerarchica analoghe a quelle dei modelli di regressione.

L' esplorazione di tutte le $2^p - 1$ combinazioni di sottomodelli non è una strada percorribile, di conseguenza è necessario ricorrere a un metodo computazionalmente più efficiente. Una soluzione è utilizzare un approccio MCMC di tipo *reversible jump* (Robert e Casella, 2004, Capitolo 11), dove i parametri (β, k) variano condizionatamente a γ, mentre l'insieme $\gamma = (\gamma_1, \ldots, \gamma_p)$ aggiorna una componente alla volta condizionandosi a (β, k) e ai dati osservati.

4.3.1 Algoritmo MCMC per la selezione delle variabili

L'algoritmo è costituito da una prima parte simile al metodo illustrato nel Paragrafo 4.2.2, dal momento che una volta generati i parametri iniziali del modello, la costruzione della catena di Markov per la coppia (β, k) avviene muovendosi di iterazione in iterazione secondo la classica procedura di aggiornamento del MH. Per ogni configurazione $(\beta_{(t)}, k_{(t)}))$, si genera la stima Monte Carlo di $\gamma_j^{(t)}$ per $j = 1, \ldots, p$, tramite la distribuzione a posteriori

$$\pi\left(\gamma_j | \mathbf{y}, \mathbf{X}, \gamma_{-j}^{(t)}, \beta^{(t)}, k^{(t)}\right). \tag{4.7}$$

La difficoltà maggiore riguarda il calcolo della (4.7). Una possibilità, ispirandosi all'approccio di Dellaportas *et al.* (2002), è calcolare la probabilità di cambiamento di stato per ciascuna variabile indicatrice, a partire da

$$\pi\left(\gamma_j = \overline{\gamma_j^{t-1}} | \mathbf{y}, \mathbf{X}, \gamma_{-j}^{(t)}, \beta^{(t)}, k^{(t)}\right), \tag{4.8}$$

ovvero calcolando la probabilità per ciascuna γ_j di assumere un valore opposto rispetto all'iterazione precedente. In questo modo si calcola la probabilità di "salto" di γ_j a partire da $\hat{f}\left(\mathbf{y}|\mathbf{X}, \beta^t, k^t, \gamma_{-j}^{(t)}, \overline{\gamma_j^{(t-1)}}\right)$ e da una distribuzione a priori per il cambiamento di stato di tipo uniforme in $[0, 1]$. Nel caso in questione si ipotizza che le distribuzioni marginali delle γ_j siano indipendenti per qualsiasi valore di j. A questo punto, si confronta la nuova probabilità a posteriori con quella ottenuta al passo $t-1$. Il passaggio allo stato $(\beta^{(t)}, k^{(t)}, \gamma_j^{(t)})$ avviene con probabilità α pari al rapporto tra le due probabilità a posteriori. I passi dell'algoritmo implementato per realizzare la procedura di selezione sono illustrati in Figura 4.5.

METROPOLIS-HASTINGS SAMPLER PER SELEZIONE DELLE VARIABILI

Al tempo 0, generare $\gamma_j^{(0)} \sim \mathcal{B}\text{in}\left(\frac{1}{2}\right)$, $\log\beta^{(0)} \sim \mathcal{N}(0, \tau^2)$ e

$k^{(0)} \sim \mathcal{U}_{\{1,\dots,K\}}$.

Al tempo $1 \leq t \leq T$,

(1) Generare $\log\tilde{\beta} \sim \mathcal{N}\left(\log\beta^{(t-1)}, \tau^2\right)$ e $\tilde{k} \sim \mathcal{U}_{\{k-r,\dots,k+r\}}$

(2) Calcolare la probabilità di accettazione Metropolis-Hastings $\rho(\tilde{\beta}, \tilde{k}, \beta^{(t-1)}, k^{(t-1)})$

(3) Muoversi a $(\beta^{(t)}, k^{(t)})$ tramite passo di aggiornamento MH

(4) Per $j = 1, \dots, p$, calcolare $\pi\left(\gamma_j = \overline{\gamma_j^{t-1}} | \mathbf{y}, \mathbf{X}, \gamma_{-j}^{(t)}, \beta^{(t)}, k^{(t)}\right)$ e

$$\alpha = \frac{\pi\left(\gamma_j = \overline{\gamma_j^{t-1}} | \mathbf{y}, \mathbf{X}, \gamma_{-j}^{(t)}, \beta^{(t)}, k^{(t)}\right)}{\pi\left(\gamma_j = \gamma_j^{t-1} | \mathbf{y}, \mathbf{X}, \gamma_{-j}^{(t)}, \beta^{(t)}, k^{(t)}\right)}$$

(5) Prendere

$$\left(\beta^{(t)}, k^{(t)}, \gamma_j = \overline{\gamma_j^{t-1}}\right) = \begin{cases} \left(\beta^{(t)}, k^{(t)}, \gamma_j = \overline{\gamma_j^{t-1}}\right) & \text{con probabilità } \alpha \vee 1, \\ \left(\beta^{(t)}, k^{(t)}, \gamma_j = \gamma_j^{t-1}\right) & \text{altrimenti} \end{cases}$$

Figura 4.5: Algoritmo *Metropolis-Hastings* di tipo *reversible jump* per la selezione delle variabili.

Effettuando un numero elevato di simulazioni, l'algoritmo riesce ad esplorare tutto il supporto di (β, k, γ), selezionando con frequenza maggiore la configurazione di variabili ritenuta ottimale per i dati a disposizione.

4.3.2 Selezione delle coordinate rilevanti

L'adattamento del campione di espressioni facciali all'algoritmo MCMC per la selezione delle variabili ha coinvolto un'insieme di sottomodelli che sono stati "accettati" dal MH un numero di volte proporzionale alla loro credibilità bayesiana alla luce dei dati osservati. L'identificazione del sottoinsieme di coordinate rilevanti per il modello si traduce quindi in un'analisi di frequenza sulle configurazioni assunte dalle variabili di stato γ_j ad ogni passo dell'algoritmo. I parametri di regolazione β_{max}, τ, r e K sono rimasti invariati rispetto alla fase di classificazione, mentre il numero di simulazioni è stato fissato a 600. Quest'ultimo valore dovrebbe assumere un valore più elevato, in modo da dare tempo all'algoritmo di esplorare tutto il supporto di β, k, γ, ma per limiti di tempo è stato calcolato un numero contenuto di simulazioni al fine di ottenere delle semplici indicazioni generali. I risultati ottenuti sono riportati nella Figura 4.6.

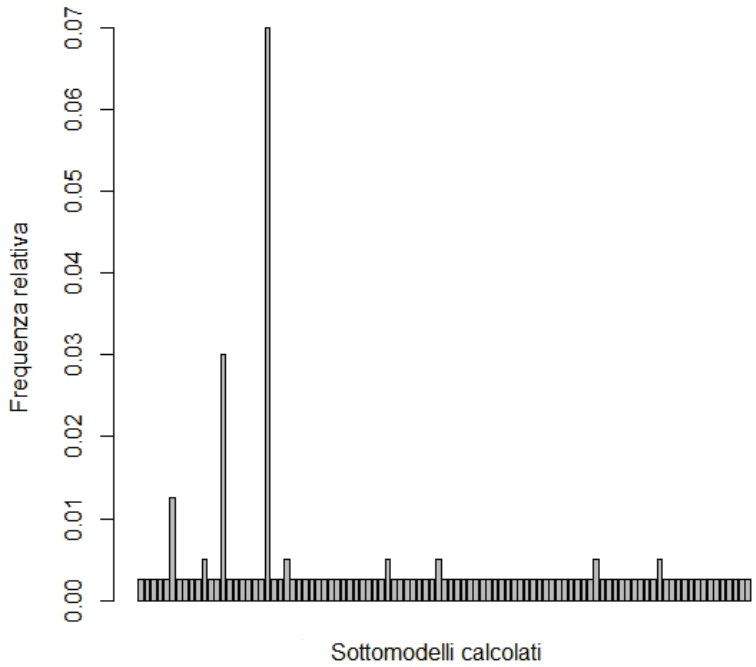

Figura 4.6: Analisi di frequenza per le configurazioni delle variabili per il modello probabilistico K-NN. Alcuni sottomodelli sono stati scelti dall'algoritmo MCMC con frequenza superiore rispetto ad altri.

