Università Politecnica delle Marche
Scuola di Dottorato di Ricerca in Ingegneria dell'Informazione
Curriculum in Ingegneria Informatica, Gestionale e dell'Automazione

Apprendimento Automatico in ambito Forense

Applicazioni di reti neurali convoluzionali in dattiloscopia, nel riconoscimento della violenza e nei rilievi segnaletici

Tesi di Dottorato di:
Paolo Contardo

Tutor:
Prof. Aldo Franco Dragoni

Coordinatore del Curriculum:
Prof. Franco Chiaraluce

XXXVI ciclo - nuova serie

Università Politecnica delle Marche
Scuola di Dottorato di Ricerca in Ingegneria dell'Informazione
Curriculum in Ingegneria Informatica, Gestionale e dell'Automazione

Apprendimento Automatico in ambito Forense

Applicazioni di reti neurali convoluzionali in dattiloscopia, nel riconoscimento della violenza e nei rilievi segnaletici

Tesi di Dottorato di:
Paolo Contardo

Tutor:
Prof. Aldo Franco Dragoni

Coordinatore del Curriculum:
Prof. Franco Chiaraluce

XXXVI ciclo - nuova serie

Università Politecnica delle Marche
Scuola di Dottorato di Ricerca in Ingegneria dell'Informazione
Facoltà di Ingegneria
Via Brecce Bianche – 60131 Ancona (AN), Italy

...a tutte le vittime di ogni forma di violenza.

Ringraziamenti

In questi tre anni ho avuto la fortuna di essere affiancato da tanti professionisti sia accademici che forensi; un connubio di competenze che mi hanno consentito di crescere sotto ogni profilo. Per questo sento il dovere di ringraziare il dirigente del Gabinetto Interregionale di Polizia Scientifica per le Marche e l'Abruzzo V.Q. dott. Massimiliano Olivieri che, unitamente al mio tutor accademico prof. Aldo Franco Dragoni, hanno dato vita all'accordo d'Intesa tra l'Università politecnica delle Marche e la Polizia di Stato; l'attuale dirigente del Gabinetto Interregionale di Polizia Scientifica per le Marche e l'Abruzzo V.Q. dott.ssa Rita Padovani per il suo impegno positivo nel voler portare avanti la ricerca scientifica all'interno dell'accordo d'intesa; il prof. Paolo Sernani, cuore pulsante all'interno del laboratorio AIRTLab, in assenza del quale questo dottorato di ricerca non avrebbe potuto volare così in alto; i prof.ri Milena Martarelli e Paolo Castellini per il supporto tecnico ricevuto nelle attività di laboratorio; il mio tutor accademico prof. Aldo Franco Dragoni per aver creduto in questo progetto, affidandomi con fiducia l'incarico a condurre la ricerca nei tre ambiti dell'accordo d'intesa, assistendomi in ogni momento con la stessa diligenza del buon padre di famiglia; tutto lo staff AIRTLab pronto ad adoperarsi per agevolare il superamento di ogni difficoltà; e alla mia famiglia per avermi sostenuto in tutto questo periodo.

Ancona, Giugno 2024

Paolo Contardo

Sommario

Lo sviluppo di questo dottorato di ricerca, trae ispirazione da un accordo d'Intesa triennale, iniziato nel 2019, fra il centro di ricerca interdipartimentale C.A.R.M.E.L.O. e la Polizia di Stato, rinnovato di recente e incrementato con un nuovo topic fino al 2026 grazie ai positivi risultati ottenuti. Il primo argomento dell'accordo è stato intitolato "Dattiloscopia 2.0", con l'obiettivo di identificare più efficacemente le persone sospettate di aver commesso dei delitti, dai frammenti d'impronte latenti rinvenute sulla scena del crimine. Nei vari esperimenti effettuati, gli algoritmi di *Computer Vision* sono stati efficaci su impronte acquisite in condizioni ideali, ma fallimentari sui frammenti di impronte latenti, pertanto nel tempo residuo in convenzione, si testerà il deep learning direttamente sui frammenti di impronte digitali, riducendo l'obsolescenza dei metodi attuali come suggerito dall'Unione Europea attraverso le guide ENFSI.

Il secondo argomento è denominato "Fotosegnalamento 2.0", con l'obiettivo di dettare un nuovo protocollo operativo per il fotosegnalamento di Polizia, risparmiando risorse tecnologiche e migliorando il riconoscimento facciale dai video e loro ricostruzione in 3D. In collaborazione con il laboratorio di Misure Meccaniche e AIRTLab, entrambi afferenti all'Università politecnica delle Marche, è stato creato un prototipo per fotosegnalamento multi prospettico con braccio motorizzato, chiamato MCMPrototype, attrezzato con 4 fotocamere angolate verticalmente e gestite da Raspberry 4PI, un impianto di videosorveglianza simulata con 5 web-cams e uno con 3 telecamere full HD. Attraverso questo sistema è stato prodotto il database di fotosegnaletiche FRMDB, unico nel suo genere, con 28 immagini multi posa del volto, fisse e invariabili e le riprese dell'impianto di videosorveglianza, per 67 soggetti. Con i dati acquisiti, attraverso la tecnologia del deep learning, si è mostrato che le CNN sono efficaci a riconoscere volti da video multi prospettiva, purché si disponga di un fotosegnalamento multi posa. In linea con l'obiettivo, tutti i test hanno mostrato che il fotosegnalamento canonico con 2 profili, ha l'accuratezza peggiore, invece c'è un elevato miglioramento a partire da 5 pose in su. Nel tempo residuo in convenzione, si doterà il prototipo di un congruo numero di telecamere Kinect per la modellazione 3D del volto.

Il terzo argomento trattato, chiamato "Riconoscimento di violenza 2.0" prevede come obiettivo l'addestramento di sistemi automatici al rilevamento di violenza da videoriprese, sia per il mantenimento dell'ordine pubblico attra-

verso il controllo in real-time di ampi spazi, sia per il processamento rapido di lunghe registrazioni a sostegno dell'attività investigativa di Polizia Giudiziaria. Preliminarmente, è stato prodotto l'AIRTLab dataset, unico nel suo genere, con 350 clips di scene violente e non violente ma confondibili. Testando 3 modelli di rete neurale convoluzionale 3D, basati su deep learning, il dataset AIRTLab si è confermato efficace per testare la robustezza delle reti neurali verso i falsi positivi. Ma l'importante richiesta di risorse computazionali e di memoria di queste reti, le ha rese poco adattabili a dispositivi mobili. Per ovviare a questo inconveniente, in linea con l'obiettivo, è stato proposto un nuovo modello combinando una rete neurale CNN 2D, progettata per dispositivi integrati, con un layer ricorrente in cascata, ottenendo un calo di prestazioni di solo 1%AUC e 2% Accuracy a vantaggio di leggerezza e memoria. Inoltre, come parte integrante del riconoscimento automatico della violenza, sono stati effettuati alcuni esperimenti per riconoscere atti violenti dai file audio, al fine di poter sviluppare un software come strumento di difesa personale a disposizione delle vittime di violenza, ad esempio utile per contrastare atti di violenza di genere, aiutare le persone che hanno un istinto violento indirizzandole verso percorsi di recupero e prevenire gli effetti negativi sulle vittime indirette.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

La stesura del presente elaborato, è il frutto di un'intensa attività di ricerca multidisciplinare mirata ad offrire uno studio che mostra come le nuove tecnologie possono aiutare le forze dell'ordine (FO) a mantenere o aumentare le garanzie di sicurezza per una migliore convivenza all'interno della Comunità.

Nell'ordinamento italiano, le fonti del diritto costituiscono l'insieme dei dettami normativi volti a garantire la libertà e la civile convivenza tra i cittadini, in tali normative sono indicati anche i "compiti" degli addetti ai lavori come ad esempio le funzioni della polizia giudiziaria (PG).

La polizia giudiziaria, che è composta dalla Polizia di Satato, dai Carabinieri, dalla Polizia Penitenziaria e la Guardia di Finanza, da sempre è impegnata al mantenimento dell'ordine e sicurezza pubblica e, al fine di garantire i diritti di tutti i cittadini, deve intervenire davanti alle violazioni di legge come dettato dall'art.55 del Codice di Procedura Penale (CPP), il quale stabilisce che: *"La polizia giudiziaria deve, anche di propria iniziativa, prendere notizia dei reati, impedire che vengano portati a conseguenze ulteriori, ricercarne gli autori, compiere gli atti necessari per assicurare le fonti di prova e raccogliere quant'altro possa servire per l'applicazione della legge penale"*[6][7].

Tra i corpi delle forze dell'ordine, con funzioni di polizia giudiziaria, la legge ha inteso ricomprendere anche la Polizia Locale e alla Polizia Provinciale.

Un compito di altissima responsabilità e per il quale, con questo studio, ho voluto confermare come l'Intelligenza Artificiale (AI) può essere un vantaggioso supporto per la pubblica sicurezza al servizio della legalità.

L'attività di ricerca qui condotta, proprio per la sua particolare caratteristica multidisciplinare, mi ha portato a frequentare sia il laboratorio di misure meccaniche e termiche dell'Università Politecnica delle Marche (UNIVPM), sia i laboratori del Gabinetto Interregionale di Polizia Scientifica (GIPS) per le Marche l'Abruzzo, ma principalmente gli esperimenti sono stati effettuati all'interno dell'Laboratorio di Intelligenza Artificiale e Sistemi in Tempo Reale (AIRTLab).

AIRTLab è un laboratorio di ricerca scientifica situato presso il Dipartimento di Ingegneria dell'Informazione (DII) dell'UNIVPM, ed è diretto dal prof. Aldo

Franco Dragoni, docente di Informatica e Sistemi Operativi in Tempo Reale presso l'Università politecnica delle Marche.

Le attività attuali del laboratorio sono focalizzate sull'AI e, nello specifico, sul ragionamento basato sulla logica, sui sistemi autonomi distribuiti e sull'uso del Deep Learning (DL) per l'elaborazione di immagini e segnali.

Una delle missioni del laboratorio è capire fino a che punto gli algoritmi possano funzionare entro precisi vincoli temporali, poiché algoritmi intelligenti potrebbero non essere sufficienti senza una soluzione in tempo.

Lo sviluppo di questo dottorato di ricerca, affonda le proprie radici all'interno di un accordo d'Intesa tra la Direzione Centrale Anticrimine della Polizia di Stato e l'Università Politecnica delle Marche e nasce grazie a una dinamica di eventi più avanti descritti.

1.1 Nascita della convenzione DAC-UNIVPM

Da anni il prof. Dragoni collabora con la Polizia di Stato attraverso il GIPS per le Marche e l'Abruzzo, conducendo alcune ricerche scientifiche in ambito biometrico.

Inoltre il prof. Dragoni è membro del centro di ricerca interdipartimentale *Center for Advanced Research on Measurements for Engineering and Life Optimization*(CARMELO), diretto dalla prof.ssa Milena Martarelli.

Grazie alla sinergia instauratasi nel tempo, nasce un accordo d'intesa siglato tra il Ministero dell'Interno, Dipartimento della Pubblica Sicurezza, Direzione Centrale Anticrimine della Polizia di Stato (DAC) attraverso il GIPS per le Marche e l'Abruzzo e l'UNIVPM, attraverso il centro di ricerca interdipartimentale CARMELO [8].

L'accordo d'intesa, firmato nel 2018 dall'allora Capo della Polizia Direttore Generale della Pubblica Sicurezza Franco Gabrielli e dall'allora Maggifico Rettore dell'UNIVPM Sauro Longhi, verte su tre macro-tematiche:

- **Dattiloscopia 2.0:** con l'obiettivo di dettare un nuovo protocollo operativo per l'identificazione dattiloscopica, affinché anche il rilievo di frammenti piccoli d'impronta rinvenuti sulla scena del crimine, con contenuto di dettagli insufficiente rispetto ai requisiti richiesti dalla magistratura italiana attraverso la sentenza n.2559 del 14.11.1959 della II° Sezione della Corte di Cassazione [1], possa essere utile e potenzialmente più discrimi-

[1] "Invero, dopo talune oscillazioni, questa Corte Suprema ha affermato il principio che le emergenze delle indagini dattiloscopiche offrono senz'altro piena garanzia di attendibilità, anche quando esse concernino solo una porzione di dito, sempre che dalle dette indagini risulti la sicurezza dell'identificazione dell'impronta attraverso l'esistenza di almeno 16-17 punti caratteristici uguali per forma e posizione [OMISSIS] conformemente ai risultati delle più moderne ricerche scientifiche, l'indagine identificativa di una persona attraverso le impronte digitali dà piena garanzia di attendibilità senza bisogno di elementi sussidiari

nante per l'identificazione di sospetti, attraverso un approccio statistico sulle ricorsività presenti nel dermatoglifo, incrementando ulteriormente le garanzie identificative con conseguente possibilità di riesame di decenni di casi irrisolti.

- **Fotosegnalamento 2.0:** con l'obiettivo di tracciare un protocollo operativo che minimizzi l'utilizzo di risorse e ottimizzi l'affidabilità del riconoscimento automatico dei volti e la loro ricostruzione tridimensionale. Di fatto attualmente il laboratorio di misure meccaniche e termiche, afferente al centro CARMELO, in cooperazione con il laboratorio AIRTLab, svolge attività di ricerca con l'obiettivo di creare un sistema integrato di fotosegnalamento tridimensionale del soggetto e ottimizzare le procedure di fotosegnalamento bidimensionale per sfruttare al meglio le tecnologie di riconoscimento facciale.

- **Violence detection in video:** con l'obiettivo di addestrare sistemi automatici di rilevamento delle azioni violente da videoriprese attraverso un duplice approccio:
  - ON-LINE: per il controllo in tempo reale di ampi spazi durante grandi affollamenti, come strumento di rapida localizzazione dei focolai violenti a supporto del tempestivo intervento delle forze dell' ordine o come azione preventiva anche in ambito antiterroristico, ai varchi di frontiera o nelle manifestazioni che richiedono una partecipazione selezionata.
  - OFF-LINE: per il processamento di lunghi filmati registrati, a supporto delle indagini di PG per una rapida individuazione degli autori di reati o elementi utili alla ricostruzione della dinamica dei fatti.

Un ventaglio di competenze in cui le parti ne condividono l'interesse comune, impegnandosi a collaborare per lo svolgimento di programmi di ricerca e di formazione, attraverso attività volte a migliorare la realizzazione dei propri compiti istituzionale.

1.2 Tesi Dattiloscopia 2.0

Alla fine del 2018, al termine del percorso di studi della mia Laurea Specialistica presso l'UNIVPM, attraverso il mio professore Nicola Paone, conobbi i professori Tomasini Enrico Primo e il prof. Aldo Franco Dragoni, i quali essendo i principali realizzatori del progetto che ha portato alla firma dell'accordo

di certezza, quando si riscontri l'esistenza di almeno 16-17 punti caratteristici uguali per forma e posizione, anche se le impronte appartengono solo alla porzione di un dito".[9]

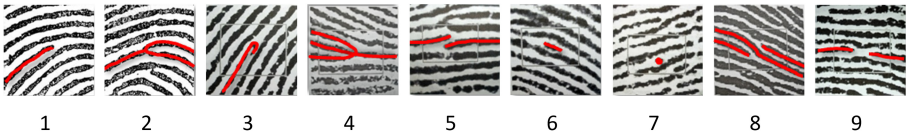


Figura 1.1: **Minutiæ del tipo puro:** *1.terminazione 2.uncinato 3.biforcatura 4.triforcatura 5.deviatura 6.tratto 7.punto 8,interdizione 9.termini speculari*

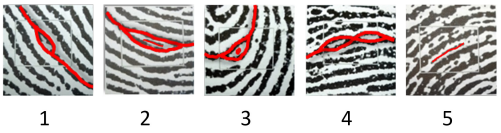


Figura 1.2: **Minutiæ del tipo composito:** *1.occhiello ansuale 2.isolotto ansuale 3.punto ansuale 4.doppio occhiello 5.interlinee ansuali*

d'intesa DAC-UNIVPM, mi proposero di sviluppare la tesi di laurea sul tema Dattiloscopia.

Accettando, condussi uno studio sulla rappresentatività statistica delle composizioni di minutiæ presenti sul dermatoglifo dei polpastrelli delle nostre dita delle mani[10].

Le minitiæ, sono delle accidentalità che si formano nello sviluppo delle linee di cresta dell'impronta digitale e sono generate in modo casuale a partire dal terzo mese di gestazione intrauterina, permangono invariate per tutta la vita anche post mortem fino al completo di disfacimento dell'epidermide [11].

È proprio questa casualità che ci rende tutti diversi e identificabili attraverso le impronte digitali.

In letteratura, nell'analisi dattiloscopica, vengono riconosciute minutiæ di tipo puro e minutiæ del tipo composito come evidenziato da Andrea Giuliano nel libro intitolato "*Dieci e tutte diverse*" [11] e riportate nelle figure 1.1 e 1.2.

Di fatto minutiæ pure e composizioni, ad oggi, vengono considerate più come dato numerico dai dattiloscopisti piuttosto che valutate per la loro rarità statistica, che invece potrebbe essere un dato molto discriminante.

L'obiettivo della tesi era quello di vedere se fosse possibile ridurre il numero dello standard minimo nazionale di corrispondenze di minutiæ per confermare un'identificazione, indicato in figura 1.3, imposto dalla magistratura con la già citata sentenza n.2559, mantenendo o aumentando a miglior garanzia la soglia di variabilità accettata dalla magistratura.

Tale soglia, che pone l'Italia al primo posto come Stato garantista[1], trae origine da alcune considerazioni della magistratura riferite ad un calcolo statistico sulla teoria della frequenza dei punti d'identità nelle impronte papillari desunta

STATO	STANDARD RICHIESTO
ITALIA	**16-17 PUNTI**
CIPRO, TANZANIA	16 PUNTI
FINLANDIA, FRANCIA, GRECIA, POLONIA, PORTOGALLO, ALBANIA, BRASILE, REP. POP. CINESE, GIBILTERRA, HONG KONG, ISRAELE, TURCHIA, SVEZIA, AUSTRIA, SPAGNA, BELGIO, IRLANDA, UCRAINA	12 PUNTI
BAHAMAS	10-16 PUNTI
SLOVENIA, DANIMARCA, OLANDA, UNGHERIA, REPUBBLICA CECA	10 PUNTI
BOSNIA, GERMANIA, ROMANIA	8-12 PUNTI
BULGARIA	8 PUNTI
STATI UNITI, SUD AFRICA	7 PUNTI
GRAN BRETAGNA, SVIZZERA, LUSSEMBURGO, MONACO, NORVEGIA, SLOVACCHIA, AUSTRALIA, CANADA, FINLANDIA, KOSSOVO, MAROCCO, NUOVA ZELANDA, SCOZIA, RUSSIA	NESSUNO STANDARD

Figura 1.3: Valore minimo di corrispondenze, richiesto tra comparazioni di natura digito papillare[1]

Author	P(Fingerprint Configuration)	N=36,R=24,M=72 (N=12,R=8,M=24)
Galton (1892)	$\frac{1}{16} \times \frac{1}{256} \times \left(\frac{1}{2}\right)^R$	1.45×10^{-11} (9.54×10^{-7})
Pearson (1930)	$\frac{1}{16} \times \frac{1}{256} \times \left(\frac{1}{36}\right)^R$	1.09×10^{-41} (8.65×10^{-17})
Henry (1900)	$\left(\frac{1}{4}\right)^{N+2}$	1.32×10^{-23} (3.72×10^{-9})
Balthazard (1911)	$\left(\frac{1}{4}\right)^N$	2.12×10^{-22} (5.96×10^{-8})
Bose (1917)	$\left(\frac{1}{4}\right)^N$	2.12×10^{-22} (5.96×10^{-8})
Wentworth & Wilder (1918)	$\left(\frac{1}{50}\right)^N$	6.87×10^{-62} (4.10×10^{-21})
Cummins & Midlo (1943)	$\frac{1}{31} \times \left(\frac{1}{50}\right)^N$	2.22×10^{-63} (1.32×10^{-22})
Gupta (1968)	$\frac{1}{10} \times \frac{1}{10} \times \left(\frac{1}{10}\right)^N$	1.00×10^{-38} (1.00×10^{-14})
Roxburgh (1933)	$\frac{1}{1000} \times \left(\frac{1.5}{10 \times 2.412}\right)^N$	3.75×10^{-47} (3.35×10^{-18})
Trauring (1963)	$(0.1944)^N$	2.47×10^{-26} (2.91×10^{-9})
Osterburg et al. (1977)	$(0.766)^{M-N}(0.234)^N$	1.33×10^{-27} (1.10×10^{-9})
Stoney (1985)	$\frac{N}{5} \times 0.6 \times (0.5 \times 10^{-3})^{N-1}$	1.2×10^{-80} (3.5×10^{-26})

Figura 1.4: Confronto della probabilità di una particolare configurazione di impronte digitali utilizzando diversi modelli[2]

da un calcolo di Victor Balthazard nel 1911 [2], la cui formula è indicata in figura 1.4, dove 4 indica il tipo di minutiæ (terminazione sinistra, terminazione destra, biforcazione sinistra e biforcazione destra) e N la quantità delle stesse rilevata sull'impronta digitale [11].

Tale riduzione di corrispondenze consentirebbe di allinearsi con gli standard più ricorrenti di altre nazioni europee, riducendo contestualmente il rischio dei "falsi positivi" e "falsi negativi" ma conservando o addirittura aumentando le garanzie probatorie.

In breve, combinando opportunamente tra loro alcune tecniche ben note in statistica, che mettono a fattor comune le proprietà del calcolo combinatorio e la regola del prodotto, elaborai la formula 1.8 in grado di stimare la variabilità probabilistica P del calcolo combinatorio ponderato sulla rarità delle composizioni di minutiæ, in cui ϑ rappresenta l'orientazione della composizione, χ i tipi di composizioni riscontrate e K è la porzione di impronta considerata in mm^2.

Definendo con t_i le quantità della composizione iesima rinvenuta sull' impronta, con la formula 1.1 ho determinato le composizioni totali rilevate.

$$N = \sum_{i=1}^{X} t_i \tag{1.1}$$

A questo punto, per poter applicare P ho dovuto determinare prima i coefficienti specifici per ogni tipologia di composizione di minutiæ, ovvero la significatività specifica S_{si} 1.2, la significatività specifica normalizzata S_{sni} 1.3, il coefficiente statistico specifico C_{ssi} 1.4, la quantità statistica ponderata della composizione iesima t_{spi} 1.5 la cui somma consente di determinare N_{sp} 1.6 cioè

la quantità di minutiæ eqivalente ponderata statisticamente.

$$S_{si} = N \cdot R_{si} \tag{1.2}$$

$$S_{sni} = \frac{S_{si}}{S_{s1}} \tag{1.3}$$

$$C_{ssi} = \frac{1}{S_{sin}} \tag{1.4}$$

$$t_{spi} = t_i \cdot C_{ssi} \tag{1.5}$$

$$N_{sp} = \sum_{i=1}^{X} t_{spi} \tag{1.6}$$

$$R_{si} = \frac{TOT_i}{TOT} \tag{1.7}$$

$$P = (\vartheta \cdot \chi)^{N_{sp}} \cdot \frac{K!}{N_{sp}! \cdot (K - N_{sp})!} \tag{1.8}$$

Prendendo come fonte uno studio[2] condotto dall'Arma dei Carabinieri sulla identificazione dattiloscopica in Italia, in cui sono riportati alcuni dati che rilevano la ricorsività delle composizioni di minutiæ, riscontrata sulle impronte digitali di 74 uomini italiani [3], andai a determinare la rappresentatività statistica R_{si} 1.7 di ogni tipo di composizione e la relativa ricorsività percentuale i cui dati sono riportati in figra 1.5.

A questo punto, ipotizzando un frammento d'impronta in cui vengano rinvenute $t_1 = 5$ biforcazioni, $t_2 = 5$ terminazioni, $t_3 = 1$ occhiello e $t_4 = 1$ tratto, si otterranno $N = 12$ punti caratteristici reali di minutiæ, minori delle 16/17 citate nella nota sentenza 2559 [9]. In realtà si potrebbe obbiettare che scorporando l'occhiello e il tratto, si otterrebbero due biforcazioni e due terminazioni che farebbero salire il valore di N da 12 a 14, che è sempre minore di 16/17 ma la composizione farebbe salire di molto il valore di N equivalente cioè N_{sp}.

Inoltre calcolando con in nuovi dati la variabilità ponderata P 1.8, otteniamo un valore di rappresentatività statistica notevolmente maggiore di quello calcolato con la formula di Victor Balthazard 1.4 preso come riferimento dalla magistratura nella sentenza della Cassazione n.2559 [9].

Nella figura 1.6 è possibile notare visivamente il divario tra la variabilità P calcolata con il calcolo combinatorio ponderato sulla rappresentatività statistica e il calcolo con la formula di Victor Balthazard [10].

[2]Disponibile on-line su: https://www.onap-profiling.org/identificazione-dattiloscopica-la-realta-italiana

PUNTI CARATTERISTICI		TOTALE	R_{si} Rappresentatività Statistica	RARITA'%
	Terminazione	12877	0,484535	48,453492
	Biforcazione	11934	0,449052	44,905178
	Occhiello	757	0,028484	2,848435
	Tratto	551	0,020733	2,073299
	Punto	257	0,009670	0,967038
	Crossover	93	0,003499	0,349940
	Sperone	88	0,003311	0,331126
	Linea interrotta	10	0,000376	0,037628
	Doppia biforcazione	8	0,000301	0,030102
	Triforcazione	1	0,000038	0,003763
	TOTALE	26576	1,000000	100,000000

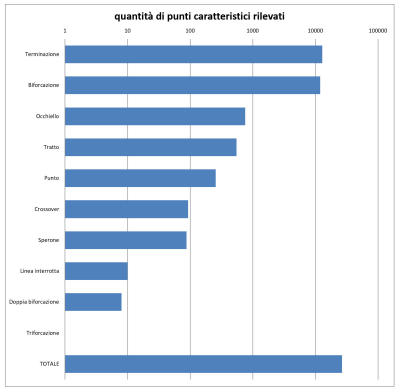


Figura 1.5: Distribuzione statistica, riportata su scala logaritmica in base 10, della rappresentatività delle composizioni di minutiæ su dati rilevati dall'Arma dei Carabinieri[3]

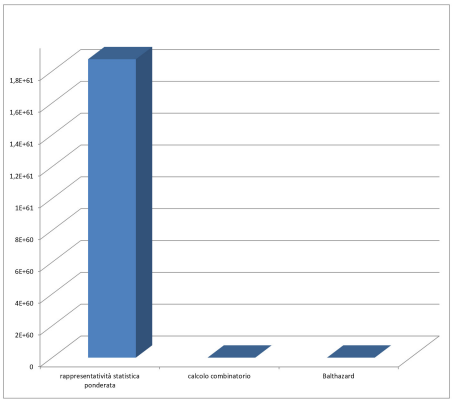


Figura 1.6: Valori della variabilità P calcolata su un ipotetico frammento di impronta digitale contenente n.5 biforcazioni, n.5 terminazioni, n.1 occhiello e n.1 tratto, con il calcolo combinatorio ponderato sulla rappresentatività statistica, il calcolo combinatorio semplice e il calcolo con la formula di Victor Balthazard.

1.3 Nascita del dottorato di ricerca

Alla luce dei risultati ottenuti con la tesi *Dattiloscopia 2.0*, appariva proficuo provare a far diventare lo studio in essere un protocollo operativo utile al dattiloscopista per facilitare un'identificazione quando ci sono pochi punti caratteristici sul frammento dell'impronta digitale da confrontare, ma potendo dare un peso statistico ponderato ad ogni composizione di minutiæ, anche un frammento prima scartabile sarebbe diventato ora probatorio.

Per far questo, per validare i risultati, bisognava dimostrare alla magistratura, tra le altre cose, che lo studio era condotto su un campione statisticamente rappresentativo e quindi ripetere il conteggio delle composizioni di minutiæ su un set di impronte digitali reali di milioni di persone.

Una missione impensabile da eseguire manualmente, ma che se eseguita automaticamente in modo efficace da algoritmi dedicati, replicando l'analisi statistica sui nuovi dati, avrebbe non solo rispettato i requisiti dettati dalla magistratura nella sentenza della Corte di Cassazione n.2559 [9], ma avrebbe facilitato l'identificazione dattiloscopica anche su frammenti d'impronte ritenuti oggi non utili, con possibilità di riapertura di casi irrisolti, consentendo all'Italia di conservare le doti garantiste che già possiede e di allinearsi con gli altri stati europei.

Da qui, il prof. Aldo Franco Dragoni, visto che il protocollo d'intesa tra la Polizia di Stato e il centro CARMELO, era di durata triennale proprio come un corso di dottorato, mi invitò a studiare un progetto sui tre punti del Protocollo d'Intesa, informandomi della possibilità di partecipare al concorso di dottorato.

Capitolo 2

Obiettivi

Come indicato nel capitolo 1.1, questa ricerca verte su tre macro tematiche che coinvolgono la dattiloscopia, il fotosegnalamento di polizia e il riconoscimento automatico della violenza, in ognuna delle quali lo scopo è quello di individuare strumenti e metodi operativi in grado di potenziare l'operato delle Forze dell'Ordine a garanzia della sicurezza della comunità.

In tutti e tre i campi, la strategia generale è quella di applicare i più avanzati sistemi di intelligenza artificiale che possano contribuire al miglioramento dell'efficienza dell'attività delle FO nello svolgimento delle loro funzioni, a beneficio della collettività, riducendo gli ostacoli che deve affrontare la risorsa umana nelle attività forensi d'Istituto.

A tal proposito, unitamente ai ricercatori del laboratorio AIRTLab, ho condotto uno studio su tutti e tre i domini dell'accordo d'intesa [12] analizzando lo stato dell'arte e la letteratura scientifica per capire come la comunità scientifica sta affrontando questi argomenti e quali soluzioni sta adoperando per aumentare la sicurezza. Tra le tecniche di apprendimento automatico per l'analisi dei dati il "deep learning" sta crescendo rapidamente, ottenendo risultati promettenti nel riconoscimento vocale, nell'elaborazione delle immagini, nel riconoscimento di "pattern" e in molti altri domini applicativi. In seguito al successo del deep learning, molte tecniche di analisi automatica dei dati stanno diventando comuni anche nelle forze dell'ordine. A riguardo, sono stati fissati gli obiettivi di questa ricerca, attraverso uno studio riportato nell'articolo che segue, che presenta una panoramica sul potenziale impatto del deep learning su tre domini applicativi peculiari delle forze dell'ordine. In particolare, analizziamo i risultati ottenuti da tecniche basate su deep learning per il riconoscimento dei volti, delle impronte digitali e di scene di violenza all'interno di video. Infatti, la combinazione di:

1. dati provenienti dalla procedura di fotosegnalamento,
2. pervasività dei dispositivi di videosorveglianza,
3. capacità di apprendere da un'enorme quantità di dati,

potrebbe supportare i prossimi passi nella prevenzione del crimine.

2.1 Panoramica su tre domini applicativi

[*pubblicato*][1]

Dalla sua nascita come disciplina, l'Intelligenza Artificiale (IA) mira a capire se siamo in grado di realizzare macchine con la capacità di pensare. Durante questa incessante ricerca, l'IA simbolica, o "Good Old-Fashioned AI" [13], cerca di modellare la conoscenza dei domini applicativi in un formalismo di alto livello che possa essere comprensibile dall'uomo. Innumerevoli applicazioni si basano sull'IA simbolica: dai sistemi di assistenza personali [14, 15] ai sistemi di supporto alle indagini di polizia [16], dalla modellazione di automi [17] e agenti autonomi [18, 19, 20], ai motori di ragionamento per le "smart home" [21, 22, 23] e molti altri. In parallelo all'IA simbolica, l'apprendimento automatico, o "machine learning", cerca di dare alle macchine la capacità di apprendere autonomamente da esempi. All'interno del machine learning, stiamo assistendo alla rapida crescita del "deep learning": esso mira a costruire modelli computazionali, composti da più strati di elaborazione, in grado di apprendere autonomamente le migliori rappresentazioni dei dati per realizzare compiti specifici, come il riconoscimento vocale, il riconoscimento automatico di oggetti in immagini, il riconoscimento di "pattern" e molti altri [24].

Seguendo i progressi raggiunti dall'IA, diversi metodi di analisi dei dati basati sia sull'IA simbolica che sul deep learning stanno diventando popolari anche tra le forze dell'ordine [25]. A tal proposito, questo articolo presenta una panoramica sull'impatto delle tecniche di deep learning su tre domini applicativi comuni nelle forze dell'ordine:

- il riconoscimento di volti, in relazione all'uso di immagini raccolte con il fotosegnalamento, cioè la procedura di acquisizione di un'immagine frontale e una di profilo di una persona, delle sue impronte digitali e delle sue informazioni personali;
- il riconoscimento delle impronte digitali e, in particolare, l'estrazione delle minuzie, cioè le caratteristiche distintive utilizzate per verificare la corrispondenza di impronte digitali;
- il rilevamento automatico di scene di violenza all'interno di video, con l'obiettivo di alleggerire le autorità giudiziarie dalla necessità di controllare manualmente ore di filmati per identificare brevi eventi.

Sebbene questi domini possano sembrare diversi, la loro informatizzazione ha radici comuni nella "Computer Vision" e si sta rapidamente evolvendo grazie al deep learning. Pertanto, l'obiettivo di questo articolo è quello di dare una

[1] Contardo, P., Sernani, P., Falcionelli, N., Dragoni, A. F. (2021, May). Deep Learning for Law Enforcement: A Survey About Three Application Domains. In RTA-CSIT (pp. 36-45).

descrizione sintetica di tale evoluzione, mostrandone il potenziale impatto nelle applicazioni di sicurezza e di prevenzione del crimine.

Il resto dell'articolo è diviso in sezioni dedicate a ognuno dei domini applicativi presentati: il riconoscimento dei volti (Sezione 2.1.2), il riconoscimento delle impronte (Sezione 2.1.1) e il rilevamento di scene di violenza nei video (Sezione 2.1.3). Infine, la Sezione 2.1.4 presenta le conclusioni di questa panoramica, evidenziando alcuni aspetti meritevoli di ulteriori approfondimenti.

2.1.1 Dattiloscopia 2.0

Le impronte digitali, cioè i dermatoglifi creati dalle creste e dai solchi epidermici sulle nostre dita, sono state utilizzate per scopi di identificazione per più di 2000 anni [26]. Poiché le impronte digitali sono una caratteristica biometrica così discriminante, la realizzazione di sistemi di identificazione automatica delle impronte ("Automatic Fingerprint Identification Systems" – AFIS) è stata un argomento di primo piano nella Computer Vision negli ultimi quattro decenni. In tali sistemi, la corrispondenza delle impronte digitali per identificare una persona o verificarne l'identità si basa sulla presenza di singolarità nelle creste epidermiche chiamate minuzie [27]. A questo proposito, gli algoritmi per estrarre le caratteristiche ed verificare la corrispondenza di immagini di impronte digitali si sono concentrati su due tipi fondamentali di minuzie: biforcazioni e terminazioni, cioè i punti in cui una cresta si divide in altre due e dove una cresta finisce [28, 29, 30]. Oltre a problemi come il rumore dell'immagine, le distorsioni, le rotazioni e gli spostamenti, la grande variabilità tra diverse impronte dello stesso dito e al contempo la somiglianza tra due immagini di dita diverse rendono la verifica della corrispondenza delle impronte digitali una sfida molto impegnativa [31].

Gli algoritmi basati su Computer Vision tradizionale hanno dimostrato la loro efficacia sulla verifica della corrispondenza delle impronte digitali, e in particolare, della corrispondenza delle minuzie, evolvendo nel corso degli anni. Per esempio, nel 1997, Maio e Maltoni [28] hanno proposto un algoritmo basato sul "line following" applicato alle linee di cresta in immagini di impronte digitali in scala di grigio, al fine di identificare terminazioni e biforcazioni. Farina et al. [29] hanno proposto di identificare le minuzie da immagini binarie scheletrate. Fronthaler et al. [30] hanno sfruttato delle caratteristiche di simmetria (lineare e parabolica) per ridurre il rumore ed estrarre le minuzie su immagini in scala di grigi. Cappelli et al. [32] hanno proposto una nuova rappresentazione per le minuzie, trattandone l'estrazione e il riconoscimento delle impronte digitali come un problema di "pattern matching" 3D invece che 2D, ottenendo ottimi risultati di accuratezza.

Naturalmente, questi sono solo alcuni esempi dei molti algoritmi e tecniche disponibili nel riconoscimento delle impronte digitali. Infatti, come evidenziato nell'indagine di Peralta et al. [31], anche se gli algoritmi più accurati sono tra loro diversi, si basano su caratteristiche comuni come le coordinate, l'angolo e il tipo di minuzie. Qual è, dunque, il ruolo del deep learning nel riconoscimento delle impronte digitali, data la maturità del settore e le buone prestazioni degli algoritmi basati sulla Computer Vision tradizionale? Negli ultimi anni, le tecniche basate sul deep learning si sono dimostrate utili per superare alcuni dei limiti delle tecniche tradizionali. Mentre gli algoritmi tradizionali, come quelli presentati, si comportano bene su impronte raccolte con sensori dedicati, in condizioni ideali, hanno fallito su impronte latenti, cioè impronte parziali impresse involontariamente sulle superfici [33, 34, 35, 36]. A questo scopo, Tang et al. [34] hanno proposto di convertire le operazioni tradizionali di estrazione delle minuzie in una CNN che può essere addestrata "end-to-end". Allo stesso modo, Cao et al. [36] hanno presentato un sistema di riconoscimento di impronte digitali latenti basato su CNN. Anche Li et al. [35] hanno proposto un'architettura basata su CNN, ma con un obiettivo diverso: migliorare le immagini di impronte digitali latenti da utilizzare per il riconoscimento, da effettuare però con uno dei tanti algoritmi esistenti.

Il riconoscimento di impronte latenti e parziali non è l'unica sfida di questo campo che viene affrontata con il deep learning. Nell'uso delle impronte digitali per l'autenticazione, Lin e Kumar [37] hanno presentato un modello basato su CNN per apprendere rappresentazioni 3D delle impronte digitali, da usare in applicazioni di riconoscimento delle impronte senza contatto. Con la disponibilità di scanner ad alta risoluzione, sono state sviluppate architetture basate su CNN per riconoscere i pori del sudore nelle immagini di impronte digitali ad alta risoluzione [38, 39]. Infine, si stanno studiando tecniche di deep learning per rilevare tentativi malevoli di autenticazione tramite impronte digitali artificiali, al fine di sviluppare metodi "anti-spoofing" [40, 41].

2.1.2 Fotosegnalamento 2.0

Il fotosegnalamento è una procedura di routine delle forze di polizia basata sulla raccolta di due fotografie, impronte digitali e informazioni personali di un soggetto. Sulla base di questa definizione, c'è una chiara connessione tra il riconoscimento automatico dei volti e il fotosegnalamento, per quanto concerne l'identificazione del volto. Non a caso, il riconoscimento del volto è una delle tecniche biometriche più naturali utilizzate per l'identificazione [42]. Ha infatti un vantaggio significativo rispetto ad altre tecniche biometriche: può essere fatto passivamente, cioè senza azioni esplicite da parte del soggetto da identificare [43]. Pertanto, a causa della vasta gamma di possibili applicazioni in

ambito di sicurezza, il riconoscimento automatico di volti ha attirato l'interesse dei ricercatori in Computer Vision per più di quarant'anni.

Pertanto, le prime tecniche di riconoscimento automatico di volti apparse in letteratura scientifica sono basate su metodologie attinenti alla Computer Vision tradizionale. In questo senso, Turk e Pentland [44] hanno proposto "Eigenfaces", cioè l'applicazione della "Principal Component Analysis" (PCA) per estrarre un vettore di caratteristiche che massimizzano la varianza in un insieme di immagini di volti di addestramento. Proiettando un'immagine di un volto nello spazio ottenuto con la PCA, l'identificazione del viso può essere eseguita con un metodo di tipo "nearest neighbor", calcolando la distanza del volto da riconoscere rispetto alle immagini di addestramento. Mentre Eigenfaces massimizza la varianza interclasse tra le immagini dei volti di soggetti diversi, non tiene conto della varianza intraclasse tra le immagini dei volti di un singolo soggetto. Al contrario, il metodo "Fisherfaces" [45] aggiunge alla PCA una "Linear Discriminant Analysis" (LDA), per minimizzare la varianza intraclasse. A differenza di Eigengaces e Fisherfaces, Ahonen et al. [46] hanno proposto di calcolare i "Local Binary Patterns Histograms" (LBPH) sulle immagini dei volti, dividendole in regioni per calcolare i "Local Binary Patterns" (LBP). Analogamente a Eigenfaces e Fisherfaces, una funzione di distanza basata sugli LBPH estratti può essere utilizzata per eseguire l'identificazione di un volto.

Le tecniche descritte (e quelle direttamente derivate) hanno dimostrato una buona accuratezza nel riconoscimento di volti appartenenti a dataset in cui alcuni parametri come la posa del volto, l'illuminazione e l'espressione sono fissati. Tuttavia si sono dimostrate insufficienti per estrarre caratteristiche discriminanti stabili e invarianti ai cambiamenti presenti in applicazioni reali [47], come nelle immagini ottenute da video e telecamere di sorveglianza. Non risultano pertanto adatte per applicazioni attinenti alle forze dell'ordine, per esempio per confrontare le due immagini del fotosegnalamento, raccolte in condizioni ideali, con le immagini ottenute da video. Al contrario, le tecniche basate sul deep learning si sono dimostrate capaci di estrarre caratteristiche che sono invarianti al cambiamento delle condizioni di espressione facciale, illuminazione e posa. Infatti, sebbene siano state presentate tecniche basate sulla combinazione di reti neurali e revisione delle conoscenze [48, 49] prima della popolarità del deep learning, solo con la proposizione di reti neurali convoluzionali ("Convolutional Neural Networks" – CNN) l'accuratezza nel riconoscimento dei volti senza alcun vincolo ha subito un netto miglioramento. A tal proposito, Taigam et al. [50] hanno presentato DeepFace, una CNN a 8 strati per elaborare immagini di volti di dimensioni 152 x 152 pixel a 3 canali, ottenendo una precisione del 97,35% sul dataset "Labeled Faces in the Wild" (LFW) [51]. Similmente, Schroff et al. [52] hanno proposto Facenet, una CNN a 22 strati

addestrata in diversi esperimenti con un numero variabile di immagini di volti, tra 100 e 200 milioni, appartenenti a 8 milioni di soggetti diversi. Hanno ottenuto il 99,63% di accuratezza su LFW, usando immagini di input di dimensione 220 x 220 pixel. Cao et al. [53] hanno mostrato l'efficacia di ResNet-50 [54], una CNN a 50 strati basata sul "residual learning" capace di ottenere un errore di identificazione top-1 del 3,9% sul dataset VGGFace2 (composto da oltre 3 milioni di immagini di più di 9 mila soggetti).

Le tecniche basate su CNN per il riconoscimento dei volti elencate in questa Sezione sono solo alcuni esempi tra i molti che hanno dimostrato la loro robustezza a condizioni mutevoli e immagini di volti senza alcun vincolo (si veda Guo e Zhang [55] per una lista dettagliata di tecniche di riconoscimento dei volti basate su deep learning). In ogni caso, per quanto ne sappiamo, in letteratura scientifica mancano studi per capire fino a che punto tali tecniche sono efficaci nell'identificare un soggetto noto quando solo le due immagini del fotosegnalamento sono disponibili in fase di addestramento dell'algoritmo di riconoscimento.

2.1.3 Riconoscimento di violenza 2.0

La crescente disponibilità di tecnologie per la videosorveglianza, unita alla necessità di alleggerire le autorità dal compito di controllare ore di registrazioni video, ha stimolato l'attenzione di gruppi di ricercatori verso il rilevamento automatico della violenza nei video. Il rilevamento di scene di violenza è considerato parte del più generale campo del riconoscimento di azioni umane: nello specifico, si tratta di un problema binario che consiste nel riconoscere la presenza o l'assenza di violenza all'interno di video [4].

I primi lavori apparsi in merito al riconoscimento di scene di violenza sono basati su tecniche di Computer Vision già sviluppate per il riconoscimento delle azioni e possono essere categorizzati in due classi [56], sulla base delle caratteristiche estratte per rappresentare le azioni stesse:

- nelle tecniche basate sull'estrazione di caratteristiche locali, la rappresentazione di un'azione è calcolata utilizzando i punti di interesse ("Point of Interests" – POI) dai singoli fotogrammi di un video;

- nelle tecniche basate sull'estrazione di caratteristiche globali, la rappresentazione di un'azione è calcolata valutando le caratteristiche di più fotogrammi nel loro insieme.

Tra le tecniche che si basano su caratteristiche locali, Chen e Hauptmann [57] hanno proposto MoSIFT, una tecnica che combina la "Scale-Invariant Feature Transform" (SIFT) [58] con l'"Optical Flow", per rappresentare il movimento

dei punti di interesse. Xu et al. [56] hanno fatto evolvere MoSIFT combinandolo con una "Kernel Density Estimation" (KDE) non parametrica per rimuovere le caratteristiche ridondanti e irrilevanti: usando lo "sparse coding" per rappresentare le caratteristiche estratte, hanno infatti ottenuto buoni risultati nel rilevare la violenza tra singoli individui nei video. Invece, Deniz et al. [59] hanno proposto di calcolare l'accelerazione dei movimenti a partire dallo spettro di potenza di fotogrammi adiacenti, al fine di rilevare una grande variazione di velocità, ottenendo risultati comparabili a MoSIFT, ma con un algoritmo più veloce in fase di esecuzione.

Per quanto riguarda le tecniche basate su caratteristiche globali, Hassner et al. [60] hanno proposto il calcolo dei descrittori "Violence Flows" (VIF), un'evoluzione dell'Optical Flow che calcola le variazioni di grandezza dei vettori di flusso, ottenendo risultati promettenti sul rilevamento della violenza nelle folle. Gao et al. [61] hanno aggiunto ai descrittori VIF l'orientamento del vettore di flusso, proponendo OVIF, migliorando le prestazioni del rilevamento della violenza tra singoli individui, ma con una minore accuratezza nel rilevamento della violenza nelle folle.

