Indice:

* Stato dell’arte:

1. Introduzione
2. Studi sull’engagement all’interno degli ambienti di apprendimento
3. Studi sul riconoscimento basati su FACS per scelta del modello da utilizzare
4. Studi che trattano campioni prelevati in scenari non controllati per una migliore precisione del modello
5. Cosa sono le AUs
6. Metodologie di tagging delle immagini del dataset
7. Estrazione delle feature facciali dalle immagini e dai video dei dataset

* Sviluppo:
  1. Scelta del modello
  2. Scelta del dataset post analisi papers
  3. Descrizione delle features del dataset
  4. Estrazione AUs:
     + Immagini
     + Video
  5. Pulizia dei dati
  6. Creazione descrizione in linguaggio naturale
  7. Strumenti utilizzati
  8. Applicazione del modello scelto
* Analisi dei risultati:
  1. ?

1. Introduzione

“*The test of successful education is not the amount of knowledge that pupils take away from school, but their appetite to know and their capacity to learn*.”

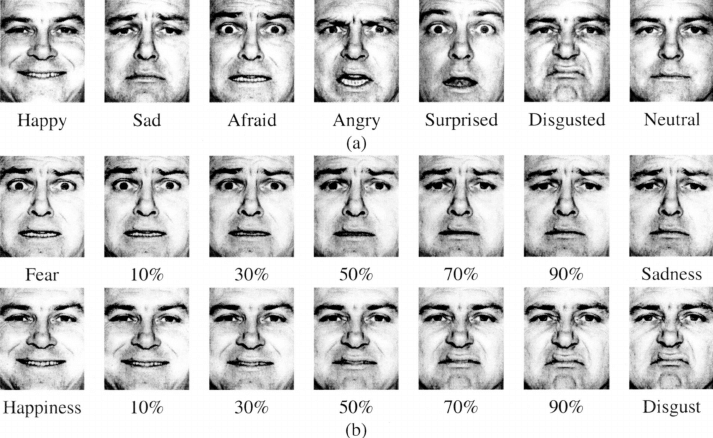
-Sir Richard Livingstone 1941. [5]

La ricerca moderna, conseguentemente al fenomeno della pandemia da *Covid-19*, ha riservato attenzione sempre maggiore al tema della fruizione di contenuti, sia dal punto di vista dell’intrattenimento che dal punto di vista della produttività individuale.

La pandemia ha quindi consentito alla ricerca l’impiego di strumenti impattanti in questo campo, in quanto ormai possibile osservare, e quindi analizzare, le attitudini comportamentali del singolo per migliorarne l’esperienza di utilizzo.

Effettuando diverse ricerche a riguardo ho potuto constatare che la maggior parte degli studi precedentemente effettuati si concentra sull’analisi delle emozioni che vengono categorizzate come primarie dal sistema FACS (Facial Action Coding System) [6]:

* Happiness,
* Anger,
* Sadness,
* Disgust,
* Fear,
* Neutral,



Di contro argomenti come, ad esempio, gli stati d’animo che influenzano l’engagement degli studenti durante l’apprendimento sono stati ampiamente trascurati.

Diversi studi hanno dimostrato che le emozioni positive sono direttamente collegate al pensiero creativo e alla capacità dell’individuo di riflettere su ciò che sta compiendo, indi una maggiore agevolezza e beneficio nell’apprendimento che ne consegue; diversamente le emozioni negative sono correlate ad una maggiore difficoltà di esercitare queste caratteristiche portando ad un minore rendimento relativamente a questo ambito.

Per quanto le emozioni primarie forniscano sicuramente un indicatore dello stato d’animo, e del conseguente miglioramento dell’esperienza di apprendimento o di lavoro, non sono però generalmente manifestate in modo esplicito, soprattutto per quanto concerne le espressioni facciali, in contesti di ufficio e studio.

Coerentemente si è ritenuto fondamentale concentrare lo studio sulle espressioni facciali più frequenti all’interno di questi ambienti.

