Emotion Prediction - Landmarks Analysis

Emanuele Fittipaldi, Paolo Plomitallo
University of Salerno,Via Giovanni Paolo II, 132, Fisciano, Salerno, Italy, 84084
e.fittipaldi@studenti.unisa.it, p.plomitallo@studenti.unisa.it

Abstract

Nel panorama della computer vision, il problema dell'emotion prediction è un tema tutt'ora caldo, in quanto pone l'enfasi su ciò che accade dal verificarsi di una micro-espressione fino alla manifestazione completa di una macro-espressione. Questo studio si concentra nell'analizzare il movimento dei landmark facciali ricavati tramite la libreria *Mediapipe*, al fine di comprendere la sequenza di frame all'interno di una sequenza video, in cui una micro-espressione ha luogo. Le variazioni di posizione dei landmark sono analizzate attraverso test statistici come il p-value. I landmark significativi vengono contati in modo da capire in quale punto della sequenza video c'è un movimento. Per la maggior parte degli individui e delle espressioni, i risultati ci indicano che nei frame iniziali c'è la più alta probabilità di trovare una micro-espressione. Tramite analisi inter-classe ed intra-classe tra i diversi soggetti e le diverse espressioni e comparando la similarità tra i vettori delle distanze dei landmark e la similarità tra i vettori dei landmark significativi, le emozioni associate alla sfera negativa sono quelle più simili in termini di movimento dei landmark, ottenendo sempre delle similarità molto vicine.

1 Introduzione

Il panorama corrente è pieno di tecniche che sfruttano il Machine Learning e il Deep Learning per risolvere il problema dell'emotion detection, ma ancora molto poco è stato fatto nella direzione dell'emotion prediction. L'emotion prediction pone il focus su ciò che accade tra un viso a riposo fino al palesarsi di una espressione ben distinta, cercando di fornire una percentuale di confidenza su quale espressione sta configurandosi. Il nostro studio è stato fatto nella direzione dell'analisi dei landmark, 468 punti estratti tramite la libreria *Mediapipe*, al fine di fornire una insight su ciò che accade tra i diversi frame per evidenziare quali sono le peculiarità delle diverse espressioni al fine di fornire delle feature utili per la predizione di una emozione.

2 Stato dell'arte

Nel lavoro di Alugupally Et al [1] è stato affrontato il problema della classificazione di una espressione facciale sulla base dell'analisi dei soli landmark. Sono stati proposti due scenari nei quali testare l'approccio:

- Scenario in cui non si dispone di frame rappresentanti il viso in condizioni di riposo.
- Scenario in cui si dispone di frame rappresentanti il viso in condizioni neutrali.

Lo scenario in cui si dispone di frame rappresentanti il viso in condizioni neutrali è ciò che ci interessa in quanto condivide diversi punti con il nostro studio. Nel lavoro di Algupally Et al [1] sono stati considerati 18 landmark e tutte le possibili distanze diverse ricavabili a partire da questi landmark, date dal coefficiente binomiale di 18 su 3 (153 misure). Sono state utilizzate tre distanze diverse:

- distanza euclidea
- distanza di manhattan
- scostamento tra landmark

Lo scostamento di un landmark i è registrato sia rispetto l'asse orizzontale ($\Delta_{i,h}$) che rispetto l'asse verticale ($\Delta_{i,v}$) rispetto al viso in stato neutrale. Di queste 153 distanze ottenute, sono state dunque ricavate le 10 più performanti per metrica attraverso analisi stepwise discriminant. Delle differenti metriche (euclidea, manhattan e scostamento) è stato quindi testato il potere predittivo singolarmente ed in combinazione attraverso l'analisi lineare discriminante (LDA). Questa analisi permetteva di ottenere per ogni frame una percentuale indicante la probabilità di quel frame di appartenere ad una specifica espressione facciale. La classificazione quindi veniva conclusa assegnando l'espressione che ha più alta probabilità a quel frame.

3 Background

3.1 Descrizione del Dataset

Il Dataset che sarà utilizzato è il Cohn-Kanade Espression Dataset (CK+). Esso contiene: 593 video a 30 FPS con risoluzione

640x490 oppure 640x480, rappresentanti una specifica espressione facciale a partire da una espressione neutrale. 327 di questi video (circa la metà) sono etichettati con una delle otto classi di espressione.

