Valutazione delle mappe di salienza basate su gaze following per la predizione dello sguardo nei video

Stefano Giacoia

Abstract—Questo studio esplora l'efficacia delle mappe di salienza basate su algoritmi di gaze following per stimare la direzione dello sguardo dei soggetti presenti in un video. L'obiettivo è costruire una "social context-gaze-behavior value map" che riflette la salienza degli oggetti o delle persone osservate dai soggetti nel video e valutare la sua utilità nella predizione del gaze in contesti dinamici e sociali. Questa ricerca offre nuove prospettive sull'interazione umana e sul comportamento collettivo, evidenziando come l'attenzione visiva sia distribuita in gruppi e quali elementi attirino maggiormente l'interesse in scenari sociali. Utilizzando il modello Gaze360 per rilevare e analizzare la direzione dello sguardo, e DensePose per il riconoscimento delle pose umane, lo studio ha elaborato i frame video per creare mappe di salienza che riflettono le aree di maggior interesse visivo. Le mappe sono state generate computando un istogramma 2D delle fissazioni e applicando una convoluzione con un kernel gaussiano per ottenere una stima di densità liscia. Le predizioni delle mappe di salienza sono state confrontate con i dati eye-tracking raccolti per valutare la loro accuratezza nel predire lo sguardo dei soggetti.

1 Introduzione

Negli ultimi decenni, il campo dell'interazione uomo-macchina ha beneficiato notevolmente degli sviluppi nell'intelligenza artificiale, che mira a anticipare e rispondere alle esigenze umane in modo sempre più efficace. Uno degli aspetti fondamentali di questa interazione è l'utilizzo dello sguardo umano per navigare e interagire con l'ambiente circostante. Nonostante i notevoli progressi tecnologici, le tecnologie esistenti spesso non riescono a catturare accuratamente o a interpretare dinamicamente le intenzioni degli utenti in ambienti complessi o densamente sociali.

La capacità di tracciare e interpretare la direzione dello sguardo umano ha consentito significative innovazioni, migliorando l'interfaccia tra uomini e macchine e facilitando una più profonda comprensione del comportamento umano. Le tecnologie di tracking dello sguardo trovano applicazione in una varietà di ambiti, dalla realtà aumentata ai sistemi di assistenza alla guida, che migliorano la sicurezza mediante l'analisi dell'attenzione dell'utente. Tuttavia, esiste una sfida continua nel migliorare l'accuratezza e la reattività di questi sistemi, soprattutto in scenari reali e non controllati.

Questo progetto ha esplorato l'applicabilità delle mappe di salienza basate sul contesto osservato nei video, valutando come i soggetti interagiscono visivamente con l'ambiente. Utilizzando tecnologie all'avanguardia come DensePose e Detectron2 per estrarre pose e identificare oggetti nei frame video, e impiegando il modello Gaze360 per analizzare le direzioni dello sguardo, è stato possibile creare mappe di salienza che indicano le aree di maggiore interesse visivo. L'efficacia di tali mappe è stata successivamente confrontata con i risultati ottenuti da sessioni di eye-tracking, con l'obiettivo finale di verificare se queste mappe possono effettivamente predire la direzione dello sguardo.

2 Modello teorico

2.1 Detectron2

Detectron2 [1], è un framework avanzato di visione artificiale sviluppato da Facebook AI Research (FAIR). Detectron2 è stato scelto per la sua flessibilità, efficienza e l'ampio supporto a diversi modelli di deep learning focalizzati sulla rilevazione di oggetti e la segmentazione istanziale. Il modello specifico utilizzato è stato COCO-InstanceSegmentation con Mask R-CNN [2], addestrato su COCO, un dataset di riferimento contenente immagini di oggetti in contesti naturali con annotazioni dettagliate (vedi esempio in fig 1).

2.2 DensePose

DensePose [3] è un potente strumento sviluppato come parte del framework Detectron2 da Facebook AI Research (FAIR). È stato scelto per il progetto per la sua capacità unica di mappare i pixel di immagini di figure umane su una superficie corporea 3D, offrendo una rappresentazione dettagliata e precisa delle pose umane.