------------------

2. Studi sull’engagement all’interno degli ambienti di studio

Gli studi ritrovati che effettuano quest’analisi o trattano un argomento simile o tangente sono:

* [1] Recognizing Cognitive Emotions in E-Learning Environment:

In questo studio gli stati di d’animo che vengono classificati dal loro sistema sono i seguenti:

* + - Entusiasmo,
    - Interesse,
    - Sorpresa,
    - Noia,
    - Perplessità,
    - Frustrazione,
    - Neutrale

Nello studio viene riportato che le emozioni positive, quindi entusiasmo, interesse e sorpresa sono spesso associate al raggiungimento del “flow state” da parte dello studente.

Più il singolo soggetto mantiene costante un ‘’mood’’ positivo tanto più permarrà in questo flow state, con esito un apprendimento più veloce ed efficace.

Il raggiungimento di questo stato è stato inoltre connesso alla capacità dei singoli studenti di percepirsi come autosufficienti nell’attività di studio.

Questo percepirsi come autosufficienti deriva dalla capacità dello studente di riuscire ad essere in controllo della sua personale situazione di studio, rispecchiandosi nella sua capacità di:

* + - Pianificare,
    - Controllare,
    - Dirigere

l’attività di apprendimento.

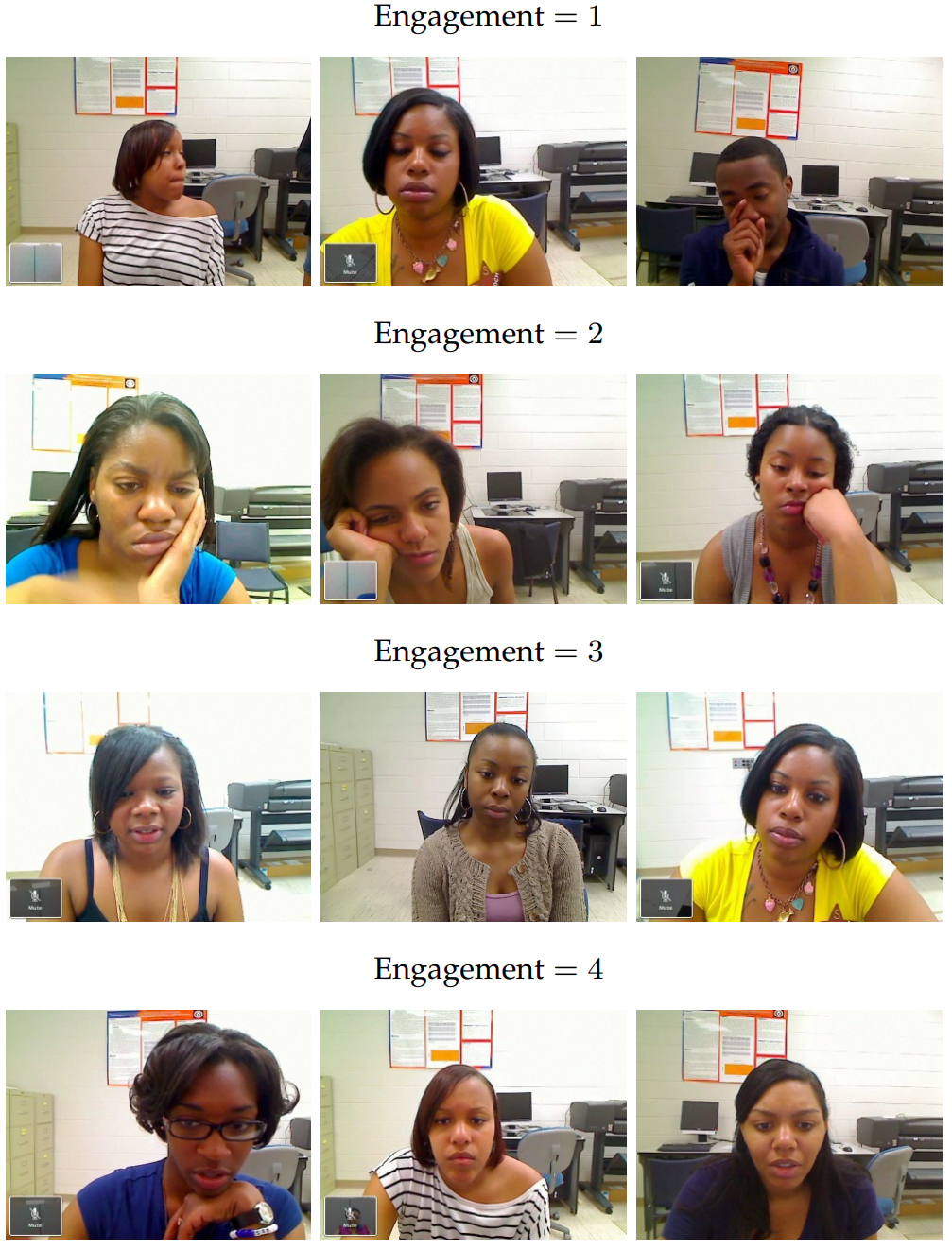
È quindi necessario che gli studenti riescano a maturare un senso di confidenza nei riguardi dello studio, e che vengano dunque supportati in vista dell’approccio ai problemi che vengono da loro individuati come difficili.

