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

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

1. Introduzione

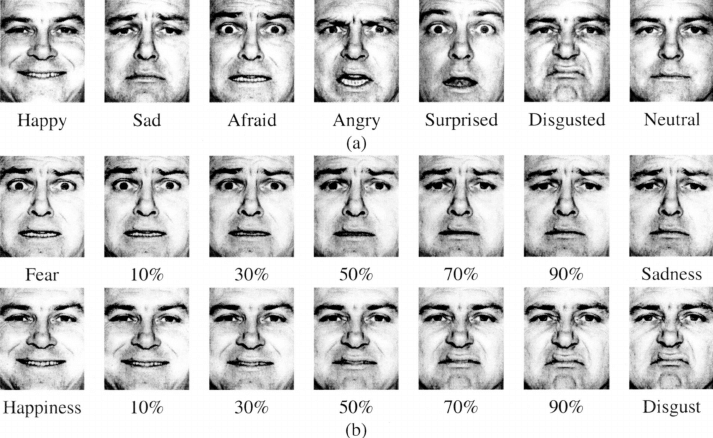
“The test of successful education is not the amount of knowledge that pupils take away from school, but their appetite to know and their capacity to learn.” Sir Richard Livingstone 1941. [5]

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 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,



mentre solo in pochi studi è presente uno studio riguardo, ad esempio, gli stati d’animo che influenzano l’engagement degli studenti durante l’apprendimento o comunque altri stati d’animo.

Diversi studi hanno dimostrato che le emozioni positive sono direttamente collegate con il pensiero creativo e la capacità dell’individuo di riflettere su ciò che sta facendo e quindi una maggiore facilità e piacere nell’apprendimento; mentre le emozioni negative sono correlate con una maggiore difficoltà di esercitare queste caratteristiche e quindi con performance minori in questi ambiti.

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 eloquente, ancor di più attraverso le espressioni facciali, in contesti di ufficio e studio.

Si è quindi ritenuto importante concentrare lo studio sulle espressioni facciali che sono 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 correlate con il raggiungimento del “flow state” da parte dello studente.

Più il singolo soggetto riesce a restare in un mood positivo tanto più riesce a restare in questo flow state che porta ad un apprendimento più veloce ed efficacie.

Hanno inoltre correlato il raggiungimento di questo stato con la capacità dei singoli studenti di percepirsi come autosufficienti nell’attività di studio.

Questo percepirsi come autosufficienti è dato dalla capacità dello studente di riuscire ad essere “in controllo” della sua personale situazione di studio e si riflette quindi nella sua capacità di:

* + - Pianificare,
    - Controllare,
    - Dirigere

l’attività di apprendimento.