Ogni soggetto è in una cartella la quale segue il formato "Sxxx" dove il numero identifica un soggetto. All'interno di queste cartelle troviamo le cartelle delle videosequenze seguenti il formato "xxx" il cui numero individua una emozione.

3.2 Landmarks

Per tenere traccia di come la morfologia di un viso cambia, abbiamo utilizzato i landmark forniti dalla libreria MediaPipe. Essi sono dei punti rappresentati in uno spazio tridimensionale che variano di posizioni a seconda di come un viso cambia. Nello specifico abbiamo utilizzato MediaPipe Face Mesh la quale è una soluzione che stima la posizione di 468 landmark in uno spazio 3D attraverso l'impiego del ML e senza la necessità di fornire diversi frame o diversi punti di vista. Questi landmark sono numerati da 0 a 467.

3.3 Test di significatività

Nell'analisi delle distanze dei landmark è stato impiegato il test di significatività two-tailed. Questo test si occupa di verificare se la media è significativamente maggiore o significativamente minore di x. La media è considerata significativamente differente da x se la statistica di test ricade oltre il 2.5% o prima del 2.5% della sua distribuzione di probabilità, risultante in un p-value inferiore a 0.05. Questo valore, rappresenta la soglia Alpha.

4 Sistema proposto

4.1 Estrazione dei landmark

Avendo già a disposizione i diversi frame componenti ciascuna videosequenza, il primo task è stato quello dell'estrazione dei 468 landmark facciali di ogni frame. Per ogni frame quindi è stato creato un csv contenente 468 righe e 3 colonne (x,y,z).

4.2 Distanze

Lo step successivo all'estrazione della posizione dei landmark per ogni frame è stato quello di ricavare le distanze dei landmark relativamente a loro stessi. A seconda dalla posizione di partenza considerata per il calcolo della distanza e dalla misura di distanza impiegata abbiamo ricavato:

- Distanze Globali: le distanze tra il landmark al frame *n* e il landmark al primo frame.
- Distanze Locali: le distanze tra il landmark al frame *n* e il landmark al frame *n-1*.

Come misura di distanza sono state impiegate la distanza euclidea e la distanza di manhattan.

4.3 Aggiunta di label

Il nostro Dataset di partenza aveva delle label mancanti, quali:

- Sesso - label mancante per tutti i soggetti
- Emozione - label mancante per 266 soggetti

Queste label sono state quindi aggiunte manualmente in modo da avere un dataset completo con il quale poter lavorare.

4.4 Analisi delle micro-espressioni

Per l'analisi delle micro-espressioni sono state prese in considerazione le distanze locali euclidee, per studiare l'andamento della posizione dei landmark rispetto al frame precedente, per ogni frame nella sequenza video.
Le micro-espressioni sono caratterizzate da una bassa intensità e breve durata. Per intercettare queste piccole contrazioni muscolari siamo andati alla ricerca dei frame in cui il maggior numero di landmark ha registrato una variazione di posizione significativa, adoperando il test di significatività two-tailed. La significatività è stata testata andando a considerare le distanze locali dei landmark come due popolazioni separate. Il punto tramite il quale effettuiamo questa separazione è stato progressivamente spostato in modo da verificare in quale sequenza di frame il landmark risultasse essere significativo. L'ipotesi nulla e l'ipotesi alternativa sono state formulate in questo modo:

- H_0: Non ci sono differenze significative tra le due popolazioni.
- H_1: Esiste una differenza significativa tra le popolazioni.

Rigettando l'ipotesi nulla si dimostra che il landmark ha subito una variazione di posizione significativa, quindi siamo in presenza di una contrazione muscolare che ci indica la presenza di una micro-espressione. Tramite questo test siamo stati in grado di esprimere la sequenza video in termini di landmark significativi per frame. Dato che le micro-espressioni sono caratterizzate da una bassa intensità e breve durata, sono stati calcolati i delta tra i landmark significativi per frame, tra tutte le coppie adiacenti di frame per i primi n=10 frame se n_frame maggiore di 10, altrimenti sono stati considerati tutti i frame nella videosequenza . I frame aventi il delta più alto sono stati memorizzati come i frame in cui la micro-espressione ha avuto luogo.