Ovviamente un sistema che permette di “leggere” con precisione lo stato d’animo di uno studente o una studentessa o persino di un’intera classe, e quindi capire se questi si presentano in uno stato di flow che possa permettere una migliore performance, è uno strumento utile per qualsiasi insegnate.

Un esempio circoscritto all’ambiente di lavoro potrebbe invece essere un’analisi del lavoratore nello svolgimento di un task e la possibilità da parte di un capo progetto o di un tutor di poter intervenire solo nel momento in cui il suo subordinato sta riscontrando dei problemi, in modo da poter permettere una crescita professionale adeguata e non seguita al 100% e sfruttare al meglio l’impiego del tempo del tutor.

* [4] The faces of Engagement: Automatic Recognition of Student Engagement from Facial Expression

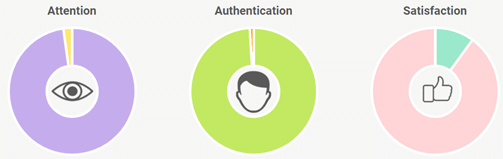
Nello studio gli stati d’animo che sono organizzati in scala, da meno attento/a a più attento/a:

* + - Not engaged at all (Non coinvolto: che guarda da un’altra parte, che sta ovviamente non pensando al compito, occhi completamente chiusi)
    - Nominally engaged (Formalmente coinvolto: occhi appena aperti, chiaramente non attento/a al task che sta svolgendo)
    - Engaged in task (Coinvolto nel task: requisito che non richiede un’ammonizione per la continuazione del task)
    - Very engaged (Altamente coinvolto: lo/a studente potrebbe essere elogiato/a per il suo livello di coinvolgimento)
    - Clip/frame non chiaro (l’immagine analizzata non contiene una persona o comunque non è possibile effettuare un’identificazione)
  + 
  + Lo studio si concentra sull’effettuare una stima dell’engagement degli studenti. Hanno inizialmente sviluppato un metodo per rilevare automaticamente l’engagement, hanno poi indagato su quali segnali siano utilizzati nel riconoscimento automatico effettiatp dao computer e hanno poi analizzato invece quali strumenti vengano utilizzati dagli insegnanti per risolvere lo stesso task.
  + In fine hanno investigato se i risultati di queste analisi si correlano effettivamente con la qualità delle performance degli studenti.
* [6] Facial coding as a mean to enable continuous monitoring of student’s behaviour in e-Learning

Il paper si focalizza sul tracciamento continuo degli studenti, sia per quanto riguarda una vera e propria identificazione degli stessi attraverso il riconoscimento facciale, sia per calcolarne il livello di attenzione ed eventualmente stimarne le emozioni provate durante i corsi MOOCs (Massive Open Online Courses).

Per effettuare l’analisi del livello di attenzione è stata utilizzata una libreria esterna Dlib che concede di creare una mappatura delle caratteristiche facciali dello/a studente; per giunta, la piattaforma include anche un gaze tracker che lascia prevedere in quale direzione gli utenti stiano guardando mentre seguono il corso.

Questi tre aspetti vengono successivamente uniti per creare un applicativo web per l’apprendimento attraverso il quale, alla fine di ogni lezione, è possibile visionare in quale percentuale della durata del corso le persone che lo utilizzano sono risultate attente.