Il deep learning ha contribuito a far progredire il campo del rilevamento della violenza superando alcune delle limitazioni dell'Optical Flow, come le discontinuità e i movimenti della videocamera, nonché ottenendo ottime prestazioni sia nel rilevamento della violenza tra singoli individui che della violenza nelle folle, utilizzando lo stesso modello. In particolare, le CNN 3D si sono dimostrate capaci di apprendere informazioni spazio-temporali, cioè caratteristiche che rappresentano le informazioni di movimento in un video, oltre alle informazioni spaziali in un singolo frame. Per esempio, Ding et al. [62] hanno presentato una CNN 3D a 9 strati per il rilevamento della violenza, ottenendo un'accuratezza del 91% sul dataset "Hockey Fight" [63]. Con un approccio simile, Li et al. [64] hanno raggiunto il 98,3% di accuratezza sul dataset "Hockey Fight" e il 97,2% sul dataset "Crowd Violence" [60], sviluppando una CNN 3D a 10 strati che alterna strati completamente connessi e di transizione a seguito di uno strato di convoluzione. Anche metodologie basate sul "transfer learning", facendo uso di CNN 3D pre-addestrate, hanno portato a ottimi risultati. Ad esempio, in ricerche precedenti [4] abbiamo ottenuto un'accuratezza del 98,5% e del 99,2% su "Hockey Fight" e "Crowd Violence" usando C3D [5], una CNN 3D pre-addestrata a classificare categorie di sport, come estrattore di caratteristiche e, in cascata, un classificatore SVM ("Support Vector Machine"). Allo stesso modo, Ullah et al. [65] hanno usato C3D come estrattore di caratteristiche, ma aggregandola a strati completamente connessi per la classificazione, con buone prestazioni in entrambi i dataset "Hockey Fight" (96% di precisione) e "Crowd Violence" (98%). Oltre alle CNN 3D, anche l'architettura di tipo ConvLSTM [66] si è dimostrata efficace nel rilevamento della violenza. A tal

fine, Sudhakaran e Lanz [67] hanno proposto di combinare le informazioni spaziali estratte dai fotogrammi da una CNN 2D con una ConvLSTM, ottenendo un'accuratezza del 97,1% sul dataset "Hockey Fight" e del 94,5% sul dataset "Crowd Violence".

Pertanto, diverse tecniche basate su deep learning hanno dimostrato la loro accuratezza sui dataset tradizionalmente usati in letteratura, come l'"Hockey Fight" e il "Crowd Violence". Tuttavia ci sono ancora ricerche in corso per validare la loro robustezza contro i falsi positivi [68], e l'accuratezza su filmati da reali circuiti di videosorveglianza [69].

2.1.4 Riflessioni sui tre domini applicativi

Abbiamo presentato una breve panoramica circa le applicazioni di deep learning su tre domini applicativi collegati alle forze dell'ordine: riconoscimento dei volti, riconoscimento delle impronte digitali e rilevamento di scene di violenza all'interno dei video. Questi tre domini hanno alcune caratteristiche comuni. Infatti, i primi metodi apparsi in letterature scientifica per l'informatizzazione e l'automazione di tali domini sono tutti radicati nella Computer Vision, utilizzando tecniche come la Principal Component Analysis, la binarizzazione e la scheletrizzazione delle immagini, l'Optical Flow, ecc. Tuttavia, l'uso di tecniche di deep learning, come le reti neurali convoluzionali (2D e 3D) e le ConvLSTM, ha migliorato significativamente l'accuratezza delle applicazioni automatiche che si occupano del riconoscimento dei volti, delle impronte digitali e della rilevazione della violenza.

Nonostante alcune di queste tecniche di deep learning siano già state integrate in sistemi commerciali, almeno per il riconoscimento dei volti e delle impronte digitali[2], c'è la necessità di studiare ancora più a fondo il loro potenziale impatto nelle applicazioni del mondo reale. Ad esempio, per quanto riguarda il riconoscimento dei volti, c'è bisogno di ricerca per capire l'efficacia dell'identificazione dei volti quando solo le due immagini derivanti dal fotosegnalamento sono disponibili per l'addestramento degli algoritmi. Per quanto riguarda il riconoscimento delle impronte digitali, sono in corso studi per ottenere un'estrazione efficace delle minuzie dalle immagini delle impronte latenti, quelle solitamente rilevate nelle scene del crimine. Per quanto riguarda il rilevamento della violenza, l'accuratezza delle tecniche di deep learning su video di reali reti di videosorveglianza e la loro robustezza ai falsi positivi sono tra gli obiettivi della ricerca del momento.

Inoltre, per essere efficaci in applicazioni reali, le tecniche basate sul deep learning (e quelle basate sull'IA in generale) hanno bisogno di prendere in

[2]Si veda, per esempio, il sistema italiano "SARI", un'estensione di un sistema di identificazione automatica delle impronte digitali (AFIS) che supporta il riconoscimento dei volti [70].

considerazione i tempi di esecuzione. Infatti, come sottolineato in [71], una risposta intelligente conserva la sua importanza solo se data in tempo. Infine, poiché le prove raccolte usando l'IA dovrebbero essere spiegabili a un giudice in un tribunale [25], anche i metodi di Explainable AI (XAI), capaci di fornire spiegazioni comprensibili agli umani dei loro risultati [72], dovrebbero essere studiati nei domini applicativi presentati, per evitare l'uso delle tecniche basate su deep learning come semplici modelli "black box".

Capitolo 3

Intelligenza Artificiale in Giurisprudenza

L'Intelligenza Artificiale, in ogni sua forma e a grande velocità, si sta imponendo con prepotenza un po' in tutti i settori sella società proprio per la sua importanza.

Questo sviluppo così diffuso, tende ad allargarsi su ampia scala, tra le piattaforme di servizi digitali, nella mobilità, nei dispositivi che portiamo sempre con noi come anche negli elettrodomestici, animando grandi dibattiti sugli spazi pubblici e comunque ovunque dove c'è trattamento di dati.

Molti nuovi mezzi di comunicazione e servizi, si dotano sempre di più di tecnologie basate sull'intelligenza artificiale, come fare acquisti o accedere alle informazioni online, e fanno ormai parte della nostra vita quotidiana con un frenetico sviluppo in costante evoluzione.

Ma che succederebbe se una di queste tecnologie fosse così evoluta da poter decidere sulla libertà degli individui, sostituendosi al ruolo del giurista?

Le macchine stanno evolvendo ad una velocità quasi incontrollabile, acquistando sempre di più certe abilità che tendono all'approssimazione umana, superandolo in termini di tempo, addestrandole su dati che si riferiscono ad esperienze passate.

Un giurista, non arriva ad una sentenza attraverso una mera analisi di dati, esso deve avere anche le competenze che gli consentono di capire lo stato attuale dei fatti, capire gli stati d'animo, le emozioni e tutti quei fattori emotivi che gli consentono di cogliere la coerenza dei ragionamenti portati in un'aula di tribunale, perché anche se i capi d'imputazione sono i medesimi, in realtà ogni fatto accaduto nasconde in se circostanze sulle quali il giudice non può esimersi dal formulare delle ipotesi con un dedicato grado di ragionevolezza [73].

Dopo gli anni ottanta, in cui si ebbe un rapido sviluppo della *logica fuzzy* nei sistemi informatici [74], proprio grazie alle sue doti di approssimazione, si iniziarono a generare dei sistemi esperti in grado di indirizzare l'uomo verso la scelta di più opzioni, defferentemente dai metodi precedenti in cui l'uso degli

algoritmi software nel processamento di dati portavano a una soluzione netta e non approssimata.

La macchina non dava soluzioni ma iniziava a supportare l'uomo nel poter prendere la decisione più vicina alla soluzione ritenuta migliore, tali sistemi presero il nome di *decision support systems* (DSS)[75].

Oggi l'uso di sistemi esperti in ambito giuridico, è già in pratica sia per reperire informazioni giuridiche, ma anche come supporto ad esempio per la redazione di atti o nell'attività giuridica, purche sussistano due condizioni fondamentali: ci deve essere razionalià nel processo di elaborazione giuridica e chiarezza nell'individuazione e traduzione in linguaggio machina delle fonti del diritto [76].

Quanto detto, nasconde delle difficoltá intrinseche, difficilmente formalizzabili, che rende arduo il compito della macchina, perchè essa dovrebbe considerare una molteplicità di variabili interpretative sia concrete (gli ordinamenti giuridici sono molto complessi nella loro struttura, i regolamenti spesso sono articolati e correlati tra loro, i linguaggi usati nelle sentenze non seguono un protocollo predefinito), cioè che appartengono a fonti scritte, sia astratte come le informazioni che derivano dall'esperienza o le articolazioni linguistiche non classificate del linguaggio naturale, cioè quelle per cui non esiste una grammatica scritta [77].

Quindi la macchina deve possedere tutta la conoscenza necessaria affinche possa operare non solo calcoli ma anche formulare deduzioni logiche sul patrimonio informativo posseduto [76].

3.1 AI nel Diritto in Italia

Anche in Italia poco prima degli anni ottanta, ha iniziato a diffondersi l'uso dei sistemi esperti nell'amministrazione della giustizia.

La fonte che costituiva il bacino di conoscenza della macchina era di solito ben circoscritta in uno specifico settore, anche se a volte poteva integrare dei concetti o casi già riferiti in giurisprudenza [75].

In quel periodo iniziarono a svilupparsi alcuni sistemi come Methodus, Sefit, Lexis, Iri, Remida, ma negli anni 90, il giudice Carmelo Asaro, all'epoca Sostituto Procuratore presso il Tribunale di Lucca, nel tempo libero sviluppò da solo *DAEDALUS*, un software per assisterlo e convalidare ogni attività intrapresa, dandogli supporto passo dopo passo in tutte le fasi del processo, sia durante la procedura d'indagine che poi nel ruolo di Pubblico Ministero [78].

DEDALUS si dimostrò nel tempo un software molto versatile, infatti dopo qualche anno dalla sua uscita, anche i giudici iniziarono un percorso di sperimentazione promuovendo il progetto *DEDALUS Cassazione*, nato proprio dall'esigenza della sezione penale della Corte di Cassazione, per gestire i dati

riguardo ai reati e per calcolare i tempi di prescrizione e sacdenza delle misure cautelari [79].

Ma solo a partire dal 2018 l'Italia ha iniziato a prevedere un piano normativo per regolamentare lo sviluppo nazionale dell'Intelligenza Artificiale, rientrante nel Piano Coordinato della Commissione Europea.

Di fatto il 24 novembre 2021, il Governo Italiano presentò il *Programma Strategico Intelligenza Artificiale 2022-2024*[1] [80], in cui dichiara il proprio impegno a spendere investimenti su undici settori ritenuti prioritari, ponendo in atto una politica strutturata sullo sviluppo dell'Intelligenza Artificiale.

Tale sviluppo affonda le proprie fondamenta in cinque princípi guida:

- L'IA italiana è un IA europea;
- L'Italia sarà un polo globale di ricerca e innovazione dell'IA;
- L'IA italiana sarà antropocentrica, affidabile e sostenibile;
- Le aziende italiane diventeranno leader nella ricerca, nello sviluppo e nell'innovazione basata sull'IA;
- La pubblica amministrazione italiana governerà l'IA e con l'IA.

Questi princípi devono guidare il paese verso il raggiungimento di sei obiettivi:

1. rafforzare la ricerca di frontiera nell'IA;
2. ridurre la frammentazione della ricerca sull'IA;
3. sviluppare e adottare un'IA antropocentrica e affidabile;
4. aumentare l'innovazione basata sull'IA e lo sviluppo della tecnologia di IA;
5. sviluppare politiche e servizi basati sull'IA nel settore pubblico
6. creare, trattenere ed attrarre ricercatori di IA in Italia.

Una strategia di sviluppo che dovrà diffondersi coinvolgendo l'area dei talenti e delle competenze, il settore della ricerca, e il campo delle applicazioni [80].

Un'altro sistema esperto, di più recente origine, in uso nel sistema giudiziario italiano, è "Toga"[2], un database che contiene tutte le fattispecie delittuose disciplinate da Codice Penale e dalle Leggi Speciali.

Consultando il sito internet del produttore, il software Toga promette un enorme risparmio di tempo velocizzando tutta una serie di attivitá correlate al

[1] Disponibile on-line su: https://assets.innovazione.gov.it/1637777289-programma-strategico-iaweb.pdf

[2] Disponibile on-line su: https://toga.cloud/

processo giuridico: come verifica della competenza, procedibilitá, ammissibilitá a riti alternativi, termini prescrizionali e può effettuare calcoli per determinare la durata delle misure cautelari e delle pene da comminare.

Un'altra tecnologia che si sta ricavando uno spazio nel campo del diritto, è l'intelligenza artificiale predittiva, ma più in forma di prevenzione dei reati.

L'assistente capo della Polizia di Stato Marco Venturi, sviluppò nel 2004 un software per predire la commissione di reati, chiamato *KeyCrime*[3] il quale, facendo una stima probabilistica sulla base di informazioni raccolte in occasioni di crimini seriali, ipotizza le zone e il periodo temporale in cui potrebbero ricommettersi certi tipi di reati.

Uno strumento digitale che aiuta gli investigatori a comprendere le potenziali mosse successive da parte dell'autore o degli autori di reati.

Ciò consente alle forze dell'ordine, deputate al controllo del territorio, di concentrare la loro attività proprio in quelle zone individuate come possibili bersagli di attacchi criminali.

Il software, ha iniziato la sua sperimentazione per la prima volta nella Questura di Milano nel 2007 [81].

In soli sette anni, ci fu un aumento di casi risolti di rapine che passò dal 27% ad oltre il 61% e fino all'81% riguardo alle rapine alle farmacie (dati Federfarma), in un'area metropolitana popolata da oltre 1.330.000 abitanti su una superficie di 181,67 km^2 [81].

Un risultato sicuramente positivo che può portare benefici sia sotto il profilo economico, attraverso l'ottimizzazione delle risorse, sia sotto il profilo sociale con l'aumento della percezione della sicurezza.

3.2 Aspetti normativi nell'Unione Europea

Nel 2010 è stata lanciata l'*agenda digitale europea* [82] che ha stabilito per la prima volta la grande importanza delle Tecnologie dell'Informazione e della Comunicazione per il raggiungimento degli obbiettivi prefissati dell'Europa.

In essa è indicata la strategia digitale per la creazione di spazi e servizi digitali sicuri, per la creazione di condizioni di parità sui mercati digitali con le grandi piattaforme, nonché per il rafforzamento della sovranità digitale dell'Europa, contribuendo nel contempo all'obiettivo europeo della neutralità climatica entro il 2050.

Nella comunicazione n.237 [83] del 25 aprile 2018, la Commissione Europea ha dato una sua definizione all'intelligenza artificiale[4] suddividendo la sua sussistenza in due tipologie: o software che agiscono nel mondo virtuale, come nel

[3] Disponibile on-line su: https://keycrime.com/

[4] "L'Intelligenza artificiale' (IA) indica sistemi che mostrano un comportamento intelligente analizzando il proprio ambiente e compiendo azioni, con un certo grado di autonomia, per raggiungere specifici obiettivi".

riconoscimento dei volti, nel comando vocale ecc., oppure come tecnologia a bordo macchina, ad esempio nella guida autonoma, o internet delle cose.

Due mesi dopo, la Commissione formò il gruppo AI-HLEG (*High Level Expert Group on Artificial Intelligence*) composto da 52 esperti selezionati in ambito accademico, civile e industriale, con mansioni di coordinamento della strategia mirata allo sviluppo dell'Intelligenza Artificiale nell'Unione Europea.

Uno dei suoi contributi, ad oggi ancora considerato come linea guida, fu quello di dettare degli orientamenti per un'AI affidabile [84], basandosi su tre principi fondamentali:

1. Legalità: l'Intelligenza Artificiale deve rispettare tutte le leggi e i regolamenti;
2. Eticità: l'Intelligenza Artificiale deve rispettare tutti i principi etici;
3. Robustezza: in senso tecnico e sociale, i sistemi di Intelligenza Artificiale non devono arrecare danni, neanche non voluti.

Sulla base di questi tre principi, con la comunicazione n.168 [85], nel 2019 la Commissione accolse il concetto di affidabilità e 7 requisiti chiave che l'IA deve soddisfare, individuati dal gruppo di esperti del AI-HLEG, che sono:

1. intervento e sorveglianza umani,
2. robustezza tecnica e sicurezza,
3. riservatezza e governance dei dati,
4. trasparenza,
5. diversità, non discriminazione ed equità,
6. benessere sociale e ambientale,
7. accountability.

Appare evidente che la strategia individuata dalla Commissione Europea per promuovere lo sviluppo dell'AI, riguarda l'individuazione di garanzie a sostegno dei valori etici più che gli aspetti meramente tecnologici dell'AI.

Infatti il 21 aprile del 2021, la Commissione Europea ha presentato al Parlamento Europeo la proposta di Regolamento n.206 [86] con cui si da esito all'impegno politico dichiarato dalla presidente Ursula Von Der Leyen, la quale annunciò, tra i vari orientamenti politici per la Commissione 2019-2024 pubblicati nel libro "*Un'Unione più ambiziosa*" [87], che la Commissione avrebbe presentato una normativa per un approccio europeo coordinato alle implicazioni umane ed etiche dell'intelligenza artificiale.

Proprio in virtù di tale annuncio, la Commissione pubblicò il Libro Bianco sull'intelligenza Artificiale, uscito il 19 febbraio 2020, nel quale si esalta proprio l'inclinazione dell'Europa all'eccellenza e alla fiducia verso l'Intelligenza Artificiale [88].

La proposta di Regolamento in parola, si sforza di conciliare le esigenze dell'Intelligenza Artificiale con le indicazioni del Garante della Privacy, che sono pubblicate nel Regolamento n.679 del 27 aprile 2016, dove al Capo II, all'articolo 5, è riportato un elenco di principi fondamentali riservati al corretto trattamento dei dati personali, lasciando comunque al titolare la responsabilizzazione, ovvero la competenza sul rispetto di tali principi e il grado di comprovarlo [89].

Al Capo 2 del Titolo III della proposta di Regolamento, la Commissione affronta un'altro importantissimo aspetto dell'IA, che riguarda i sistemi ad alto rischio [86].

In tale Capo sono dettati una serie di requisiti volti a definire un processo gestionale che i sistemi di Intelligenza Artificiale ad alto rischio devono possedere.

In linea generale, l'articolo n.9 indica che ogni sistema d'Intelligenza Artificiale ad alto rischio, deve prevedere un processo iterativo continuo di varie misure gestionali, suddivise per fasi, che deve essere costantemente aggiornato al fine di indentificare tempestivamente ogni eventuale rischio.

Questo processo gestionale prevede quattro fasi:

1. ogni sistema di Intelligenza ad alto rischio, possiederà dei rischi prevedibili noti che devono essere identificati e analizzati;

2. quando un sistema di Intelligenza Artificiale ad alto rischio è in uso come previsto e quando lavora in condizioni improprie ma ragionevolmente prevedibili, deve essere operata comunque una valutazione e stima dei rischi che eventualmente potrebbero emergere;

3. successivamente all'immissione sul mercato, deve essere operata anche una valutazione sugli eventuali rischi derivanti dall'analisi dei dati raccolti dal sistema di monitoraggio definito nell'articolo 61 del Capo 1 "*Monitoraggio successivo all'immissione sul mercato*";

4. devono essere adottate anche adeguate misure di gestione dei rischi conformemente alle disposizioni presenti negli altri paragrafi dell'articolo n.9 [86].

Alla luce di quanto finora osservato, appare evidente come la politica europea approcci verso l'incentivazione allo sviluppo dell'Intelligenza Artificiale tra gli stati membri, tendendo a spingere forte sul progresso tecnologico ma a fronte

però di un elevato processo etico nella sua massima espressione, tema che verrà approfondito nel capitolo 7.

Capitolo 4

Dattiloscopia 2.0

Lo scopo principale dell'analisi dattiloscopica forense, è quello di attribuire, con "assoluta certezza", un'identità ad un'impronta digitale.

Come indicato in 2.1.1 ci sono un certo numero di istanze in letteratura, in cui gli algoritmi evolutivi vengono utilizzati per confrontare i dettagli di un'impronta digitale con quello di un database di immagini di impronte digitali.

I risultati di tutte queste tecniche, come al solito, dipendono dalla qualità dell'immagine in ingresso, pertanto le tecniche di miglioramento dell'immagine vengono spesso impiegate per ridurre il rumore e migliorare la definizione tra le creste e le valli in modo che non vengano identificate minuzie spurie [11].

In effetti, la corrispondenza delle impronte digitali latenti, rinvenute sulla scena del crimine, è difficile proprio a causa della loro scarsa qualità in quanto vengono depositate dal reo in condizioni inconsce e l'accuratezza dell'analisi nella fase di comparazione, spesso, è migliorata combinando minuzie marcate manualmente con quelle estratte automaticamente.

Sono stati proposti in letteratura molteplici metodi per il miglioramento delle immagini delle impronte digitali, che si basano sulla normalizzazione dell'immagine e sull'utilizzo di svariate tecniche di filtraggio come i filtri di Gabor, il filtro di Fourier direzionale, il metodo di binarizzazione, il miglioramento mediante filtro mediano direzionale, il recupero d'immagini basato sugli istogrammi colore e caratteristiche testuali e molti altri [90].

Diverse altre tecniche di miglioramento presenti in letteratura sono basate su logica fuzzy e reti neurali.

Ryu et al. in uno studio pubblicato nel 2011, proposero un nuovo approccio per migliorare l'estrazione delle caratteristiche per immagini d'impronte digitali di bassa qualità, utilizzando la risonanza stocastica [91].

La risonanza stocastica si riferisce a un fenomeno in cui una quantità appropriata di rumore, aggiunto al segnale originale, può aumentare il rapporto segnale-rumore.

Nella figura 4.1, sono riassunte in un grafo alcune tecniche di detection delle minutiæ che sfruttano in ingresso le due possibili combinazioni di immagine,

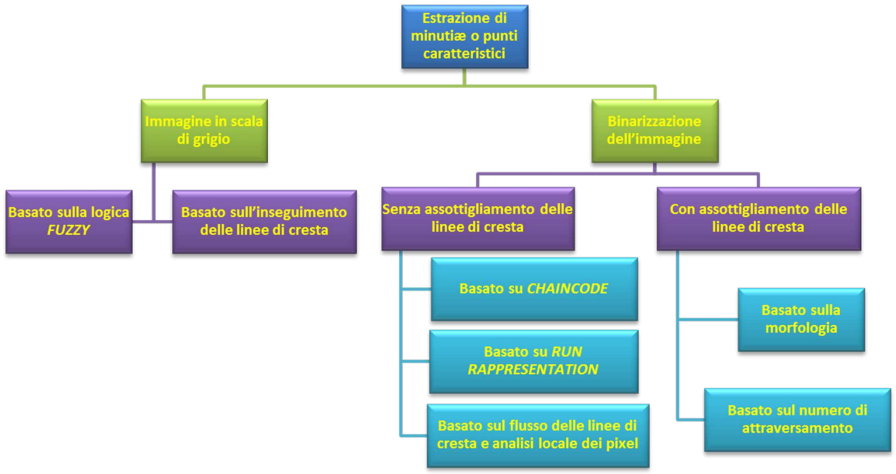


Figura 4.1: Flusso di alcune tecniche di estrazione delle minutiæ da immagini post processate

ovvero direttamente in scale di grigio o attraverso la binarizzazione della stessa [10].

Proprio a causa della scarsa qualità delle immagini, dentificare più efficacemente gli autori di un delitto a partire dai frammenti d'impronte latenti rinvenute sulla scena del crimine, è una missione ardua.

Gli algoritmi di Computer Vision, sono efficaci su impronte acquisite in condizioni ideali, ma fallimentari sui frammenti di impronte latenti spesso affette da un sensibile contenuto di rumore, dove la maggior parte dell'informazione utile, spesso è racchiusa in una piccola regione dell'immagine.

Per questo, poter dare un peso statistico ad ogni accidentalità presente sul dermadoglifo, come già mostrato in 1.2, potrebbe essere una potenziale soluzione mirata a non considerare scartabile la traccia rinvenuta sulla scena del crimine.

In un recente studio, Mengting Gao et al., hanno utilizzato la rete neurale convoluzionale YOLOv5 per progettare un algoritmo di rilevamento automatico in grado di classificare automaticamente alcune minuzie pure e composte delle impronte digitali [92].

L'esito della classificazione nella loro ricerca, simile per certi versi a quella operata dall'Arma dei Carabinieri [3], è stato raggiunto valutando le frequenze statistiche di occorrenza delle minuzie su un dataset di 619.297 immagini di impronte digitali.

I risultati mostrano, per ogni dito, i seguenti intervalli di frequenza:

- terminazioni [68.49% 70.81%]

- biforcazioni [26.37% 27.26%]
- creste indipendenti [1.533% 1.626%]
- speroni [1.129% 1.198%]
- laghi [0,4588% 0,4963%]
- crossover [0,3034% 0,3256%]

Si tratta di solo sei tipi tra minuzie e composizioni di esse, valutate su ogni dito, quando abbiamo visto che in letteratura ne vengono riconosciute ben più di sei, come mostrato in figura 1.5, un approccio sicuramente interessante su cui poter validare eventuali sviluppi della ricerca citata nel capitolo 1.2.

4.1 Stato dell'arte dell'analisi dattiloscopica

4.1.1 In Italia

Relativamente agli accertamenti biometrici, operati dalle forze dell'ordine per identificare una persona, rientrano i rilievi dattiloscopici, i rilievi fisionomici e la più recente profilazione del DNA.

Mentre per gli ultimi due può essere emesso solo un giudizio di compatibilità, il rilievo dattiloscopico rimane l'unico accertamento biometrico che consente di dichiarare con assoluta certezza un'identità personale.

Come più volte detto, l'identificazione dattiloscopica in Italia è ancorata alla sentenza della Suprema Corte di Cassazione n.2559 [9], ma come avviene la valutazione dei 16/17 punti richiesti?

A. Giuliano [11], nel confronto tra due impronte digitali, fa una distinzione tra identità assoluta e identità relativa: la prima che si ottiene tramite una esatta sovrapposizione delle porzioni da confrontare; l'altra che tiene in considerazione le caratteristiche plastico-dinamiche della cute, secondo le quali l'impronta tende a deformarsi elasticamente quando entra in contatto con qualsiasi superficie.

Pertanto, è pacifico che l'identificazione dattiloscopica debba basarsi sulla identità relativa attraverso il raffronto dei caratteri generali e particolari della impronta digitale.

I caratteri generali, sono delle configurazioni geometriche ricorrenti che descrivono l'andamento dei fasci delle linee di cresta sull'impronta digitale e si possono raggruppare in quattro archetipi fondamentali:

- adelta
- monodelta

- bidelta
- composta

I caratteri particolari, invece, sono proprio le minutiæ che descrivono i punti di identità dell'impronta digitale, ed è proprio a questi che la Suprema Corte di Cassazione si riferisce, richiedendo una corrispondenza di 16/17 punti uguali per forma e posizione.

Dal punto di vista probatorio, A. Giuliano [11] sottolinea che non si può ritenere sufficiente la propria certezza circa la validità delle metodologie utilizzate per compiere un dato accertamento, lasciando così aperta una fessura nell'accertamento dattiloscopico.

Tale lacuna oggi è colmata, poiché la Polizia Scientifica Italiana ha ottenuto la Certificazione Bureau Veritas Italia sulle procedure e sulle metodologie di accertamento per indagini e processi, che dichiara la totale affidabilità e competenza della la Polizia Scientifica Italiana e di tutte le procedure e le metodologie di lavoro [93]. In generale, l'Indagine dattiloscopica si fonda su due procedure: una preventiva ed una giudiziaria. Nella prima, viene redatto il cartellino fotosegnaletico, sul quale è riportata acquisizione di tutte le impronte digitali e palmari, le generalità della persona, motivo della segnalazione, connotati fisici, foto (fronte e profilo destro). Tutto ciò è poi archiviato in un sistema denominato APFIS (*Automated Palmar Fingerprints Identification Identification System*), introdotto nel 1998 [94], capace di memorizzare non solo tutti i cartellini segnaletici, ma anche di evidenziare i punti caratteristici di tutte le impronte e, attraverso un sofisticato software, di proporre i candidati da confrontare per giungere all'identità dattiloscopica di uno specifico soggetto. La seconda procedura d'indagine è quella giudiziaria che consiste nell'individuazione di una o più impronte dell'autore di un reato, o frammenti di queste, trovate sul luogo del reato stesso, e nel loro inserimento in APFIS. Eseguita tale fase si procede alla valutazione dei confronti proposti dall'APFIS tra le impronte del reo e le impronte di un soggetto già archiviate nel sistema a seguito di una procedura di fotosegnalamento. Quest'ultima operazione è alquanto complessa e necessita di un'attenta e scrupolosa analisi [95]. La gestione operativa del sistema è affidata alla Sezione AFIS della II Divisione del Servizio Polizia Scientifica [96]. Nel corso degli anni, vista l'esigenza di scambio di informazioni istituzionali esternamente alla Direzione Centrale Anticrimine, sono state prodotte delle applicazioni web per consentire lo scambio di informazioni sia con il Servizio Cooperazione Internazionale presso la Direzione Centrale di Polizia Criminale, sia con il Dipartimento delle Libertà Civili. Una applicazione molto importante tra le varie applicazioni web, è rappresentata dal "Sotto Sistema Anagrafico" (SSA), un sistema che mette a disposizione degli uffici territoriali il patrimonio informativo che l'APFIS rappresenta [94].

4.1.2 Nell'Unione Europea

A livello europeo, il Regolamento numero 603 del 2013, disciplina il sistema per il confronto dei dati relativi alle impronte digitali raccolte negli Stati membri: l'EURODAC [97]. Tale sistema garantisce l'immediato e celere confronto delle impronte digitali dei soggetti richiedenti protezione internazionale e delle persone (cittadini di uno stato terzo o apolidi) che abbiano attraversato irregolarmente le frontiere dell'Unione Europea. Tramite il confronto delle impronte presenti nel sistema, i paesi dell'Unione Europea possono verificare se un richiedente protezione internazionale o un cittadino straniero, che si trova illegalmente sul suo territorio, ha già presentato una domanda in un altro paese dell'UE o se un richiedente protezione internazionale è entrato irregolarmente nel territorio dell'Unione [98]. Tale sistema garantisce l'efficace applicazione del regolamento 604 del 2013 [99]che stabilisce i criteri e i meccanismi di determinazione dello Stato membro competente per l'esame di una domanda di protezione internazionale presentata in uno degli Stati membri da un cittadino di un paese terzo o da un apolide.

L'EURODAC risolve una serie di problemi connessi al ritardo nella trasmissione dei dati da parte di alcuni Stati Membri (che ora è fissata in 72 ore dalla presentazione della domanda di protezione internazionale o dall'attraversamento irregolare di una frontiera dell'Unione Europea) e recepisce l'esigenza che la banca dati possa essere funzionale agli obiettivi di contrasto del terrorismo e della criminalità organizzata [98]. Lo scopo dell'intero sistema EURODAC è quello di implementare l'efficace applicazione della Convenzione di Dublino III per realizzare uno spazio di liberta, sicurezza e giustizia aperto a quanti, spinti dalle circostanze, cercano protezione internazionale nell'Unione [99].

Nel 1995, è stata fondata la Rete Europea degli Istituti Forensi (ENFSI), composta da oltre settanta istituti forensi dei vari paesi europei e riconosciuta dalla Commissione Europea come organizzazione monopolistica nel campo delle scienze forensi.

Il "motore" che muove tutta l'attività più importante dell'ENFSI, è costutuito da 17 gruppi di esperti che operano distintamente nei seguenti campi:

- Animal, Plant and Soil Traces
- Digital Imaging
- DNA
- Documents
- Drugs
- Explosives

- Fingerprint
- Firearms/GSR
- Fire and Explosions Investigation
- Forensic Information Technology
- Forensic Speech and Audio Analysis
- Handwriting
- Marks
- Paint, Glass and Taggants
- Road Accident Analysis
- Scene of Crime
- Textile and Hair

Per quanto riguarda le impronte digitali, l'ENFSI ha previsto un gruppo di esperti dedicato che è l'European Fingerprint Working Group (EFP-WG), il quale attraverso l'organizzazione di incontri, workshop formativi e collaborazione nella ricerca, promuove lo sviluppo e il miglioramento nei campi del rilevamento, dell'imaging e del confronto delle impronte digitali.

Il 27 settembre 2016, l'EFP-WG ha rilasciato un manuale pratico denominato "Best Practice Manual for Fingerprints Examination" che mira a fornire un quadro di procedure, principi di qualità, processi di formazione e approcci all'esame forense per addetti ai lavori [100].

La realizzazione di questo manuale è stata supportata dal Programma di Prevenzione e Lotta contro il Crimine della Commissione Europea [101].

L'obiettivo principale del progetto era che usando i manuali pratici, si sarebbe ottenuto il naturale miglioramento della qualitá dei servizi forensi a disposizione della giustizia e delle forze dell'ordine in tutta Europa, un percorso che sarebbe maturato incoraggiando così la standardizzazione forense e la cooperazione transfrontaliera tra i paesi.

Una delle novità, rispetto all'approccio italiano, è che al punto n.6 del manuale, capitolo in cui viengono trattati i metodi di analisi, confronto, valutazione e verifica, si riconoscono tre diversi approcci per valutare la forza delle prove delle impronte digitali:

1. *approccio numerico*: in sede di confronto tra un frammento e un'impronta digitale, è richiesto un numero di corrispondenze fissato dalla legislazione del paese;

2. *approccio olistico*: il dattiloscopista deve valutare la qualità e la quantità delle caratteristiche del frammento, se è sufficiente allora può essere confrontata con l'immagine di riferimento;

3. *approccio probabilistico*: deve essere riportato anche il valore probatorio del confronto.

É proprio il terzo approccio che ha risvolti interessanti in quanto prevede che la forza probatoria può essere calcolata utilizzando un modello statistico o può essere stimata sulla base delle conoscenze e dell'esperienza dell'esaminatore.

La forza probatoria è riportata come un rapporto di verosimiglianza che può essere calcolato o espresso utilizzando una scala verbale e/o una scala numerica.

La conclusione si basa sulla interpretazione soggettiva dell'esaminatore ma per essere validata richiede la verifica da parte di un altro esaminatore.

Un altro punto importante previsto nel manuale è la validazione e stima della incertezza di misura [100], valutabile se i laboratori utilizzano procedure e metodi convalidati per l'esame delle impronte digitali e dovrebbero considerare:

- l'accuratezza;
- la precisione;
- un intervallo di misura;
- ripetibilità;
- riproducibilità;
- robustezza.

Esistono numerosi metodi di analisi biometrica tra le nazioni europee, pertanto l'ENFSI, attraverso questo manuale pratico, suggerisce di adottare la migliore procedura sicura e solida [100].

4.1.3 Nei paesi extra Unione Europea

Fuori dai confini europei, tra i più antichi laboratori di scienze fisiche, troviamo sicuramente il *National Institute of Standards and Technology* (NIST), fondato nel 1901.

Il NIST è un'agenzia governativa americana nata per sviluppare la competitività industriale degli Stati Uniti in quel momento: un'infrastruttura di misurazione di second'ordine che era rimasta indietro rispetto alle capacità del Regno Unito, della Germania e di altri rivali economici.

Molti settori che operano o si servono di tecnologie di misurazione, si basano in qualche modo sulla tecnologia, sulle misurazioni e sugli standard forniti dal National Institute of Standards and Technology.

Favorite anche dal trascorso periodo pandemico, nella biometria delle impronte digitali, si stanno sviluppando sempre pù velocemente le tecnologie di acquisizione senza contatto, però a causa della modalità di raccolta, le impronte digitali senza contatto solitamente hanno una bassa risoluzione e sono soggette a incoerenze nella postura [102, 103].

Questo problema porta a scarsi risultati sul confronto per la maggior parte degli attuali algoritmi per le impronte digitali relativamente ai metodi tradizionali basati su individuazione di minuzie.

Per risolvere questi problemi, Shy et al. propongono un nuovo metodo di corrispondenza delle impronte digitali senza contatto 2D denominato *fingerprint triplet-GAN* (FTG), in cui sfruttando un metodo end-to-end, utilizzano una rete tripla per eseguire il bilanciamento e l'aumento dei dati e migliorare la robustezza a bassa risoluzione, progettando poi un GAN per aiutare a estrarre caratteristiche indipendenti dalla postura per migliorare la robustezza della rotazione del modello [66].

Alla qualità dell'immagine acquisita, si aggiunge anche l'esigenza di una efficace interoperabilità tra sensori di impronte digitali basati sia su contatto che senza contatto.

In una recente pubblicazione, Ruzicka et al. propongono un nuovo approccio in grado di combinare la correzione della posa con ulteriori operazioni di miglioramento, utilizzando modelli di deep learning per guidare la correzione dell'angolo di visione, migliorando così le caratteristiche di corrispondenza delle impronte digitali senza contatto [104].

Sono sistemi che sfruttano tecnologie all'avanguardia e come tali, dal punto di vista della sicurezza del dato, possiedono intrinsecamente la necessità di garantire la sicurezza dell'identità associata all'impronta digitale.

Uno dei metodi utilizzati dalle associazioni criminali per sfidare la sicurezza biometrica è sicuramente lo spoofing delle immagini, pertanto c'è sicuramente un esigenza urgente di proteggere i sistemi di acquisizione senza contatto contro lo spoofing biometrico.

Rajaram at al. affrontano il problema attraverso uno studio in cui viene proposto l'approccio Contact Less Network (CLNet) per rilevare la falsità nelle impronte digitali senza contatto, con una rete neurale profonda che utilizza immagini di impronte digitali senza contatto seguite da un sistema *transfer learning* chiamato SpoofDetNet che si basa sul modello MobileNetV2 [105].

Nel 2021, il NIST ha rilasciato alcune raccomandazione sulle migliori pratiche per l'acquisizione di impronte digitali senza contatto e lo scambio di dati, definendo le procedure per la valutazione e la certificazione dei dispositivi di acquisizione delle impronte digitali senza contatto [102].

Questo protocollo consente agli sviluppatori di tecnologie contactless per le impronte digitali, di richiedere la certificazione dei propri dispositivi e di

eseguire test automatici del proprio dispositivo utilizzando lo strumento NIST Fingerprint Registration and Comparison Tool (NFRaCT).

Di più recente pubblicazione, il NIST ha rilasciato una un protocollo di certificazione formale per i dispositivi di acquisizione senza contatto con l'obiettivo di garantire sia l'interoperabilità con i dispositivi di raccolta dei contatti legacy sia con altri dispositivi di acquisizione delle impronte digitali senza contatto, stabilendo criteri di misurazione utilizzando come riferimento un dispositivo già certificato e accettabile.

Una guida che si prefigge continui aggiornamenti per stare al passo con i progressi tecnologici garantendo così la continua fedeltà alle norme, ai dispositivi legacy, oltre a includere potenzialmente nuovi progressi tecnologici non appena emergono [106].

4.2 Prospettive future

In un recentissimo studio sulle pubblicazioni scientifiche in ambito forense, C. Weyermann et al. hanno rilevato un notevole aumento del numero delle pubblicazioni legate alle scienze forensi [107].

Questo trend crescente è stato particolarmente marcato negli ultimi 20 anni con oltre 9000 pubblicazioni uscite nel 2019 rispetto alle circa 2000 pubblicate nel 1999 (fonte:Scopus 2022), un segno tangibile dell'interesse crescente della comunità scientifica verso le scienze forensi.

L'ENFSI, in quanto rete di oltre settanta istituti forensi dei paesi europei, coerentemente con il suo obiettivo di garantire che la qualità, lo sviluppo e la diffusione della scienza forense in tutta Europa sia all'avanguardia nel mondo, ha previsto per il futuro il piano "Vision2030".

Il progetto, che mira a migliorare l'affidabilità, la validità delle scienze forensi e promuovere l'implementazione delle tecnologie emergenti [108], si regge su tre pilastri fondamentali:

1. incontro con il futuro: incoraggiare gli studi ad adottare una procedura sicura e solida per l'uso e lo scambio di dati biometrici; promuovere l'uso dell'Intelligenza artificiale nei processi rilevanti; sfruttare nuovi metodi per l'analisi della scena del crimine; mantenere l'aggiornamento sulle tecniche emergenti;

2. rafforzare l'impatto dei risultati forensi: comprendere il comportamento dei materiali durante il trasferimento; sostenere l'armonizzazione per la condivisione dei dati; sensibilizzare e promuovere le capacità multidisciplinari per contrastare la migrazione clandestina, la tratta e il traffico.

3. dimostrare affidabilità nei risultati forensi: sostenere studi sui fondamenti delle scienze forensi; comprendere l'influenza dell'interazione umana nella decisione al processo; promuovere la garanzia della qualità e delle competenze quando si esplorano nuove tecniche e procedure.

Per le forze dell'ordine, è fondamentale avere la lungimiranza nel riconoscere le competenze scientifiche e i trend tecnologici che interesseranno la comunità forense negli anni a venire, per questo, in linea con le proiezioni suggerite dall'ENFSI, nel tempo residuo in convenzione, all'interno del laboratorio AIRTLab dell'UNIVPM si testerà il deep learning direttamente sui frammenti delle impronte digitali latenti, al fine di comprendere se tale tecnologia può essere di supporto al dattiloscopista esperto incaricato di effettuare le comparazioni per associare una paternità al frammento rinvenuto sulla scena del crimine.

Capitolo 5

Fotosegnalamento 2.0

Il riconoscimento biometrico, sta diventando sempre più frequentemente oggetto d'interesse per i sistemi di Intelligenza Artificiale, in particolare quelli che sfruttano tecniche di Deep Learning grazie alla loro migliore accuratezza. In un recente survey, Minaee et al. hanno presentato una panoramica di oltre 150 lavori promettenti sul riconoscimento biometrico (tra cui riconoscimento di volti, impronte digitali, iride, impronte palmari, orecchie, voce, firma e andatura), che implementano modelli di deep learning e mostrano i loro punti di forza in diverse applicazioni [109].

I sistemi di riconoscimento facciale, sinteticamente FR (*Facial Recognition*) possono essere utilizzati per cercare e confrontare volti (estratti da immagini o video) con un database contenete immagini facciali. La precisione dei sistemi FR è oggigiorno elevata per una vasto intervallo di qualità dell'immagine, principalmente grazie all'introduzione dell'intelligenza artificiale o delle reti neurali convoluzionali. Sia nel settore pubblico che in quello privato, questa tecnologia ha molti usi, ad esempio nei confronti ove si cerca l'identificazione di un soggetto sconosciuto attraverso il confronto "uno a molti" (1:N), o uno a uno (1:1) per la verifica dell'identità dichiarata, o N:M per raggruppare gli individui sotto un'unica identità [110].

Scendendo nel particolare dell'attività di ricerca in questo ambito, rientrante nei punti dell'Accordo d'Intesa tra la Polizia di Stato e l'UNIVPM, l'obiettivo è quello di trovare la soluzione a un problema che sinteticamente ho nominato "QFQ" (Quantity, Framing, Quality).

Sono termini che nascondono tre quesiti non banali e interdipendenti:

1. Qual è il *numero minimo* di foto utile a consentire alle reti neurali di operare un efficace riconoscimento facciale di soggetti ripresi dai sistemi di videosorveglianza?

2. Quali sono le *inquadrature migliori* che consentono la più ampia cattura di particolari del volto utile alle stesse reti sia per il riconoscimento automatico sia per la sua ricostruzione 3D?

3. Qual è la *qualità minima* dei dati prodotti tale da garantire sia l'efficacia dei sistemi di riconoscimento automatico sia di ricostruzione 3D, ma anche e/o soprattutto tale da giustificare eventuali investimenti da parte delle forze dell'ordine per ampliare la capacità di storage delle banche dati nazionali?

I sistemi di intelligenza artificiale dedicati al riconoscimento di immagini, utilizzano grandi quantità di dati per potersi addestrare e lavorare al meglio, pertanto riguardo al fotosegnalamento dovevamo cercare dei dati su cui orientare i vari esperimenti.

Ma nei vari repository disponibili in rete, non ci sono database di immagini di volti ripresi da varie direzioni nello spazio, tanto meno registrazioni video che li inquadrano, invece sono disponibili database che riprendono i volti da più angolazioni solo su un piano principale.

Per rispondere ai primi due quesiti, avevamo bisogno di un dataset di immagini del volto inquadrato da più prospettive nello spazio e di videoriprese su ogni individuo paragonabili ad una videosorveglianza virtuale, mentre per rispondere al terzo era necessaria una indagine di mercato incrociata all'analisi dei dati ottenuti.

Il vincolo di partenza è stato quello di rispettare la legacy del fotosegnalamento attuale e quindi costruire intorno ad esso l'upgrade necessario a ottenere i dati adeguati per testare le reti neurali sugli obiettivi prefissati e rispondere così ai tre quesiti.

5.1 Stato dell'arte dei rilievi segnaletici

Come già ricordato nella sezione 4.1.1, l'identificazione personale eseguita dalle Forze di Polizia, può derivare da esigenze di carattere giudiziario o preventivo. L'identificazione tramite fotosegnalamento rientra istituzionalmente tra le attività di competenza della Polizia Scientifica. Dal punto di vista normativo, il fotosegnalamento viene effettuato nei casi disciplinati dall'articolo 4 del Testo Unico di Pubblica Sicurezza [111], l'articolo 349 del codice di procedura penale [112], l'articolo 5 della Legge 30 luglio 2002 numero 189 [113], e dagli articoli 9 [114], 14 [115] e 17 [116]del Regolamento dell'Unione Europea n.603 del 2013 [98].

In sede di indagini preliminari, ai sensi dell'art.349 co2 cpp[1], la PG può procedere al fotosegnalamento della persona nei cui confronti vengono svolte le indagini [112].

[1] art.349 co2° Alla identificazione della persona nei cui confronti vengono svolte le indagini [Dispositivo dell'art. 61 Codice di procedura penale] può procedersi anche eseguendo, ove occorra, rilievi dattiloscopici, fotografici e antropometrici nonché altri accertamenti. [OMISSIS]

Il fotosegnalamento, sin dagli inizi del 900, consiste nell'attività volta a rilevare e riportare sui cartellini segnaletici i dati biografici dell'individuo, uniti ai connotati, i contrassegni, le impronte digitali/palmari e alle foto, frontale e profilo destro del volto.

Nel 1907, il Commissario della Polizia di stato Umberto Ellero ideò un complesso ottico per fotosegnalamento chiamato "*Gemelle Ellero*" in cui l' apparecchio eseguiva una doppia posa simultanea su due lastrine di vetro con formato differente [117][118][119].

La fotografia in doppia posa simultanea, consentiva di congelare visivamente due profili del volto i un'istante di tempo altrimenti irripetibile.

Ancora oggi il fotosegnalamento conserva la metodologia di allora, seppur con apparecchiature moderne, catturando due immagini del volto, ovvero la posa frontale e il profilo destro.

Vista oggi la possibilità di sfruttare le grandi doti dell'Intelligenza Artificiale, c'è la necessità di effettuare studi per vedere come le nuove tecnologie possono contribuire a ridurre il livello di obsolescenza dei metodi attuali.

5.1.1 In Italia

L'art.651 C.P.[2], consente al Pubblico Ufficiale di chiedere a una persona di identificarsi [120] e, ai sensi dell'art. 294 del regolamento per l'esecuzione del TULPS, può chiedere di fornire la carta d'Identità o un titolo equipollente [121]. Tra i titoli equipollenti, come sancito nell'art.35 DPR 445/2000 rientrano anche le tessere rilasciate da un'Amministrazione dello Stato, purché siano munite di fotografia e signatura equivalente al timbro [122]. Dalla disamina di questi articoli di Legge, non si evincono particolari requisiti circa le fotografie da mettere sui documenti utili all'identificazione. Tuttavia, l'Italia è membro dell'International Civil Aviation Organization (ICAO), un'agenzia delle Nazioni Unite che aiuta 193 paesi a cooperare insieme e a condividere il proprio spazio aereo con reciproco vantaggio. L'Italia è inclusa nel gruppo di 11 Stati che rivestono primaria importanza nel settore del trasporto aereo, e ha recepito la regolamentazione ICAO che definisce i requisiti per i documenti d'Identità validi per viaggiare fuori dai confoni nazionali. Per la Carta d'Identità e il Passaporto Elettronico, si possono accettare solo ed esclusivamente fotografie conformi alle norme ICAO. Le caratteristiche tecnico-qualitative che deve possedere la foto per il passaporto elettronico e la carta d'identità sono riassunte qui di seguito [123]:

- La fotografia deve *essere*:

[2] art.651 C.P. Chiunque, richiesto da un pubblico ufficiale nell'esercizio delle sue funzioni, rifiuta di dare indicazioni sulla propria identità personale, sul proprio stato, o su altre qualità personali, è punito con l'arresto fino a un mese o con l'ammenda fino a euro 206.