4.5 Analisi delle macro-espressioni

Per l'analisi delle macro-espressioni sono state prese in considerazioni le distanze globali, ovvero l'andamento della posizione dei landmark rispetto al frame del viso a riposo, per ogni frame nella sequenza video.
L'obiettivo di questa analisi è stato quello di estrarre per ogni soggetto e per ogni emozione, una lista di landmark che hanno dato prova di essere significativi sulla base di una treshold calcolata ad-hoc per ciascun individuo.

Per il calcolo di questa threshold sono stati tentati due approcci diversi:

- Un primo approccio è stato quello di considerare la distanza globale massima, ovvero lo spostamento massimo del landmark rispetto al primo frame. L'idea è stata quella di ricavare un upper-bound tra gli spostamenti in modo da poter prelevare il 30% dei landmark associati alle distanze subito sotto questo limite superiore. L'ipotesi dietro questo approccio, è che queste distanze potessero essere associate a dei landmark significativi considerando la loro variazione di posizione che li ha portati ad essere tra il 30% dei landmark che si sono spostati di più. Questo approccio, è stato abbandonato in quanto i landmark significativi, risultavano essere quasi sempre gli stessi, anche considerando emozioni diverse.

- Il secondo approccio consiste nel calcolare una threshold rappresentata come la media tra le differenze delle distanze dei landmark dell'ultimo frame rispetto al primo frame. La significatività dei landmark è stata dunque ricavata considerando soltanto i landmark la cui distanza globale nell'ultimo frame, superasse questa threshold.

Una volta ottenute queste liste di landmark significativi per ogni soggetto/espressione, sono state confrontate tramite il coseno di similitudine per verificare quanto simili siano le espressioni intra-classe ed inter-classe. Per ottenere una similitudine più alta sono stati considerati due approcci:

- Primo approccio: alla lista dei landmark significativi, viene aggiunta una ulteriore informazione riguardo la direzione dello spostamento del landmark rispetto l'asse X e l'asse Y. Queste direzioni sono state calcolate confrontando le posizioni dei landmark nell'ultimo rispetto al primo frame. Se la differenza di queste posizioni è positiva, allora lo spostamento è avvenuto seguendo la stessa direzione. Per indicare ciò è stato assegnato il valore +1 al landmark. Se la differenza di queste posizioni è negativa, allora lo spostamento è avvenuto nella direzione opposta. É stato assegnato il valore -1.

- Secondo approccio: piuttosto che considerare la lista dei landmark significativi come lista di partenza, viene considerata la lista delle distanze significative. Queste distanze vengono ricavate nello stesso modo in cui sono stati ricavati i landmark significativi, ovvero adoperando la threshold calcolata con l'approccio 2. Questo approccio ha dato dei risultati migliori.

4.6 Predizione delle emozioni

Per testare la capacità di discriminazione di una emozione rispetto ad un'altra, basandosi sulle distanze global, è stato valutato un approccio di classificazione. L'approccio è articolato nei seguenti punti:

1. Per ogni emozione è stato creato un training set e un test set (80%-20%) dei soggetti.

2. Sui soggetti del training set è stata realizzata una lista, per ogni emozione, contenente la media delle distanze globali per ogni landmark. Denominate 'L1',...,L7'.

3. Con il training set è stata realizzata una tabella A di similarità tra le distanze dell'ultimo frame, calcolate combinando le diverse emozioni.

4. Su un soggetto del test set, è stata predetta l'emozione calcolando le similarità (nominate: 'S1',...,'S7') tra le distanze globali del soggetto e le liste 'L1',...,'L7', pesando i risultati con i valori contenuti nella tabella A.

