UNIVERSITÀ
POLITECNICA
DELLE MARCHE

Facoltà di Ingegneria
Corso di Laurea in Ingegneria dell'informazione e
dell'automazione

Analisi metriche di un sistema di riconoscimento facciale attraverso l'uso di reti neurali convoluzionali sul FRMDB.

Metric analysis of a facial recognition system using convolutional neural networks on the FRMDB

Candidato:
Yassir Flavio Suarez Sanchez

Relatore:
Prof. Aldo Franco Dragoni

Correlatore:
Prof. Paolo Sernani
Dott. Paolo Contardo

Anno Accademico 2023-2024

Sommario

"Negli ultimi anni, il campo del riconoscimento facciale ha ottenuto enormi progressi, in parte attribuibili alla crescente richiesta da una vasta gamma di settori; in particolare, questi significativi sviluppi sono stati resi possibili grazie alla diffusione del Deep Learning e delle reti neurali, le quali hanno prodotto risultati di maggiore precisione rispetto alle metodologie precedentemente impiegate.

Al passo con tali progressi, l'Università Politecnica delle Marche ha collaborato con le Forze dell'Ordine in uno studio finalizzato all'identificazione di soggetti attraverso l'utilizzo di un dataset di foto segnaletiche con immagini provenienti da diverse angolazioni (FRMDB [1]). In questo studio, tali immagini sono state confrontate con fotogrammi casuali, provenienti da sistemi di videosorveglianza, attraverso l'impiego di diverse reti neurali (VGG16, ResNet50, SeNet50); sebbene i risultati non abbiano mostrato un'elevata precisione, si evidenzia una tendenza in cui l'uso di un maggior numero di immagini risponde in maniera più favorevole rispetto a quelle tradizionalmente impiegate dalla Polizia.

Lo scopo di questa tesi è valutare la possibilità di migliorare l'accuratezza del sistema; questa valutazione inizia con una serie di test differenti da quelli precedentemente condotti, e successivamente prevede l'analisi dell'impatto sulle metriche in seguito all'introduzione di fotogrammi specifici, maggiormente simili alle immagini presenti nel FRMDB."

Indice

Capitolo 1

Introduzione

Nel campo Biometrico, il volto rappresenta uno dei principali ambiti di studio, poiché è in grado di identificare individui in base alle proprie caratteristiche facciali. Nel corso degli anni, si sono sviluppati diversi sistemi per il riconoscimento facciale, impulsati dalla crescente richiesta nei vari settori che vanno dalla sicurezza informatica alla gestione dell'identità digitale.
Il concetto di riconoscimento facciale, in biometria, può essere suddiviso in due modelli principali: la verificazione facciale (face verification/ authentication) e l'identificazione facciale (face identification/recogniton) [2].

- Face Recognition: questo processo sfrutta una relazione di tipo 1: n; il cui scopo è associare un volto estratto da un'immagine a un'identità univoca già presente nella base di dati di riferimento
- Face Verification: questo processo segue la relazione di tipo 1:1, cercando di stabilire una corrispondenza univoca tra l'immagine presentata e l'identità memorizzata nel sistema.

Di particolare interesse per le forze dell'ordine, soprattutto nell'ambito della sicurezza cittadina, risulta essere il sistema di Face Identification; questa tecnologia permette di individuare soggetti sospetti e di prevenire potenziali minacce terroristiche attraverso frame acquisiti da sistemi di videosorveglianza. In Italia, è in funzione un sistema di riconoscimento facciale noto come SARI (Sistema Automatico di Riconoscimento Immagini) Enterprise, gestita dall'azienda Parsec 3.26; tale sistema usufruisce la Base di Dati A.F.I.S. (Automated Fingerprint Identification System), che conserva i dati biometrici di quasi 10 milioni di persone (datato al 5 febbraio 2020 [3]), incluse foto-segnaletiche frontali e di profilo. Queste immagini vengono confrontate con i frame provenienti dalle telecamere di sorveglianza, che attraverso degli algoritmi del SARI enterprise si producono un elenco di immagini classificate in base al grado di similarità; inoltre, questo sistema è dotato di un filtro in cui sono specificate alcune caratteristiche dell'individuo (sesso, colore della pelle, colore dei capelli, ecc.) al fine di ridurre il numero di identità coinvolte nel confronto, cosi da migliorare l'accuratezza dei risultati [4].

L'Università Politecnica delle Marche, in collaborazione con le forze dell'ordine, ha condotto una ricerca intitolato 'FRMDB: Face Recognition Using Multiple Points

of View' [1], il cui obiettivo era la costruzione di una dataset in cui fossero presenti varie angolazioni di foto-segnaletiche per ciascuna identità, superando i limiti delle immagini comunemente usate dall'AFIS; in questo dataset sono stati registrati 39 soggetti, di cui 17 femmine e 22 maschi, di età media 24,6. In particolare, lo studio si è concentrato sull'analisi del comportamento di alcune reti neurali convoluzionali (VGG16, Resnet50 e successivamente SeNet50), le quali hanno estratto le caratteristiche facciali da diverse configurazioni delle immagini del database (test), per poi confrontarle con i fotogrammi casuali provenienti dai video di sorveglianza. Nonostante l'accuratezza non elevata, i risultati ottenuti hanno mostrato una maggiore efficienza nel riconoscimento facciale per gruppi di immagini più ampi rispetto a quelli tradizionalmente utilizzati dall'AFIS; tuttavia, non è ancora chiara la quantità ideale per ottenere risultati ottimali.

Basandosi su questa ricerca, la tesi si propone di individuare l'insieme ideale di foto-segnaletiche che produca il miglior risultato, cercando di massimizzare l'accuratezza utilizzando il minor numero di foto-segnaletiche. Al fine di raggiungere tale obiettivo, eseguirò una serie di test sul FRMDB composto da 39 identità, per poi replicarli su un dataset più ampio contenente 67 identità; e, infine, si verificherà la variazione delle metriche attraverso il confronto tra i gruppi di foto-segnaletiche con fotogrammi specifici, cercando di individuare quelle più simili a quelle presenti nel database, anziché utilizzare fotogrammi casuali.

1.1 Evoluzione Riconoscimento Facciale

Una delle principali motivazioni per lo sviluppo e lo studio dell'identificazione dei soggetti tramite il volto è che, a differenza di altri sistemi biometrici (come impronte digitali, DNA, iris, ecc), esso può essere considerato non-intrusivo, ossia non richiede una partecipazione diretta, da parte dei soggetti, per la raccolta di immagini e quindi permette una disponibilità quasi immediata delle risorse; a differenza di altri sistemi che obbligano a un'interazione diretta e, in certi casi, parecchio invasiva che di conseguenza comportano tempi lunghi prima di poter essere usufruita; inoltre, sono universali, cioè ad ogni individuo è associato un volto unico e distinto [5].
I risultati della FR in termini di accuratezza raggiungono percentuali significative, molto vicine come precisione agli altri sistemi biometrici e quasi paragonabili alle capacità umane[6].

I primi tentativi di implementazione del riconoscimento facciale in ambito tecnologico risalgono alla metà degli anni '60 grazie a Woody Bledsoe, che ideò un primo sistema semi-automatico, che consisteva nell'inserimento manuale in un computer, attraverso un RAND tablet, delle distanza tra le varie caratteristiche facciali di un volto e successivamente un software confrontava queste misurazioni con quelle presenti nel database[7].
La svolta in questo campo si è ottenuta negli anni '90 con l'introduzione dell'algoritmo Eigenface, sviluppato da Alex Pentland e Matthew Turk, che si basa sul metodo

dell'Analisi delle Componenti Principali (PCA), che è considerato un approccio olistico, in quanto prende in considerazione tutto il volto nel suo complesso anziché focalizzarsi sulle singole caratteristiche che lo compongono; quindi il volto viene rappresentato come una combinazione lineare di diversi autovettori che catturano le principali variazioni nell'aspetto globale dei volti. L'algoritmo Eigenface proietta l'immagine di un volto nello spazio delle eigenfaces, ottenendo un set di coefficienti che rappresentano il volto come una combinazione delle eigenfaces; questi coefficienti vengono poi confrontati con quelli di altre immagini registrati nel database per trovare la corrispondenza più vicina. Il problema principale di questo algoritmo, dovuto al fatto che è un metodo olistico, è la sua incapacità nell'affrontare i cambiamenti facciali incontrollabili (come l'illuminazione e il cambio di espressione)[8].
A causa di queste problematiche, agli inizi degli anni 2000 sono stati sviluppati nuovi sistemi basati su descrittori di caratteristiche locali come i filtri di Gabor e i pattern binari locali (LBP); queste tecniche hanno permesso di migliorare l'accuratezza del riconoscimento facciale, soprattutto in condizioni di variazioni di illuminazione, espressioni facciali e angolazioni diverse. Tuttavia, tali approcci presentavano alcune limitazioni: non erano in grado di adattarsi efficacemente a condizioni impreviste e mancavano di distintività e compattezza nell'estrazione delle caratteristiche[9].
Nel 2012, l'introduzione di AlexNet ha segnato una svolta rivoluzionaria nel campo del riconoscimento facciale mediante il deep learning; in particolare AlexNet è una rete neurale convoluzionale (CNN) composta da 8 strati (5 di convoluzione e 3 completamente connessi), dei quali ognuno apprendeva livelli di rappresentazioni differenti, andando a garantire una grande precisione e stabilità alle variazioni presenti tra le immagini. I risultati significativi di AlexNet diedero inizio all'adozione di tecniche di deep learning nel campo del riconoscimento facciale[10]. Partendo da questa tecnologia, nel 2014 Facebook ha sviluppato DeepFace, una rete neurale convoluzionale formata da 9 stratti di convoluzione, la quale è stata addestrata su 4 millioni di immagini appartenenti a circa 4 mila identità distinte, tale CNN è riuscito ad ottenere un'accuratezza del 97,35%, sul set di dati di riferimento LFW(Labeled Faces in the Wild), risultati quasi simili alle prestazioni umane (97,53%)[11].
Attualmente, il riconoscimento facciale si sta rapidamente diffondendo; aziende come Apple e Mastercard hanno già implementato questa tecnologia, e numerose altre imprese si stanno avvicinando con investimenti per integrarla nelle loro attività, mirando a rendere le operazioni più efficienti e soprattutto più sicure per l'impresa e per i propri clienti[12][13].

(a) sviluppo FR (b) ultimi anni

Figura 1.1: Evoluzione del riconoscimento facciale

Capitolo 2

Materiali e Metodi

2.1 Reti Neurali Convoluzionali (CNN)

Le reti neurali sono una branca del Machine Learning fondamentale per gli algoritmi di Deep Learning (apprendimento profondo), un metodo avanzato di apprendimento automatico che, mediante l'addestramento attraverso dati, permette ai computer di riconoscere modelli e correlazioni nascoste presenti in dati non strutturati, con l'intento di classificarli e raggrupparli; migliorando sempre di piu le capacità di previsione e di azione con l'aumento continuo dei dati. La struttura degli algoritmi delle reti neurali è ispirata al cervello umano e in particolare alla funzione dei neuroni; la struttura base è formata da nodi (neuroni) interconnessi tra loro e organizzati in strati (layers), quindi l'output di un neurone diviene l'input del neurone successivo fino a quando non si ottiene l'output del layer finale. Ad ogni nodo sono associati due parametri: il peso (weight), un valore numerico che determina la forza della conessione e regola l'impatto dell'input sull'output finale, e un bias che è un valore costante aggiunto al calcolo degli input ponderati, che consente di perfezionare e regolare le previsioni e garantendo al nodo di fare previsione anche quando l'input è pari a zero.