  - *recente*, non può essere di più di 6 mesi;
  - *larga* tra 35 e 40 mm;
  - *in primo piano*, deve inquadrare viso e spalle, con il viso che copre il 70-80% della foto dalla base del mento alla fronte;
  - *messa a fuoco* e deve essere nitida;
  - *di alta qualità*;
  - *senza macchie* d'inchiostro o pieghe.

- La fotografia deve *avere*:
  - *lo sguardo diretto* della persona verso l'obiettivo della macchina fotografica;
  - *il colorito naturale* della persona ritratta;
  - *luminosità e contrasto ottimale*;
  - *alta risoluzione* e devono essere stampate su carta fotografica di alta qualità.

- Relativamente allo stile ed all'illuminazione devono:
  - avere una colorazione neutra;
  - riprendere con gli occhi aperti;
  - riprendere la persona con gli occhi chiaramente visibili e non coperti dai capelli;
  - riprendere la persona frontalmente, né di lato (stile ritratto) né inclinata, mostrando chiaramente entrambi i lati del viso;
  - essere su sfondo chiaro e a tinta unita;
  - essere riprese con luce uniforme e senza ombre, né riflessi né effetto occhi rossi.

- Per coloro che indossano occhiali da vista, la foto deve:
  - mostrare chiaramente gli occhi senza riflessi sugli occhiali;
  - le lenti non devono essere colorate (se possibile, evitare le montature pesanti e indossare occhiali con montatura più leggera);
  - la montatura non deve coprire nessuna parte degli occhi.

- Per coloro che indossano un copricapo:
  - non sono consentite foto con copricapo se non per motivi religiosi, ma devono essere chiaramente visibili i tratti del viso dalla punta del mento all'intera fronte ed entrambi i lati del viso.

- Per le fotografie ritraenti i bambini, le foto devono:
  - mostrare soltanto la persona ritratta (senza schienale, giocattoli o altre persone visibili) mentre guarda l'obiettivo con un'espressione neutra e la bocca chiusa.

A tal proposito, Franco et al. hanno proposto il framework BioLab-ICAO, un ampio database composto da ground truth, un protocollo di test ben definito e algoritmi di base per agevolare in modo efficace e speditivo la verifica della conformità delle immagini [124].

Riguardo alle fotosegnaletiche, da anni la Polizia Scientifica utilizza una apparecchiatura denominata IDENTISYSTEM, di proprietà della SECOM S.r.l., in grado di scattare simultaneamente la fotografia frontale e del profilo destro del volto, servendosi di una tecnologia basata su sistema ottico. La SECOM S.r.l., ha dotato le nuove apparecchiature per fotosegnalamento di un software proprietario in grado di controllare la qualità della foto conformemente alle norme ICAO [125].

Riguardo al riconoscimento automatico dei volti, da alcuni anni la Direzione Centrale Anticrimine della Polizia di Stato, Servizio Polizia Scientifica, si è dotata di un sistema automatico di ricerca immagini attraverso il quale un operatore di polizia può ricercare l'identità di un volto all'interno di una banca dati.

Questo sistema, denominato SARI, è un software di analisi video in grado di elaborare immagini/video e identificare volti confrontandoli in tempo reale (<1,5 sec) con un'immagine di riferimento. Il software di 'verifica' confronta il volto identificato nella scena con quello di riferimento, generando un ok in caso di verifica positiva. Il software di 'identificazione', invece, confronta il volto individuato nella scena con quelli presenti all'interno di un ampio database e generando un avviso in tempo reale in caso di corrispondenza positiva [126].

Il SARI è stato prodotto in due versioni:

- SARI-Enterprise: combinando diversi algoritmi di riconoscimento facciale, può confrontare un volto presente in un flusso video, con immagini tra 20 milioni di volti.

- SARI-Real Time: in grado di analizzare lo streaming video in tempo reale proveniente da telecamere IP rilevando volti e identificandli rispetto a una lista predefinita di sospetti.

La gestione operativa dei due sistemi di riconoscimento automatico, è affidata rispettivamente alla II Divisione [96] e alla IV Divisione [127] inquadrati nell'organigramma del Servizio Polizia Scientifica [128]. Tuttavia l'esito dell'identificazione finale, è rimesso al parere dell'esperto forense che esprime un

parere di compatibilità sulla base della verosimiglianza analizzata in sede di comparazione fisionomica.

5.1.2 Nell'Unione Europea

In ambito europeo la Commissione ha riconosciuto l'ENFSI come organizzazione momopolistica nel campo della scienza forense, per migliorare lo scambio reciproco di informazioni tra gli stati membri e garantirne la qualità della distribuzione in Europa. E anche riguardo al trattamento di immagini digitali, l'ENFSI ha rilasciato una serie di manuali di buone pratiche (BPM), finalizzati a garantire elevata affidabilità dei risultati, massima qualità delle informazioni contenute, robustezza delle prove e armonizzazione delle metodologie. Detti manuali sono:

1. *Best Practice Manual for Facial Image Comparison*[3]:
   ENFSI-BPM-DI-01 Version 01 - January 2018

2. *Best Practice Manual for Forensic Image and Video Enhancement*[4]:
   ENFSI-BPM-DI-02 Version 01 – June 2018

3. *Best Practice Manual for Digital Image Authentication*[5]:
   ENFSI-BPM-DI-03 Issue 01 – October 2021

4. *Guideline for Facial Recognition System End Users*[6]:
   ENFSI-DI-GDL-001 Version 001 – November 2022

Il primo manuale, affronta l'elaborazione, l'esame e il confronto di immagini raffiguranti volti e la valutazione dei risultati in un contesto forense, definendo l'analisi morfologica come metodo di confronto che si basa sulla valutazione della corrispondenza della forma, dell'aspetto, della presenza e della posizione dei tratti del viso. Il metodo di analisi soggettiva prevede 6 principali fasi procedurali:

- *Analisi*: mirata a valutare se la qualità delle immagini interrogate possiede un grado di dettaglio facciale utile e quali caratteristiche sono disponibili per il confronto;

- *Confronto*: attraverso un approccio morfologico standardizzato, questa fase è volta a individuare sistematicamente le caratteristiche facciali at-

[3] Disponibile su: https://enfsi.eu/wp-content/uploads/2017/06/ENFSI-BPM-DI-01.pdf
[4] Disponibile su: https://enfsi.eu/wp-content/uploads/2017/06/Best-Practice-Manual-for-Forensic-Image-and-Video-Enhancement.pdf
[5] Disponibile su: https://enfsi.eu/wp-content/uploads/2022/12/1.-BPM_Image-Authentication_ENFSI-BPM-DI-03-1.pdf
[6] Disponibile su: https://enfsi.eu/wp-content/uploads/2023/02/DI-GDL-001_GDL-for-Facial-Recognition-System-End-Users_20221111.pdf

traverso l'osservazione scrupolosa del confronto tra le immagini interrogate e di riferimento, per stabilire eventuali somiglianze o differenze tra le caratteristiche osservate;

- *Valutazione*: durante la fase di valutazione le somiglianze e le differenze osservate dal confronto vengono valutate dall'esaminatore, giungendo ad una conclusione che indica il peso probatorio come livello di supporto per una delle proposizioni concorrenti. La FIC (*Facial Image Comparison*) è un processo soggettivo e attualmente non è possibile attribuire una probabilità quantitativa ai risultati dell'esame. Pertanto le conclusioni si baseranno sulla formazione, conoscenza ed esperienza dell'esaminatore.

- *Verifica*: i punti precedenti definiscono il metodo chiamato ACE (*Analysis, Comparison, Evaluation*), che necessita di verifica attraverso la valutazione di un secondo esaminatore che, utilizzando lo stesso metodo, non conosce l'esito del primo esaminatore (verifica cieca) oppure conosce l'esito del primo esaminatore (verifica non cieca);

- *Risoluzione dei disaccordi*: trattandosi di metodi di valutazione comunque soggettivi, devono essere definiti dei protocolli scritti per risolvere i giudizi discordanti in quelle valutazioni che differiscono in modo sostanziale (per esempio con una terza verifica);

- *Revisione tra pari*: l'esito delle comparazioni deve comunque essere sottoposto a revisione per valutare gli aspetti critici e gli aspetti tecnici.

Quando si analizzano le immagini che saranno oggetto d'interrogazione, è necessario esaminare tutto il materiale a disposizione pertinente all'indagine che si va ad operare. Con modalità analoghe a quelle che abbiamo proposto nella sezione 5.6, è quindi possibile selezionare immagini specifiche che forniscono il maggior numero di dettagli facciali per il confronto e con una posa e un angolo di ripresa simili alle immagini di riferimento [129][130]. Al punto n.6 di questo manuale è prevista anche una procedura di validazione e stima dell'incertezza della misura che deve indagare sugli aspetti fondamentali della comparazione [129]. Lo scopo del processo di validazione è stabilire se il metodo soddisfa l'accuratezza, precisione, ripetibilità, riproducibilità e robustezza richieste. Riguardo all'incertezza, ogni laboratorio dovrebbe individuare le potenziali fonti di incertezza tra tutti i fattori che interessano se comparazioni, come l'errore umano, la qualità delle immagini, la formazione strutturata, le esercitazioni ecc. Nell'appendice A del manuale sono infine riportate le linee guida per l'istruzione e la formazione degli esaminatori, pensate per fornire un elenco completo di argomenti di formazione rilevanti per il confronto facciale forense [129].

Il secondo manuale di buone pratiche, rilasciato dall'ENFSI, si concentra sugli aspetti tecnici del miglioramento dei dati di immagini e video digitali [130].

Il metodo suggerito, denominato FIVE (*Forensic Image and Video Enhancement*), nel caso di elaborazioni di immagini singole, prevede una selezione delle immagini, che possono essere singole o fotogrammi provenienti da registrazioni video. Partendo dalla destinazione d'uso e dalla carenza di immagini di input, si dovrebbe selezionare l'immagine più promettente tra quelle disponibili. Poiché è prevedibile che questa operazione di selezione sull'immagine sarà in grado di fornire il miglior risultato desiderato senza effetti collaterali inaccettabili e con perdite minime altrove (nella regione di interesse, ROI) [130]. Un effetto tangibile del miglioramento derivante dalla selezione dedicata delle immagini, lo abbiamo riscontrato dai risultati dei test effettuati sull'FRMDB e riportati nella sezione 5.6.

Il terzo manuale di buone pratiche proposto dall'ENFSI, affronta il processo forense per l'autenticazione dei file di immagini digitali, vale a dire, valuta la misura in cui le domande e le affermazioni fornite, riguardanti la genesi e il ciclo di vita (provenienza) dei dati di immagini digitali, possono essere supportate o risposte [131].

Più specificatamente, il documento si occupa di fornire una metodologia robusta rivolta ad ottenere risultati affidabili riferiti a:

- analisi del contesto;
- analisi della fonte;
- analisi dell'integrità;
- analisi dell'elaborazione;
- rilevamento della manipolazione.

I metodi di analisi discussi includono l'analisi ausiliaria dei dati e l'analisi del contenuto dell'immagine, sia tramite metodi algoritmici che tramite ispezione visiva. L'arco temporale di validazione previsto dal manuale, copre l'intero processo forense, dal sequestro di file di immagini digitali alla presentazione dei risultati in tribunale. Comprende gli aspetti specifici relativi alle risorse, alla gestione degli elementi, alla valutazione iniziale, ai metodi, alla sequenza degli esami, alla ricostruzione, alla convalida, alla valutazione della garanzia della qualità e alla presentazione dei risultati [131].

L'ultimo dei quattro manuali, mira a suggerire le procedure guida utili al riconoscimento facciale, individuando 4 fasi principali:

- Rilevamento del volto;
- Allinamento del viso;
- Estrazione delle caratteristiche;

- Riconoscimento/ricerca del volto.

Il caso d'uso principale per le applicazioni investigative è la ricerca di un soggetto sconosciuto rispetto a un database di riferimento, allo scopo di ottenere l'identificazione del soggetto. Sebbene si parli di "Identificazione", va notato che, a differenza delle ricerche di impronte digitali, il risultato di una ricerca FR non è un'identificazione positiva (che può essere, ad esempio, prodotta come prova) ma un elenco di potenziali candidati che possono essere successivamente oggetto di ulteriori indagini [110], come il sistema SARI in uso alla Polizia Scientifica [126], citato nella sezione 5.1.1. Un'immagine facciale sconosciuta, denominata immagine sonda o immagine di query, viene ricercata in un database, generalmente contenente immagini di riferimento. Il risultato del sistema FR è nella maggior parte dei casi un elenco di candidati, ordinato dal punteggio di somiglianza più alto a quello più basso secondo i criteri dell'algoritmo. I punteggi di somiglianza sono proprietari del sistema FR, generalmente con un punteggio di somiglianza più elevato che indica un maggiore grado di corrispondenza tra le immagini facciali. Talvolta vengono riportati in percentuale ma questo non deve essere considerato come un punteggio di probabilità che due immagini raffigurino lo stesso individuo, è necessaria comunque una revisione operata dall'esperto forense che nell'elenco dei candidati proposti determinerà se è presente il potenziale sospettato [110]. Relativamente alla creazione del database delle immagini di riferimento, l'ENFSI suggerisce di acquisire le immagini in condizioni controllate e che soddisfino gli standard di qualità appropriati per il riconoscimento facciale automatizzato. Una ulteriore guida su come acquisire immagini che soddisfano questi criteri è stata rilasciata dal *Facial Identification Scientific Working Group* (FISWG) [132] che verrà trattata nella sezione che segue.

5.1.3 Nei paesi extra Unione Europea

Fuori dai confini europei, da sempre impegnato nella standardizzazione dei sistemi di misurazione, il NIST ha definito lo standard ANSI/NIST ITL 1-2011 Type 10 che definisce il contenuto, il formato e le unità di misura per lo scambio di impronte digitali, impronte palmari, plantari, facciali/foto segnaletiche, segni di cicatrici e tatuaggi (Scar Mark & Tattoo, SMT), iride, DNA e altri campioni biometrici e informazioni forensi che possono essere utilizzati nel processo di identificazione o verifica di un soggetto [133]. Queste informazioni sono destinate principalmente allo scambio tra la giustizia penale, le amministrazioni o organizzazioni che si affidano e/o effettuano marcature forensi di dati di immagine sui sistemi di identificazione automatizzati o utilizzano altri dati biometrici per l'identificazione. Tra l'altro, il NIST ha fornito un dataset di immagini di fotosegnaletiche sviluppato per essere utilizzato durante la produzione e test

di sistemi automatizzati di identificazione attraverso le foto segnaletiche [134]. Proprio riferendosi a tali sistemi, il NIST nel 2014 ha effettuato un test sui software di 16 fornitori di sistemi di riconoscimento facciale, del tipo "uno a molti" (1:N), rilevando che il miglioramento della qualità dell'immagine è il fattore che contribuisce maggiormente alla precisione del riconoscimento [135]. Il database è costituito da un file compresso, contenente un totale di 3.248 immagini di dimensioni variabili utilizzando il formato PNG [134]. Sempre in America, dopo i quattro attentati suicidi dell'11 settembre 2001, le agenzie federali americane si sono rese conto che mancavano registrazioni di impronte digitali per identificare i terroristi facendo particolare attenzione sull'identificazione dei volti, così nel 2009 l'FBI americana fondò il *Facial Identification Scientific Working Group* (FISWG), con l'obiettivo di sviluppare standard condivisi, linee guida e migliori pratiche per la disciplina dei confronti basati sulle immagini delle caratteristiche facciali umane, fornire raccomandazioni per attività di sviluppo della ricerca e far avanzare lo stato della scienza in modo etico [136]. Il FISWG, nel 2019, ha rilasciato una guida pratica per gli utenti che utilizzano apparecchiature fotografiche progettate per catturare immagini facciali da utilizzare poi con i sistemi di riconoscimento facciale automatizzati o utilizzate per confronti manuali da parte di un esaminatore facciale qualificato [132]. Nella guida sono definite 3 macro aree che contengono le informazioni obbligatorie:

- Descrizione di un ambiente di acquisizione controllato:
  - Illuminazione;
  - Posizione della telecamera;
  - Sfondo;
- Aspetto ottimale del soggetto:
  - Di fronte;
  - Copricapo;
  - Capelli;
  - Occhiali;
  - Espressione;
  - Bocca;
  - Posizione della spalla;
  - Accessori;
  - Trucco e pulizia;
  - Conteggio dei volti (uno solo nella foto);
  - Condizioni mediche;

- Parametri ottimali per l'acquisizione controllata:
  - Tipo di fotocamera;
  - Risoluzione della fotocamera;
  - Classificazione ISO;
  - Messa a fuoco;
  - Funzione di rilevamento del volto nella fotocamera (consigliata);
  - Bilanciamento del bianco regolabile;
  - Apertura del diaframma;
  - Velocità dell'otturatore;
  - Lunghezza focale;
  - Flash;
  - Formato file;
  - Compressione;
  - Spazio del colore (RGB, LAB);
  - Supporto con treppiede;
  - Interfaccia di memoria;
  - Acquisizione remota (Opzionale);
  - Orientamento;
  - Metadati.

Fornendo anche, a titolo di esempio, una raccolta di pose improprie che sono da evitare in un ambiente di acquisizione controllato, cioè in un ambiente dove sono ben definiti tutti i parametri d'acquisizione, come tonalità e colore dello sfondo, orientamento e distanza dell'illuminazione, livello di quota degli occhi, distanza dell'obiettivo dal volto, distanza del capo dal pannello di sfondo e così via[7].

5.2 *MCMPrototype* - Un prototipo di banco per fotosegnalamento

Grazie alla sinergia ormai consolidata tra il laboratorio AIRTLab e il laboratorio di misure Meccaniche e Termiche dell'UNIVPM, abbiamo messo a punto un sistema di acquisizione automatica delle pose del volto da molteplici prospettive, che abbiamo chiamato *MCMPrototype*, figura 5.1.

[7] Disponibile on-line su: https://www.tn.gov/content/dam/tn/tbi/documents/Mug%20Shot%20Best%20Practices%20Poster.pdf

Figura 5.1: MCMPrototype, composto da un braccio motorizzato su cui è installato un sistema d'illuminazione dissipata a led e 4 fotocamere, posizionate ad angolature zenitali di +60°, +30°, 0° e -30°, gestite da Raspberry Pi Zero W.

L'MCMProrotype è composto principalmente da un braccio motorizzato su cui è installato un sistema d'illuminazione dissipata a led e 4 fotocamere, nominate $Z1$, $Z2$, $Z3$ e $Z4$, posizionate rispettivamente ad angolature di +60°, +30°, 0° e -30°, gestite da Raspberry Pi Zero W.

Il setup relativo al fissaggio delle quattro fotocamere sul braccio robotico è stato eseguito inclinandole di quattro angoli, come indicato in figura 5.2(A), ed è stato pensato per cogliere la maggior superficie del volto, escludendo l'asse zenitale in quanto l'immagine prodotta ritrarrebbe quasi totalmente la capigliatura che, eccezion fatta per il suo colore originale, potrebbe variare la forma in tempi anche molto brevi (basti pensare per esempio a una forte folata di vento o una rasatura strategica da parte di un malfattore).

Il braccio compie una rotazione a velocità controllata, spazzando un arco simmetrico rispetto al centro del volto in un intervallo compreso tra -135° e +135°, durante il quale le quattro fotocamere catturano sette pose del volto ognuna in posizioni predefinite del braccio, come indicato in figura 5.2(B).

Il soggetto, durante l'acquisizione, è seduto e fermo al centro dell'asse di rotazione e il suo volto è illuminato da luce diffusa a led installata direttamente sul braccio rotante, in modo da non avere zone d'ombra al passaggio del braccio.

Per ogni soggetto, vengono acquisite le stesse pose, approssimativamente equidistanti dall'obbiettivo delle camere, consistenti in 28 distinte immagini multi prospettive del volto, come mostrato in figura 5.3 dove riquadrate in rosso sono evidenziate le due pose del fotosegnalamento canonico.

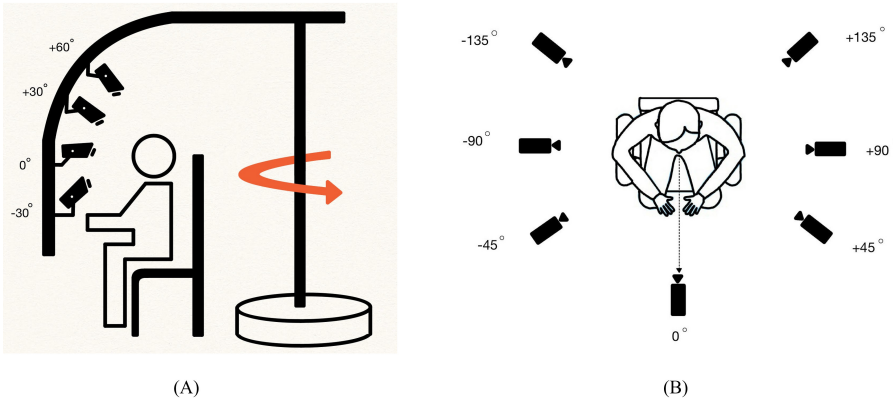


Figura 5.2: (A)Piazzamento Zenitale; (B)Piazzamento Azimutale

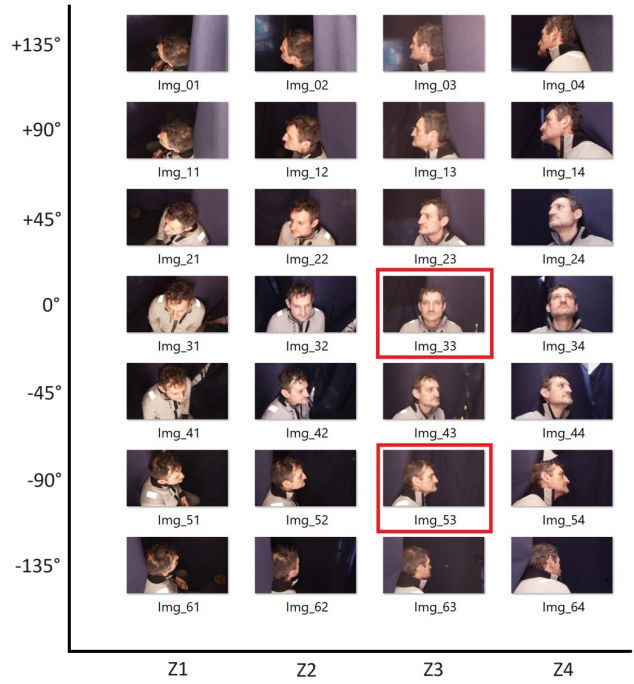


Figura 5.3: Esempio di acquisizione del MCMPrototype; le immagini Img_{33} e Img_{53} riproducono le pose tipiche del fotosegnalamento di Polizia.

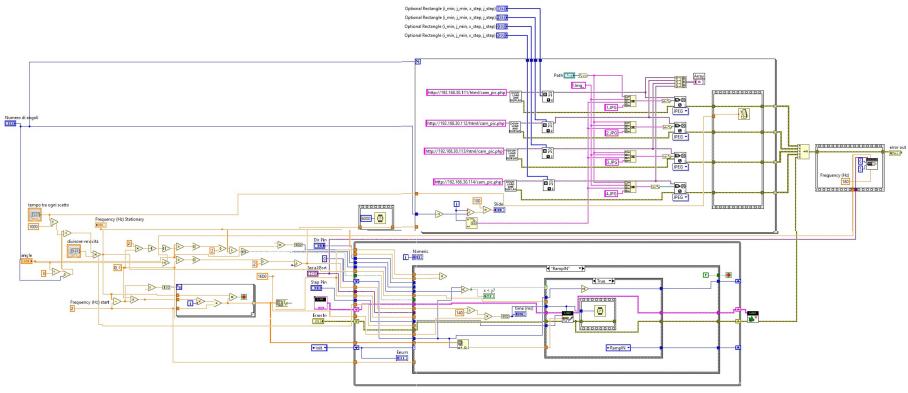


Figura 5.4: Block Biagram LabVIEW, il sistema di controllo è progettato in retroazione per stabilizzare automaticamente i parametri a ogni acquisizione.

Tutti i parametri di mobilità del braccio motorizzato, l'illuminazione e il sincronismo dell'acquisizione delle immagini del volto, sono gestiti tramite software *LabVIEW*.

In figura 5.4 è riportato il *Block Diagram Labview*, al centro del quale è posizionato il clock che entra nel blocco di gestione delle quattro fotocamere, sincronizzandole con il moto del braccio rotante.

L'uscita dai quattro Raspberry Pi è poi retroazionata con il clock stesso, in modo tale da ottenere un controllo autostabilizzante dopo ogni singola acquisizione; in questo modo i parametri vengono sempre aggiornati e migliorati.

La simultaneità degli scatti è assente, la qualità delle foto è molto bassa e non sono stati considerati gli standard ICAO, ma questi non erano requisiti che potevano essere soddisfatti con il prototipo a disposizione, tuttavia per i fini di questo studio il sistema nel suo complesso è risultato assolutamente funzionale ed efficace, almeno in prima istanza.

Vale la pena di rilevare che tutto il materiale meccanico, elettronico e informatico utilizzato, non è provento di finanziamenti dedicati, ma è materiale di recupero adattato ai nostri scopi, di conseguenza nell'insieme può apparire rudimentale ma si è dimostrato molto adatto per la ricerca in atto.

Per poter acquisire database più corposi, e soddisfare anche il requisito della simultaneità, l'ideale sarebbe progettare un sistema di acquisizione multiprospettiva statico, ottico o digitale, che velocizzi e semplifichi la fase del fotosegnalamento, integrato di tutti i requisiti canonici ma anche dotato di buone doti di flessibilità per essere aggiornato alle tecnologie emergenti.

5.2.1 Sistema di simulazione videosorveglianza

Parallelamente al prototipo di acquisizione, in un luogo riservato dell'UNIVPM per non inquadrare zone di passaggio in comune, abbiamo riprodotto un duplice sistema di simulazione di videosorveglianza multiprospettiva, inizialmente a bassa risoluzione con 5 webcams, poi ad alta risoluzione con 3 videocamere professionali.

Il posizionamento delle telecamere, è stato pensato per riprendere i soggetti da direzioni simili a quelle adottate dagli impianti disseminati un po' ovunque sul territorio, simulando riprese di videosorveglianza orizzontali e oblique.

Orizzontalmente vengono effettuate riprese frontali e laterali, a una quota di circa 140 cm dal suolo per simulare le telecamere installate sugli sportelli di prelievo automatico ATM delle banche, tipico bersaglio dei borseggiatori che tentano di usare le tessere bancomat trafugate o attaccare gli utenti rapinandoli all'atto del prelievo di contante.

Le riprese dall'alto e oblique, sono state effettuate da un'altezza tale per cui quando il soggetto ripreso passa per il centro dell'inquadratura, l'asse di ripresa forma un angolo di circa 60° con il pavimento.

La scelta è dettata dal fatto che negli impianti di videosorveglianza distribuiti sul territorio, solitamente, le telecamere sono poste a un'altezza tale da non poter essere manomesse facilmente da un malintenzionato e inclinate in modo tale da non aver un campo di ripresa molto ampio per non perdere i dettagli del target protetto.

Nel nostro caso abbiamo scelto di posizionare le telecamere ad una quota di 250cm dal suolo e con un angolo d'inclinazione rispetto alla zona di passaggio di $45° \leq \beta \leq 75°$.

5.3 *FRMDB* - Un database di fotosegnaletiche e videosorveglianza per l'identificazione automatica

[*pubblicato*][8]

Grazie alla gentile disponibilità di studenti e docenti dell'UNIVPM, abbiamo speso alcune tesi di laurea per questo scopo, producendo così un database di immagini di fotosegnalamento totalmente anonimo chiamato FRMDB, ottenuto acquisendo i volti di 67 soggetti (28 immagini ognuno, a -135°, -90°, -45°, 0°, +45°, +90°, +135° azimutali e +60°, +30°, 0°, -30° zenitali) 39 dei quali corredati di 5 videoclip a bassa risoluzione per ogni persona e i restanti

[8] Contardo, P., Sernani, P., Tomassini, S., Falcionelli, N., Martarelli, M., Castellini, P., Dragoni, A. F. (2023). FRMDB: Face Recognition Using Multiple Points of View. Sensors, 23(4), 1939.

28 corredati di 3 videoclip ad alta risoluzione, previa dichiarazione di consenso formale sottoscritta dalle persone che si sono volontariamente sottoposte alla raccolta[9].

I test iniziali, sono stati effettuati sul set di immagini dei primi 39 soggetti e i video a bassa risoluzione, avendo così già dei riscontri sull'impatto dell'uso di foto segnaletiche prese da più punti di vista nel riconoscimento dei volti sui fotogrammi dei video di sorveglianza.

L'intelligenza artificiale e l'apprendimento profondo (DL) hanno fatto grandi progressi in molti settori applicativi [24].

Il settore delle forze dell'ordine è uno di questi, che sfrutta l'IA e il DL per le indagini sui crimini [25] e per implementare applicazioni sempre più in grado di rilevare autonomamente attività sospette [137].

Ad esempio, in applicazioni come il rilevamento di armi [138], il rilevamento di violenza [139] e il rilevamento di incidenti stradali [140], le applicazioni basate sulla DL sono state sviluppate e sviluppate.

Le tecniche sfruttano la disponibilità dei sistemi di videosorveglianza, fornendo informazioni accurate e ricche per raggiungere la sicurezza [141].

Essendo una delle tecniche biometriche più naturali per l'identificazione [42], il riconoscimento del volto può essere considerato un'applicazione di legge.

Infatti, la naturale variazione tra gli individui porta a una buona separazione tra le classi, rendendo le caratteristiche facciali interessanti per il riconoscimento biometrico [142].

L'interesse della comunità di ricerca sulla Visione artificiale è stato catalizzato per oltre quattro decenni, con le prime metodologie basate sull'analisi delle componenti principali (PCA) e sull'analisi discriminante lineare (LDA) (ad es, Eigenfaces [44] e Fish- erfaces [45]), il riconoscimento dei volti è diventato maturo con i risultati ottenuti dalle Reti Neurali Convoluzionali (CNN) nella verifica dei volti, ovvero il compito di verificare se due immagini di volti appartengono alla stessa persona, e nell'identificazione, ovvero il compito di identificare un volto in un insieme di soggetti noti [55].

[9] **Dichiarazione di consenso informato:** Tutti i soggetti coinvolti nello studio hanno ottenuto il consenso informato e hanno liberamente compilato e sottoscritto una dichiarazione di consenso alla pubblicazione e diffusione di immagini, prendendo atto dell'informativa ricevuta relativamente al D.Lgs. n.196/2003 e successive modifiche, al Regolamento UE2016/679 e l'informativa dell'Università Politecnica delle Marche sulla protezione dei dati relativa all'uso di immagini, foto e riprese audio-video (UNIVPM-Informativa protezione dati relativa all uso di immagini foto).

Tutte le immagini e i video contenuti nell'FRMDB sono anonimi, acquisiti in tempi diversi, ordine casuale e in nessun modo riconducibili ai dati personali contenuti nelle dichiarazioni sottoscritte dai soggetti partecipanti. Tutti moduli sottoscritti delle dichiarazioni sono depositati agli atti presso l'Università Politecnica delle Marche e conservati per gli usi consentiti dalla legge.

Grazie a questi progressi, il riconoscimento facciale è ampiamente utilizzato per l'autenticazione biometrica in applicazioni come lo sblocco degli smartphone [143] e la verifica dei passaporti [144].

Inoltre, il riconoscimento facciale è considerato sufficientemente maturo per integrare i sistemi di identificazione automatica delle impronte digitali (AFIS): i sistemi di identificazione biometrica multimodale che ne derivano [145] sono utili quando sono disponibili solo le immagini di una persona sospettata di reato, invece delle impronte digitali.

Un esempio è il sistema SARI ("Sistema Auto- matico Riconoscimento Immagini") implementato dalla Polizia italiana che, tra le altre funzionalità, permette di verificare l'autenticità delle immagini utilizzate nei documenti di identificazione [146].

L'integrazione del riconoscimento facciale nei sistemi di supporto decisionale esistenti per le indagini criminali, come il SARI, dimostra il suo livello di preparazione.

Tuttavia, mancano ricerche sui sistemi di riconoscimento facciale che possono essere utilizzati per l'identificazione o la verifica attraverso il confronto delle immagini riprese dalle telecamere a circuito chiuso con il database disponibile di foto segnaletiche [147].

Infatti, nonostante i progressi nel Pose-Invariant Face Recognition (PIFR), cioè l'identificazione o la verifica di individui con immagini di volti catturati in pose arbitrarie, la corrispondenza tra due pose arbitrarie è ancora una sfida aperta [148][149].

Inoltre, durante la procedura di fotosegnalamento le forze di polizia nazionali, come già detto, raccolgono abitualmente due immagini, quella frontale e quella del profilo destro (comunemente note come foto segnaletiche), insieme alle impronte digitali e alle informazioni personali di un soggetto.

Tuttavia, mancano ricerche per capire fino a che punto le CNN per il riconoscimento facciale siano efficaci nell'identificare una persona nota in filmati di videosorveglianza quando sono disponibili solo le due immagini standard di fotosegnalazione come immagini di riferimento [12].

A tal fine, il presente lavoro estende il nostro precedente lavoro [150] proponendo il Face Recognition from Mugshot Database (FRMDB), un set di dati di immagini di volti e video per testare l'uso di immagini di foto segnaletiche, scattate da più punti di vista (POV), come immagini di riferimento nel riconoscimento di volti su fotogrammi di videosorveglianza.

Il set di dati proposto può essere utilizzato per misurare l'accuratezza del riconoscimento dei volti con diversi sottoinsiemi di foto segnaletiche.

L'obiettivo è capire se l'utilizzo di immagini del volto da più punti di vista può avere un impatto positivo sulle prestazioni del riconoscimento facciale, giustificando lo sforzo necessario per scattare più immagini e archiviarle.

In particolare, questo lavoro aggiunge i seguenti contributi originali allo stato dell'arte del riconoscimento facciale:

- Propone un nuovo dataset, l'FRMDB, composto da 39 soggetti con foto segnaletiche scattate da 28 prospettive diverse e 5 video di sorveglianza ripresi da 5 prospettive diverse. Il dataset è ad accesso libero e rilasciato gratuitamente in un repository GitHub[10].

- Presenta una rassegna della letteratura sui database esistenti per il riconoscimento dei volti, analizzando il loro potenziale nel benchmarking delle tecniche di verifica e identificazione in scenari di sorveglianza. Nonostante le indagini e le recensioni esistenti sul riconoscimento dei volti includano anche una descrizione dettagliata dei database disponibili, come ad esempio in [151][152], noi analizziamo i set di dati considerando la disponibilità di immagini e clip adatte a testare il riconoscimento in condizioni di videosorveglianza.

- Il progetto confronta i risultati di due CNN consolidate per il riconoscimento dei volti sul dataset proposto e sul database Surveillance Cameras Face (SCFace) [153]. Tale confronto è utile per convalidare l'obiettivo del FRMDB, ovvero testare il riconoscimento dei volti su fotogrammi di telecamere di sicurezza quando sono disponibili diverse foto segnaletiche per l'identificazione.

- Fornisce un benchmark iniziale per il dataset proposto, iniziando ad analizzare le prestazioni del riconoscimento facciale quando sono disponibili diversi sottoinsiemi di foto segnaletiche, prese da vari punti di vista, come riferimento. Il codice sorgente degli esperimenti è pubblicato in un repository GitHub ad accesso libero[11].

Infatti, come osservato nella Sezione 5.3.1, nonostante la disponibilità di molti database per la verifica e l'identificazione dei volti, SCFace è l'unico che include foto segnaletiche e immagini di telecamere di sorveglianza per effettuare il riconoscimento dei volti nei fotogrammi delle telecamere a circuito chiuso utilizzando come immagini di riferimento immagini da 9 punti di vista diversi.

Tuttavia, tutti i volti nei fotogrammi delle telecamere di sorveglianza sono per lo più frontali.

Pertanto, abbiamo costruito un nuovo set di dati, il FRMDB, contenente più foto segnaletiche per ogni soggetto (28) e video di telecamere di sorveglianza ripresi da 5 diversi punti di vista.

[10] Il dataset è disponibile al seguente indirizzo: https://github.com/airtlab/face-recognition-from-mugshots-database

[11] Il codice sorgente degli esperimenti è disponibile al seguente indirizzo: https://github.com/airtlab/tests-on-the-FRMDB

Per quanto riguarda le CNN testate, abbiamo confrontato VGG16 [153] e ResNet50 [54], pre-addestrate sui dataset VGGFace [154] e VGGFace2 [53] per l'estrazione delle caratteristiche del volto.

Inoltre, testando queste CNN sul database SCFace e sul dataset proposto, intendiamo comprendere l'impatto che diversi set di foto segnaletiche possono avere sull'identificazione dei soggetti sospetti registrati nei filmati delle telecamere di sicurezza.

Le foto segnaletiche sono scattate da più punti di vista, oltre alle immagini frontali e di profilo standard raccolte dalle forze di polizia durante la procedura di fotosegnalamento.

Inoltre, i risultati riportati in questo articolo sono completamente riproducibili, dato che sia il set di dati proposto che il codice sorgente dei test sono pubblicati in repository GitHub ad accesso aperto.

La sezione 5.3.1 comprende una rassegna della letteratura sugli insiemi di dati disponibili per il riconoscimento facciale, evidenziando le differenze con quello proposto in questo lavoro, e sulle tecniche di riconoscimento facciale, giustificando la scelta delle CNN per il nostro confronto.

La sezione 5.3.2 descrive il dataset costruito per la nostra ricerca e la metodologia implementata per eseguire i nostri test comparativi.

La Sezione 5.3.6 presenta i risultati dei nostri test, analizzando le prestazioni di accuratezza sul database SCFace e sul nostro dataset utilizzando diversi set di foto segnaletiche come immagini di riferimento. Infine, la Sezione 5.3.8 fornisce le conclusioni di questa ricerca e propone lavori futuri.

5.3.1 Rassegna della letteratura scientifica

Per spiegare la necessità di un nuovo dataset e giustificare la scelta delle CNN utilizzate negli esperimenti, descriviamo le caratteristiche dei database di volti disponibili in letteratura e presentiamo l'evoluzione delle tecniche di riconoscimento dei volti nel corso degli anni.

Sebbene siano disponibili diversi database, la maggior parte di essi non include caratteristiche adeguate per valutare le prestazioni di riconoscimento in filmati provenienti da telecamere di sicurezza, utilizzando come immagini di riferimento set di foto segnaletiche diverse dalle foto frontali e di profilo scattate durante la procedura standard di fotosegnalamento.

Tuttavia, le tecniche basate sulle CNN hanno dimostrato la loro superiorità quando condizioni come l'illuminazione, l'espressione facciale e la posa non sono fisse [55][152].

Per questi motivi, proponiamo un nuovo set di dati e confrontiamo due diverse CNN su di esso.

Database per il riconoscimento facciale

Dato che il riconoscimento dei volti ha suscitato l'interesse dei ricercatori di Computer Vision per oltre quarant'anni, sono disponibili diversi database di immagini di volti per confrontare le diverse tecniche.

Uno dei primi database apparsi per confrontare le diverse metodologie di riconoscimento è l'*AT&T Database of Faces*, precedentemente noto come *Olivetti Research Laboratory* (ORL) *Database of Faces* [155].

Nonostante includa 10 diverse immagini in scala di grigi ($92 \times 112 pixel$) per ciascuno dei 40 soggetti inclusi nel database, variando le espressioni facciali e l'illuminazione, tutte le immagini sono in posizione frontale, senza video di sicurezza o fotogrammi da telecamere di sicurezza con cui confrontarsi.

Il database era gratuito e di libero accesso, anche se, al momento in cui scriviamo, il sito web ufficiale sembra essere stato chiuso.

Con il miglioramento delle tecniche di riconoscimento facciale e l'ottenimento di risultati eccezionali sul database AT&T e su insiemi di dati simili, la ricerca si è concentrata su scenari non vincolati, ossia con condizioni variabili relative a illuminazione ambientale, risoluzione dell'immagine, ingombro dello sfondo, posa del volto, espressione e occlusione [156].

Di conseguenza, sono comparsi database di immagini di volti dedicati al riconoscimento di volti senza vincoli, come il *Labeled Face in the Wild* (LFW) [157][158] e il *YouTube Faces Database* [159].

Il database LFW comprende 13.233 immagini a colori ($250 \times 250 pixel$) di 5.749 persone uniche, con 1.680 soggetti che hanno due o più immagini.

Le immagini dei volti sono state raccolte da varie fonti sul web, utilizzando il rilevatore di volti *Viola-Jones* [160].

Il database LFW è gratuito e ad accesso libero, tuttavia è stato concepito per la verifica dei volti senza vincoli e quindi non include set di foto segnaletiche e video da confrontare in modo sistematico.

Pertanto, non è adeguato per valutare le prestazioni delle tecniche di riconoscimento dei volti testando le immagini da più punti di vista.

Allo stesso modo, il database dei volti di YouTube comprende 3.425 video a colori di YouTube di 1.595 persone diverse.

Pertanto, anche questo database è destinato alla verifica dei volti senza vincoli, senza set di foto segnaletiche scattate sistematicamente da utilizzare in scenari di videosorveglianza.

Come la LFW, anche il database dei volti di YouTube è di libero utilizzo e ad accesso aperto.

Con i risultati ottenuti dalle CNN nel riconoscimento delle immagini e dei volti, sono comparsi database con un numero maggiore di immagini di volti e identità uniche, al punto da rendere possibile l'addestramento e la valutazione delle CNN su scala milionaria.

A tal fine, il database *CASIA-Webface* [161] comprende 494.414 immagini di volti di 10.575 identità uniche, le cui immagini sono state raccolte dal web a varie risoluzioni.

Il database è disponibile su richiesta, anche se il sito web ufficiale sembra non essere più attivo al momento in cui scriviamo.

Il *Megaface Challange Dataset* [162][163] comprende 4,7 milioni di foto a colori di 672.057 soggetti unici, a varie risoluzioni.

Poiché la Megaface Challenge è terminata, il database è stato interrotto e i dati Megaface non sono più distribuiti ufficialmente.

Il dataset *VGGFace* [154] contiene 982.803 immagini a colori (95% frontali, 5% di profilo) di 2.622 identità uniche, mentre il dataset *VGGFace2* [53] comprende 3,31 milioni di immagini a colori di 9.131 soggetti unici.

Entrambi i dataset VGGFace e VGGFace2 sono gratuiti e ad accesso libero.

I dataset Megaface Challenge, VGGFace e VGGFace2 includono volti raccolti dal Web in diverse condizioni di illuminazione, posa, espressione e occlusione, analogamente a LFW e YouTube Face Dataset.

La quantità di immagini disponibili in questi database li rende ideali per l'addestramento di tecniche basate sulla DL come le CNN, anche per l'utilizzo in modalità transfer learning, come abbiamo fatto con i dataset VGGFace e VGGFace2 in questo lavoro.

Tuttavia, poiché questi database non includono set di foto segnaletiche e video di sicurezza da confrontare sistematicamente, non sono adatti a valutare l'impatto dell'uso di foto segnaletiche da più punti di vista sulle prestazioni di riconoscimento dei volti in scenari di sorveglianza.

Nel corso degli anni sono stati pubblicati anche alcuni database che includono soggetti con pose diverse, cioè foto segnaletiche da più prospettive.

Ad esempio, il database *Facial Recognition Technology* (FERET) [164][165] comprende 14.051 immagini a colori ($512 \times 768 pixel$) di 1.199 soggetti.

Per 200 soggetti tra quelli che compongono il database, sono disponibili 9 foto segnaletiche scattate sistematicamente da diversi punti di vista (da -60° a +60°).

Il set di dati è disponibile su richiesta con un accordo di rilascio dedicato.

Il *Max Planck Institute for Biological Cybernetics Face Database* [166] comprende immagini a colori ($256 \times 256 pixel$) scattate da 7 diversi punti di vista (da -90° a +90°, con un passo di 30°) circa 200 identità uniche, per un totale di 1.400 immagini, tuttavia, il database non è più disponibile.

L'*Extended Yale Face Database B* [167][168] comprende 16.128 immagini in scala di grigi ($640 \times 480 pixel$) di 28 identità uniche ottenute combinando 9 diverse pose (un volto frontale, 5 immagini a 12° e 3 immagini a 24°) con 64 diverse condizioni di illuminazione.

Il database è gratuito e ad accesso libero.

Anche il *Korean Face Data Base* (KFDB) [169] comprende immagini di volti da diversi punti di vista.

In particolare, dispone di 52.000 immagini a colori ($640 \times 480 pixel$) di 1.000 soggetti unici, con diverse condizioni di illuminazione, espressioni facciali e sistematicamente da 7 angolazioni diverse (da -45° a +45°, con un passo di 15°).

Al momento in cui scriviamo, il database non è disponibile.

Il database *CAS-PEAL* [170] contiene 30.900 immagini di volti a colori ($360 \times 480 pixel$) di 1.040 identità uniche.

Sono disponibili immagini facciali da 21 punti di vista diversi, che combinano 7 angolazioni diverse sul piano orizzontale (da -67,5° a +67,5°, con un passo di 22,5°) e 3 angolazioni diverse sul piano verticale (da -30° a +30°, con un passo di 30°).

Per alcuni sottoinsiemi di soggetti, sono disponibili altre immagini con diverse caratteristiche facciali come espressioni, luci e accessori vari.

Il database è disponibile su richiesta.

Il set di dati *Multi-PIE* [171] contiene 755.370 immagini a colori ($3072 \times 2048 pixel$) di 337 soggetti unici registrate in diverse sessioni per includere variazioni di posa, illuminazione ed espressione.

Per ogni sessione, 13 immagini che vanno da -90° a +90° con un passo di 15° sul piano orizzontale sono state scattate da diverse telecamere poste all'altezza della testa.

Sono state scattate altre due immagini a 30° sul piano orizzontale e sopra l'altezza della testa.

Il set di dati è disponibile per la distribuzione su richiesta.

Il *NIST Mugshot Identification Database* (MID) [134] comprende 3.228 immagini in scala di grigi (di dimensioni variabili) di 1.573 individui, 1.333 soggetti hanno sia la foto segnaletica frontale che quella di profilo, 131 soggetti hanno due o più foto frontali e 89 soggetti hanno due o più foto di profilo.

Il database è disponibile su richiesta.

Nonostante i database FERET, Yale, MPI, KFDB, CAS-PEAL, Multi-PIE e MID includano foto segnaletiche da più punti di vista, non contengono fotogrammi o video dei soggetti provenienti da telecamere di sicurezza che consentano di analizzare l'impatto dell'utilizzo di sottoinsiemi di immagini da diverse angolazioni sulle prestazioni del riconoscimento facciale in scenari di sorveglianza.