Un primo passo per la classificazione è considerare il rapporto più grande tra le similarità 'S1',...,'S7', assegnando ad esso un premio e agli altri rapporti una penalità.
Per pesare queste similarità con i valori contenuti nella tabella A, sono stati impiegati tre diversi metodi:

1. Moltiplicare ogni valore 'S1',...,'S7' con i valori sulla diagonale.
   $S_i * A(i,i)$

2. Calcolare la media delle moltiplicazioni tra i valori 'S1',...,'S7' con la colonna di ogni emozione.
   $\frac{1}{7} \sum S_i * A(j,i)$

3. Sottrarre dalla moltiplicazione tra i valori 'S1',...,'S7' e la diagonale, i restanti valori delle colonne corrispondenti.
   $(S_i * A(i,i)) - \sum A(j,i)$

Questi approcci sono stati utilizzati in combinazione per fornire un'accuratezza migliore nella classificazione. Ogni approccio fornisce in output l'emozione che ha la percentuale di similarità più alta rispetto alle altre. Se la maggioranza degli approcci concordano sulla stessa emozione, il soggetto viene classificato sotto quella emozione. Invece, se non c'è una maggioranza ne viene creata una nel seguente modo:

- Per ogni emozione data in output vengono presi i valori delle similarità calcolate dai diversi approcci.

- Vengono confrontati in coppia andando a calcolare la somma delle differenze tra i valori calcolati.

- L'emozione che ha le differenze maggiori rispetto alle altre, viene assegnata al soggetto.

Oltre all'emozione più probabile sono stati applicati gli approcci sulla seconda emozione più probabile. Come risultato della classificazione vengono forniti due emozioni che corrispondono a quelle più probabili, aumentando anche l'accuratezza della predizione.

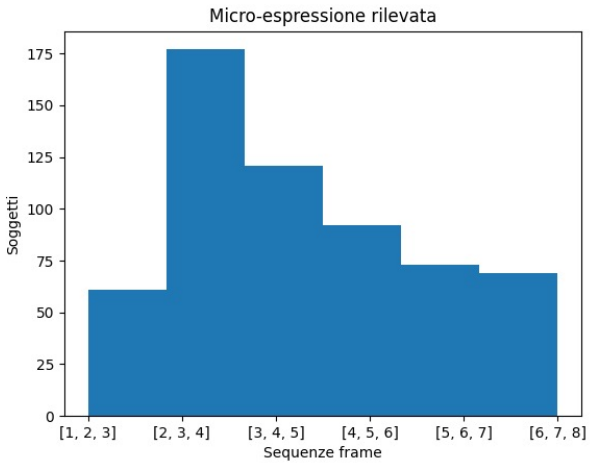


Figure 1: sequenza di frame in cui è stata rilevata una micro-espressione

5 Risultati sperimentali

5.1 Risultati analisi micro-espressioni

Come prima statistica, abbiamo analizzato su tutte le emozioni, in quale intervallo di frame una micro-espressione viene registrata. Come si può vedere dall'Istogramma 1, su 173 soggetti aventi mediamente a disposizione circa 20 frame, la micro-espressione è stata rilevata intorno ai frame (2,3,4) mentre su circa 125 soggetti la micro-espressione è stata rilevata intorno ai frame (3,4,5). Come ci si potrebbe aspettare l'intervallo di frame per i quali sono state registrate meno micro-espressioni è l'intervallo (1,2,3) in quanto, partendo da una condizione di riposo, è altamente improbabile che ci sia un movimento importante da rilevare.

Nella Tabella 1 sono riportate le sequenze di frame più frequenti in cui è stata registrata una micro-espressione, suddivise per emozione. Come si può notare, in tutte le emozioni, i frame (2,3,4) sono i frame in cui quasi sicuramente si verifica una micro-espressione. É interessante notare che soltanto per l'emozione contempt (disprezzo), i frame in cui viene rilevata una micro-espressione sono i frame (3,4,5)

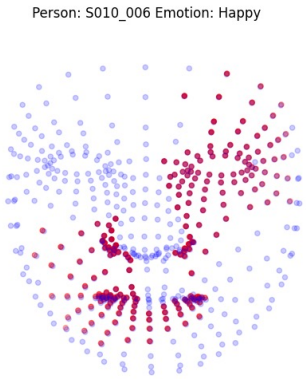
Nella Figura 2c sono evidenziati i landmark che hanno rilevato una micro-espressione nei frame (2,3,4). I punti blu rappresentano i landmark nella condizione di viso a riposo, mentre i punti rossi rappresentano i landmark che hanno subito una variazione di posizione importante. É facile osservare a partire dai punti rossi quale espressione sta configurandosi (happy).

