Università Politecnica Delle Marche

Dipartimento di Ingegneria dell'Informazione

LAUREA MAGISTRALE IN INGEGNERIA INFORMATICA E DELL'AUTOMAZIONE

Progettazione e sviluppo di una soluzione per la rilevazione e l'analisi automatica dei landmark facciali in pazienti affetti da malattie neurologiche

Design and development of a solution for automatic detection and analysis of facial landmarks in patients with neurological diseases

Relatore:

Prof. Emanuele Frontoni

Correlatori:

Dott.ssa Lucia Migliorelli

Sara Moccia, PhD

Candidato: 1096444

Francesco Alborino

Anno Accademico 2020 / 2021

Mentre negli anni '90 Internet abbracciava i primi settori,

oggi è impossibile prescindere da esso.

Allo stesso modo, ora, il Deep Learning sta abbracciando i primi settori...

- Un mio semplice pensiero -

Sommario

Il presente elaborato di tesi presenta un possibile sistema, sviluppato con un approccio Deep Learning, in grado di effettuare il rilevamento dei punti facciali di interesse clinico in maniera automatica nei pazienti affetti da patologie neurologiche, nello specifico Ictus e SLA, al fine di migliorare le attuali pratiche di valutazione cliniche.

L'analisi dei movimenti e delle espressioni facciali per applicazioni sanitarie è un'area di ricerca in rapida crescita, che ha visto importanti progressi negli ultimi anni. Alcune delle applicazioni in cui tale task è impiegato sono il riconoscimento del dolore a partire da immagini e video del viso; l'analisi automatica delle espressioni in pazienti con disturbi neurologici (ad esempio, Parkinson, Ictus, SLA, Alzheimer, ecc.) e il rilevamento automatico di sintomi legati a condizioni psichiatriche, come la depressione e la schizofrenia. Indipendentemente dallo specifico contesto, l'obiettivo generale è quello di fornire informazioni accurate, oggettive e standardizzate ai clinici, associate alla dinamica cinematica e dinamica facciale, al fine di migliorare le attuali pratiche di valutazione e per valutare gli effetti del trattamento. In molti casi, il rilevamento dei landmark facciali è usato come punto di partenza della pipeline di elaborazione, al fine di estrarre robuste caratteristiche spazio-temporali dei gesti e delle espressioni che, a loro volta, possono essere utilizzate per dedurre la condizione clinica di interesse.

Il modello presentato dovrà essere robusto alle menomazioni oro-facciali indotte dalla progressione o dalla severità della malattia, e robusto alle particolari espressioni compiute dai pazienti durante l'esecuzione dei task appositamente preparati dai clinici. Questa soluzione permetterà ai clinici di avere un sistema oggettivo

che fornisca valutazione riguardo l'evoluzione delle patologie neurologiche del paziente. Inoltre, verranno presentate delle proposte per il calcolo di indici clinici specifici (ad esempio, l'indice di simmetria facciale e labiale).

Viene, in seguito, approfondito il task di facial landmark detection (o rilevamento dei punti di riferimento del viso), che è un sottoinsieme del problema della predizione della forma del corpo di un essere umano. L'obiettivo è, quindi, quello di rilevare strutture facciali utilizzando metodi di predizione suddivisi in due fasi: il primo step consiste nella localizzazione del volto all'interno dell'immagine, mentre il secondo prevede il rilevamento dei landmark facciali.

Successivamente, sono descritti i metodi di interesse, affrontando prima il task generico della detection degli oggetti, fino a raffinare il problema in quello più specifico di rilevamento dei landmark facciali. Infatti, tale task presuppone che venga, per prima cosa, rilevato il bounding box corrispondente al volto del soggetto, per poi effettuare la predizione dei landmark all'interno della regione individuata.

In seguito, sono descritti i dataset e i modelli con le relative configurazioni di parametri utilizzati per le fasi di training. Inoltre, vengono mostrati gli sperimentati compiuti con i relativi risultati.

Infine, vengono discussi i risultati ottenuti dagli esperimenti effettuati e i possibili sviluppi futuri dell'intero sistema.

Indice

Capitolo 1

Introduzione

Il seguente capitolo presenta una introduzione all'elaborato, presentando le malattie del sistema del nervoso e l'obiettivo della presente tesi.

1.1 Le malattie del sistema nervoso

Le malattie neurologiche sono le patologie che hanno per oggetto il sistema nervoso, ossia encefalo, midollo spinale e/o nervi.
Le malattie neurologiche scoperte sono circa 600, alcune delle quali molto conosciute quali l'ictus, le demenze, l'emicrania, la sindrome del tunnel carpale, morbo di Parkinson e Sclerosi Laterale Amiotrofica (SLA).
Le patologie neurologiche non sono tutte gravi allo stesso modo, ma si collocano tra modesta gravità clinica ed alta gravità. Le patologie rientranti in quest'ultima categoria sono capaci di avere effetti altamente debilitanti o, peggio, possono condurre con un lento declino verso la morte del soggetto. Frequentemente si riscontrano tali patologie nei soggetti anziani; tuttavia, per le malattie del sistema nervoso figurano anche condizioni congenite (quindi manifeste fin dalla nascita) e condizioni che compaiono molto prima dell'età avanzata.
Le patologie neurologiche sono oggetto di studio per neurologi (medici con specializzazione in neurologia), neurochirurghi (medici con specializzazione in neurochirurgia) e neuropsichiatri (medici con specializzazione in psichiatria).

1.2 Incidenza della malattie neurologiche

Le malattie del sistema nervoso che richiedono l'intervento dello specialista neurologo mostrano un'incidenza del 7,5% l'anno e una prevalenza del 30%. A questi numeri vanno aggiunte quelle situazioni di malattia del sistema nervoso che non arrivano – per qualsiasi ragione – allo specialista neurologo, come per esempio le cefalee, le demenze, il low-back-pain (che rappresenta la maggiore causa di assenza dal posto di lavoro nel mondo occidentale) e altre ancora.
In Italia gli studi evidenziano una prevalenza del dolore cronico, con una percentuale di diffusione del 27% secondo le ultime stime. In realtà, però, non sono riportati dati precisi circa il dolore neuropatico. Se ipotizziamo il rapporto tra dolore cronico e neuropatico pari a quello rilevato in Europa, allora la prevalenza di dolore neuropatico dovrebbe aggirarsi attorno al 6% nella popolazione italiana. In questo elaborato, nello specifico, verranno trattate la SLA e l'Ictus.
La Sclerosi Laterale Amiotrofica (SLA) è una malattia neurodegenerativa che porta ad una degenerazione dei motoneuroni, fino a causare una paralisi totale. Attualmente non esiste cura e l'esito è infausto. L'incidenza è di circa 1-3 casi ogni 100.000 abitanti all'anno. In Italia si stimano almeno 3.500 malati e 1.000 nuovi casi ogni anno. La prevalenza, cioè il numero di casi presenti sulla popolazione, è in aumento: questo grazie alle cure che permettono di prolungare la vita del malato[1].
L'ictus cerebrale, invece, è la più frequente malattia neurologica, per la quale il cervello, a seguito della chiusura o della rottura di un'arteria, non riceve più sangue (ischemia) o viene inondato da sangue stravasato da un'arteria rotta (emorragia). Il primo caso rappresenta circa l'85% di tutti i casi di ictus. In Italia si verificano ogni anno circa 200.000 casi di questa patologia: di questi, circa l'80% è rappresentato da nuovi episodi. Gli studi statistici mostrano che l'ictus acuto causa più morti dell'infarto del miocardio (7,28 vs 4,95 per 10.000 abitanti), inoltre la mortalità a 30 giorni dopo un ictus ischemico è pari al 20% e dopo un ictus emorragico al 50%. È, inoltre, evidenziato che il tasso di prevalenza

[1] https://www.osservatoriomalattierare.it/malattie-rare/sla

di ictus nella popolazione anziana (età 65-84 anni) italiana è del 6,5%, leggermente più alto negli uomini (7,4%) rispetto alle donne (5,9%). I tassi grezzi di incidenza sulla popolazione italiana in diverse località variano tra 1,54 e 2,89 per 1.000, anche in rapporto alla variabilità dell'età media delle popolazioni considerate. Nel momento in cui si manifesta un evento cerebrovascolare acuto, questo va affrontato in un'ottica globale e appropriata, partendo dal riconoscimento dei primi sintomi, passando all'attivazione del servizio di emergenza-urgenza, all'individuazione e allertamento delle strutture adeguate, al trattamento dell'ictus e del TIA, fino alla gestione intraospedaliera, al progetto e al trattamento riabilitativo, alla prevenzione secondaria, all'auspicabile rientro al domicilio, alla presa in carico da parte del MMG, dello specialista territoriale o dell'ADI[2].

1.3 Il panorama clinico attuale

Attualmente, la valutazione per la progressione delle patologie neurologiche è eseguita direttamente dai clinici (es. esame del nervo cranico) oppure usando tecniche basate su sensori (ad es. metodi di tracciamento elettronico, articolografia elettromagnetica) [1]. Tuttavia, le valutazioni dei clinici risultano essere molto soggettive e mostrano una ridotta affidabilità, mentre le tecniche basate su sensori richiedono strumenti costosi, complessi e molto invasivi. Questi inconvenienti vietano la traduzione di tale tecnologia in pratica clinica quotidiana, limitando, di conseguenza, l'efficacia del monitoraggio della progressione della malattia [1].

La ricerca improntata sulla Computer Vision e sul Deep Learning può aiutare a migliorare le valutazioni cliniche, rendendole meno invasive, più precise e, sopratutto, oggettive. Lo studio del volto umano attraverso la Computer Vision per scopi clinici ha prosperato nel ultimi anni, trovando molte applicazioni in neurologia, patologie del linguaggio e psichiatria[1]. La disponibilità di approcci di allineamento del viso efficienti e accurati costituisce un passo importante verso

[2] http://www.rssp.salute.gov.it/rssp/paginaParagrafoRssp.jsp?sezione=situazione&capitolo=malattie&id=2655

lo sviluppo di strumenti intelligenti che innovino la modalità di seguire e curare il paziente. Studi recenti hanno riportato che misure semplici e clinicamente interpretabili (ad es. velocità, accelerazione, range di movimento) estratte dal labbro e dai movimenti della mascella consentono di mappare i sintomi e segni bulbari di patologie neurologiche e neurodegenerative come la Sclerosi Laterale Amiotrofica (SLA) e la Sclerosi Muscolare Spinale (SMA) oppure quantificare i progressi compiuti dai pazienti sottoposti a protocolli riabilitativi post Ictus.