DensePose utilizza una rete neurale convoluzionale avanzata che trasforma le immagini di persone in un insieme di coordinate UV, corrispondenti a punti su un modello tridimensionale del corpo umano. Questo consente di comprendere non solo la posizione delle

Fig. 1: Esempio di estrazione di oggetti tramite un modello di instance segmentation di detectron2

varie parti del corpo in 2D, ma anche di inferire la disposizione tridimensionale del corpo nello spazio.

Il modello specifico adottato è densepose_rcnn_R_101_FPN_s1x che si basa sull'architettura ResNet-101 con un Feature Pyramid Network (FPN). Questo modello è noto per la sua efficacia nel rilevare dettagli a diverse scale di risoluzione, un aspetto fondamentale per garantire precisione nella mappatura delle pose. Questa precisione è essenziale per le analisi successive nel progetto.

2.3 Gaze360

Gaze360 [4], sviluppato dal MIT CSAIL, è riconosciuto per la sua capacità di fornire stime accurate della direzione dello sguardo in ambienti non controllati. Questo modello impiega un'architettura avanzata che integra una versione modificata di ResNet-18 con capsule LSTM bidirezionali, ideale per l'analisi di sequenze video. Il trattamento bidirezionale dei frame consente al modello di utilizzare informazioni sia passate che future, migliorando significativamente la precisione della predizione dello sguardo nel frame centrale di sequenze di 7 frame.

Una peculiarità di Gaze360 è l'uso delle coordinate sferiche per rappresentare la direzione dello sguardo, rendendo i risultati intuitivamente interpretabili in relazione alla posizione della telecamera. Questa caratteristica rende il modello particolarmente adatto per applicazioni in ambienti dinamici, dove le variazioni continue dello sguardo sono indicatori cruciali per l'analisi comportamentale.

3 Simulazione e esperimenti

3.1 Dataset

Il dataset FIND ("Find-Who-to-Look-at" [5]) è una collezione video di eye-tracking che comprende dati di fissazione di 65 video multiplo-facciali provenienti da YouTube e Youku. È progettato per facilitare la ricerca sulle dinamiche dello sguardo e sull'interazione visiva umana, in particolare nello studio di come le persone fissano diversi volti nei video.

3.2 Dettagli implementativi

3.2.1 Identificazione e segmentazione degli oggetti nei video

Sebbene in contesti sociali i soggetti tendano spesso a rivolgere lo sguardo verso altre persone, esistono circostanze in cui gli oggetti presenti nell'ambiente catturano significativamente la loro attenzione.

In questa fase del progetto, è stato impiegato il framework Detectron2 per identificare e segmentare gli oggetti presenti nei frame estratti dai video. L'obiettivo era isolare gli oggetti non-persona permettendo di focalizzare l'attenzione sugli elementi del contesto che potrebbero influenzare la direzione dello sguardo dei soggetti.

Il modello Mask R-CNN con ResNet-50 e Feature Pyramid Network (FPN) è stato configurato per identificare diverse istanze di oggetti basandosi sul dataset COCO, che fornisce un ampio riconoscimento di oggetti in contesti naturali. La soglia di riconoscimento è stata impostata a 0.6 per assicurare che solo gli oggetti con un alto grado di confidenza fossero considerati, riducendo così il rischio di falsi positivi.

Una volta configurato, il modello è stato utilizzato per processare i frame e ottenere predizioni precise sulle posizioni e le classi degli oggetti. Il focus era sull'identificazione di oggetti non-persona, utilizzando l'indice di classe specifico per escludere le persone dalle analisi. Questo passaggio è cruciale per concentrarsi sugli oggetti che i soggetti potrebbero guardare.

Per verificare l'accuratezza della segmentazione e facilitare la revisione visuale, le bounding box degli oggetti identificati come non-persona sono state visualizzate sulle immagini originali per confermare l'efficacia del modello nell'isolare gli oggetti di interesse.

Le coordinate delle bounding box estratte sono state quindi convertite in un formato utilizzabile successiva.

3.2.2 Estrazione, ottimizzazione e tracciamento delle bounding box dei volti tramite Densepose

L'utilizzo del modello DensePose ha permesso una mappatura precisa delle superfici corporee 3D dei soggetti nei video, fondamentale per l'analisi delle pose umane. Tuttavia, la tendenza del sistema a generare falsi positivi ha richiesto l'introduzione di un processo di selezione ottimizzato delle bounding box.