Generalmente una rete neurale e composta da tre strati: uno strato di input, uno o più strati nascosti (hidden layer) ed uno strato di output Fig. 2.1[14].

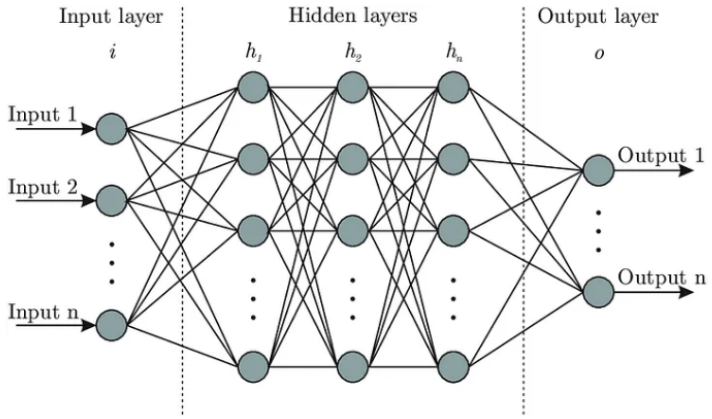


Figura 2.1: struttura base rete neurale

Le reti neurali convoluzionali (Convolutional Neural Network) rappresentano uno dei modelli più conosciuti e utilizzati nel campo del deep learning per la Computer Vision, specializzate nel riconoscimento di immagini e nell'identificazione di oggetti. Nelle CNN, sia l'input che l'output sono rappresentati su tre dimensioni: ampiezza, altezza e profondità; l'ampiezza e l'altezza si riferiscono alle dimensioni spaziali dell'immagine, ovvero il numero di pixel che la compongono; mentre la profondità si riferisce al numero di canali colore.

Le CNN sono reti neurali composte da diversi tipi di strati: strati convoluzionali, strati di pooling, strati di attivazione (come ReLU), strati completamente connessi e strati di loss Fig.2.2[15] [16].

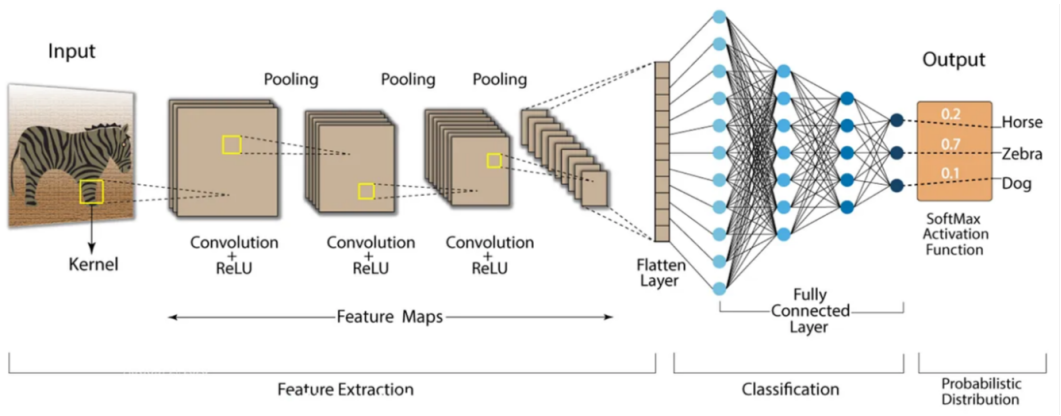


Figura 2.2: struttura rete neurale convoluzionale

Strato Convoluzionale

Lo strato convoluzionale è lo strato principale di una CNN ed è quello dove avviene la maggior parte dei calcoli; qui una volta ricevuta l'immagine di input, vengono applicati dei filtri, detti anche kernel, per estrarre specifiche caratteristiche dell'immagine. Ogni kernel è una matrice che ha dimensioni spaziali ridotte rispetto all'immagine originale, ma con profondità identica a quella di input; questi filtri vengono fatti scorrere in lunghezza e in larghezza dell'intera immagine e, una volta percorsa completamente, producono una matrice, chiamata feature map, che rappresenta la risposta dell'input ai filtro in ogni posizione[17].

La dimensione dell'output dipende da tre iper-parametri: la profondità, lo stride e lo zero padding.

- **Profondità** si riferisce al numero di filtri che compongono lo strato, da non confondere con la profondità della rete neurale.
- **Stride (passo)** corrisponde al passo con il quale il filtro si sposta, ossia il numero di pixel che si considera od ogni operazione; aumentado questo parametro l'output diminuisce.

- **Zero padding** rappresenta lo spesore di una cornice di zero aggiuntiva che si inserice nell'input in modo tener invariata la dimensione dell'output.

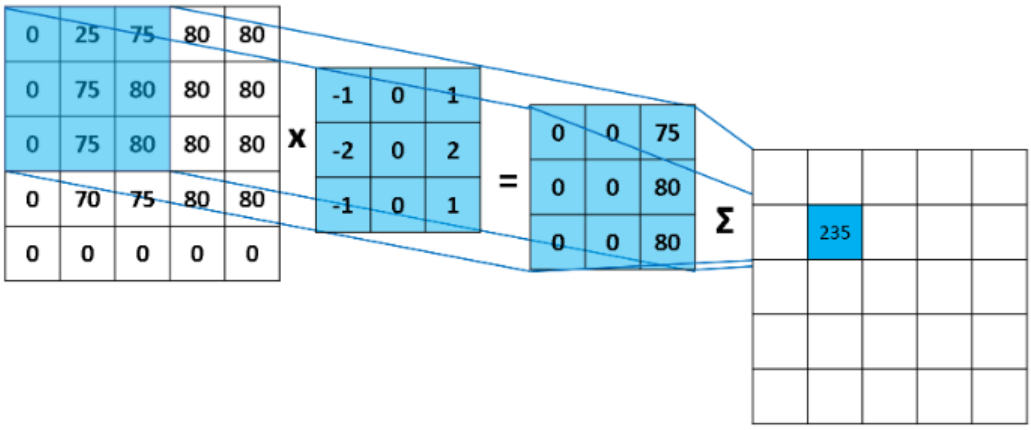


Figura 2.3: strato convoluzionale: kernel

Una volta passato attraverso lo strato convoluzionale l'input di dimensione $W_i x W_i x D$, dove W_i si riferisce all'altezza e all'ampiezza dell'immagine e D alla profondità di colore; le dimensioni dell'output prodotto puo essere ricavato dalla seguente formula:

$$W_o = \frac{(W_i - F + 2P)}{S} + 1,$$

dove F si riferisce alle dimensione spaziale del filtro, P quella di zero-padding e S lo stride; quindi le dimesnione nell'ouput è uguale $W_o x W_o x D$ [18].

Strato ReLu

Lo strato ReLu(Rectifier Linear Unit) è uno strato di attivazione ampiamente utilizzato per la sua semplicità, che opera sugli output degli strati convoluzionali introducendo non linearità, permettendo di apprendere relazioni complesse tra i dati. In particolare ReLu è una funzione che applica un'operazione elemento per elemento definita come:

$$f(x) = \max(0, x),$$

dove x è la variabile di input Fig.2.4; è una funzione che ha come scopo principale quello di annulare tutti i valori negativi di input e lasciare inalterati quelli positivi; inoltre la funzione non presenta saturazione, continuando a crescere per valori positivi.

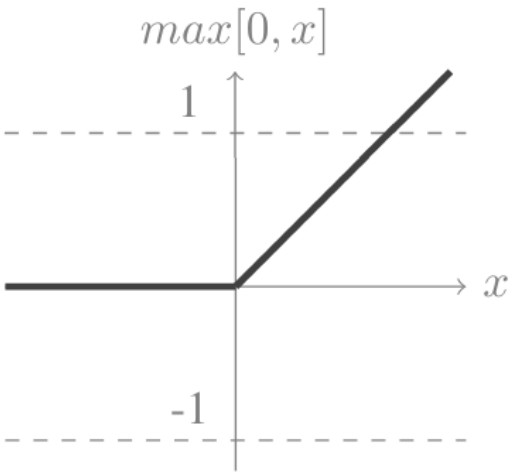


Figura 2.4: funzione di attivazione ReLu

Strato Pooling

Nello strato di pooling avviene un ridimesionamento spaziale e parametrico delle feacture map, mantenendo le caratteristiche principali; tale operazione si finalizza mediante filtri, che agiscono gradualmente su porzioni di feacture map.

In particolare l'output è prodotto della dimensione di input, dalla tipologia di pooling utilizzato e dalla dimensioni dei filtri.

Due sono le principali tecniche di pooling:

- **Max pooling:** all'interno della porzione considerata, seleziona il valore massimo.
- **Average pooling:** all'interno della porzione considerata, calcola e seleziona il valore medio.

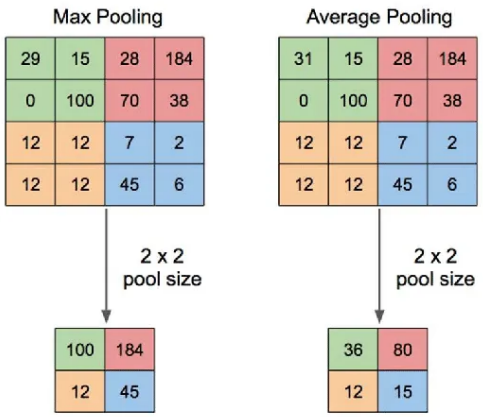


Figura 2.5: Max pooling e Average pooling

Considerando un input $W_i x W_i x D$, dove W_i le dimesioni e D la pronfodità di colore; il volume dell'ouput prodotto si puo ricavare dalla seguetne formula:

$$W_o = \frac{W_i - F}{S} + 1,$$

dove F la dimensione del filro di pooling e S il passo; quindi l'ouput ha dimensioni $W_o x W_o x D$.

Strato completamente connesso

Nello strato completamente connesso(fully connected), ogni nodo è collegato a tutti i nodi della feature maps derivanti dal risultato dell'ultimo strato di pooling nella CNN; questo strato è spesso utilizzato come strato di output, ossia per la classificazione o la regressione delle caratteristiche estratte.

Lo strato completamente connesso permette di combinare le caratteristiche estratte per produrre un vettore di output che rappresenta le probabilità delle diverse classi o i valori previsti.

Inoltre, questo puo essere considerato come una rete neurale base, infatti ogni nodo possiede un peso e un bias che aggirano sulle feature estrate.