1.4 Obiettivo della tesi

Il presente elaborato si pone come obiettivo quello di presentare una possibile soluzione, sviluppata con un approccio Deep Learning, in grado di effettuare un rilevamento dei punti facciali di interesse clinico in maniera automatica. Il modello sviluppato dovrà essere robusto alle menomazioni oro-facciali indotte dalla progressione o dalla severità della malattia, e robusto alle particolari espressioni compiute dai pazienti durante l'esecuzione dei task appositamente preparati dai clinici. Inoltre, dovrà essere in grado di fornire ad altri algoritmi specifici le coordinate precise dei punti di rilevamento, così da poter valutare l'evoluzione delle patologie neurologiche del paziente e monitorare gli eventuali progressi compiuti. Inoltre, verranno presentate delle proposte per il calcolo di indici clinici specifici (ad esempio, l'indice di simmetria facciale e labiale).
Come mostrato in Figura 1.1, il workflow del presente elaborato è suddiviso in due percorsi principali. Il primo, in giallo, mostra la fase di training: i video forniti dal dataset vengono sottocampionati (per evitare ridondanze di frame consecutivi, riducendo, così, la complessità computazionale), e vengono utilizzati per allenare il modello. Il secondo percorso, in verde, mostra come viene utilizzato il suddetto modello: si parte da un video, si sottocampiona, ed attraverso il modello avviene la predizione dei landmark, i quali verranno processati per ricavare i diversi indici di interesse clinico.

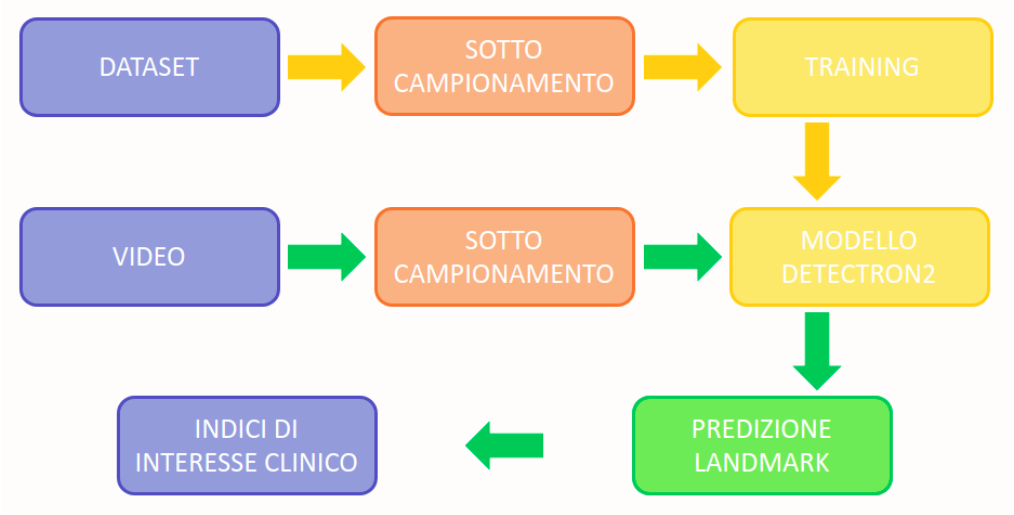

Figura 1.1: *Workflow ad alto livello dell'obiettivo della presente tesi.*

Capitolo 2

Stato dell'arte

Nel seguente capitolo viene mostrato lo stato dell'arte degli algoritmi di facial landmark detection e per le valutazioni cliniche attuali per valutare la progressione delle patologie neurologiche.

2.1 Facial Landmark Detection

2.1.1 Descrizione del task

Il task di facial landmark detection (o rilevamento dei punti di riferimento del viso) è un sottoinsieme del problema della predizione della forma del corpo di un essere umano. L'obiettivo è quindi quello di rilevare strutture facciali utilizzando metodi di predizione. Il rilevamento dei punti di riferimento facciali è quindi un processo in due fasi: localizzare del volto all'interno dell'immagine e rilevamento dei landmark del volto.

In generale, data un'immagine di input I di dimensioni W x H x C, dove W è la larghezza, H è l'altezza e C è il numero di canali colore dell'immagine (solitamente 3), il task di facial landmark detection è quello di trovare una funzione $\phi : I \rightarrow L$ che parta dall'immagine di input I e predica un vettore di riferimento L, tale che, per ogni punto di riferimento, contenga le coordinate (x, y). Il numero di landmark può essere variabile a seconda del contesto di utilizzo e dell'annotazioni presenti all'interno del dataset utilizzato, in questo elaborato è

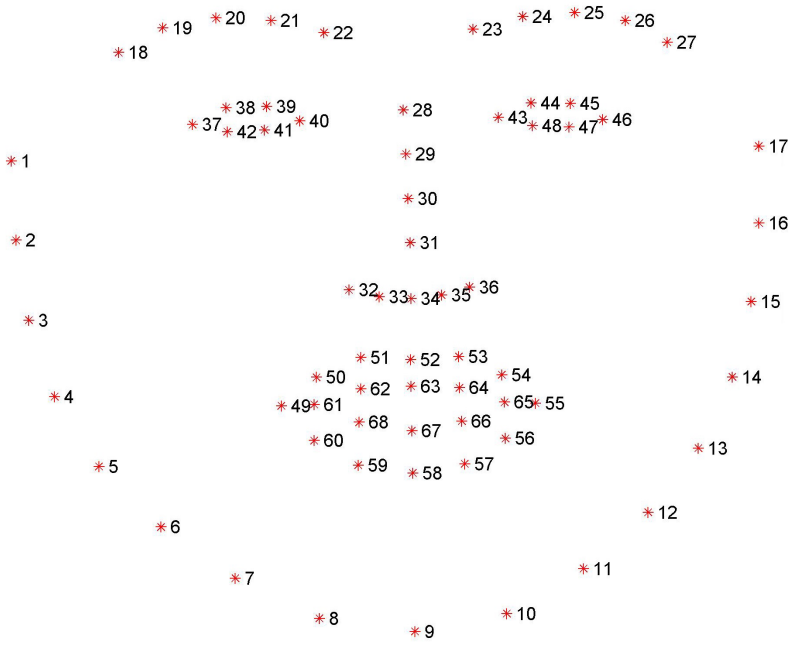

Figura 2.1: *Configurazione delle etichette che identificano i 68 landmark del volto*[2].

preso come riferimento la configurazione a 68 landmark, disposti in una delle principali configurazioni proposte da Dlib[1] (Figura 2.1).
Per valutare la qualità della funzione costruita, utilizzando i dati del set di test, devono essere definite delle metriche per il confronto degli algoritmi. Allo stato dell'arte viene utilizzata la metrica del Normalized Mean Error (NME):

$$NME = \frac{1}{K}\sum_{k=1}^{k} NME_k \rightarrow NME_k = \left[\frac{1}{N_L}\sum_{i=1}^{N_L}\frac{||y_i - p||_2}{d}\right] \cdot 100$$

dove y è il ground truth, p è la predizione, d il coefficiente di normalizzazione, N_L il numero di landmarks e K è il numero immagini contenute nel dataset.

2.1.2 Primi algoritimi per la detection di landmark facciali

I primi algoritmi si basavano principalmente sull'adattamento di una mesh facciale deformabile. Più algoritmi importanti includono Active Shape Model (ASM)[3], Active Appearance Model (AAM)[4] and Constrained Local Model (CLM)[5]. In

[1] http://dlib.net/

molti casi tali algoritmi utilizzano metodi statistici come base e prevedono con buona precisione in ambienti controllati (con illuminazione adeguata e faccia frontale). Tuttavia, in condizioni di ripresa reali, quelle di cui abbiamo bisogno in molte più applicazioni, la loro qualità è insufficiente.
Attualmente, gli algoritmi basati sull'utilizzo di reti neurali mostrano un errore più basso per il task di riconoscimento dei landmark facciali rispetto agli algoritmo che non impiegano le reti neurali per lo svolgimento del task. L'errore risulta essere accettabile anche con l'utilizzo di un ampio angolo di ripresa e in presenza di occlusioni. Questi algoritmi includono:

- **metodi di regressione diretta**, quando il modello predice coordinate (x, y) direttamente dall'immagine per ciascun landmark;

- **metodi di regressione di heatmap**, in cui viene creata una heatmap 2D per ciascun landmark. I valori nella mappa termica possono essere interpretati come probabilità che in una determinata posizione dell'immagine sia presente il landmark.

Inoltre, alcuni algoritmi sono implementati in una forma sequenziale, in cui una previsione viene affinata in più passaggi [6].

2.1.3 Algoritmi basati su reti neurali per la detection di landmark facciali

Lo stato dell'arte per la facial landmark detection presenta, per la maggior parte, soluzioni recenti le quali non sono ancora in produzione, ma sono realizzate e presentate semplicemente come articoli scientifici. Dei pochi algoritmi consolidati, ne vengono riportati i due principali:

- **Dlib**[2] è una libreria di Machine Learning open source. Essa ha un'elevata velocità di riconoscimento (secondo gli autori, circa 1 millisecondo per volto) e fornisce un modello addestrato con il dataset 300W. L'algoritmo è ancora attivamente utilizzato nella ricerca moderna.