Il *ChokePoint Dataset* [172] si differenzia dai dataset sopra citati, infatti con due insiemi di soggetti di 25 e 29 (gli insiemi si sovrappongono) e 48 sequenze video, il dataset intende riprodurre le condizioni di videosorveglianza per la verifica da video a video.

Tuttavia, il dataset non include foto segnaletiche da utilizzare per la identificazione o la verifica dei soggetti nei video.

Il dataset è ad accesso libero.

Per quanto ne sappiamo, l'unico database che include foto segnaletiche scattate sistematicamente da più punti di vista e immagini di volti provenienti da telecamere di sicurezza è il database *Surveillance Cameras Face* (SCFace) [153].

Il database contiene infatti 4.160 immagini di 130 soggetti unici, ogni soggetto ha 9 foto segnaletiche a colori ($2048 \times 3072 pixel$) scattate da -90° a +90° con un passo di 22,5°, un'altra foto frontale a colori ($2048 \times 3072 pixel$), una foto segnaletica frontale a infrarossi (IR) ($320 \times 426 pixel$) e 21 immagini (15 a colori e 6 IR) di dimensioni variabili scattate con 7 telecamere di sicurezza a 3 diverse distanze.

Il database è disponibile su richiesta, con un accordo di rilascio dedicato.

Date le sue caratteristiche, il database SCFace era l'unico disponibile per testare la capacità delle CNN di eseguire il riconoscimento di volti su immagini di sorveglianza utilizzando diversi sottoinsiemi di foto segnaletiche, come abbiamo fatto nella nostra precedente ricerca [150].

Tuttavia, tutte le immagini delle telecamere di sicurezza sono quasi frontali, mentre nella vita reale un soggetto può essere inquadrato da diverse prospettive.

Mentre alcuni dei set di dati esistenti, come VGGFace e VGGFace2, permettono di addestrare.

La maggior parte dei database di volti osservati fin'ora, non è adeguata per valutare le capacità di riconoscimento su filmati di videosorveglianza quando sono disponibili diversi set di foto segnaletiche, da più punti di vista, come immagini di riferimento.

Per ovviare a questa limitazione, proponiamo un nuovo set di dati che comprende 39 soggetti, con 28 foto segnaletiche e 5 video ripresi da telecamere di sicurezza posizionate in cinque punti diversi, in cui le immagini vengono scattate combinando 7 angoli sul piano orizzontale (da -135° a +135°con un passo di 45 ∘) e 4 angoli sul piano verticale (da +60° a -30° con un passo di 30°).

Pertanto, come mostrato nella Tabella 5.1 che riassume le caratteristiche dei database discussi in questa sottosezione, il dataset proposto consente l'utilizzo di foto segnaletiche da più punti di vista sia sul piano orizzontale che verticale per l'identificazione di soggetti in filmati provenienti da telecamere di sicurezza.

Evoluzione delle tecniche di riconoscimento facciale

Le prime tecniche per il riconoscimento automatico dei volti nelle immagini digitali si basavano sull'analisi delle componenti principali (PCA) e sull'analisi lineare discriminante (LDA in particolare, Turk e Pentland [44] hanno proposto di calcolare gli autogeni delle facce, cioè di estrarre un vettore di caratte-

Tabella 5.1: Riassunto delle caratteristiche dei database di volti discussi nella sottosezione *Database per il riconoscimento facciale* confrontati con il dataset proposto. Quello proposto in questo lavoro è l'unico dataset che include foto segnaletiche da più punti di vista sia sul piano orizzontale che su quello verticale, insieme a video di telecamere di sicurezza ripresi da più punti di vista.

Database	**# Subjects**	**# Face Images**	**Posed/In the wild**	**Multiple POVs (°)**	**Images/Videos From Security Cams**	**Availability**
AT&T [155]	40	400 (grayscale)	Posed	none	none	Not available
LFW [157, 158]	5,749	13,233 (color)	In the wild	none	none	Open-access
YouTube Faces [159]	1,688	3,425 (color videos)	In the wild	none	none	Open-access
CASIA-Webface [161]	10,575	494,414 (color)	In the wild	none	none	Upon request
Megaface [162, 163]	672,057	4.7 million (color)	In the wild	none	none	Not available
VGGFace [154]	2,622	982,803 (color)	In the wild	none	none	Open-access
VGGFace2 [53]	9,131	3.31 million (color)	In the wild	none	none	Open-access
FERET [164, 165]	1,199	14,051 (color)	Posed	Horizontal plane: -60°, -40°, -25°, -15°, 0°, 15°, 25°, 40°, 60° Vertical plane: none	none	Upon request
MPI [166]	200	1,400 (color)	Posed	Horizontal plane: from -90° to +90°, 30° step Vertical plane: none	none	Not available
Extended Yale [167, 168]	28	16,128 (grayscale)	Posed	Horizontal plane: 0°, 12°, 24° Vertical plane: none	none	Open-access
KFDB [169]	1,000	52,000 (color)	Posed	Horizontal plane: from -45° to +45°, 15° step Vertical plane: none	none	Not available
CAS-PEAL [170]	1,040	30,900 (color)	Posed	Horizontal plane: from -67.5° to +67.5°, 22.5° step Vertical plane: -30° to +30°, 30° step	none	Upon request
Multi-PIE [171]	337	755,370 (color)	Posed	Horizontal plane: from -90° to +90°, 15° step Vertical plane: 2 pictures on a different unknown angle	none	Upon request
NIST MID [134]	1,573	3,288 (color)	Posed	Horizontal plane: -90°, 0°, +90° Vertical plane: none	none	Upon request
ChokePoint [172]	25-29	48 (color videos)	Security Cams	Horizontal plane: 3 unknown angles Vertical plane: none	48 Videos from 3 POVs in total	Open-access
SCFace [153]	130	4,160 (color and IR)	Posed + Security Cams	Horizontal plane: from -90° to +90°, 22.5° step Vertical plane: none	23 Frontal Face Images per subject	Upon request
FRMDB (proposed)	**39**	**1,092 (color) 195 (color videos)**	**Posed + Security Cams**	**Horizontal plane: from -135° to +135°, 45° step Vertical plane: -60° to +30°, 30° step**	**5 Videos from multiple POVs per subject**	**Open-access**

ristiche che massimizzano la varianza interclasse in un insieme di immagini di addestramento).

Proiettando l'immagine di un volto nello spazio ottenuto con la PCA, la identificazione del volto può essere eseguita con un metodo di *nearest neighbor*, calcolando la distanza dalle immagini di addestramento.

Belhumer et al. [45] hanno invece proposto di aggiungere alla PCA l'Analisi Discriminante Lineare (LDA), al fine di minimizzare la varianza intra-classe, chiamando questa tecnica Fisherfaces.

A differenza di Eigengaces e Fisherfaces, Ahonen et al. [46] calcolano gli istogrammi dei modelli binari locali (LBPH) sulle immagini dei volti per descrivere le regioni dei volti con i modelli binari locali (LBP).

In questo modo, una funzione di distanza basata sugli LBPH può essere utilizzata per eseguire l'identificazione del volto.

Nonostante i risultati promettenti ottenuti in database come l'AT&T, dove alcune variabili tra posa, espressione e illuminazione sono fisse, queste tecniche sono insufficienti per estrarre caratteristiche invarianti ai cambiamenti del mondo reale [165], come ad esempio nei filmati di videosorveglianza.

Dopo gli straordinari risultati nell'elaborazione delle immagini ottenuti da AlexNet nella competizione ImageNet 2012 [173], le CNN hanno anche mostrato risultati robusti nel riconoscimento dei volti con le mutevoli condizioni di illuminazione, espressione e posa tipiche delle immagini di volti non affaticati [55].

Ad esempio, DeepFace [50], una CNN a 8 strati per l'elaborazione di immagini di volti a 3 canali 152×152, ha ottenuto un'accuratezza del 97,35% sul dataset LFW [51].

Parkhi et al. [154] hanno addestrato VGG16 [174], una CNN a 16 strati, sul dataset VGGFace e l'hanno testata sul dataset LFW, ottenendo un'accuratezza del 98,95%.

Analogamente, FaceNet [52], una CNN a 22 strati addestrata in diversi esperimenti con un numero variabile di immagini di volti, tra 100 e 200 milioni, appartenenti a 8 milioni di soggetti diversi, ha ottenuto un'accuratezza del 99,63% su LFW, utilizzando immagini di input 220×220.

Cao et al. [53] hanno testato ResNet50 [54], una CNN a 50 strati basata sull'apprendimento residuale, e SE-ResNet50 (cioè ResNet50 con i blocchi Squeeze ed Excitation [175]) sul dataset VGGFace2, ottenendo un errore di identificazione top-1 del 3,9% con ResNet50 sul dataset VG- GFace2.

You et al. [176] hanno confrontato diverse CNN su LFW (e su altri dataset) applicando il transfer learning: le CNN sono state preaddestrate sul database CASIA-Webface.

I modelli migliori sono stati VGG16 [174] e ResNet50, che hanno ottenuto un'accuratezza del 98,94% e del 98,52% sul database LFW.

Inoltre, nonostante il Pose-Invariant Face Recognition (PIFR), cioè la identificazione o la verifica di individui con immagini di volti catturati in pose arbitrarie, sia ancora una sfida aperta, le tecniche recenti hanno mostrato progressi incoraggianti [148][149].

La maggior parte di queste tecniche si basa sulla generazione di immagini sintetiche (o parzialmente sintetiche), per frontalizzare il volto o creare immagini in qualsiasi posa.

Ad esempio, Hassner et al. [177] hanno proposto di allineare i punti delle caratteristiche facciali, nonostante la posa del soggetto, a una superficie facciale 3D unica per tutti i volti.

Retroproiettando il colore dell'immagine del volto sulla superficie 3D e prendendo in prestito le apparenze dai lati simmetrici corrispondenti del volto, producono l'immagine frontale finale.

Hanno testato la loro metodologia sul set di dati LFW, ottenendo un'accuratezza del 91,62%.

Tran et al. [178][179] hanno presentato un'estensione delle reti avversarie generative (GAN) per generare un numero arbitrario di volti sintetici in qualsiasi posa.

Hanno ottenuto un'accuratezza di identificazione del 90,8% sul set di dati Multi-PIE.

Zhao et al. [180] hanno introdotto un Pose Invariant Model (PIM) basato sull'uso di una GAN per la frontalizzazione del volto e di una CNN per l'apprendimento delle caratteristiche del volto.

Testando l'accuratezza del riconoscimento sulle immagini del dataset Multi-PIE a 15°, 30°, 45°, 60°, 75°, e 90°, hanno ottenuto una media del 96%.

Queste metodologie sono state applicate su set di dati per il riconoscimento dei volti senza vincoli, senza considerare il confronto tra le immagini delle foto segnaletiche e i fotogrammi delle telecamere di sorveglianza.

Nel dataset proposto, invece, abbiamo combinato questi due aspetti: ciò può essere utile per confrontare le CNN tradizionali (senza frontalizzazione e immagini sintetiche) come nel nostro approccio, o le tecniche PIFR come quelle elencate.

Date le capacità dimostrate dalle CNN nel riconoscimento dei volti, in questo lavoro confrontiamo VGG16 e ResNet50 per compiere un primo passo nella valutazione del riconoscimento.

5.3.2 Materiali e metodi

Data la necessità di disporre di database di volti per valutare la capacità di riconoscere i volti in fotogrammi di sicurezza da foto segnaletiche prese da più

punti di vista, proponiamo un nuovo dataset che comprende 28 diverse foto segnaletiche più 5 video da telecamere di sicurezza per ogni soggetto.

Tale set di dati sarà utile per valutare se l'utilizzo di più pose, oltre alla foto frontale e al profilo destro solitamente disponibili nei database delle forze dell'ordine, possa avere un impatto positivo sulle prestazioni di riconoscimento dei volti.

Per stabilire un benchmark iniziale per il dataset proposto, abbiamo testato due diverse CNN, VGG16 e ResNet50, pre-addestrate per il riconoscimento dei volti su database di grandi dimensioni, ovvero VGGFace e VGGFace2.

A tal fine, nelle sottosezioni che seguono, si descrivono il dataset di immagini di volti proposto, vengono forniti i dettagli delle CNN utilizzate per i nostri test e una spiegazione del protocollo sperimentale e le metriche calcolate nei nostri test.

5.3.3 Il database proposto

Il Face Recognition from Mugshots Database (FRMDB) comprende 39 identità uniche, 17 femmine e 22 maschi.

L'età media dei soggetti è di 24,6 anni, con l'individuo più giovane di 19 anni e il più anziano di 52 anni (deviazione standard σ7,8%).

Per ogni soggetto, il set di dati comprende:

- 28 foto segnaletiche, cioè 28 immagini a colori scattate da diversi punti di vista con il soggetto in posa durante l'acquisizione.
- 5 video di telecamere di sicurezza, ripresi da 5 punti di vista. Inoltre, è disponibile un video a mosaico che include tutti e 5 i filmati contemporaneamente.

La Figura 5.5 include le 28 foto segnaletiche del soggetto "031" presenti nel database (ogni identità è un codice a 3 cifre, per preservare l'anonimato).

Ogni foto segnaletica è un'immagine JPEG di $972 \times 544 pixel$.

Abbiamo raccolto le foto segnaletiche scattando immagini da 7 angolazioni sul piano orizzontale e da 4 angolazioni sul piano verticale.

In particolare, sul piano orizzontale le immagini sono state scattate da -135° a +135°, con un passo di 45° (con 0° di fronte al soggetto).

Sul piano verticale, le immagini sono state scattate da -30° a +60° (con lo 0° della fotocamera sul piano degli occhi del soggetto) con un passo di 30°.

A questo proposito, la Figura 5.2 mostra i diversi punti di vista sui piani orizzontale e verticale utilizzati per scattare le foto segnaletiche.

Per gli esperimenti presentati in questo articolo, abbiamo ritagliato manualmente il volto in ogni foto segnaletica per ogni soggetto.

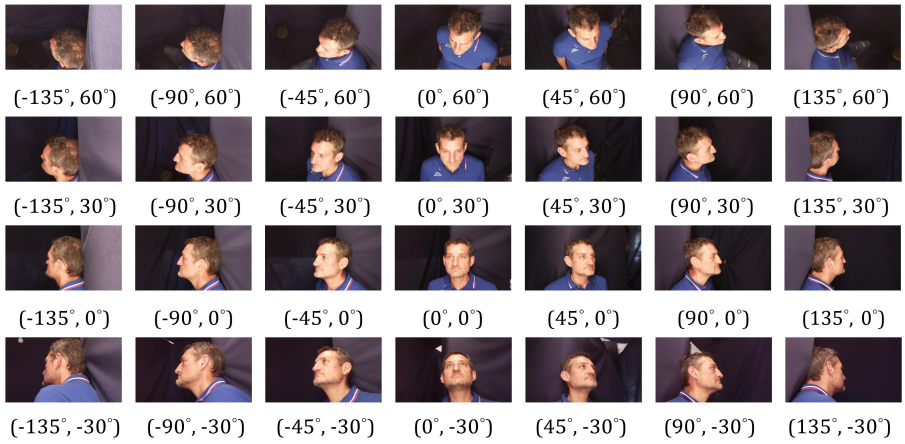


Figura 5.5: Un campione delle foto segnaletiche disponibili per ogni soggetto nell'FRMDB. Per ogni foto segnaletica, gli angoli da cui è stata scattata la foto sono riportati come coppia (h, v): h è l'angolo sul piano orizzontale da -135° a +135°, con un incremento di 45° tra un angolo e il suo adiacente (da sinistra a destra); v è l'angolo sul piano verticale da 60° a -30°, con un passo di -30° tra un angolo e il suo adiacente (dall'alto in basso).

Pertanto, abbiamo pubblicato le foto segnaletiche ritagliate nell'archivio dei dati.

Per scattare le foto segnaletiche, abbiamo chiesto al soggetto di sedersi in una camera ricoperta di teli per produrre uno sfondo scuro e le immagini sono state scattate da 4 fotocamere posizionate in 4 punti su un braccio robotico che ruotava intorno all'asse verticale, come mostrato nella Figura 5.1.

Tale rotazione ha permesso di acquisire le foto segnaletiche da 7 angolazioni sul piano orizzontale e da 4 angolazioni sul piano verticale, una scelta arbitraria ma pensata per catturare il maggior numero di dettagli del volto, trascurando le parti maggiormente coperte dal cuoio capelluto in quanto potenzialmente variabile in tempi brevi (un malfattore, per camuffarsi dopo la commissione di un delitto, potrebbe per esempio tingere i capelli, fare una pettinatura diversa o una rasatura improvvisa).

In questo modo, le immagini sono state scattate quattro a quattro: tutte le immagini con lo stesso angolo orizzontale, cioè nella stessa posizione di rotazione (nella stessa colonna della Figura 5.5), sono state scattate contemporaneamente.

L'illuminazione è stata fornita da una striscia di led posta sul braccio rotante, con luce diffusa per non sovraesporre le fotocamere e per non creare zone d'ombra sul viso.

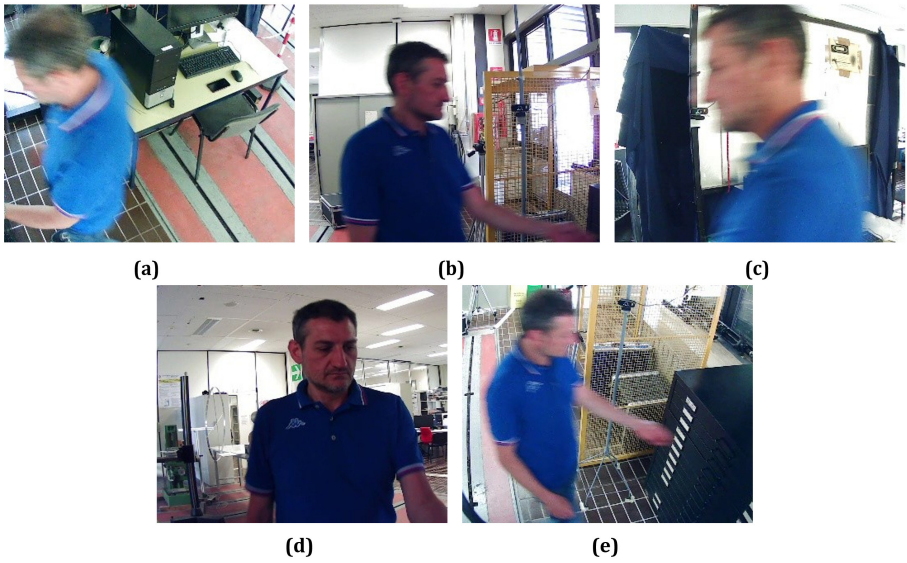


Figura 5.6: Fotogrammi dei video delle telecamere di sicurezza del database proposto. I video sono stati registrati contemporaneamente da 5 punti di vista diversi. Durante la registrazione dei video, ai soggetti è stato chiesto di camminare fino a una cassettiera, aprire un cassetto, estrarre un foglio, firmare il foglio e tornare al punto di partenza.

La Figura 5.6 include un fotogramma per ciascuno dei 5 video di telecamere di sicurezza appartenenti al soggetto "031" presenti nel database.

I video sono codificati con il codec H.264 (il formato del contenitore è Matroska - mkv) e registrati a 60 fotogrammi al secondo.

La dimensione dei fotogrammi è di $352 \times 288 pixel$ (la dimensione del mosaico che include tutte e 5 le clip è di $1280 \times 720 pixel$).

La durata media dei video è di 18,5 s (minimo 15 s, massimo 29 s, deviazione standard $\sigma = 2,9s$).

Per registrare i video delle telecamere di sicurezza, a ogni soggetto è stato chiesto di camminare fino a una cassettiera, aprire un cassetto, estrarre un foglio, firmare il foglio e tornare al punto di partenza.

Durante l'esecuzione di questi compiti, 5 telecamere posizionate in 5 punti diversi hanno ripreso il soggetto.

I 5 video di ogni soggetto sono stati registrati contemporaneamente.

Per gli esperimenti presentati in questo articolo, abbiamo selezionato manualmente un fotogramma per ogni video e ritagliato il volto, per testare le prestazioni di riconoscimento su tali fotogrammi utilizzando diversi set di foto segnaletiche.

I fotogrammi selezionati e i volti ritagliati sono disponibili nel repository del dataset proposto.

Oltre alle foto segnaletiche descritte e ai video delle telecamere di sicurezza, l'FRMDB include:

- Una foto frontale aggiuntiva (1920x1080 pixel, JPEG) per ogni soggetto, scattata con una luce diversa da una fotocamera posta di fronte al soggetto.

- Per 12 dei 39 soggetti è disponibile un secondo set di 5 video delle telecamere di sicurezza (più il mosaico). Per questi soggetti, il secondo set di video di sicurezza varia perché il soggetto indossa diversi accessori sulla testa, come occhiali, occhiali da sole, cappelli e bandane. I soggetti non indossano tali accessori nelle foto segnaletiche.

- Per 3 dei 39 soggetti, una seconda serie di 28 foto segnaletiche scattate con il soggetto sorridente.

- Un file di testo per ogni soggetto contenente il sesso, l'età e gli accessori indossati nella seconda serie di video di sicurezza, se disponibili.

Questi file potrebbero essere utili per ulteriori test di riconoscimento in condizioni diverse, tuttavia, non abbiamo utilizzato tali file negli esperimenti presentati in questo lavoro.

5.3.4 Le CNN confrontate

Utilizzando il dataset proposto e il database SCFace, abbiamo testato le capacità di riconoscimento di due diverse CNN, ovvero VGG16 e ResNet50, quando diversi sottoinsiemi di fotoso segnaletiche sono disponibili.

Le foto segnaletiche sono utilizzate come immagini di riferimento.

In particolare, le CNN estraggono per ogni volto un embedding, cioè un vettore di caratteristiche che descrivono l'immagine del volto.

Gli embedding delle foto segnaletiche e quelli dei volti nelle telecamere di sicurezza possono essere confrontati mediante una misura di distanza o di somiglianza, come la distanza euclidea o la somiglianza del coseno per la identificazione e la verifica dei volti.

Per quanto riguarda VGG16, abbiamo utilizzato la stessa architettura di [154], mentre per ResNet50 abbiamo utilizzato l'architettura descritta in [53].

In particolare, in entrambe le reti, l'input è un'immagine del volto di 224×224 e l'embedding è calcolato applicando il *Global Average Pooling* all'output dell'ultimo blocco convoluzionale della rete.

Ciò significa che con VGG16 l'embedding è un vettore di caratteristiche a 512 elementi, mentre per ResNet50 è un vettore di caratteristiche a 2048 elementi.

Seguendo i risultati ottenuti da [53], abbiamo normalizzato in $L2$ le incorporazioni calcolate con entrambe le CNN.

L'addestramento delle reti è lo stesso descritto in [154] per il modello VGG16 e in [53] per i modelli ResNet50.

Pertanto, VGG16 è stato addestrato da zero sul dataset VGGFace, utilizzando la funzione di perdita tripla e la *Stochastic Gradient Descent* (SGD) per l'ottimizzazione, con lotti di 64 campioni e un tasso di apprendimento iniziale pari a 0,01, diminuito tre volte di un fattore 10 quando l'accuratezza sul set di validazione ha smesso di aumentare.

ResNet50 è stato addestrato da zero sul dataset VGGFace2, utilizzando la funzione di perdita *soft-max* e l'SGD per l'ottimizzazione, con lotti di 256 campioni e un tasso di apprendimento iniziale pari a 0,1, diminuito due volte di un fattore 10 quando l'errore ha smesso di diminuire.

Invece di ripetere l'addestramento, abbiamo applicato i pesi originali della rete [12] [13], utilizzando la conversione Keras dei modelli Caffe originali[14] .

[12] I pesi originali di VGG16 possono essere scaricati da: *https://www.robots.ox.ac.uk/vgg/software/vgg_face/*

[13] I pesi originali di ResNet50 possono essere scaricati da: *https://github.com/ox-vgg/vgg_face2*

[14] La conversione in Keras delle CNN è disponibile in: *https://github.com/rcmalli/keras-vggface*

5.3.5 Protocollo sperimentale e metriche di valutazione

Abbiamo testato la capacità di riconoscimento di VGG16 e ResNet50, addestrati su VGGFace e VGGFace2, sulle immagini del database SCFace e sul dataset proposto.

In particolare, abbiamo definito diversi sottoinsiemi di foto segnaletiche da utilizzare come immagini di riferimento per il riconoscimento.

A questo proposito, il database SCFace comprende 130 soggetti, con 9 immagini in posa per soggetto (le foto segnaletiche) più 21 immagini per soggetto ritagliate dai fotogrammi delle telecamere di sicurezza.

Come spiegato nella Sezione 5.3.1, il database SCFace è il più adeguato in aggiunta al dataset proposto per studiare l'impatto dell'uso di foto segnaletiche prese da più punti di vista.

Infatti, comprende 9 immagini in posa per soggetto, prese sistematicamente da diverse angolazioni sul piano orizzontale, che vanno da -90° a 90° (cioè, dal profilo sinistro a quello destro, con 0° che è l'immagine frontale) con passi di 22,5°.

La Tabella 5.2 elenca i sottoinsiemi di foto segnaletiche utilizzati come immagini di riferimento per il riconoscimento dei volti sulle immagini delle telecamere di sicurezza del database SCFace e del FRMDB.

Per ciascun database, la tabella descrive gli angoli da cui sono state scattate le foto segnaletiche come una coppia (h, v), dove h è l'angolo sul piano orizzontale e v è l'angolo sul piano verticale.

Per il database SCFace, l'angolo v è sempre $0\circ$, poiché non sono disponibili angoli diversi sul piano verticale.

I sottoinsiemi di foto segnaletiche sono composti come segue:

- Solo l'immagine frontale, cioè quella a (0°, 0°) per entrambi i database. Abbiamo chiamato questi sottoinsiemi "Test F" (poiché, nel database SCFace, "F" è l'etichetta data alle immagini frontali).

- L'immagine frontale, l'angolo sinistro più vicino all'immagine frontale (che è (-22.5°, 0°) per (-45°, 0°) per il database SCFace e (-45°, 0°) per il FRMDB), e l'angolo destro più vicino all'immagine frontale (SCFace: (22,5°, 0°); FRMDB: (45°, 0°). Abbiamo chiamato questi sottoinsiemi "Test F-L1-R1", utilizzando le etichette delle immagini definite nel database SCFace per queste immagini.

- La foto frontale e la foto del profilo destro, cioè (90°, 0°), per entrambi i database, simulano le foto segnaletiche della polizia attualmente disponibili nella maggior parte delle forze di polizia. Abbiamo chiamato questo sottoinsieme "Test 1".

- L'immagine frontale, l'immagine del profilo destro e il profilo sinistro, cioè (-90°, 0°). Noi abbiamo chiamato questi sottoinsiemi "Test 2".

- Le immagini del "Test 2" più le immagini più vicine alla foto frontale a partire dal profilo destro e dal profilo sinistro, che sono (77,5°, 0°) e (-77,5°, 0°) per il database SCFace, e (45°, 0°) e (-45 ∘ , 0 ∘) per l'FRMDB. Abbiamo chiamato questi sottoinsiemi "Test 3".

- Le immagini del "Test 3" più le immagini a (45°, 0°) e (-45°, 0°) per il database SCFace e le immagini a (135°, 0°) e (-135°, 0°) per il FRMDB. Abbiamo chiamato questi sottoinsiemi "Test 4". Infatti, il "Test 4" comprende tutte le immagini con 0 ∘ sul piano verticale del dataset proposto.

- Tutte le 9 foto segnaletiche per il database SCFace e le immagini del "Test 4" più tutte le foto segnaletiche con un angolo di 30° sul piano verticale per l'FRMDB. Abbiamo chiamato questi sottoinsiemi "Test 5".

- Tutte le 28 foto segnaletiche dell'FRMDB. Chiamiamo questo sottoinsieme "Test 6".

Per quanto riguarda le immagini dei volti delle telecamere di sicurezza da riconoscere, abbiamo utilizzato le immagini scattate a 1 m di distanza con le 5 telecamere a colori del database SCFace, escludendo tre soggetti, in quanto il loro volto è per lo più coperto dai capelli (ottenendo così 635 immagini di volti, 5 per soggetto), utilizzando le *Multi-Task Cascaded Convolutional Networks* (MTCNN) [181] e il Viola-Jones detection framework [160] (implementato nel classificatore a cascata OpenCV) per l'estrazione dei volti.

Per l'FRMDB, invece, abbiamo usato i volti ritagliati manualmente dai fotogrammi dei video delle telecamere di sicurezza.

Pertanto, nel dataset proposto ci sono 5 immagini per soggetto da riconoscere (coerentemente con il database SCFace), per un totale di 210 immagini di volti.

Per ogni sottoinsieme di foto segnaletiche, abbiamo registrato la capacità di ogni CNN di identificare il soggetto in ogni immagine proveniente dalle telecamere di sicurezza, registrando se il soggetto era incluso nella *top-1*, *top-3*, *top-5* e *top-10* delle foto segnaletiche più simili e nella *top-1*, *top-3*, *top-5* e *top-10* delle identità più vicine.

Per valutare la somiglianza tra un volto in una foto segnaletica e un volto nell'immagine di una telecamera di sicurezza, utilizziamo le CNN per calcolare le incorporazioni dei volti e la distanza euclidea come misura di somiglianza tra due incorporazioni di volti.

Dato tale criterio di somiglianza, misuriamo l'accuratezza come segue:

Tabella 5.2: Sottoinsiemi di foto segnaletiche del database SCFace e del FRMDB utilizzati come immagini di riferimento nei test. La tabella elenca il nome che diamo a ogni sottoinsieme e, per ogni database, gli angoli da cui sono state prese le foto segnaletiche incluse come coppia (h, v), dove h è l'angolo sul piano orizzontale e v è l'angolo sul piano verticale.

Subset name	**Mugshots (SCFace)**	**Mugshots (FRMDB)**
"Test F"	(0°, 0°)	(0°, 0°)
"Test F-L1-R1"	(0°, 0°), (-22.5°, 0°), (22.5°, 0°)	(0°, 0°), (-45°, 0°), (45°, 0°)
"Test 1"	(0°, 0°), (90°, 0°)	(0°, 0°), (90°, 0°)
"Test 2"	(0°, 0°), (90°, 0°), (-90°, 0°)	(0°, 0°), (90°, 0°), (-90°, 0°)
"Test 3"	(0°, 0°), (90°, 0°), (-90°, 0°), (77.5°, 0°), (-77.5°, 0°)	(0°, 0°), (90°, 0°), (-90°, 0°), (45°, 0°), (45°, 0°)
"Test 4"	(0°, 0°), (90°, 0°), (-90°, 0°), (77.5°, 0°), (-77.5°, 0°), (45°, 0°), (-45°, 0°)	(0°, 0°), (135°, 0°), (-135°, 0°), (90°, 0°), (-90°, 0°), (45°, 0°), (45°, 0°)
"Test 5"	(0°, 0°), (90°, 0°), (-90°, 0°), (77.5°, 0°), (-77.5°, 0°), (45°, 0°), (-45°, 0°), (-22.5°, 0°), (22.5°, 0°)	(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"	None	(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°)

- Per le foto segnaletiche più simili, calcoliamo l'accuratezza come il numero di immagini delle telecamere di sicurezza per le quali il soggetto corretto era nella top-1, top-3, top-5 e top-10 delle foto segnaletiche più simili rispetto al numero totale di immagini delle telecamere di sicurezza.

- Per le identità più simili, calcoliamo l'accuratezza come il numero di immagini delle telecamere di sicurezza per le quali il soggetto corretto era nella top-1, top-3, top-5 e top-10 delle identità più vicine rispetto al numero totale di immagini delle telecamere di sicurezza.

Ovviamente, l'identità top-1 e la foto segnaletica top-1 si sovrappongono.

Ad esempio, se l'immagine di una telecamera di sicurezza contiene il soggetto "003" e le nove foto segnaletiche più vicine sono quelle riportate nella Tabella 5.3, il soggetto corretto è nella top-5 delle foto segnaletiche più simili (non è nella top-1, poiché la foto segnaletica più simile è quella frontale del soggetto 001 e non è nemmeno nella top-3).

Tuttavia, il soggetto corretto è nella top-3 delle identità più vicine, poiché la prima identità riconosciuta è "001", la seconda è "002" e "003" è la terza.

La top-1 è la stessa anche se si considerano le foto segnaletiche o le identità, dato che si basa sull'immagine di riferimento che ha l'incorporamento più vicino al volto nell'immagine di una telecamera di sicurezza.

5.3.6 Risultati e discussione

Per convalidare l'FRMDB, fornire un benchmark iniziale per il dataset proposto e valutare l'impatto dell'uso di diversi sottoinsiemi di foto segnaletiche

Tabella 5.3: Esempio di misura dell'accuratezza: data la classifica della tabella, nell'ipotesi che il soggetto da riconoscere sia "003", il soggetto corretto si trova nella top-5 delle foto segnaletiche più simili e nella top-3 delle identità più vicine ("003" è la terza identità riconosciuta).).

1. 008 (0°, 0°)	**6.** 001 (0°, 0°)
2. 009 (0°, 0°)	**7.** 005 (45°, 0°)
3. 009 (45°, 0°)	**8.** 005 (45°, 30°)
4. 008 (-45°, 30°)	**9.** 002 (0°, 0°)
5. 005 (0°, 0°)	...

per il riconoscimento dei volti nei fotogrammi delle telecamere di sicurezza, abbiamo eseguito dei test comparativi utilizzando la metodologia descritta nella Sezione 5.3.2.

In particolare, abbiamo eseguito i test su un taccuino *Jupyter*, disponibile nel repository pubblico GitHub degli esperimenti, in un ambiente cloud (Google Colab), utilizzando Keras 2.8.0 e TensorFlow 2.8.2 per costruire le CNN e caricare i pesi della rete.

In questa sezione, quindi, discutiamo i risultati sul database SCFace e sul dataset proposto, inoltre, elenchiamo i limiti della ricerca descritta.

Risultati sul database SCFace

La Figura 5.9, mostra i risultati ottenuti da VGG16 e ResNet50, pre-addestrati sui dataset VGGFace e VGGFace2, sulle 5 immagini di telecamere di sicurezza scattate a 1 m di distanza del database SCFace.

ResNet50 ottiene un'accuratezza migliore di VGG16 nelle top-1, top-3, top-5 e top-10, indipendentemente dal fatto che si considerino le identità o le foto segnaletiche più importanti.

Infatti, con ResNet50, il soggetto corretto si trova nelle 10 foto segnaletiche più vicine e i soggetti nel 99% delle immagini delle telecamere di sicurezza, per qualsiasi sottoinsieme delle immagini di riferimento (Figura 5.7h).

In realtà, considerare le prime tre identità è sufficiente per il riconoscimento del volto con ResNet50 (Figura 5.7d), dato che l'accuratezza è superiore al 98% nella maggior parte dei test (l'accuratezza è del 97% nel "Test 1", che si basa solo sulla foto segnaletica frontale e sul profilo destro come immagini di riferimento per il riconoscimento del volto).

Valutando l'impatto dell'uso di diverse foto segnaletiche, i risultati di accuratezza peggiorano quando, invece di usare solo la foto frontale per il riconoscimento del volto ("Test F"), si aggiungono la foto di profilo destra e sinistra

("Test 2") o anche più immagini di riferimento ("Test 3", "Test 4", "Test 5") prese da diverse angolazioni sul piano orizzontale.

Questo è evidente con VGG16 in tutte le classifiche "Top" (Figure 5.7a, 5.7c, 5.7e e 5.7g) indipendentemente dalla considerazione delle foto segnaletiche o delle identità più vicine. ResNet50 (Figure 5.7b, 5.7d, 5.7f e 5.7h) presenta la stessa tendenza, anche se i risultati sono migliori in generale.

Con entrambe le CNN, i risultati migliori si ottengono utilizzando solo la foto segnaletica frontale come immagine di riferimento ("Test F"), mentre i risultati peggiori si ottengono utilizzando la foto frontale e la foto del profilo destro ("Test 1"), cioè l'immagine attualmente raccolta dalla maggior parte delle forze di polizia durante il fotosegnalamento.

L'accuratezza aumenta leggermente negli altri test rispetto al "Test 1", ma non è mai migliore di quella ottenuta con la sola foto frontale.

L'unica eccezione a questa tendenza è l'accuratezza top-1 per VGG16 (Figura 5.7a), dove il sottoinsieme che utilizza la foto frontale (F), la foto a -22,5° (L1) e la foto a 22,5° (R1) ottiene un'accuratezza del 72,14%, contro il 71,21% dell'utilizzo della sola foto frontale.

Pertanto, i risultati sembrano suggerire che l'utilizzo di più immagini scattate da diverse angolazioni non sia adatto al riconoscimento dei volti: la procedura abitualmente attuata dalle forze di polizia per la raccolta delle foto segnaletiche potrebbe non essere degna di essere cambiata, dato che la migliore accuratezza si ottiene utilizzando come immagine di riferimento solo la foto segnaletica frontale.

Tuttavia, questo non può essere considerato un risultato generale e conclusivo.

Infatti, una spiegazione per questo comportamento contro intuitivo delle prestazioni di riconoscimento dei volti quando si utilizzano più foto segnaletiche è la natura delle immagini delle telecamere di sicurezza disponibili nel database SCFace.

Le immagini delle telecamere di sicurezza sono quasi frontali (Figura 5.7), mentre le 9 foto segnaletiche utilizzate come immagini di riferimento sono prese sistematicamente da -90° a 90° con incrementi di 22,5° sul piano orizzontale.

Pertanto, le foto segnaletiche diverse da quella frontale aggiungono rumore al compito di riconoscimento del volto, peggiorando l'accuratezza.

Le immagini delle telecamere di sicurezza non sono pienamente rappresentative della realtà.

Infatti, le immagini delle telecamere di sicurezza possono includere volti sotto diverse prospettive, vicino a una foto di profilo o anche con un'angolazione più elevata.

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

Figura 5.7: Immagini a colori delle telecamere di sicurezza riprese a 1 m di distanza per il soggetto 001 del database SCFace. Le immagini delle prime quattro telecamere (a-d) includono un'immagine frontale del volto, mentre la quinta (e) è leggermente a destra del soggetto.

Risultati sul FRMDB

La figura 5.8 mostra i risultati sul dataset proposto.

Mentre ResNet50 ottiene punteggi migliori di VGG16 in tutte le classifiche "Top" anche sul dataset proposto, l'accuratezza è significativamente inferiore a quella ottenuta su SCFace per entrambe le CNN.

Questo risultato conferma le caratteristiche difficili dell'FRMDB: come mostrato nella figura 5.6, i fotogrammi delle telecamere di sicurezza provengono da diverse prospettive (invece di includere solo volti frontali come nell'SCFace).

Inoltre, i video sono a bassa risoluzione ($352 \times 288 pixel$), emulando le telecamere di sicurezza pubbliche di bassa qualità, che possono includere un volto molto piccolo (come $85 \times 85 pixel$) da riconoscere dalle foto segnaletiche.

Pertanto, data la minore accuratezza, il dataset proposto sembra una migliore rappresentazione delle sfide che si presentano nella vita reale per il riconoscimento dei volti attraverso le telecamere di sicurezza.

A differenza del database SCFace, il sottoinsieme composto dalla sola immagine frontale ("Test F") non ottiene mai la migliore accuratezza con ResNet50.

Invece, il sottoinsieme composto dall'immagine frontale, dall'immagine a (-45°, 0°) e dall'immagine a (45°, 0°), cioè "Test F-L1-R1", ottiene i migliori risultati in tutte le classifiche "Top" (Figure 5.9b, 5.9d, 5.9f, 5.9h).

Ad esempio, l'identità corretta è nella top-10 (Figura 5.9h) delle identità più vicine per il 74,87% delle telecamere di sicurezza, utilizzando come immagini di riferimento le immagini del sottoinsieme "Test F-L1-R1".

Tale percentuale diminuisce al 71,28% utilizzando la sola immagine frontale come immagine di riferimento. Anche con VGG16 (Figure 5.9a, 5.9c, 5.9e, 5.9g), non c'è una chiara predominanza del sottoinsieme composto dalla sola immagine frontale, a differenza del database SCFace.

Ad esempio, con il sottoinsieme "Test F-L1-R1", il soggetto corretto è nella top-3 (Figura 5.9c) delle identità e delle foto segnaletiche per il 44,62% delle

immagini delle telecamere di sicurezza, mentre con la sola immagine frontale questa percentuale scende al 43,08%.

I test sul dataset proposto non presentano la stessa tendenza mostrata con il database SCFace.

Anche se i sottoinsiemi "Test 1", cioè le immagini di fotosegnalamento attuali, e "Test 2" (che aggiunge il profilo sinistro al sottoinsieme precedente) ottengono i risultati peggiori con entrambe le CNN in entrambi i database, con il dataset proposto l'aumento del numero di immagini migliora i risultati in alcuni casi.

In particolare, con VGG16, utilizzando tutte le immagini a 0° e 30° sul piano verticale ("Test 5") si ottiene quasi lo stesso risultato del "Test F-L1-R1", riuscendo a riconoscere il soggetto nella top-3 delle identità più vicine ((Figura 5.9c) nel 44,1% dei fotogrammi.

In generale, i risultati ottenuti utilizzando sottoinsiemi di foto segnaletiche con più immagini ("Test 3 e 6") sono migliori rispetto all'utilizzo della foto frontale o della foto frontale e del profilo destro.

Questi risultati e la minore accuratezza ottenuta rispetto al database SCFace convalidano il dataset proposto come adeguato per studiare l'effetto dell'utilizzo di foto segnaletiche da più punti di vista per il riconoscimento dei volti nelle telecamere di sorveglianza.

5.3.7 Limitazioni

I risultati presentati in questo lavoro sono promettenti, ma presentano alcune limitazioni.

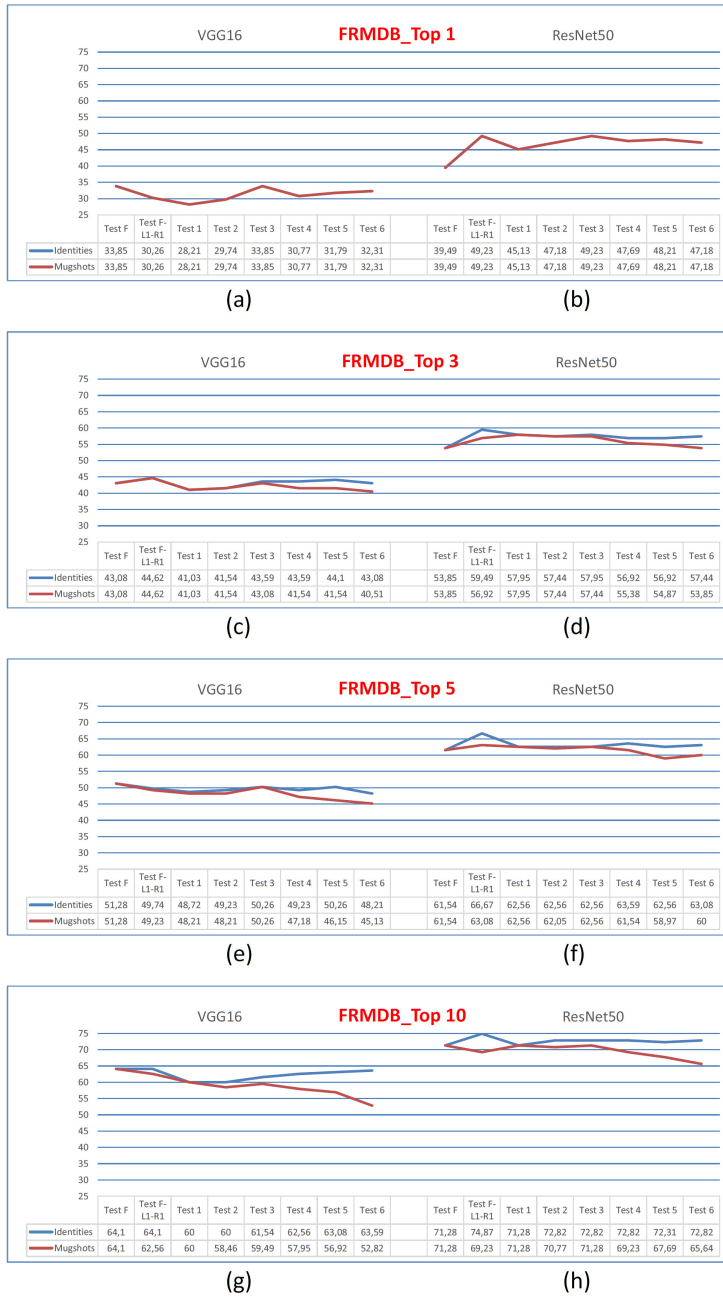
Per quanto riguarda il set di dati proposto, la limitazione principale è il numero di soggetti unici, 39, che potrebbe sembrare basso.

Tuttavia, il dataset non è destinato all'apprendimento di caratteristiche del volto.

Per questo compito, in letteratura sono disponibili database con un numero adeguato di immagini, fino a un milione. Il dataset ha invece lo scopo di mettere alla prova le tecniche di riconoscimento facciale nel riconoscere i soggetti nei fotogrammi dei video delle telecamere di sicurezza, utilizzando le foto segnaletiche come immagini di riferimento. Pertanto, l'FRMDB può essere utilizzato per i test, piuttosto che per l'apprendimento.

Inoltre, le dimensioni delle foto segnaletiche ($972x544pixel$) e dei video delle telecamere di sicurezza ($352x288pixel$) possono sembrare piccole. Tuttavia, i filmati delle telecamere a circuito chiuso sono solitamente a bassa risoluzione e di bassa qualità, al punto che stanno emergendo tecniche di miglioramento della qualità basate sulla DL [182].

Pertanto, riteniamo che il set di dati proposto sia rappresentativo di uno scenario di vita reale.



	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Identities	33,85	30,26	28,21	29,74	33,85	30,77	31,79	32,31	39,49	49,23	45,13	47,18	49,23	47,69	48,21	47,18
Mugshots	33,85	30,26	28,21	29,74	33,85	30,77	31,79	32,31	39,49	49,23	45,13	47,18	49,23	47,69	48,21	47,18

	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Identities	43,08	44,62	41,03	41,54	43,59	43,59	44,1	43,08	53,85	59,49	57,95	57,44	57,95	56,92	56,92	57,44
Mugshots	43,08	44,62	41,03	41,54	43,08	41,54	41,54	40,51	53,85	56,92	57,95	57,44	57,44	55,38	54,87	53,85

	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Identities	51,28	49,74	48,72	49,23	50,26	49,23	50,26	48,21	61,54	66,67	62,56	62,56	62,56	63,59	62,56	63,08
Mugshots	51,28	49,23	48,21	48,21	50,26	47,18	46,15	45,13	61,54	63,08	62,56	62,05	62,56	61,54	58,97	60

	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Identities	64,1	64,1	60	60	61,54	62,56	63,08	63,59	71,28	74,87	71,28	72,82	72,82	72,82	72,31	72,82
Mugshots	64,1	62,56	60	58,46	59,49	57,95	56,92	52,82	71,28	69,23	71,28	70,77	71,28	69,23	67,69	65,64

Figura 5.8: Misure di accuratezza percentuale in top-1 (a-b), top-3 (c-d), top-5 (e-f) e top-10 (g-h) per VGG16 e ResNet50 sul dataset FRMDB proposto. L'asse delle ordinate è stato tagliato tra il 25% e il 75% per apprezzare meglio visivamente gli scostamenti percentuali

Per quanto riguarda i risultati presentati, le CNN testate si basano sui risultati della ricerca sul riconoscimento dei volti presentati nella letteratura scientifica, come spiegato nella Sezione 2. Tuttavia, per ottenere risultati più generali sull'uso di diversi sottoinsiemi di foto segnaletiche per il riconoscimento di volti in fotogrammi provenienti da telecamere di sicurezza, è necessario effettuare uno studio sistematico su modelli alternativi e un confronto su più set di dati.