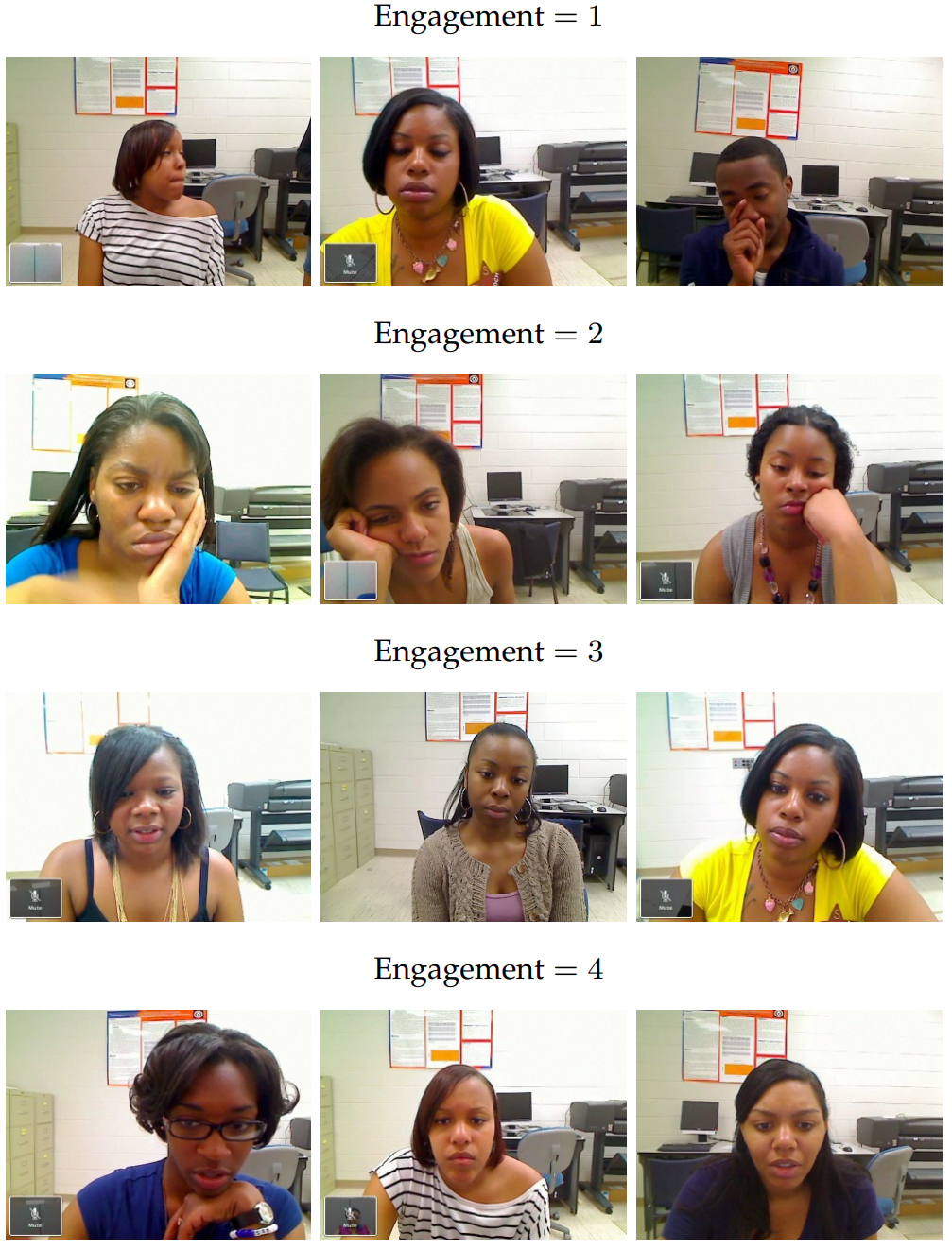
È quindi necessario che gli studenti riescano a creare un senso di confidenza nei riguardi dello studio ed è quindi necessario che vengano supportati quando si approcciano a problemi che vengo da loro percepiti come difficili.

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 sono in uno stato di flow che permette performance migliori, è uno strumento utile per qualsiasi insegnate.

Un esempio all’interno di un ambiente di lavoro potrebbe invece essere un’analisi del lavoratore nello svolgimento di una task e la possibilità da parte di un capo progetto o di un tutor di poter intervenire solo quando il suo subordinato sta riscontrando dei problemi in modo da poter permettere una crescita professionale adeguata e non seguita o viceversa al 100% e efficienza nell’impiego del tempo del tutor.

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

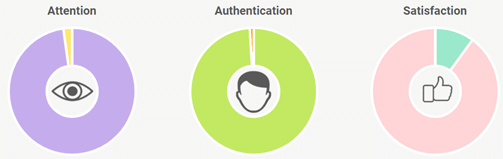
Nello studio gli stati d’animo che sono organizzati in scala, da meno attento/a a più attento/a:

* + - Not engaged at all (Del tutto non coinvolto: che guarda da un’altra parte, che sta ovviamente non pensando al compito, occhi completamente chiusi)
    - Nominally engaged (coinvolto per dire: occhi appena aperti, chiaramente non attento/a alla task che sta svolgendo)
    - Engaged in task (coinvolto nella task: che non richiede un’ammonizione per continuare la task)
    - Very engaged (molto coinvolto: lo/a studente potrebbe essere elogiata 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)
  + 
  + TODO scrivere altro a riguardo del paper
* [6] Facial coding as a mean to enable continuous monitoring of student’s behaviour in e-Learning

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

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

Questi 3 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 state attente.



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

Questo studio, a differenza di altri, ha come premessa il fatto di utilizzare 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 fare questo hanno analizzato molti studi precedentemente effettuati ed hanno scelto di raccogliere i campioni attraverso la visione, da parte dei soggetti, di video educazionali, ed hanno categorizzato 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 effettua il riconoscimento dell’engagement e della localizzazione degli studenti.