Strato loss

L'ultimo stratto è quello di loss, che serve a calcolare e minimizzare il piu possibile la deviazione tra l'output previsto e l'output effetivo. La funzione piu diffusa è Softmax:

$$softmax(z_i) = \frac{exp(z_i)}{\sum_j exp(z_j)}$$

$\mathbf{z} = (z_1, z_2, \ldots, z_K)$ rappresenta i valori dei ultimi nodi dello strato completamente connesso.

Gli output della trasformazione Softmax sono sempre compresi nell'intervallo [0,1] e la loro somma è sempre uguale a 1, quindi essi costituiscono una distribuzione di probabilità delle classi di appartenza dell'input. [19]

2.1.1 VGG16

VGG-16 è un modello di rete neurale convoluzionale proposta nel 2014 da Karen Simonyan e Andrew Zisserman della Visual Geometry Group(VGG) dell'università di Oxford [20]; pre-addestrata per l'estrazione di caratteristiche facciali per mezzo dataset di VGGFace[21], dove è presente un elevato numero di campioni, circa 2.6 milioni di immagini di 2600 identità; che hanno riportato risultati di accuratezza ottimi.

Questa CNN è composta da 13 strati convoluzionali, 5 strati di Max Pooling, 3 strati completamente connessi e uno strato finale di Softmax Fig.2.15.

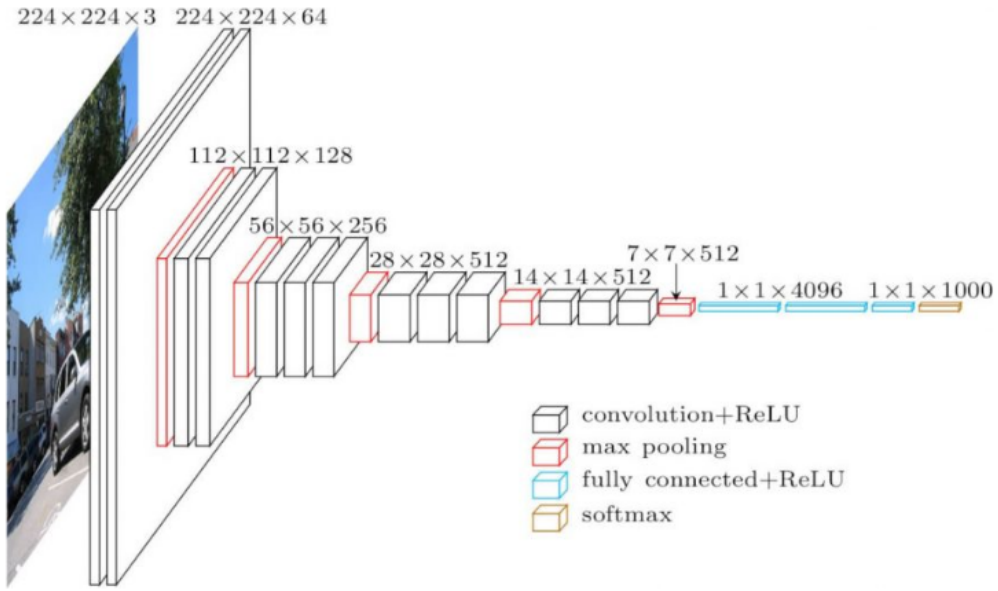


Figura 2.6: struttura VGG16

Dalla figura si puo notare che VGG16 in input riceve un'immagine di dimensioni 224x224x3, successivamente sono presenti una serie strati convoluzionali, dove vengono applicati kernel con dimensioni spaziali di 3x3 e con stride pari ad 1, all'interno della quale per mantere la dimensione spaziale dell'immagine viene applicato il zero-padding; inoltre insieme a ogni strato di convoluzione viene impiegato un strato di attivazione ReLu. Successivamente segue uno strato di Max Pooling, dove vengono utilizzati filtri di dimensione 2x2 con stride pari a 2, che riducono le dimensioni dell'immagine (tale schema si ripete 5 volte); infine, la CNN presenta tre strati completamente connessi e una funzione d'attivazione softmax.

In fase di addestramento viene impiegato la funzione triplet loss; Questa funzione ha lo scopo di ridurre la distanza tra gli embaddings di immagine che presentano la stessa identità e incrementare la distanza tra immagini con identità diverse; inoltre viene imposto un margine tra le immagini con la stessa identità, cosi da evitare una rapppresentazione equivalente nello spazio euclideo [22].

Sia x una immagine di input e $f(x) \in \mathbb{R}^d$ la funzione che rappresenta l'embadding dell'immagine nello spazio euclideo d-dimensionale; lo scopo, data una immagine àncora x_i^a, è quello di avvicinare tutte le immagine che risultano positive x_i^p (che hanno la stessa identità) e distanziare quelle che risultano negative x_i^n(identità differenti). Matematicamente si puo rappresentare come segue:

$$\|f(x_{a_i}) - f(x_{p_i})\|_2^2 + \alpha < \|f(x_{a_i}) - f(x_{n_i})\|_2^2 \quad \forall f(x_{a_i}), f(x_{p_i}), f(x_{n_i}) \in T,$$

dove α rappresenta il margine, e T l'insieme di tutti i terzetti possibili nel set di

addestramento; da cui la funzione di errore da minimizzare è:

$$L = \sum_{i=1}^{N} \Big[\|f(x_{a_i}) - f(x_{p_i})\|_2^2 - \|f(x_{a_i}) - f(x_{n_i})\|_2^2 + \alpha \Big]_+$$

In realtà, tra tutti i possibili terzetti di immagini ottenibili da un dataset, figurano molti che gia soddisfano la condizione della disequazione precedente; cio implica che contribuiscono poco all'apprendimento da parte della rete, rallentando la convergenza. Quindi per rendere piu efficiente l'addestramento della rete, è neccessario selezionare i terzetti che violano la condizione in maniera piu netta. Idealmente fissata un'àncora x_i^a, bisognerebbe individuare l'immagine positiva x_i^p che massimizza $\|f(x_{a_i}) - f(x_{p_i})\|_2^2$ (hard positive) e l'immagine negativa x_i^n che minimizzi $\|f(x_{a_i}) - f(x_{n_i})\|_2^2$ (hard negative). Nella pratica applicare questa tecnica all'intero dataset di addestramento non è computazionalmente realizzabile; infatti la soluzione individuata è quella di selezionare casualmente un campione negativo tra quelli che violano la condizione di vicinanza; cosi da ridurre la complessità computazionale delle operazioni di ricerca.

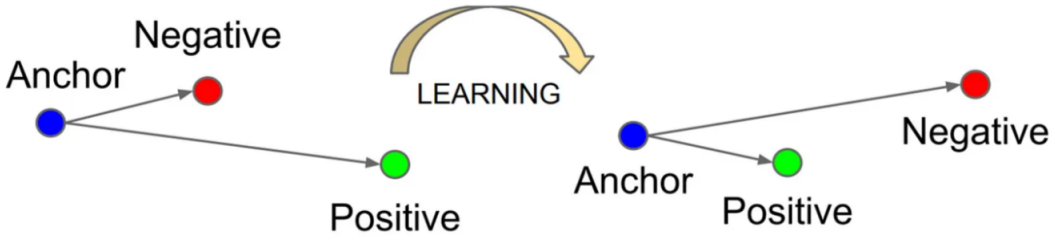


Figura 2.7: funzione triplet loss

2.1.2 ResNet50

ResNet (Residual Network) è rete neurale convoluzionale introdotta da MIcrosoft Reasearch nel 2015[23], pre-addestrata per l'estrazione di caratteistiche facciali per mezzo del dataset di VGGFace2[24]; che contiene circa 3,31 milioni di immagini con 9131 identità.

L'architterua di Resnet è caraterizzata dalla sua profondità, ossia dal numero elevato di strati che presenta, cio è stato fatto per evitare i fenomeni di decadimento del gradiente che si incontrava durante l'addestramento; il gradiente è quel componente che determina il tasso di apprendimento della rete: più è alto, più efficiente è la rete neurale; viceversa, se è basso, implica una minore efficienza, sino a raggiungere uno stato di saturazione che blocca l'appredimento della CNN.

Quindi la soluzione implementata è una tecnica di apprendimento residuale: piuttosto che far apprendere a un certo numero di strati una funzione $H(x)$, si ricerca la sua funzione residuale $F(x) = H(x) - x$; allora la funzione originale diviene $H(x) = F(x) + x$, questo è reso possibile delle shortcut connection, che permette

di collegare strati successivi saltando alcuni strati intermedi, mantenendo l'input della funzione e sommando l'output degli strati saltati; il blocco così ottenuto viene nominato blocco residuale Fig.2.8. Questo è valido se si considera implicitamente che $F(x)$ e x hanno stessa dimensione, quindi per la somma non richiede ulteriori operazioni; ma nella realtà non e cosi, infatti è necessario applicare una proiezione lineare nello shortcut del tipo $y = F(x, W_i) + W_s x$, con W_s matrice quadrata.

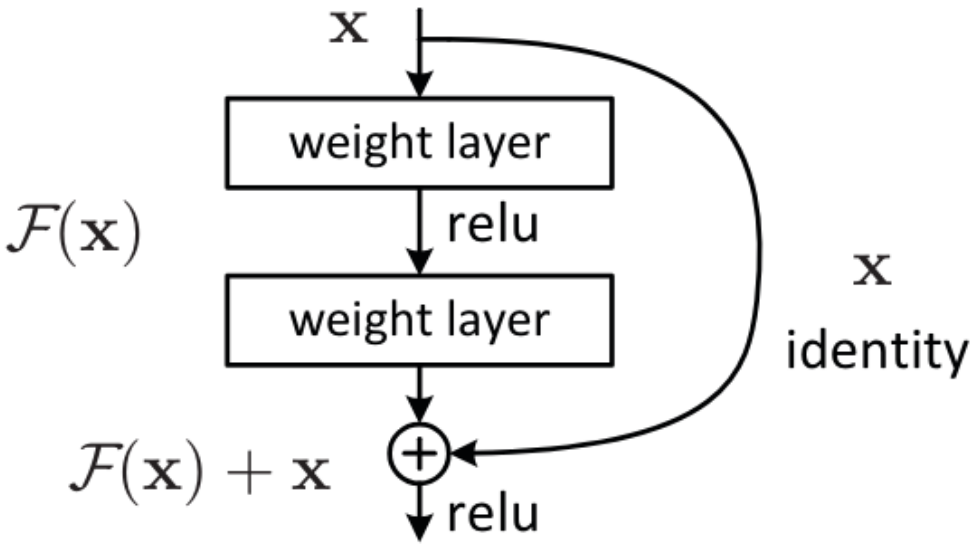


Figura 2.8: blocco residuale[25]

Le architteture di ResNet sono diverse, tra cui la piu peculiare risulta quellla a 152 strati, ma per l'esperimentazione in questa tesi considereremo quella a 50 strati; di queste 48 sono strati convoluzionali, 1 di Max Pooling e 1 di Average Pooling.