5.3.8 Conclusioni sul dataset FRMDB

Abbiamo presentato il FRMDB, ovvero un dataset che comprende 28 foto segnaletiche e 5 video provenienti da telecamere di sicurezza di 39 soggetti unici.

Il dataset proposto ha lo scopo di mettere alla prova le tecniche di riconoscimento facciale per l'identificazione dei soggetti nei video utilizzando le foto segnaletiche disponibili.

Sul dataset proposto e sul database SCFace, abbiamo testato due CNN consolidate, VGG16 e ResNet50, pre-addestrate sui dataset VGGFace e VGGFace2 per l'estrazione di caratteristiche del volto. Tali esperimenti consentono di trarre le seguenti conclusioni principali:

- Il set di dati proposto è adeguato per valutare le tecniche di riconoscimento facciale per l'identificazione dei soggetti nei video utilizzando le foto segnaletiche, tenendo conto dei diversi punti di vista.
- La minore accuratezza rispetto al database SCFace ha evidenziato la natura impegnativa del dataset.
- Il sottoinsieme di foto segnaletiche composto solo dal volto frontale non ha mostrato la stessa predominanza ottenuta con SCFace, poiché FRMDB include video di sorveglianza da più punti di vista.
- In entrambi i set di dati, le tradizionali immagini di segnalazione fotografica, cioè l'immagine frontale e il profilo destro, sono superate da altri sottoinsiemi di foto segnaletiche. In particolare, con l'FRMDB proposto, il sottoinsieme composto dall'immagine frontale e dalle immagini a $45\circ$ sul piano orizzontale raggiunge la migliore accuratezza nella maggior parte dei test.

Sono necessarie ulteriori ricerche per ottenere risultati sul numero ideale di foto segnaletiche, cercando un compromesso con la necessità di strumenti aggiuntivi (e di spazio di archiviazione) necessari alle forze dell'ordine per raccogliere più foto segnaletiche.

Per ottenere risultati più generali, è necessario testare altre tecniche, tra cui quelle per il Pose-Invariant Face Recognition (PIFR) e la stima della posa, al fine di scegliere la foto segnaletica con la posa più vicina ai fotogrammi della telecamera di sicurezza prima del confronto.

I lavori futuri sul dataset proposto affronteranno le limitazioni descritte aggiungendo un maggior numero di soggetti, con video a risoluzione più elevata, in modo da avere una maggiore variabilità e quindi costruire un database ancora più rappresentativo della videosorveglianza nella vita reale.

5.4 Analisi dell'impatto delle foto segnaletiche sulla verifica di volti nell'ambito di indagini criminali

[*pubblicato*][15]

Con la crescita dei sistemi di sorveglianza in tutto il mondo, in grado di fornire informazioni accurate e ricche in molte applicazioni di sicurezza [141], il settore della pubblica sicurezza sta assistendo a un'applicazione senza precedenti dell'Intelligenza Artificiale (IA) e del Deep Learning (DL) per le indagini sui crimini e il supporto alle forze dell'ordine [25].

A questo proposito, l'IA e il DL includono applicazioni come il rilevamento automatico della violenza [68, 183], l'analisi delle reti sociali [184], il rilevamento automatico di armi [185] e molte altre.

Tuttavia, alcune caratteristiche fisiche dell'ambiente ripreso, come le variazioni di posa e illuminazione, possono avere effetti negativi sulla capacità di riconoscimento automatico. A tal proposito, Franco et al. propongono una nuova tecnica per il riconoscimento facciale, basata sull'uso congiunto di modelli 3D e immagini 2D, progettata appositamente per sopperire ai cambiamenti di posa e illuminazione. A partire da un modello 3D dell'utente, generano una molteplicità di immagini 2D per ogni persona, in cui varia sia la posa che l'illuminazione del volto. Usando tali immagini per l'addestramento supervisionato della rete, essa genera un modello per ogni utente che servirà per il riconoscimento automatico attraverso l'abbinamento con le immagini 2D [186].

Il riconoscimento automatico di volti è una di queste applicazioni e una delle tecniche biometriche più naturali utilizzate per l'identificazione [42].

Ha un vantaggio significativo rispetto ad altre tecniche biometriche: può essere effettuata passivamente, cioè senza azioni esplicite da parte del soggetto da identificare [43].

[15] Contardo, P., Di Lorenzo, E., Falcionelli, N., Dragoni, A. F., Sernani, P. (2022, October). Analyzing the impact of police mugshots in face verification for crime investigations. In 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) (pp. 236-241). IEEE.

Per questo, data l'ampia gamma di possibili applicazioni in ambito di sicurezza, il riconoscimento facciale ha attirato l'interesse della comunità della Computer Vision per oltre 40 anni.

In effetti, le funzionalità di riconoscimento facciale sono implementate anche nei sistemi di supporto alle indagini criminali già esistenti, ampliando le potenzialità di indagine delle forze di polizia e delle agenzie di contrasto.

In questo modo, il tradizionale e valido Autonated Fingerprint Identification System (AFIS) viene esteso con funzionalità di riconoscimento delle immagini utili quando, al posto delle impronte digitali, sono disponibili solo immagini della persona che ha commesso un reato, implementando sistemi di identificazione biometrica multimodale [145].

Il sistema SARI (Sistema Automatico Riconoscimento Immagini) utilizzato dalla Polizia Scientifica italiana [70] è un esempio di tale integrazione.

La maggior parte delle forze di polizia di diversi paesi raccolgono abitualmente due immagini, comunemente note come foto segnaletiche (la foto frontale e il profilo destro), insieme alle impronte digitali e alle informazioni personali di un soggetto, per vari scopi, che vanno dal rilascio di documenti alla registrazione di criminali.

Esiste quindi una chiara connessione tra il riconoscimento facciale e questa procedura di raccolta dati in ambito di verifica dei volti, cioè il compito di verificare se due immagini di volti appartengono alla stessa persona, e in ambito di identificazione dei volti, cioè il compito di identificare un volto in un insieme di soggetti noti.

Nonostante il riconoscimento facciale sia promettente per l'identificazione di potenziali criminali nei video di sicurezza, mancano ricerche per capire fino a che punto le CNN siano efficaci nell'identificazione di un soggetto noto quando sono disponibili solo le due immagini standard del volto derivanti dalla procedura di fotosegnalamento come immagini di riferimento [147, 187].

Il presente lavoro compie un primo passo per colmare tale lacuna, analizzando l'uso di più di due foto segnaletiche scattate da punti di vista diversi per la verifica delle prestazioni delle CNN nel riconoscimento facciale.

A tal fine, abbiamo misurato le metriche di accuratezza nel processo di verifica del volto quando diversi sottoinsiemi di foto segnaletiche sono disponibili come campioni di riferimento.

L'obiettivo è capire se la modifica della procedura di acquisizione delle foto segnaletiche della polizia può avere un impatto positivo sulla verifica del volto e sulla garanzia d'identità, giustificando lo sforzo necessario per scattare più foto e memorizzarle nei database.

A questo proposito, la ricerca presentata in questo lavoro aggiunge i seguenti contributi allo stato dell'arte del riconoscimento facciale:

- fornisce un confronto tra due CNN per la verifica di immagini di volti provenienti dalle telecamere di videosorveglianza utilizzando foto segnaletiche;
- analizza le prestazioni della verifica del volto quando sono disponibili diversi sottoinsiemi di foto segnaletiche, prese da vari punti di vista, come riferimento.

In particolare, abbiamo confrontato le prestazioni delle reti VGG16 [174] e ResNet50 [54], pre-addestrati sul dataset VGGFace2 [53] per l'estrazione delle caratteristiche del volto, nel compito di verifica dei volti sul database "Surveillance Cameras Face" (SCFace) [153].

Questi contributi preliminari forniscono una panoramica sull'impatto che un insieme di foto segnaletiche, diverse dalle immagini frontali e di profilo standard raccolte dalle forze di polizia durante la procedura di fotosegnalamento, potrebbe avere sull'identificazione di soggetti sospetti registrati in filmati di telecamere di sicurezza.

Di seguito si fornisce una breve *rassegna della letteratura* sull'evoluzione delle tecniche di riconoscimento facciale, giustificando la scelta delle CNN per il nostro confronto; quindi si descrive la *metodologia* implementata per eseguire il nostro esperimento comparativo.

La sezione *Risultati e discussione* presenta i risultati del nostro test, analizzando le prestazioni di accuratezza sul database SCFace utilizzando diversi set di foto segnaletiche come immagini di riferimento.

Infine, la sezione *Conclusioni* elenca le osservazioni conclusive di questa ricerca, suggerendo lavori futuri.

5.4.1 Rassegna della letteratura

Le prime metodologie per il riconoscimento dei volti apparse in letteratura si basavano su tecniche pure di Computer Vision.

Ad esempio, Turk e Pentland [44] hanno proposto Eigenfaces, applicando l'analisi delle componenti principali (PCA) per estrarre un vettore di caratteristiche che massimizzasse la varianza in un insieme di immagini di addestramento. Proiettando l'immagine di un volto nello spazio ottenuto con la PCA, l'identificazione del volto può essere eseguita con un metodo di tipo "nearest neighbor", calcolando la distanza dalle immagini di addestramento. Mentre Eigenfaces massimizza la varianza interclasse tra immagini di volti di soggetti diversi, non tiene conto della varianza intraclasse tra le immagini di volti di un singolo soggetto. Il metodo Fisherfaces [45] aggiunge invece alla PCA l'Analisi Discriminante Lineare (LDA), al fine di minimizzare la varianza intraclasse. A differenza di Eigenfaces e Fisherfaces, Ahonen et al. [46] hanno proposto di calcolare i "Local Binary Patterns Histograms" (LBPH) sulle immagini dei volti,

dividendole in regioni per calcolare i "Local Binary Patterns" (LBP). Analogamente a Eigenfaces e Fisherfaces, una funzione di distanza basata sugli LBPH può essere utilizzata per l'identificazione dei volti.

Mentre queste tecniche (e quelle derivate) hanno ottenuto una buona precisione su insiemi di dati in cui alcuni parametri come la posa, l'illuminazione e l'espressione del volto sono fissi, sono insufficienti per estrarre caratteristiche di identità stabili e invarianti ai cambiamenti del mondo reale [47], come nelle immagini ottenute da video e telecamere di sorveglianza. Pertanto, non sono adatte ad ambiti di pubblica sicurezza reale, quando si confrontano le sole due foto segnaletiche (una frontale e una di profilo) raccolte dalle agenzie di polizia in condizioni ideali, con le immagini ottenute sul campo. Al contrario, le tecniche basate sul Deep Learning si sono dimostrate in grado di estrarre caratteristiche invarianti al variare delle condizioni di espressione facciale, illuminazione e posa. Anche se prima dell'affermazione del Deep Learning sono stati studiati metodi che combinano le reti neurali con formalismi logici, come la Belief Revision [49], sono le CNN ad aver migliorato significativamente l'accuratezza nel riconoscimento dei volti senza vincoli. A tal fine, Taigam et al. [50] hanno presentato DeepFace, una CNN a 8 strati per l'elaborazione di immagini di volti a 3 canali 152x152, in grado di ottenere un'accuratezza del 97,35% sul dataset Labeled Faces in the Wild (LFW) [51]. Analogamente, Schroff et al. [52] hanno proposto Facenet, una CNN a 22 strati addestrata in diversi esperimenti con un numero variabile di immagini di volti, tra 100 e 200 milioni, appartenenti a 8 milioni di soggetti diversi. Hanno ottenuto un'accuratezza del 99,63% su LFW, utilizzando immagini di input 220 x 220. Cao et al. [53] hanno dimostrato l'efficacia di ResNet-50 [54], una CNN a 50 strati basata sull'apprendimento residuale in grado di ottenere un errore di identificazione top-1 del 3,9% sul dataset VGGFace2 (composto da oltre 3 milioni di immagini di più di 9 mila soggetti).

Le tecniche di riconoscimento facciale basate su CNN hanno dimostrato di essere robuste alle condizioni mutevoli tipiche delle immagini facciali senza vincoli di posa, illuminazione, espressione [55]. Per questo motivo, abbiamo confrontato due CNN ben consolidate, ovvero VGG16 e ResNet50 , per compiere un primo passo nella valutazione delle capacità di riconoscimento quando si utilizzano diversi sottoinsiemi di foto segnaletiche come immagini di riferimento per un compito di verifica del volto.

5.4.2 Metodologia di confronto

Per ottenere gli obiettivi descritti nell'Introduzione, abbiamo confrontato sul database SCFace VGG16 e ResNet50, addestrati sul dataset VGGFace2, utilizzando le CNN per calcolare gli embeddings dei volti (cioè i vettori di ca-

Tabella 5.4: I sottoinsiemi di foto segnaletiche del database SCFace utilizzati nel nostro confronto. Sono riportati il nome che diamo a ciascun sottoinsieme (prima colonna), il nome delle foto segnaletiche incluse secondo il database SCFace originale (seconda colonna) e l'angolo sul piano orizzontale (terza colonna) delle foto segnaletiche incluse.

Nome del sottoinsieme	**Foto segnaletiche**	**Angoli**
"Test F"	F	0°
"Test F-L1-R1"	F, L1, R1	0°, -22.5°, 22.5°
"Test 1"	F, R4	0°, 90°
"Test 2"	F, R4, L4	0°, 90°, -90°
"Test 3"	F, R4, L4, R3, L3	0°, 90°, -90°, 77.5°, -77.5°
"Test 4"	F, R4, L4, R3, L3, R2, L2	0°, 90°, -90°, 77.5°, -77.5°, 45°, -45°
"Test 5"	F, R4, L4, R3, L3, R2, L2, R1, L1	0°, 90°, -90°, 77.5°, -77.5°, 45°, -45°, 22.5°, -22.5°

ratteristiche che descrivono le immagini) per un compito di verifica dei volti. Il database SCFace comprende 130 soggetti, con 9 foto segnaletiche per soggetto e 21 immagini per soggetto ritagliate dai fotogrammi delle telecamere di sicurezza. Per quanto ne sappiamo, il database SCFace è il più adeguato tra quelli disponibili in letteratura per studiare l'impatto dell'uso di foto segnaletiche scattate da diversi punti di vista. Infatti, le 9 foto segnaletiche sono sistematicamente prese da diverse angolazioni sul piano orizzontale, che vanno da -90° a 90° (cioè, dal profilo sinistro a quello destro, con 0° come immagine frontale) con incrementi di 22,5°.

Abbiamo definito sette diversi sottoinsiemi di foto segnaletiche come immagini di riferimento per il compito di verifica del volto sulle immagini delle telecamere di sicurezza del database SCFace. Le foto segnaletiche provengono dalle 9 foto segnaletiche del database SCFace e ogni sottoinsieme è composto secondo la Tabella 5.4, che include il nome dato a ciascuna foto segnaletica nel database SCFace originale.

In particolare, i sottoinsiemi sono composti come segue:

- Solo l'immagine frontale (0°). Abbiamo chiamato questo sottoinsieme "Test F".

- L'immagine frontale (0°), l'angolo sinistro più vicino all'immagine frontale (-22,5°) e l'angolo destro più vicino all'immagine frontale (22,5°). Abbiamo chiamato questo sottoinsieme "Test F-L1-R1", utilizzando le etichette delle immagini definite nel database SCFace.

- L'immagine frontale (0°) e l'immagine del profilo destro (90°), che simulano le foto segnaletiche della polizia attualmente disponibili nella maggior parte delle forze di polizia. Abbiamo chiamato questo sottoinsieme "Test 1".

- la foto frontale (0°), la foto di profilo destra (90°) e il profilo sinistro (-90°). Abbiamo chiamato questo sottoinsieme "Test 2".
- le cinque immagini a 0°, 90°, -90°, 77,5° e -77,5°. Abbiamo chiamato questo sottoinsieme "Test 3".
- le sette immagini a 0°, 90°, -90°, 77,5°, -77,5°, 45° e -45°. Abbiamo chiamato questo sottoinsieme "Test 4".
- le nove immagini a 0°, 90°, -90°, 77,5°, -77,5°, 45°, -45°, 22,5° e -22,5°. Abbiamo chiamato questo sottoinsieme "Test 5".

Abbiamo scelto di tilizzare questi sottoinsiemi come immagini di riferimento per ogni soggetto del database SCFace sui cui eseguire la verifica del volto sulle immagini delle telecamere di sicurezza. A tal fine, per ogni soggetto, utilizziamo le immagini scattate a 1 m di distanza con le cinque telecamere a colori, escludendo tre soggetti, poiché il loro volto è per lo più coperto dai capelli (ovvero, eseguiamo la verifica del volto su 635 immagini di volti, 5 per soggetto). Quindi, per ogni sottoinsieme di foto segnaletiche, abbiamo registrato la capacità di ogni CNN di identificare il soggetto in ogni immagine proveniente dalle telecamere di sicurezza, registrando se il soggetto era incluso nella top-1, top-3, top-5 e top-10 delle foto segnaletiche più simili e nella top-1, top-3, top-5 e top-10 delle identità più vicine. Per valutare la somiglianza tra un volto in una foto segnaletica e un volto nell'immagine di una telecamera di sicurezza, utilizziamo le CNN per calcolare gli embedding dei volti e la distanza euclidea (calcolata sugli embedding normalizzati dei volti) come misura di somiglianza tra due embedding di volti. Dato questo criterio di somiglianza, misuriamo l'accuratezza come segue:

- per le foto segnaletiche più simili, calcoliamo l'accuratezza come il numero di immagini delle telecamere di sicurezza per le quali il soggetto corretto era nella top-1, top-3, top-5 e top-10 delle foto segnaletiche più simili rispetto al numero totale di immagini delle telecamere di sicurezza;
- Per le identità più simili, calcoliamo l'accuratezza come il numero di immagini delle telecamere di sicurezza per le quali il soggetto corretto era nella top-1, top-3, top-5 e top-10 delle identità più vicine rispetto al numero totale di immagini delle telecamere di sicurezza.

Ad esempio, se l'immagine di una telecamera di sicurezza contiene il soggetto "003" e le nove foto segnaletiche più vicine sono quelle riportate nella Tabella 5.5, il soggetto corretto è nella top-5 delle foto segnaletiche più simili (non è nella top-1, poiché la foto segnaletica più simile è quella frontale del soggetto 001 e non è nemmeno nella top-3). Tuttavia, il soggetto corretto è nella top-3

Tabella 5.5: Esempio di misura dell'accuratezza: data la classifica della tabella, nell'ipotesi che il soggetto da riconoscere sia "003", il soggetto corretto si trova nella top-5 delle foto segnaletiche più simili e nella top-3 delle identità più vicine ("003" è la terza identità riconosciuta).

1. 001_F	**6.** 007_F
2. 002_F	**7.** 003_R1
3. 002_R1	**8.** 003_L1
4. 001_L1	**9.** 009_F
5. 003_F	...

delle identità più vicine, poiché la prima identità riconosciuta è "001", la seconda è "002" e "003" è la terza. La top-1 è la stessa anche se si considerano le foto segnaletiche o le identità, dato che si basa sull'immagine di riferimento che ha l'incorporamento più vicino al volto nell'immagine di una telecamera di sicurezza.

Per quanto riguarda l'architettura di VGG16 e ResNet50, l'input è stato ridimensionato a 224 x 224, seguendo la stessa procedura di addestramento di [154] per VGG16 e di [53] per ResNet50. Nel caso di VGG16, un face embedding è composto dalle 512 caratteristiche generate come output dell'ultimo strato completamente connesso. Nel caso di ResNet50, un face embedding è composto dalle 2048 caratteristiche generate come output dell'ultimo strato completamente connesso. Per rilevare ogni volto disponibile nelle immagini del database SCFace, ovvero per eseguire l'estrazione del volto, abbiamo utilizzato le reti convoluzionali multi-task a cascata (MTCNN) [181]. Poiché con MTCNN non siamo riusciti a estrarre tutte le immagini di volti disponibili, abbiamo utilizzato il framework di rilevamento Viola-Jones [160], utilizzando l'implementazione del "cascade classifier" di OpenCV, sulle immagini rimanenti. Abbiamo escluso dagli esperimenti l'unica immagine delle camere di videosorveglianza su cui non è stato possibile rilevare il volto automaticamente. Per questo, gli esperimenti hanno riguardato un totale di 634 immagini da telecamere di sicurezza.

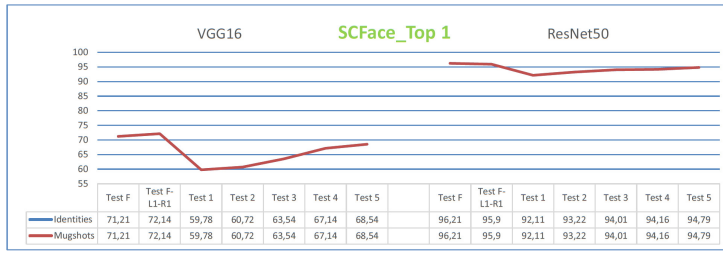
5.4.3 Risultati e discussione

La Figura 5.9 mostra i risultati ottenuti da VGG16 e ResNet50, addestrati sul dataset VGGFace2, nella verifica di volti su 634 immagini a colori di telecamere di sicurezza scattate a 1 m di distanza dal soggetto. In generale, ResNet50 ottiene un'accuratezza migliore di VGG16 nelle Top-1, Top-3, Top-5 e Top-10, indipendentemente dal fatto che si considerino le identità o le foto segnaletiche più importanti. Infatti, con ResNet50, il soggetto corretto si trova

nelle dieci foto segnaletiche più vicine e nei soggetti nel 99% delle immagini delle telecamere di sicurezza, per qualsiasi sottoinsieme di immagini di riferimento (Figura 5.9(h)). In realtà, considerare le prime tre identità è sufficiente con ResNet50 (Figura 5.9(d)), dato che l'accuratezza è superiore al 98% nella maggior parte dei test (l'accuratezza è del 97% nel test 1, utilizzando come immagini di riferimento solo la foto frontale e il profilo destro).

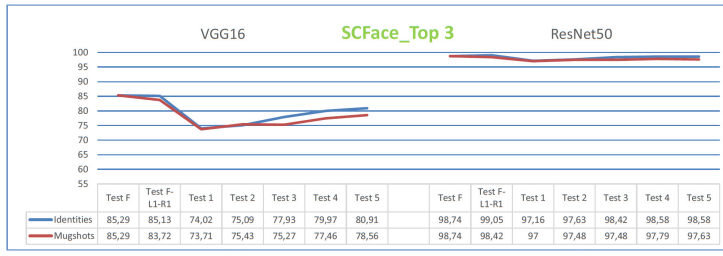
Per quanto riguarda l'impatto dell'uso di diverse foto segnaletiche, i risultati sembrano controintuitivi: i risultati di accuratezza peggiorano quando, invece di usare solo la foto frontale per la verifica del volto (Test F), si aggiungono la foto di profilo destra e sinistra (Test 2) o addirittura più immagini di riferimento (Test 3, Test 4, Test 5) prese da diverse angolazioni sul piano orizzontale. Questo è evidente con VGG16 in tutti i test (Figure 5.9 (a), (c), (e) e (g)) indipendentemente dal prendere in considerazione le foto segnaletiche o le identità più vicine. ResNet50 (Figure 5.9 (b), (d), (f) e (h)) presenta la stessa tendenza, anche se i risultati sono migliori in generale. Tuttavia, con entrambe le CNN, i risultati migliori si ottengono utilizzando solo la foto frontale come immagine di riferimento (Test F), mentre i risultati peggiori si ottengono utilizzando la foto frontale e la foto del profilo destro (Test 1), cioè l'immagine attualmente raccolta dalla maggior parte delle forze di polizia durante il fotosegnalamento. L'accuratezza aumenta leggermente negli altri test rispetto al Test 1, ma non è mai migliore di quella ottenuta con la sola foto frontale. L'unica eccezione a questa tendenza è l'accuratezza del Top 1 per il VGG16 (Figura 5.9 (a)), dove il sottoinsieme che utilizza l'immagine frontale (F), l'immagine a -22,5° (L1) e l'immagine a 22,5° (R1) ottiene un'accuratezza del 72,14%, contro il 71,21% dell'utilizzo della sola foto segnaletica frontale.

Pertanto, i risultati sembrano suggerire che l'utilizzo di più immagini scattate da diverse angolazioni non sia adatto alla verifica dei volti: la procedura abitualmente usata dalle forze di polizia per la raccolta delle foto segnaletiche potrebbe non dover essere cambiata, dato che la migliore accuratezza si ottiene utilizzando come immagine di riferimento solo la foto segnaletica frontale. Tuttavia, questo non può essere considerato un risultato generale e conclusivo. Infatti, una spiegazione per questo comportamento controintuitivo delle prestazioni nella verifica di volti quando si utilizzano più foto segnaletiche è nella natura delle immagini delle telecamere di sicurezza disponibili nel database SCFace. A tal fine, la Figura 5.7 mostra l'immagine a colori di cinque telecamere di sicurezza, a una distanza di 1 m, del soggetto 001 del database SCFace. Tutte le immagini del volto nelle foto delle telecamere di sicurezza sono quasi frontali, mentre le 9 foto segnaletiche utilizzate come immagini di riferimento sono prese sistematicamente da -90° a 90° con incrementi di 22,5°. Pertanto, le foto segnaletiche diverse da quella frontale aggiungono rumore alla verifica di volti, peggiorando le prestazioni di riconoscimento. Questo è il limite



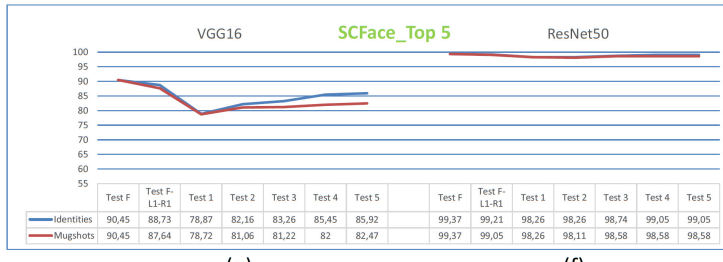
	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5
Identities	71,21	72,14	59,78	60,72	63,54	67,14	68,54	96,21	95,9	92,11	93,22	94,01	94,16	94,79
Mugshots	71,21	72,14	59,78	60,72	63,54	67,14	68,54	96,21	95,9	92,11	93,22	94,01	94,16	94,79

(a) (b)



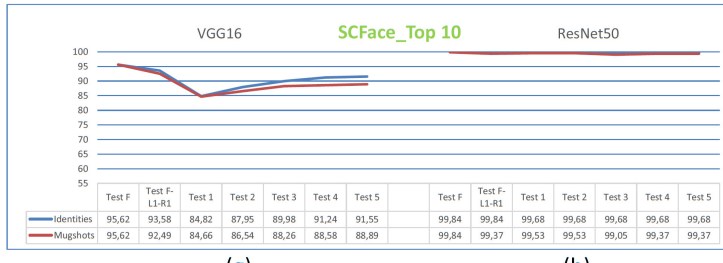
	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5
Identities	85,29	85,13	74,02	75,09	77,93	79,97	80,91	98,74	99,05	97,16	97,63	98,42	98,58	98,58
Mugshots	85,29	83,72	73,71	75,43	75,27	77,46	78,56	98,74	98,42	97	97,48	97,48	97,79	97,63

(c) (d)



	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5
Identities	90,45	88,73	78,87	82,16	83,26	85,45	85,92	99,37	99,21	98,26	98,26	98,74	99,05	99,05
Mugshots	90,45	87,64	78,72	81,06	81,22	82	82,47	99,37	99,05	98,26	98,11	98,58	98,58	98,58

(e) (f)



	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5
Identities	95,62	93,58	84,82	87,95	89,98	91,24	91,55	99,84	99,84	99,68	99,68	99,68	99,68	99,68
Mugshots	95,62	92,49	84,66	86,54	88,26	88,58	88,89	99,84	99,37	99,53	99,53	99,05	99,37	99,37

(g) (h)

Figura 5.9: Misure di accuratezza percentuale su SCFace, Top-1 (a-b), Top-3 (c-d), Top-5 (e-f) e Top-10 (g-h) per VGG16 e ResNet50, considerando le migliori identità (blu) e le migliori foto segnaletiche (arancione). I migliori risultati nella verifica del volto si ottengono sull'immagine frontale, tranne in Top-1 con VGG16 (a), dove il sottoinsieme di foto segnaletiche "F-L1-R1" ha ottenuto la migliore accuratezza (il sottoinsieme che utilizza solo la foto frontale è il secondo migliore in questo test). L'asse delle ordinate è stato tagliato tra il 55% e il 100% per apprezzare meglio visivamente gli scostamenti percentuali

principale dello studio proposto: mentre le immagini delle foto segnaletiche sono ideali per analizzare il loro impatto sulla verifica di volti utilizzando diversi sottoinsiemi di immagini prese da diversi punti di vista, le immagini delle telecamere di sicurezza non sono pienamente rappresentative della realtà. Infatti, le immagini delle telecamere di sicurezza potrebbero includere volti in una prospettiva diversa, simile a una foto di profilo o addirittura con un'angolazione più elevata.

Un'altra limitazione del lavoro di ricerca proposto riguarda il numero di CNN confrontate, poiché abbiamo limitato il nostro studio a ResNet50 e VGG16, senza includere altre architetture di reti neurali.

5.4.4 Conclusioni

Per quanto ne sappiamo, questo è il primo studio che cerca di valutare l'impatto delle foto segnaletiche della polizia in un compito di verifica del volto, comprendendo se l'uso di immagini prese da più punti di vista possa aumentare le prestazioni delle tecniche di riconoscimento rispetto alle tradizionali immagini frontali e di profilo. A tal fine, abbiamo calcolato le metriche di accuratezza nella verifica di volti con le CNN sulle immagini del database SCFace, utilizzando come immagini di riferimento diversi sottoinsiemi di foto segnaletiche, scattate da diversi punti di vista.

Abbiamo ottenuto i migliori risultati di accuratezza utilizzando solo la foto segnaletica frontale: l'utilizzo di più immagini da più punti di vista peggiora i risultati. I risultati peggiori sono stati ottenuti con il sottoinsieme di foto segnaletiche composto dalla foto frontale e dal profilo destro del soggetto, come nella normale procedura di fotosegnalamento attuata dalle forze di polizia. L'utilizzo di un maggior numero di foto segnaletiche prese da diverse angolazioni migliora i risultati rispetto a questo scenario, ma senza alcun miglioramento rispetto all'utilizzo della sola foto frontale. Tuttavia, questi risultati controintuitivi sono spiegabili con il fatto che quasi tutte le immagini delle telecamere di sicurezza del database SCFace includono solo immagini frontali del volto.

Pertanto, i risultati ottenuti non possono essere considerati generali e conclusivi per valutare se una modifica della procedura di fotosegnalamento, per raccogliere più di due foto segnaletiche da diversi punti di vista, valga gli investimenti necessari in termini di acquisizione e archiviazione delle immagini. A tal fine, i lavori futuri includeranno ulteriori test. Inoltre, per ottenere risultati più generali sull'uso delle foto segnaletiche in un compito di verifica dei volti, è necessario un nuovo database di immagini, che includa le immagini dei volti nelle foto delle telecamere di sicurezza scattate da più punti di vista. Un tale database sarebbe più rappresentativo della realtà, dove un soggetto può essere

inquadrato in una telecamera di sicurezza da una prospettiva diversa da quella frontale.

5.5 Valutazione delle reti neurali profonde per il riconoscimento dei volti con diversi sottoinsiemi di foto segnaletiche

[*pubblicato*][16]

Il riconoscimento automatico dei volti è considerato una tecnologia matura per l'identificazione biometrica, in effetti dato che il Deep Learning e, nello specifico, le reti neurali convoluzionali (CNN) hanno dimostrato risultati eccezionali nella verifica e nell'identificazione dei volti [55], sono tecnologie oggi disponibili in applicazioni comuni come nella verifica dei passaporti [144, 188], nello sblocco degli smartphone [143, 189, 190] e nell'autenticazione biometrica in generale [191].

Tuttavia alterazioni di morphing, volontarie o casuali, possono mettere alla prova la robustezza delle reti neurali verso il riconoscimento automatico. A tal proposito, Franco et al. hanno condotto uno studio riguardo l'impatto dell'alterazione digitale delle immagini del volto, dimostrando che alcuni algoritmi sono efficaci verso le alterazioni limitate ma c'è necessità di metodi più robusti verso alterazioni più sofisticate [192].

Principalmente, ciò che viene chiesto a una rete neurale convoluzionale si può riassumere in due compiti, ovvero valutare se l'immagine di un volto appartiene a un'identità specifica in un insieme di soggetti noti (1:N) e nella verifica dei volti, ossia nel valutare se più immagini di un volto appartengono alla stessa persona [55].

Mentre il loro utilizzo negli AFIS esistenti e nei sistemi di supporto alle indagini criminali dimostra il livello di preparazione tecnologica delle CNN per il riconoscimento dei volti, la corrispondenza tra pose arbitrarie, ovvero il Pose-Invariant Face Recognition (PIFR), è considerata una sfida aperta [149, 148].

Richiamando il precedente lavoro descritto nella sezione 5.4, i test in esso effettuati si basano sul riconoscimento di poche identità all'interno di video catturati da telecamere di sicurezza a bassa risoluzione. Per ottenere risultati più generali sono invece necessari test con un maggior numero di identità, confrontando video di telecamere di sicurezza a bassa e ad alta risoluzione.

[16] Contardo, P., Rossini, N., Tomassini, S., Falcionelli, N., Dragoni, A. F., & Sernani, P. (2023, October). Evaluating Deep Neural Networks for Face Recognition with Different Subsets of Mugshots From the Photo-Signaling Procedure. In 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) (pp. 543-548). IEEE.

Per questi motivi, in questo lavoro estendiamo la nostra ricerca precedente aggiungendo nuovi soggetti al dataset FRMDB, per superare i limiti precedenti, ed effettuando un'ulteriore valutazione sperimentale. In particolare, il lavoro contribuisce allo stato dell'arte della valutazione dell'uso delle foto segnaletiche della polizia nel riconoscimento dei volti:

- aggiungendo 28 nuove identità all'FRMDB. Oltre a includere 28 foto segnaletiche, sono stati aggiunti 3 video di telecamere di sicurezza ad alta risoluzione, ripresi da 3 diversi punti di vista, invece dei 5 video di telecamere di sicurezza a bassa risoluzione, disponibili nella versione precedente dell'FRMDB.

- per confrontare l'impatto dell'uso di diversi sottoinsiemi di foto segnaletiche, riprese da diversi punti di vista, nel compito di riconoscimento facciale con i video di sicurezza.

In effetti, i video ad alta risoluzione consentono di effettuare ulteriori test rispetto alle nostre ricerche precedenti. In particolare, in questo lavoro, confrontiamo l'accuratezza del riconoscimento, utilizzando diversi sottoinsiemi di foto segnaletiche con il protocollo presentato in [193], sui nuovi soggetti e sui loro video ad alta risoluzione, con l'accuratezza sui soggetti e sui video originali del FRMDB e con l'accuratezza sull'intero dataset (composto da nuove e vecchie identità).

Nella valutazione sperimentale proposta, abbiamo confrontato le capacità di riconoscimento di due diverse reti CNN, ovvero VGG16 [174] e ResNet50 [54], pre addestrate rispettivamente sui dataset VGGFace [154] e VGGFace2 [53].

5.5.1 Rassegna della letteratura

Come evidenziato nell'introduzione, il riconoscimento dei volti è un'importante area di ricerca nella computer vision, con applicazioni nei sistemi di sicurezza, nell'autenticazione biometrica e nei social media. Analogamente ad altri compiti della computer vision, le reti neurali convoluzionali (CNN) sono emerse come un potente strumento per il riconoscimento dei volti, grazie alla loro capacità di apprendere caratteristiche di alto livello dai dati grezzi delle immagini. Per questo motivo, sono disponibili molti database di immagini di volti per addestrare e testare le architetture basate sul Deep Learning per il riconoscimento dei volti. Ad esempio, il database Labeled Face in the Wild (LFW) [157, 158] contiene 13.233 immagini a colori (250 $\times$ 250 pixel) di 5.749 soggetti, senza alcun vincolo in termini di posa, illuminazione, sfondo, ecc. Come tale, il database è destinato al riconoscimento dei volti senza vincoli. Tuttavia, non c'è un numero fisso di immagini per soggetto (1.680 soggetti hanno solo due o più immagini) e, essendo in natura, non ci sono punti di

vista sistematici da cui vengono prese le immagini come accade, invece, con le foto segnaletiche della polizia. Lo stesso vale per il database CASIA Webface [154], con 494.414 immagini di 10.575 soggetti. Inoltre, nel corso degli anni, si sono resi disponibili database su scala milionaria. Il Megaface Challenge Dataset [162, 163] contiene 4,7 milioni di immagini a colori di 672.057 soggetti unici. Come la LFW, tuttavia, non è disponibile una raccolta sistematica di foto da punti di vista definiti. Inoltre, il database è stato interrotto. VGGFace [154] e VGGFace2 [53], su cui si basa il pre-training delle CNN utilizzate in questo studio, sono database di volti su scala milionaria. Il VGGFace comprende 928.803 immagini di volti, di cui il 95% frontali e il 5% di profili, corrispondenti a 2.622 soggetti. Il VGGFace2 contiene invece 3,31 milioni di immagini a colori di 9.131 identità. Con un tale numero di immagini, VGGFace e VGGFace2 sono ideali per addestrare CNN e architetture basate sul Deep Learning per estrarre le caratteristiche del volto. Tuttavia, anche questi database non contengono immagini scattate sistematicamente a punti di vista fissi come nelle foto segnaletiche della polizia. Un database che include soggetti ripresi sistematicamente è il database SCFace [153] che abbiamo utilizzato nei test presentati alla conferenza MetroXRAINE del 2022 [194]. Include 9 immagini a punti di vista fissi per 130 soggetti, più 21 immagini per soggetto scattate dalle telecamere di sicurezza. Tuttavia, queste 21 immagini sono tutte frontali, invece di essere scattate da diversi punti di vista, come accadrebbe in natura. Per tutti questi motivi, nel nostro precedente lavoro [193], abbiamo presentato l'FRMDB, che comprende 28 foto segnaletiche scattate sistematicamente più 5 video di sicurezza a bassa risoluzione (da diversi punti di vista) di 39 soggetti. Tuttavia, come evidenziato nel documento FRMDB, il database necessita di un maggior numero di soggetti e di video ad alta risoluzione. Per questo motivo, abbiamo ampliato il database con 28 nuovi soggetti e video ad alta risoluzione e abbiamo effettuato i test aggiuntivi presentati in questo articolo.

Per quanto riguarda la valutazione sperimentale discussa in questo lavoro, abbiamo utilizzato VGG16 e ResNet50, due note CNN che hanno dimostrato accuratezza nella verifica dei volti in natura. In effetti, diversi studi hanno dimostrato l'efficacia delle CNN per il riconoscimento dei volti. A questo proposito, DeepFace proposta da Taigman et al. [50] ha ottenuto un'accuratezza del 97,35% sul database LFW, con i suoi 8 strati applicati ai tre canali di immagini a colori da 152 × 152 pixel. L'architettura FaceNet proposta da Schroff et al. [52], una CNN a 22 strati, ha ottenuto un'accuratezza del 99,63% su LFW. Analogamente, il modello DeepID3 sviluppato da Sun et al. [195] ha ottenuto un'accuratezza del 99,53% sullo stesso database. Nel complesso, questi studi hanno dimostrato il potenziale delle CNN per il riconoscimento dei volti e hanno evidenziato l'importanza della ricerca in questo settore per migliorare l'accuratezza, l'efficienza e la robustezza di questi sistemi. Tuttavia, le CNN

pre-addestrate e, nello specifico, VGG16 e ResNet50, hanno mostrato capacità straordinarie con l'apprendimento per trasferimento, ossia hanno mostrato accuratezza su nuovi set di dati quando sono state addestrate su un diverso insieme di immagini. In particolare, Parkhi et al. [154] hanno dimostrato che i 16 strati di VGG16, addestrati sul database VGGFace, erano in grado di ottenere un'accuratezza del 98,95% sulle immagini della LFW. Cao et al [53]. hanno dimostrato che i 50 strati di ResNet50, basati sulle connessioni residue, hanno ottenuto un errore di identificazione del 3,9% sulle immagini del database VGGFace2. You et al. [176] hanno dimostrato che ResNet50 (tra le altre CNN testate), pre-addestrata sul set di dati CASIA-Webface, ha ottenuto il 98,58% sulle immagini del database LFW. Per questi motivi, e per confrontare la valutazione sperimentale con la nostra ricerca precedente, esaminiamo i risultati ottenuti da VGG16 e ResNet50, considerando le capacità esibite da queste CNN nell'apprendimento del trasferimento e nel riconoscimento dei volti. Si tratta di un ulteriore passo avanti nella valutazione delle capacità di riconoscimento quando si utilizzano vari sottoinsiemi di foto segnaletiche come immagini di riferimento.

5.5.2 Materiali e metodi

Per eseguire test con diversi sottoinsiemi di foto segnaletiche e capire se il riconoscimento del volto con le CNN trae vantaggio dall'uso di più immagini rispetto alle tradizionali foto frontali e di profilo scattate nella procedura di fotosegnalamento delle agenzie di polizia, abbiamo utilizzato VGG16 e ResNet50 per estrarre le caratteristiche del volto. Queste CNN generano un embedding per ogni volto, che rappresenta un vettore di caratteristiche che descrive l'immagine del volto. L'architettura di VGG16 è quella descritta in Parkhi et al. [154], mentre per ResNet50 abbiamo seguito il modello descritto in Cao et al [53]. Questi modelli elaborano un'immagine del volto 224 x 224 come input e l'embedding viene calcolato con il Global Average Pooling applicato all'output del blocco convoluzionale finale. L'embedding estratto da VGG16 è un vettore di caratteristiche a 512 elementi; l'embedding calcolato da ResNe50 è invece composto da 2048 valori. Nella nostra valutazione sperimentale, abbiamo utilizzato la versione Keras dei due modelli[17]. Come in Cao et al., abbiamo normalizzato in L2 gli embeddings ottenuti da entrambe le reti. Il processo di addestramento di queste reti si è attenuto alle metodologie presentate da Parkhi et al. per VGG16[18] e da Cao et al. per ResNet50[19], utilizzando i dataset

[17] Le conversioni Keras di entrambe le reti sono pubblicamente disponibili su https://github.com/rcmalli/keras-vggface

[18] Invece di riallenare la rete da zero, abbiamo utilizzato i pesi della rete originale disponibili pubblicamente su https://www.robots.ox.ac.uk/~vgg/software/vgg_face/

[19] Invece di riallenare la rete da zero, abbiamo utilizzato i pesi della rete originale disponibili pubblicamente su https://github.com/ox-vgg/vgg_face2

VGGFace e VGGFace2.

I test della valutazione sperimentale sono stati effettuati sul FRMDB [193], che originariamente comprendeva 39 soggetti, con 28 foto segnaletiche e 5 video di sicurezza, sistematicamente ripresi da diversi punti di vista, per ogni soggetto. In questo lavoro, estendiamo la ricerca precedente aggiungendo nuovi soggetti all'FRMDB, per superare i limiti precedenti. In particolare, oltre ai 39 soggetti già disponibili, aggiungiamo 28 nuove identità. Per ogni nuova identità:

- È disponibile un totale di 28 foto segnaletiche, ovvero 28 foto a colori scattate da diversi punti di vista con il soggetto in posa durante l'acquisizione.
- Un totale di 3 video di telecamere di sicurezza, ripresi da 3 punti di vista. A differenza delle identità raccolte in precedenza, che avevano 5 video di telecamere di sicurezza a bassa risoluzione, i nuovi video sono ad alta risoluzione (1920 × 1080 pixel).

Come per i 39 soggetti originali, le foto segnaletiche dei 28 nuovi soggetti hanno una risoluzione di 972 × 544 pixel (JPEG). Le foto segnaletiche sono state scattate da 7 angolazioni sul piano orizzontale e da 4 angolazioni sul piano verticale: sul piano orizzontale, la gamma va da -135° (-90° è il profilo sinistro) a +135° (90° è il profilo destro), con un passo di 45° (con 0° che si trova davanti al soggetto); sul piano verticale, la gamma va da -30° (sotto gli occhi del soggetto) a +60° (sopra gli occhi, con 0° che è l'immagine presa dal piano degli occhi del soggetto) con un passo di 30°. I video di sicurezza ad alta risoluzione consentono di effettuare ulteriori test rispetto alle nostre ricerche precedenti. In questo lavoro, infatti, confrontiamo l'accuratezza del riconoscimento sui nuovi soggetti e sui loro video ad alta risoluzione, con l'accuratezza sui soggetti e sui video originali del FRMDB e con l'accuratezza sull'intero dataset (composto da nuove e vecchie identità).

Le due CNN, VGG16 e ResNet50, sono utilizzate per il riconoscimento dei soggetti nei video di sicurezza utilizzando diversi sottoinsiemi di foto segnaletiche. Tali sottoinsiemi sono definiti nella Tabella 5.2.

Ogni foto segnaletica di un sottoinsieme è descritta come una coppia (h, v): h è l'angolo sul piano orizzontale e v è l'angolo sul piano verticale. In particolare, i sottoinsiemi di foto segnaletiche utilizzati nei test sono:

- "Test F", che include solo l'immagine frontale (0°, 0°).
- "Test F-L1-R1", comprendente l'immagine frontale, l'angolo sinistro più vicino al frontale (−45°, 0°) e l'angolo destro più vicino all'immagine frontale (45°, 0°).

- "Test 1", che comprende l'immagine frontale e l'immagine di profilo destra (90°, 0°). Questo sottoinsieme comprende semplicemente le foto segnaletiche tradizionalmente raccolte dalle agenzie di polizia durante la procedura di fotosegnalamento.
- "Test 2", che include l'immagine frontale, l'immagine del profilo destro e del profilo sinistro, ovvero (−90°, 0°).
- "Test 3", comprendente le immagini del "Test 2" e quelle più vicine all'immagine frontale a partire dai profili destro e sinistro, ossia (45°, 0°) e (−45°, 0°).
- "Test 4", che include le immagini del "Test 3" e quelle a (135°, 0°) e (135°, 0°). In questo modo, il sottoinsieme "Test 4" contiene tutte le immagini a 0° sul piano verticale.
- "Test 5", che comprende le immagini del "Test 4" e tutte le immagini con un angolo di 30° sul piano verticale.
- "Test 6", comprendente tutte le 28 foto segnaletiche disponibili per ogni soggetto nel database.