Per ridurre gli errori, è stato implementato un filtro che conserva solo le n bounding box più grandi per ogni frame, con n corrispondente al numero effettivo di persone presenti. Questo approccio assume che le bounding box di dimensioni maggiori rappresentino più probabilmente soggetti reali. Successivamente, è stato implementato un algoritmo di tracciamento dell'identità basato su IoU per mantenere la coerenza dell'identificazione dei soggetti attraverso i frame successivi. Utilizzando l'IoU come metrica, il sistema cerca di abbinare ogni bounding box rilevata con l'identità già nota più vicina, aggiornando o assegnando nuovi identificatori se necessario. Questo metodo assicura che le stime sulla direzione dello

sguardo siano collegate in modo affidabile ai soggetti corretti nel tempo, facilitando analisi comportamentali precise.

3.2.3 Uso di Gaze360 e generazione delle mappe di salienza

Per analizzare la direzione dello sguardo dei soggetti nei video, è stato impiegato l'algoritmo Gaze360.

Il modello riceve in input sequenze di frame di immagini. Ogni sequenza è composta da 7 frame, con il frame centrale come target per la previsione della direzione dello sguardo. Da ciascun frame viene estratta una porzione che contiene la testa del soggetto la quale viene quindi passata attraverso una rete neurale convoluzionale (CNN) per l'estrazione delle caratteristiche. Le caratteristiche estratte dalla CNN per ciascun frame vengono quindi passate a una rete LSTM bidirezionale che è in grado di modellare le sequenze di dati considerando sia le informazioni passate che future. Nel contesto del modello Gaze360, l'uso di LSTM bidirezionali permette di sfruttare le informazioni temporali contenute nei 7 frame per migliorare la stima della direzione dello sguardo, risolvendo eventuali ambiguità presenti in singoli frame. L'output delle LSTM è una rappresentazione compatta che viene passata attraverso un livello completamente connesso per produrre due output: la direzione dello sguardo e una stima dell'incertezza. La direzione dello sguardo è rappresentata in coordinate sferiche, che sono considerate più interpretabili nel contesto della stima 3D dello sguardo.

Una volta ottenuto l'output, la sfida principale è stata interpretare la direzione dello sguardo dei soggetti e determinare verso quali elementi nella scena essi fossero rivolti. Il primo passo è stata la conversione delle coordinate sferiche prodotte dall'algoritmo in coordinate cartesiane, semplificando il confronto spaziale tra la direzione dello sguardo e gli oggetti o le persone presenti nel video.

Una volta trasformate le coordinate, sono state identificate tre possibili focali dell'attenzione dei soggetti:

- **Verso la telecamera:** Stabilire se lo sguardo di un soggetto è diretto verso la telecamera è cruciale per valutare l'interazione diretta con l'osservatore. Secondo il principio dell'"eye contact effect" [6], gli spettatori tendono a essere naturalmente attratti dagli sguardi che incontrano direttamente la telecamera. Per determinare se uno sguardo è orientato verso la telecamera, si assume che questa sia posizionata frontalmente al soggetto, rappresentata dal vettore [0, 0, -1]. Il vettore della direzione dello sguardo viene normalizzato per ridurlo a una lunghezza unitaria, permettendo un confronto basato solo sulla direzione. Successivamente, viene calcolato il prodotto scalare tra il vettore normalizzato dello sguardo e il vettore della telecamera per valutare la loro similarità direzionale. L'angolo tra i due vettori viene calcolato tramite l'arcocoseno del prodotto scalare. Se l'angolo risultante è inferiore a una soglia predeterminata di 30 gradi, si conclude che il soggetto sta guardando direttamente verso la telecamera.
- **Verso un altro soggetto:** Per comprendere se un soggetto sta guardando un altro individuo, è stato sviluppato un algoritmo specifico. Questo metodo verifica se un soggetto si trova nella direzione generale dello sguardo. Le coordinate dell'occhio e la direzione dello sguardo vengono utilizzate per determinare se la linea di vista include la posizione di un altro soggetto. Successivamente, la direzione dello sguardo viene confrontata con la posizione precisa del soggetto, convertendo le posizioni degli occhi e del soggetto in vettori bidimensionali. Questi vettori vengono normalizzati e viene calcolato il prodotto scalare tra la direzione dello sguardo e la direzione verso il soggetto. Anche in questo caso l'arcocoseno di questo prodotto viene utilizzato per ottenere l'angolo tra i due vettori. Se l'angolo è inferiore a una soglia predefinita di 60 gradi, si interpreta che lo sguardo è effettivamente diretto verso quel soggetto. Vengono confrontate le distanze tra la posizione degli occhi e i soggetti nei dintorni, identificando il soggetto con la distanza minima come il focus dell'attenzione visiva.
- **Verso un oggetto:** Se la direzione dello sguardo di un soggetto non è rivolta né verso la telecamera né verso un altro soggetto, viene valutato se lo sguardo può essere diretto verso un oggetto nella scena. Il processo utilizza un approccio simile a quello adottato per determinare se lo sguardo è diretto verso un altro soggetto. Viene verificato se lo sguardo interseca la posizione di uno o più oggetti, analizzando se ciascun oggetto potrebbe essere quello che il soggetto sta guardando. Si valuta quale di questi oggetti ha il centro di massa più vicino alla linea di sguardo del soggetto.