La struttura di ResNet-50 può essere considerata in 5 stadi (stage); in particolare, il primo stadio è quello che riceve l'immagine non elaborata (con dimensione 224x224x3), successivamente i 4 stadi presentano uno strato di convoluzione e più strati di identità, questi strati presentano i sistemi residuali descritti precedentemente; infine, dopo tutti gli stadi, sono presenti uno strato di Average Pooling e lo strato completamente connesso il quale esegue la classificazione usufruendo della funzione softmax.

I blocchi identità si attivano quando la dimensione dell'input è uguale a quella di output, dato che permette la somma. Invece quando presentano due dimensioni diverse allora o si compensa con l'aggiunta di zero oppure si usa la proiezione lineare.

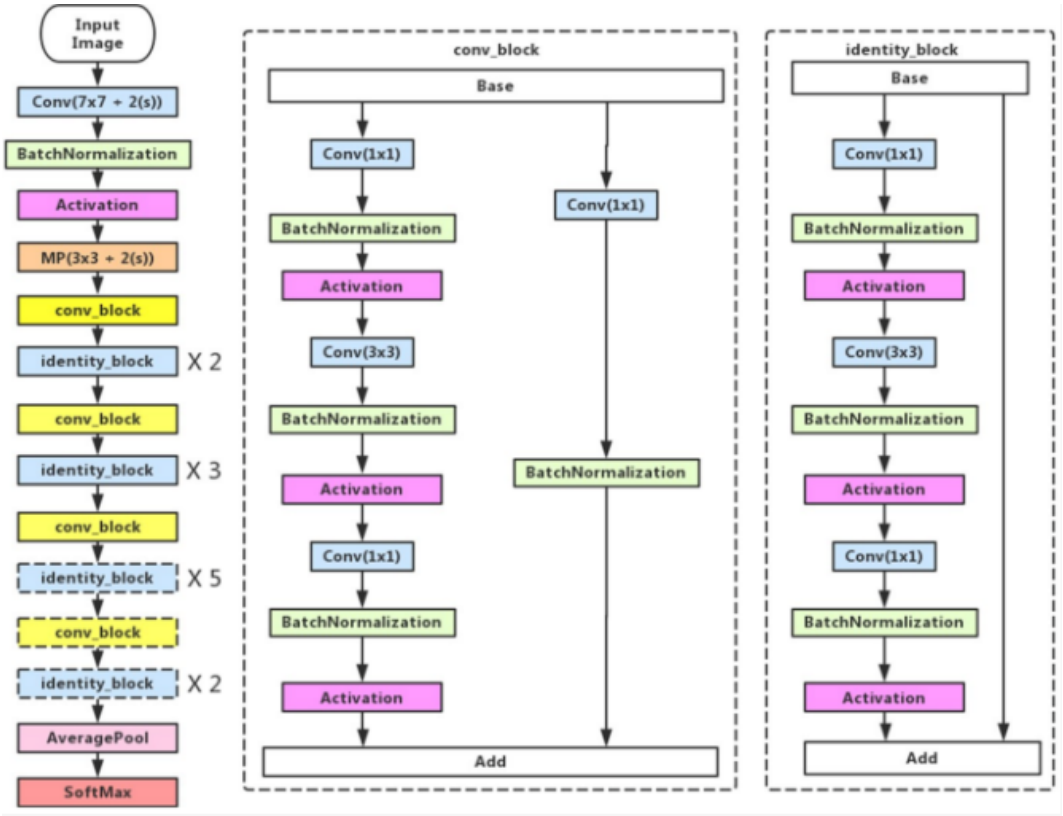


Figura 2.9: Struttura ResNet

2.1.3 SeNet50

SeNet-50 (Squeeze and Excitation-ResNet50) è una rete neurale convoluzionale, la cui architettura si basa su quella di ResNet, ma con la prensenza di blocchi Squeeze and Excitation(SE) [26]. Questo blocco agisce nello strato convoluzionale, in particolare interviene nella modulazione dei pesi nei canali che creano le feature map; a differenza dei modelli di CNN tradizionali, il blocco SE analizza singolarmente i canali e attribuisce un peso maggiore a quelli piu rilevanti e riduce quelli meno significativi; garantendo così una rappresentazione piu efficiente dell'informazione.

Il modo in cui il blocco SE agisce può essere suddiviso in due fasi:

- **Squeeze (compressione):** in questa fase avviene una compressione delle feature map prodotte da uno strato convoluzionale, fino ad ottenere un vettore monodimensionale con numero di elementi pari al numero di filtri dello strato convoluzionale; questi elementi sono scalari.

- **Excitation (eccitazione):** in questa fase il vettore ottenuto nella fase precedente, passa attraverso due strati completamente connessi, così si ottiene un vettore con la stessa dimensione, che rappresenta una collezione di pesi da applicare alla feature map originaria.

La costruzione di un blocco SE parte da una trasformazione F_tr, che mappa l'input $X \in \mathbb{R}^{H' \times W' \times C'}$ nella feature map $U \in \mathbb{R}^{H \times W \times C}$. Se consideriamo F_tr un operazione convoluzionale e sia $V = [v_1, v_2, ..., v_C]$ il vettore che definisce i filtri; allora il vettore dell'output prodotto $U = [u_1, u_2, ..., u_C]$; e il c-esimo elemento puo essere calcolato come segue:

$$u_c = v_c * X = \sum_{s=1}^{C'} v_c^s * x^s,$$

dove $*$ si intende l'operazione di convoluzione, $v_c = [v_c^1, v_c^2, \ldots, v_c^{C'}]$, $X = [x^1, x^2, \ldots, x^{C'}]$ e $u_c \in \mathbb{R}^{H \times W}$; non consideriamo il bias così da semplificare la notazione.

Nella fase di Squeeze viene applicato uno strato di global average pooling, in modo da comprimere i C canali della feature map in un descrittore di canale; quest'ultimo matematicamente $z \in \mathbb{R}^c$ è generato contraendo U nelle sue dimensioni spaziali H e W in modo tale che il suo c-esimo elemento sia calcolato come:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{K} u_c(i, j)$$

Nella fase di Excitation viene aggiunta una funzione di non-linearità per mezzo di uno strato completamento connesso, seguito da una funzione di ReLu δ, a cui segue un altro strato completamente connesso e, infine, una funzione di attivazione Sigmoide:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)),$$

dove $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ e $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$; dove W_1 realizza una riduzione dimensionale basata sul fattore di riduzione r, mentre W_2 opera un incremento dimensionale. L'output del blocco $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \ldots, \tilde{v}_C]$ è calcolato scalando il vettore U con s, allora il c-esimo elemento si ricava come segue:

$$\tilde{x_c} = F_{\text{scale}}(u_c, s_c) = s_c u_c$$

La Fig.2.17 mostra il confronto tra blocco SE (destra) e blocco residuale(sinistra); qui si evidenzia che il blocco SE viene inserito prima di quello residuale il che ha permesso l'architettura di SE-Net50.

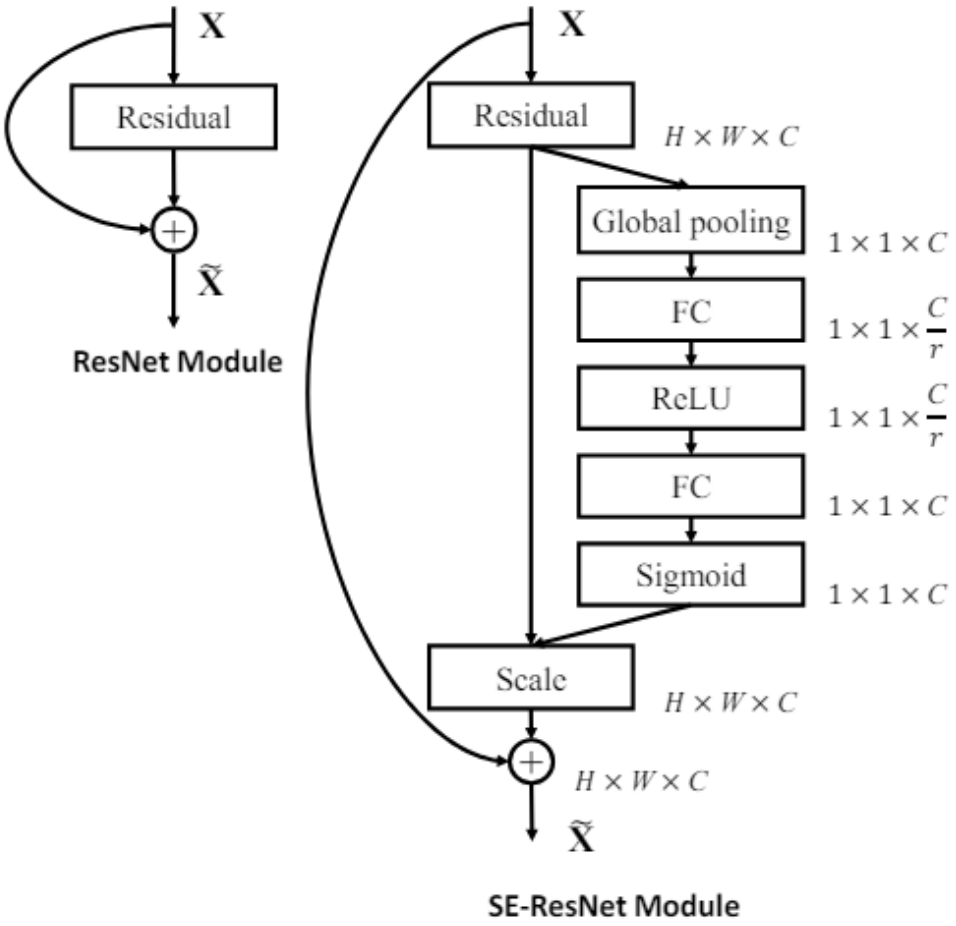


Figura 2.10: confronto blocco SE con il blocco residuale

2.2 FRMDB

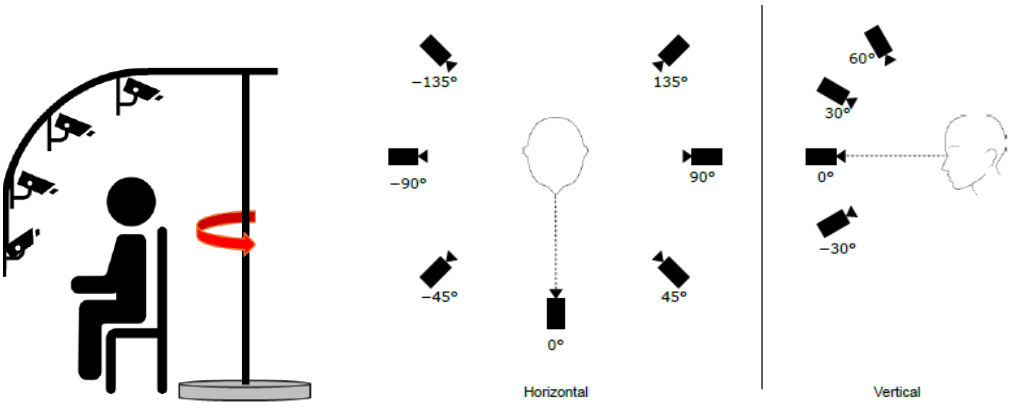
Il Face Recognition from Mugshot Data Base(FRMDB) è un dataset sviluppato dall'Università Politecnica delle Marche, con lo scopo di verificare che l'impiego di gruppi di foto-segnaletiche prese da diverse propettive, diverse da quelle tradizionalmente usate(frontale e profilo destro), migliorano l'accuratezza da parte dei sistemi per il ricoscimento facciale dei soggetti; questo dataset contiene nella prima pubblicazione[1] un numero di 39 soggetti unici di cui 17 femmine e 22 maschi, a cui successivamente sono stati aggiunti 28 soggetti nuovi, di cui 14 maschi e 14 femmine; per ogni soggetto nel database sono registrati:

- Per tutte le 67 identità sono presenti 28 foto-segnaletiche con risoluzione 972x544 pixels, prese da diverse angolazioni(Fig. 2.11), in formato JPEG images.
- Per i primi 39 soggetti sono presnti 5 video da telecamere catturati da 5 punti di vista diversi; i quali sono codificati in H.264 codec, con risoluzione di 352x288 pixels e 60 fotogrammi al secondo. Per i 28 soggetti restanti sono presenti 3

video ad alta definizione prese da telecamere posti in 3 punti di vista differenti, i video sono codificati con il codec H.264 e registra 25 fotogrammi al secondo con risoluzione pari 1920x108 pixels.