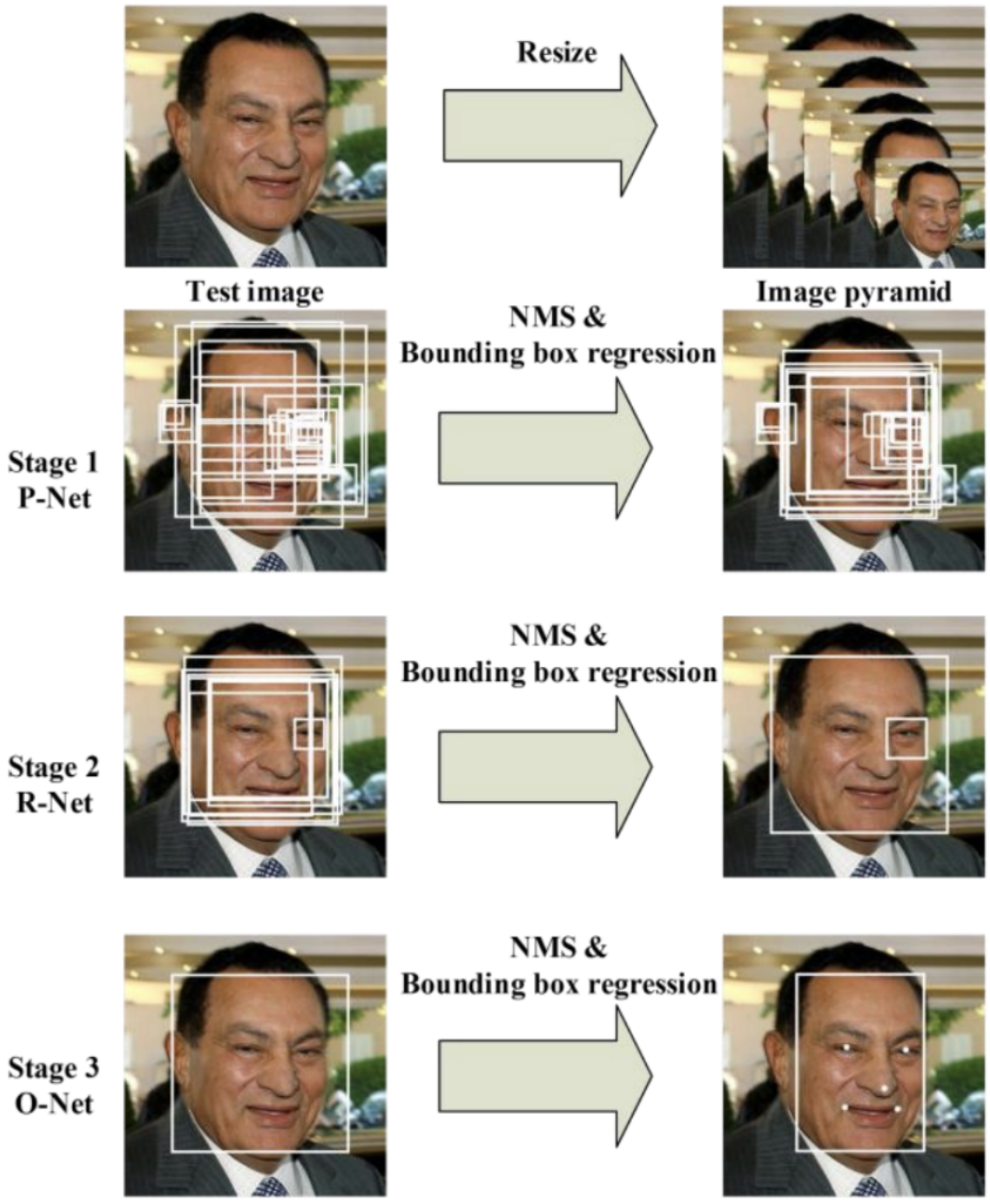

Figura 2.2: *P-Net è una rete elabora l'immagine originale in più risoluzioni e produce molte previsioni rettangolari della faccia. Successivamente, R-Net perfeziona i rettangoli predetti ed infine O-Net effettua il perfezionamento finale*[7].

- **Multi-task Cascaded Convolutional Networks (MTCNN)**[7] è un modello addestrato congiuntamente per riconoscere volti e landmark, nello specifico 5 punti quali occhi, punta del naso, angoli della bocca, mostrando un miglioramento su entrambi i compiti. La rete è costruita sotto forma di sequenza di tre reti: Proposal Network (P-Net), Refine Network (R-Net), Output Network (O-Net). Ognuna di loro predice il rettangolo di delimitazione della faccia, la probabilità che un particolare rettangolo contenga una faccia e i cinque punti di riferimento, come mostrato in Figura 2.2. P-Net è una rete veloce e completamente convolutiva, che elabora l'immagine originale in più risoluzioni (la cosiddetta piramide dell'immagine). Questa rete produce molte previsioni rettangolari della faccia che vengono, quindi, filtrate dall'algoritmo di Non-Maximum Suppression (NMS). Successivamente, R-Net perfeziona i rettangoli predetti, senza rielaborare l'intera immagine, in modo tale da risparmiare tempo. L'NMS viene, quindi, applicato nuovamente. Infine, O-Net effettua il perfezionamento finale. Quest'ultima è la rete più lenta, ma elabora un numero ridotto di rettangoli facciali [6].

Negli ultimi tre anni la ricerca si è concentrata sul cercare di individuare le principali problematiche sorte con il task di facial landmarks detection, per poi strutturare dataset e modelli robusti ad almeno uno di questi inconvenienti. Di questi, i principali sono:

- **Posizione della testa**: i volti da analizzare spesso non sono collocati di fronte alla telecamera, oppure hanno un'inquadratura non centrale, per esempio dal basso o dall'alto;
- **Luminosità**: la luminosità delle immagini è molto variabile;
- **Occlusioni**: Spesso le immagini da analizzare presentano elementi in primo piano che occludono parte del volto;
- **Espressioni**: alcune parti del viso, in particolare la bocca, possono presentare conformazioni particolari, difficili da predirre se non riscontrate in

fase di training dal modello;

- **Computazione**: la maggior parte dei modelli utilizzati per questo task, o perlomeno i più accurati, risultano essere molto onerosi.

Nell'articolo [6] sono riportate alcune soluzioni per cercare di risolvere alcuni di questi problemi. Innanzitutto, sono state realizzate diverse tipologie di dataset, alcune delle quali focalizzate su un problema specifico, come ad esempio Caltech Occluded Faces in the Wild (COFW), dataset che si concentra sull'etichettatura delle immagini del viso che sono parzialmente occluse da oggetti del mondo reale (microfono, occhiali, ecc.) o dalla persona stessa (capelli, mano, ecc.). Per quanto riguarda i modelli, l'articolo [6] riporta diverse tipologie di approcci, dipendenti anche dal principale inconveniente che si vuole risolvere, per esempio Practical Facial Landmark Detector (PFLD) [8] (Figura 2.3), la quale presenta, in fase di training, una sezione specifica per risultare più robusta a volti non posizionati di fronte all'inquadratura. Un'altra alternativa proposta è Deep Adaptive Graph (DAG) [9] (Figura 2.4), che, secondo gli autori, è in grado di comprende la struttura del viso nonostante la presenza di occlusioni parziali. L'articolo [6] presenta, inoltre, un confronto tra i vari NME dello stato dell'arte. Nonostante le differenze tra le tipologie di modelli, questi valori variano tra un minimo di circa il 3,5% e un massimo di circa il 5,5%. L'articolo [6] si conclude discutendo un altro aspetto molto importante: nonostante una significativa crescita della qualità dei metodi, pochi di si concentrano sull'applicabilità nel mondo reale, il che significa che in molti casi, anche quando eseguito su una GPU, gli algoritmi funzionano più lentamente che in tempo reale (circa 30 fps o 33 millisecondi). Inoltre, molte delle soluzioni presentate richiedono prestazioni elevate su dispositivi mobili o dispositivi portatili, ma, per quanto mostrato nell'articolo [6], solo in un caso gli autori di un algoritmo hanno realizzato un modello nato per applicazioni mobili, quindi più "leggero".

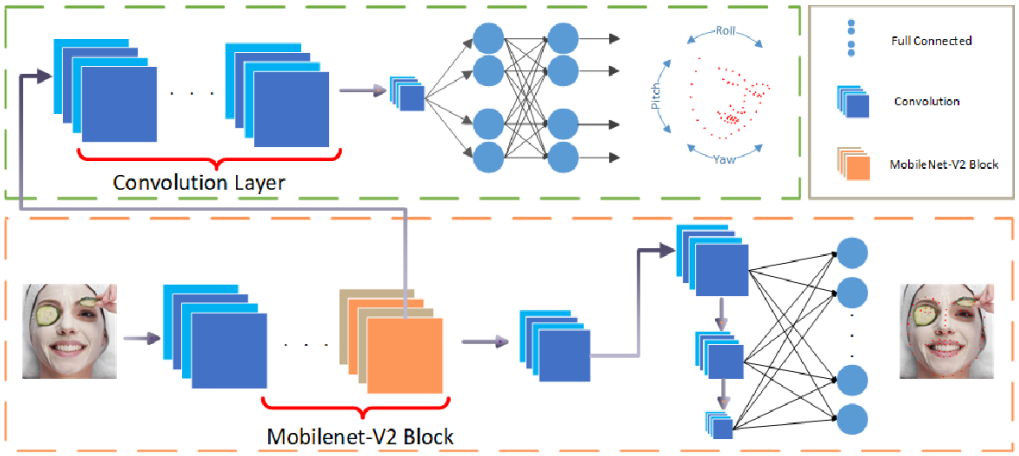

Figura 2.3: *Illustrazione dell'architettura Practical Facial Landmark Detector (PFLD). Questa è formata da due reti, la backbone network (struttura mostrata in basso) per prevedere le coordinate dei landmark e quella ausiliaria (struttura mostrata in alto), allenata per stimare correttamente le informazioni geometriche*[8].