Per ciascuna simulazione, è stato generato un valore dalla distribuzione a posteriori di β, k, γ. Analizzando le frequenze delle varie combinazioni dei γ_j, è stato ottenuta una configurazione che ottiene una frequenza più alta rispetto a tutte le altre, pari circa al 7% del totale. Il corrispondente sottomodello include 30 coordinate geometriche, di cui 3 ascisse e 27 ordinate. La rappresentazione facciale delle coordinate selezionate è presentata nella Figura 4.7.

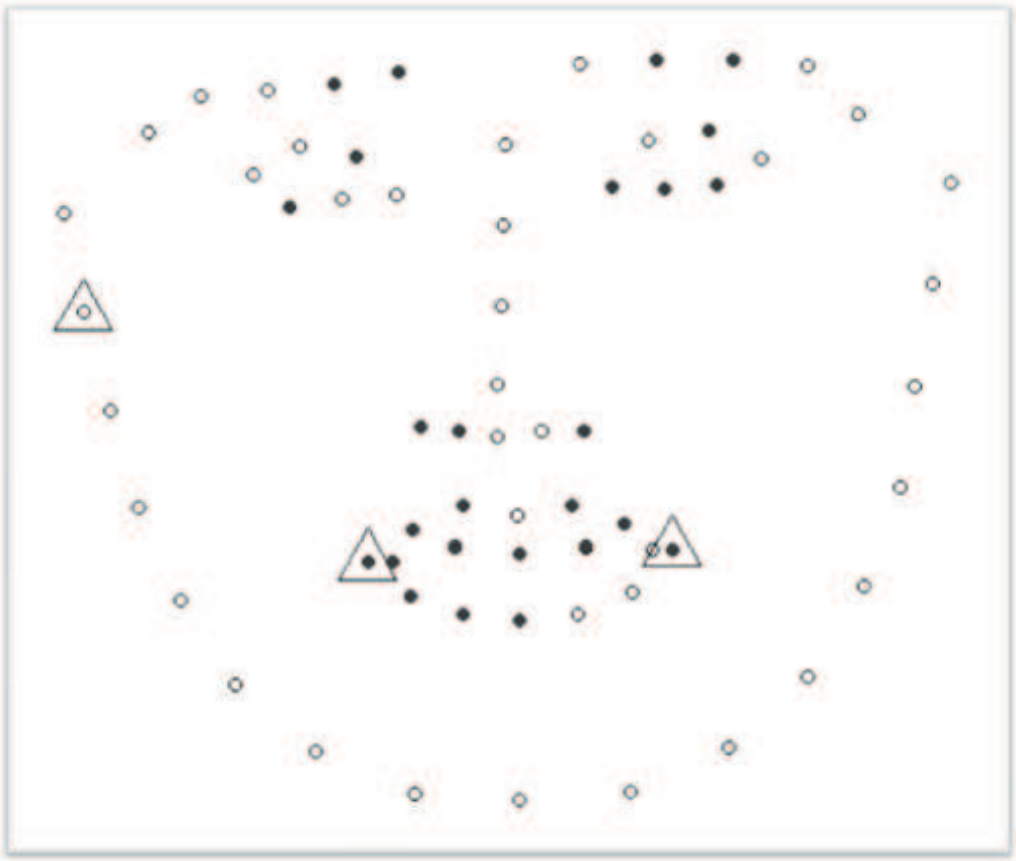

Figura 4.7: Rappresentazione su espressione neutra delle coordinate facciali attivate nella configurazione selezionata. Il pallino pieno identifica le coordinate verticali, mentre il triangolo quelle orizzontali. Si noti che sono attive entrambe le coordinate dei punti agli estremi della bocca, associati al muscolo zigomatico maggiore.

Una valutazione dell'importanza marginale di ciascuna variabile si può ottenere calcolando la media per ciascuna variabile γ_j, che corrisponde alla frazione con cui la variabile viene inclusa nei sottomodelli, rispetto al numero

totale di simulazioni. Ad esclusione della coordinata Y_{23}, tutte le altre con tasso di inclusione superiore all'80% appartengono alla configurazione precedente con frequenza più alta. Di seguito nella Tabella 4.2 viene riportato l'elenco completo per frazioni di inclusione superiori a 0.60.

	Variabile	*Tasso di inclusione*
1	Y.Landmark55	99.75%
2	Y. Landmark58	98.75%
3	Y. Landmark50	90%
4	Y. Landmark22	86.5%
5	Y. Landmark49	85.5%
6	Y. Landmark68	84.5%
7	Y. Landmark51	81.75%
8	Y. Landmark23	80.75%
9	Y. Landmark36	79.25%
9	X. Landmark49	79.25%
11	Y. Landmark54	70.5%
12	Y. Landmark24	70%
13	Y. Landmark45	69.75%
14	Y. Landmark66	63.25%
15	Y. Landmark67	60.75%

Tabella 4.2: Importanza marginale delle variabili per il modello bayesiano. Il tasso di inclusione è stato calcolato eliminando le 10 simulazioni per la fase di *burn in*.

Le indicazioni sulle variabili fornite dal modello bayesiano coincidono con quelle dei modelli LASSO, LDA, *boosting* e foreste casuali. Per ulteriori considerazioni a riguardo e la presentazione dei risultati complessivi, alla luce degli studi precedenti, si faccia riferimento al capitolo conclusivo.

Capitolo 5

Conclusioni

5.1 Premessa

In questo elaborato sono stati presentati molteplici approcci per il riconoscimento delle espressioni facciali contenute nell'*Extended Cohn-Kanade Dataset (CK+)*, di proprietà della *Carnegie Mellon University* (CMU) di Pittsburgh. Il database CK+ associa ad ogni immagine le coordinate geometriche (x, y) normalizzate relative a 68 punti facciali (*landmarks*), estratti tramite un algoritmo di computer vision chiamato *Active Appearance Model*. In aggiunta, seguendo la metrica FACS, vengono indicate nel database le unità d'azione (AU) attivate e la relativa emozione universale rappresentata, laddove sia univocamente definita. Raccogliendo i *landmarks* dei fotogrammi a cui è stata associata un'emozione ad alta intensità espressiva, si è ottenuto un campione di 1881 pose per effettuare il riconoscimento delle emozioni, tramite l'applicazione di tecniche statistiche per la classificazione in presenza di variabile risposta categoriale. Per realizzare questo studio, sono state utilizzate immagini raffiguranti espressioni neutre o una delle 7 emozioni universali identificate da Ekman, ovvero rabbia, disprezzo, disgusto, paura, felicità, tristezza e sorpresa, cercando di identificare la classe emozionale in funzione delle 68 coppie di coordinate del viso estratte da ciascuna immagine. In questa fase di lavoro è stato affrontato il problema della selezione dei punti più rilevanti per il corretto riconoscimento delle emozioni, e alla luce del ridotto numero di soggetti coinvolti per la realizzazione delle sequenze, si è valutata la possibilità di utilizzare campioni di stima e verifica indipendenti.