Per quanto riguarda le immagini dei volti da riconoscere nella valutazione sperimentale, le abbiamo ritagliate manualmente dai video di sicurezza. Ciò significa che ci sono 3 immagini di volti da riconoscere per ognuno dei nuovi 28 soggetti (per un totale di 84 immagini), mentre ci sono 5 immagini di volti da riconoscere per ognuno dei vecchi 39 soggetti del FRMDB (per un totale di 195 immagini). Abbiamo valutato le prestazioni di VGG16 e ResNet50 nel riconoscimento di individui ripresi nelle immagini delle telecamere di sicurezza utilizzando le foto segnaletiche dei sottoinsiemi descritti. Abbiamo raccolto i risultati in base all'identificazione del soggetto corretto tra le foto segnaletiche più simili (top-1, top-3, top-5 o top-10) e le identità più vicine (top-1, top-3, top-5 o top-10). Dato che le CNN calcolano gli embeddings dei volti per rappresentare ogni immagine del volto, la somiglianza tra due embeddings di volti è stata valutata utilizzando la distanza euclidea. Di conseguenza, abbiamo calcolato la seguente accuratezza:

- Il numero di volti di telecamere di sicurezza per i quali il soggetto corretto è stato identificato nella top-1, top-3, top-5 e top-10 delle foto segnaletiche più simili rispetto al numero totale di volti di telecamere di sicurezza (per la classifica delle foto segnaletiche più simili).
- Il numero di volti di telecamere di sicurezza per i quali il soggetto corretto è stato identificato nella top-1, top-3, top-5 e top-10 delle identità più simili rispetto al numero totale di volti di telecamere di sicurezza (per la classifica delle identità più simili).

Tabella 5.6: Un esempio di calcolo dell'accuratezza nel nostro studio. Nell'ipotesi che il soggetto corretto sia "013", con il seguente elenco di foto segnaletiche più vicine, l'immagine giusta è nella top-5 delle foto segnaletiche più simili e nella top-3 delle identità più vicine (la terza identità riconosciuta è "013", dopo "005" e "021").

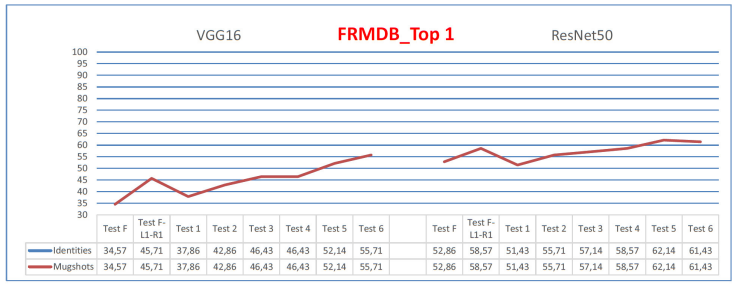
1. 005 (0°, 0°)	**6.** 032 (0°, 0°)
2. 021 (0°, 0°)	**7.** 013 (45°, 0°)
3. 021 (45°, 30°)	**8.** 013 (45°, 30°)
4. 005 (0°, 30°)	**9.** 015 (0°, 0°)
5. 013 (0°, 0°)	...

Naturalmente, la classifica top-1 basata sulle identità e la classifica top-1 basata sulle foto segnaletiche sono identiche. Allo stesso modo, quando si considera il sottoinsieme composto da una sola foto segnaletica, cioè la foto frontale, l'accuratezza è la stessa in tutte le classifiche, sia che si considerino le identità che le foto segnaletiche.

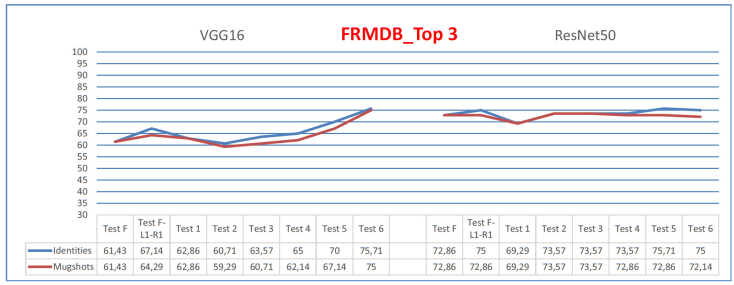
In realtà, considerare sia le identità che le foto segnaletiche può essere rilevante, in quanto può portare a risultati diversi. A titolo di esempio, la Tabella 5.6 elenca una possibile classifica delle dieci foto segnaletiche più vicine all'immagine del volto di un soggetto ripresa da una telecamera di sicurezza. Nell'ipotesi che l'identità corretta sia "013", tale soggetto non è nella top-1, dato che la foto segnaletica più simile è un'immagine dell'identità "005"; non è nella top-3 delle foto segnaletiche più simili, dato che le due foto segnaletiche successive appartengono all'identità "021". Il soggetto corretto rientra invece nella top-5 delle foto segnaletiche più simili, poiché la prima immagine corretta è esattamente la quinta. Tuttavia, il soggetto corretto si trova nella top-3 delle identità più vicine, dato che "013" è la terza riconosciuta, dopo "005" e "021".

5.5.3 Valutazione sperimentale e discussione

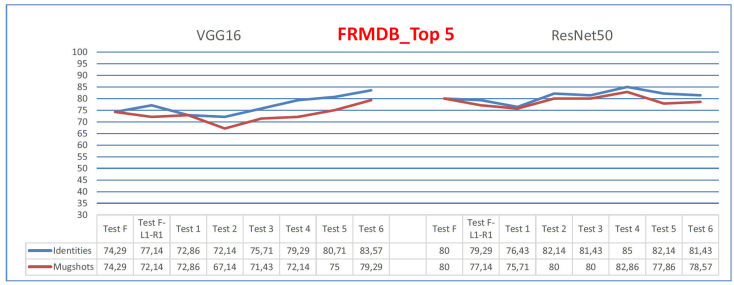
Per la valutazione sperimentale, abbiamo eseguito VGG16 e ResNet50 sui 28 nuovi soggetti, calcolando gli embeddings dei volti, estraendo un fotogramma per ciascuno dei 3 video di sicurezza ad alta risoluzione disponibili per ogni soggetto, nonché sull'intero dataset FRMDB. Per i vecchi soggetti, i volti da riconoscere erano 5, poiché per ogni soggetto sono disponibili 5 video di sicurezza a bassa risoluzione. Per eseguire il riconoscimento dei volti e registrare l'accuratezza, abbiamo confrontato ciascun embedding con quelli calcolati sui sottoinsiemi di foto segnaletiche presentati nella Sezione *Materiali e metodi*. A tal fine, la Figura 5.10 include i risultati dell'esecuzione del riconoscimento facciale solo sui 28 nuovi soggetti aggiunti per questa ricerca. Ciò significa che le immagini delle telecamere di sicurezza da riconoscere e i sottoinsiemi di foto



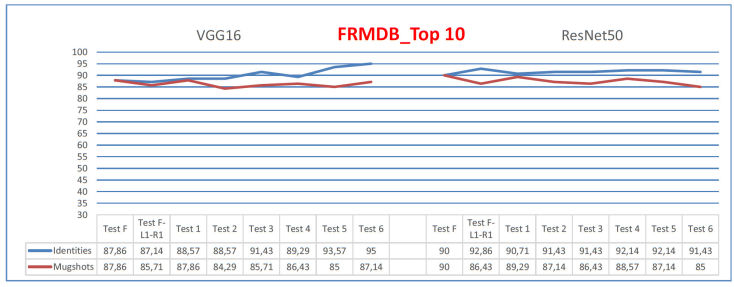
(a) (b)



(c) (d)

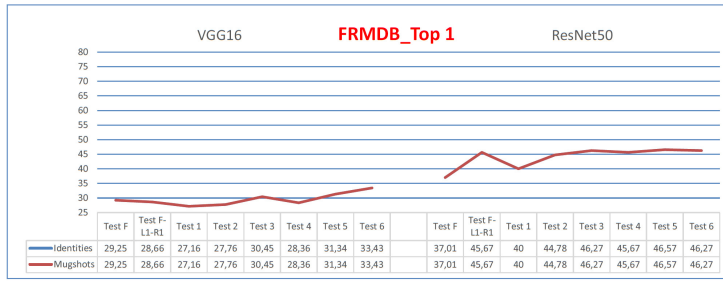


(e) (f)

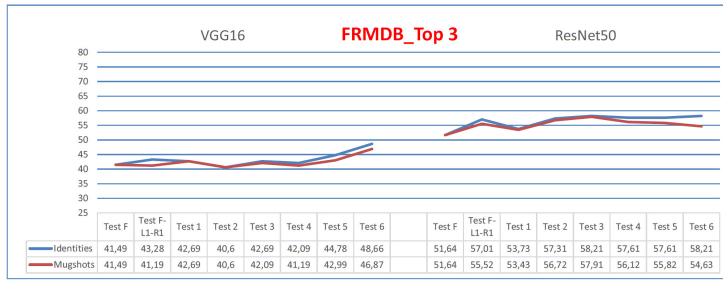


(g) (h)

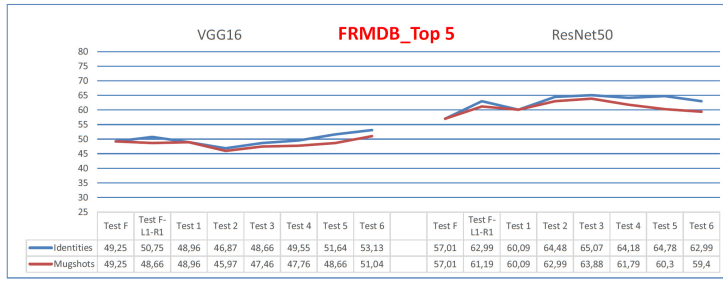
Figura 5.10: Misure di accuratezza percentuale dei modelli VGG16 e ResNet50 sui 28 nuovi soggetti dell'FRMDB: (a-b)top-1, (c-d)top-3, (e-f)top-5 e (g-h)top-10. Queste classifiche considerano sia le migliori identità (blu) sia le migliori foto segnaletiche (arancione). L'asse delle ordinate è stato tagliato tra il 30% e il 100% per apprezzare meglio visivamente gli scostamenti percentuali.



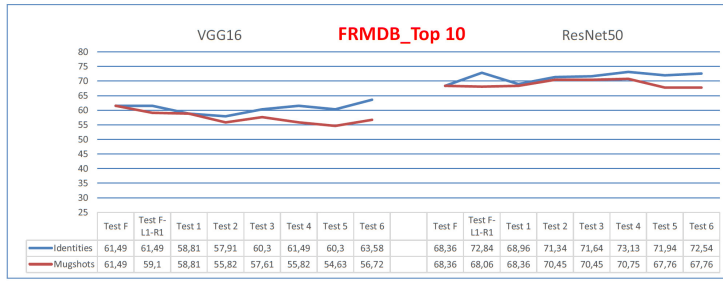
(a) (b)



(c) (d)



(e) (f)



(g) (h)

Figura 5.11: Misure dell'accuratezza percentuale dei modelli VGG16 e ResNet50 sull'intero set di dati FRMDB: (a-b)top-1, (c-d)top-3, (e-f) top-5 e (g-h)top-10. Queste classifiche considerano sia le migliori identità (blu) sia le migliori foto segnaletiche (arancione). L'asse delle ordinate è stato tagliato tra il 25% e l'80% per apprezzare meglio visivamente gli scostamenti percentuali.

segnaletiche da diversi punti di vista includevano solo questi soggetti. A differenza della nostra precedente ricerca [193], basata sui soli 39 soggetti originali, si nota una chiara tendenza a ottenere i migliori risultati con i sottoinsiemi composti da più foto segnaletiche, ovvero "Test 5" e "Test 6". Anche se, in generale, VGG16 ha ottenuto un'accuratezza inferiore rispetto a ResNet50, in tutte le classifiche (Figure 5.10 (a), (c), (e) e (g)) il sottoinsieme di foto segnaletiche "Test 6", cioè quello con tutte le 28 foto segnaletiche per ogni soggetto, ha ottenuto l'accuratezza più alta, sia per la classifica delle identità che per quella delle foto segnaletiche. Per quanto riguarda VGG16, il sottoinsieme "Test 1", composto dalla foto frontale e dal profilo destro, ovvero le foto segnaletiche raccolte al momento, ha ottenuto il peggior risultato nella top-1 e nella top-5, e il terzo peggior risultato nella top-3 e nella top-10. ResNet50 (Figure 5.10 (b), (d), (f) e (h)) ha mostrato una tendenza simile. Nella top-3 e nella top-10, il sottoinsieme con l'accuratezza più elevata è stato quello con tutte le foto segnaletiche (75,71% di accuratezza dell'identità nella top-3, 95% di accuratezza dell'identità nella top-10); nella top-1 i risultati migliori sono stati quelli del sottoinsieme "Test 5" (62,14% di accuratezza dell'identità); nella top-5, "Test 4" ha ottenuto la migliore accuratezza (85% nella classifica dell'identità), mentre "Test 5" è stato il secondo migliore. Analogamente a quanto accade con VGG16, ResNet50 ha ottenuto la peggiore accuratezza con il sottoinsieme composto dalla foto frontale e dal profilo destro ("Test 1") nelle top-1, top-3 e top-5, mentre nella top-10 non c'è stata una chiara distinzione tra i sottoinsiemi, poiché l'accuratezza è stata compresa tra il 90 e il 92,14% per tutti i sottoinsiemi considerando le identità più vicine, e tra l'86,43 e l'88,57% considerando le foto segnaletiche più vicine.

La Figura 5.11 mostra le metriche di accuratezza quando si considera l'intero FRMDB, invece di testare con i nuovi 28 soggetti con video ad alta risoluzione. Considerare 59 soggetti, di cui 39 con video di telecamere di sicurezza a bassa risoluzione, diminuisce l'accuratezza, sia per VGG16 che per ResNet50, in tutte le classifiche, nonostante si considerino le foto segnaletiche o le identità più vicine. Ciò è dovuto al maggior numero di identità da riconoscere e alla prevalenza di fotogrammi a bassa risoluzione dei video di sicurezza (195) rispetto a quelli ad alta risoluzione (84) da riconoscere. Anche con l'intero database, con VGG16 (Figure 5.11 (a), (c), (e), e (g)) l'accuratezza trae vantaggio dall'utilizzo di tutte le foto segnaletiche: in tutte le classifiche, il sottoinsieme migliore è stato "Test 6". Tuttavia, in tutte le classifiche, "Test 5" e "Test 6" hanno ottenuto risultati simili. Ad esempio, nella top-5, "Test 5" ha ottenuto il 51,64% di accuratezza in termini di identità più vicine e il 48,66% di accuratezza in termini di foto segnaletiche più vicine, mentre "Test 6" ha ottenuto il 53,13% e il 51,04%. Tuttavia, il "Test 6" contiene 28 foto segnaletiche, mentre il "Test 5" ne contiene solo quattordici, e quindi un maggiore spazio di archiviazione, per

ospitare il doppio delle foto segnaletiche, potrebbe non valere l'investimento necessario, visto l'esiguo aumento di accuratezza. Diversamente, con ResNet50 (Figure 5.11 (b), (d), (f), e (h)) non c'è stata una chiara tendenza in termini di sottoinsiemi di foto segnaletiche, anche se il "Test F", che include solo la foto frontale, e il "Test 1", che include le immagini della fotosegnalamento tradizionale, hanno ottenuto l'accuratezza più bassa in tutte le classifiche. Tutti gli altri sottoinsiemi hanno ottenuto valori di accuratezza molto simili: ad esempio, nella top-3, "Test F-L1-R1" ha ottenuto l'accuratezza minima dell'identità tra gli altri sottoinsiemi, ovvero il 57,01%; "Test 6", con tutte le foto segnaletiche, ha ottenuto il massimo, ovvero il 58,21%.

In generale, l'ampliamento dell'FRMDB e l'aggiunta dei nuovi test, rispetto al nostro lavoro precedente, ha evidenziato che l'aumento del numero di immagini di riferimento, con l'aggiunta di più punti di vista rispetto alla foto frontale e al profilo destro solitamente raccolti durante la fotosegnalamento, migliora l'accuratezza del riconoscimento del volto eseguito sui fotogrammi dei video delle telecamere di sicurezza. Abbiamo ottenuto la migliore accuratezza utilizzando tutte le 28 foto segnaletiche per ogni soggetto, e valori simili utilizzando le 14 foto segnaletiche da -135° a 135° sul piano orizzontale, e da 0° a 30° sul piano verticale.

Infine, è opportuno sottolineare che i risultati della nostra valutazione sperimentale presentano alcune limitazioni. Per quanto riguarda il database utilizzato, una limitazione fondamentale potrebbe essere il numero relativamente basso di soggetti distinti, che ammonta a 59. Anche se questo numero può sembrare basso, si tratta di un numero di soggetti che non è stato considerato. Anche se questo numero può sembrare basso, è importante notare che il database non è stato progettato per l'apprendimento dei tratti del volto. La letteratura offre già database specificamente concepiti per questo scopo, con milioni di immagini disponibili. L'intento del database utilizzato è invece quello di fungere da benchmark per la valutazione delle tecniche di riconoscimento facciale nel contesto dell'identificazione di soggetti all'interno di fotogrammi di telecamere di sicurezza, utilizzando come immagini di riferimento le foto segnaletiche. Pertanto, l'FRMDB è più adatto a scopi di prova che di apprendimento. Inoltre, i risultati presentati si basano su CNN pre-addestrate, che si basano su ricerche esistenti nel campo del riconoscimento facciale, come discusso nella sezione *Rassegna della letteratura*. Tuttavia, è necessario condurre un esame completo di modelli alternativi e confrontare diversi set di dati per ottenere una visione più ampia dell'influenza dell'utilizzo di diversi sottoinsiemi di foto segnaletiche sul riconoscimento dei volti nei fotogrammi dei video di sicurezza.

5.5.4 Conclusioni

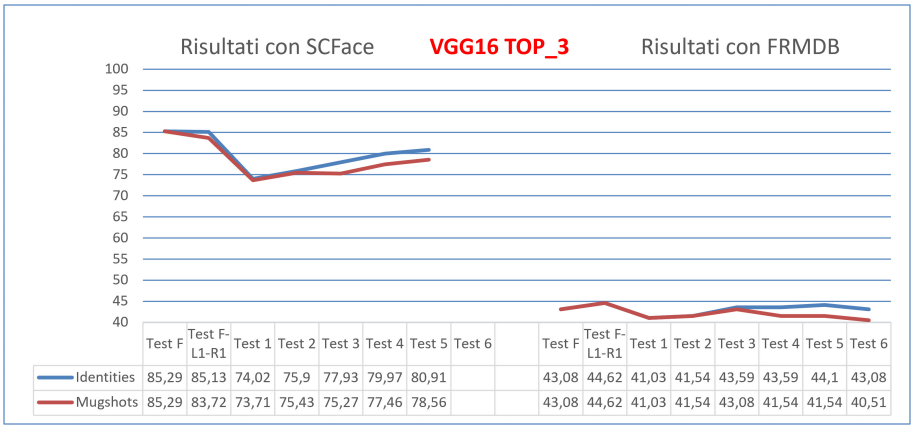
In questo lavoro abbiamo presentato la valutazione dell'impatto che le foto segnaletiche della polizia prese sistematicamente da diversi punti di vista hanno su un compito di riconoscimento facciale su fotogrammi di video di telecamere di sicurezza. In particolare, abbiamo condotto una valutazione sperimentale utilizzando VGG16 e ResNet50 per estrarre embeddings di volti dal FRMDB, ampliato con 28 soggetti e 3 video ad alta risoluzione per soggetto. Abbiamo confrontato i risultati con diversi sottoinsiemi di foto segnaletiche, prese da diversi punti di vista, come immagini di riferimento, cercando di capire la variazione nell'accuratezza del riconoscimento.

I migliori risultati di accuratezza sono stati ottenuti utilizzando il sottoinsieme contenente tutte le foto segnaletiche e il sottoinsieme contenente la metà delle foto segnaletiche. I risultati peggiori sono stati invece ottenuti utilizzando una combinazione di foto frontale e profilo destro, comunemente utilizzata nelle procedure di fotosegnalamento della polizia, e con la sola foto frontale. Pertanto, quando sono state aggiunte più foto segnaletiche da diverse angolazioni, i risultati sono migliorati rispetto allo scenario tradizionale di fotosegnalamento, suggerendo che l'aumento del numero di foto segnaletiche è utile per aumentare l'accuratezza del riconoscimento facciale.

Sono necessarie ulteriori indagini, con test aggiuntivi. In particolare, sono necessari test orientati a capire quale potrebbe essere il numero corretto di foto segnaletiche, per trovare un compromesso tra l'accuratezza del riconoscimento e i costi dello spazio di archiviazione aggiuntivo necessario e la modifica delle procedure di fotosegnalamento. Inoltre, per ottenere risultati più ampi, è necessario valutare tecniche aggiuntive come il Pose-Invariant Face Recognition (PIFR) e la stima della posa. Questa valutazione mira a selezionare le foto segnaletiche che si allineano maggiormente con la posa dei soggetti nei fotogrammi delle telecamere di sicurezza, prima di eseguire il confronto.

5.6 Stima sull'utilità di pose aggiuntive rispetto al fotosegnalamento canonico

Alla luce delle ricerche sopra descritte, il fine ultimo dell'attività di ricerca in questo ambito è quello di trovare un protocollo innovativo per il fotosegnalamento, a supporto delle Forze dell'Ordine, per migliorare il riconoscimento automatico di persone sospettate di aver commesso degli illeciti giuridicamente perseguibili. Una possibile via di soluzione, potrebbe concretizzarsi nel trovare risposta al problema "QFQ" indicato nel capitolo n.5 Attraverso il sistema di acquisizione MCMPrototype, indicato nella sezione 5.1 e l'impianto di simulazione della videosorveglianza multi-prospettiva, abbiamo prodotto l'FRMDB



	Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6		Test F	Test F-L1-R1	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Identities	85,29	85,13	74,02	75,9	77,93	79,97	80,91			43,08	44,62	41,03	41,54	43,59	43,59	44,1	43,08
Mugshots	85,29	83,72	73,71	75,43	75,27	77,46	78,56			43,08	44,62	41,03	41,54	43,08	41,54	41,54	40,51

Figura 5.12: Confronto sull'accuratezza percentuale della rete VGG16 in top.3, testata sui dataset SCFace e FRMDB (l'asse delle ordinate è stato tagliato tra il 40% e 100% per apprezzare meglio visivamente gli scostamenti percentuali).

5.3. I primi esperimenti effettuati sulle "fotosegnaletiche" dei 39 soggetti, tramite due reti neurali convoluzionali (ResNet50 e VGG16 pre-addestrate su set di dati VGGFace per l'estrazione delle caratteristiche facciali), hanno evidenziato che l'utilizzo di fotosegnaletiche con più immagini è migliore rispetto all'utilizzo di una sola immagine o delle sole due immagini canoniche, e quindi l'FRMDB si è rivelato adatto a confrontare le varie tecniche di riconoscimento facciale[194], ma appare ovviamente limitato a causa del ridotto numero di soggetti e la bassa qualità dei video. In figura 5.12 sono messi a confronto i risultati di figura 5.9(c) con quelli di figura 5.8(c) riguardante l'accuratezza della rete VGG16 in top 3 rispettivamente sul SCFace e sul FRMDB, scegliendo arbitrariamente uno solo dei risultati ottenuti nei test effettuati sui dataset in quanto la tendenza degli altri esperimenti si ricalca approssimativamente su questi. Dalla lettura dei valori si possono trarre alcune prime considerazioni generali:

- L'aumento di immagini con diversi profili del volto della fotosegnaletica, rispetto al fotosegnalamento canonico individuato nel Test 1, apporta un leggero miglioramento su SCFace e rimane pressoché stabile sul FRMDB;
- L'aumento di immagini di fotosegnaletiche rispetto al Test 1, non introduce rumore;
- Il Test 1, è quello che fa registrare l'accuratezza peggiore in entrambi i dataset;
- L'accuratezza sulle Identities rispetto alle Mugshots è sempre maggiore;

- L'accuratezza sul dataset SCFace è quasi doppia di quella sul dataset FRMDB;
- L'accuratezza migliore sul dataset SCFace, si ottiene nel riconoscere le sole immagini frontali (Test F) o con l'aggiunta dei profili ruotati di soli 22,5° (Test F-L1-R1);
- L'accuratezza dei test sull'FRMDB è pressoché costante in tutti i test.

Da una prima osservazione si potrebbe erroneamente concludere che le reti neurali utilizzate reagiscono positivamente sul database SCFace all'aumentare delle fotosegnaletiche, rimanendo invece indifferenti sul database FRMDB. Ma se effettuiamo una disanima sulla configurazione dei database e sulle modalità dei test, si può fornire un'interpretazione ragionevole. Di fatti, riguardo alla videosorveglianza dei due dataset, l'SCFace si compone di immagini per lo più frontali estratte da video a diverse distanze 5.13, confrontate con 9 fotosegnaletiche in posa da -90° a +90° con passo di sfasamento di 22,5°, mostrate in figura 5.14. Per la selezione delle immagini della videosorveglianza dell'FRMDB, invece, non è stato utilizzato nessun criterio specifico, ma sono state campionate le immagini dei vari profili del volto casualmente, durante la riproduzione del video. Di conseguenza, il confronto sul dataset SCFace è avvenuto tra pose simili, invece il confronto sul dataset FRMDB è avvenuto tra pose casuali.

Al fine di avvicinare le metodologie degli esperimenti ai casi realmente trattati dalle Forze dell'Ordine, abbiamo ripetuto i test ma attraverso una simulazione dell'attività dell'investigatore forense. Più in dettaglio, seguendo le indicazioni dell'ENFSI indicate nel manuale di buone pratiche per il confronto delle immagini facciali [129], trattato nella sezione 5.1.2, abbiamo utilizzato una tecnica di selezione delle immagini non casuale, facendo scorrere i vari video della sorveglianza e selezionando manualmente i fotogrammi che mostravano le pose che approssimativamente si avvicinano ai profili presenti nel dataset del fotosegnalamento. Proprio come farebbe un investigatore durante la fase di indagini preliminari, che deve visionare le riprese di una videosorveglianza ed operare una comparazione fisionomica, siamo andati a selezionare i fotogrammi che abbiamo ritenuto più utili al confronto con le fotosegnaletiche, come se dovessimo individuare un sospettato. In post produzione, la selezione dei fotogrammi è stata ritagliata intorno al capo della persona ripresa, un esempio di campionamento è riportato in figura 5.15.

Quindi abbiamo ripetuto i test sull'FRMDB, mettendo a confronto le capacità di riconoscimento automatico di tre reti neurali, ovvero una VGG16, una Resnet50 e una Senet, preaddestrate su VGGFace2 e testate sui nuovi dati. Riprendendo i risultati dei test effettuati sul dataset SCFace, per la VGG16 e ResNet50, li abbiamo messi a confronto con i nuovi risultati dei test sull'FRMDB, che a differenza dei precedenti sono cambiati totalmente.

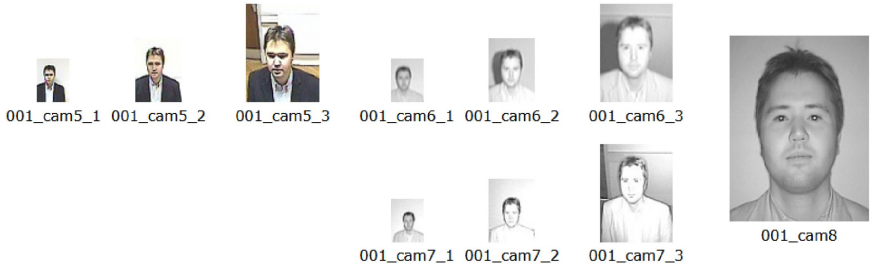


Figura 5.13: Esempio di immagini estratte dalla videosorveglianza del database SCFace.

Figura 5.14: Esempio di fotosegnaletiche del database SCFace.

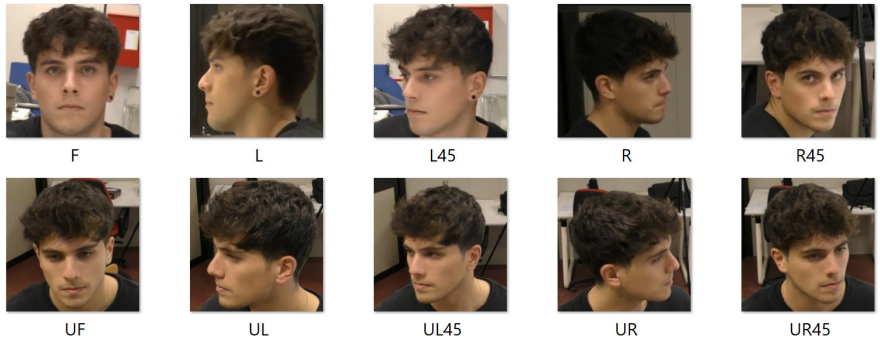


Figura 5.15: Esempio di campionamento delle immagini della videosorveglianza in HD dell'FRMDB, ritagliate intorno al capo della persona ripresa.

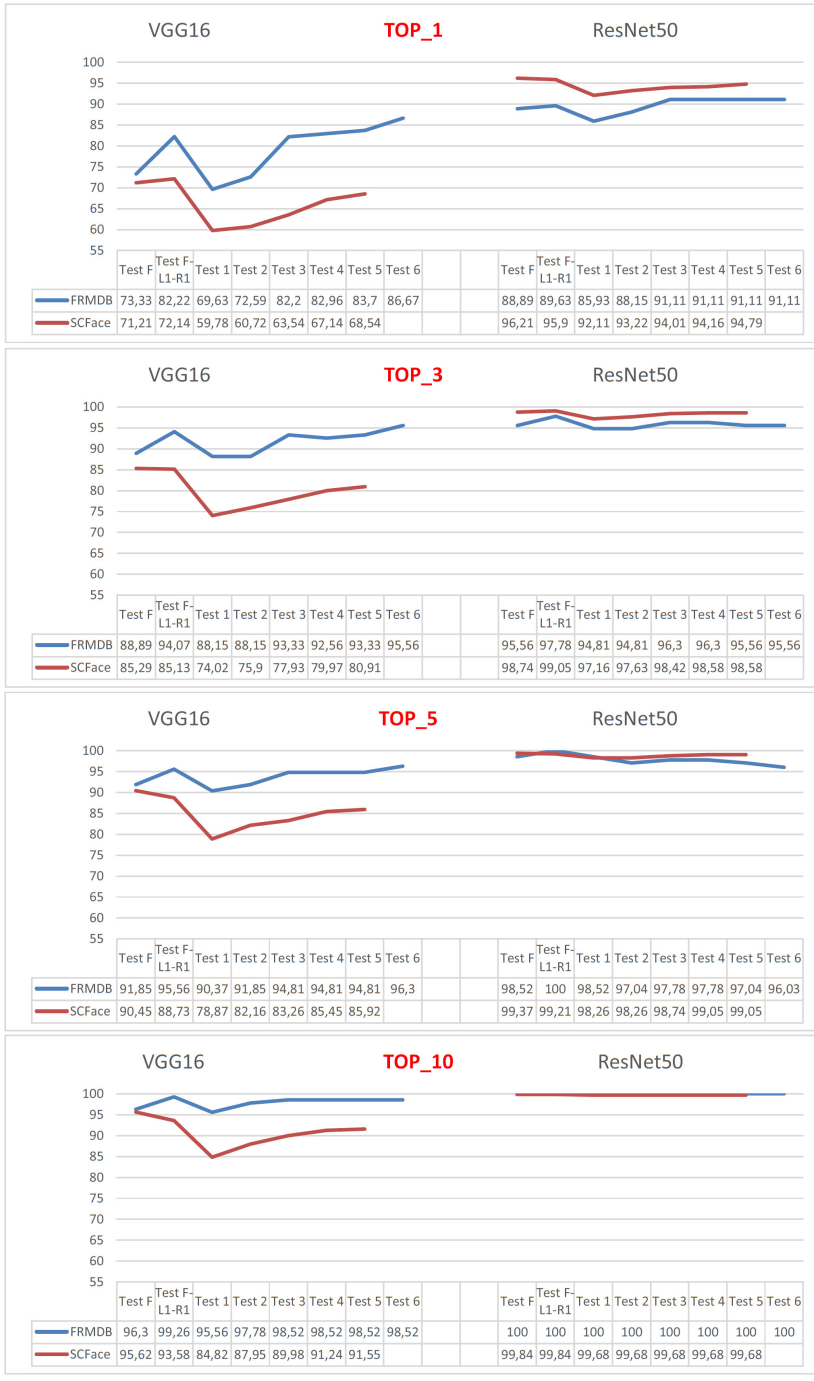


Figura 5.16: Confronto dei valori di accuratezza percentuale del riconoscimento facciale delle reti neurali VGG16 e ResNet50, riguardanti le identities, effettuati sul database SCFace e FRMDB.

Osservando l'andamento delle accuratezze per le due reti neurali riportate sui grafici di figura 5.16, riguardanti i risultati sulle *identities* in tutte le classifiche Top_1, Top_3, Top_5 e Top_{10}, si possono dedurre alcune importanti considerazioni:

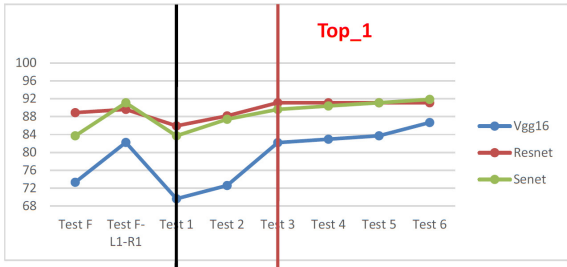
- in tutte le classifiche, la rete VGG16 restituisce valori di accuratezza nettamente maggiori sul database FRMDB;
- in tutte le classifiche e per ogni database, la rete ResNet50 restituisce valori di accuratezza maggiori della rete VGG16;
- in tutte le classifiche, le accuratezze della rete ResNet50 sono molto vicine tra i due dataset e diminuiscono la differenza reciproca all'aumentare della classifica, fino a quasi coincidere in Top_{10};
- in tutte le classifiche, la differenza media tra le accuratezze delle due reti neurali sul FRMDB, è minore della differenza tra le accuratezze sul SCFace.

Tale confronto, rafforza ancor di più la validità dell'FRMDB come database per il riconoscimento automatico dei volti e conferma l'adeguatezza della rete ResNet50 nel riconoscimento automatico dei volti in queste condizioni operative. Nella figura 5.17, invece, sono riportati i valori di accuratezza delle tre reti neurali Vgg16, ResNet50 e Senet, sul dataset FRMBD, in tutte le classifiche, dove sono stati selezionati i fotogrammi della videosorveglianza. Da una disamina dei risultati di figura 5.17, si possono dedurre ulteriori importanti considerazioni:

- per ogni test e in ogni classifica, la rete ResNet50 si è rivelata mediamente più accurata delle altre;
- per ogni test e in ogni classifica, la rete VGG16 è stata meno accurata delle altre;
- per ogni test e in ogni classifica, le reti ResNet50 e Senet hanno prodotto accuratezze mediamente confrontabili;
- i $TestF - L1 - R1, 3, 4, 5$ e 6 fanno registrare i valori di accuratezza più elevati per tutte e tre le reti neurali in ogni classifica;
- il $Test1$ è quello che fa registrare l'accuratezza peggiore per tutte e tre le reti neurali in ogni classifica;
- a partire dal $Test3$ in poi, l'aumento di fotosegnaletiche non apporta un aumento di accuratezza considerevole.

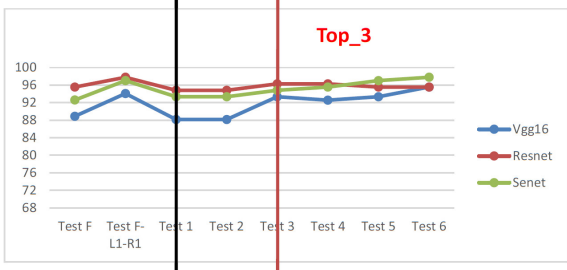
top1

prove	Vgg16	Resnet	Senet
Test F	73,33	88,89	83,7
Test F-L1-R1	82,22	89,63	91,11
Test 1	69,63	85,93	83,7
Test 2	72,59	88,15	87,41
Test 3	82,2	91,11	89,63
Test 4	82,96	91,11	90,37
Test 5	83,7	91,11	91,11
Test 6	86,67	91,11	91,85



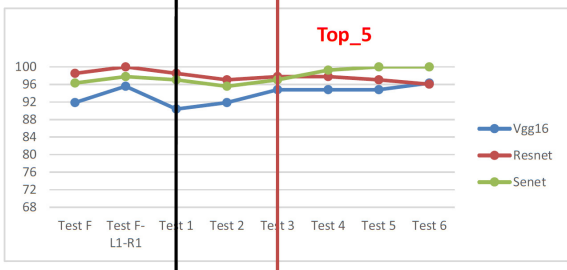
top3

prove	Vgg16	Resnet	Senet
Test F	88,89	95,56	92,59
Test F-L1-R1	94,07	97,78	97,04
Test 1	88,15	94,81	93,33
Test 2	88,15	94,81	93,33
Test 3	93,33	96,3	94,81
Test 4	92,56	96,3	95,56
Test 5	93,33	95,56	97,04
Test 6	95,56	95,56	97,78



top5

prove	Vgg16	Resnet	Senet
Test F	91,85	98,52	96,3
Test F-L1-R1	95,56	100	97,78
Test 1	90,37	98,52	97,04
Test 2	91,85	97,04	95,56
Test 3	94,81	97,78	97,04
Test 4	94,81	97,78	99,26
Test 5	94,81	97,04	100
Test 6	96,3	96,03	100



top10

prove	Vgg16	Resnet	Senet
Test F	96,3	100	99,26
Test F-L1-R1	99,26	100	100
Test 1	95,56	100	98,52
Test 2	97,78	100	100
Test 3	98,52	100	100
Test 4	98,52	100	100
Test 5	98,52	100	100
Test 6	98,52	100	100

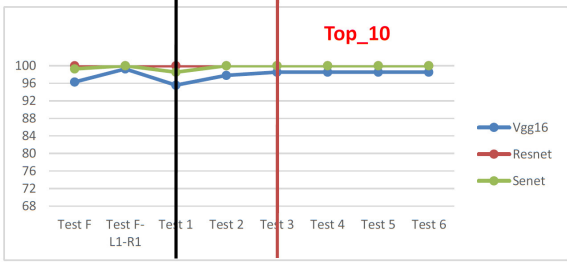


Figura 5.17: Risultati delle accuratezze percentuali delle tre reti neurali VGG16, ResNet50 e Senet, ottenute nel riconoscimento automatico del volto effettuato sul database FRMDB, in cui sono stati selezionati i fotogrammi della videosorveglianza. La linea verticale di colore nero in corrispondenza del $Test1$ evidenzia mediamente il peggior valore di accuratezza. La linea verticale di colore rosso, in corrispondenza del $Test3$, evidenzia il valore di accuratezza più alto oltre il quale l'aumento ulteriore, seppur presente, non fornisce un incremento considerevole.

Figura 5.18: Pose tipiche del fotosegnalamento canonico.

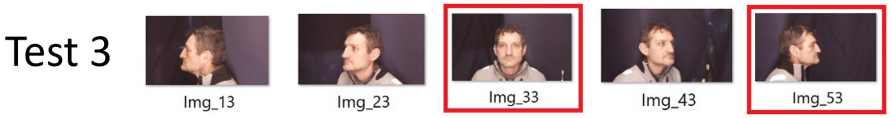


Figura 5.19: Pose del fotosegnalamento canonico, incrementato con i profili inclinati a -45°, +45° e il profilo sinistro a 90°.

I nuovi test hanno evidenziato il netto miglioramento dell'accuratezza sul riconoscimento dei volti eseguito sui fotogrammi delle videosorveglianze, in cui sono stati selezionati i profili prossimi alle pose del fotosegnalamento. Di fatto, osservando i grafici di figura 5.17, rispetto al fotosegnalamento canonico del $Test1$, le cui pose sono riportate in figura 5.18, tutti i test fanno registrare mediamente valori di accuratezza più alti per ogni rete neurale testata e in ogni classifica. Un valore di accuratezza considerevolmente maggiore del $Test1$, si registra a partire dal $Test3$, le cui pose sono riportate in figura 5.19, tale configurazione aggiunge al fotosegnalamento canonico solo due profili ruotati di +45° e -45° e il profilo sinistro ruotato di 90°. Ulteriori incrementi di immagini, dei test proposti, producono mediamente aumenti di accuratezza per tutte le reti testate e in ogni classifica. Essendo questa ricerca mirata a dettare un nuovo protocollo operativo per il fotosegnalamento che rispetti la legacy della procedura attuale, come vincolo che ci eravamo posti inizialmente, possiamo trascurare il $TestF$, che comprende solo la foto frontale, e il $TestF - L1 - R1$, riportato in figura 5.20, in quanto, seppur dotati di accuratezza abbastanza alta, non ricomprendono la posa destra del volto inclinata di 90°.

Per generalizzare il ragionamento, siamo andati a determinare la media totale delle accuratezze di tutte e tre le reti neurali inglobando tutte le quattro

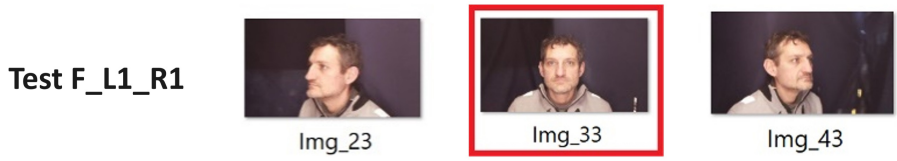


Figura 5.20: Configurazione di fotosegnalamento con profilo frontale, e inclinazioni a +45° e -45°.

Prove	Accuratezza media %
Test F	92,09916667
Test F-L1-R1	95,37083333
Test 1	91,29666667
Test 2	92,2225
Test 3	94,6275
Test 4	94,93583333
Test 5	95,185
Test 6	95,78166667

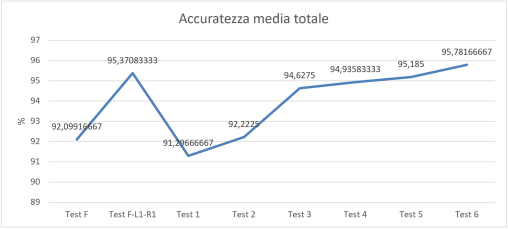


Figura 5.21: Andamento dell'accuratezza media calcolata totalmente sui risultati delle reti VGG16, ResNet50 e Senet, unendo tutte le classifiche Top1, Top3, Top5 e Top10.

Prove	Incremento %
Test 1	0
Test 2	0,925833333
Test 3	2,405
Test 4	0,308333333
Test 5	0,249166667
Test 6	0,596666667

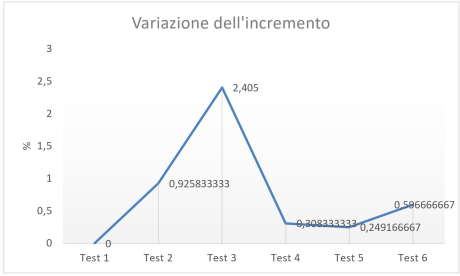


Figura 5.22: Variazione dell'incremento dell'accuratezza media valutato tra un Test e il successivo, azzerata sul Test1.

classifiche per ogni test, ottenendo i risultati riportati in figura 5.21. Trattandosi di valori via via crescenti mediamente dal $Test1$ in poi, siamo andati a calcolare la variazione percentuale dell'incremento tra un Test e il successivo, assumendo come valore di soglia per impostare lo 0% l'accuratezza del $Test1$. Dal risultato degli incrementi di accuratezza media, mostrati nel grafico di figura 5.22, si nota che la variazione più significativa, si ottiene in corrispondenza del $Test3$, effetto che ci consente di poter assumere questa configurazione come quella con il minor numero di pose che apporta un considerevole aumento di accuratezza nel riconoscimento automatico del volto.

Alla luce delle considerazioni fatte, possiamo accettare in prima approssimazione la configurazione del $Test3$, le cui pose sono riportate in figura 5.19, come possibile candidata per risolvere il quesito "QFQ" posto nel capitolo n.5.

Tuttavia, sono necessarie ulteriori ricerche in questa direzione, mirate a consolidare le conferme fin qui ottenute, il cui scopo preliminare era vedere se CNN addestrate tradizionalmente (come quelle usate nei sistemi di supporto tipo il SARI in uso attualmente alla Polizia Scientifica [126]) beneficiano del fatto che in fase di comparazione possono usare foto da più angolazioni. Per questo scopo, il dataset FRMDB, seppur limitato rispetto ad altri dataset tradizionalmente usati nel riconoscimento facciale, si è dimostrato efficace

per comprendere l'utilità di un fotosegnalamento multi-posa per il riconoscimento automatico dei volti dai sistemi di videosorveglianza. Alcuni ulteriori accorgimenti da adottare in sede di fotosegnalamento, utili ad aumentare le prestazioni del riconoscimento automatico, potrebbero essere rappresentati dai seguenti aspetti:

- Usare un sistema di acquisizione simultanea del volto, per garantire la stabilità espressiva;
- Parametrizzare le caratteristiche ambientali, come l'illuminazione, le distanze dall'obiettivo e l'altezza;
- Parametrizzare la postura del soggetto e le caratteristiche biometriche;
- Stimare l'influenza del pigmento della pelle;

Riguardo alle riprese dei sistemi di videosorveglianza, alcuni accorgimenti per migliorare il riconoscimento automatico potrebbero rappresentare:

- Utilizzare sistemi di filtraggio per la rimozione del rumore;
- Utilizzare dei sistemi automatici per stimare le pose del volto e catturare le più prossime a quelle presenti nel fotosegnalamento;

In fine sarebbe utile valutare automaticamente la minima qualità delle immagini che garantisce un'efficace riconoscimento facciale, per consolidare i risultati fin qui ottenuti e poter così definire un nuovo protocollo operativo e innovativo per le attività di fotosegnalamento per le Forze dell'Ordine.

Capitolo 6

Riconoscimento di violenza 2.0

La crescente disponibilità di tecnologie per la videosorveglianza, unita alla necessità di alleggerire le autorità dal compito di controllare ore di registrazioni video, ha stimolato l'attenzione di gruppi di ricercatori verso il rilevamento automatico della violenza nei video. Il rilevamento di scene di violenza è considerato parte del più generale campo del riconoscimento di azioni umane: nello specifico, si tratta di un problema binario che consiste nel riconoscere la presenza o l'assenza di violenza all'interno di video [4].

I primi lavori apparsi in merito al riconoscimento di scene di violenza sono basati su tecniche di Computer Vision già sviluppate per il riconoscimento delle azioni e possono essere categorizzati in due classi [56], sulla base delle caratteristiche estratte per rappresentare le azioni stesse:

- nelle tecniche basate sull'estrazione di caratteristiche locali, la rappresentazione di un'azione è calcolata utilizzando i punti di interesse ("Point of Interests" – POI) dai singoli fotogrammi di un video;
- nelle tecniche basate sull'estrazione di caratteristiche globali, la rappresentazione di un'azione è calcolata valutando le caratteristiche di più fotogrammi nel loro insieme.