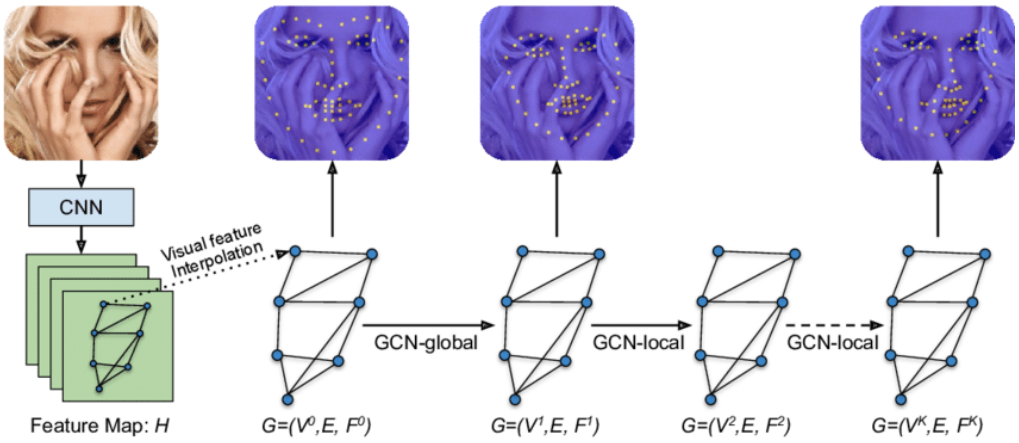

Figura 2.4: *Illustrazione dell'architettura Deep Adaptive Graph (DAG). Il grafo iniziale è inizializzato con il valore medio calcolato dai dati di training, poi il grafo di riferimento viene deformato da GCN-global attraverso una serie di trasformazioni prospettiche. Le caratteristiche visive e le caratteristiche di forma, infatti, sono reinterpolate dalla mappa delle caratteristiche e ricalcolate dopo ogni modulo GCN* [9].

2.2 Valutazioni della muscolatura facciale nella pratica clinica

2.2.1 Analisi automatica del viso per applicazioni cliniche

L'analisi dei movimenti e delle espressioni facciali per applicazioni sanitarie è un'area di ricerca in rapida crescita, che ha visto importanti progressi negli ultimi anni. Alcune delle applicazioni sono: il riconoscimento del dolore da immagini e video del viso; l'analisi automatica delle espressioni in pazienti con disturbi neurologici (ad esempio, Parkinson, Ictus, SLA, Alzheimer, ecc.); e il rilevamento automatico di sintomi legati a condizioni psichiatriche, come la depressione e la schizofrenia. Indipendentemente dalla specifica condizione, l'obiettivo generale è quello di fornire informazioni accurate, oggettive e standardizzate ai clinici, associate alla dinamica cinematica e dinamica facciale, al fine di migliorare le attuali pratiche di valutazione e per valutare gli effetti del trattamento. In molti casi, il rilevamento dei landmark facciali è usato come punto di partenza della pipeline di elaborazione, al fine di estrarre robuste caratteristiche spazio-temporali dei gesti e delle espressioni che, a loro volta, possono essere utilizzate per dedurre la condizione clinica di interesse. Recentemente, lo stato degli approcci all'avanguardia basati sul Deep Learning, come il Face Alignment Network (FAN), sono stati applicati su pazienti con demenza e paralisi facciale, dimostrando una maggiore precisione di localizzazione rispetto ai tradizionali approcci di allineamento del volto [1].

Wang et al.[10] hanno implementato diverse architetture di apprendimento profondo (3DCNN e multi-stream CNN) per estrarre le caratteristiche spazio-temporali dall'intera regione del viso, al fine di classificare diverse attività facciali in pazienti con l'Alzheimer.

Un altro approccio, presentato da Bishay et al.[11], consiste nell'implementara una rete VGG16 con l'obiettivo di rilevare le unità d'azione facciali (UA) da specifiche aree del viso. Tali regioni sono state usate come rappresentazione di basso livello per stimare la gravità della schizofrenia. Tuttavia, quando l'obiettivo è l'analisi della cinematica facciale in condizioni neurologiche che influenzano i ge-

sti e movimenti (per esempio, Parkinson, Ictus, SLA, ecc.), la rappresentazione facciale attraverso il rilevamento dei punti sarebbe preferibile, in quanto permette l'estrazione di misure clinicamente interpretabili, che possono essere correlate alla presenza, gravità e sviluppo della sintomatologia.

2.2.2 Dataset per la ricerca clinica

Poiché il principale ostacolo per ottenere un buon punto di riferimento per popolazioni cliniche è la disponibilità limitata di dati di allenamento annotati, gli autori dell'articolo [1] hanno rilasciato il primo dataset di video di individui con SLA e Ictus, con relativi punteggi clinici e con l'annotazione di 68 landamark facciali su oltre 3300 immagini. La disponibilità di questi dati mira a favorire lo sviluppo di approcci innovativi e robusti per la detection dei landmark facciali, che possono essere utilizzati per tracciare e analizzare movimenti facciali in queste popolazioni cliniche.

Capitolo 3

Metodi

Nel seguente capitolo viene descritta l'evoluzione degli approcci per la detection degli oggetti. Il task di rilevamento dei landmark facciali presuppone che venga rilevato prima il bounding box corrispondente al volto del soggetto, per poi effettuare la predizione dei landmark all'interno della regione individuata.

3.1 Reti R-CNN per la detection

3.1.1 Region-based CNN

Le reti neurali convoluzionali vengono utilizzate in diversi ambiti per molteplici obiettivi, per risolvere problemi di classificazione ma anche, ad esempio, problemi di segmentazione semantica delle istanze, ovvero il task di dividere un'immagine in diversi insiemi di pixel che devono essere opportunamente etichettati e classificati. Per effettuare la segmentazione semantica, però, è necessario compiere un primo stap, che consiste nell'identificazione dell'oggetto tramite object detection, per poi classificarlo.
Le reti region-based sono state proposte proprio per risolvere problemi di object detection, cioè task comprendenti sia la classificazione delle immagini sia il rilevamento di oggetti. La Region-based Convolutional Neural Network [12](Figura 3.1), in questo contesto, può rivelarsi molto utile, in quanto trasforma un problema di object detection in un problema di classificazione: a partire da un'im-

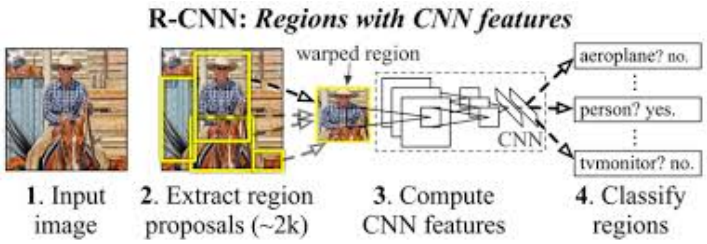

Figura 3.1: *Workflow di una R-CNN*

magine, vengono selezionate circa 2000 possibili regioni d'interesse (RoI) tramite l'algoritmo Selective Search [13], per poi estrarre le caratteristiche di ogni singola regione utilizzando una CNN. A questo punto, si effettuerà una classificazione delle regioni sulla base delle caratteristiche estratte applicando una SVM per la classificazione ed una regressione lineare, al fine di restringere il bounding box dell'oggetto. Viene, dunque, realizzata una SVM per ogni classe e, quindi, in fase di training, bisogna allenarle singolarmente, una per una.

3.1.2 Fast R-CNN

Dal momento che la rete R-CNN risulta essere molto costosa computazionalmente, principalmente a causa dell'estrazione delle regioni d'interesse che successivamente vengono classificate da una CNN, è stata sviluppata la rete Fast R-CNN [14], in grado di effettuare prima un'estrazione delle caratteristiche dall'immagine, per poi individuare le regioni d'interesse. In Fast R-CNN, i primi strati convoluzionali hanno lo scopo di evidenziare le caratteristiche più importanti dell'immagine, così da rendere più efficiente la determinazione delle regioni d'interesse. Per la classificazione, al posto delle SVM, viene utilizzato un singolo classificatore softmax.

3.1.3 Faster R-CNN

Successivamente viene proposta la rete Faster R-CNN [15], in grado di migliorare Fast R-CNN modificando il processo di selezione delle regioni d'interesse. I metodi descritti sinora utilizzavano un algoritmo di selective search, esterno alla rete neurale, che non poteva, quindi, essere allenato. In Fast R-CNN erano

stati implementati i primi strati in modo che si occupassero di estrarre la mappa delle caratteristiche dell'immagine, per poi andare ad individuare le regioni d'interesse. Faster R-CNN propone, invece, la Region Proposal Network (RPN), che permette di eliminare l'algoritmo Selective Search e di implementare tutte le componenti all'interno della rete, senza più richiedere contributi esterni.
La rete Region Proposal Network (RPN) prende in input un'immagine e restituisce in output un insieme di regioni proposte con associata la probabilità che un oggetto specifico si trovi effettivamente in tale area. Si utilizza il termine "regione" per identificare un'area avente dimensione rettangolare in cui esiste una certa probabilità di rilevare un oggetto.

3.1.4 Deep Residual Networks

Le reti neurali formate da molti strati hanno l'inconveniente che, aumentando la profondità della rete, l'accuratezza del modello diminuisce, con un conseguente aumento dell'errore di training. Per risolvere questo problema, nel 2015, vengono introdotte le reti neurali residuali, ovvero le Residual Neural Networks. Finora le CNN descritte prendevano un input x, questo veniva sottoposto a diverse operazioni all'interno del layer, e veniva restituito un valore in output $F(x)$, che veniva a sua volta dato come input al layer successivo, e così via. Nelle reti neurali residuali invece, come suggerisce il nome, l'output assume il valore del suo residuo ($H(x) = F(x) + x$). In questo modo, è stato possibile aumentare in modo significativo la "profondità" di una rete neurale, senza compromettere l'accuratezza dell'intero modello. Questo tipo di reti sono state integrate nei modelli di Detectron2 (che verrano mostrati successivamente) così da migliorarne le prestazioni.

3.2 Detectron2

3.2.1 Il nuovo framework della Facebook AI Research

Dal suo rilascio nel 2018, Detectron è diventato uno dei progetti open source più adottati di Facebook AI Research (FAIR). Per costruire e far progredire questo

progetto, è stata rilasciata la seconda generazione della libreria, che presenta importanti miglioramenti sia per la ricerca, che per l'uso in produzione. Detectron2 è una riscrittura da zero di Detectron, implementata in PyTorch con un nuovo design più modulare: Detectron2 risulta, infatti, flessibile ed estensibile, ed è in grado di supportare un training veloce su macchine GPU singole o multiple. Detectron2 include implementazioni di alta qualità di algoritmi di rilevamento degli oggetti all'avanguardia, tra cui DensePose, panoptic feature pyramid networks e numerose varianti della pionieristica Mask R-CNN, anch'essa sviluppata da FAIR. Il suo design estensibile rende facile l'implementazione di progetti di ricerca all'avanguardia, senza dover forzare l'intera base di codice.

