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 a identificare i problemi nell’apprendimento delle conoscenze, ho ritenuto corretto studiare e scegliere fra i modelli da loro proposti per l’elaborazione delle informazioni per il mio caso di studio.

Il paper con i migliori risultati che ho trovato è stato [12], in questo studio viene utilizzata una random forest a due livelli:

* il primo livello è utilizzato per determinare le action units nelle sequenze di espressioni che vengono analizzate.
* Il secondo livello invece prende in input le action units estratte ed effettua la classificazione delle espressioni finali.

Il modello riesce ad ottenere una media di riconoscimento del 100% per le action units e del 96.38% per le espressioni facciali (emozioni) che risulta essere più alto dei prossimi studi riportati.

Prima ancora di arrivare al primo livello della random forest viene utilizzato il modello AAM che identifica i punti di riferimento sulla faccia del soggetto ripreso nel video analizzato (o nella sequenza di immagini analizzate) e successivamente viene utilizzato il Lucas-Kanade optical flow tracker per tracciare i punti di riferimento nei frame successivi.

Viene poi calcolato il vettore di spostamento fra i punti di riferimento trovati sul frame con un’espressione neutrale e i punti di riferimento trovati nel frame dove l’espressione facciale è al suo picco.

Questo vettore di spostamento viene utilizzato nel primo layer della random forest per la prelevazione delle Action Units.

Le espressioni facciali sono processi dinamici che derivano, come conseguenza, dell'attività muscolare facciale di varie parti di questa:

* fronte,
* occhi,
* naso,
* bocca,
* mascella,
* contorno

Le caratteristiche dinamiche possono essere rappresentate dalla differenza tra fotogrammi in termini di punti. Le sei espressioni facciali di base:

* felicità,
* tristezza,
* rabbia,
* disgusto,
* paura,
* sorpresa

possono essere descritte utilizzando le AUs, dove ogni UA è codificata in base ai coinvolgimenti muscolari facciali.

Il processo di estrazione delle caratteristiche di movimento di espressione facciale è illustrato in figura

Immagine che contiene Sito Web

Descrizione generata automaticamente

Processo di estrazione dinamica delle feature delle espressioni facciali

Random Forest è un algoritmo di machine learning che combina molti alberi decisionali, dove ogni albero dipende dai valori di un vettore casuale campionato in modo indipendente e con la stessa distribuzione.

Questo algoritmo è in grado di gestire grandi spazi di caratteristiche grazie alle due fonti di casualità, ovvero input e feature casuali, che lo rendono robusto.

Random Forest è diventato molto potente per l'estimazione della posa, la rilevazione di oggetti e altre aree in computer vision grazie alla sua bassa complessità di calcolo, facilità di implementazione, accuratezza nella classificazione e capacità di gestire grandi set di dati di formazione.

Il Random Forest è composto da molti alberi decisionali, dove ogni albero viene costruito ricorsivamente assegnando un test binario ad ogni nodo non foglia in base ai campioni di formazione.

Per la classificazione, il Random Forest combina i risultati degli alberi decisionali per votare per la classe più popolare.

Questo algoritmo ha la capacità di gestire errori di bilanciamento nella popolazione delle classi in dati sbilanciati.

Per validare i risultati i ricercatori hanno degli esperimenti sui datasets esteso Cohn-Kanade e Oulu-CAISA VIS utilizzando solo le sequenze di immagini che contengono una vista frontale o girata a 30 gradi dei soggetti.

Nel primo dataset sono categorizzate 7 espressioni (anger, contempt, disgust, fear, happy, sadness and surprise) e le persone riprese hanno un’età compresa fra i 18 e i 50 anni.

Il secondo dataset contiene samples ripresi in diverse condizioni di luci ma sono state analizzate solo le immagini con condizioni di illuminazione normali (luce forte o buona illuminazione).

In entrambi i dataset i video riprendono delle persone con, inizialmente, delle espressioni neturali e alla fine le espressioni che più esaltano le emozioni che vogliono rappresentare.