5.2 Risultati analisi macro-espressioni

Com'è possibile osservare nella Figura 5 emozioni diverse possono risultare molto simili. Questo risultato è emerso anche dall'analisi effettuata sulle distanze dei landmark. Le Tabelle

(a)

(b)



(c)

Figure 2: Sequenze dei frame (2,3,4) in cui è stata rilevata la micro-espressione.

Emozione	Sequenza frame micro-espressione
anger	[2,3,4]
contempt	**[3,4,5]**
disgust	[2,3,4]
happy	[2,3,4]
fear	[2,3,4]
sadness	[2,3,4]
surprise	[2,3,4]

Table 1: Numero occorrenze sequenza frame per ogni emozione.

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	**0.7843**	0.7099	0.7816	0.6725	0.6413	0.7405	0.6230
2	0.7099	**0.6708**	0.7001	0.6395	0.5695	0.6953	0.5672
3	0.7816	0.7001	**0.7794**	0.6660	0.6613	0.7289	0.6296
4	0.6725	0.6395	0.6660	**0.6083**	0.5794	0.6561	0.5636
5	0.6413	0.5695	0.6613	0.5794	**0.7674**	0.5769	0.6429
6	0.7405	0.6953	0.7289	0.6561	0.5769	**0.7198**	0.5804
7	0.6230	0.5672	0.6296	0.5636	0.6429	0.5804	**0.5829**

Figure 3: Tabella similarità landmark.

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	**0.8698**	0.8198	0.8673	0.7931	0.7911	0.8466	0.7685
2	0.8198	**0.7846**	0.8136	0.7588	0.7364	0.8083	0.7218
3	0.8673	0.8136	**0.8652**	0.7890	0.7972	0.8400	0.7701
4	0.7931	0.7588	0.7890	**0.7355**	0.7325	0.7791	0.7137
5	0.7911	0.7364	0.7972	0.7325	**0.8195**	0.7560	0.7564
6	0.8466	0.8083	0.8400	0.7791	0.7560	**0.8326**	0.7424
7	0.7685	0.7218	0.7701	0.7137	0.7564	0.7424	**0.7222**

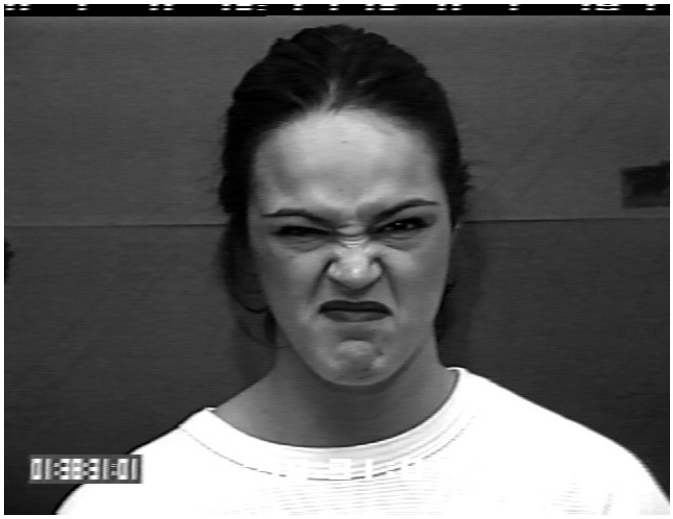
Figure 4: Tabella similarità distanze.

3, 4 rappresentano i rapporti di similarità tra tutte le possibili combinazioni di emozioni, che vengono espresse tramite indici in cui: 1 - 'Anger', 2 - 'Contempt', 3 - 'Disgust', 4 - 'Fear', 5 - 'Happy', 6 - 'Sadness', 7 - 'Surprise'.
La Tabella 3 visualizza i risultati dei rapporti di similarità sulla base dei landmark significativi (Approccio 1). La Tabella 4, invece, le similarità sulla base delle distanze significative (Approccio 2).
L'ipotesi iniziale era che il rapporto di similarità tra le stesse emozioni fosse più alto rispetto ai rapporti con le altre emozioni, ma indipendetemente dall'approccio attuato, l'ipotesi è stata smentita. È possibile notare che le emozioni **1-3-6**, quindi Rabbia, Disgusto e Tristezza sono spesso sovrapponibili.
Ci sono tre possibili spiegazioni a questo fenomeno:

1. Il dataset CK+ è stato costruito chiedendo ai soggetti di simulare una emozione.

2. Espressioni facciali simili tra soggetti diversi ed emozioni diverse.

3. L'approccio basato sull'analisi delle distanze/landmark non è sufficiente.



(a)

(b)

Figure 5: Rabbia (a) vs Disgusto (b)

Analizzando i diversi soggetti del dataset sono state rilevate delle discrepanze rispetto all'etichetta dell'emozione associata. Questo era largamente dovuto al fatto che l'espressione è stata mimata dal soggetto, quindi non era molto rappresentativa per l'emozione ad esso assegnata. In contesti simulati non è sempre facile per un soggetto mimare l'espressione in modo realistico. Questo ha portato il dataset a non fornire dei campioni rappresentativi per le varie emozioni e dunque ha fatto sì che delle sovrapposizioni avessero luogo.
Un altro problema riscontrato è che soggetti diversi possono esprimere emozioni diverse con espressioni simili. Facendo riferimento alla Figura 5 possiamo notare come due soggetti completamente diversi hanno manifestato la **Rabbia** (5a) e il **Disgusto** (5b) con una espressione molto simile. Questo fa si che un approccio basato sull'analisi dello scostamento dei landmark non è sensibile a sufficienza per poter percepire la differenza semantica tra le due emozioni.

Emozione	Accuratezza
anger	63,84%
contempt	40,00%
disgust	73,15%
happy	96,52%
fear	44,60%
sadness	39,40%
surprise	43,74%

Table 2: Percentuale accuratezza predizione per ogni emozione.

5.3 Risultati predizione delle emozioni

Data la natura sovrapponibile di alcune emozioni, nella fase di predizione, sono state prese in considerazione due possibili esiti, corrispondenti alle emozioni più probabili. I tre approcci sono stati combinati considerando a partire dai singoli output di similarità una maggioranza. Se 2 su 3 approcci concordano sulla stessa emozione allora il soggetto viene classificato sotto quella emozione. Sono inoltre stati gestiti i casi in cui maggioranze non fossero presenti attraverso l'utilizzo di pesi estratti dalla matrice delle similarità 4. Questi passi sono stati eseguiti su dieci iterazioni. Per ogni iterazione il dataset è stato mischiato e sono stati estratti l'80% degli individui formando il training set, i restanti inseriti nel test set. Ogni iterazione ha prodotto una percentuale di accuratezza. L'accuratezza è stata ricava come una media di queste e corrisponde al 60,13%. L'accuratezza della predizione di ogni emozione sono mostrate nella Tabella 2.

6 Conclusioni

Il nostro studio ha dimostrato che analizzando lo scostamento dei landmark frame per frame è possibile individuare, con un certo grado di approssimazione, la sequenza di frame nella quale si verifica una micro-espressione. Questo approccio ha mostrato dei limiti. Un primo limite è la impossibilità di esprimere un grado di confidenza per il quale possiamo affermare che siamo in presenza di una micro-espressione. Questo limite deriva dal fatto che sono stati considerati gli spostamenti dei landmark rispetto a loro stessi, sia frame per frame che rispetto alle condizioni a riposo. Un evoluzione di questo approccio potrebbe essere quello di considerare gli scostamenti tra tutte le possibili coppie di landmark, al fine da catturare possibili relazioni esistenti tra i diversi punti del volto. Questa tecnica potrebbe essere utile nel caso in cui si dispone di pochi landmark. Le tecniche più performanti per la classificazione e predizione di una emozione, rimangono comunque quelle costituenti l'attuale stato dell'arte. Un possibile sviluppo futuro dell'approccio qui presentato è quello di adoperare il FACS [2]. Il FACS permette di ottenere una rappresentazione delle condizioni di un volto tramite un insieme di Action Units. Le Action Units rappresentano logicamente i muscoli facciali, fornendo anche l'intensità della contrazione per muscolo. Data una sequenza video è facile intuire come il FACS possa dare delle insight più approfondite in termini di contrazioni muscolari frame per frame. Sulla base di quali Action Units si attivano e sulle relative percentuali, potrebbe essere addestrato un classificatore.

References

[1] N Alugupally, A Samal, D Marx, and S Bhatia. Analysis of landmarks in recognition of face expressions. *Pattern Recognition and Image Analysis*, 21(4):681–693, 2011.

[2] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.