Figura 2.11: Foto-segnaletiche

Le foto segnaletiche sono scattate per mezzo di un braccio meccanico in cui sono state poste 4 telecamere(Fig. 2.12a), questo braccio meccanico ruota intorno ad un asse verticale e si ferma in 7 punti diversi, dove vengono scattate le foto da tutte le telecamere, in tal modo si ottengono le 28 foto segnaletiche da 7 angolazioni orizzontali, che va da -135° a +135° con un passo di 45°, e da 4 verticali, da -30° a +60° con un passo di 30°(Fig. 2.12b).



(a) braccio meccanico

(b) Asse verticale e orizzontale

Figura 2.12: descrizione foto

All'interno del database le foto segnaletiche sono identificabili per mezzo del loro nome; in particolare, ogni foto segnaletica è denominata "**Img_XY.jpg**", dove X rappresenta l'angolo orizzontale (con 0 = -135° a 6 = +135°) e Y l'angolo verticale (con 1 = -30° a 4 = +90°). Mentre, i video nel database sono nominati "**yyyy-mm-dd hh-mm-ss-CamX.mkv**" e "**yyyy-mm-dd CamX.mts**", utilizzando il timestamp all'inizio del video, e un indice X che indica la telecamera che ha registrato il video (da 1 a 5, per i primi 39 soggeti, e 1 a 3, per i restanti).
La struttura del Database è la seguente Fig. 2.13 :

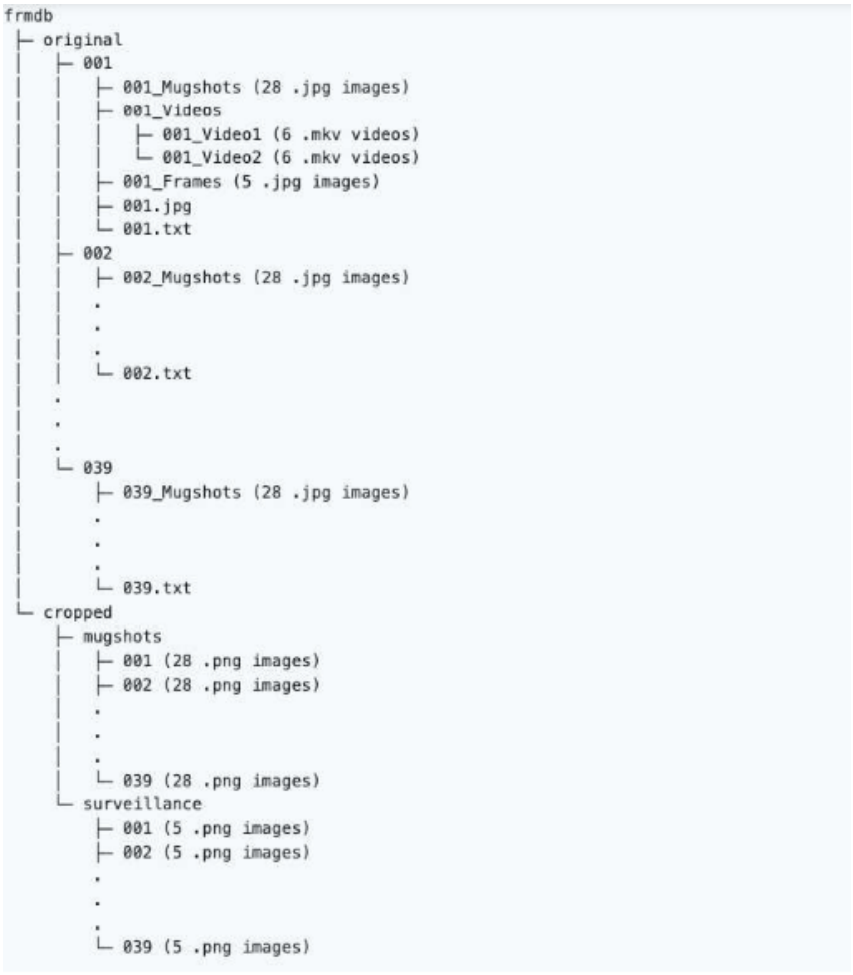


Figura 2.13: Struttura database 39 soggetti

Dall'immagine si nota che sono presenti due directory **orginal** e **cropped**. All'interno della prima troviamo sottocartelle con codice a 3 cifre che si riferiscono ad ogni soggetto (es. "001","002", ...), e ogni cartella del soggetto presenta:

- Una cartella "XYZ_Mugshots" che contiene le 28 foto segnaletiche di ciascun soggetto.
- Una cartella "XYZ_Videos", che contiene i video di sorveglianza; all'interno

della quale è presente una sottocartella "XYZ_Video1" dove stanno i 5 di video di sorveglianza, e per alcuni soggetti è presente un'altra cartella "XYZ_Video2", in cui sono presenti i video dove i soggetti indossano gli occhiali.

- Una cartella "XYZ_Frames" che contiene 5 fotogrammi presi ciascuno da un video di sorveglianza, in formato (.jpg).
- un file "XYZ.jpg" un mugshot frontale in full HD(1920 x 1080).
- Un file "XYZ.txt", dove sono presenti informazioni del soggetto(età, sesso e se indossa gli occhiali).

Nella seconda directory cropped sono presenti due sottocartelle:

- una cartella **mugshots** "XYZ" per ciascun soggetto, che contiene tutte le foto sengaletica croppati del soggeto e il nome dei file rimane lo stesso, in formato PNG.
- una cartella **surveillace** "XYZ" per ciascun soggetto, dove sono presenti frame croppati, presi dai fotogrammi in "XYZ_Frames", 5 per i primi 39 soggetti, presi dai 5 video, e tre fotogrammi per i restani, tutte in formato PNG.

Dove XYZ è il codice del soggetto.

2.2.1 Risultati

Con il nuovo database a disposizione, è stata eseguita una serie di test utilizzando insiemi di foto segnaletiche prese dal FRMDB; tra questi è stato considerato un insieme che presenta le prospettive di profilo destro e frontale (test-1) cosi da disporre di un confronto immediato. Per questi test e i successivi si sono considerati due tipologie di metriche per misurare l'affidabilità e l'efficacia delle reti neurali:

- **Identities:** indica la capacità, dato un qualunque fotogramma di videosorveglianza, di trovare nella classifica di somiglianza, restituita dalla rete neurale, l'identità del soggetto in questione. Nello specifico se l'identità associata alla prima foto segnaletica restituita corrisponde a quella del soggetto del fotogramma (top-1), se sta tra le prime 3 (top-3), se sta tra le prime 5 (top-5) e se sta tra le prime 10 (top-10); in caso di ripetizione di identità nelle foto segnaletiche la consideriamo come unica.
- **Mugshots:** indica la capacità, dato un qualunque fotogramma di videosorveglianza, di trovare nella classifica di somiglianza, restituita dalla rete neurale, la prima delle 28 foto segnaletiche del soggetto in questione. Nello specifico se la prima foto segnaletica restituita corrisponde all'identità del soggetto presente nel fotogramma (top-1), se sta tra le prime 3 (top-3), se sta tra le prime 5 (top-5) e se sta tra le prime 10 (top-10).

I vari test effettuati sono i seguenti e la loro descrizione fa riferimento alla Fig. 2.11; in particolare tale immagine mostra l'insieme di foto segnaletiche che compongono i test, nelle quali ogni foto segnaletica è riconoscibile attraverso le angolazioni verticali e orizzontali inserite nelle parentesi:

Test	Mugshots
Test F	(0°, 0°)
Test F-L1-R1	(0°, 0°), (-45°, 0°), (45°, 0°)
Test 1	(0°, 0°), (90°, 0°)
Test 2	(0°, 0°), (90°, 0°), (-90°, 0°)
Test 3	(0°, 0°), (90°, 0°), (-90°, 0°), (45°, 0°), (45°, 0°)
Test 4	(0°, 0°), (135°, 0°), (-135°, 0°), (90°, 0°), (-90°, 0°), (45°, 0°), (45°, 0°)
Test 5	(0°, 0°), (135°, 0°), (-135°, 0°), (90°, 0°), (-90°, 0°), (45°, 0°), (45°, 0°), (0°, 30°), (135°, 30°), (-135°, 30°), (90°, 30°), (-90°, 30°), (45°, 30°), (45°, 30°)
Test 6	(0°, 0°), (135°, 0°), (-135°, 0°), (90°, 0°), (-90°, 0°), (45°, 0°), (45°, 0°), (0°, 30°), (135°, 30°), (-135°, 30°), (90°, 30°), (-90°, 30°), (45°, 30°), (45°, 30°), (0°, 60°), (135°, 60°), (-135°, 60°), (90°, 60°), (-90°, 60°), (45°, 60°), (45°, 60°), (0°, -30°), (135°, -30°), (-135°, -30°), (90°, -30°), (-90°, -30°), (45°, -30°), (45°, -30°)

Figura 2.14: I test effettuati sul FRMDB

Nella ricerca madre [1], i vari test sopra descritti sono stati eseguiti solo sulle prime 39 identità, mentre qui di seguito presenteremo i risultati riguardanti il numero totale di identità disponibili[27]:

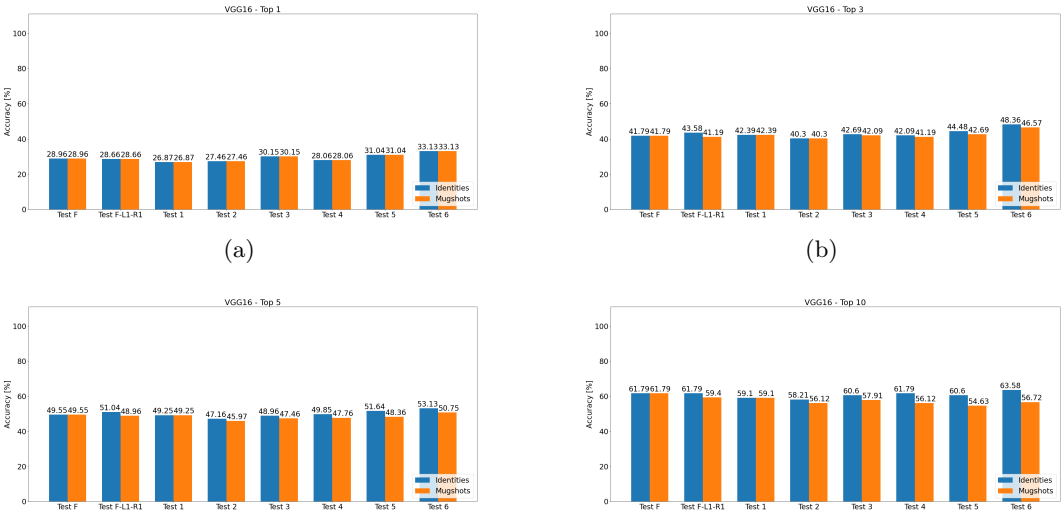


Figura 2.15: risultati VGG16

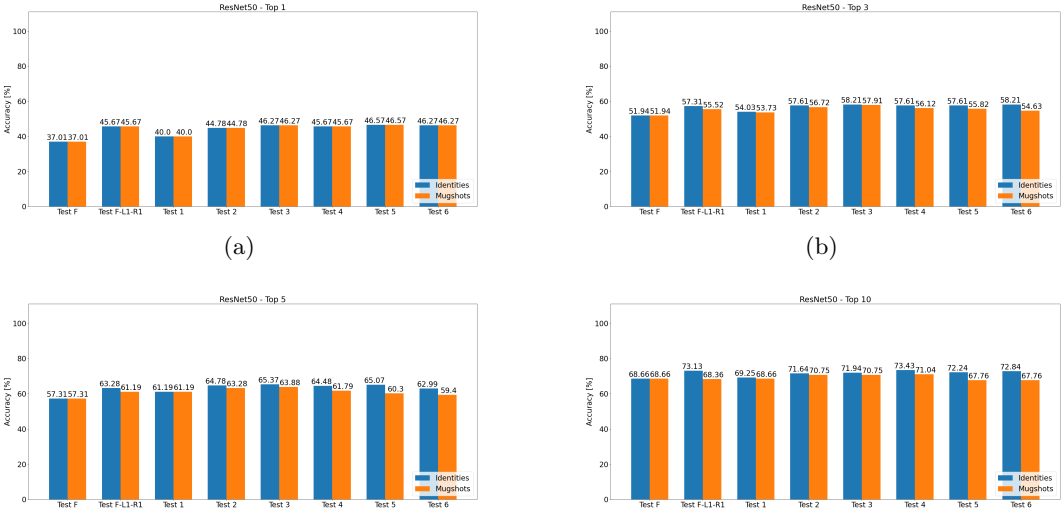


Figura 2.16: risultati ResNet50

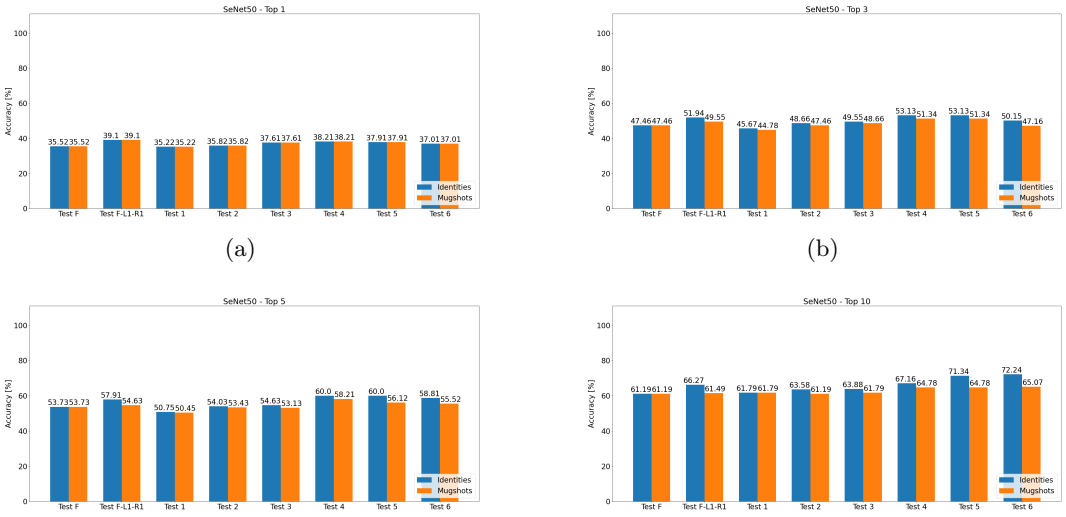


Figura 2.17: risultati SeNet50

Dai risultati si è concluso che, rispetto al test-1, l'impiego di gruppi maggiori di foto segnaletiche producono risultati migliori; tale tendenza si riscontra in tutti i casi analizzati per i due parametri considerati nelle tre reti neurali. Nello specifico, il test-5 e il test-6, che rispettivamente comprendono 14 e 28 foto segnaletiche, sono quelle che restituiscono le migliori percentuali, anche se tali percentuali non sono molto elevate.
Partendo da questi risultati, cercheremo innazitutto di verificare che effettivamente sia vera l'ipotesi che gruppi di immagini maggiori producano migliore accuratezza; in seguito tenteremo di individuare un gruppo che produca la migliore percentuale generale; e infine, tenteremo di migliorare l'accuratezza dei sistemi prendendo nuovi fotogrammi con miglior risoluzione e/o considerando diverse pose dei soggetti all'interno dei fotogrammi.

Capitolo 3

Risultati

Come anticipatamente spiegato, la tesi ha lo scopo di analizzare il comportamento delle tre reti neurali in due specifici contesti. Nel primo, denominato "insiemi diversi di Foto-segnaletiche", saranno eseguiti nuovi test sul FRMDB e successivamente sul FRMDB aggiornato, al fine di verificare e confermare che l'impiego di un numero maggiore di fotosegnaletiche migliora l'accuratezza nell'individuazione di un soggetto. Mentre nel secondo, denominato "frame specifici", replicheremo i test gia' effettuati e presentati nella ricerca pubblicata, ma considereremo fotogrammi di videosorveglianza con immagini che riprendono i soggetti in pose piu simili a quelle presenti nelle foto segnaletiche.

3.1 Insieme diverso di foto-segnaletiche

In questa fase, sono stati elaborati dei nuovi test, dove abbiamo considerato gruppi diversi di fotosegnaletiche per l'addestramento delle reti neurali; cercando di coprire un vasto range di casi, in modo da garantire dei risultati e delle conclusioni il piu possibile oggettive .
I test sono i seguenti e le loro descrizione fanno riferimento a Fig. 2.11:

- **Test 7**: tutte le foto rispetto all'asse verticale e per l'asse orizzontale -/+90 e 0.
- **Test 8**: tutte le foto rispetto all'asse verticale e per l'asse orizzontale +90 e 0 (questa conseguente dal fatto che le Forze dell'Ordine spesso posseggono solo quella prospettiva).
- **Test 9**: tutte le foto rispetto all'asse verticale e per l'asse orizzontale -/+90, -/+45 e 0.
- **Test 10**: tutte le foto dalle prospettive verticale e per l'asse orizzontale -/+45 e 0.
- **Test 11**: per l'asse verticale -/+30 e 0, e per l'asse orizzontale -/+90, -/+45 e 0.
- **Test 12**: per l'asse verticale 30 e 0 , e per l'asse orizzontale -/+90, -/+45 e 0.

- **Test 13**: asse verticale -/+ 30 e 0, asse orizzontale -/+ 45 e 0.
- **Test 14**: asse verticale - 30 e 0, asse orizzontale -/+ 45 e 0.
- **Test 15**: asse verticale + 30 e 0, asse orizzontale -/+ 45 e 0.
- **Test 16**: asse verticale -/+ 30 e 0, asse orizzontale -/+ 45.
- **Test 17**: asse verticale –30 e 0, asse orizzontale -/+ 45.
- **Test 18**: asse verticale +30 e 0, asse orizzontale -/+ 45.
- **Test 19**: asse verticale 0 e asse orizzontale -/+ 45; e asse verticale -30 e asse orizzontale + 45.
- **Test 20**: asse verticale 0 e asse orizzontale -/+ 45, 0 e 90; asse verticale -30 e asse orizzontale -/+ 45.
- **Test 21**: asse verticale 0 e asse orizzontale -/+ 45 e 0; asse verticale -30 e asse orizzontale – 45.
- **Test 22**: asse verticale 0 e asse orizzontale -45; asse verticale -30 e asse orizzontale + 45 .
- **Test 23**: asse verticale 0, asse orizzontale -/+ 45, 0 e 90. (vgg16)
- **Test 24**: asse verticale 0, asse orizzontale -/+ 45, 0 e 90; e asse verticale -30, asse orizzontale -/+ 45.
- **Test 25**: test9 sostituendo nell'asse verticale 90 la parte orizzontale con -/+45, 90 e 0.
- **Test 26**: test14 aggiungendo l'asse verticale 60 la parte orizzontale con -/+45, 90 e 0.
- **Test 27**: test17 aggiungendo l'asse verticale 60 la parte orizzontale con -/+45, 90 e 0.
- **Test 28**: test19 aggiungendo l'asse verticale 60 la parte orizzontale con -/+45, 90 e 0.
- **Test 29**: test22 aggiungendo l'asse verticale 60 la parte orizzontale con -/+45, 90 e 0.
- **Test 30**: test21 aggiungendo l'asse verticale 60 la parte orizzontale con -/+45, 90 e 0.
- **Test 31**: tutte le foto dell'asse verticale 60°.
- **Test 32**: tutte le foto dell'asse verticale 30°.

- **Test 33**: tutte le foto dell'asse verticale 0°.
- **Test 34**: tutte le foto dell'asse verticale -30°.

I test possono essere suddivisi in due gruppi, il primo che comprende i test dal 7 al 30 sono stati impiegati per verificare l'ipotesi principale; la loro costruzione è partita considerando casi genereci e sulla base dei miglior risultati si sono sviluppati i successivi test. Mentre, i test dal 31 a 34 sono stati utilizzati per confrontare tutti i gruppi di foto presenti ad ogni livello verticale, in modo da verificare l'incidenza nell'accuratezza delle varie prospettive.

3.1.1 FRMDB

Il FRMDB, lo stesso dataset presentato nella ricerca, contiene 39 individui unici, di cui 17 femmine e 22 maschi con età media 24,6; il confronto delle foto segnaletiche viene effettuato con fotogrammi casuali dei video di sorveglianza. I risultati dei vari test eseguiti nelle tre reti neurali, sono riportati di seguito, ogni risultato dei test dal 7 al 30 è messo in confronto con quello del test-1(frontale e profilo destro), indicato dal colore arancione e rappresentato parallelamente all'asse orizzontale, sia per quanto riguarda Indetities che Mugshot in ogni Top.