Lo studio di relazioni dirette tra punti facciali ed emozioni è un elemento di novità rispetto agli studi precedenti, dato che in passato le analisi sono state strutturate con lo scopo di rilevare l'attivazione, misurata in AU, di determinati muscoli facciali e solo in un secondo momento, servendosi di regole

interpretative, sono stati associati gli stati emotivi. In secondo luogo, i più attuali sistemi di analisi facciale, descritti nel Capitolo 2, classificano le caratteristiche del volto basandosi quasi esclusivamente su un modello singolo, sottovalutando i vantaggi derivanti dall'utilizzo di molteplici tecniche statistiche: calibrando o combinando quelle che si adattano meglio al problema, si può migliorare la capacità di riconoscimento complessiva.

Congiuntamente ai classificatori classici, è stata realizzata una versione probabilistica dei *K Nearest-Neighbors* (K-NN), utilizzando un approccio bayesiano. Per ottenere un modello probabilistico completo è stata resa simmetrica la relazione fra le osservazioni y_i e i k punti più vicini. Tale formulazione si basa essenzialmente su metodi Monte Carlo che forniscono stime a posteriori dei parametri del modello, simulando sequenzialmente dalle distribuzioni condizionate. La modellazione è stata realizzata in una duplice ottica: (1) effettuare la classificazione delle coordinate facciali, tramite un modello bayesiano stimato attraverso un algoritmo *Markov Chain Monte Carlo* (MCMC) che genera catene di Markov dalla distribuzione a posteriori dei parametri del modello e permette di assegnare le osservazioni alla classi, calcolando una stima della probabilità a posteriori di appartenenza a ciascuna classe; (2) identificare la configurazione di variabili con probabilità a posteriori più alta, in modo da selezionare quali punti del volto umano assumono maggiore importanza per la classificazione delle espressioni. In questa seconda specificazione è stato implementato un algoritmo MCMC di tipo *reversible jump* che permette di calcolare ad ogni passo la probabilità, per ogni variabile, di entrare o uscire dal modello. Tutta la parte relativa all'approccio bayesiano costituisce una formulazione nuova e originale per il riconoscimento delle espressioni.

5.2 Risultati ottenuti

Per quanto concerne la classificazione con tecniche di data mining, i risultati migliori sono stati raggiunti da regressione logistica con regolazione LASSO, analisi discriminante lineare, reti neurali, alberi di classificazione combinati tramite *boosting* e foreste casuali. Un'elevata capacità previsiva è stata ottenuta anche dal modello probabilistico bayesiano. Questi modelli hanno ottenuto un tasso di riconoscimento superiore sia a quello di studi orientati all'identificazione delle AU, che a quello condotto da Lucey *et al.* (2010), in cui la classificazione intermedia in AU è stata successivamente convertita in emozioni. Una sintesi dei risultati è presentata nella Tabella 5.1.

Sistemi	Metodo	Precisione	Database
CMU	AAM + SVM	83.33%	CK+ (2010)
CMU1	Rete neurale	95.5%	Cohn-Kanade (2000)
UCSD1	SVM + MLR	91.5%	Cohn-Kanade (2000)
UIUC1	BN + HMM	73.22%	Cohn-Kanade (2000)
UIUC2	NN + GMM	71%	Cohn-Kanade (2000)
Attuale	LDA	95.8%	CK+ (2010)
Attuale	Rete neurale	98.4%	CK+ (2010)
Attuale	LASSO	95.2%	CK+ (2010)
Attuale	Boosting	93.7%	CK+ (2010)
Attuale	Random Forest	96.7%	CK+ (2010)
Attuale	Probabilistic K-NN	93%	CK+ (2010)

Tabella 5.1: Modelli per la classificazione delle espressioni facciali. Per le sigle dei gruppi di lavoro, si veda il Paragrafo 2.1. Per le sigle dei metodi utilizzati, si fa riferimento alla didascalia della Tabella 2.2.

Una migliore capacità di classificazione è probabilmente dovuta alla scelta di effettuare il riconoscimento diretto delle emozioni, oltre che alla varietà di approcci considerati. A sostegno di questa scelta operativa, è necessario evidenziare come tutti gli altri studi non abbiano considerato le espressioni neutre, che rappresentano la principale fonte di errata classificazione per le procedure utilizzate.

Escludendo le reti neurali, che per instabilità dei risultati e scarsa interpretabilità non costituivano un valido strumento, con i modelli più precisi si è cercato di individuare le variabili più rilevanti. In questa fase è stata utlizzata una procedura *stepwise* per le variabili del modello LDA, mentre per la regressione logistica LASSO il processo di selezione è implicito. Nel caso dei classificatori combinati, è stata misurata l'importanza intrinseca delle covariate all'interno del modello. I risultati ottenuti, sebbene ogni approccio presenti delle peculiarità, consentono di trarre alcune conclusioni complessive: (1) la **prevalenza assoluta**, nel sottoinsieme delle variabili rilevanti, di **coordinate verticali**; (2) una buona porzione dei punti tracciati sul volto è considerata importante per l'economia dei modelli, ma una concentrazione particolare è riscontrabile lungo l'**asse centrale del volto**, in corrispondenza di bocca, **naso**, **occhi** e **arcata sopraccigliare**; al contrario, i punti di contorno della sagoma facciale raramente vengono inclusi dalle procedure di selezione.

Per quanto concerne la versione probabilistica dei K-NN, il modello ha fornito una stima dell'importanza marginale di ciascuna variabile e al tempo stesso un sottomodello con più alta probabilità a posteriori, dal momento che la configurazione di variabili che lo compone è stata selezionata un numero di volte superiore a tutti gli altri. Le indicazioni fornite da questo metodo confermano sostanzialmente quelle fornite dagli altri modelli, sebbene il numero ridotto di simulazioni calcolate non potesse garantire dei risultati consistenti. Sebbene la natura delle procedure di selezione non sia equiparabile, nella Figura 5.1 sono presentati i punti del volto che complessivamente sono stati indicati come i più rilevanti dalle metodologie studiate. I punti contrassegnati sono quelli selezionati più volte dai vari modelli.