Nel caso in cui la direzione dello sguardo di un soggetto non rientri in nessuno dei tre scenari precedentemente descritti, non viene generato alcun punto di fissazione. È frequente che nei video i soggetti guardino verso elementi o persone fuori dalla scena visibile. Generare un punto di fissazione sul bordo del frame per rappresentare tali casi avrebbe prodotto risultati inaccurati e fuorvianti.

Una volta che sono state memorizzate le coordinate dei punti di fissazione viene innescato un processo di generazione di mappe di densità visiva per ciascun frame. Inizialmente, viene creato un istogramma 2D dai punti di fissazione rilevati, che viene poi trattato con un filtro gaussiano. La scelta del parametro sigma per il filtro gaussiano è cruciale per modulare l'effetto di smussamento; esso viene calcolato basandosi sulla dimensione media dei bounding boxes dei soggetti nel frame divisa per sei, permettendo così di adattare dinamicamente l'intensità del filtro alle variazioni di scala delle figure nell'immagine. L'applicazione del filtro gaus-

siano serve a evidenziare le aree di maggiore interesse visuale, producendo una mappa di calore con transizioni morbide tra zone ad alta e bassa densità di fissazione. La mappa risultante viene normalizzata per assicurare che la somma dei suoi valori sia unitaria, riflettendo proporzionalmente l'intensità delle fissazioni relative al contesto del frame. Infine, la mappa di calore normalizzata viene scalata a valori compresi tra 0 e 255 e salvata come immagine in scala di grigi, facilitando l'analisi e la presentazione visiva delle dinamiche di attenzione visiva osservate.

4 Risultati ottenuti

Per valutare l'efficacia delle mappe di salienza, generate utilizzando un algoritmo di gaze following, nel predire il gaze dei soggetti eye-tracked mentre guardano i video, è stato impiegato un approccio basato sull'analisi di regressione [7]. Questo approccio permette di quantificare l'importanza relativa di diverse mappe di salienza nel modellare la distribuzione delle fissazioni oculari, poiché probabilmente la social value map è solo uno dei fattori che contribuiscono a spiegare l'allocazione dell'attenzione visiva. Le mappe di salienza considerate includono quindi la Social Value Map, la Speaker Map, la Non-Speaker Map, la STS Map (salienza di basso livello), la Center Bias (CB) Map e la Uniform Map. Queste mappe sono state generate per ogni frame di ogni video e analizzano differenti aspetti della scena:

- **Social Value Map:** Rappresenta i punti in cui i soggetti presenti all'interno del video rivolgono il proprio sguardo.
- **Speaker Map:** Evidenzia il/i soggetto/i parlanti.
- **Non-Speaker Map:** Evidenzia i soggetti non parlanti.
- **STS Map:** Mappa di salienza di basso livello che considera caratteristiche visive come contrasto e movimento.
- **Center Bias Map:** Rappresenta il bias centrale, ovvero la tendenza degli osservatori a fissare il centro dello schermo.
- **Uniform Map:** Mappa uniforme utilizzata come baseline.

