VIDEO SUMMARIZATION BASATO SU ALGORITMI DI SENSOR FUSION

INDICE

ELENCO DELLE FIGURE

ELENCO DELLE TABELLE

1. INTRODUZIONE

2. MATERIALI E METODI

In questa sezione verranno presentati i materiali e i metodi utilizzati in questo lavoro. Come primo elemento verrà introdotto il dataset utilizzato. Successivamente, verranno presentati i metodi di preprocessing dei segnali fisiologici considerati e come si è ovviato al problema della loro asincronicità durante l'estrazione delle feature. Verrà illustrato il processo di selezione delle feature estratte, per conservare quelle più efficaci alla classificazione delle emozioni. In ultimo, verrà presentato il modello delle mappe di salienza fisiologiche e il loro impiego nel processo di video summarization.

2.1. **Dataset.** Lo studio utilizza i dati contenuti in MAHNOB-HCI [SLPP11]: un database, disponibile online, che fornisce campionamenti di segnali fisiologici, e valori di self-report, in risposta alla sollecitazione emotiva di un soggetto.

I segnali fisiologici raccolti includono 6 registrazioni video del soggetto, riprese da angolature diverse, segnali audio, dati di eye-tracking e segnali fisiologici del sistema centrale e periferico, tra cui: 32 canali di EEG, 3 canali ECG, temperatura corporea e conduttanza cutanea.

Alla raccolta dei dati hanno partecipato 16 femmine e 11 maschi, con differenti backgroud culturali.

L'intero database contiene dati relativi a due diversi setup sperimentali. La sessione sperimentale di un soggetto è stata successivamente suddivisa, catalogata e identificata univocamente in un file XML di metadati. Il file descrive il contenuto della singola sessione, indicandone i riferimenti a: soggetto partecipante, tipo di esperimento e clip visualizzata.

Di seguito sono descritti brevemente i due setup sperimentali:

(1) Emotion recognition: in cui ai partecipanti è stato chiesto di guardare 20 clip video, estratte da film e show televisivi, con lo scopo di indurre una risposta emotiva. Le clip sono scelte a caso da un insieme più ampio e prima di ogni clip, è stato mostrata una piccola clip neutrale, al fine di ridurre il bias dovuto allo stato emotivo del soggetto. Al termine di ogni clip, i soggetti hanno compilato, utilizzando i valori da 1 a 9 di un tastierino numerico, il questionario di self-reporting per annotare la propria risposta emotiva in termini di:
   - *feltEmo*, label dell'emozione provata, la codifica numerica è riportata in tabella 1
   - *feltArsl*, Arousal percepito, 1 per nessuna attivazione, 9 per massima attivazione
   - *feltVlnc*, Valence percepito, 1 per molto negativo, 9 per molto positivo, 5 per neutrale
   - *feltCtrl*, Control percepito, 1 per senza controllo, 9 per pieno controllo
   - *feltPred*, Predictability percepita, 1 per imprevedibile, 9 per completamente prevedibile

   questi valori, insieme ad altri, sono riportati nel file di metadati associato ad ogni sessione.

(2) Implicit tagging: che prevede di mostrare una sequenza di clip video o immagini, prima senza tag e successivamente con un tag che descriva, talvolta in modo corretto talvolta in modo errato, l'emozione che questa rappresenta. Ai partecipanti è stato chiesto di annotare se fossero in accordo o in disaccordo con la descrizione dell'emozione indicata.

In questo lavoro sono state considerate tutte le sessioni relative al primo tipo di esperimento per un totale di 401 sessioni. La tabella 2 riassume il numero di sessioni per clip video, con relativo valore medio e deviazione standard delle risposte dei partecipanti.

Codifica Numerica	Nome Emozione
0	Neutrale
1	Rabbia
2	Disgusto
3	Paura
4	Gioa, felicità
5	Tristezza
6	Sorpresa
11	Divertimento
12	Ansia

Tabella 1. Codifica numerica delle emozioni selezionabili nei questionari di self-report