* + - Inizialmente vengono identificate la faccia e dei punti di riferimento all’interno di queste all’interno di ognuno dei frame analizzati
    - Successivamente i video vengono suddivisi in segmenti più piccoli e le feature vengono estratte “sommando” i 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.

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 le gli stati d’animo che possono, direttamente, aiutare ad identificare i problemi nell’apprendimento delle conoscenze, ho ritenuto corretto studiare e basare l’estrapolazione delle feature facciali e, conseguenzialmente, 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 quello che mi è risultato il più ben strutturato è stato [2]:

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

Nello studio effettuato in [2] sono stati utilizzati i dataset FER 2013 e RAF DB per l’analisi delle emozioni FACS (felicità, tristezza, sorpresa, paura, disgusto, rabbia, stato neutrale) e sono stati utilizzati diversi metodi per analizzare i dati estratti e i risultati di ognuno di questi sono stati confrontati fra di loro:

in particolare i metodi utilizzati sono:

* + STN (SpatiaL Transformer Networks): sono 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): sono una tecnica di rete neurale che si concentra sullo sfruttare la correlazione tra i canali delle feature map, al fine di migliorare la loro rappresentazione. In pratica, le reti SE "spremono" (squeeze) i dati di input in un singolo vettore, calcolano l'importanza di ogni canale e "eccitano" (excite) i canali più rilevanti, migliorando così la qualità delle feature map.
  + BAM (Bottleneck Attention Module): è un 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 utilizzata 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 ottiene le 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 possono essere un buon metodo per valutare il 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] e che invece ho ritrovato più prominente all’interno della ricerca riguardo le emozioni FACS è quello dei campioni denominati come “in-the-wild”.

Molti degli studi effettuati si concentra su campioni prodotti in ambienti controllati, con nessun tipo di problema per quanto riguarda i problemi usuali che 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 controllate, di luce uniforme, con sfondo statico, con insufficienti variazioni nella posa della persone, poca occlusione e dove le persone svolgono lo stesso compito non sono validi per creare un modello che possa effettuare delle predizioni veritiere in ogni ambiente.

Per quanto riguarda la raccolta di immagini, negli studi “in the wild” è stata effettuata 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 in ambienti controllati.

Per quanto riguarda la classificazione delle immagini dei dataset vari studi hanno utilizzato tecniche diverse:

* Self report sul proprio stato emotivo da parte delle persone dalle quali vengono prelevati i campioni per le analisi, attraverso dei 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 valutatore o utilizzando come riferimento lo stato emotivo riportato dalla persona
* Attraverso la piattaforma FEELTRACE (o software simili) è possibile aggiungere annotazioni ai video in tempo reale utilizzando un joystic [3]

Per quanto riguarda l’estrazione dei dati dalle immagini utilizzate:

[2]

Gli approcci che utilizzano il deep learning rappresentano la soluzione allo stato dell’arte per quanto riguarda la classificazione delle emozioni basate sulle espressioni facciali.

* [3] prediction and localization of student engagement in the wild
  + Lo studio si propone di effettuare un’analisi simile a quella effettuata da [2] ma prendendo in campione delle immagini estrapolate fuori da contesti controllati.
  + Le difficoltà principali affrontate in questo studio sono date dal fatto che le espressioni facciali sono caratterizzate da minime deformazioni facciali che sono difficili da tracciare e per questo lo studio si è predisposto di avere un approccio con particolare attenzione nel trovare quali feature facciali sono le più importanti per effettuare le analisi.
  + Questo problema, secondo i redattori dello studio, è dato dal fatto che la maggior parte dei dataset che vengono utilizzati di solito viene generato in ambienti controllati e con un numero limitato di pose facciali e occlusione ambientale; questo porta ovviamente i modelli generati negli studi che utilizzano questo tipo di dati ad essere meno affidabili quando si vogliono effettuare delle valutazioni in contesti non controllati.
  + I ricercatori hanno quindi deciso di ritrovare delle immagini da analizzare attraverso clip “dal mondo reale” e da scene di film, serie tv e programmi tv.