Poiché i video sono composti da frame sequenziali, al posto di eseguire una regressione per ogni video, per tenere conto della dinamica temporale nei valori di beta, è stato utilizzato un filtro di Kalman. La regressione lineare è stata applicata solo per il primo frame di interesse, per inizializzare i valori di beta. Successivamente, il filtro di Kalman è stato impiegato per aggiornare questi valori frame per frame, tenendo conto delle variazioni temporali e riducendo la rumorosità nei dati. Questo approccio consente di ottenere una stima più stabile e accurata dei valori di beta lungo il video.

Per ridurre la complessità computazionale, l'analisi è stata limitata ai frame dal 100 al 300 per ogni video. Questo intervallo è stato scelto per rappresentare una porzione significativa del video senza rendere eccessiva la computazione.

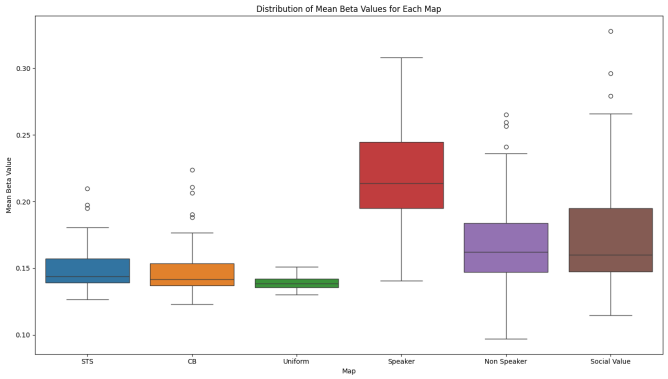

Fig. 2: Distribuzione dei valori medi di beta per ciascuna mappa

È stata considerata la media dei valori di beta per ogni mappa piuttosto che l'ultimo valore di beta ottenuto per l'ultimo frame. Questo perché la media fornisce una stima più robusta e rappresentativa delle prestazioni complessive di ciascuna mappa lungo l'intero intervallo di frame analizzato, mitigando l'impatto di eventuali fluttuazioni o anomalie temporanee. Questi valori medi sono stati utilizzati per generare due tipi di visualizzazioni: un boxplot e cinque diverse area chart che confrontano l'andamento della social value map con tutte le altre mappe nei vari video.

Il boxplot presentato nella figura 2 mostra la distribuzione dei valori medi di beta per ciascuna delle sei mappe di salienza utilizzate nello studio. Questa rappresentazione grafica consente di analizzare la centralità, la dispersione e la presenza di outliers nei dati relativi ai beta medi. Analizzando i risultati, si può notare che le mappe STS, CB e Uniform mostrano valori mediani di beta simili, inferiori a 0.15, con una variabilità relativamente bassa. La presenza di alcuni outliers suggerisce che in alcuni video queste mappe hanno valori di beta leggermente più alti o più bassi rispetto alla media, ma generalmente la loro efficacia è consistente un esito atteso data la natura di queste mappe.

La STS Map si basa su caratteristiche visive di basso livello come il contrasto e il movimento, che sono generalmente presenti in tutte le scene video, spiegando così la sua performance costante e l'assenza di grandi variazioni nei valori di beta.

La CB Map rappresenta il bias centrale, una tendenza universale degli osservatori a fissare il centro dello schermo. La natura invariabile di questo fenomeno visivo porta a una distribuzione dei beta molto stretta e consistente.

L'Uniform Map, utilizzata come baseline, non tiene conto di nessuna caratteristica specifica della scena, risultando in valori di beta medi relativamente uniformi e una variabilità minima.

La Speaker Map ha il valore mediano di beta più alto, attorno a 0.2, con un IQR ampio che si estende fino a

circa 0.3. Questo indica una grande variabilità nella sua efficacia come predittore del gaze. La maggiore altezza del box e la presenza di outliers più elevati suggeriscono che in alcuni video la presenza di parlanti è un fattore estremamente importante per la predizione del gaze. La variabilità significativa può essere attribuita al fatto che l'attenzione degli osservatori varia notevolmente a seconda della presenza e della posizione dei parlanti nei video, rendendo la Speaker Map molto sensibile alle dinamiche sociali della scena.