Tra le tecniche che si basano su caratteristiche locali, Chen e Hauptmann [57] hanno proposto MoSIFT, una tecnica che combina la "Scale-Invariant Feature Transform" (SIFT) [58] con l'"Optical Flow", per rappresentare il movimento dei punti di interesse. Xu et al. [56] hanno fatto evolvere MoSIFT combinandolo con una "Kernel Density Estimation" (KDE) non parametrica per rimuovere le caratteristiche ridondanti e irrilevanti: usando lo "sparse coding" per rappresentare le caratteristiche estratte, hanno infatti ottenuto buoni risultati nel rilevare la violenza tra singoli individui nei video. Invece, Deniz et al. [59] hanno proposto di calcolare l'accelerazione dei movimenti a partire dallo spettro di potenza di fotogrammi adiacenti, al fine di rilevare una grande variazione di velocità, ottenendo risultati comparabili a MoSIFT, ma con un algoritmo più veloce in fase di esecuzione.

Per quanto riguarda le tecniche basate su caratteristiche globali, Hassner et al. [60] hanno proposto il calcolo dei descrittori "Violence Flows" (VIF), un'evoluzione dell'Optical Flow che calcola le variazioni di grandezza dei vettori di flusso, ottenendo risultati promettenti sul rilevamento della violenza nelle folle. Gao et al. [61] hanno aggiunto ai descrittori VIF l'orientamento del vettore di flusso, proponendo OVIF, migliorando le prestazioni del rilevamento della violenza tra singoli individui, ma con una minore accuratezza nel rilevamento della violenza nelle folle.

Il deep learning ha contribuito a far progredire il campo del rilevamento della violenza superando alcune delle limitazioni dell'Optical Flow, come le discontinuità e i movimenti della videocamera, nonché ottenendo ottime prestazioni sia nel rilevamento della violenza tra singoli individui che della violenza nelle folle, utilizzando lo stesso modello. In particolare, le CNN 3D si sono dimostrate capaci di apprendere informazioni spazio-temporali, cioè caratteristiche che rappresentano le informazioni di movimento in un video, oltre alle informazioni spaziali in un singolo frame. Per esempio, Ding et al. [62] hanno presentato una CNN 3D a 9 strati per il rilevamento della violenza, ottenendo un'accuratezza del 91% sul dataset "Hockey Fight" [63]. Con un approccio simile, Li et al. [64] hanno raggiunto il 98,3% di accuratezza sul dataset "Hockey Fight" e il 97,2% sul dataset "Crowd Violence" [60], sviluppando una CNN 3D a 10 strati che alterna strati completamente connessi e di transizione a seguito di uno strato di convoluzione. Anche metodologie basate sul "transfer learning", facendo uso di CNN 3D pre-addestrate, hanno portato a ottimi risultati. Ad esempio, in ricerche precedenti [4] abbiamo ottenuto un'accuratezza del 98,5% e del 99,2% su "Hockey Fight" e "Crowd Violence" usando C3D [5], una CNN 3D pre-addestrata a classificare categorie di sport, come estrattore di caratteristiche e, in cascata, un classificatore SVM ("Support Vector Machine"). Allo stesso modo, Ullah et al. [65] hanno usato C3D come estrattore di caratteristiche, ma aggregandola a strati completamente connessi per la classificazione, con buone prestazioni in entrambi i dataset "Hockey Fight" (96% di precisione) e "Crowd Violence" (98%). Oltre alle CNN 3D, anche l'architettura di tipo ConvLSTM [66] si è dimostrata efficace nel rilevamento della violenza. A tal fine, Sudhakaran e Lanz [67] hanno proposto di combinare le informazioni spaziali estratte dai fotogrammi da una CNN 2D con una ConvLSTM, ottenendo un'accuratezza del 97,1% sul dataset "Hockey Fight" e del 94,5% sul dataset "Crowd Violence".

Pertanto, diverse tecniche basate su deep learning hanno dimostrato la loro accuratezza sui dataset tradizionalmente usati in letteratura, come l'"Hockey Fight" e il "Crowd Violence". Tuttavia ci sono ancora ricerche in corso per validare la loro robustezza contro i falsi positivi [68], cioè in grado di riconoscere scene violente da scene non violente ma apparentemente confondibili come

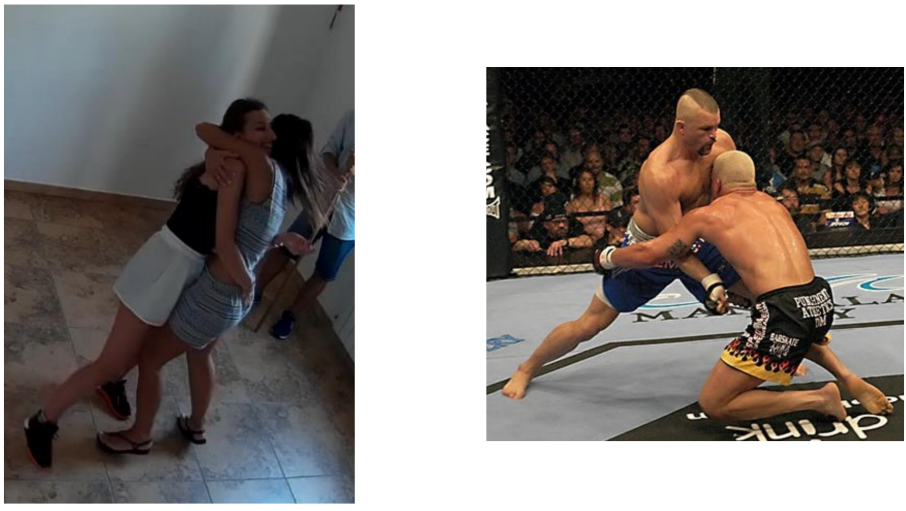


Figura 6.1: Esempio di immagini per testare la robustezza delle reti neurali nel discriminare i falsi positivi: l'immagine (a) mostra un abbraccio di natura non violenta, l'immagine (b) mostra un abbraccio di natura violenta

potrebbe essere il caso indicato in figura 6.1, e l'accuratezza su filmati da reali circuiti di videosorveglianza [69].

6.1 Techiche di Deep Learning per il rilevamento automatico della violenza nei video

Negli ultimi anni sono apparse diverse tecniche per il rilevamento automatico di scene di violenza in video e filmati di sicurezza che in ambito forense hanno l'obiettivo principale di sollevare le autorità dalla necessità ad analizzare ore di filmati a circuito chiuso (CCTV). A questo proposito, le tecniche basate sul Deep Learning, come le Reti Neurali Ricorrenti (RNN) e le Reti Neurali Convoluzionali (CNN), si sono rivelate efficaci per il rilevamento della violenza. Tuttavia, la maggior parte di queste tecniche richiede notevoli risorse computazionali e di memoria per eseguire il rilevamento automatico della violenza, caratteristiche che rendono problematica l'applicazione di tali tecnologie a bordo macchina. Uno strumento che invece potrebbe favorire il controllo, da parte delle Forze dell'Ordine, di aree sensibili agli atti violenti. Nei capitoli che seguono sono riportate alcune ricerche che ho sviluppato all'interno del laboratorio AIRTLab, già oggetto di pubblicazione, che fissano le basi verso la

progettazione di una rete neurale adeguata all'implementazione a bordo macchina. Tale tecnologia sfrutta le proprietà che derivano dalla combinazione di una CNN consolidata, MobileNetV2, progettata per l'uso in dispositivi integrati, con uno strato ricorrente per estrarre le caratteristiche spazio-temporali nei video di sicurezza. Un modello leggero può essere eseguito in dispositivi integrati, nella modalità tipica del cd. "edge computing", ad esempio per consentire l'elaborazione dei video vicino alla telecamera che li registra, per preservare la privacy. Nello specifico, sfruttiamo il transfer learning, utilizzando una versione pre-addestrata di MobileNetV2, e proponiamo due diversi modelli combinandoli con una rete "Long Short-Term Memory" bidirezionale (Bi-LSTM) e una LSTM convoluzionale (ConvLSTM). L'articolo presenta test di accuratezza dei due modelli sul dataset AIRTLab e un confronto con modelli più complessi sviluppati nel nostro lavoro precedente, al fine di valutare il calo di accuratezza necessario per utilizzare un modello compatibile con risorse limitate. La rete composta da MobileNetV2 e ConvLSTM ottiene un'accuratezza del 94,1%, contro il 96,1% di un modello basato su una CNN 3D più complessa.

A seguito della crescente disponibilità di telecamere di videosorveglianza e la necessità di tecniche per identificare automaticamente gli eventi nei filmati, c'è un crescente interesse verso il rilevamento automatico della violenza nei video. Le architetture basate sul deep learning, come le reti neurali convoluzionali 3D, hanno dimostrato la loro capacità di estrarre caratteristiche spazio-temporali dai video, risultando efficaci nel rilevamento della violenza. Tuttavia, comportamenti amichevoli o movimenti rapidi come abbracci, piccoli colpi, applausi, battimani, ecc. possono ancora causare falsi positivi, risultando nell'interpretazine di un'azione innocua come violenta. Scendendo nel particolare dell'attività di ricerca in questo ambito, rientrante nei punti dell'Accordo d'Intesa tra la Polizia di Stato e l'UNIVPM, l'obiettivo è addestrare sistemi automatici al rilevamento di violenza da videoriprese, sia per il controllo real-time di ampi spazi, sia per il processamento rapido di lunghe registrazioni. È stato prodotto l'AIRTLab dataset, unico nel suo genere, con 350 clips di scene violente e non violente ma confondibili. Testando 3 modelli 3D basati su deep learning, il dataset si è confermato efficace per testare la robustezza delle reti neurali sui falsi positivi. Ma dato il peso computazionale, in linea con l'obiettivo, è stata combinata una CNN, progettata per dispositivi integrati, con un layer ricorrente ottenendo la perdita di solo 1%AUC e 2% Accuracy a vantaggio di leggerezza e memoria. Nel tempo residuo, rientrante nell'accordo d'Intesa tra l'UNIVPM e la Polizia di Strato, si replicherà l'approccio basato su Deep Learning per cercare di riconoscere la violenza nei file audio, al fine di contrastare gli atti di violenza di genere, purtroppo sempre molto numerosi, e di conseguenza anche limitare gli effetti negativi sulle vittime indirette, come per esempio sui figli che assistono alle liti tra i genitori.

6.2 *AIRTLab* - Un dataset per il riconoscimento automatico della violenza nei video

[*pubblicato*][1]

Il rilevamento automatico di violenza e crimini all'interno di video è un argomento che sta destando interesse, specialmente nell'ottica di alleggerire dal bisogno di controllare ore di registrazioni per identificare eventi di pochi secondi. I dataset tradizionalmente disponibili per testare soluzioni di rilevamento automatico di violenza sono composti da poche clip a bassa risoluzione, spesso raffiguranti casi molto specifici (ad esempio risse sui campi da hockey). Nonostante la comparsa di nuovi dataset per superare tali limiti, c'è ancora bisogno di video pensati per mettere alla prova la robustezza ai falsi positivi, causati spesso da azioni che sembrano violente, ma non lo sono. In questo articolo proponiamo dunque un nuovo dataset composto da 350 clip (file MP4, 1920 x 1080 pixel, 30 fps), annotate come "non-violent" (120 clip) quando rappresentano comportamenti non violenti e "violent" (230 clip) quando, al contrario, includono scene di violenza. In particolare, le clip non violente includono azioni (abbracci, battiti di mani, esultanze, ecc.) che possono risultare dei falsi positivi durante il rilevamento automatico della violenza, a causa di movimenti veloci e somiglianza con comportamenti violenti. Tutte le clip incluse nel dataset sono state girate da attori non professionisti (ogni clip contiene dai 2 ai 4 attori).

6.2.1 Valore dei dati

- Rispondendo al crescente interesse verso il rilevamento automatico di violenza e crimini all'interno di video, il dataset proposto è pensato per addestrare e mettere alla prova tecniche di riconoscimento automatico della violenza.
- Nel breve e nel medio periodo, i ricercatori interessati a queste tecniche potranno usare le clip Full HD per addestrare e testare i loro algoritmi. Nel lungo periodo forze dell'ordine, autorità e l'intera comunità di cittadini potrebbe beneficiare di algoritmi sempre migliori, capaci di ridurre il tempo di reazione alle riprese di scene di violenza e crimini.
- Uno specifico obiettivo di questo dataset è testare la robustezza ai falsi positivi da parte delle tecniche di riconoscimento automatico di violenza. Per questa ragione, il dataset può essere preso in considerazione negli esperimenti per la valutazione dell'accuratezza degli algoritmi.

[1] Contardo, P., Bianculli, M., Falcionelli, N., Sernani, P., Tomassini, S., Lombardi, M., Dragoni, A. F. (2020). A dataset for automatic violence detection in videos. Data in brief, 33, 106587.

Tabella 6.1: Descrizione delle specifiche

Disciplina	Visione Artificiale e Riconoscimento di Pattern
Area specifica	Rilevamento Automatico della Violenza nei Video.
Tipologia dei dati	File Video (mp4) e File di Testo (csv)
Come sono stati acquisiti i dati	Le clip sono state registrate da due telecamere sistemate in due punti diversi di una stanza, al fine di costruire un dataset contenente video ripresi da due differenti angolazioni. Le telecamere usate sono: • La fotocamera frontale di un Asus Zenfone Selfie ZD551KL (13 MP, Auto Focus, f/2.2). • La TOPOP Action Cam OD009B (12 MP, obiettivo fisheye 170°).
Formato dei dati	Raw
Parametri per la raccolta dei dati	Le clip sono in formato MP4, codec H.264, a risoluzione 1920 x 1080 pixel e 30 fps. La durata media delle clip è 5,63 secondi (il video più corto dura 2 secondi, il più lungo 14). 230 clip su 350 sono annotate come violente ("violent") mentre le restanti 120 sono annotate come non violente ("non-violent").
Descrizione della raccolta dei dati	Il dataset contiene 350 clip, divise in due cartelle "violent" e "non-violent". Tali cartelle sono ulteriormente divise in due sottocartelle, "cam1" e "cam2": • "violent/cam1" contiene 115 clip con comportamenti violenti; • "violent/cam2" contiene 115 clip con gli stessi comportamenti violenti disponibili in "violent/cam1", ma registrati con un'altra telecamera e da un differente punto di vista; • "non-violent/cam1" contiene 60 clip di comportamenti non violenti; • "non-violent/cam2" contiene 60 clip con gli stessi comportamenti non violenti disponibili in "non-violent/cam1", ma registrati con un'altra telecamera e da un differente punto di vista. Le clip sono state registrate da un gruppo di attori non professionisti (da 2 a 4 per clip) che recitano le azioni violente e non violente.
Posizione della raccolta di dati	Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, Ancona, Italy.
Accessibilità ai dati	Repository pubblico: GitHub (`https://github.com`) Nome del repository: A Dataset for Automatic Violence Detection in Videos URL diretto ai dati: https://github.com/airtlab/A-Dataset-for-Automatic-Violence-Detection-in-Videos

```
violence-detection-dataset
  ├─ non-violent
  │    ├─ cam1 (60 .mp4 clip)
  │    └─ cam2 (60 .mp4 clip)
  └─ violent
       ├─ cam1 (115 .mp4 clip)
       └─ cam2 (115 .mp4 clip)
```

Figura 6.2: La struttura del repository contenente le 350 clip del dataset, divise in non violente (120 clip) e violente (230 clip)

6.2.2 Descrizione dei dati

La grande disponibilità di telecamere di sicurezza e l'esigenza di prendere decisioni tempestive nonostante la necessità di visionare ore di filmati [185] hanno catalizzato l'interesse dei ricercatori verso lo sviluppo di tecniche di rilevamento automatico di violenza e crimini all'interno di video. Recentemente, tanto le tecniche basate su feature selezionate [196][61] quanto quelle basate su "deep learning" [197][4] si sono dimostrate accurate nella rilevazione della violenza, in particolare nei video disponibili grazie a dataset come l'"Hockey Fight Dataset" [63], il "Movie Fight Dataset" [63] e il "Crowd Violence Dataset" [60]. Tuttavia, tali dataset mettono a disposizione pochi video, a bassa risoluzione, spesso registrati in ambienti molto specifici (per esempio stadi per l'hockey). Un tentativo di superamento di queste criticità è "RWF-2000" [69], un dataset che contiene 2000 clip prese da circuiti di videosorveglianza. Ciononostante, c'è ancora il bisogno di capire l'accuratezza delle tecniche di rilevamento automatico della violenza in video raffiguranti azioni rapide, non violente (battiti di mani, abbracci, esultanze, ecc..), che possono causare falsi positivi. Per tale ragione, questo articolo presenta un dataset specificatamente progettato per contenere video non violenti che possono causare falsi positivi. Il dataset è composto da 350 clip nella forma di file video MP4 (con codec H.264) della durata media di 5,63 secondi, dove il video più corto dura 2 secondi, mentre il più lungo 14. Tutte le clip sono a risoluzione $1920 \times 1080 pixel$, con il frame rate pari a 30 fps. I video sono organizzati in cartelle come mostrato in Figura 6.2. I video del dataset sono divisi in due cartelle principali: "*non-violent*" e "*violent*". Tale divisione fornisce l'annotazione dei video, distinguendoli rispettivamente in non violenti e violenti. Per ogni cartella è presente un'ulteriore suddivisione in due sottocartelle: "cam1" e "cam2". In particolare:

- "non-violent/cam1" contiene 60 clip che rappresentano comportamenti non violenti;
- "non-violent/cam2" contiene 60 clip con gli stessi comportamenti non violenti disponibili in "non-violent/cam1", ma registrati con un'altra telecamera e da un differente punto di vista;

Figura 6.3: Un frame preso da una clip violenta (telecamera 1

- "violent/cam1" contiene 115 clip che rappresentano comportamenti violenti;
- "violent/cam2" contiene 115 clip con gli stessi comportamenti violenti disponibili in "violent/cam1", ma registrati con un'altra telecamera e da un differente punto di vista.

Le clip sono state registrate da un gruppo di attori non professionisti. Ogni clip contiene un minimo di 2 attori e un massimo di 4. Nei video violenti 6.3, agli attori è stato chiesto di simulare azioni tipiche delle risse come calci, pugni, schiaffi, percosse con un bastone, pugnalate e colpi di pistola. Nei video non violenti 6.4, agli attori è stato chiesto di simulare azioni che possono causare falsi positivi nel rilevamento della violenza a causa della velocità dei movimenti o della somiglianza con le azioni violente. Per esempio, i video non violenti includono azioni come abbracci, esultanze, soggetti che si scambiano il "cinque" o applaudono, e soggetti che gesticolano.

All'interno del repository sono disponibili tre file csv con un'annotazione aggiuntiva per i video:

- il file "action-class-occurrences.csv" fornisce l'elenco di tutte le azioni registrate nelle clip, insieme al numero di volte che ciascuna azione compare nel dataset, oltre ad un'etichetta per spiegare se l'azione è violenta (y) o no (n). Tutte le azioni presenti nelle clip sono elencate in Tabella 1;
- il file "non-violent-action-class.csv" elenca tutte le azioni delle clip non violente;
- il file "violent-action-class.csv" elenca tutte le azioni delle clip violente.

Figura 6.4: Un frame preso da una clip non-violenta (telecamera 2)

Tabella 6.2: Lista delle azioni (e del relativo numero di occorrenze) presenti nel dataset.

Azioni violente		**Azioni non violente**	
Azione	**Occorrenza**	**Azione**	**Occorrenza**
fight	46	greet	33
club	36	hug	16
punch	23	handgesture	15
push	22	jump	10
kick	21	highfive	6
slap	18	handshake	3
stab	15	walk	1
gunshot	14		
choke	13		

6.2.3 Setting Sperimentale, Materiali e Metodi

Come evidenziato in una precedente ricerca [4], le tecniche di rilevamento della violenza possono erroneamente interpretare alcune azioni come violente, a causa di movimenti rapidi e della somiglianza con comportamenti violenti. Per questo, le clip non violente sono state registrate per mettere alla prova gli algoritmi di rilevamento della violenza e verificarne la robustezza ai falsi positivi, anche quando il dataset non è bilanciato, come nel nostro caso. Nelle registrazioni di clip violente, oltre a simulare calci, pugni e schiaffi, sono stati usati una pistola di plastica, un coltello giocattolo e un bastone di legno (ricoperto di carta a bolle) per rappresentare in maniera verosimile le azioni che coinvolgono armi. Le clip sono state registrate con due diverse telecamere piazzate in due punti differenti di una stanza, cosicché i video rappresentassero azioni da due punti di vista differenti. Le telecamere usate sono:

- La fotocamera frontale di un Asus Zenfone Selfie ZD551KL (13 MP, Auto Focus, f/2.2).
- La TOPOP Action Cam OD009B (12 MP, obiettivo fisheye 170°).

Tutte le clip sono state girate nella stessa stanza, in condizioni di luce naturale. L'Asus Zenfone è stato posizionato nell'angolo in alto a sinistra di fronte la porta, mentre l'Action Cam nell'angolo in alto a destra di fianco la porta. Tutte le azioni sono state registrate contemporaneamente da entrambe le telecamere. Pertanto, le clip con nome identico, ma in cartelle diverse (per esempio "non-violent/cam1/1.mp4" e "non-violent/cam2/1.mp4") mostrano la stessa azione registrata da angolazioni differenti (la cartella "cam1" contiene i video registrati con l'Asus Zenfone, la cartella "cam2" contiene i video ripresi con l'Action Cam). Inoltre, in aggiunta all'annotazione dei video in violenti e non-violenti, abbiamo annotato manualmente le azioni contenute in ciascuna clip. Tale annotazione può essere usata in esperimenti di classificazione più dettagliata, al fine di addestrare e testare algoritmi capaci di riconoscere le singole azioni presenti nei video.

6.3 Deep Learning per il riconoscimento automatico della violenza, test sul dataset AIRTLab

[*pubblicato*][2]

[2] Contardo, P., Sernani, P., Falcionelli, N., Tomassini, S., Dragoni, A. F. (2021). Deep learning for automatic violence detection: Tests on the AIRTLab dataset. IEEE Access, 9, 160580-160595.

I sistemi di videosorveglianza pubblici sono comuni in tutto il mondo, essendo in grado di fornire informazioni accurate e ricche in molte applicazioni connesse alla sicurezza [141]. Tuttavia, la necessità di guardare ore di filmati compromette la possibilità di prendere decisioni in tempi brevi, elemento essenziale nella videosorveglianza per la prevenzione della criminalità e della violenza [185]. A questo proposito sono stati presentati diversi studi sulla rilevazione automatica di scene violente nei video, con l'obiettivo di sgravare le autorità dalla necessità di guardare ore di video per identificare eventi della durata di pochi secondi. Mentre i primi lavori di ricerca utilizzavano caratteristiche artigianali e descrittori di flusso tipici dei metodi tradizionali di riconoscimento dell'azione [60, 56, 61], lavori recenti hanno evidenziato l'accuratezza degli approcci basati sull'apprendimento profondo nel rilevamento della violenza [198, 199, 65].

Tra le tecniche basate sul deep learning per il rilevamento della violenza, nella nostra precedente ricerca [4] abbiamo mostrato l'efficacia della combinazione di reti neurali convoluzionali 3D (3D CNN) e Support Vector Machine (SVM) per rilevare nei video sia i combattimenti da persona a persona che la violenza della folla. Tuttavia, abbiamo evidenziato che ci sono ancora falsi positivi, come comportamenti amichevoli o mosse rapide (abbracci, piccoli colpi, applausi, batti cinque, ecc.) rilevati come violenti. Per indagare ulteriormente in tale direzione, questo articolo presenta un confronto di tre diversi modelli di deep learning su un nuovo insieme di dati, il dataset AIRTLab [68], che abbiamo creato per includere, come esempi non violenti, video clip che possono causare falsi positivi. Nello specifico, questo documento aggiunge i seguenti contributi allo stato dell'arte del rilevamento automatico della violenza nei video:

- descrive un nuovo dataset da usare per confrontare le tecniche per il rilevamento automatico della violenza nei video. Il dataset è specificamente progettato per testare le prestazioni rispetto ai falsi positivi;

- propone due modelli basati sul transfer learning e un modello "addestrato da zero", testandoli sul dataset presentato. I risultati servono come base per confrontare le prestazioni delle tecniche di rilevamento della violenza applicate al dataset proposto;

- confronta le prestazioni dei modelli proposti con note reti neurali convoluzionali 2D (2D CNN) pre-addestrate. Nello specifico, verifica le prestazioni di VGG16 e VGG19 [174], ResNet50 versione 2 [54], Xception [200] e NASNet Mobile [201]. Essendo 2D, queste reti sono state adattate per essere applicate frame per frame ai video ed elaborare informazioni spazio-temporali, al fine di essere confrontate con i modelli proposti. In questo modo, il presente articolo fornisce anche un confronto sul compito

del rilevamento della violenza di reti 2D già note in letteratura per essere efficaci in altri contesti;

- fornisce l'implementazione di tutti i modelli e gli esperimenti descritti, in quanto il codice sorgente dei test è pubblicamente disponibile in un repository GitHub[3], per garantire la riproducibilità del esperimenti. Inoltre, il dataset AIRTLab è disponibile anche in un repository pubblico[4].

Infatti, i dataset tradizionalmente utilizzati per confrontare le tecniche di rilevamento della violenza, come l'Hockey Fight Dataset [63], il Movie Fight Dataset [63] e il Crowd Violence Dataset [60], di solito includono pochi video, registrati a bassa risoluzione; in molti casi tali video sono registrati in ambienti troppo specifici (come arene di hockey e stadi di calcio). Invece, il dataset proposto in questo documento include 350 clip Full HD, a 30 fotogrammi al secondo.

Inoltre, in seguito alla nostra precedente ricerca, in questo articolo usiamo C3D, una CNN 3D pre-addestrata per classificare sport nei video [5], come estrattore di caratteristiche nei due modelli basati su transfer learning. Nel primo modello presentato, l'attività di classificazione viene eseguita da un classificatore SVM. Nel secondo modello, la classificazione avviene grazie a strati completamente connessi. Al contrario, il modello che è stato addestrato da zero si basa sull'architettura Convolutional Long Short-Term Memory (ConvLSTM) [66] per estrarre le caratteristiche spazio-temporali dei video; l'attività di classificazione viene poi eseguita da strati completamente connessi che seguono la ConvLSTM. Abbiamo testato questi tre modelli sul dataset AIRTLab, per confrontare le loro prestazioni. In ogni caso, più che proporre nuove architetture e modelli di riconoscimento, il contributo maggiore della ricerca presentata è la conferma che il dataset AIRTLab sia in grado di sfidare la robustezza ai falsi positivi delle tecniche di rilevamento della violenza, testando architetture che hanno già mostrato buone performance con dataset diversi. Abbiamo dunque testato i nostri modelli anche sui dataset Hockey Fight e Crowd Violence, costruendo un confronto sulla letteratura esistente. Infine, confrontando i modelli proposti con le prestazioni ottenute da modelli end-to-end basati su note CNN 2D pre-addestrate, forniamo anche dei risultati di reti 2D sul dataset AIRTLab, nonché sull'Hockey Fight Dataset e il Crowd Violence Dataset.

Il resto di questo documento è organizzato come segue. La sezione *Stato dell'arte* elenca ricerche di interesse rispetto al lavoro descritto in questo documento, evidenziando somiglianze e differenze con la ricerca presentata. La

[3] Codice sorgente degli esperimenti: https://github.com/airtlab/violence-detection-tests-on-the-airtlab-dataset

[4] Dataset: https://github.com/airtlab/A-Dataset-for-Automatic-Violence-Detection-in-Videos

sezione *Materiali e metodi* descrive le tecniche di deep learning utilizzate, fornendone le basi necessarie, dettagliando la struttura del dataset e spiegando le architetture dei tre modelli proposti. La sezione *Valutazione Sperimentale* presenta e discute i risultati sperimentali e i principali risultati, descrivendo l'esecuzione dei test sul dataset proposto così come i risultati ottenuti su altri dataset. Infine, la Sezione *Conclusioni* trae le conclusioni di questa ricerca.

6.3.1 Stato dell'arte

Per quanto riguarda le tecniche di rilevamento della violenza, negli ultimi anni diversi modelli basati su deep learning hanno mostrato le loro potenzialità sui dataset menzionati nella sezione precedente, raggiungendo prestazioni di altissimo livello in termini di accuratezza della classificazione. Tra queste tecniche, le CNN 3D e le ConvLSTM si sono dimostrate efficaci nell'apprendere le informazioni spazio-temporali contenute nei video [202]. A questo proposito, la Tabella 6.3 elenca le performance in termini di accuratezza delle tecniche di rilevamento della violenza basate su Deep Learning, sui dataset Hockey Fight e Crowd Violence. Tra i dataset per confrontare tecniche di rilevamento di violenza, l'Hockey Fight Dataset [63], il Movie Fight Dataset [63], e il Crowd Violence Dataset [60] sono stati utilizzati ampiamente. L'Hockey Fight Dataset include 1000 clip equamente suddivise in "fight" e "no-fight". Ogni clip ha tra 41 e 50 fotogrammi (come riportato anche in [203]), originariamente ad una risoluzione di $720 \times 576 pixel$, anche se la versione $320 \times 240 pixel$, come proposto in [204] è comunemente usata. Il dataset Movie Fight include 200 clip estratte da film, 100 etichettate come "fight" e 100 come "no-fight". Analogamente all'Hockey Fight Dataset, ogni clip è composta da 50 fotogrammi a 720×576 e $720 \times 480 pixel$. Il Crowd Violence Dataset è composto da 246 clip scaricate da Youtube (123 violente, 123 non violente), con una risoluzione di $320 \times 240 pixel$ e con una durata media di 3,6 secondi. Tutti questi dataset includono video a bassa risoluzione; i dataset Hockey Fight e Crowd Violence includono clip registrate in ambienti molto specifici (come arene di hockey e stadi di calcio); il Movie Fight Dataset contiene pochi fotogrammi (10000) e la maggior parte degli studi recenti ha già raggiunto 100% precisione su di esso (vedi, per esempio, [67] e [64]). Recentemente, Cheng et al. [69] hanno proposto un dataset, RFW-2000, pensato per superare questi problemi. RWF2000 è composto da 2000 clip da filmati di videosorveglianza, a varie risoluzioni. Analogamente al RWF-2000, il dataset che proponiamo supera i limiti dei dataset tradizionali, offrendo 350 clip di varia durata (media 5,36 secondi), in risoluzione Full HD ($1920 \times 1080 pixel$). Tuttavia, a differenza degli altri dataset, AIRTLab include nelle clip non violente azioni come abbracciarsi, dare il cinque e applaudire, esultare e gesticolare che potrebbero risultare in falsi

Tabella 6.3: Accuratezza delle tecniche di rilevamento della violenza basate sul deep learning sui dataset Hockey Fight e Crowd Violence. L'ultima riga riporta i risultati del nostro precedente lavoro [4], basato sulla combinazione della rete C3D (pre-addestrata) con un classificatore SVM.

Autori	Architettura	Hockey Fight	Crowd Violence
Ding et al. (2014) [62]	3D CNN	91.0%	-
Song et al. (2019) [203]	3D CNN	99.6%	94.3%
Li et al. (2019) [64]	3D CNN	98.3%	97.2%
Ullah et al. (2019) [65]	C3D + Fully connected layers	96.0%	98.0%
Sudhakaran and Lanz. (2017) [67]	2D CNN + ConvLSTM	97.1%	94.5%
Hanson et al (2018) [205]	2D CNN + ConvLSTM	98.1%	96.3%
Accattoli et al. (2020) [4]	C3D + SVM	98.5%	99.2%

positivi, a causa di movimenti bruschi e somiglianza con alcuni comportamenti violenti. Pertanto, il dataset AIRTLab è progettato per testare la robustezza delle tecniche di rilevamento della violenza contro i falsi positivi. A questo proposito, 350 clip potrebbero sembrare poche rispetto ad altri ambiti tipici della computer vision e alle 1000 clip dell'Hockey Fight. Tuttavia, dobbiamo evidenziare che la durata media della clip (5,63 secondi) è superiore all'Hockey Fight (circa 2 secondi) e al Crowd Violence (3,6 secondi). Sono pertanto inclusi più fotogrammi rispetto ai dataset tradizionalmente usati in letteratura. In particolare, Ding et al. [62], hanno proposto una CNN 3D a 9 strati per il rilevamento della violenza: processando 40 fotogrammi alla volta, ad risoluzione di 60 frame di x di 90 pixel, tre strati convoluzionali (3D) alternati a due strati di pooling, due strati completamente connessi e uno strato softmax per la classificazione hanno raggiunto un'accuratezza del 91% sul dataset Hockey Fight. Più di recente, Song et al. [203] hanno raggiunto 99.6% di accuratezza sul dataset Hockey Fight, e 94.3% sul Crowd Violence, addestrando da zero una CNN 3D che riproduce l'architettura C3D e migliorandone il metodo di campionamento. Allo stesso modo, Li et al. [64], con CNN 3D a 10 strati che alterna strati di transizione e strati completamente connessi dopo uno strato convoluzionale, ha raggiunto il 98.3% di accuratezza sull'Hockey Fight, e il 97.2% sul Crowd Violence. Tuttavia, gli approcci di tipo transfer learning basati su CNN 3D hanno ottenuto risultati ancora migliori utilizzando modelli pre-addestrati su attività di classificazione diverse dal rilevamento della violenza. Ullah et al. [65] hanno implementato un modello pre-addestrato basato su C3D, usandone fino al secondo strato completamente connesso ("FC7"), usando uno strato SoftMax per effettuare la classificazione, ottenendo buone prestazioni sia nell'Hockey Fight (96% di accuratezza) che nel Crowd Violence (98%). Analogamente, nel nostro lavoro precedente [4], abbiamo usato C3D, fino a al primo strato completamente connesso ("FC6"), come un estrattore di caratteristiche, e un classificatore SVM per il rilevamento della violenza, raggiungendo ottime prestazioni sia nell'Hockey Fight (98.5% di accuratezza) che nel Crowd Violence (99.2%). Anche l'uso dell'architettura ConvLSTM nei modelli

di rilevamento della violenza ha ottenuto risultati promettenti. Ad esempio, Sudhakaran e Lanz. [67] hanno proposto di aggregare le informazioni spaziali estratte dai fotogrammi con CNN 2D con le informazioni temporali estratte da una ConvLSTM. Con tale architettura hanno raggiunto il 97.1% di accuratezza sul dataset Hockey Fight, e il 94.5% sul Crowd Violence. Un approccio simile è stato proposto da Hanson et al. [205] che ha unito la CNN 2D VGG13 [174] con uno strato ConvLSTM per raggiungere il 98.1% di accuratezza sull'Hockey Fight, e 96.3% sul Crowd Violence.

6.3.2 Materiali e metodi

Come sottolineato nelle sezioni Introduzione e Stato dell'arte, le architetture basate sul deep learning e, in particolare, CNN 3D e ConvLSTM, sono in grado di modellare le caratteristiche spazio-temporali dei video e hanno dimostrato la loro accuratezza nel rilevamento della violenza. Per questo motivo, abbiamo basato i classificatori proposti in questo articolo su tali tipologie di rete neurale, confrontando:

- reti end-to-end, cioè composti da un unico modello per eseguire la classificazione, con un modello composto da una CNN 3D e una SVM;
- una rete addestrata da zero sul rilevamento della violenza con due reti basate su transfer learning, cioè addestrate su un dataset per scopi diversi dal rilevamento della violenza.

La figura 6.5 mostra il flusso di lavoro alla base della ricerca descritta in questo documento. Proponiamo tre modelli di deep learning (due sono basati sul transfer learning e uno è addestrato da zero) per eseguire il rilevamento della violenza nei video. Introduciamo un dataset, il dataset AIRTLab, e presentiamo le prestazioni dei modelli proposti su tale dataset. Per costruire un confronto sulla letteratura esistente, testiamo le loro prestazioni anche sui dataset Hockey Fight e Crowd Violence, tradizionalmente utilizzati in letteratura per confrontare le tecniche di rilevamento della violenza. Infine, per dimostrare l'importanza dei classificatori proposti, confrontiamo i nostri modelli con le prestazioni ottenute da ben note CNN 2D pre-addestrate come VGG16, VGG19 e ResNet50, adattate per essere applicate ai video (che sono 3D, essendo composti di più fotogrammi).

A tal fine, forniamo alcune nozioni di base su CNN 3D e l'architettura ConvLSTM (6.3.3), presentiamo il dataset AIRTLab, utilizzato per testare i classificatori proposti (6.3.4), e descriviamo l'architettura dei modelli proposti (6.3.5).

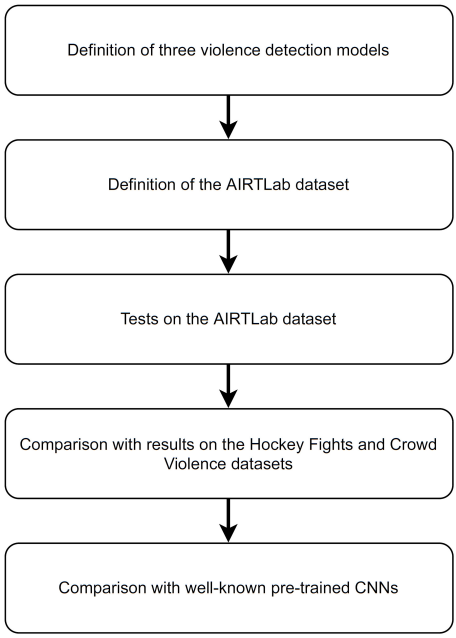


Figura 6.5: Il flusso di lavoro dello studio proposto in questo documento.

6.3.3 Nozioni di base: CNN 3D and ConvLSTM

Come evidenziato nel lavoro seminale di LeCun e Bengio [206], in una CNN 2D ogni unità di uno strato riceve l'input da un insieme di unità (il local receptive field) dello strato precedente. Ji et al [207] hanno esteso questo concetto proponendo di utilizzare CNN 3D. La convoluzione 3D si ottiene utilizzando un kernel 3D sul cubo formato impilando insieme più frame adiacenti. In questo modo, la mappa delle caratteristiche risultante rappresenta le informazioni temporali disponibili nei dati campione, oltre alle informazioni spaziali solitamente modellate da una CNN 2D.

In questo lavoro, usiamo un CNN 3D, C3D [5], già addestrata sul dataset Sports-1M [208] per riconoscere categorie di sport nei video. Poiché si è dimostrato utile estrarre caratteristiche spazio-temporali dai video, utilizziamo C3D come estrattore di caratteristiche, usando i pesi fino al primo strato completamente connesso ("fc6"), con un approccio di tipo transfer learning.

Mentre C3D è utilizzato in due dei tre modelli proposti, l'ultimo modello si basa su ConvLSTM che si è parimenti dimostrata utile per rappresentare le caratteristiche spazio-temporali. Specificamente, usiamo la formulazione di Shi et al. [66], che hanno esteso l'architettura LSTM [209] aggiungendo strutture convoluzionali di transizione di stato. n'unità nascosta LSTM è composta da una cella auto-ricorrente, chiamata cella di memoria, il cui ingresso/uscita è

regolato da tre porte moltiplicative, ovvero la porta di ingresso, la porta di uscita e la porta "forget" [210].

Come Shi et al. hanno sottolineato, l'architettura LSTM è adeguata per estrarre le caratteristiche temporali, ma include troppa ridondanza per le caratteristiche spaziali. A questo proposito, hanno proposto di aggiungere strutture convoluzionali nelle transizioni tra la porta di ingresso e la cella di memoria, e nell'autoricorrenza della cella di memoria, regolata dalla porta "forget".

6.3.4 Il dataset AIRTLab

Per valutare i tre modelli di deep learning proposti, abbiamo sviluppato un dataset, chiamato AIRTLab, per testare specificamente la robustezza delle tecniche di rilevamento della violenza contro i falsi positivi in clip non violente con movimenti rapidi (come abbracci, applausi, batti cinque, ecc.). Il dataset è disponibile pubblicamente in un repository GitHub.

Il dataset è composto da 350 clip che sono file video MP4 (codec H.264) della durata media di 5,63 secondi, con il video più breve della durata di 2 secondi e il video più lungo di 14 secondi. Per tutte le clip, la risoluzione è di 1920 x 1080 pixel e il frame rate di 30 fps. Il dataset è suddiviso in due directory principali, "non-violent" e "violent", etichettando le clip incluse come contenenti rispettivamente comportamenti non violenti e comportamenti violenti. Le directory sono suddivise in due sottodirectory, "cam1" e "cam2":

- "non-violent/cam1" include 60 clip che rappresentano comportamenti non violenti;
- "non-violent/cam2" include 60 clip con gli stessi comportamenti non violenti inclusi in "non-violent/cam1", ma ripresi da un'altra prospettiva;
- "violent/cam1" include 115 clip che rappresentano comportamenti violenti;
- "violent/cam2" include 115 clip con gli stessi comportamenti violenti in "non-violent/cam1", ma ripresi da un'altra prospettiva.

Tutte le clip sono state registrate nella stessa stanza, con condizioni di luce naturale, posizionando due telecamere in due punti diversi (l'angolo in alto a sinistra davanti alla porta della stanza e l'angolo in alto a destra sul lato della porta).

Le clip sono state eseguite da un gruppo di attori non professionisti, che variavano da 2 a 4 per clip. Per le clip violente, agli attori è stato chiesto di simulare azioni frequenti nelle risse, come calci, pugni, schiaffi, bastonate (colpire con un bastone), pugnalate e colpi di pistola. Per le clip non violente, agli attori è stato chiesto di simulare azioni che possono risultare in falsi

positivi mediante tecniche di rilevamento della violenza a causa della velocità dei movimenti o della somiglianza con azioni violente. Nello specifico, le clip non violente includono azioni come abbracciarsi, dare il cinque e applaudire, esultare e gesticolare. Tutte le azioni nelle clip violente e non violente sono state annotate manualmente. La specifica completa del dataset è disponibile in un documento dedicato [68], ad accesso aperto.

In termini di lunghezza media delle clip e numero totale di fotogrammi, il dataset proposto è più grande dei dataset generalmente utilizzati per il confronto delle tecniche di rilevamento della violenza, come i dateset Hockey Fight e Crowd Violence. Tuttavia, sembra relativamente piccolo se confrontato con altri dataset utilizzati nella visione artificiale e nella classificazione dei video, come Sports-1M. Sebbene sia possibile estrarre più video da Internet, con un pesante lavoro manuale di annotazione, il dataset proposto è concepito per testare specificamente la robustezza contro i falsi positivi. Pertanto, il dataset deve essere appositamente progettato per raggiungere tale obiettivo.

6.3.5 Modelli proposti

In questo articolo proponiamo tre diversi modelli basati sul deep learning per classificare il rilevamento della violenza nei video e testiamo la loro accuratezza sul dataset AIRTLab. Nello specifico:

1. il primo modello consiste in C3D, usato come estrattore di caratteristiche, e una SVM lineare per classificare le clip in violente e non;
2. il secondo modello utilizza C3D come estrattore di caratteristiche, ma la classificazione è eseguita grazie a due strati completamente connessi, componendo un modello end-to-end;
3. il terzo modello è addestrato da zero e si basa su ConvLSTM.

In due dei tre modelli proposti, la rete C3D addestrata sul dataset Sports-1M viene utilizzata come estrattore di caratteristiche, applicando un approccio di tipo transfer learning. In effetti, il transfer learning può ottenere una migliore generalizzazione rispetto a una addestramento da zero e prevenire l'overfitting [211, 212]. Nella definizione originale di Tran et al. [5], C3D utilizza un kernel 3x3x3 (con passo uguale a 1) in un totale di otto strati convoluzione alternati a cinque strati di pooling , seguito da due strati completamente connessi e uno strato di output softmax per calcolare la distribuzione di probabilità tra le categorie di sport. Tutti i neuroni negli strati di convoluzione utilizzano la funzione di attivazione lineare rettificata (ReLu). La tabella 6.4 elenca gli strati di C3D usati in questo lavoro: abbiamo preso tutti gli strati convoluzionali e di pooling originali e preso l'output del primo strato completamente connesso (chiamato "fc6" dagli autori di C3D) come descrittore dell'input originale,

Tabella 6.4: Strati di C3D usati come estrattore di caratteristiche in due dei modelli proposti. Abbiamo usato il C3D fino al primo strato completamente connesso (cioè denso), chiamato dai suoi autori "fc6" [5].

Strato	**Architettura**	**Output Shape**	**Params #**
Conv3D	64 filters, 3x3x3 (stride 1), ReLu	(16, 112, 112, 64)	5248
MaxPooling3D	1x2x2	(16, 56, 56, 64)	0
Conv3D	128 filters, 3x3x3 (stride 1), ReLu	(16, 56, 56, 128)	221312
MaxPooling3D	2x2x2	(8, 28, 28, 128)	0
Conv3D	256 filters, 3x3x3 (stride 1), ReLu	(8, 28, 28, 256)	884992
Conv3D	256 filters, 3x3x3 (stride 1), ReLu	(8, 28, 28, 256)	1769728
MaxPooling3D	2x2x2	(4, 14, 14, 256)	0
Conv3D	512 filters, 3x3x3 (stride 1), ReLu	(4, 14, 14, 512)	3539456
Conv3D	512 filters, 3x3x3 (stride 1), ReLu	(4, 14, 14, 512)	7078400
MaxPooling3D	2x2x2	(2, 7, 7, 512)	0
Conv3D	512 filters, 3x3x3 (stride 1), ReLu	(2, 7, 7, 512)	7078400
Conv3D	512 filters, 3x3x3 (stride 1), ReLu	(2, 7, 7, 512)	7078400
ZeroPadding3D	(0, 0), (0, 1), (0, 1)	(2, 8, 8, 512)	0
MaxPooling3D	2x2x2	(1, 4, 4, 512)	0
Flatten	-	(8192)	0
Dense	4096 units, ReLu	(4096)	33558528

rimuovendo il secondo strato completamente connesso e lo strato softmax finale. C3D prende come input sequenze di 16 fotogrammi con una risoluzione di 112x112 pixel. Pertanto, abbiamo mantenuto questo formato di input per tutti i modelli di rilevamento della violenza proposti in questo documento.

La figura 6.6 mostra lo schema dei modelli proposti. Le clip fornite come input ai tre modelli sono suddivise in blocchi di 16 fotogrammi e ridimensionate alla risoluzione di 112x112 pixel, per essere conformi all'input C3D.

Nel primo dei tre modelli proposti (Figura 6.6a), il descrittore di caratteristiche di 4096 elementi fornito in uscita dal primo strato completamente connesso di C3D alimenta una SVM, con kernel lineare e $C = 1$, per classificare la sequenza di 16 fotogrammi come violenta o meno. Nel nostro lavoro precedente [4], abbiamo già dimostrato la capacità di questo modello sull'Hockey Fight e il Crowd Violence, ottenendo una precisione 98.51% e 99.29% rispettivamente.