Il progresso nell'IA è uno sforzo comunitario che include privati, università e industria. I problemi che FAIR mira a risolvere vanno ben oltre ciò che ogni individuo o gruppo può raggiungere da solo. Per questo motivo, tale gruppo crede fortemente nella condivisione del codice, in modo da incentivare la ricerca riproducibile, la sperimentazione rapida e lo sviluppo di nuove idee. Rilasciando Detectron2, essi sperano di accelerare ulteriormente la ricerca nelle aree di rilevamento degli oggetti, segmentazione e comprensione della posa umana.

Una nuova ricerca inizia con la comprensione, la riproduzione e la verifica dei risultati precedenti in letteratura. Con Detectron2, si mira a fornire implementazioni di riferimento di alta qualità per molti algoritmi all'avanguardia, al fine di democratizzare questa fase del processo di ricerca. Il design modulare della libreria permette anche ai ricercatori di implementare nuovi progetti con una separazione netta dalla funzionalità standard della libreria di rilevamento. Per esempio, Mesh R-CNN, il recente lavoro di FAIR sulla previsione di maglie 3D per istanza di oggetto da immagini 2D, è stato sviluppato in Detectron2. Il design modulare di Detectron2 ha permesso ai ricercatori di estendere facilmente Mask R-CNN per lavorare con complesse strutture di dati che rappresentano mesh 3D, integrare nuovi set di dati e progettare nuove metriche di valutazione[1].

[1] https://ai.facebook.com/blog/-detectron2-a-pytorch-based-modular-object-detection-library-/

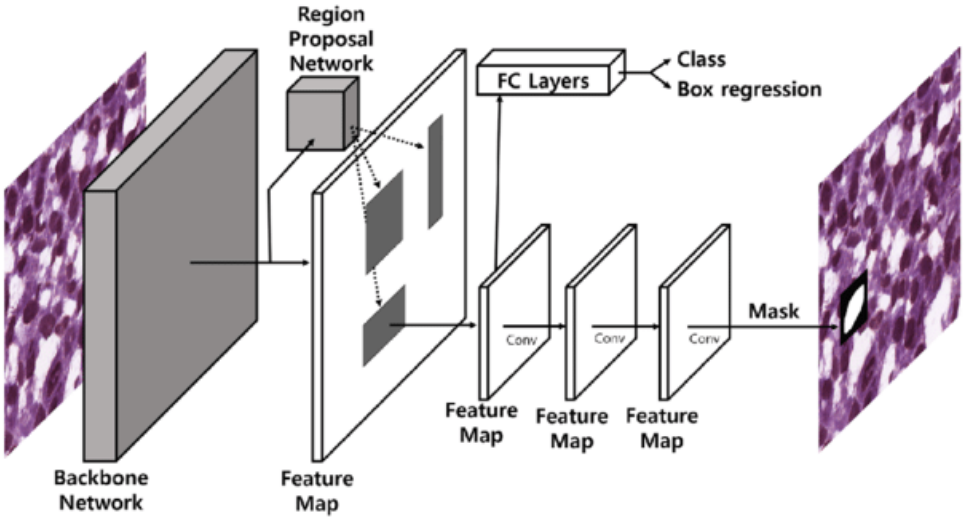

Figura 3.2: *Architettura Mask R-CNN*[16].

3.2.2 Mask R-CNN

Tra i possibili modelli proposti dal framework Detectron2[2], viene descritta la Mask R-CNN, la quale ha come variante la Keypoint R-CNN la quale verrà utilizzata per la soluzione proposta da questo elaborato. La rete Mask R-CNN [16], introdotta come estensione di Faster R-CNN, presenta un componente aggiuntivo in grado di predire la maschera di un'istanza. Mentre in Faster R-CNN erano presenti solo il regressore dei bounding box ed il classificatore, Mask R-CNN presenta, comunque, questi due componenti, ma, in parallelo, prevede anche la maschera dell'istanza. In particolare, per ogni Region Of Interest (RoI), Mask R-CNN genera una maschera di dimensione 28x28 che viene espansa per adattarsi alle dimensioni del bounding box corrispondente. Mask R-CNN restituisce in output, oltre alla classe d'appartenenza e al bounding box, anche una maschera binaria da applicare al bounding box corrispondente. In Figura 3.2 è raffigurata l'architettura di una Mask R-CNN utilizzata in campo biomedico per segmentare i nuclei direttamente su immagini istopatologiche[3].

[2] https://github.com/facebookresearch/detectron2/blob/master/MODEL_ZOO.md

[3] https://www.researchgate.net/figure/The-overall-network-architecture-of-Mask-R-CNN_fig1_336615317

3.2.3 Backbone e Feature Pyramid Network

Mask R-CNN, presente in Detectron2, è stato riproposto con diverse versioni di backbone utili per estrarre le feature, per esempio ResNet e ResNeXt, entrambe con 50 o 101 layer, in cui l'estrazione delle feature avviene a partire dall'ultimo layer convoluzionale, appartenente al quarto o al quinto blocco. Per migliorare l'accuratezza di queste feature è stata inserita una seconda rete convoluzionale, poszionata dopo il primo backbone: la Feature Pyramid Network (FPN). Quest'ultima, grazie alla sua struttura piramidale, permette l'estrazione di diverse feature aventi differenti scale. Questo approccio combinato di ResNet e FPN ha permesso un miglioramento sia della precisione che della velocità di estrazione delle feature.

3.3 Indici di interesse clinico

3.3.1 Scale di valutazione

I clinici che seguono pazienti affetti da patologie neurologiche, come Ictus e SLA, hanno bisogno di valutare l'evoluzione della malattia nel tempo. Questo avviene attraverso visite periodiche, di solito con cadenza mensile, durante le quali il clinico utilizza delle scale di valutazione per analizzare, attraverso dei task compiuti dal paziente, l'evoluzione della malattia.

3.3.2 Utilizzo di interpolazioni geometriche

I clinici che utilizzano le suddette scale valutano uno specifico item per ogni task svolto dal paziente. Questo significa che, con l'aumentare dei task da svolgere, il paziente si stanchi, influenzando negativamente la valutazione. Purtroppo non sono stati ancora realizzati algoritmi con approccio Deep Learning in grado di restituire direttamente valutazioni cliniche corrispondenti a delle scale. Questo è dovuto al fatto che esistono pochi dataset disponibili contenenti immagini facciali di popolazioni con patologie neurologiche; un'altra motivazione è la complessità implicita del task, nonché la difficoltà nel validare clinicamente una soluzione di

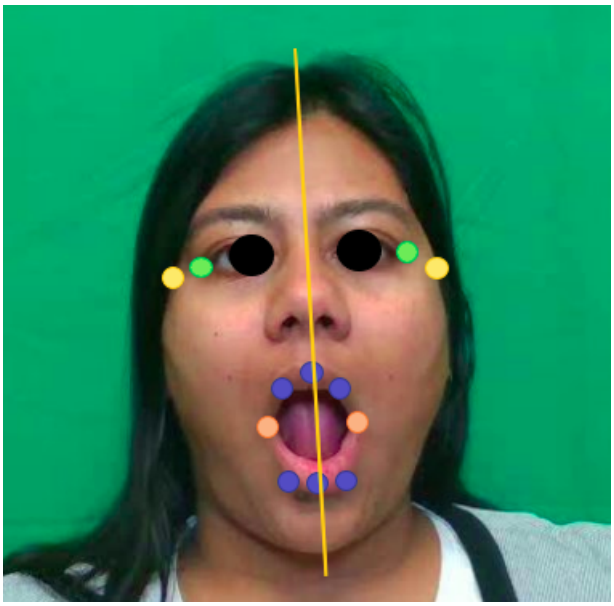

Figura 3.3: *Illustrazione dimostrativa dell'utilizzo di alcuni landmark facciali per ricavare informazioni cliniche rilevanti. L'asse in giallo rappresenta l'asse di simmetria verticale del volto, i punti blu e arancioni sono utili per calcolare indici relativi alla muscolatura boccale e i punti verdi servono per calcolare la distanza di riferimento.*

questo tipo. Per questo motivo, nel presente elaborato, viene presentato un possibile indice clinico basato su un approccio geometrico che considera la posizione dei landmark predetti sul volto.

Come mostrato in Figura 3.3, questo approccio geometrico si basa sui seguenti fondamentali:

- **Distanza di riferimento**: corrisponde alla distanza tra i punti più esterni dei due occhi (in verde). Gli indici potrebbero utilizzare questa misura, che si suppone immutabile nel tempo, come riferimento valido. Inoltre, con l'inserimento una tantum di questa distanza effettiva in millimetri da parte del clinico, si potranno ricavare dei valori assoluti in millimetri.
- **Asse di simmetria verticale del volto**: utile per rilevare una posizione del volto inclinata. Può essere utilizzato dagli indici che necessitano di

valutare la simmetria del volto durante un task. A partire dai punti gialli, si calcola il punto medio tra essi e l'asse perpendicolare passante per il punto medio del segmento (rappresentato in giallo).

A partire dalla distanza di riferimento e dall'asse di simmetria verticale del volto è possibile realizzare diversi tipi di indici, per esempio: indici per valutare la simmetria di alcuni landmark specifici (variabili a seconda della patologia), indici per valutare misure assolute come massima apertura buccale (punti blu) e massimo stiramento delle labbra (punti arancioni), o indici che monitorano la stanchezza analizzando le variazioni di inclinazione del volto.
L'approccio descritto è presentato solamente a livello teorico. Infatti, la conversione delle suddette scale in indici di interesse clinico è un lavoro complesso che deve essere svolto in collaborazione con uno o più reparti clinici specializzati e che necessita di una validazione clinica prima di poter essere effettivamente utilizzato.

Capitolo 4

Dataset e Protocollo Sperimentale

Il seguente capitolo descrive i dataset utilizzati per le fasi di training e i modelli utilizzati nei diversi esperimenti eseguiti.