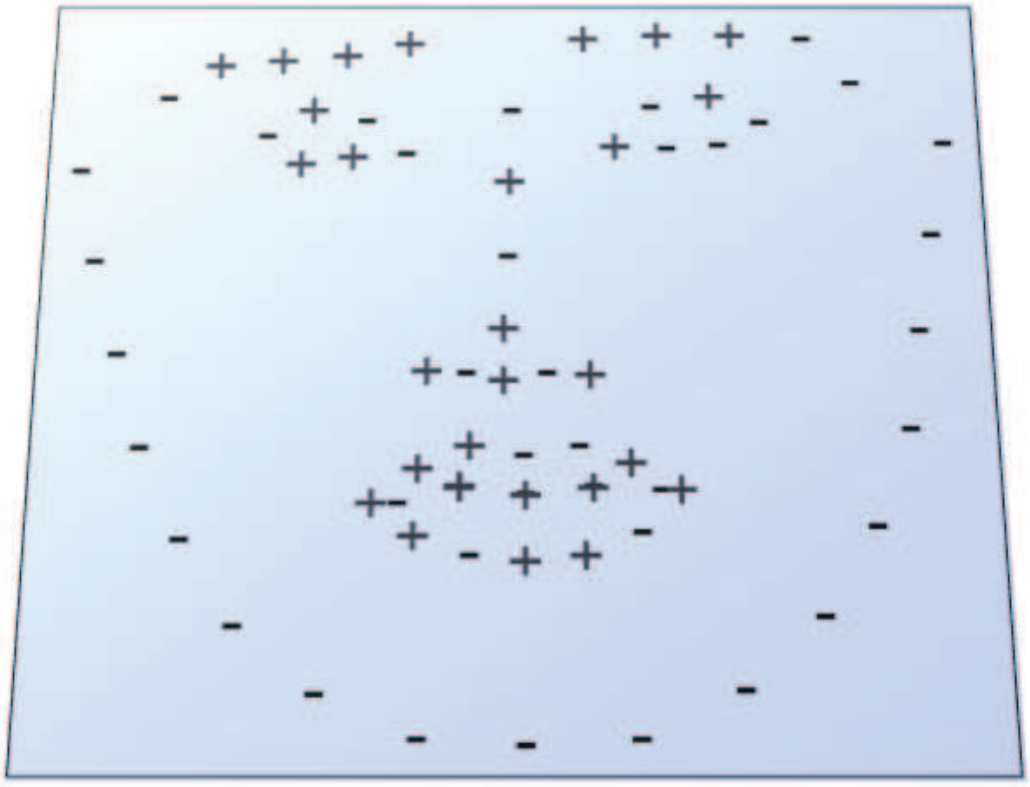

Figura 5.1: Selezione dei punti facciali più rilevanti per il riconoscimento delle espressioni. I punti segnati con un "+" sono i più importanti per gli approcci considerati.

5.3 Considerazioni finali

L'esito complessivo analisi è stato positivo, raggiungendo risultati in linea o superiori a quelli ottenuti dagli studi precedenti. La selezione di un campione costituito da pose frontali con alta intensità espressiva ha reso più semplice conseguire percentuali di riconoscimento elevate. All'interno del lavoro svolto, costituisce un elemento di innovazione lo studio della relazione tra punti ed emozioni primarie da un punto di vista statistico, che rappresenta una piccola porzione delle espressioni riproducibili dal volto umano, ma che viene giustificato dal principio di universalità. La possibilità di applicare le metodologie utilizzate a un campione proveniente da qualsiasi popolazione e cultura, assegna ai risultati ottenuti un valore sociologico che formulazioni più generali non avrebbero garantito.

Un ulteriore punto di novità è l'introduzione di un modello probabilistico completo di natura bayesiana, che non era stato formalizzato e implementato in precedenza. L'interpretazione in termini di probabilità a posteriori e le capacità di identificazione dimostrate dalle procedure MCMC, rappresentano un punto a favore per l'approccio metodologico utilizzato. Il calcolo di un numero superiore di simulazioni avrebbe garantito una maggiore consistenza dei risultati, ciò nonostante le indicazioni preliminari fornite sono sostanzialmente coerenti con quelle dei modelli di classificazione realizzati in precedenza.

Sebbene possa sembrare insoddisfacente da un punto di vista logico, la scelta di studiare in modo indipendente le coordinate geometriche dei punti (x, y) è giustificata dalla volontà di concedere massima flessibilità agli strumenti di classificazione, in funzione di spostamenti sia verticali che orizzontali. In secondo luogo, l'individuazione di una direzione di movimento principale per ciascun punto è un'informazione di interesse per la pianificazione di eventuali studi futuri. Al contempo, la necessità di identificare le aree del volto più informative da un punto di vista espressivo, ha portato alla selezione del punto completo qualora una sola delle due coordinate risultasse significativa.

Le possibilità di sviluppo per attività di ricerca come questa sono molteplici: l'estensione dell'analisi a immagini a bassa intensità espressiva, l'introduzione di espressioni miste, la rilevazione di stati emotivi non spontanei e non ultimo, l'utilizzo di fotogrammi in cui alcune componenti facciali non siano perfettamente rilevabili. Si pensi al caso di fotografie danneggiate, pose in prospettiva non frontale o parti del volto parzialmente coperte. In questi contesti, le difficoltà per le procedure di classificazione aumenteranno e necessariamente dovranno essere riadattate in funzione della nuova situazione studiata, tuttavia gli strumenti presentati rappresentano un valido punto di

partenza per organizzare analisi future. Considerata l'efficacia delle tecniche statistiche applicate al riconoscimento espressivo, l'integrazione di queste metodologie in sistemi complessi e automatizzati può fornire un valido supporto per l'individuazione, la conoscenza e la rappresentazione di molteplici aspetti della comunicazione non verbale che le singole capacità umane non sono in grado di gestire in modo efficace.

Appendice

Codice Java per l'estrazione delle immagini

```
import java.io.*;
import java.util.*;

public class Landmarks
{
   public static int somma(int[] v)
   {
         int tot=0;
         for(int i=0;i<v.length;i++)
         tot+=v[i];
         return tot;
   }

   public static void main(String[] argv) throws IOException
   {
        FileWriter fw1 = new FileWriter ('Landmarks.xls');
        FileWriter fw2 = new FileWriter ('Selection.xls');
        File root = new File("C:/Users/.../Landmarks");
        ArrayList<File> lista = new ArrayList<File>();
        lista.addAll(Arrays.asList(root.listFiles()));
        int[] subdir = new int [593];
        int i = 0;
        int count = 0;
        int index = 0;
        int tmp = 0;

        while(lista.size() > 0)
        {
           File f = lista.remove(0);
           if(f.isDirectory())
           {
             lista.addAll(Arrays.asList(f.listFiles()));
             if(i >= 123)
             {
             subdir[i-123]=(Arrays.asList(f.listFiles()).size());
```

```
                }
                i++;
            }
            else
            {
              FileReader fr = new FileReader(f);
              BufferedReader br = new BufferedReader(fr);
              String s;
              int j = 0;
              while((s = br.readLine()) != null)
              {
                if (count == tmp + subdir[index])
                {
                  fw2.write(s);
                  j++;
                  if(j%68 == 0)
                  {
                    fw2.write("\n");
                    tmp += subdir[index];
                    index++;
                  }
                }
                fw1.write(s);
              }
              fw1.write("\n");
              count++;
              fr.close();
            }
          }
          fw1.close();
          fw2.close();
    }
}
```

Analisi dei dati

Caricamento dei dati e Analisi Esplorativa

```
database <- read.table(file.choose(), header=T)
head(database)
missing <- which(is.na(database$Emotion)==TRUE)
data <- database[-c(missing),-c(1,2)]
dim(data)
emo <- c("Neu","Ang","Con","Dis","Fea","Hap","Sad","Sur")

# Divisione del data-set in insieme di stima e di verifica

n <- nrow(data)
```

```
set.seed(100)
ind <- sample(1:n)
trainval <- ceiling(n*.8)
testval <- ceiling(n*.1)
train <- data[ind[1:trainval],]
test <- data[ind[(trainval+1):(trainval+testval)],]
valid <- data[ind[(trainval+testval+1):n],]

f1 <- as.formula(paste("Emotion~", paste(names(train)[-c(1)],
                        collapse="+"),  collapse=NULL))

# Analisi preliminare dei dati

plot(sd(data1[,x]), type="o", col=1, ylim=c(23,32),pch=19,
        ylab="Standard Deviation", xlab="Landmarks")
abline(h=mean(sd(data1[,x])), lty=6)
abline(h=mean(sd(data1[,y])), lty=2)
points(sd(data1[,y]), type="o", pch="*", col=1)
identify(sd(data1[,x]), type="o", col=1, ylim=c(23,32),pch=19)
identify(sd(data1[,y]), type="o", pch="*", col=1)

summary(data)
plot(train[,1], xlab="Emozione", ylab="Frequenza Assoluta",
        freq=FALSE, col=8, font=2, col.axis="black")
par(mfrow=c(3,3))
for(i in 2:9)
{
        if(is.factor(data[,i]))
          boxplot(data[,1]~data[,i])
          else
          plot(data[,i]~data[,1], xlab=variable.names(data)[i],
          ylab=variable.names(data)[1])
}
```