Nome clip	sessioni	feltEmo	feltArsl	feltVlnc	feltCtrl	feltPred
107.avi	18	5.33 ± 4.06	6.22 ± 2.24	3.22 ± 1.59	3.28 ± 1.71	3.39 ± 1.04
111.avi	24	4.62 ± 1.31	6.00 ± 1.47	2.12 ± 0.99	3.08 ± 1.91	5.25 ± 2.54
138.avi	19	3.95 ± 2.09	4.21 ± 1.84	3.05 ± 1.31	3.53 ± 1.90	6.79 ± 1.90
146.avi	19	3.95 ± 2.09	3.05 ± 1.72	3.26 ± 1.24	5.16 ± 1.95	6.37 ± 2.06
30.avi	25	5.56 ± 3.84	6.36 ± 1.70	2.84 ± 1.70	3.68 ± 2.50	4.12 ± 2.35
52.avi	19	6.26 ± 4.85	3.68 ± 1.92	6.47 ± 1.93	5.89 ± 1.76	5.26 ± 1.76
53.avi	24	7.71 ± 4.55	5.62 ± 1.47	4.38 ± 2.04	3.83 ± 1.88	6.88 ± 1.51
55.avi	18	3.56 ± 3.48	5.72 ± 2.08	1.72 ± 0.89	3.00 ± 2.35	5.89 ± 2.30
58.avi	17	5.35 ± 3.98	3.76 ± 1.95	6.88 ± 1.41	6.24 ± 2.33	6.06 ± 1.56
69.avi	26	3.19 ± 3.25	5.73 ± 2.01	2.23 ± 1.07	3.65 ± 2.38	5.31 ± 2.09
73.avi	19	5.68 ± 5.09	4.95 ± 2.07	3.58 ± 1.68	3.74 ± 2.10	5.84 ± 2.01
79.avi	18	3.33 ± 1.53	3.56 ± 2.18	7.00 ± 1.91	7.50 ± 1.72	6.28 ± 1.78
80.avi	19	4.26 ± 1.94	5.63 ± 1.71	7.79 ± 0.85	6.11 ± 2.26	4.95 ± 2.37
90.avi	26	7.62 ± 4.28	4.12 ± 1.97	6.69 ± 1.41	6.27 ± 1.91	5.27 ± 2.57
cats_f.avi	19	9.26 ± 3.02	5.05 ± 1.78	6.89 ± 1.70	6.58 ± 2.48	4.63 ± 2.71
dallas_f.avi	18	0.56 ± 1.65	1.94 ± 1.47	5.00 ± 1.08	5.89 ± 2.37	7.11 ± 2.08
detroit_f.	19	0.58 ± 2.52	1.42 ± 0.69	4.68 ± 1.29	6.42 ± 2.65	6.79 ± 2.12
earworm_f.	18	4.17 ± 4.42	4.72 ± 2.44	3.61 ± 1.85	5.67 ± 2.40	4.72 ± 2.76
funny_f.avi	19	9.05 ± 3.57	5.32 ± 2.11	6.26 ± 1.56	6.42 ± 1.95	4.21 ± 2.78
newyork_f.	17	0.76 ± 1.56	1.76 ± 1.20	4.59 ± 1.54	6.06 ± 2.86	6.94 ± 2.01

Tabella 2. Numero di sessioni per clip video, con relativo valore medio e deviazione standard di emozione, valenza, arousal, control e predictability

Nella fase di selezione delle feature necessarie alla costruzione delle mappe di salienza è stato utilizzato l'intero insieme di sessioni. I test di video summarization invece sono stati applicati ai media riportati in tabella 3, in quanto rappresentativi delle differenti classi di emozioni.

Le clip contenute nel database sono lunghe tra i 34.9 e i 117s $(M = 81.4s; SD = 22.5s)$ con una risoluzione di 1280x800 pixel e frame rate non omogeneo. Per comodità di implementazione, le clip selezionate sono state scalate a una risoluzione di 320x200 pixel e ricodificate a 24pfs.

Gaze. i dati di eye-tracking (Gaze) ....

Nome clip	Emozione	Classe originale	Classe media self-report
30.avi	Paura	3	4.62 ± 1.31
53.avi	Divertimento	11	7.71 ± 4.55
69.avi	Disgusto	2	3.19 ± 3.25
90.avi	Gioia	4	7.62 ± 4.28
111.avi	Tristezza	5	4.62 ± 1.31

Tabella 3. Media clip utilizzate durante i test di video summarization. Viene indicata anche la relativa classe di emozione a cui appartiene la clip e il valore medio riportato dai partecipanti

2.2. **Segnali fisiologici.** Tra i diversi segnali fisiologici presenti nel dataset, in questo lavoro vengono utilizzati: ECG, EDA, temperatura corporea (SKT) e ampiezza della respirazione (Resp). Ogni segnale fisiologico è stato campionato a 1024Hz e successivamente sottocampionato a 256Hz per ridurre i tempi computazionali. Per il denoising e il preprocessing di ECG, SKT e Resp è stato utilizzato il toolbox per l'elaborazione di segnali biologici *BioSPPy*[CAL+18].

Al termine di tutte le elaborazioni, ai segnali risultanti è stato applicato un algoritmo di resampling per ottenere serie temporali a 24 sample/s.

Le caratteristiche principali e i trattamenti specifici dei segnali utilizzati sono di seguito descritti:

ECG. è stato registrato usando tre sensori posizionati sul petto del partecipante. Il segnale è misurato in microvolt (μV). Il primo step di elaborazione è l'applicazione un filtro bassa-banda con banda passante 3-45Hz per sopprimere rumore e interferenze. Il secondo step prevede la ricerca degli *R-peak* all'interno dei complessi QRS. Il tempo che intercorre tra due R-peak è definito come intervallo RR. Il reciproco di RR, moltiplicato per 60, fornisce una misura dei battiti per minuto (bpm) e quindi del heart rate (HR), secondo la relazione:

$$HR(bpm) = \frac{60}{RR(s)} \tag{1}$$

L'HR tipico di una persona a riposo può variare tra i 60 e i 100bpm. Il segnale così ottenuto è stato poi utilizzato nelle fasi successive del lavoro

EDA.

Resp.

SKT.

2.3. **Estrazione delle feature.**

3. RISULATI

4. CONSLUSIONI

RIFERIMENTI BIBLIOGRAFICI

[CAL+18] Carlos Carreiras, Ana Priscila Alves, André Lourenço, Filipe Canento, Hugo Silva, Ana Fred, et al. Biosppy: Biosignal processing in python, 2018.

[SLPP11] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2011.