4.1 Dataset Neuroface

Il dataset realizzato dagli autori dell'articolo [1] presenta trentasei partecipanti: 11 pazienti con SLA (4 uomini, 7 donne), 14 pazienti con Ictus (10 uomini, 4 donne) e 11 soggetti sani (7 uomini, 4 donne). Tutti i partecipanti, che non erano cognitivamente compromessi, quindi con un punteggio *Montreal Cognitive Assessment score*[1] superiore a 26, hanno superato uno screening dell'udito. I pazienti con SLA sono stati diagnosticati secondo i criteri di *El Escorial Criteria* della World Federation of Neurology[2]. Nove partecipanti avevano sintomi spinali all'esordio, mentre due partecipanti presentavano una SLA ad insorgenza bulbare. Lo studio è stato approvato dai comitati etici di ricerca presso il Sunnybrook Research Institute e UHN: Toronto Rehabilitation Institute. Tutti i partecipanti hanno firmato un consenso secondo i requisiti della Dichiarazione di Helsinki[3],

[1] https://en.wikipedia.org/wiki/Montreal_Cognitive_Assessment
[2] https://pubmed.ncbi.nlm.nih.gov/7807156/
[3] https://it.wikipedia.org/wiki/Dichiarazione_di_Helsinki

permettendo, quindi, la condivisione dei media realizzati per la ricerca.
Ad ogni soggetto è stato chiesto di eseguire una serie di compiti vocali e compiti non vocali comunemente usati durante un esame clinico. Essi comprendevano:

- 10 ripetizioni della frase: "Buy Bobby a Puppy" a una velocità e ad un volume di voce confortevoli (BBP);
- ripetizioni della sillaba /pa/ il più velocemente possibile, in un solo respiro (PA);
- ripetizioni delle sillabe /pataka/ il più velocemente possibile, in un solo respiro (PATAKA);
- arricciare le labbra (per esempio fingere di soffiare una candela 5 volte e fingere di baciare un bambino 5 volte - BLOW e KISS);
- massima apertura della mascella 5 volte (OPEN);
- fingere di sorridere con labbra strette 5 volte (SPREAD);
- fare un grande sorriso 5 volte (BIGSMILE);
- alzare le sopracciglia 5 volte (BROW).

Durante i compiti, i partecipanti erano seduti di fronte alla telecamera, con una distanza viso-camera tra i 30 e i 60 cm. Una fonte di luce continua è stata posta dietro la telecamera per illuminare uniformemente il viso. Sono state realizzate un totale di 261 registrazioni video che sono state incluse nel dataset: 80 da soggetti sani, 76 da pazienti con SLA e 105 da pazienti con Ictus.
Dalla totalità dei video sono stati estratti 3306 fotogrammi (1015 soggetti sani, 920 pazienti SLA e 1371 pazienti Ictus). Su questi fotogrammi, i ground truth dei 68 landmark facciali sono stati annotati seguendo la configurazione della Figura 2.1. Per ogni compito non vocale, sono stati considerati 3 fotogrammi per ripetizione:

1. inizio del gesto (cioè, posizione di riposo);
2. picco del gesto (ad esempio, massima apertura della mascella);

3. punto medio tra i due precedenti.

Per gli esperimenti effettuati con questo dataset, sono stati creati 3 set: uno di training, uno di validazione e uno di testing. Il dataset di training conteneva fotogrammi appartenenti a 25 soggetti, il set di validazione conteneva fotogrammi appartenenti a 5 soggetti e il set di test conteneva fotogrammi appartenenti a 4 soggetti.
Per aumentare la numerosità del dataset *Neuroface* [1] in fase di training sono state attuate delle tecniche di data augmentation, riuscendo a triplicare la numerosità del dataset, passando da circa 2.300 fotogrammi per il training a circa 7000. Le modifiche applicate sono state: variare la luminosità delle immagini in maniera casuale ed invertire l'immagine lungo l'asse verticale (invertendo anche le relative annotazioni).

4.2 Dataset 300W

Per far fronte alla scarsità di fotogrammi forniti dal dataset *Neuroface* [1], e più in generale, dai dataset nell'ambito delle patologie neurologiche, in questo elaborato è stato deciso di utilizzare il dataset *300W*[4], menzionato da diversi articoli allo stato dell'arte [6], per effettuare un pre-training sul modello per poi utilizzare il dataset *Neuroface* per effettuare fine-tuning. *300W* è un dataset realizzato per migliorare la ricerca nell'ambito del riconoscimento dei landmark facciali su immagini e video. Il focus del dataset è sull'annotazione di immagini del mondo reale con tutti i suoi inconvenienti precedentemente descritti (luce, occlusioni, ecc...). A partire da questo dataset, sono state realizzate delle vere e proprie "challange" per confrontare i vari tipi di approcci proposti dallo stato dell'arte.
La versione del dataset utilizzata per questo elaborato contiene più di 100 video, con una media di circa 2.000 frame per video, rappresentanti casi del mondo reale in cui, salvo qualche occlusione parziale, il volto da rilevare è ben visibile come, ad esempio, video estrapolati da interviste, monologhi di politici, spiegazioni di

[4] https://ibug.doc.ic.ac.uk/resources/300-W/

professori, ecc.

Per ridurre il carico computazionale, come suggerito nell'articolo [17], è stato effettuato un sottocampionamento dei video, selezionando 1 frame ogni 5, così da avere un totale da più di 40.000 annotati.

4.3 Protocollo Sperimentale

4.3.1 Modelli forniti da Detectron2

Il framework Detectron2 fornisce una serie di modelli, descritti nella pagina GitHub della repository ufficiale[5]. Questi modelli sono stati allenati con il dataset ImageNet[6].

Per il task di facial landmark detection si è deciso di utilizzare i modelli presenti nella categoria "COCO Person Keypoint Detection Baselines with Keypoint R-CNN" i quali sono stati realizzati per effettuare la stima della posa di immagini rappresentati delle persone [16].

L'idea è, quindi, quella di sfruttare la struttura di Keypoint R-CNN per predire la posizione dei landmark facciali. Per far questo, il bounding box e la classe predetta corrisponderanno al volto della persona invece che al corpo e, a partire da questo bounding box, la Keypoint R-CNN restituirà in output le coordinate (x, y) dei landmark facciali.

4.3.2 Configurazione dei parametri

Tra i parametri configurabili per il training di ogni modello, vengono descritti i più rilevanti:

- **Congelamento pesi (CP)**: il backbone dei modelli forniti da Detectron2 sono suddivisi in 5 strati, ognuno contenente numerosi layer. In questo modo è possibile congelare i pesi per un numero arbitrario di strati. Di default gli sviluppatori di Detectron2 hanno impostato tale numero pari a

[5] https://github.com/facebookresearch/detectron2/blob/master/MODEL_ZOO.md
[6] https://www.image-net.org/

2, in modo che i primi strati fossero congelati, così da effettuare un fine-tuning sugli altri. Al contrario, fissando a zero questo parametro si ottiene un training ex novo;

- **Base learning rate (LR)**: corrisponde al learning rate di partenza. Di default gli sviluppatori di Detectron2 lo hanno impostato a 0,001;
- **Batch size (BS)**: corrisponde al numero di immagini che vengono prelevate dal dataset per effettuare un'iterazione. Di default gli sviluppatori di Detectron2 lo hanno impostato a 512;
- **Max iter e epoche(MI)**: il max iter corrisponde al numero massimo di iterazioni. Questo parametro, combinato con la batch size e con il numero di immagini del training set, ci fornirà il numero di epoche attraverso la seguente formula: $N_{ep} = \frac{I_{max}Bs}{N_{tr}}$ dove N_{ep} è il numero di epoche, $I_{max}Bs$ è il prodotto tra il numero massimo di iterazioni e la batch size ed, infine, N_{tr} è la numerosità delle immagini di training;
- **Numero di keypoint**: corrisponde al numero di keypoint con il quale si vuole allenare Keypoint R-CNN. Il massimo suggerito dagli sviluppatori di Detectron2 era impostato a 17, ma con alcune modifiche nel codice è stato portato a 68, ovvero il numero di landmark facciali di interesse per perseguire l'obiettivo della tesi.

4.3.3 Metriche di valutazione

Per valutare la precisione con la quale vengono predetti i landmark facciali in fase di testing, è stata utilizzata la metrica del Normalized Mean Error descritta nel Capitolo 2, che usa come parametro d di normalizzazione la diagonale del bounding box che individua il volto, come mostrato nella seguente formula:

$$NME = \frac{1}{K}\sum_{k=1}^{k} NME_k \rightarrow NME_k = \left[\frac{1}{N_L}\sum_{i=1}^{N_L}\frac{\sqrt{(x_i - x_p)^2 + (y_i - y_p)^2}}{Diag_{bbox}}\right] \cdot 100$$

dove (x_i,y_i) sono le coordinate del ground truth, (x_p,y_p) sono le coordinate predette dal modello, $Diag_{bbox}$ è il modulo della diagonale del bounding box contenente

il volto del soggetto, N_L è il numero di landmark e K è il numero di immagini contenute nel dataset. Inoltre, per ottenere maggiori informazioni riguardo la localizzazione dell'errore di predizione nel volto, quest'ultimo è stato suddiviso per regioni. Sono state individuate, quindi, cinque regioni: mento, naso, sopracciglia, occhi e bocca, suddivise per landmark come mostrato in Figura 4.1.
Nel seguente elaborato vengono proposte le seguenti variazioni della metrica NME:

- NME_{diag}: rappresenta il valore dell'NME calcolato sulla totalità dei 68 landmark;
- NME_{chin}: rappresenta il valore dell'NME calcolato sui 17 landmark del mento;
- $NME_{eyebrows}$: rappresenta il valore dell'NME calcolato sui 10 landmark delle sopracciglia;
- NME_{nose}: rappresenta il valore dell'NME calcolato sui 9 landmark del naso;
- NME_{eyes}: rappresenta il valore dell'NME calcolato sui 12 landmark degli occhi;
- NME_{mouth}: rappresenta il valore dell'NME calcolato sui 20 landmark della bocca.

4.3.4 Ablation study

Per prima cosa, è stata effettuata un'indagine esplorativa per decidere quale modello utilizzare tra i 4 presenti nella categoria "COCO Person Keypoint Detection Baselines with Keypoint R-CNN"[7]. Il dataset utilizzato per il training è il *Neuroface* su cui non è stata utilizzata alcuna tecnica di data augmentation, pertanto il numero di immagini presenti è di circa 3300. La configurazione dei parametri è mostrata nella prima sezione della Tabella 4.1.