Analisi delle componenti principali

```
pc.train<-princomp(train[,-1], cor=T)
screeplot(pc.train, main="Componenti principali", type="lines")
summary(pc.train)
pc.train$sd^{2}
sum(pc.train$sd^{2})   # uguale alla varianza totale
pc.train$loadings[1:10,1:3]     # coefficenti comp
pc.train$scores[1:10,1:3]
biplot(pc.train, choices = 1:2, scale = 1, pc.biplot=TRUE)
library(plotpc)
plotld(pc.train$loadings,lty=1, lwd=1:3)
fmod1 <-  as.formula(paste("~", paste(names(train)[-c(1)],
                        collapse="+"),  collapse=NULL))
x.stima <- model.matrix(fmod1, train)
```

```
y.stima <- train[,1]
x.verifica <- model.matrix(fmod1, test)
y.verifica <- test[,1]
x.pc <- model.pcr$scores[,1:3]
x.stimapc <- data.frame(y.stima,x.pc)
f.pc <- as.formula(paste("y.stima~", paste(names(x.stimapc)[-1],
collapse="+"), collapse=NULL))
pc.load <- model.pcr$loadings[,1:3]
pc.test <- x.verifica[,-1] %*% pc.load
head(pc.test)
x.testpc <- data.frame(y.verifica,pc.test)
```

Modello multilogit

```
library(VGAM)
glm.model <- vglm(f1, multinomial(refLevel = 1), train)
coef(glm.model, matrix=TRUE)
summary(glm.model)
weights(glm.model, type="prior", matrix=FALSE)
glm.model@y
str(glm.model)
glm.model@qr$rank

p.glm <- predict(glm.model, newdata=test, type = "response")
p1.glm <- rep(NA, nrow(p.glm))
for (i in 1:nrow(p.glm))
{
        p1.glm[i] <- which.max(p.glm[i,])
}

source("lift-roc1.R") # funzione personalizzata
tabella.sommario(p1.glm,test$Emotion)
```

Modello loglineare

```
library(nnet)
multi.model <- multinom(f1, data=train, MaxNWts = 1200, maxit=300)
names(multi.model)

p.log <-predict(multi.model,newdata=test,type="class")
tabella.sommario(p.log,test$Emotion)
```

Analisi discriminante lineare

```
library(MASS)
system.time(lda(f1, data=train))
names(d1)
plot(d1, dimen=1, type="both")
```

```
p.lda <-predict(d1,newdata=test)
tabella.sommario(p.lda$class,test$Emotion)
```

Albero di classificazione

```
library(tree)

# Crescita dell'albero:

t1 <- tree(f1, data=train,
           control=tree.control(nobs=nrow(train),minsize=2,
        mindev=0.0001))
plot(t1)
text(t1,cex=0.6)

# Potatura dell'albero:

t2 <- prune.tree(t1,newdata=valid)
plot(t2)
J<-t2$size[t2$dev==min(t2$dev)]
t3<-prune.tree(t1,best=J)
plot(t3)
text(t3,cex=0.2)
summary(t3)

# Previsioni

p.tree <-predict(t3,newdata=test, type="class")
tabella.sommario(p.tree,test$Emotion)
```

Classificatori combinati

```
# Bagging

library(ipred)
fit.bag <- bagging(f1,data=train,nbagg=50, coob=TRUE)
fit.bag

p.bag <- predict(fit.bag, newdata=test)
tabella.sommario(p.bag,test$Emotion)

# Random Forest

library(randomForest)
fit.rf <- randomForest(f1, data=train,importance=TRUE,
                                  proximity=TRUE)
round(importance(fit.rf), 2) # importanza delle variabili

p.rf <- predict(fit.rf, newdata=test, type="response")
```

```
tabella.sommario(p.rf,test$Emotion)

# Boosting

library(adabag)
stump <- rpart.control(cp=0.01, maxdepth=4, minsplit=0)
fit.ada <- boosting(f1, train, boos = TRUE, mfinal = 100,
                                coeflearn = 'Zhu')
names(fit.ada)

p.ada <- predict.boosting(fit.ada, newdata=test)
tabella.sommario(p.ada1$class,test1$Emotion)
```

Rete neurale

```
library(nnet)

decay<- 10^(seq(-1, -0.7, length=10))
err <- rep(0,10)
for(k in 1:10) {  n1<- nnet(f1, data=train, decay=decay[k], size=8,
                  maxit=1200, MaxNWts = 1200,  trace=FALSE)
                  p1n<- predict(n1, newdata=valid, type="class")
                  a<- tabella.sommario(p1n,valid$Emotion)
                  err[k] <- 1-sum(diag(a))/sum(a)
                  print(c(err[k], decay[k]))
                       }
cbind(err,decay)

n2<- nnet(f1, data=train,decay=0.107,size=8,MaxNWts = 1200,
                maxit=2000)
p.nnet <- predict(n2, newdata=test, type="class")
plot(p.nnet,test$Emotion)
summary(n2)
tabella.sommario(p.nnet,test$Emotion)
```

GAM

```
library(VGAM)
fg1 <- as.formula(paste("Emotion ~s(", paste(names(train[-1]),
        collapse=")+s("),")"))
gam.model <- vgam(fg1, family=multinomial(refLevel = 1), data=train,
                                trace=TRUE, maxit=30)

p1.gam <- predict(gam.model,newdata=test,type= "response")

pgam <- NULL
for (i in 1:nrow(p1.gam))
{
        pgam[i] <-which.max(p1.gam[i,])
```

```
}

tabella.sommario(pgam,test$Emotion)
```

Regressione logistica LASSO

```
train2 <- model.matrix(f1, data=train)

library(glmnet)
fit.glmnet <- glmnet(train2[,-1],train$Emotion, family="multinomial",
                          alpha=1, maxit=200000)

summary(fit.glmnet)
par(mfrow=c(3,3))
plot(fit.glmnet)
names(fit.glmnet)

# cross validation per il lambda
cv.glmnet <- NULL
p.gnet <- NULL
nlambda <- length(fit.glmnet$lambda)
for(i in 1:nlambda) {
p.cv <- predict(fit.glmnet,test2[,-1],s=fit.glmnet$lambda[i],
                     type="response")
for(j in 1:nrow(p.cv))
{
        p.gnet[j] <- which.max(p.cv[j,,])    }
        tab <- table(p.gnet,test$Emotion)
        cv.glmnet[i] <- 1-sum(diag(tab))/sum(tab)
}

# previsione per con il lambda migliore

p.glmnet <- predict(fit.glmnet, test2[,-1], s=fit.glmnet$lambda[77],
                          type="response")
for(j in 1:nrow(p.glmnet))
p.gnet[j] <- which.max(p.glmnet[j,,])
tabella.sommario(p.gnet,test$Emotion)
```

MARS

```
library(earth)
fit.mars <- earth(train2,train$Emotion)
summary(fit.mars)
plot(fit.mars)
p.mars <- predict(fit.mars,newdata=test2, type="response")
summary(p.mars)

pmars <- NULL
```

```
for(j in 1:nrow(p.mars)){
pmars[j] <- which.max(p.mars[j,])    }
tabella.sommario(pmars,test$Emotion)
```

SVM

```
library(e1071)
# cross validation per il parametro di tuning
cv.svm <- tune(svm,train2,train$Emotion, ranges=list(cost=2^(-2:4)),
               tunecontrol=tune.control(sampling="cross", cross=20))
summary(cv.svm)

# scelgo la costante migliore

fit.svm <- svm(train[,-1],train$Emotion,kernel="radial",cost=2)
summary(fit.svm)
p.svm <- predict(fit.svm, newdata=test[,-1])

tabella.sommario(p.svm,test$Emotion)
```