* [7] Prediction and Localization of student engagement in the wild

Questo studio, a differenza di altri, ha come premessa l’utilizzo delle immagini raccolte in ambienti non controllati per la creazione del modello che andrà successivamente ad effettuare la predizione per i nuovi campioni.

Per ambienti controllati si intende setup di acquisizione dei video e delle immagini grazie ai quali non è possibile trovare problemi quali, scarsa illuminazione, occlusione ambientale, etc…

Per attuare ciò, hanno analizzato molti studi precedentemente effettuati, ed hanno scelto di raccogliere i campioni attraverso la visione, da parte dei soggetti, di video educazionali categorizzando poi i vari video ed immagini ottenute in una scala, con valore da 0 a 3:

* + - 0 🡪 per niente interessato (il soggetto non sembra interessato e guarda spesso al di fuori dello schermo)
    - 1 🡪 poco interesse (il soggetto apre a malapena gli occhi, si muove in modo irrequieto sulla sedia)
    - 2 🡪 interessato/a al contenuto (sembra che al soggetto il contenuto riprodotto risulti interessante ed esso interagisce con questo)
    - 3 🡪 altamente interessato/a (il soggetto ha “gli occhi attaccati allo schermo” e risulta concentrato/a)

Hanno successivamente utilizzato un framework che esegue il riconoscimento dell’engagement e della localizzazione degli studenti.

* + - Inizialmente vengono identificate la faccia e dei punti di riferimento all’interno di queste in ognuno dei frame analizzati
    - Procedendo, i video vengono suddivisi in segmenti più piccoli e le feature vengono estratte “effettuando una media” dei risultati di ognuno dei frame.
    - Si passa poi alla sequenza di frame successiva per effettuare la stessa analisi.
    - Una volta raccolti tutti i dati, questi vengono elaborati per calcolarne l’engagement e la localizzazione attraverso la deep MIL network utilizzando la media e la top-k pooling per calcolarne la regressione.



Esempio di campioni dal loro dataset

------------------------

3. Studi sul riconoscimento delle emozioni FACS per scelta del modello da utilizzare

Essendo l’ammontare di studi che trattano l’analisi delle emozioni FACS maggiore rispetto a quelle che cercano di creare sistemi di riconoscimento automatico per gli stati d’animo che possono, direttamente, aiutare ad identificare i problemi nell’apprendimento delle conoscenze ho ritenuto corretto studiare e basare i modelli da utilizzare per effettuare un’elaborazione di queste informazioni per il mio caso di studio, attraverso i modelli proposti dai paper relativi alle FACS.

Fra i vari studi analizzati, per arrivare ad una conclusione circa la scelta del modello, quello risultato il più utile è stato [2]:

in questo studio vengono utilizzate le CNN (Convolutional Neural Networks), queste estraggono le feature facciali dalle immagini che successivamente vengono date in input a classificatori standard per eseguire una classificazione di queste emozioni.

Nello studio sono stati utilizzati i dataset FER 2013 e RAF DB per l’analisi delle emozioni FACS (felicità, tristezza, sorpresa, paura, disgusto, rabbia, stato neutrale), si è fatto ricorso a diversi metodi per analizzare i dati estratti e i risultati di ognuno di questi sono stati confrontati fra di loro:

Nello specifico i metodi utilizzati sono:

* + STN (Spatial Transformer Networks): reti neurali utilizzate per effettuare la trasformazione geometrica degli input, ovvero per eseguire operazioni di rotazione, traslazione e scaling sui dati di input. Queste reti sono in grado di apprendere in maniera automatica tali trasformazioni e di applicarle direttamente ai dati di input.
  + SE (Squeeze and Excitation Networks): tecnica di rete neurale che si concentra sullo sfruttare la correlazione tra i canali delle feature map, al fine di migliorare la loro rappresentazione. Sostanzialmente, le reti SE "estraggono" (squeeze) i dati di input in un singolo vettore, calcolano l'importanza di ogni canale e "stimolano" (excite) i canali più rilevanti, migliorando così la qualità delle feature map.
  + BAM (Bottleneck Attention Module): modulo di attenzione che utilizza una tecnica di "bottleneck" per ridurre il numero di feature map da elaborare, rendendo il processo più efficiente. In particolare, il BAM sfrutta un'operazione di pooling per creare una rappresentazione ridotta dei dati di input, che viene poi sfruttata per calcolare l'attivazione di ogni canale delle feature map originali.
  + Immagine che contiene testo, ricevuta, schermata