Nella fase finale del metodo proposto, viene utilizzato l’ultimo layer della random forest per identificare l'etichetta finale dell'espressione facciale in base ai risultati della rilevazione delle AU.

Viene utilizzata una rete bayesiana come confronto al metodo proposto; questo modello predittivo otteiene un tasso di riconoscimento delle caratteristiche facciali di solo l'86,3%.

Il metodo proposto dal paper, basato sulle AUs, può invece aumentare il tasso di riconoscimento medio al 89,37%.

I risultati mostrati nella tabella riportata nell’immagine qui sotto sono il tasso di riconoscimento per la rilevazione dell'espressione facciale basata sulle AU utilizzando il random forest.

Immagine che contiene tavolo

Descrizione generata automaticamente Confrontando i risultati della Tabella riportata nell’immagine qui sotto, le prestazioni sono migliorano significativamente.

Immagine che contiene tavolo

Descrizione generata automaticamente Il classificatore con rete bayesiana ha un tasso di riconoscimento relativamente basso per le emozioni di tristezza e paura (rispettivamente 68,75% e 88,89%), mentre il random forest migliora queste prestazioni al 93,75% e al 100%.

Una possibile spiegazione è che la tristezza e la paura sono di solito poco evidenti e possono essere facilmente confuse con altre emozioni. Tuttavia, il random forest ha maggiore capacità discriminativa e può trovare la differenza tra di esse.

Inoltre, i tassi di successo di emozioni come la rabbia, la sorpresa e il disgusto sono stati significativamente migliorati utilizzando il framework proposto.

Per confermare l'efficacia del metodo, sono stati selezionati casualmente set di training e testing dal database e l'esperimento è stato ripetuto per 9 volte. Tutti i risultati sono mostrati nella Tabella qui sotto.

Immagine che contiene tavolo

Descrizione generata automaticamente È stato dimostrato che il sistema ha una prestazione stabile utilizzando set di training e testing diversi.

Il tasso di riconoscimento medio è significativamente più alto del metodo di confronto [15].

Un altro studio analizzato per la predizione delle emozioni attraverso l’analisi delle Action Units è [13].

In questo studio viene proposto un metodo per mappare le AUs a 6 emozioni:

* anger,
* fear,
* happy,
* sad,
* surprise,
* disgust

Vengono scelte delle AUs che poi vengono mappate alle emozioni attraverso relazioni sttistiche e delle tecniche di match.

Le relazioni fra le emozioni e le AUs sono collezionate come stringhe template che comprendono le più descrittive delle AUs per ognuna delle emozioni trattate.

Le stringhe di template vengono poi computate utilizzando un concetto chiamato discriminative power:

l’LCS, un metodo per approssimare quanto combaciano le sottostringhe alla stringa di template, è applicato per calcolare la vicinanza delle stringhe di test alle AUs con la stringa di template per la singola emozione.

Lo studio ha trovato che LCS è un metodo efficiente per gestire particolari problemi come il rilevamento errato delle AUs e aiuta a ridurre predizioni false.

Il metodo proposto è stato testato su vari datasets:

* CK+,
* ISL,
* FACS,
* JAFFE,
* Mind Reading,
* altri frame ripresi in condizioni in-the-wild

è stata poi effettuata una comparizione fra il metodo proposto e altri precedentemente proposti e si sono notati dei miglioramenti, sia sui datasets di benchmark che per quelli in-the-wild.

Il metodo proposto basa il rilevamento elle AUs sul metodo proposto da [16] e successivamente le AU vengono mappate con il metodo proposto nel paper.

Immagine che contiene diagramma

Descrizione generata automaticamente

Per l’allenamento del modello per il riconoscimento delle AU sono state utilizzate 580 immagini dal CK+ dataset, e, per ognuna di queste, sono state effettuate varie modifiche alle immagini per renderle utilizzabili; è stato applicato l’AdaBoost per la selezione delle feature per ridurre la dimensione del vettore delle features ed è stato utilizzato il Support Vector Machines (SVM) per il modello delle AUs.