K Nearest-Neighbors

```
library(ipred)
fit.nn <- ipredknn(f1,data=train,k=2)

p.nn <- predict.ipredknn(fit.nn, newdata=test, type="class")
tabella.sommario(p.nn,test$Emotion)
```

Selezione delle variabili

```
# Analisi discriminante lineare

library(klaR)
sel.lda <- stepclass(f1, train, method="lda", improvement = 0.01)
sel.lda$model

# Regressione LASSO

lasscoef <- coef(fit.glmnet, s=fit.glmnet$lambda[77])
matlass <- cbind(as.matrix(lasscoef$Neu),as.matrix(lasscoef$Ang),
                 as.matrix(lasscoef$Con),as.matrix(lasscoefDis),
                 as.matrix(lasscoef$Fea),as.matrix(lasscoef$Hap),
                 as.matrix(lasscoef$Sad),as.matrix(lasscoef$Sur))
colnames(matlass) <- c("Neutral","Anger","Contempt","Disgust","Fear",
                       "Happiness","Sadness", "Surprise")
round(matlass, 1)

# Boosting e Random Forest
```

```
sort(round(fit.ada$importance, 3), decreasing=TRUE)
sort(round(fit.rf$importance, 3), decreasing=TRUE)
```

Confronto tra i modelli

```
# Indice AUC

library(pROC)
roc.glm <- multiclass.roc(p1.glm, as.numeric(test$Emotion),
                percent=TRUE)
ci.glm <- ci(p1.glm, as.numeric(test$Emotion)) # CI
roc.lda <- multiclass.roc(as.numeric(p.lda$class),
                as.numeric(test$Emotion), percent=TRUE)
p.net <- factor(p.nnet, levels=emotions)
roc.nnet <- multiclass.roc(as.numeric(p.net),
                    as.numeric(test$Emotion), percent=TRUE)
roc.lasso <- multiclass.roc(p.gnet, as.numeric(test$Emotion),
                     percent=TRUE)
p.boost <- factor(p.ada$class, levels=emotions)
roc.ada <- multiclass.roc(as.numeric(p.boost),
                  as.numeric(test$Emotion), percent=TRUE)
roc.rf <- multiclass.roc(as.numeric(p.rf),
              as.numeric(test$Emotion), percent=TRUE)
```

K Nearest Neighbors

MCMC per la classificazione

```
probknn <- function(data,bmax,K,tau,r,Nsim)
{
response <- data[,1]
class <- levels(data[,1])
nclass <- length(class)
n <- nrow(data)
theta=k=beta=log.post=error.rate <- rep(0,Nsim)
theta[1] <- rnorm(1,0,tau)
k[1] <- runif(1,1,K)
beta[1] <- (bmax*exp(theta[1]))/(exp(theta[1])+1)
log.post[1] <- -9999
cond.prob <- array(rep(0,n*nclass*(Nsim-1)),dim=c(n,nclass,Nsim-1),
                   dimnames=list(1:n,class,1:(Nsim-1)))
p.knn <- matrix(rep(0,n*nclass),ncol=nclass, dimnames=list(1:n,class))
# Nearest Neighbours Matrix
distance.matrix <- as.matrix(dist(data[,-1],up=T,diag=T))
nearest <- t(apply(distance.matrix,1,order))[,-1] #nearest neighbours

for (i in 2:Nsim)
{
```

```
theta.til <- rnorm(1,theta[i-1],tau)
k.til <- round(runif(1,max(1,k[i-1]-r),min(k[i-1]+r,K)))
beta.til <- (bmax*exp(theta.til))/(exp(theta.til)+1)
knn.matrix <- nearest[,1:k.til]  # k elements
prop <- matrix(0,n,nclass, dimnames=list(1:n,class)) # proportion
for (j in 1:n)
{
      if(is.null(ncol(knn.matrix)))
     {  prop[j,] <- response[knn.matrix[j]]    }
     else
     {  prop[j,] <- summary(response[knn.matrix[j,]])/k.til   }
}
for (v in 1:n)
{
     for(j in 1:nclass)
     {
          cond.prob[v,j,i-1] <- (exp(beta.til*prop[v,j])) /
                                  sum(exp(beta.til*prop[v,]))
     }
}
product <- matrix(0,n,1)
log.prior <- log(runif(1,0,beta.til)*runif(1,1,k.til))
for (t in 1:n)
{
      product[t] <- cond.prob[t,as.numeric(response)[t],i-1]
}
log.lik <- sum(log(product))
log.post[i] <- log.lik+log.prior
log.rho <- log.post[i]-log.post[i-1]+theta.til-theta[i-1]+
              log((1+exp(theta[i-1]))^2)-log((1+exp(theta.til))^2)+
                dunif(k.til,k[i-1]-r,k[i-1]+r)-dunif(k[i-1],k.til-r,k.til+r)
if(runif(1) < exp(log.rho))
{
      beta[i] <- beta.til
      theta[i] <- theta.til
      k[i] <- k.til
}
else
{
     beta[i] <- beta[i-1]
     theta[i] <- theta[i-1]
     k[i] <- k[i-1]
}
}
for(e in 1:n)
p.knn[e,] <- apply(cond.prob[e,,],1,mean)
list(pred=data.frame(p.knn),par=data.frame(beta,theta,k, log.post))
}
```

```
# Matrice di confusione

p.mcmc <- c(rep(0,nrow(data)))
for(i in 1:nrow(data))
p.mcmc[i] <- which.max(knn1.mcmc$pred[i,])
tabella <- table(p.mcmc,data$Emotion)
error <- 1-sum(diag(tabella))/sum(tabella)
```