    Descrizione generata automaticamenteCBAM (Convolutional Bottleneck Attention Module): è una versione migliorata del BAM che utilizza sia l'attenzione spaziale che quella di canale. In pratica, il CBAM esegue prima un'operazione di attenzione spaziale per calcolare l'importanza delle diverse regioni dell'immagine, e successivamente utilizza un'operazione di attenzione di canale per calcolare l'importanza dei diversi canali delle feature map. Questo rende il CBAM particolarmente utile per il riconoscimento di oggetti in immagini complesse.

Effettuando un confronto fra questi, il modello BAM è quello che offre una performance migliore sui due dataset, e a seguire l’STN.

Le emozioni che lo studio si propone di valutare non sono esattamente quelle predisposte per lo studio di questa tesi, ma le valutazioni estratte da questo studio si possono ritenere un buon metodo di valutazione del modello da scegliere.

--------------------------

4. Studi che trattano campioni prelevati in scenari non controllati per una migliore precisione del modello

Un altro problema trattato solo da [7] (fra i paper ritrovati a tema engagement e stati d’animo) e che invece ho notato più prominente all’interno della ricerca riguardo le emozioni FACS è quello dei campioni denominati come “in-the-wild”.

Molti degli studi effettuati si concentrano su campioni prodotti in ambienti controllati, con nessun tipo di interferenza per quanto concerne i problemi che abitualmente si possono riscontrare quando invece si lavora con riprese che potrebbero avere una valenza anche quando le migliori condizioni non si presentano.

Come riporta [3] “The videos are recorded under controlled conditions, e.g. illumination is uniform, background is static, and there is a limited amount of head pose variation and occlusion. Although a range of affective states are displayed and recorded, emotions are elicited by a limited number of tasks, e.g., in [50], all subjects underwent exactly the same tasks.”

Quindi i video creati in condizioni verificate, di luce uniforme, sfondo statico, con insufficienti variazioni nella posa delle persone, poca occlusione e dove le persone svolgono lo stesso compito non sono validi per creare un modello che possa effettuare delle predizioni veritiere che si possano applicare ad ogni ambiente.

Per quanto riguarda la raccolta di immagini, negli studi “in the wild” è stata messa in atto la prelevazione e categorizzazione delle immagini da film, serie tv, e altri media che non rispettano gli standard imposti durante la creazione dei dataset prelevati invece in ambienti controllati [3].

Per quanto riguarda il tagging delle immagini sono stati consultati due annotatori esperti FACS AU che hanno individuato lo stato d’animo della o delle persone presenti all’interno della scena presentata attraverso una piattaforma TODO.

Entrambi gli annotatori erano sempre presenti durante l’analisi dei singoli campioni ed hanno quindi potuto discutere ognuno dei casi incerti insieme; una volta finite le prime analisi sono state inoltre ripresentati alcuni dei frame già analizzati agli stessi annotatori in modo da effettuare un’ulteriore verifica sulla loro consistenza e essi hanno dato la stessa annotazione l’87% delle volte. Le annotazioni dello studio sono quindi considerabili altamente affidabili.

Per costruire il modello hanno poi utilizzato l’estrazione delle AU, attraverso un sistema semi automatico, ed hanno utilizzato questi dati estratti per la costruzione del modello.

I risultati di queste analisi hanno portato alla conclusione che i modelli sviluppati attraverso raccolte di immagini prelevate in scenari controllati non portano necessariamente a risultati puntuali (o accurati) quando vengono applicati al mondo reale o “in the wild”.

----------------

5. Cosa sono le AUs e il sistema FACS:

[8] Le FACS, acronimo di Facial Action Coding System, sono un sistema di codifica delle espressioni facciali sviluppato dallo psicologo Paul Ekman e dal collega Wallace V. Friesen negli anni '70. Questo sistema si basa sull'identificazione di movimenti muscolari specifici del viso che si verificano durante le espressioni facciali.