[7] https://github.com/facebookresearch/detectron2/blob/master/MODEL_ZOO.md

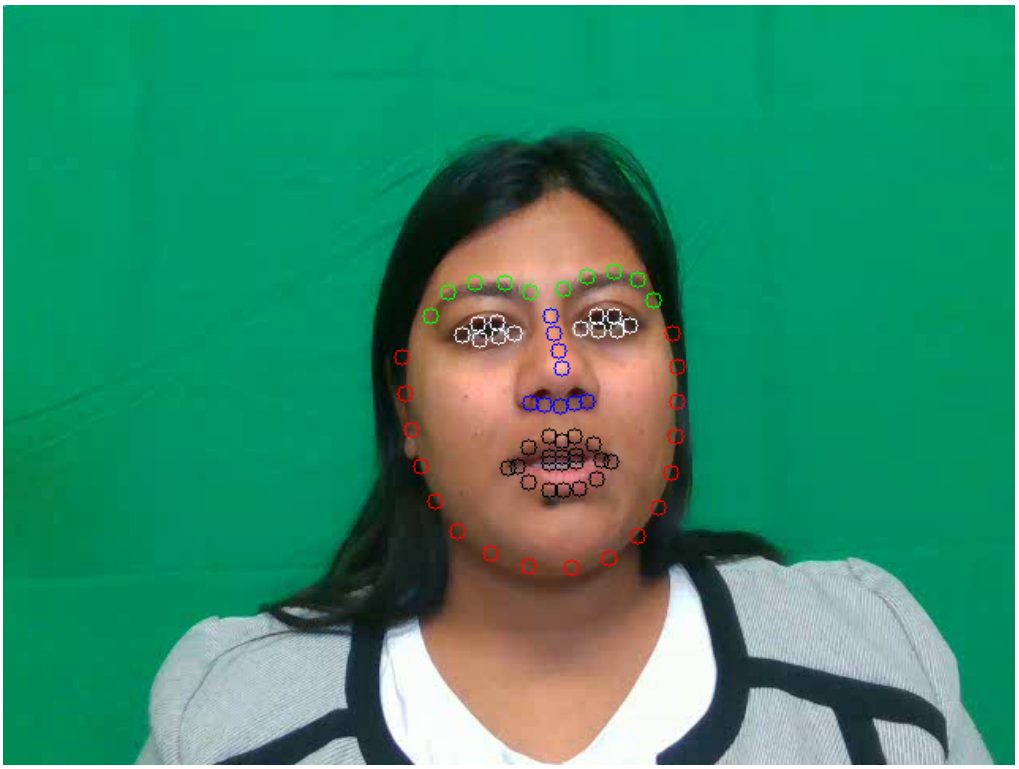

Figura 4.1: *Suddivisione delle cinque regioni del volto: mento, naso, sopracciglia, occhi e bocca*

In seguito, i modelli che hanno ottenuto migliori risultati (R50-FPN-1x e R101-FPN-3x) sono stati allenati, mantenendo base learing rate invariato, ma con il congelamento dei pesi impostato a zero (training ex novo), la batch size a 128, il numero massimo di iterazioni a 25000 e con il dataset *Neuroface* ampliato grazie alle tecniche di data augmentation, quindi il triplo delle immagini (circa 9900). La configurazione dei parametri è mostrata nella seconda sezione della Tabella 4.1, in corrispondenza dei modelli A-R50-FPN-1x e A-R101-FPN-3x.
Come spiegato precedentemente, i modelli forniti dagli sviluppatori di Detectron2, appartenenti alla categoria COCO Person Keypoint Detection Baselines with Keypoint R-CNN, non sono stati ideati per il task di facial landmark detection. La Keypoint R-CNN è stata realizzata per il task di stima della posa, per individuare, quindi, all'interno del bounding box, un numero massimo di 17 keypoint. Per questo motivo, oltre che allenare i modelli da zero, è stato deciso di

modificare la risoluzione della maschera, aggiungendo un modulo di convoluzione trasposta 2D come ulteriore strato alla Keypoint R-CNN. Quindi è stato avviato un esperimento configurato come il precedente per testare quest'ultima modifica. La configurazione dei parametri è mostrata nella terza sezione della Tabella 4.1, in corrispondenza del modello A-R50-FPN-1x-M.

Infine, utilizzando il backbone R50-FPN-1x e conservando la suddetta all'ultimo strato della Keypoint R-CNN, è stato eseguito un training utilizzando il dataset *300W*. La configurazione dei parametri è rimasta pressochè invariata, sono stati cambiati solo il base learning rate, impostato a 0,005, e il numero di iterazioni massime, impostato a 100.000, per un totale di 400 epoche (considerando batch size a 128 e numero di immagini di training pari a 31.000). Successivamente è stato effettuato del fine-tuning, a partire dai pesi ottenuti dal modello allenato con il dataset 300W, utilizzando il dataset *Neuroface* ampliato. Per questo esperimento è stato impostato il congelamento dei pesi a 2, in modo che i primi due strati di backbone non venissero modificati, e il base learning rate a 0,001. La configurazione dei parametri è mostrata nell'ultima sezione della Tabella 4.1.

Modello	Cp	Lr	Bs	Mi	Dataset
R50-FPN-1x	1	0,01	256	5.000	Neuroface
R50-FPN-3x	1	0,01	256	5.000	Neuroface
R101-FPN-3x	1	0,01	256	5.000	Neuroface
X101-FPN-3x	1	0,01	256	5.000	Neuroface
A-R50-FPN-1x	0	0,01	128	25.000	A-Neuroface
A-R101-FPN-1x	0	0,01	128	25.000	A-Neuroface
A-R50-FPN-1x-M	0	0,01	128	25.000	A-Neuroface
300W-R50-FPN-1x-M	0	0,005	128	100.000	300W
300W-NF-R50-FPN-1x-M	2	0,001	128	25.000	300W-A-Neuroface

Tabella 4.1: *La prima colonna mostra il modello utilizzato, la seconda (Cp) il parametro di congelamento dei pesi, la terza (lr) il base learning rate utilizzato, la quarta (Bs) la batch size scelta, la quinta (Mi) il numero massimo di interazioni richieste al modello e l'ultima (Dataset) descrive il dataset utilizzato. Si precisa che la presenza di "A-" sia nel modello, che nel dataset, indica l'attuazione di tecniche di data augmentation sul set di training.*

Capitolo 5

Risultati

Nel seguente capitolo vengono discussi i risultati ottenuti dagli esperimenti mostrati nel capitolo precedente e i possibili sviluppi futuri.

5.1 Confronto tra i risultati ottenuti

La prima indagine, descritta nel precedente capitolo, ha permesso di individuare i due modelli più promettenti con il quale proseguire, poi, con gli esperimenti. Questi modelli presentano tutti una struttura simile. Il loro backbone, infatti, è formato da ResNet con 50 layer, ResNet con 101 layer, ResNeXt con 101 layer. La distinzione tra i 4 modelli è data dalla tipologia specifica di backbone, con l'aggiunta di un'ulteriore suddivisione di ResNet50 in due modelli pre-allenati con un numero di epoche diverse, indicati con "lr sched" uguale a 1x e 3x. Come

	NME					
Modello	diag	chin	eyebrows	nose	eyes	mouth
R50-FPN-1x	**7,84**	12,66	0,06	2,02	7,46	7,31
R50-FPN-3x	8,00	13,96	0,04	2,69	5,74	8,52
R101-FPN-3x	**7,92**	12,96	0,05	1,33	4,94	8,62
X101-FPN-3x	8,04	14,52	0,04	2,22	7,24	3,45

Tabella 5.1: *Indagine esplorativa eseguita sui 4 modelli forniti dalla categoria COCO Person Keypoint Detection Baselines with Keypoint R-CNN.*

	NME					
Modello	diag	chin	eyebrows	nose	eyes	mouth
A-R50-FPN-1x	**5,52**	8,09	0,07	1,6	6,78	3,54
A-R101-FPN-3x	7,32	13,16	0,84	1,61	8,85	3,45

Tabella 5.2: *Risultati ottenuti dai due migliori modelli in Tabella 5.1, a seguito di un training effettuato con il dataset Neuroface ampliato con tecniche di data augmentation.*

si può vedere in Tabella 5.1, i backbone selezionati sono R50-FPN-1x e R101-FPN-3x. Si è deciso di conservare un backbone con 50 layer intermedi e uno con 101. Tra i due modelli con backbone con 50 layer, si è deciso di tenere il più preciso, ovvero quello con NME_{diag} pari a 7,84, mentre il backbone X101-FPN-3x, oltre ad essere il più impreciso (NME_{diag} pari a 8,04), è anche il più oneroso computazionalmente, per questo è stato scartato e scelto R101-FPN-3x.
In Tabella 5.2 sono mostrati i risultati relativi al confronto dei due precedenti modelli selezionati, allenati entrambi sul dataset di *Neuroface* ampliato con tecniche di data augmentation. I risultati evidenziano come il backbone R50-FPN-1x, essendo meno complesso, sia migliorato in precisione a seguito di un training con maggior numero di immagini (il triplo rispetto al precedente esperimento) e maggior numero di iterazioni (il quintuplo rispetto al precedente esperimento). Questo esperimento presenta un divario di quasi 2 punti percentuali tra i NME_{diag} dei due modelli, e questo è dovuto alla minor complessità del modello R50-FPN-1x. Tale risultato ha come conseguenza il fatto che tale modello sia più adattabile al cambio di task, da quello di stima della posa a quello di rilevamento dei landmark facciali.
La successiva modifica applicata alla Keypoint R-CNN (per aumentare la risoluzione della maschera) ha condotto ad un miglioramento significativo della precisione e il relativo esperimento, infatti, ha dato come risultato un NME_{diag} di 4,54. Il modulo convoluzionale trasposto 2D aggiunto all'ultimo strato della Keypoint R-CNN fornisce alla maschera di predizione dei keypoint una risoluzione maggiore, in modo che la predizione della posizione dei landmark risulti più

	NME					
Modello	diag	chin	eyebrows	nose	eyes	mouth
300W-R50-FPN-1x-M	14,85	27,94	0,16	3,38	14,59	8,16
300W-NF-R50-FPN-1x-M	**8,86**	15,02	0,11	2,57	7,52	6,34

Tabella 5.3: *Risultati del modello allenato con 300W e dello stesso modello con fine-tuning su Neuroface.*

accurata. Per questo motivo, nell'ultimo esperimento si è deciso di mantenere questa modifica.