(a) (b) (c) (d) (e) (f)

Figura 3.1: Risultati test su Dataset1 7-30

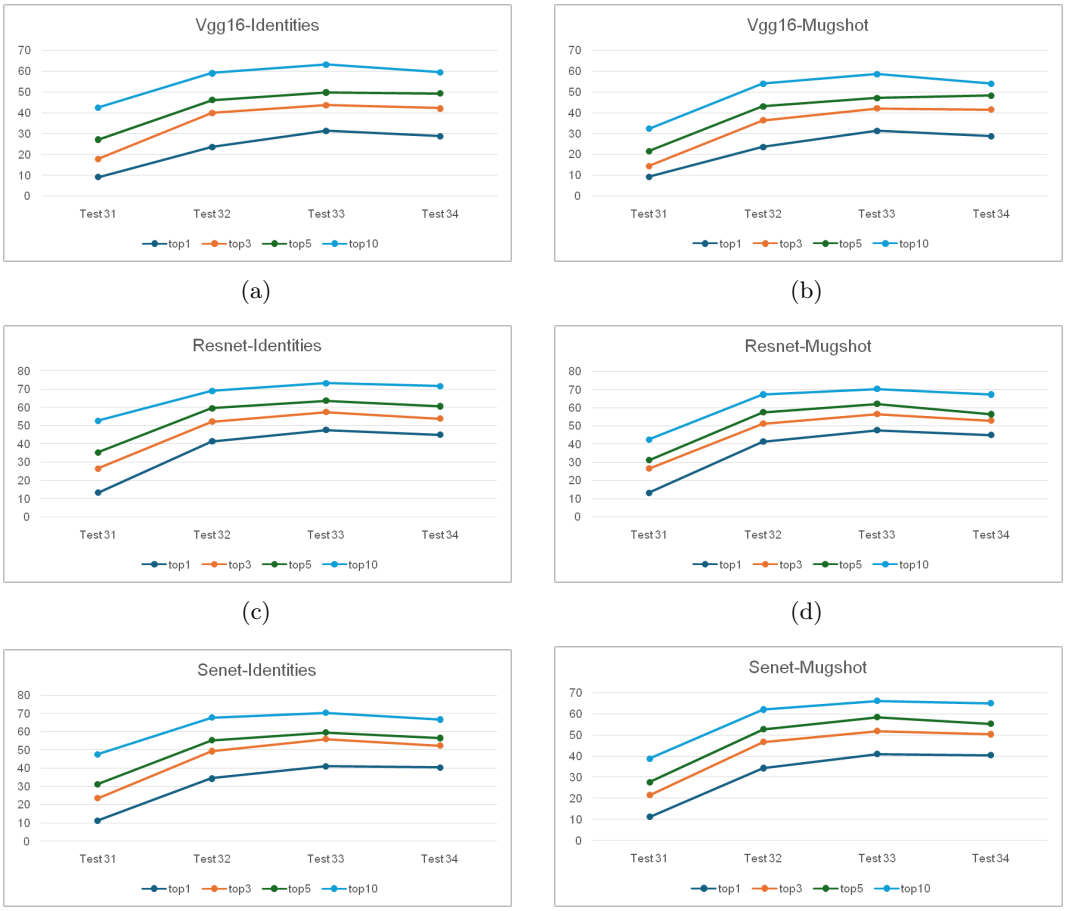


Figura 3.2: Risultati test su Dataset1 31-34

I risultati del primo gruppo di test (da 7-30) per quanto riguardano le Identities, come si vede da Fig.3.1a, Fig.3.1c, e Fig.3.1e, mostrano nella maggior parte dei test un miglioramento nell'accuratezza rispetto al test-1, tale tendenza appare piu' evidente in vgg16, mentre Senet è la rete neurale in cui si nota di meno. Per quanto riguardo il parametro di Mugshot, Fig.3.1b, Fig.3.1d, e Fig.3.1f , l'accurateza nelle tre reti neurali, sia in top1 che in top3 riportano delle percetuali, nella maggior parte dei casi, migliori rispetto al test-1; mentre, in top-5 e top-10 il test-1 mostrano risultati migliori; in particolare, tale fenomeno potrebbe dipendere dal numero elevato di foto segnaletiche impiegate nel confronto dei nuovi test, anzi i risultati in top-1 e top-3 confermano che l'impiego di un numero di fotosegnaletiche, restituisce con migliore precisione l'identità del soggetto.

Mentre i test 31-34, Fig.3.2, mostrano che la prospettiva 0° è quella migliore, inoltre tra -30 e +30 si puo' notare una similitudine e, infine, +60 è quella con i risultati peggiori, cio' vale sia in identities e mugshot, quindi si puo' concludere che quest'ultima prospettiva potrebbe essere ininfluente. Per tale ragione nei test 7-30 buona parte non comprende la prospettiva +60, e in quelli in cui sono pressenti,

riportano i peggiori risultati (test da 26 a 30).

3.1.2 FRMDB aggiornato

Il FRMDB aggiornato analizza gli stessi soggetti del FRMDB ai quali vengono aggiunti 28 nuove indentità dei quali 14 maschi e 14 femmine, per un totale di 67 soggetti unici; per alcuni dei nuovi individui sono presenti due tipologie di video, uno in cui non indossano alcun accessorio e un altro in cui indossano almeno un accessorio (nei test attuali saranno considerati solo i video in cui i soggetti non indossano alcun accessorio). Abbiamo replicato tutti i test condotti sul FRMDB, per verificare che le ipotesi siano ancora valide; i risultati sono di seguito riportati:

(a) (b)

(c) (d)

(e) (f)

Figura 3.3: Risultati test su Dataset2 7-30

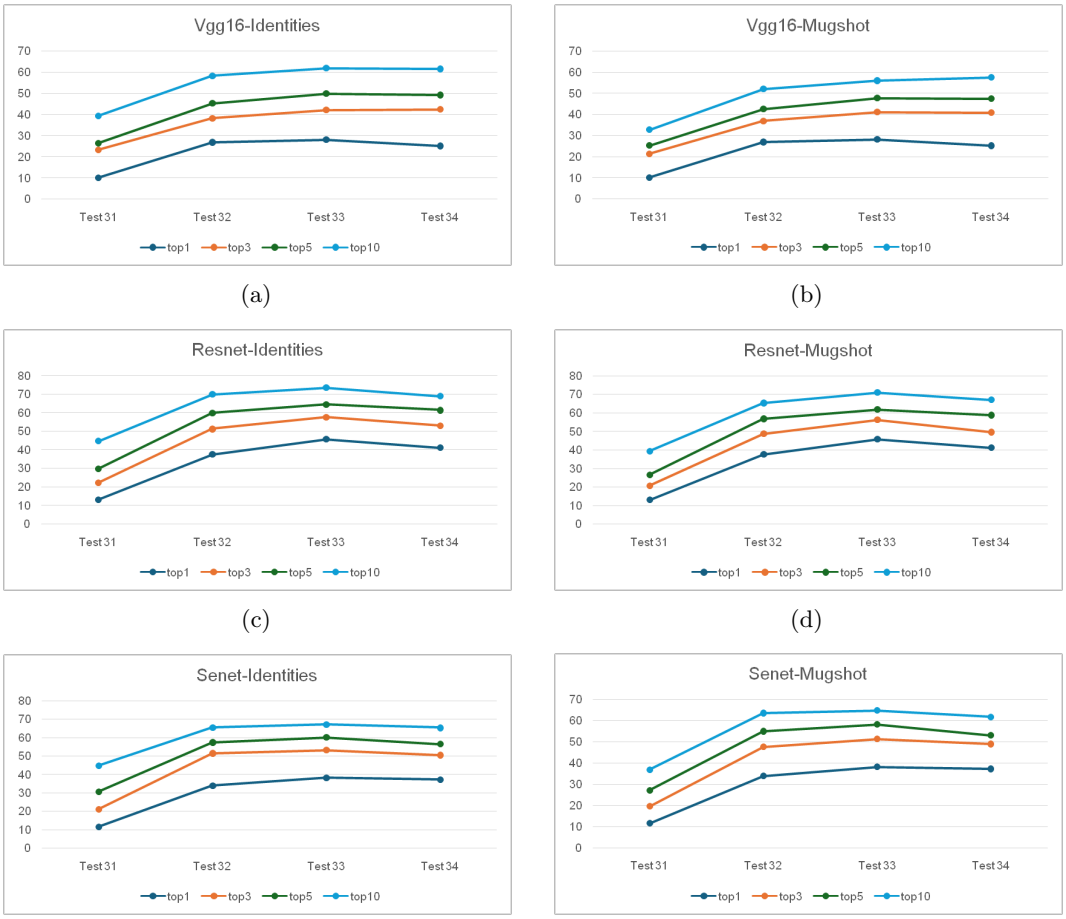


Figura 3.4: Risultati test su Dataset2 31-34

I risultati mostrano che l'accuratezza peggiora di una percentuale diversa per ogni Top di ogni rete neurale, quella che subisce una minore variazione è vgg16, mentre quella che ne risente di piu è Senet, tali risultati sono semplificati nella Tabella 3.1, qui sono riportati la media di peggiornameto dell'accuratezzadi ogni rete neurale per ogni Top, tale peggioramento è dovuto al nuovo numero maggiore di soggetti presenti; inoltre, la variazione è dovuta anche dalla diversa risoluzione presenti nei video dei nuovi soggetti.
Inoltre si puo notare che i risultati(da 7 a 30), Fig.3.3a, Fig.3.3c, e Fig.3.3e, mostrano una tendenza piu chiara, infatti, rispetto ai risultati sul primo FRMDB, i nuovi risultati mostrano che vi sono piu test con accuratezza migliore e tale discostamento in percetuale è maggiore, cio si verifica nelle tre reti neurali. Si puo concludere che il confronto di gruppi maggiori di immagine migliora l'accuratezza di individuazione di un soggetto.
I risulati dei test 31 a 34, Fig.3.4a, Fig.3.4c, e Fig.3.4e, hanno lo stesso andamento, e confermano che tra tutte le prospettive verticali, quella a +60 si rivela la meno utile.