La Non-Speaker Map mostra una variabilità moderata con un valore mediano attorno a 0.16. L'intervallo IQR è più ampio rispetto alle mappe di STS, CB e Uniform, ma non raggiunge l'ampiezza della Speaker Map, indicando una performance più consistente ma comunque influenzata dalla presenza di non parlanti nei video. La variabilità può essere giustificata dalla natura dinamica delle interazioni sociali nei video, dove i soggetti non parlanti possono attirare l'attenzione degli osservatori in modo variabile.

La Social Value Map presenta un valore mediano di beta attorno a 0.15, con un IQR ampio che riflette una variabilità significativa nella sua efficacia. Questo suggerisce che la Social Value Map cattura meglio le dinamiche sociali presenti nei video, rendendola una mappa particolarmente utile per predire il gaze dei soggetti eye-tracked. Tuttavia, la presenza di outliers indica che la sua performance può variare notevolmente a seconda del contenuto del video. La variabilità della Social Value Map è dovuta alla sua capacità di rappresentare i punti in cui i soggetti nel video rivolgono il proprio sguardo, un aspetto che può variare notevolmente a seconda delle dinamiche sociali e delle interazioni presenti in ogni scena.

Da questa prima analisi possiamo concludere che la Speaker Map ha la maggiore variabilità e i valori mediani di beta più alti rispetto alle altre mappe, suggerendo una maggiore capacità predittiva in contesti specifici. La Non Speaker Map e la Social Value Map mostrano valori mediani di beta simili, con la Non Speaker Map leggermente superiore. Tuttavia, la Social Value Map presenta una maggiore variabilità, indicata dall'estensione dell'IQR e dalla lunghezza dei baffi, il che riflette una maggiore sensibilità alle dinamiche sociali presenti nei video. Questo suggerisce che, sebbene la Social Value Map sia efficace nel catturare le dinamiche sociali, la sua performance varia notevolmente in base al contenuto specifico dei video. Le mappe di STS, CB e Uniform mostrano una performance più consistente ma meno influenzata dai fattori specifici dei video, evidenziando un ruolo complementare nelle analisi di salienza visiva.

Vediamo quindi ora le area chart per avere una visione più dettagliata del comportamento della Social Value Map rispetto alle altre mappe.

Nell'area chart che confronta la Social Value Map con la Uniform Map (Fig 3), possiamo osservare una chiara distinzione tra i due set di dati. La Social Value Map (linea arancione) mostra una variabilità significativa