MCMC per selezione delle variabili

```
var.sel <- function(data,bmax,K,tau,r,Nsim)
{
response <- data[,1]
class <- levels(data[,1])
nclass <- length(class)
covariates <- colnames(data[,-1])
n <- nrow(data)
p <- ncol(data[,-1])
k=beta=log.post <- rep(0,Nsim)
k[1] <- runif(1,1,K)
beta[1] <- exp(rnorm(1,0,tau))
log.post[1] <- -9999
data1 <- data[,-1]
gamma <- matrix(rep(0,Nsim*p),nrow=Nsim,ncol=p)
for(j in 1:p)
{
      gamma[1,j] <- rbinom(1,1,0.5)
}

for (i in 2:Nsim)
{
zero <- which(gamma[i-1,]== 0)
# Nearest Neighbours Matrix
distance.matrix  <- as.matrix(dist(data1[,-c(zero)],up=T,diag=T))
nearest <- t(apply(distance.matrix,1,order))[,-1] #nearest neighbours
beta.til <- exp(rnorm(1,log(beta[i-1]),tau))
k.til <- round(runif(1,max(1,k[i-1]-r),min(k[i-1]+r,K)))
knn.matrix <- nearest[,1:k.til]  # k elements
prop <- matrix(0,n,nclass, dimnames=list(1:n,class)) # proportion
for (j in 1:n)
{
      if(is.null(ncol(knn.matrix)))
      {  prop[j,] <- response[knn.matrix[j]]    }
      else
      {  prop[j,] <- summary(response[knn.matrix[j,]])/k.til   }
}
cond.prob <- matrix(NA,n,nclass, dimnames=list(1:n,class))
for (v in 1:n)
{
```

```
    for(j in 1:nclass)
    {
     cond.prob[v,j] <- (exp(beta.til*prop[v,j])) /
                            sum(exp(beta.til*prop[v,]))
     }
}
product <- matrix(0,n,1)
log.prior <- 0
for (t in 1:n)
{
    product[t] <- cond.prob[t,as.numeric(response)[t]]
}
log.lik <- sum(log(product))
log.post[i] <- log.lik+log.prior
log.rho <- log.post[i]-log.post[i-1]+
              pnorm(k.til,k[i-1])-pnorm(k[i-1],k.til)
if(runif(1) <= exp(log.rho))
{
    beta[i] <- beta.til
    k[i] <- k.til
}
else
{
    beta[i] <- beta[i-1]
    k[i] <- k[i-1]
}

# reversible jump
for(w in 1:p)
{
if(w %in% zero)
{
    zero.p <- c(zero[!zero == w])
}
else
{
     zero.p <- c(zero,w)
}
distance.p  <- as.matrix(dist(data1[,-c(zero.p)],up=T,diag=T))
nearest.p <- t(apply(distance.p,1,order))[,-1] #nearest neighbours
knnmat.p <- nearest.p[,1:k[i]]  # k elements
prop.p <- matrix(0,n,nclass, dimnames=list(1:n,class)) # proportion
for (j in 1:n)
{
     if(is.null(ncol(knnmat.p)))
    {   prop.p[j,] <- response[knnmat.p[j]]          }
    else
    {   prop.p[j,] <- summary(response[knnmat.p[j,]])/k[i]   }
}
```

```
cond.p <- matrix(NA,n,nclass, dimnames=list(1:n,class))
for (v in 1:n)
{
    for(j in 1:nclass)
   {
     cond.p[v,j] <- (exp(beta[i]*prop.p[v,j])) /
                          sum(exp(beta[i]*prop.p[v,]))
   }
}
product.p <- rep(0,n)
logprior.p <- log(runif(1,0,1))
for (t in 1:n)
{
    product.p[t] <- cond.p[t,as.numeric(response)[t]]
}
loglik.p <- sum(log(product.p))
logpost.p <- loglik.p+logprior.p
alpha <- logpost.p - log.post[i]
if(runif(1) <= exp(alpha))
{
    if (w %in% zero)
    {
        gamma[i,w] <- 1
    }
    else
    {
        gamma[i,w] <- 0
    }
    log.post[i] <- logpost.p
}
else
{
    gamma[i,w] <- gamma[i-1,w]
}
if(gamma[i,w] == 0)
{
    zero <- c(zero,w)
}
else
{
    zero <- c(zero[!zero == w])
}
}
}
data.frame(beta,k,log.post,gamma)
}
```

```
# Frequenza dei sottomodelli

inc <- matrix(NA, Nsim*Nsim,1)
riga=1
for(i in 1:Nsim)
{
    for(j in 1:Nsim)
    {
        inc[riga] <- (all(simulation[i,]==simulation[j,]))
        riga = riga +1
    }
    riga <- riga +1
}

match <- matrix(inc[!is.na(inc)],Nsim,Nsim)
diag(match) <- NA
freq <- summary(factor(trunc((which(match==TRUE)/Nsim)+1)))
barplot(freq)
```

Riferimenti bibliografici

Ahuja, N., Kriegman, D. e Yang, M. (2002). Detecting faces in images: a survey. *Pattern Analysis and Machine Intelligence. IEEE Transaction on*, **24**, 34-58.

Ahn, S. J., Bailenson, J. e Jabon, M. (2008). Facial Expressions as Predictors of Online Buying Intention. In *Proceedings of the 58th Annual International Communication Association Conference*.

Arya, A., Di Paola, S. e Parush, A. (2009). Perceptually Valid Facial Expressions for Character-Based Applications. *International Journal of Computer Games Technology*, **ed. 2009**, 13 pagine.

Azzalini, A. e Scarpa, B. (2004). *Analisi dei dati e data mining.* Springer-Verlag, Milano.

Bartlett, M., Donato, G., Ekman, P., Hager, J. e Sejnowski, T. (1999). Classifying facial actions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **21**, 974–989.

Bartlett, M.,Fasel, I.,Frank, M., Littlewort, G., Whitehill, J. e Wu, T. (2011). The Computer Expression Recognition Toolbox. *IEEE International Conference on Automatic Face, Gesture Recognition and Workshops*, 298 - 305.

Bartlett, M., Fasel, I., Littlewort, G., Movellan, J. e Susskind, J. (2006). Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, **24**, 615-625.

Bartlett, M., Lee, K. e Littlewort, G. (2009). Automatic coding of Facial Expressions displayed during Posed and Genuine Pain. *Image and Vision Computing*, **27**, 1797-1803.

Bartlett, M., Littlewort-Ford, G., Braathen, B., Fasel, I., Hershey, J., Marks, T., Movellan, J.R., Sejnowski, T. e Smith, E. (2001). Automatic analysis of spontaneous facial behavior: a final project report. *Technical Report INC-MPLab-TR*.

Bartlett, M. e Whitenill, J. (2010). *Automated Facial Expression Measurement: Recent Applications to Basic Research in Human Behavior, Learning and Education*. Oxford University Press, Oxford.

Beauchemin, S.S. e Barron, J.L. (1995). The computation of optical flow. *ACM Computing Surveys*, **27**, 433-466.

Black, M.J. e Yacoob, Y. (1997). Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion. *lnternational Journal of Computer Vision*, **25**, 23-48.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, **24**, 123–140.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A. e Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Cassel, T.D., Cohn, J.F., Messinger, D.S. (2008). The dynamics of infant smiling and perceived positive emotion. *Journal of Nonverbal Behavior*, **32**, 133-155.

Chen, L. (2000). *Joint processing of ausio-visual information for the recognition of emotional expressions in human-computer interaction*. In PhD thesis, University of Illinois.

Cohen, I.,Sebe, N., Cozman, F.G., Cirelo, M.C. e Huang, T. (2003). Learning Bayesian network classifiers for facial expression recognition using both labeled and unlabeled data. In *Proceedings of the IEEE Int. Conf. on Computer Vision and Pattern Recognition*, **1**, 595-601.

Cohn, J.F., Kanade, T., Moriyama, T., Ambadar, Z., Xiao, J., Gao, J. e Imamura, H. (2001). *A comparative study of alternative facs coding algorithms*. Robotics Institute, Carnegie Mellon University, Pittsburgh.

Cootes, T. F., Edwards, G. J. e Taylor, C. J. (1998). Active Appearance Models. *Lecture Notes in Computer Science*, **1407**, 484-498.

Cootes, T. F., Edwards, G. J. e Taylor, C. J. (2001). Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**, 681-685.

Cootes, T. F., Taylor, C. J., Cooper, H. e Graham, J. (1995). Active shape models-their training and application. *Computer Vision and Image Understanding*, **61**, 38-59.

Cortes, C. e Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**, 273-297.

Darwin, C. (1872). *The Expression of Emotions in Man and Animals*. D. Appleton, New York.

Dasarathy, B. V. (1991). *Nearest Neighbour (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, **2**, 1160–1169.

Dellaportas, P., Forster, J.J. e Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, **12**, 27-36.

Dinges, D. F., Rider, R. L., Dorrian, J., McGlinchey, E. L. e Rogers, N. L. (2005). Optical computer recognition of facial expressions associated with stress induced by performance demands. *Aviation, Space, and Environmental Medicine*, **76**, 172-182.

Ekman, P. e Frank, M. (1997). The ability to detect deceit generalizes across different types of high-stake lies. *Personality and Social Psychology*, **72**,1429–1439.

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

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

Ekman, P. e Friesen, W. V. (1982). *Rationale and reliability for EMFACS Coders*. University of California, San Francisco.