Dopo un’attenta analisi del sistema FACS sono state scelte 15 AUs, sufficienti per la rappresentazione delle 6 emozioni che si è scelto di rappresentare.

Le AUs rilevate sono poi processate attraverso il modulo di mappatura per predire lo stato d’animo combaciante.

La prefasi al modello sviluppato in questo paper è che molti dei metodi utilizzati per la rilevazione delle emozioni si basano su input delle AU che sono inaffidabili in quanto ci sono degli errori non evitabili nella localizzazione delle facce all’interno delle immagini.

È per questo che il loro metodo si concentra molto sul trovare la correlazione fra le singole AUs e le emozioni che sulla diretta predizione delle emozioni.

La relazione tra le Action Units e le sei emozioni viene ottenuta attraverso un'analisi statistica del dataset di riferimento (CK+), etichettato sia per le emozioni che per le AU. Le relazioni vengono ottenute sotto forma di una matrice relazionale derivata utilizzando un concetto chiamato discriminative power [17]. Il discriminative power è definito come

Immagine che contiene diagramma, testo

Descrizione generata automaticamenteImmagine che contiene diagramma, testo

Descrizione generata automaticamente

Immagine che contiene diagramma, testo

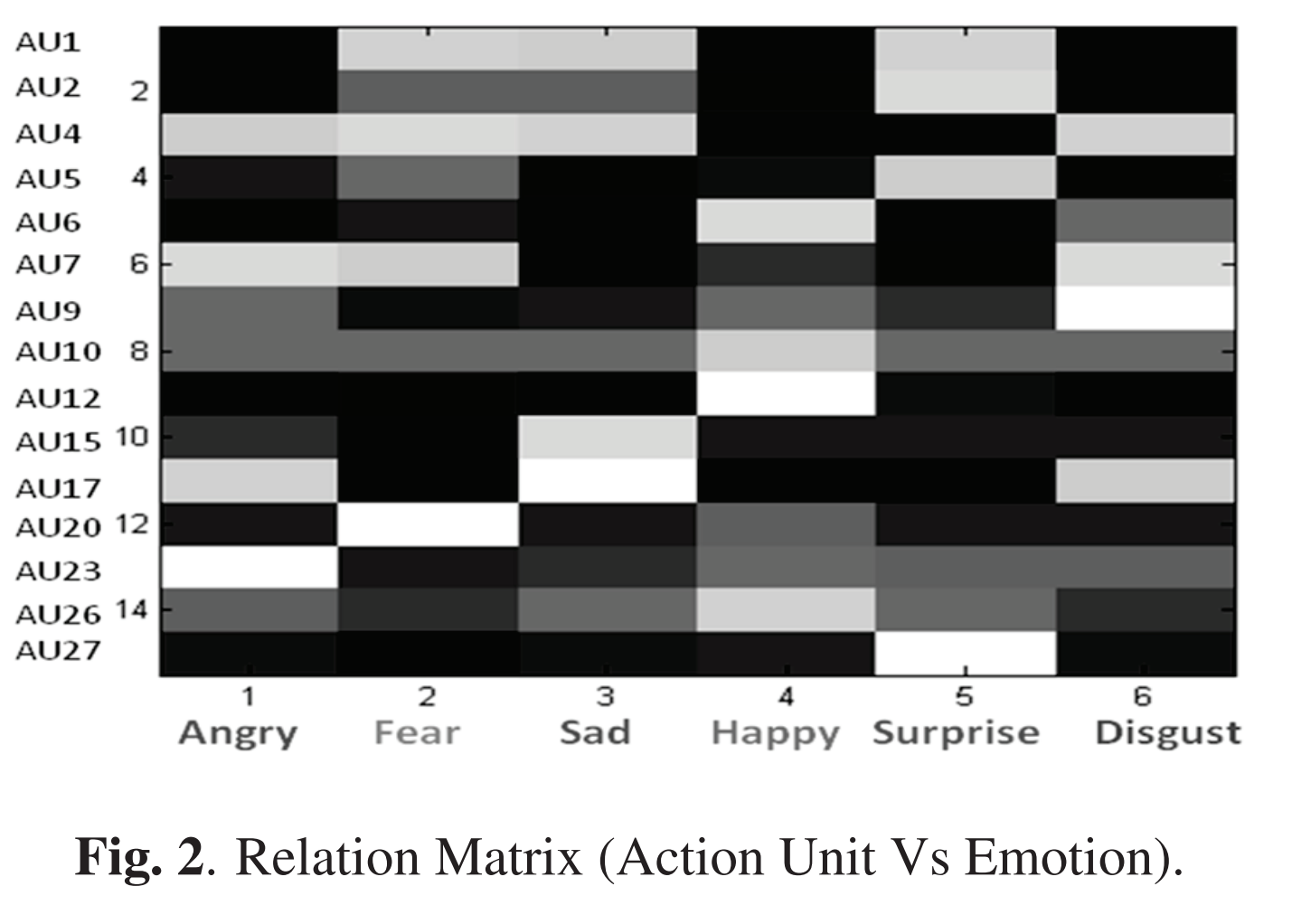
Descrizione generata automaticamentedove P(Yj |Xi) è la probabilità dell'azione Yj, dopo che Xi è avvenuta, e P(Yj | ) è la probabilità dell'azione Yj, dato che l'emozione non è avvenuta. La dimensione di H rappresenta il potere discriminante di un'AU rispetto ad un'emozione.