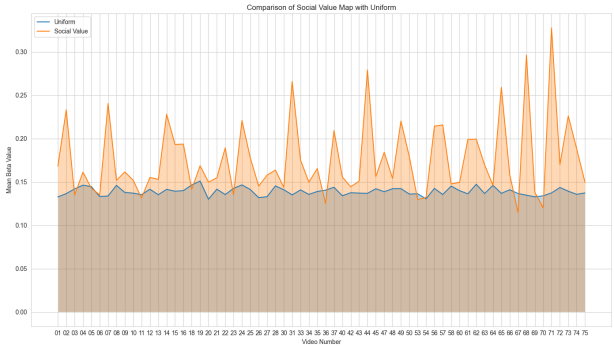

Fig. 3: Comparison of social value map with Uniform

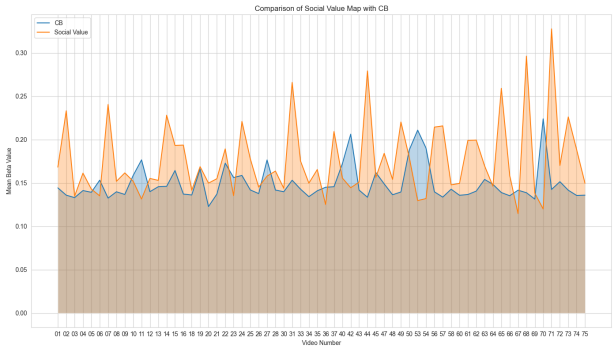

Fig. 4: Comparison of Social value map with CB

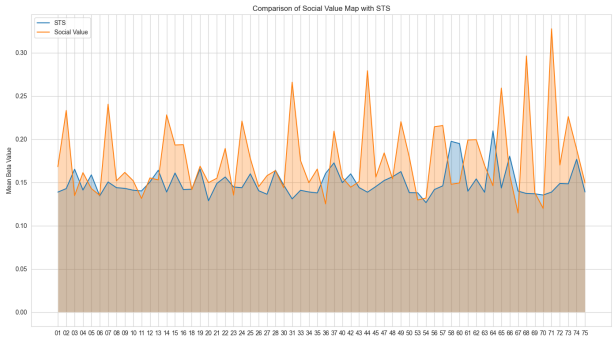

Fig. 5: Comparison of Social value map with STS

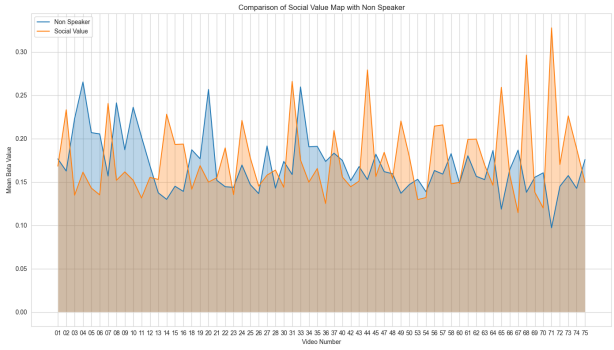

Fig. 6: Comparison of Social value map with Non Speaker

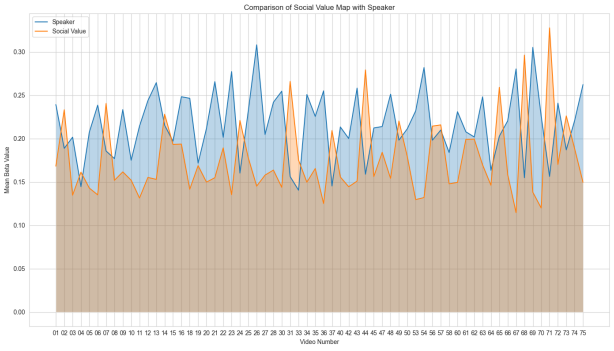

Fig. 7: Comparison of Social value map with Speaker

con picchi ben definiti, indicativi di un'interazione dinamica e contestuale all'interno dei video. Al contrario, la Uniform Map (linea blu) mantiene un andamento molto più stabile e basso, suggerendo una distribuzione uniforme e costante dei valori di beta. Questa stabilità è prevista, dato che la Uniform Map non tiene conto delle caratteristiche specifiche della scena, ma serve solo come baseline.

Quando si confronta la Social Value Map con la CB Map (Fig 4), notiamo che la CB Map (linea blu) mostra valori leggermente superiori rispetto alla Uniform Map, ma ancora piuttosto stabili. Questa stabilità è dovuta al bias centrale che rappresenta, poiché gli osservatori tendono a fissare il centro dello schermo. Tuttavia, la Social Value Map evidenzia maggiore sensibilità alle dinamiche sociali presenti nei video, mostrando picchi di beta più alti in corrispondenza di specifici contesti sociali.

Nel confronto con la STS Map (Fig 5), la STS Map (linea blu) presenta valori stabili e bassi di beta, riflettendo la sua base su caratteristiche visive di basso livello come il contrasto e il movimento, che sono generalmente presenti in tutte le scene video.

Il confronto tra la Social Value Map e la Non-Speaker Map (Fig 6) rivela che entrambe le mappe presentano una significativa variabilità nei valori di beta. La Non-Speaker Map mostra un andamento più variabile rispetto alle mappe STS, CB e Uniform, ma meno estremo rispetto alla Social Value Map, indicando che i non parlanti possono attirare l'attenzione visiva, ma in modo meno consistente.

La Speaker Map, quando confrontata con la Social Value Map (Fig 7), mostra un comportamento altamente variabile. Entrambe le mappe presentano picchi elevati, ma la Speaker Map (linea blu) tende a superare la Social Value Map in diverse occasioni. Questo indica che i soggetti parlanti sono un fattore cruciale per la predizione del gaze.