Ekman, P. Friesen, W. V. e Tomkins, S. S. (1971). Facial Affect Scoring Technique. A first validity study. *Semiotica*, **3**, 37-58.

Ekman, P., Hager, J., Irwin, W. e Methvin, C. (1969). *Ekman-Hager Facial Action Exemplars*. Human Interaction Laboratory, University of California, San Francisco.

Essa, I.A. e Pentland, A. P. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *Pattern Analysis and Machine Intelligence. IEEE Transactions on*, **19**, 757-763.

Fasel, B. e Luettin, J. (2003). Automatic facial expression analysis: Survey. *Pattern Recognition*, **36**, 259–275.

Feldt, K.S., Hamamoto, D.T., Hodges, J.S., Hsu, K.T. e Shuman, S.K. (2007). *The application of facial expressions to the assessment of orofacial pain in cognitively impaired older adults*. Department of Primary Dental Care, University of Minnesota.

Ford, G. (2002). Fully automatic coding of basic expressions from video. *Technical Report INC-MPLab*.

Freund, Y. e Schapire, R.E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, 148–156.

Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. Academic Press, San Diego.

Fridlund, A. J., Oster, H. e Ekman, P. (1987). Facial Expressions of Emotion. *Nonverbal behavior and communication*, **2nd ed.**, 143-223.

Friedman, J. (1991). Multivariate Adaptive Regression Splines. *Annals of Statistics*, **19**, 1–141.

Friedman, J., Hastie, T. e Tibshirani, R. (2008). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**, 1-22.

Friesen, W. V e Ekman, P. (1992). *Changes in FACS Scoring: Instruction manual*. Human Interaction Lab., San Francisco.

Frijda, N. H., Kuipers, P. e Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, **57**, 212-228.

Gross, R., Jain, K. e Li, S.Z. (2005). *Face Databases, Handbook of Face Recognition*. Springer-Verlag, New York.

Hand D. J. e Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, **45**, 171-186.

Hastie, T., Tibshirani, R e Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition*. Springer-Verlag, New York.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97-109.

Heisele, B., Serre, T., Poggio, T. e Pontil, M. (2001). Component-based face detection. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, **1**, 657-662.

Hjortsjö, C. H. (1969). *Man's Face and Mimic Language*. Nordens Boktryekeri, Malmö.

Huang, C.-L. e Huang, Y.-M. (1997). Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification. *Journal of Visual Communication and Image Representation*, **8**, 278-290.

Izard, C. E. (1977). *Human Emotions*. Springer, New York.

Izard, C.E. (1979). *The maximally discriminative facial movement coding system (MAX)*. Instructional Technology Center, University of Delaware.

Izard, C.E. e Dougherty, L.M. (1980). *System for identifying affect expressions by holistic judgments (Affex)*. Instructional Resources Center, University of Delaware.

Jones, M. e Viola, P. (2001). Robust real-time object detection. In *International Workshop on Statistical and Computational Theories of Vision - Modeling, Learning, Computing, and Sampling*, **CRL-2001-1**, 24 pagine.

Kanade, T., Cohn, J.F. e Tian, Y. (2000). Comprehensive database for facial expression analysis. *Automatic Face and Gesture Recognition. Proceedings. Fourth IEEE International Conference on*, 46-53.

Kanade, T. e Tian, Y. (2005). *Facial Expression Analysis*. University of Pittsburgh, Pittsburgh.

Li, S. e Gu, L. (2001). Real-time multi-view face detection, tracking, pose estimation, alignment, and recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition Demo Summary*.

Li, S. Z. e Jain, A. K. (2005). *Handbook of Face Recognition*. Springer Science Business Media, New York.

Lien, J.J., Kanade, T., Cohn, J. F. e Li, C.C. (1998). Automated Facial Expression Recognition Based on FACS Action Units. *Conf. Automatic Face and Gesture Recognition. Proceedings. Third IEEE International Conference on*, 390-395.

Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z. e Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *Computer Vision and Pattern Recognition Workshops. IEEE Computer Society Conference on.*, 94-101.

Mardia, K. V., Kent, J.T. e Bibby J.M. (1979). *Multivariate Analysis*. Academic Press, Londra.

Marin, J.M. e Robert, C.P. (2007). *Bayesian core: a practical approach to computational Bayesian statistics*. Springer Science+Business Media, New York.

Matthews, I e Baker, S. (2004). Active Appearance Models Revisited. *International Journal of Computer Vision* **60**, 135-164.

Osgood, C.H., Hastorf, A. H. e Ono, H. (1966). The Semantics of Facial Expressions and the Prediction of the Meanings of Stereoscopically Fused Facial Expressions. *Scandinavian Journal of Psychology*, **7**, 179–188.

Pantic, M. e Rothkrantz, L. J. M. (2000). Expert System for Automatic Analysis of Facial Expressions. *Image and Vision Computing*, **18**, 881–905.

Pentland, A., B. Moghaddam, B. e Starner, T. (1994). View-based and modular eigenspaces for face recognition. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 84–91.

R Development Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Robert, C.P. e Casella, G. (2004) .*Monte Carlo statistical methods*. Springer-Verlag, New York.

Robert, C.P. e Casella, G. (2010) .*Introducing Monte Carlo Methods with R.*. Springer-Verlag, New York.

Rowley, H.A., Baluja, S. e Kanade, T. (1998). Neural network-based face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on.*, **20**, 23-38.

Schneiderman, H. e Kanade, T. (2000). A statistical method for 3d object detection applied to faces and cars. *In IEEE Conference on Computer Vision and Pattern Recognition*, **1**, 746-751.

Stone, C. J, Kooperberg, C. e Bose, S. (1997). Polychotomous regression. *Journal of the American Statistical Association*, **92**, 117–127.

Tian, Y.-L., Brown, L., Hampapur, A., Pankanti, S., Senior, A. e Bolle, R. (2003). Real world real-time automatic recognition of facial expressions. *Proceedings of IEEE Workshop on performance evaluation of tracking and surveillance*, 9-16.

Tian, Y., Kanade, T. e Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **23**, 97-115.

Tian, Y., Kanade, T. e Cohn, J. F. (2002). Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. *Proceedings of the 5th IEEE International Conference on*, **ed. 2002**, 229 - 234.

Wang, P., Barrett, F., Martin, E., Milonova, M., Gurd, R. E., Gurb, R. C., Kohler, C. e Verma, R. (2008). Automated video- based facial expression analysis of neuropsychiatric disorders. *Journal of Neuroscience Methods*, **168**, 224–238.

Wen, Z e Huang, T. (2003). Capturing subtle facial motions in 3d face tracking. *Proc. of Int. Conf. On Computer Vision*, **2**, 1343 - 1350.

Whitehill, J., Bartlett, M. e Movellan, J. (2008). Automatic facial expression recognition for intelligent tutoring systems. *IEEE Conference on Computer Vision and Pattern Recognition*, 1-6.

Woodworth, R. S. (1938). *Experimental psychology*. Holt, Oxford.

Yacoob, Y. e Davis, L.S. (1996). Recognizing human facial expressions from long image sequences using optical flow. *Pattern Analysis and Machine Intelligence, IEEE Transactions on.*, **18**, 636 - 642.

Yee, T. W. e Wild, C. J. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society, Series B, Methodological*, **58**, 481–493.

Zhang, Z., Lyons, M., Schuster, M. e Akamatsu, S. (1998). Comparison Between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-layer Perceptron. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, **ed. 2008**, 454 - 459.

Zhu, J., Rosset, S., Zou, H. e Hastie, T. (2009). Multi-class AdaBoost. *Statistics and Its Interface*, **2**, 349–360.