In tale esperimento, i modelli sono stati allenati e, successivamente, testati sul test set di *Neuroface*. In Tabella 5.3 sono riportati i risultati di quest'ultimo esperimento: la precisione è diminuita notevolmente, ma questo è dovuto alla generalizzazione fornita dal dataset *300W*. Infatti, a seguito del fine-tuning, la precisione è diminuita nuovamente, con un NME_{diag} che è passato da 14,85 prima del fine-tuning a 8,86 a seguito di quest'ultimo, conservando, in parte, la capacità di generalizzare.

Inoltre, si nota come l'errore di precisione più elevato lo si ha in corrispondenza delle regioni di mento e bocca, le quali sono le parti che muscolarmente risentono di più dell'impatto della malattia. Per questo motivo, si ritiene che si potrebbe mediare questo problema aumentando la numerosità del dataset clinico con il quale effettuare fine-tuning. Al contrario si nota come le regioni di sopracciglia e naso abbiano una precisione molto elevata e, per questo motivo, potrebbero essere utilizzate in fase di post-processing come punti di riferimento per mediare eventuali errori ottenuti sugli altri landmark.

5.2 Conclusioni

Nei capitoli di questa tesi sono state introdotte ed affrontate le metodologie impiegate per valutare lo stato clinico dei pazienti affetti da patologie neurologiche, in particolare da Ictus e SLA. Dopo una breve panoramica di queste patologie,

sono state approfondite le attività relative alla valutazione della muscolatura facciale, per poi presentare lo stato dell'arte relativo al rilevamento di landmark facciali. Nel terzo capitolo sono stati descritti i metodi utilizzati per la detection di oggetti e keypoint tramite l'utilizzo di algoritmi di intelligenza artificiale. In seguito, sono stati descritti i dataset e i modelli, con le relative configurazioni di parametri utilizzati per le fasi di training. Infine sono stati mostrati e discussi i risultati ottenuti.

I modelli forniti dagli sviluppatori di Detectron2 sono stati realizzati per risolvere il task di stima della posa. Questo comporta che i primi esperimenti realizzati abbiano errori significativi. Tuttavia, con l'aumentare del numero di iterazioni e della numerosità del dataset, è stato possibile migliorare l'efficacia del modello in modo significativo. Nonostante i frame del dataset *Neuroface* siano stati suddivisi in training-evaluation-test per pazienti, le caratteristiche generali del setup di acquisizione dei video sono identiche, il che significa che il modello risulterà più preciso ma meno robusto ad un cambiamento del setup di acquisizione (quindi ad una generalizzazione). L'ultimo esperimento, che consisteva nell'effettuare un training sul dataset *300W*, per poi eseguire del fine-tuning con il dataset *Neuroface*, è stato utile per cercare di generalizzare il modello, anche se ciò ha portato ad un lieve degrado delle prestazioni in termini di precisione.

In generale, la precisione ottenuta finora non è ottima, ma questo non significa che una soluzione come questa non possa essere utilizza come strumento di supporto per clinici, dal momento che, in ogni caso, le valutazioni "occhiometriche" compiute attualmente dai clinici sono comunque imprecise ed, oltretutto, molto soggettive. Tuttavia, uno strumento come quello presentato permetterebbe di uniformare quello che è lo standard clinico, nonostante la precisione non sia, ancora, elevata.

Il lavoro presentato in questa tesi necessita di diversi stadi di sviluppo prima di poter esser candidato alla produzione. Infatti, deve essere realizzata l'ultima parte della pipeline dell'intero sistema, ovvero l'estrazione degli indici di interesse clinico. Oltretutto, il dataset clinico utilizzato ha una numerosità molto bassa, e questo si pone come un limite all'incremento delle prestazioni. Per potere mi-

gliore le performance del sistema è necessario avere a disposizione un dataset più ampio, altrimenti il modello non riuscirà ad essere sufficientemente robusto a cambiamenti sostanziali delle espressioni compiuti dai pazienti.
Per quanto riguarda il miglioramento dei processi sviluppati finora, c'è bisogno di implementare una serie di algoritmi da attuare in fase di post-processing, in seguito alla predizione dei landmark, per mediare eventuali imprecisioni compiute dal modello. Inoltre, gli algoritmi presentati fin'ora devono essere ottimizzati, con l'obiettivo di renderli più efficienti.
Dal punto di vista del lato clinico, è importante riuscire a realizzare una conversione delle scale di valutazione utilizzate attualmente dai clinici in indici di rilevanza clinica che sfruttino la posizione dei landmark predetti dal modello, così da poter, poi, validare l'intero sistema.
In ogni caso, i risultati raggiunti in questo elaborato di tesi dimostrano che le tecnologie moderne sono pronte per supportare i clinici nelle le valutazioni compiute durante le visite, anche a distanza. Quest'ultimo aspetto è molto importante, basti pensare alle difficoltà compiute da persone affette da queste patologie per spostarsi verso il centro dove effettuare la visita. Tecnologie come quella presentata in questa tesi, oltre a rendere le prognosi oggettive, rendono meno frustranti le visite dei pazienti, permettendo lo svolgimento di alcune di queste da remoto e fornendo ai clinici resoconti da consultare generati in automatico senza che il clinico assista in diretta durante l'esecuzione dei task.

Elenco delle figure

Bibliografia

[1] IEEE Sia Rezaei Diego Guarın Madhura Kulkarni Derric Lim Mark I. Boulos Lorne Zinman Yana Yunusova Andrea Bandini, Member and Babak Taati. A new dataset for facial motion analysis in individuals with neurological disorders. 2020.

[2] Nataliya Boyko, Oleg Basystiuk, and Nataliya Shakhovska. Performance evaluation and comparison of software for face recognition, based on dlib and opencv library. In *2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP)*, pages 478–482, 2018.

[3] B. van Ginneken, A.F. Frangi, J.J. Staal, B.M. ter Haar Romeny, and M.A. Viergever. Active shape model segmentation with optimal features. *IEEE Transactions on Medical Imaging*, 21(8):924–933, 2002.

[4] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[5] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2610–2617, 2012.

[6] Kostiantyn Khabarlak and Larysa Koriashkina. Fast facial landmark detection and applications: A survey. *CoRR*, abs/2101.10808, 2021.

[7] Li Zhang, Guan Gui, Abdul Mateen Khattak, Minjuan Wang, Wanlin Gao, and Jingdun Jia. Multi-task cascaded convolutional networks based intelligent fruit detection for designing automated robot. *IEEE Access*, 7:56028–56038, 2019.

[8] Xiaojie Guo, Siyuan Li, Jinke Yu, Jiawan Zhang, Jiayi Ma, Lin Ma, Wei Liu, and Haibin Ling. Pfld: A practical facial landmark detector. 2019.

[9] Weijian Li, Yuhang Lu, Kang Zheng, Haofu Liao, Chihung Lin, Jiebo Luo, Chi-Tung Cheng, Jing Xiao, Le Lu, Chang-Fu Kuo, and Shun Miao. Structured landmark detection via topology-adapting deep graph learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 266–283, Cham, 2020. Springer International Publishing.

[10] Yaohui Wang, Antitza Dantcheva, Jean-Claude Broutart, Philippe Robert, Francois Bremond, and Piotr Bilinski. Comparing methods for assessment of facial dynamics in patients with major neurocognitive disorders. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[11] S. Priebe M. Bishay, P. Palasek and I. Patras. Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis. *CoRR*, abs/1808.02531, 2018.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[13] T. Gevers J.R.R. Uijlings, K.E.A. van de Sande† and A.W.M. Smeulders. Selective search for object recognition. 2013.

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[15] Huaizu Jiang and Erik Learned-Miller. Face detection with the faster r-cnn. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 650–657, 2017.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[17] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Ringraziamenti

Vorrei ringraziare *in primis* la mia famiglia, che mi è stata sempre vicina in questi anni di università che solo grazie a loro ho avuto la possibilità di frequentare.
Ringrazio i miei compagni di corso che mi hanno supportato e accompagnato durante questo percorso, rendendolo più piacevole. Tra questi ringrazio Pietro, con il quale, oltre all'aspetto universitario, mi sono allenato a corpo libero durante questi due anni, raggiungendo eccellenti risultati.
In particolar modo ringrazio la mia fidanzata Chiara Mazzucchelli, con la quale ho avuto la fortuna di condividere questo percorso di studi, così da poterci aiutare a vicenda negli studi e nella realizzazione di diversi progetti. Mi è sempre stata accanto, per qualsiasi problema, per qualsiasi esame, per qualsiasi consegna ravvicinata e, sopratutto, in qualsiasi orario. Non avrei potuto chiedere di meglio in questi anni.
Ringrazio in particolar modo la mia correlatrice e futura collega Lucia Migliorelli, fondamentale collegamento tra il mio team ed *Ospedali riuniti di Ancona*, la quale mi ha aiutato con i suoi consigli e il suo sostegno.
Un ulteriore ringraziamento va anche agli altri membri, ma sopratutto amici, del team con il quale ho avviato una startup, AIDAPT S.r.l., con il quale sono sicuro che riusciremo a realizzare grandi progetti.
Un ringraziamento speciale al professor Emanuele Frontoni, Relatore della mia tesi e mio mentore, al professor Adriano Mancini e alla dottoressa Sara Moccia che in egual misura stimo, i quali mi hanno sempre supportato con fiducia nella realizzazione di questo lavoro e dei progetti collaterali ad esso collegati.
Infine, ma non per importanza, un ringraziamento speciale alla dottoressa Mi-

chela Coccia e alla dottoressa Laura Villani con le quali l'intero team è entusiasta di collaborare. Loro hanno introdotto me e i miei colleghi all'aspetto clinico del progetto con grande professionalità, competenza e pazienza nei nostri confronti.