Il sistema di codifica delle FACS è composto da un insieme di codici numerici, o "Action Units" (AU), che rappresentano le azioni muscolari specifiche che avvengono durante le espressioni facciali. Ci sono 66 AU, ciascuno dei quali rappresenta un'azione muscolare specifica, e ogni AU ha un codice numerico univoco.

La codifica delle FACS può essere utilizzata per identificare e descrivere le espressioni facciali in modo oggettivo e dettagliato. Ad esempio, un codificatore FACS può utilizzare il sistema per identificare l'AU corrispondente a un sollevamento delle sopracciglia, il codice 1, e l'AU corrispondente a un sorriso, il codice 12. Questi codici possono essere utilizzati per creare una descrizione dettagliata dell'espressione facciale di un individuo.

Le FACS sono state utilizzate in una vasta gamma di contesti, tra cui la ricerca scientifica, la valutazione clinica e la produzione di effetti speciali per il cinema e la televisione.

In sintesi, le FACS sono un sistema di codifica delle espressioni facciali che si basa sull'identificazione di movimenti muscolari specifici del viso. Questo sistema è stato utilizzato con successo in molti contesti diversi ed è un metodo utile per descrivere le espressioni facciali in modo oggettivo e dettagliato.

Sulla documentazione della libreria che ho utilizzato per l’estrazione delle AUs dal dataset aggregato ed utilizzato, è presente una tabella dove viene mostato:

il numero associato alla singola Action Unit, il loro relativo nome FACS, a quali muscoli della faccia sono correlate, la loro categoria FACS, le espressioni che vengono spesse legate a determinati valori di esse e i modelli che le utilizzano.

Riporto qui un estratto della tabella consultabile interamente al link presentato nel riferimento [8] con le AUs che vengono estratte per la creazione del mio modello:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AU1 | Inner Brow Raiser | Frontalis (medial) | main | sadness, surprise, fear |
| AU2 | Outer Brow Raiser | Frontalis (lateral) | main | surprise, fear |
| AU4 | Brow Lowerer | Procerus, Depressor Supercilii, Corrugator Supercilii | main | sadness, fear, anger |
| AU5 | Upper Lid Raiser | Levator Palpebrae Superioris, Superior Tarsal Muscle | main | surprise, fear, anger |
| AU6 | Cheek Raiser | Orbicularis Oculi (orbital) | main | happiness, disgust, contempt |
| AU7 | Lid Tightener | Orbicularis Oculi (palpebral) | main | fear, anger |
| AU9 | Nose Wrinkler | Levator Labii Superioris Alaeque Nasi | main | disgust |
| AU10 | Upper Lip Raiser | Levator Labii Superioris | main |  |
| AU11 | Nasolabial Deepener | Zygomaticus Minor | main | disgust, fear |
| AU12 | Lip Corner Puller | Zygomaticus Major | main | happiness, contempt |
| AU14 | Dimpler | Buccinator | main | contempt |
| AU15 | Lip Corner Depressor | Depressor Anguli Oris | main | sadness, disgust |
| AU17 | Chin Raiser | Mentalis | main | disgust |
| AU20 | Lip Stretcher | Risorius, Platysma | main | fear |
| AU23 | Lip Tightener | Orbicularis Oris | main | anger |
| AU24 | Lip Pressor | Orbicularis Oris | main |  |
| AU25 | Lip Part | Depressor Labii Inferioris | main | happiness, surprise, fear |
| AU26 | Jaw Drop | Masseter, Temporalis, Medial Pterygoid | main | fear, surprise |
| AU28 | Lip Suck | Orbicularis Oris | main |  |
| AU43 | Eyes Closed | Levator Palebrae Superioris (relaxation) | behavioral |  |

----------------------------------

6. Metodologie di tagging delle immagini del dataset

Per quanto riguarda la classificazione delle immagini dei dataset, vari studi hanno fatto ricorso a tecniche diverse:

* Self report sul proprio stato emotivo da parte delle persone dalle quali vengono prelevati i campioni per le analisi, attraverso questionari o in modo arbitrario
* Valutazione da parte di esperti, spesso più di uno, per assegnare uno stato emotivo alle singole immagini o video; o scelto in modo interamente arbitrario da parte del/dei valutatore/i o utilizzando come riferimento lo stato emotivo riportato dalla persona
* Classificazione attraverso metodi di clustering [9]

Strumenti utilizzati per effettuare la classificazione:

* Attraverso la piattaforma FEELTRACE (o altri software di annotazione manuale) che permette di registrare e analizzare i segnali EMG (Elettromiografici) e che permette di misurare la forza e la durata delle risposte emotive dei partecipanti, attraverso una scala da 0 a 10 mentre i partecipanti osservano immagini, video o altre forme di stimoli emotivi. [3]
* Questionari da sottoporre alle persone che vengono riprese per la creazione del dataset

-------------------------------

7. Estrazione delle feature facciali dalle immagini e dai video dei dataset

Per quanto riguarda l’estrazione dei dati dalle immagini utilizzate possono essere utilizzate diverse librerie e strumenti fra cui:

* OpenCV: una libreria open-source di computer vision che può essere utilizzata per il riconoscimento delle espressioni facciali. OpenCV contiene diverse funzioni per l'estrazione delle AUs, come la rilevazione delle linee facciali e la stima dei parametri di deformazione.
* dlib: una libreria di machine learning open-source che fornisce funzioni per la rilevazione e l'analisi delle facce nelle immagini. Dlib è in grado di rilevare le AUs attraverso l'utilizzo di una rete neurale convoluzionale (CNN) addestrata su un grande dataset di espressioni facciali.
* FaceReader: un software commerciale sviluppato dalla società olandese Noldus Information Technology che utilizza algoritmi di analisi dell'emozione per rilevare e classificare le AUs nelle immagini facciali. FaceReader è in grado di rilevare fino a 20 AUs e di classificarle in base alle emozioni associate.
* OpenFace: un framework open-source di computer vision sviluppato dall'Università di Carnegie Mellon che fornisce funzioni per l'estrazione delle AUs e la rilevazione delle espressioni facciali. OpenFace utilizza una combinazione di tecniche di machine learning e di analisi geometrica per rilevare le AUs.
* FACET: un software commerciale sviluppato dalla società americana Emotient (acquisita da Apple nel 2016) che utilizza una combinazione di tecniche di computer vision e di analisi dell'emozione per rilevare e classificare le AUs nelle immagini facciali. FACET è in grado di rilevare fino a 20 AUs e di classificarle in base alle emozioni associate.
* DeepFaceLab: un software open-source che utilizza reti neurali profonde per l'analisi dell'immagine e la manipolazione della faccia. Può essere utilizzato per l'estrazione delle AUs da immagini facciali.
* Pyfeat: una libreria open-source sviluppata in Python per l'estrazione delle feature facciali dalle immagini. Pyfeat si basa sull'utilizzo di un insieme di funzioni matematiche, chiamate funzioni di base, per descrivere la forma e l'aspetto delle espressioni facciali. Queste funzioni di base sono rappresentate da immagini di espressioni facciali standardizzate, come il sorriso o la rabbia, che sono state create e validate da esperti di psicologia e neuroscienze.

La libreria Pyfeat utilizza queste funzioni di base per calcolare un insieme di feature facciali, come la forma del viso, la posizione degli occhi, la forma delle sopracciglia e la posizione della bocca. Pyfeat è in grado di estrarre un insieme completo di 280 feature facciali, chiamate 2D FACS (Facial Action Coding System), sempre validate da esperti di psicologia e neuroscienze.

Pyfeat può essere utilizzato in diverse applicazioni di analisi facciale, come l'analisi delle espressioni facciali per la rilevazione delle emozioni, l'analisi del comportamento non verbale e la diagnostica medica.

Ho deciso di utilizzare questa libreria per l’estrazione delle AUs in quanto è affidabile ed è semplice nell’utilizzo, soprattutto attraverso le funzioni CUDA della libreria che mi hanno permesso di velocizzare di molto il processing delle immagini contenute nei dataset ritrovati.