Modello	**Top1 (%)**	**Top3 (%)**	**Top5 (%)**	**Top10 (%)**
Vgg16	≈ -2,0	≈ -0,3	≈ -1,4	≈ -2,5
Resnet	≈ -3,5	≈ -3,0	≈ -1,3	≈ -2,7
Senet	≈ -4,0	≈ -4,0	≈ -3,0	≈ -5,0

Tabella 3.1: differenza risultati Dataset1 e Dataset2

3.2 Frame Specifici

Per quanto riguarda questo contesto, abbiamo considerato nel confronto, specifici frame di videosorveglianza, cercando di individuare immagini con pose il piu possibile simili a quelle delle foto segnaletiche allo scopo di migliorare l'accuratezza delle reti neurali.
Questa ipotesi è stata eseguita sull'ultimo gruppo di soggetti, ossia i 28 individui aggiungitivi nel FRMDB aggiornato; per alcuni di questi soggetti sono disponibili due video: uno in cui indossano gli occhiali e l'altro in cui non li indossano. I nuovi fotogrammi che sono stati considerati per tutti gli indivudi coprivano l'asse verticale di 0°, e in alcuni individui si era riusciti a prendere anche fotogrammi dall'asse +30. Partendo da questi mezzi abbiamo considerato 4 casi, i primi due in cui si consideravano solo i fotogrammi dell'asse 0°, in cui i fotogrammi mostrano i soggeti prima con gli occhiali e poi senza, negli ultimi due abbiamo preso tutti i fototogrammi che avevamo a disposizione per il confronto; quindi ampliando,per questo caso, il numero di immagini di fotogrammi da 5 a 10. Infine i test che abbiamo esegutito sono stati quelli presentati nel foglio.
I risultati riportati di seguito sono solo della seconda prova, quella in cui si consideravano i fotogrammi dell'asse 0° e in cui gli accessori indossati dagli individui coincideva con quelli presenti nelle fotosegnaletiche, dato che tra tutte era quella che aveva dato le percentuali migliori.

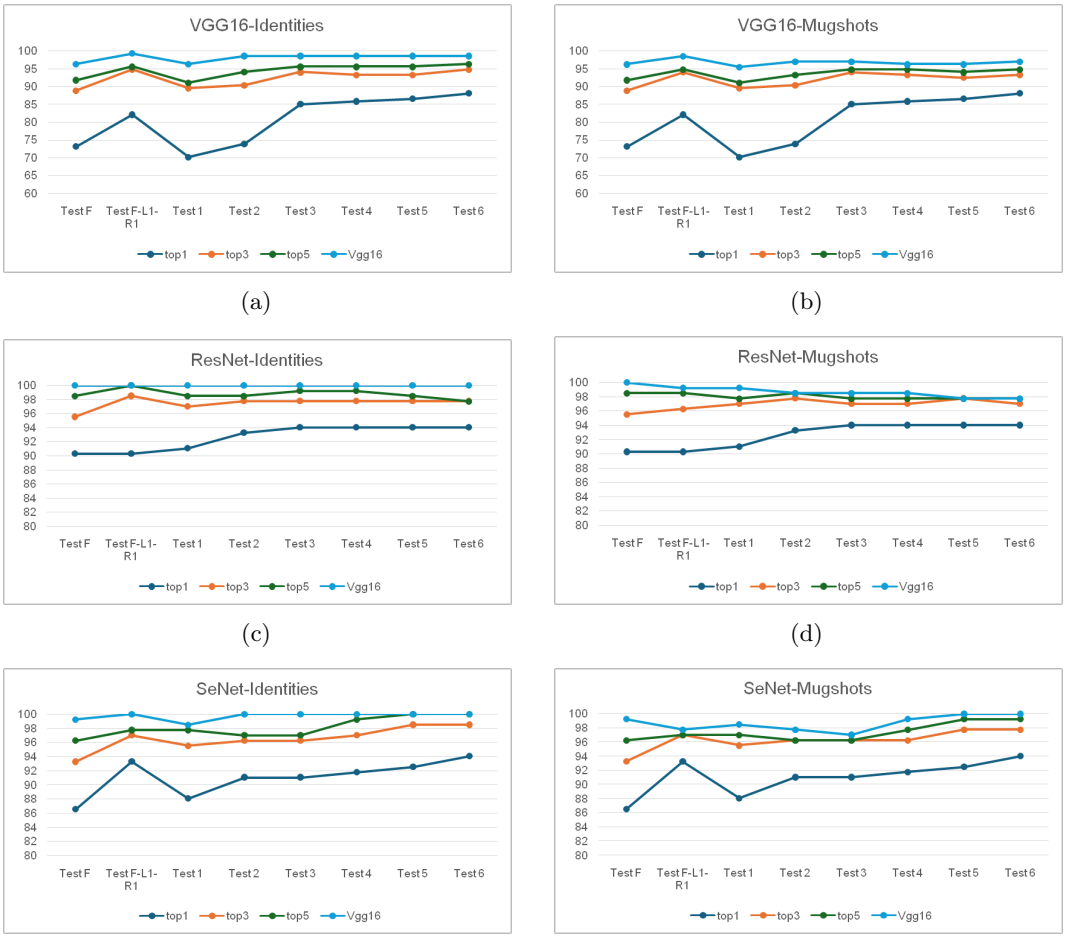


Figura 3.5: Risultati test su Dataset2

Dai risultati si puo concludere, che la ricerca di immagini specifiche per il confronto migliora l'accuratezza in tutte le reti neurali, generando risultati con percentuali elevate, inoltre si conferma l'ipotesi dei test precedenti ossia che gruppi di immagine maggiori sono migliori rispetto alle tradizionali, si puo notare che l'aumento di immagini migliora gradualmente la percentuale. In particolare in queste prove risulta essere SeNet la miglior rete neurale.
In tutte le reti neurali il test con i risultati migliori sono i test-6, ossia quella che comprende il numero maggiori di immagini; tuttavia, dal test 3, sia in vgg16 e resnet, rappresenta il punto di svolta, ossia il test dal quale il numero di immagine aggiunte e la percentuale dell'accuratezza ottenuta inizia a non essere cosi rilevante.

Capitolo 4

Conclusioni

La prima fase dei risultati ha dimostrato che l'impiego di un numero maggiore di foto-segnaletiche per l'addestramento delle reti neurali, indipendentemente da quale si consideri, permette un'individuazione e una classificazione più precisa di un soggetto; mentre la seconda parte dei risultati evidenzia che l'uso di fotogrammi dei video più simili alle fotosegnaletiche per il confronto accrescono significativamente la percentuale di accuratezza. In particolare, dai due risultati emerge che, tra tutti i test effettuati nella tesi, il Test-3 risulta il piu efficace, poiché raggiunge un'elevata precisione con un numero non elevato di foto-segnaletiche; infatti, da queste conclusioni vi sono state delle prime implementazioni da parte dell'impresa SECOM in sistemi di videosorveglianza in cui si considerano il Test3.(implementazione di sistemi in base al)
Infine, sempre nella prima fase dei risultati abbiamo verificato che tra le prospettive verticali che abbiamo considerato (-30, 0, +30, +60),quella a +60 gradi mostra la peggiore precisione(test da 31 a 34) ,e l'inclusione di questa prospettiva ha un impatto negativo sull'accuratezza complessiva, come evidenziato dai test condotti da 26 a 30.

I princpiali sviluppi futuri sono:

- Verificare se tali ipotesi e conclusioni siano ancora valide considerando un dataset più ampio, che includa un maggior numero di soggetti; inoltre, esaminando i cambiamenti di accuratezza con nuove prospettive delle foto-segnaletiche per l'addestramento delle reti neurali che non si discostino molto da +30 e -30 gradi; infine, vedere cosa accade quando si migliora la risoluzione dei video di sorveglianza che a sua volta migliora la risoluzione dei fotogrammi.

- La costruzione di un sistema che automatizzi il processo di individuazione di fotogrammi dai video di sorveglianza, in modo da rilevare pose simili alle foto-segnaletiche [28], dato che tuttora esiste una vasta gamma di studi in questo campo e diverse metodologie, bisognerebbe rilevare quella che garantisca le migliori prestazioni [29]. Una volta che tali fotogrammi vengano individuati, verranno poi utilizzati per il confronto attraverso le reti neurali, finora tale procedura era stata eseguita manualmente.

Bibliografia

[1] Paolo Contardo, Paolo Sernani, Selene Tomassini, Nicola Falcionelli, Milena Martarelli, Paolo Castellini, and Aldo Franco Dragoni. Frmdb: Face recognition using multiple points of view. *Sensors*, 23(4), 2023.

[2] Anil k Jain Stan Z Li. *Handbook of Face Recognition*. Springer, 2st edition, 2011.

[3] Sito Web della Camera dei Deputat. Resoconti delle giunte e commissioni del 5 febbraio 2020. URL = https://www.camera.it/leg18/824?tipo=A&anno=2020&mese=02&giorno=05&view=&commissione=01.

[4] Ministero degli Interni. Procedura volta alla fornitura della soluzione integrata per il sistema automatico di riconoscimento immagini s.a.r.i. https://www.poliziadistato.it/statics/06/20160627-ct-sari–4-.pdf.

[5] Reem Alrawili, Ali Abdullah S. AlQahtani, and Muhammad Khurram Khan. Comprehensive survey: Biometric user authentication application, evaluation, and discussion, 2024.

[6] Anshul Khairwa, Kumar Abhishek, Surya Prakash, and Tej Pratap. A comprehensive study of various biometric identification techniques. In *2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12)*, pages 1–6, 2012.

[7] Shaun Raviv. The secret history of facial recognition. *Wired*, 2023. https://www.wired.com/story/secret-history-facial-recognition/.

[8] Matthew Turk and Alex Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 01 1991.

[9] P. K. Modi and S. I. Patel. A state-of-the-art survey on face recognition methods. *International Journal of Computer Vision and Image Processing*, 12:1–19, 2021.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[11] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE*

Conference on Computer Vision and Pattern Recognition, pages 1701–1708, 2014.

[12] Insaf Adjabi, Abdeldjalil Ouahabi, Amir Benzaoui, and Abdelmalik Taleb-Ahmed. Past, present, and future of face recognition: A review. *Electronics*, 9(8), 2020.

[13] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, March 2021.

[14] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11):e00938, 2018.

[15] Raymond Erz Saragih and Quynh Huong To. A survey of face recognition based on convolutional neural network. *Indonesian Journal of Information Systems*, 4(2), Feb. 2022.

[16] Moacir Antonelli Ponti, Leonardo Sampaio Ferraz Ribeiro, Tiago Santana Nazare, Tu Bui, and John Collomosse. Everything you wanted to know about deep learning for computer vision but were afraid to ask. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, pages 17–41, 2017.

[17] Geet Pithadia. Convolutional neural network: Understanding the intuition behind the layers. *Towards Data Science*, 2020. https://towardsdatascience.com/convolutional-neural-network-1368ee2998d3.

[18] Mayank Mishra. Convolutional neural networks, explained. *Towards Data Science*, August 26 2020. https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939.

[19] Shipra Saxena. Introduction to softmax for neural network. *Analytics Vidhya*, May 24 2024. https://www.analyticsvidhya.com/blog/2021/04/introduction-to-softmax-for-neural-network/.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[21] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[22] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[24] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 67–74, 2018.

[25] Qingge Ji, Jie Huang, Wenjie He, and Yankui Sun. Optimized deep convolutional neural networks for identification of macular diseases from optical coherence tomography images. *Algorithms*, 12(3), 2019.

[26] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[27] Nicolò Rossini. Valutazione di algoritmi per il riconoscimento facciale con diversi sottoinsiemi di foto segnaletiche dalla procedura di foto-segnalazione. https://hdl.handle.net/20.500.12075/16171.

[28] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1087–1096, 2019.

[29] Changxing Ding and Dacheng Tao. A comprehensive survey on pose-invariant face recognition, 2016.