La matrice di relazione viene ottenuta normalizzando H su tutte le AU per ciascuna delle emozioni. Pertanto, si ottengono pesi di associazione non lineari per ciascuna delle AU, in base alla loro rilevanza per le emozioni calcolate utilizzando l'equazione sopra.

La figura sotto mostra la matrice di relazione calcolata per le sei emozioni. Qui, i valori positivi, rappresentati in bianco, indicano l'alta probabilità per un'AU di appartenere ad un'emozione; mentre i valori negativi, rappresentati in nero, indicano l'alta probabilità per un'AU di non essere associata ad un'emozione.

Ad esempio, l'emozione happy è associata positivamente alle AU6, AU7, AU12, AU26 e negativamente alle AU1, AU2, AU5, AU9.

I risultati sperimentali presenti dallo studio dimostrano che la matrice di relazione derivata è efficiente nell'identificazione delle azioni facciali altamente rilevanti e i loro pesi di associazione per le varie emozioni, e può quindi predire correttamente le emozioni.



Per ciascuna emozione, sono state selezionate le prime N entrate da AU altamente discriminanti e queste sono state salvate come stringhe di template per i futuri abbinamenti delle AU.

La lunghezza di stringa di template utilizzata è di 5 per gli esperimenti condotti dai ricercatori dello studio.

Immagine che contiene tavolo

Descrizione generata automaticamente

Date le stringhe con le Action Units per le immagini analizzate, queste vengono confrontate con le stringhe di template per trovare l'emozione corrispondente.

Si utilizza la sotto sequenza comune più lunga (LCS) [18] per misurare la similarità tra le stringhe. LCS è un metodo per trovare la sotto sequenza più lunga fra tutte le sequenze date.

In questo caso, una sotto sequenza è definita come una sequenza in cui le unità appaiono nello stesso ordine una rispetto all’altra, ma non necessariamente in modo contiguo.

Ad esempio, nella stringa “ACTTGCG”, “ACT”, “ATTC” e “ACTTGC” sono tutte sotto sequenze.

LCS permette solo "inserzioni" ed "eliminazioni", ma non "sostituzioni" delle singole AUs tra le stringhe. Questa caratteristica di LCS è risultata molto adatta per la predizione delle emozioni da stringhe di Action Units.

Ad esempio, data un'immagine di test, possono essere prodotte le combinazioni di AUs riportate nell’immagine sotto.