Le area chart dimostrano chiaramente che la Social Value Map è efficace nel catturare le dinamiche sociali presenti nei video, ma non è la mappa che performa meglio in assoluto. La Speaker Map infatti mostra una capacità predittiva superiore in diversi contesti, indicando che i soggetti parlanti sono un fattore determinante per l'attenzione visiva. Altre mappe, come la Uniform, CB, e STS, presentano una performance più stabile ma meno influenzata dai dettagli sociali. La Social Value Map, quindi, è particolarmente utile in combinazione con altre mappe per fornire una visione completa e accurata delle dinamiche dell'attenzione visiva nei video.

5 Commenti conclusivi

Analizzando complessivamente i risultati ottenuti, è evidente che la Social Value Map rappresenta un interessante strumento per predire il gaze dei soggetti eye-tracked mentre guardano i video. Questa mappa ha dimostrato un'efficacia significativa in vari contesti video, sebbene la sua performance vari a seconda delle caratteristiche specifiche delle scene rappresentate.

Analizzando i video in cui la Social Value Map presenta i picchi più alto risulta che ha performato meglio nei video in cui i soggetti guardano la telecamera. Questo è probabilmente dovuto al fatto che tali contesti creano una forte connessione tra i soggetti del video e gli osservatori, inducendoli a dirigere il proprio sguardo verso il punto di salienza indicato dalla mappa.

Tuttavia, la Social Value Map ha mostrato performance inferiori nei video in cui i soggetti tendono a guardarsi tra di loro. Questo può essere attribuito a due principali limitazioni dell'algoritmo utilizzato per generare i punti di salienza:

- **Mancanza di considerazione della profondità della scena:** L'algoritmo non tiene conto della distanza relativa tra i soggetti. Questo può portare a una rappresentazione inaccurata dei punti di salienza, con l'algoritmo che posiziona i punti verso il soggetto più vicino quando, in realtà, l'osservatore potrebbe essere più interessato a un soggetto più lontano.
- **Mancanza di un algoritmo di speaker identification:** Dall'analisi è emerso che i soggetti parlanti attirano maggiormente lo sguardo degli osservatori. Questo è in linea con le teorie della psicologia cognitiva, che suggeriscono che gli esseri umani tendono a focalizzare l'attenzione sui soggetti che stanno comunicando attivamente, poiché essi forniscono informazioni socialmente rilevanti. L'algoritmo di gaze following utilizzato non è in grado di distinguere chi sta parlando, una caratteristica che si è rivelata essere la più importante. Infatti, la Speaker Map, che identifica i soggetti parlanti, ha mostrato le performance migliori, evidenziando l'importanza di questa informazione nella predizione del gaze.

In conclusione, la Social Value Map ha mostrato una buona capacità predittiva, particolarmente in contesti dove i soggetti guardano direttamente la telecamera. I risultati suggeriscono che migliorando l'algoritmo di gaze following, in modo che consideri la profondità della scena e includa algoritmi di speaker identification, si possono ottenere risultati significativamente

migliori e più performanti. Implementare queste modifiche potrebbe potenziare ulteriormente la capacità della Social Value Map di predire con precisione il gaze degli osservatori, rendendola uno strumento ancora più efficace per l'analisi dei video.

REFERENCES

[1] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2018.

[3] I. K. R{iza Alp Güler, Natalia Neverova, "Densepose: Dense human pose estimation in the wild," 2018.

[4] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, , and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[5] M. Xu, Y. Liu, R. Hu, and F. He, "Find who to look at: Turning from action to saliency," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4529–4544, 2018.

[6] M. Senju, A.; Johnson, "The eye contact effect: mechanisms and development," in *Trends in Cognitive Sciences 13 (3), pp.127-134*, 2009.

[7] G. Boccignone, V. Cuculo, A. D'Amelio, G. Grossi, and R. Lanzarotti, "On gaze deployment to audio-visual cues of social interactions," *IEEE Access*, pp. 1–25, 2020. [Online]. Available: https://doi.org/10.1109/access.2020.3021211