In questo caso, la proprietà di "eliminazione" diventa importante per mappare le diverse stringhe di AU come {AU12} o {AU6, AU12} o {AU6, AU12, AU26} all'emozione happy, poiché tutte indicano la felicità anche se alcune delle AU mancano in alcune delle stringhe.

Allo stesso modo, l'"inserzione" gioca un ruolo nell’eliminazione delle AU rilevate erroneamente come {AU1, AU4} ma che comunque mappano l'emozione nella figura all’emozione happy.

Inoltre, permettere le "sostituzioni" aiuta a correggere gli errori di rilevazione che coinvolgono AU visivamente simili come AU12, AU20, AU23, e a predire l'emozione con un alto valore di confidenza. Viene evitato anche il mappamento errato di molte emozioni quando la stringa di test ha errori considerevoli.

Immagine che contiene testo

Descrizione generata automaticamente

Ad esempio, in sostituzioni non condizionali, una stringa di test come

{AU4, AU6, AU12, AU17, AU23}, può essere mappata a

{AU1, AU4, AU10, AU15, AU17} (tristezza) e

{AU4, AU6, AU7, AU9, AU17} (disgusto)

con un costo di sostituzione "3".

I risultati e l’accuratezza ottenuta dal metodo proposto dallo studio sui vari dataset utilizzati per valutarlo sono presenti in questa tabella:  
Immagine che contiene tavolo

Descrizione generata automaticamente

Ho analizzato anche altri papers ma nessuno di questi mi è sembrato particolarmente rilevante in quanto nessuno di quelli da me ritrovati proponeva metodologie che aggiungevano informazioni e metodi utili da aggiungere ai lavori già ritrovati e/o i risultati di accuratezza e precisione sono ottenuti da questi sono più bassi rispetto agli studi già analizzati e/o i risulati ottenuti sono di poco migliori e necessitano di più risorse (tempo, quantità di dati e potere computazionale).

Ad esempio, in [14] vengono confrontate le due principali tecniche di previsione delle emozioni facciali, ovvero le tecniche di apprendimento automatico (ML) e di deep learning (DL), in termini di accuratezza e livello di accountability. Immagine che contiene diagramma

Descrizione generata automaticamente

Analizzando le due tecniche, i ricercatori hanno dimostrato che i modelli ML possono ottenere risultati comparabili rispetto ai modelli DL.

Nonostante recentemente la maggior parte delle ricerche si siano concentrate sulla costruzione di tecniche di DL per la predizione delle emozioni, si è scoperto che i modelli DL raggiungono prestazioni migliori a fronte di un costo elevato in termini di potenza di calcolo, di set di dati più grandi e di tempi di elaborazione più lunghi.

Nei loro esperimenti, sono stati ottenuti livelli di accuratezza del 86,66% e dell'80% rispettivamente per gli approcci DL e ML in termini di previsione delle emozioni, ma il DL necessita del supporto di una GPU ad alte prestazioni e le tecniche ML possono funzionare più velocemente anche senza il supporto di una GPU.

Il beneficio più importante nell'utilizzo delle tecniche ML con le Action Units è la capacità di questi metodi di poter giustificare l'emozione prevista in termini di contributo di ciascuna AU.

[12] Facial expression recognition from image sequences using twofold random forest classifier

[13] A method to infer emotions from facial action units

[14] Facial emotion prediction through action units and deep learning

[15] M. Liu, S. Shan, R. Wang, X. Chen, Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1749–1756.

[16] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, , and J. Movellan, “Recognizing facial expression: Machine learning and application to spontaneous behavior,” in Proc. of IEEE Conf. on Compter Vision and Pat. Recog. (CVPR), pp. 568–573, 2005.

[17] R. el Kaliouby, “Mind-reading machines: the automated inference of complex mental states from video,” Ph.D. Thesis, University of Cambridge, 2005.

[18] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein, “Introduction to algorithms,” 3rd ed., MIT Press, 2009.