La tabella 6.5 mostra l'architettura del secondo modello proposto. Diversamente dal modello precedente, abbiamo costruito un'architettura end-to-end, estendendo con strati aggiuntivi la porzione di C3D utilizzata come estrattore di caratteristiche (Figura 6.6b). Nello specifico, abbiamo aggiunto uno strato di dropout, con un tasso di 0,5, per prevenire l'overfitting [213]. Successivamente abbiamo aggiunto uno strato completamente connesso con 512 neuroni utilizzando la funzione di attivazione lineare rettificata. Dopo un altro dropout di 0,5, lo strato finale composto da un neurone con l'attivazione sigmoide esegue l'effettiva classificazione delle clip da 16 fotogrammi in violente o meno. Anche in questo modello, gli strati di C3D utilizzati sono addestrati sul dataset

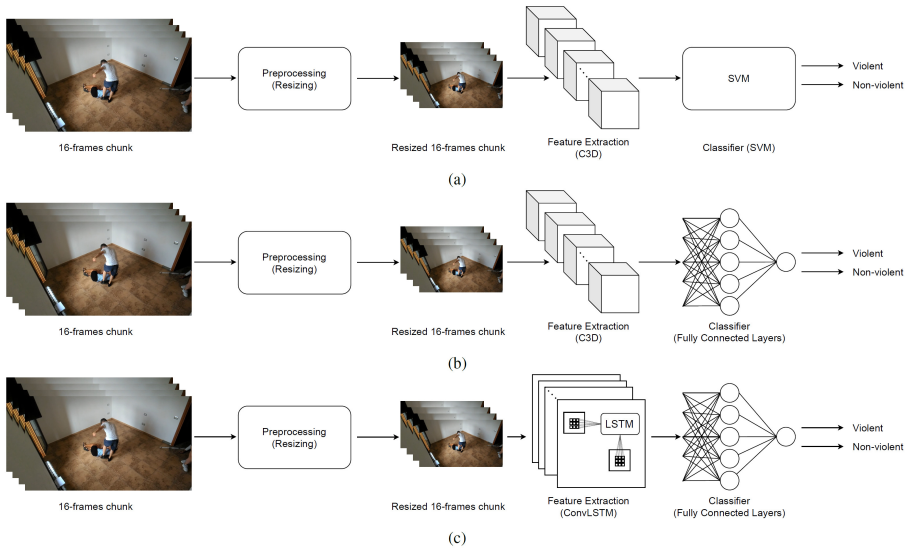


Figura 6.6: Lo schema dei tre modelli proposti in questo articolo. Tutti i modelli elaborano sequenze composte da 16 fotogrammi ridimensionati a $112 \times 112 pixel$. Il primo modello proposto (a) usa la rete C3D pre-addestrata come estrattore di caratteristiche e un classificatore SVM per etichettare le sequenze come violente o no. Il secondo modello (b) usa anch'esso la C3D come estrattore di caratteristiche, e il classificatore è composto da strati completamente connessi. Il terzo modello (c) usa uno strato ConvLSTM addestrato da zero, con strati completamente connessi per la classificazione finale.

Tabella 6.5: Il secondo modello proposto. È un modello end-to-end che aggiunge due strati completamente connessi al C3D (fino a "fc6"). C3D non viene addestrato di nuovo, quindi il numero totale di parametri addestrati è 2.098.177 che sono i pesi degli strati finali completamente connessi.

Strato	**Architettura**	**Output Shape**	**Params #**
C3D until "fc6"	(see Table 6.4)	(4096)	61214464
Dropout	0.5 rate	(4096)	0
Dense	512 units, ReLu	(512)	2097664
Dropout	0.5 rate	(512)	0
Dense	1 unit, Sigmoid	(1)	513

Tabella 6.6: Il terzo modello proposto. È un modello end-to-end basato sull'architettura ConvLSTM. È addestrato da zero e il numero totale di parametri addestrati è 198.401.537.

Strato	**Architettura**	**Output Shape**	**Params #**
ConvLSTM2D	64 filters, 3x3	(110, 110, 64)	154624
Dropout	0.5 rate	(110, 110, 64)	0
Flatten	-	(774400)	0
Dense	256 units, ReLu	(256)	198246656
Dropout	0.5 rate	(256)	0
Dense	1 unit, Sigmoid	(1)	257

Sports-1M. Invece, gli strati aggiunti sono stati addestrati da zero sulle clip disponibili, come spiegato nella Sezione *Valutazione Sperimentale*.

Il terzo modello proposto (Figura 6.6c) è basato sull'architettura ConvLSTM ed è addestrato in modalità end-to-end da zero. Gli strati sono elencati nella Tabella 6.6. Il primo strato è una ConvLSTM composta da 64 filtri 3x3, con un totale di 154.624 parametri addestrabili. Dopo un dropout di 0,5 per prevenire l'overfitting, appiattiamo l'output della ConvLSTM e aggiungiamo uno strato completamente connesso con 256 neuroni, utilizzando la funzione di attivazione lineare rettificata. Infine, dopo altri 0,5 dropout, la classificazione finale in sequenza violenta o no viene calcolata da un neurone con funzione di attivazione sigmoide. Per consentire un confronto con i modelli basati su C3D, anche l'input della rete basata su ConvLSTM è composto da sequenze di 16 frame video alla risoluzione di 112x112 pixel.

Per dimostrare l'importanza dei modelli proposti, abbiamo effettuato un confronto con le prestazioni di CNN 2D ben note, vale a dire VGG16, VGG19, ResNet50V2, Xception, e NASNet mobile, pre-addestrate sul database ImageNet [214]. La figura 6.7 descrive lo schema dei modelli basati su tali CNN 2D. Per essere applicate ai video ed elaborare le relative informazioni spazio-temporali, le CNN 2D sono distribuite nel tempo sui 16 frame che compongono un blocco di input e combinate con uno strato ricorrente, mentre due strati completamente connessi implementano la classificazione finale. La ConvLSTM e la Bidirectional-LSTM (Bi-LSTM) sono stati entrambi testati come strato ricorrente. In particolare, la tabella 6.7 include gli strati che compongono i modelli formati dalle CNN 2D pre-addestrate e la ConvLSTM. Lo strato ConvLSTM è composto da 64 filtri 3x3, seguito da uno strato completamente connesso con 256 neuroni ReLu, un dropout 0,5 e un neurone completamente connesso con attivazione sigmoide per eseguire la classificazione finale. La tabella 6.8 descrive gli strati che compongono i modelli con le CNN 2D pre-addestrate e la Bi-LSTM come modulo ricorrente. La Bi-LSTM è composta da 128 unità

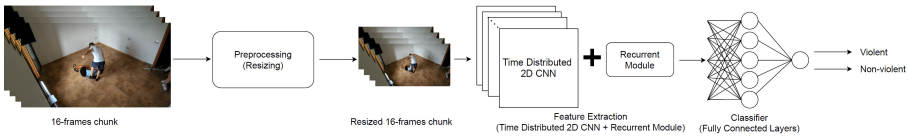


Figura 6.7: La rappresentazione schematica dei modelli basati su CNN 2D, sviluppati per confrontare i modelli proposti con le prestazioni di note CNN 2D pre-addestrate su ImageNet, come VGG16, VGG19, e ResNet50. Per applicarle ai video, le CNN 2D sono state distribuite nel tempo su spezzoni di 16 fotogrammi usati come input e combinate a strati ricorrenti (ConvLSTM e Bi-LSTM).

Tabella 6.7: Il modello basato su CNN 2D pre-addestrate e ConvLSTM. La ConvLSTM e due strati completamente connessi sono stati aggiunti a ben note CNN 2D (VGG16, VGG19, ResNet50V2, Xception, NASNet Mobile), pre-addestrate su ImageNet. Le CNN 2D sono state distribuite nel tempo per essere applicate a un input 3D, cioè i video dei dataset. Si noti che il numero di parametri dello strato ConvLSTM dipende dalla precedente CNN 2D.

Strato	**Architettura**	**Output Shape**	**Params #**
Time Distr. 2D CNN	-	-	-
ConvLSTM2D	64 filters, 3x3	(5, 5, 64)	-
Flatten	-	(1600)	0
Dense	256 units, ReLu	(256)	409856
Dropout	0.5 rate	(256)	0
Dense	1 unit, Sigmoid	(1)	257

nascoste, seguite da un dropout 0,5, uno strato completamente connesso con 128 neuroni ReLu, un altro dropout 0,5 e un neurone sigmoideo completamente connesso per la classificazione finale.

Con i modelli basati sulle CNN 2D pre-addestrate, i frame di input sono stati ridimensionati a 224x224 pixel anziché 112x112. In effetti, la maggior parte delle CNN 2D testate utilizza 224x224 come dimensione di input predefinita; inoltre, una dimensione di input di 112x112 ha portato a una accuratezza significativamente inferiore con le CNN 2D pre-addestrate.

6.3.6 Valutazione Sperimentale

Abbiamo valutato i modelli di deep learning proposti raccogliendo i risultati della classificazione sul dataset AIRTLab, oltre ai test sui dataset Hockey Fight e Crowd Violence. L'obiettivo è duplice: da un lato, vogliamo confrontare l'accuratezza dei nostri modelli nell'individuare scene violente; dall'altro, vogliamo creare un benchmark con metriche di base sul dataset proposto e convalidare il

Tabella 6.8: Il modello basato su CNN 2D pre-addestrate e Bi-LSTM. La Bi-LSTM e due strati completamente connessi sono stati aggiunti a ben note CNN 2D (VGG16, VGG19, ResNet50V2, Xception, NASNet Mobile), pre-addestrate su ImageNet. Le CNN 2D sono state distribuite nel tempo per essere applicate a un input 3D, cioè i video dei dataset. Si noti che la forma dell'output dello strato flatten distribuito nel tempo e il numero di parametri della Bi-LSTM dipendono dalla precedente CNN 2D.

Strato	**Architettura**	**Output Shape**	**Params #**
Time Distr. 2D CNN	-	-	-
Time Distr. Flatten	-	-	0
Bi-LSTM	128 units	(256)	-
Dropout	0.5 rate	(256)	0
Dense	128 units, ReLu	(128)	32896
Dropout	0.5 rate	(128)	0
Dense	1 units, Sigmoid	(1)	129

suo design inteso a verificare la robustezza della tecniche di rilevamento della violenza contro i falsi positivi. A tal fine, nei paragrafi seguenti presentiamo il *setup sperimentale e metriche di valutazione* ed i risultati della valutazione in *risultati e discussione*. Naturalmente, i risultati ottenuti presentano alcune limitazioni, come spiegato nella sottosezione *Limitazioni*.

6.3.7 Setup sperimentale e metriche di valutazione

Abbiamo testato i tre modelli proposti sul dataset AIRTLab, applicando uno schema di convalida incrociata stratificato a mescolamento randomico (stratified shuffle split). A tal fine, si è ripetuta una divisione 80-20 randomizzata per cinque volte, usando l'80% dei dati come insieme di addestramento, e il 20% come l'insieme di test, preservando la percentuale di campioni da ciascuna classe, in ogni gruppo. La suddivisione dei dati è identica per tutti i modelli testati, per implementare un confronto equo. Dato che gli input per i modelli sono sequenze composte da 16 fotogrammi e le clip nel dataset includono un totale di 3537 di tali sequenze, sono stati utilizzati 2829 campioni (cioè blocchi di 16 fotogrammi) per l'addestramento e 708 per i test, in ogni suddivisione della convalida incrociata. Il 12.5% dei dati di addestramento, vale a dire il 10% dell'intero dataset, è stato usato come insieme di validazione per l'addestramento delle due reti neurali end-to-end basate su C3D e ConvLSTM. Inoltre, per confrontare i modelli proposti con la letteratura sull'individuazione della violenza presentata nella Sezione *Stato dell'arte*, abbiamo usato la stessa convalida incrociata con divisione 80-20 per testare sui dataset Hockey Fight e Crowd Violence. Infine, abbiamo confrontato i risultati dei modelli proposti

Tabella 6.9: Numero di epoche di addestramento in ogni divisione (S1-S5) del dataset AIRTLab per i due modelli end-to-end cioè C3D con due strati completamente connessi (C3D + FC) e l'architettura basata su ConvLSTM.

	S1	**S2**	**S3**	**S4**	**S5**	**Mean**
C3D + FC	19	26	21	30	21	23.40 ± 4.03
ConvLSTM	10	8	6	15	8	9.40 ± 3.07

sui dataset AIRTLab, Hockey Fight e Crowd Violence, con quelli ottenuti dai modelli basati su CNN 2D pre-addestrate. Anche per questi test, abbiamo utilizzato la stessa convalida incrociata stratificata 80-20 utilizzata per misurare le prestazioni dei modelli proposti.

Per i due modelli end-to-end, abbiamo utilizzato l'ottimizzatore Adam per ridurre al minimo la funzione di loss "Binary Cross Entropy" durante lo addestramento delle reti neurali. Il numero di epoche di addestramento è variato per ogni suddivisione, poiché abbiamo interrotto anticipatamente l'addestramento dopo 5 epoche senza un miglioramento della loss minima di convalida, ripristinando i pesi corrispondenti alla migliore loss sull'insieme di validazione. La tabella 6.9 mostra il numero di epoche di addestramento per ogni divisione del dataset AIRTLab. Il numero medio di epoche di addestramento è 23,4 (± 4,03) per il modello basato su C3D e gli strati completamente connessi, e 9.4 (± 3.07) per la rete basata su ConvLSTM. Il batch size per l'addestramento è di 32 campioni per il modello basato su C3D e 8 per il modello basato su ConvLSTM.

Come evidenziato nella sezione 6.3, due notebook Jupyter con gli esperimenti descritti sono disponibili in un repository pubblico GitHub, al fine di garantire la riproducibilità dei test. I test sono stati eseguiti su Google Colab nel il runtime con GPU, utilizzando Keras 2.4.3, TensorFlow 2.4.1 e scikit-learn 0.22.2.post1.

Etichettando come 0 (negativo) i frammenti di 16 fotogrammi delle clip non violente e come 1 (positivo) i frammenti delle clip violente, abbiamo calcolato le seguenti metriche sull'insieme di test in ogni suddivisione dello schema di convalida incrociata:

- sensibilità (True Positive Rate – TPR), cioè la porzione dei positivi correttamente identificata (tra tutti i positivi disponibili nell'insieme di test);

- specificità (True Negative Rate – TNR), cioè la porzione dei negativi correttamente identificata (tra tutti i negativi disponibili nell'insieme di test);

- l'accuratezza, cioè i campioni correttamente identificati (tra tutti i campioni disponibili nell'insieme di test);
- lo score F_1, cioè la media armonica di precisione (il rapporto fra i positivi correttamente identificati e tutti i positivi individuati) e sensibilità.

Queste metriche possono essere formulate in termini di veri positivi (TP), veri negativi (TN), falsi positivi (FP) e falsi negativi (FN) secondo le seguenti equazioni, che consentono di determinare la *specificity* o TNR (True Negative Rate) e di conseguenza anche il FPR (False Positive Rate):

$$sensitivity = \frac{TP}{TP + FN} \tag{6.1}$$

$$specificity = \frac{TN}{TN + FP} \tag{6.2}$$

$$FPR = 1 - specificity \tag{6.3}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.4}$$

$$scoreF_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{6.5}$$

Inoltre, in ogni suddivisione, abbiamo calcolato la curva Receiver Operating Characteristic (ROC) e l'area sotto la curva (AUC), mostrando il TPR contro il tasso di falsi positivi ($FPR = 1 - TNR$) al variare della soglia di classificazione, per comprendere il capacità diagnostica di ciascun modello. Infine, per ogni modello end-to-end (ovvero tutti i modelli proposti tranne il modello C3D + SVM), riportiamo anche il valore della funzione di loss Binary Cross Entropy calcolata sull'insieme di test.

6.3.8 Risultati e discussione

Analizziamo in questa sottosezione le metriche ottenute dai tre modelli proposti sul dataset AIRTLab, nonché sui dataset Hockey Fight e Crowd Violence (per confrontarli con la letteratura esistente). Infine, nell'ultima parte di questa sottosezione, presentiamo i risultati ottenuti dalle CNN 2D pre-addestrate, al fine di evidenziare l'importanza dei tre modelli proposti.

6.3.9 Test sul dataset AIRTLab

Per ciascuno dei modelli proposti riportiamo i risultati ottenuti su ogni suddivisione del dataset AIRTLab nelle tabelle 6.10, 6.11, e 6.12, in aggiunta alla

Tabella 6.10: I risultati del modello composto da C3D e SVM, calcolati per ogni split dello schema di validazione incrociata stratificata shuffle-split, sul dataset AIRTLab.

	Split 1	**Split 2**	**Split 3**	**Split 4**	**Split 5**
Sensitivity	**97.90%**	97.06%	**97.90%**	95.80%	96.64%
Specificity	93.53%	92.24%	**96.12%**	95.69%	93.10%
FPR	6.47%	7.76%	**3.88%**	4.31%	6.90%
Accuracy	96.47%	95.48%	**97.32%**	95.76%	95.48%
F_1 score	97.39%	96.65%	**98.00%**	96.82%	96.64%
AUC	99.44%	98.89%	**99.46%**	99.45%	99.15%

media calcolata per tutte le metriche (Tabella 6.13). I risultati su ciascuna suddivisione consentono di comprendere fino a che punto il modello è stabile e indipendente da una particolare suddivisione dei dati.

La tabella 6.10 mostra le metriche calcolate sul dataset AIRTLab per il modello composto da C3D e il classificatore SVM, in ognuna delle suddivisioni calcolate con la validazione incrociata. La specificità è inferiore alla sensibilità in ogni gruppo, essendo compresa tra 92.24% nella seconda suddivisione e 96.12% nella terza suddivisione. Questi risultati confermano che la maggior parte degli errori sono nella classe non violenta, con il classificatore che fornisce alcuni falsi positivi in output. Ad esempio, nella seconda suddivisione, ci sono 18 falsi positivi, con 214 su 232 sequenze di 16 fotogrammi non violente classificate correttamente come non violente. Invece, nella stessa suddivisione, 462 su 476 sequenze violente sono correttamente etichettate come violente, con solo 14 falsi negativi. Naturalmente, questi risultati potrebbero essere parzialmente influenzati dal fatto che le due classi sono sbilanciate nel dataset di AIRTLab. L'accuratezza è in linea con il nostro lavoro precedente, ottenendo un 97.32% nella migliore suddivisione.

La tabella 6.11 elenca i risultati sul dataset AIRTLab ottenuti dal modello composto da C3D e due strati completamente connessi. L'andamento delle metriche è simile al modello che usa C3D e il classificatore SVM. Tuttavia, con C3D e gli strati completamente collegati, la differenza tra la sensibilità e la specificità è maggiore rispetto al modello precedente. Ad esempio, nella quarta suddivisione, che ha la più alta differenza (98.74% di sensibilità, 87.93% di specificità), 204 sequenze non violente sono correttamente classificate, mentre ci sono 28 falsi positivi. Invece, 470 sequenze violenti su 476 sono classificate correttamente. Una differenza significativa tra sensibilità e specificità è visibile anche nelle altre suddivisioni dello schema di convalida incrociata. In effetti, l'AUC di questo modello è leggermente inferiore al modello basato su C3D e sul classificatore SVM. Inoltre, lo split 1 e lo split 4 presentano il maggior numero di falsi positivi e la minore accuratezza come confermato dal valore di loss,

Tabella 6.11: I risultati del modello composto da C3D e dagli strati completamente connessi per la classificazione, calcolati per ogni split dello schema di validazione incrociata stratificata shuffle-split, sul dataset AIRTLab.

	Split 1	**Split 2**	**Split 3**	**Split 4**	**Split 5**
Loss	0.1471	**0.0996**	0.1135	0.1358	0.1113
Sensitivity	97.90%	98.32%	97.27%	**98.74%**	96.85%
Specificity	89.66%	92.24%	92.24%	87.93%	**93.53%**
FPR	10.34%	7.76%	7.76%	12.07%	**6.47%**
Accuracy	95.20%	**96.33%**	95.62%	95.20%	95.76%
F_1 score	96.48%	**97.30%**	96.76%	96.51%	96.85%
AUC	98.32%	**99.21%**	99.05%	98.98%	99.08%

Tabella 6.12: I risultati del modello basato sull'architettura ConvLSTM, calcolati per ogni split dello schema di convalida incrociata stratificata shuffle-split, sul dataset AIRTLab.

	Split 1	**Split 2**	**Split 3**	**Split 4**	**Split 5**
Loss	0.1041	0.1004	**0.0511**	0.0759	0.0576
Sensitivity	97.48%	**99.58%**	98.53%	97.90%	97.90%
Specificity	91.38%	90.09%	97.41%	**97.84%**	97.41%
FPR	8.62%	9.91%	2.59%	**2.16%**	2.59%
Accuracy	95.48%	96.47%	**98.16%**	97.88%	97.74%
F_1 score	96.67%	97.43%	**98.63%**	98.42%	98.31%
AUC	99.40%	99.47%	**99.83%**	99.77%	**99.83%**

maggiore degli altri split.

La tabella 6.12 mostra i risultati ottenuti dal modello basato su ConvLSTM sul dataset AIRTLab. La differenza tra sensibilità e specificità dipende molto di più dalla suddivisione dei dati rispetto ai due modelli precedenti, dimostrando che la ConvLSTM potrebbe avere troppi parametri data la quantità di dati di addestramento. Infatti, nei primi due split, questa differenza è significativa, essendo intorno al 6% e al 9%. Invece, nel terzo, quarto e quinto split, la sensibilità e la specificità sono molto più vicini, e il modello sembra robusto sia ai falsi positivi che ai falsi negativi. I valori di loss sull'insieme di test sono leggermente inferiori a quelli ottenuti dal modello composto da C3D e dai layer completamente connessi, evidenziando che la ConvLSTM tende ad adattarsi al dataset (infatti il modello viene addestrato da zero).

I risultati sul dataset AIRTLab sono riassunti nella tabella 6.13, che comprende un confronto tra i tre modelli, mostrando la media e la deviazione standard di sensibilità, specificità, accuratezza, score F_1, e AUC, oltre alla loss dei due modelli end-to-end. In termini di accuratezza, score F_1 e AUC, il modello basato su ConvLSTM è leggermente migliore rispetto agli altri, con rispettivamente

Tabella 6.13: I valori medi delle metriche calcolate sul dataset AIRTLab sulle cinque suddivisioni dello shuffle stratificato, per ciascuno dei modelli proposti.

	Loss	**Sensitivity**	**Specificity**
C3D + SVM	-	97.06 ± 0.80%	94.14 ± 1.51%
C3D + FC	0.1215 ± 0.0174	97.82 ± 0.69%	91.12 ± 2.03%
ConvLSTM	**0.0779 ± 0.0215**	**98.28 ± 0.73%**	**94.83 ± 3.37%**
	Accuracy	**F_1 score**	**AUC**
C3D + SVM	96.10 ± 0.71%	97.10 ± 0.53%	99.30 ± 0.23%
C3D + FC	95.62 ± 0.42%	96.78 ± 0.30%	98.94 ± 0.31%
ConvLSTM	**97.15 ± 1.02%**	**97.89 ± 0.74%**	**99.67 ± 0.19%**

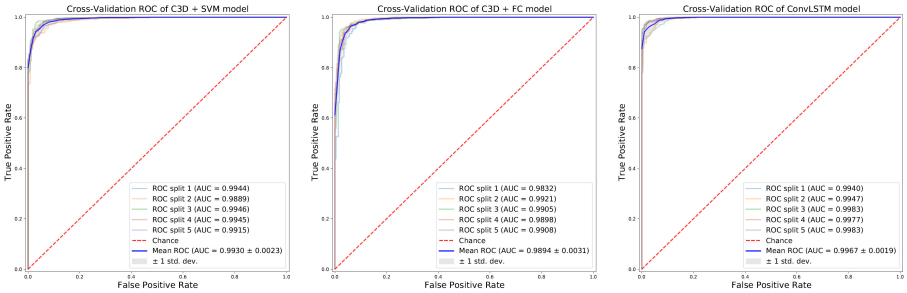


Figura 6.8: Curva ROC e AUC per i modelli C3D + SVM (a), C3D + FC (b) e ConvLSTM (c), sul dataset AIRTLab.

97.15%, 97.89% e 99.67%. Tuttavia, vale la pena notare che questo è l'unico modello addestrato da zero sul dataset AIRTLab e, pertanto, potrebbe soffrire di overfitting sul dataset, anche se la convalida incrociata e l'early stopping dovrebbero limitare il fenomeno. Ciò è confermato anche dalla loss più bassa sull'insieme di test. Gli altri due modelli mostrano un'accuratezza leggermente inferiore, ma, essendo basati su una metodologia di transfer learning, i loro risultati possono essere interpretati come più generali. Il modello composto da C3D e SVM è quello con la differenza minore tra la sensibilità (97.06%) e specificità (94.14%), mostrando risultati stabili sia con video violenti che con non violenti. Il modello composto C3D e gli strati completamente connessi presenta le metriche più basse (solo la sensibilità è leggermente migliore del classificatore SVM). Tuttavia, i tre modelli mostrano una capacità diagnostica simile nell'identificazione dei video violenti, esibendo curve ROC e AUC molto simili, come evidenziato in Fig. 6.8.

La sensibilità è maggiore della specificità per tutti e tre i modelli, convalidando lo scopo del dataset proposto. Infatti, i video non violenti contengono movimenti veloci e contatti tra i soggetti con l'obiettivo di testare la robustez-

za delle tecniche di rilevamento della violenza, pur conservando la capacità di identificare scene violente.

6.3.10 Test sui dataset Hockey Fight e Crowd Violence

Per confrontare i modelli proposti con quelli disponibili in letteratura e descritti nella sezione *Stato dell'arte*, abbiamo effettuato anche i test sui dataset Hockey Fight e Crowd Violence. A tal fine, la tabella 6.14 elenca i risultati medi ottenuti dai tre modelli sull'Hockey Fight, mentre la tabella 6.15 include i risultati sul Crowd Violence. Il modello basato su C3D e SVM conferma il buon andamento mostrato nel nostro lavoro precedente: l'accuratezza è di circa 98% sul dataset Hockey Fight e soprattutto 99% sul Crowd Violence. Il modello end-to-end composto da C3D e due strati completamente connessi ottiene risultati simili: l'accuratezza è quasi del 97% su Hockey Fight, e del 99% su Crowd Violence. Infatti, il Crowd Violence è il più piccolo dataset testato: include 1265 frammenti di 16 fotogrammi, mentre l'Hockey Fight e l'AIRTLab includono rispettivamente 2007 e 3537 frammenti. Pertanto, nonostante la strategia di suddivisione stratificata 80-20, i modelli potrebbero mostrare overfit sul Crowd Violence. Il modello basato su ConvLSTM si comporta in modo simile agli altri modelli sul dataset Hockey Fight, ottenendo una precisione media del 96,57%. Tuttavia, su Crowd Violence, l'accuratezza dei modelli basati su ConvLSTM scende all'84,19%. Tale modello ha più di 198 milioni di parametri, e potrebbe essere troppo complesso per convergere su un dataset relativamente piccolo come il Crowd Violence, come evidenziato anche dal valore di loss calcolato sul test set, che è significativamente maggiore di loss del modello composto da C3D e strati completamente connessi. Inoltre, la maggior parte degli errori della ConvLSTM sono falsi positivi: sul Crowd Violence, la specificità media è del 69.16% mentre la sensibilità media è del 95.21%. Infatti, i video del Crowd Violence sono abbastanza simili sulle due classi e, a causa del minor numero di campioni, la ConvLSTM fatica a convergere.

Pertanto, sui dataset Hockey Fight e Crowd Violence, il modello basato su C3D e SVM ottiene le migliori prestazioni. Inoltre, i due modelli basati su C3D si sono dimostrati in grado di eseguire la classificazione su diversi dataset, confermando la capacità di generalizzazione delle metodologie di transfer learning. I risultati dettagliati su Hockey Fight e Crowd Violence sono disponibili rispettivamente in *Appendice A* e *Appendice B*, includendo le metriche su ogni split dei dataset.

I risultati sperimentali raccolti sui dataset Hockey Fight e Crowd Violence permettono di confrontare i nostri modelli con i lavori di ricerca correlati descritti nella Sezione *Stato dell'arte*. A tal fine, abbiamo riportato la loro accuratezza nella Tabella 6.3. Mentre le CNN 3D addestrate da zero proposte da

Tabella 6.14: I valori medi delle metriche calcolate sul dataset Hockey Fight sulle cinque suddivisioni dello shuffle stratificato, per ciascuno dei modelli proposti.

	Loss	**Sensitivity**	**Specificity**
C3D + SVM	-	**97.82 ± 0.80%**	**97.90 ± 1.24%**
C3D + FC	**0.1276 ± 0.0662**	96.93 ± 1.75%	96.40 ± 0.97%
ConvLSTM	0.1492 ± 0.0839	96.44 ± 1.19%	96.70 ± 1.54%
	Accuracy	**F_1 score**	**AUC**
C3D + SVM	**97.86 ± 0.56%**	**97.87 ± 0.55%**	**99.62 ± 0.30%**
C3D + FC	96.67 ± 1.15%	96.69 ± 1.16%	99.27 ± 0.40%
ConvLSTM	96.57 ± 0.79%	96.58 ± 0.78%	99.31 ± 0.32%

Tabella 6.15: I valori medi delle metriche calcolate sul set di dati Crowd Violence sulle cinque suddivisioni dello shuffle stratificato, per ciascuno dei modelli proposti.

	Loss	**Sensitivity**	**Specificity**
C3D + SVM	-	**100.00 ± 0.00%**	**99.07 ± 1.18%**
C3D + FC	**0.0356 ± 0.0323**	99.59 ± 0.55%	98.32 ± 0.70%
ConvLSTM	0.3535 ± 0.0726	95.21 ± 1.37%	69.16 ± 15.15%
	Accuracy	**F_1 score**	**AUC**
C3D + SVM	**99.60 ± 0.50%**	**99.66 ± 0.43%**	**100.00 ± 0.01%**
C3D + FC	99.05 ± 0.54%	99.18 ± 0.46%	99.94 ± 0.11%
ConvLSTM	84.19 ± 5.95%	87.63 ± 3.89%	94.43 ± 2.15%

Song et al. [203] e Li et al. [64] hanno raggiunto un'ottima accuratezza sul dataset Hockey Fight (99.6% e 98.3% rispettivamente), hanno ottenuto risultati inferiori sul dataset Crowd Violence (94.3% e 97.2%). Invece, il nostro modello basato sul transfer learning con C3D combinato con SVM ottiene un'eccellente performance di accuratezza su entrambi i dataset, con il 97.9% sul Hockey Fight e il 99.6% sul Crowd Violence, confermando i risultati del nostro precedente lavoro [4]. Anche l'altro modello proposto basato su C3D ha buoni risultati su entrambi i dataset, con il 96.7% su Hockey Fight e il 99% su Crowd Violence. Tuttavia, il modello CNN 3D proposto da Li et al. ha meno parametri del modello C3D che abbiamo usato e, quindi, richiede meno risorse computazionali per rilevare la violenza nei video. In modo simile al nostro lavoro, Ullah et al. [65] hanno proposto di utilizzare il transfer learning con C3D, ma usano l'output del secondo strato completamente connesso ("fc7") come descrittore di caratteristiche invece del primo ("fc6") come abbiamo fatto nel nostro lavoro. La loro accuratezza (96% su Hockey Fight e 98% su Crowd Violence) è leggermente inferiore a quella ottenuta dai nostri modelli basati su C3D. Inoltre, il nostro modello basato su C3D e SVM è più equilibrato su entrambi i dataset rispetto alle reti multi-stream, come quella proposta da Sudhakaran e Lanz [67] che utilizza CNN 2D pre-addestrate (AlexNet [215]) per elaborare i video frame per frame e di una ConvLSTM addestrata da zero per rilevare la violenza nella sequenza dei frame. Infatti, hanno ottenuto il 97.1% di accuratezza sul dataset Hockey Fight, e il 94.5% sul dataset Crowd Violence. Inoltre, la rete multi-stream basata su VGG13 pre-addestrata e sulla ConvLSTM proposta da Hanson et al. [205] ottiene punteggi simili ai nostri modelli sull'Hockey Fight (98,1% di accuratezza), ma è ancora lontana dai nostri risultati sul dataset Crowd Violence (96,3%).

Pertanto, i nostri modelli confermano l'efficacia delle architetture basate sul transfer learning per il rilevamento della violenza, anche rispetto alla letteratura esistente.

6.3.11 Confronto con modelli basati su CNN 2D pre-addestrate

Per dimostrare l'efficacia dei tre modelli proposti in questo articolo, confrontiamo le loro prestazioni con le metriche ottenute da cinque CNN 2D ben note, pre-addestrate su ImageNet e combinate con uno strato ricorrente per l'estrazione delle caratteristiche. Le CNN 2D pre-addestrate sono VGG16, VGG19, ResNet50V2, Xception, e NASNet Mobile. Come spiegato nella sottosezione 6.3.5, abbiamo costruito cinque modelli combinando queste CNN 2D con uno strato Bi-LSTM (e due strati completamente connessi per la classificazione finale) e altri cinque modelli combinando le CNN 2D con uno strato

Tabella 6.16: I valori medi delle metriche calcolate sul dataset AIRTLab sulle cinque suddivisioni dello shuffle stratificato, per i modelli composti da CNN 2D pre-addestrate e Bi-LSTM per l'estrazione delle caratteristiche.

	Loss	**Sensitivity**	**Specificity**
VGG16 + Bi-LSTM	**0.1314 ± 0.0142**	96.93 ± 0.90%	**90.78 ± 2.34%**
VGG19 + Bi-LSTM	0.3554 ± 0.0910	94.03 ± 3.30%	71.63 ± 15.52%
ResNet50V2 + Bi-LSTM	0.6331 ± 0.0006	**100.00 ± 0.00%**	0.00 ± 0.00%
Xception + Bi-LSTM	0.6298 ± 0.0034	**100.00 ± 0.00%**	0.00 ± 0.00%
NASNet + Bi-LSTM	0.3776 ± 0.0461	91.85 ± 3.27%	64.48 ± 17.73%

	Accuracy	**F_1 score**	**AUC**
VGG16 + Bi-LSTM	**94.92 ± 0.51%**	**96.25 ± 0.36%**	**98.91 ± 0.19%**
VGG19 + Bi-LSTM	86.69 ± 6.07%	90.57 ± 3.97%	92.12 ± 3.44%
ResNet50V2 + Bi-LSTM	67.23 ± 0.00%	80.41 ± 0.00%	51.07 ± 1.67%
Xception + Bi-LSTM	67.23 ± 0.00%	80.41 ± 0.00%	55.47 ± 4.11%
NASNet + Bi-LSTM	82.88 ± 3.92%	87.93 ± 2.11%	90.41 ± 2.42%

ConvLSTM (e due strati completamente connessi per la classificazione finale). Pertanto, oltre ai nostri tre modelli, abbiamo testato altri dieci modelli sui dataset AIRTLab, Hockey Fight, e Crowd Violence.

La tabella 6.16 include i valori medi delle metriche calcolate per i modelli composti dalle CNN 2D e dallo strato Bi-LSTM, testati sul dataset AIRTLab; invece, la tabella 6.17 elenca le metriche per i modelli composti dalle CNN 2D e dallo strato ConvLSTM. Tra i modelli basati sulle CNN 2D, quelli che usano VGG16 ottengono la migliore accuratezza: VGG16 e ConvLSTM hanno un'accuratezza media del 95,62% (± 0,56%) mentre VGG16 e Bi-LSTM ottiene il 94,92% (± 0,51%). Tuttavia, nessuno dei modelli basati su CNN 2D ha prestazioni migliori dei modelli proposti in questo articolo, sul dataset AIRTLab. Infatti, sia il modello C3D più SVM che il modello basato su ConvLSTM ottengono accuratezza, score F_1 e AUC migliori di tutti i modelli basati su CNN 2D. Anche il modello composto da C3D e strati completamente connessi ha la stessa accuratezza di VGG16 con ConvLSTM.

Tra gli altri modelli basati su CNN 2D, quello che utilizza ResNet50V2 fallisce completamente il training sul dataset AIRTLab, come evidenziato dall'alto

Tabella 6.17: I valori medi delle metriche calcolate sul dataset AIRTLab sulle cinque suddivisioni dello shuffle-split stratificato, per i modelli composti da CNN 2D pre-addestrate e ConvLSTM per l'estrazione delle caratteristiche.

	Loss	**Sensitivity**	**Specificity**
VGG16 + ConvLSTM	0.1169 ± 0.0201	97.48 ± 1.09%	91.81 ± 3.17%
VGG19 + ConvLSTM	**0.1105 ± 0.0221**	96.63 ± 1.39%	**93.01 ± 2.01%**
ResNet50V2 + ConvLSTM	0.6331 ± 0.0006	**100.00 ± 0.00%**	0.00 ± 0.00%
Xception + ConvLSTM	0.2450 ± 0.0344	95.13 ± 2.27%	80.17 ± 2.68%
NASNet + ConvLSTM	0.2972 ± 0.0192	91.93 ± 3.62%	79.05 ± 7.63%

	Accuracy	**F_1 score**	**AUC**
VGG16 + ConvLSTM	**95.62 ± 0.56%**	**96.77 ± 3.88%**	99.11 ± 0.24%
VGG19 + ConvLSTM	95.45 ± 0.85%	96.62 ± 2.01%	**99.14 ± 0.27%**
ResNet50V2 + ConvLSTM	67.23 ± 0.00%	80.41 ± 0.00%	51.07 ± 1.67%
Xception + ConvLSTM	90.23 ± 1.92%	92.89 ± 1.44%	95.61 ± 1.10%
NASNet + ConvLSTM	87.71 ± 1.17%	90.95 ± 0.91%	94.77 ± 0.48%

Tabella 6.18: I valori medi delle metriche calcolate sul dataset Hockey Fight sulle cinque suddivisioni dello shuffle-split stratificato, per i modelli composti da CNN 2D pre-addestrate e Bi-LSTM per l'estrazione delle caratteristiche.

	Loss	**Sensitivity**	**Specificity**
VGG16 + Bi-LSTM	**0.1389 ± 0.0280**	**96.63 ± 0.73%**	94.30 ± 2.06%
VGG19 + Bi-LSTM	0.1538 ± 0.0398	95.05 ± 1.66%	**94.90 ± 3.10%**
ResNet50V2 + Bi-LSTM	0.6906 ± 0.0061	46.93 ± 45.07%	59.40 ± 48.51%
Xception + Bi-LSTM	0.3767 ± 0.0456	86.14 ± 7.61%	87.50 ± 5.27%
NASNet + Bi-LSTM	0.3336 ± 0.0355	86.93 ± 3.71%	87.60 ± 3.87%

	Accuracy	**F_1 score**	**AUC**
VGG16 + Bi-LSTM	**95.47 ± 1.17%**	**95.55 ± 1.11%**	**98.75 ± 0.49%**
VGG19 + Bi-LSTM	94.98 ± 1.80%	95.02 ± 1.73%	98.42 ± 0.74%
ResNet50V2 + Bi-LSTM	53.13 ± 6.27%	36.95 ± 30.40%	53.45 ± 7.12%
Xception + Bi-LSTM	86.82 ± 3.08%	86.65 ± 3.66%	92.51 ± 1.75%
NASNet + Bi-LSTM	87.26 ± 1.69%	87.27 ± 1.77%	93.33 ± 1.26%

valore di loss (0,63). Infatti, il modello classifica erroneamente tutti i campioni negativi, etichettandoli come positivi, in quanto la specificità è uguale a 0 in tutti gli split dello schema di cross-validazione stratificata shuffle split. La specificità è inferiore alla sensibilità per tutte le CNN 2D, il che significa che la maggior parte degli errori sono falsi positivi.

La tabella 6.18 e la tabella 6.19 elencano i risultati delle CNN 2D con lo strato Bi-LSTM e le CNN 2D con lo strato ConvLSTM sul dataset Hockey Fight. Tra le CNN 2D, VGG16 con lo strato ConvLSTM ottiene la migliore accuratezza (97,31% ± 0,40%), seguita da VGG19 con lo strato ConvLSTM (96,36% ± 1,39%) e VGG16 con lo strato Bi-LSTM (95,47% ± 1,17%). Tuttavia, come è successo sul dataset AIRTLab, il modello composto da C3D e il classificatore SVM ha la più alta accuratezza. Il modello composto da C3D e strati completamente connessi e il modello basato su ConvLSTM ottengono un'accuratezza leggermente inferiore rispetto a VGG16 più lo strato ConvLSTM, ma si comportano meglio di qualsiasi altra CNN 2D testata).

ResNet50V2 ottiene cattivi risultati anche sul dataset Hockey Fight, con un'accuratezza simile a un classificatore casuale quando combinato con Bi-

Tabella 6.19: I valori medi delle metriche calcolate sul set di dati Hockey Fight sulle cinque suddivisioni dello shuffle-split stratificato, per i modelli composti da CNN 2D pre-addestrate e ConvLSTM per l'estrazione delle caratteristiche.

	Loss	**Sensitivity**	**Specificity**
VGG16 + ConvLSTM	**0.0965 ± 0.0290**	**98.12 ± 1.58%**	**96.51 ± 1.14%**
VGG19 + ConvLSTM	0.1129 ± 0.0337	97.82 ± 1.61%	94.90 ± 2.20%
ResNet50V2 + ConvLSTM	0.5925 ± 0.0348	76.93 ± 14.95%	78.90 ± 10.33%
Xception + ConvLSTM	0.2491 ± 0.0244	92.47 ± 2.45%	92.70 ± 1.21%
NASNet + ConvLSTM	0.2787 ± 0.0615	91.58 ± 1.63%	91.00 ± 3.00%

	Accuracy	**F_1 score**	**AUC**
VGG16 + ConvLSTM	**97.31 ± 0.40%**	**97.34 ± 0.42%**	**99.63 ± 0.15%**
VGG19 + ConvLSTM	96.36 ± 1.39%	96.44 ± 1.35%	99.50 ± 0.32%
ResNet50V2 + ConvLSTM	77.91 ± 3.55%	77.11 ± 5.96%	86.78 ± 1.22%
Xception + ConvLSTM	92.59 ± 1.02%	92.60 ± 1.11%	96.01 ± 0.80%
NASNet + ConvLSTM	91.29 ± 1.63%	91.33 ± 1.74%	96.21 ± 0.77%

Tabella 6.20: I valori medi delle metriche calcolate sul dataset Crowd Violence sui cinque split dello shuffle-split stratificato, per i modelli composti da CNN 2D pre-addestrate e Bi-LSTM per l'estrazione delle caratteristiche.

	Loss	**Sensitivity**	**Specificity**
VGG16 + Bi-LSTM	**0.0703 ± 0.0185**	99.58 ± 0.82%	94.39 ± 3.30%
VGG19 + Bi-LSTM	0.0817 ± 0.0193	97.53 ± 0.93%	**96.07 ± 1.91%**
ResNet50V2 + Bi-LSTM	0.6817 ± 0.0006	**100.00 ± 0.00%**	0.00 ± 0.00%
Xception + Bi-LSTM	0.5838 ± 0.0384	82.47 ± 6.32%	58.88 ± 12.23%
NASNet + Bi-LSTM	0.4427 ± 0.0489	89.32 ± 4.32%	66.17 ± 8.39%

	Accuracy	**F_1 score**	**AUC**
VGG16 + Bi-LSTM	**97.39 ± 1.02%**	**97.79 ± 0.84%**	**99.84 ± 0.10%**
VGG19 + Bi-LSTM	96.92 ± 1.13%	97.34 ± 0.97%	99.61 ± 0.19%
ResNet50V2 + Bi-LSTM	57.71 ± 0.00%	73.18 ± 0.00%	51.27 ± 2.02%
Xception + Bi-LSTM	72.49 ± 3.76%	77.56 ± 2.87%	78.36 ± 2.66%
NASNet + Bi-LSTM	79.53 ± 3.72%	83.45 ± 2.92%	86.72 ± 3.42%

LSTM, aumentando fino al 77,91% quando combinato con il ConvLSTM. I modelli basati su VGG16 e VGG19 ottengono una specificità inferiore alla sensibilità (con più falsi positivi che falsi negativi). Invece, i modelli basati su Xception e NASNet Mobile hanno sensibilità e specificità simili sull'Hockey Fight, anche se sono in generale classificatori peggiori, in termini di accuratezza, score F_1, e AUC rispetto ai modelli basati su VGG. La tabella 6.20 e la tabella 6.21 elencano i risultati dei modelli CNN 2D con Bi-LSTM e ConvLSTM sul dataset Crowd Violence. I modelli basati su VGG16 e VGG19 ottengono un'accuratezza significativamente migliore rispetto alle altre CNN 2D. Il miglior modello è quello basato su VGG19 e lo strato ConvLSTM, con un'accuratezza pari al 98,74% (± 0,81%). Tuttavia, i nostri modelli basati su C3D si sono comportati, come classificatori, meglio di tutte le CNN 2D, ottenendo una migliore accuratezza, punteggio F_1, e AUC sul dataset Crowd Violence. Solo il modello basato su ConvLSTM ha faticato ad allenarsi su tale dataset, con una precisione simile a quella ottenuta da Xception e NASNet Mobile con lo strato ConvLSTM. Il modello basato su ResNet50V2 non è riuscito ad apprendere il compito di classificazione anche sul dataset Crowd Violence, ottenendo il valore

Tabella 6.21: I valori medi delle metriche calcolate sul dataset Crowd Violence sui cinque split dello shuffle-split stratificato, per i modelli composti da CNN 2D pre-addestrate e ConvLSTM per l'estrazione delle caratteristiche.

	Loss	**Sensitivity**	**Specificity**
VGG16 + ConvLSTM	0.0781 ± 0.0265	97.26 ± 1.44%	96.64 ± 3.95%
VGG19 + ConvLSTM	**0.0434 ± 0.0214**	**98.63 ± 1.06%**	**98.88 ± 1.50%**
ResNet50V2 + ConvLSTM	0.5634 ± 0.0714	92.33 ± 6.36%	44.85 ± 26.35%
Xception + ConvLSTM	0.3691 ± 0.0742	89.45 ± 5.81%	82.05 ± 3.26%
NASNet + ConvLSTM	0.3961 ± 0.0640	90.27 ± 3.61%	82.05 ± 4.24%

	Accuracy	**F_1 score**	**AUC**
VGG16 + ConvLSTM	97.00 ± 0.96%	97.41 ± 0.76%	99.80 ± 0.16%
VGG19 + ConvLSTM	**98.74 ± 0.81%**	**98.90 ± 0.70%**	**99.93 ± 0.05%**
ResNet50V2 + ConvLSTM	72.25 ± 8.44%	79.69 ± 4.20%	76.93 ± 14.14%
Xception + ConvLSTM	86.32 ± 3.77%	88.23 ± 3.45%	93.19 ± 2.96%
NASNet + ConvLSTM	86.80 ± 1.57%	88.73 ± 1.46%	93.99 ± 1.70%

di perdita più alto e non riuscendo a identificare correttamente i campioni negativi. Per riassumere, tra le CNN 2D pre-addestrate testate in questo articolo, VGG16 e VGG19 ottengono i migliori risultati su tutti i dataset, in particolare se combinati con uno strato ConvLSTM. Xception e NASNet Mobile hanno ottenuto risultati significativamente inferiori, mentre ResNet50V2 ha avuto una performance molto scarsa. In generale, le CNN 2D con il ConvLSTM ottengono risultati leggermente migliori delle CNN 2D con lo strato Bi-LSTM. Confrontando i modelli basati sulle CNN 2D con i modelli disponibili in letteratura ed elencati nella tabella 6.3, si conferma l'efficacia del transfer learning nel compito di rilevazione della violenza. Per esempio, il modello che combina VGG19 e ConvLSTM ottiene punteggi leggermente migliori rispetto al modello proposto da Ullah et al. [65]. Tuttavia, i modelli originariamente proposti in questo articolo ottengono risultati migliori sul dataset AIRTLab, dove il peggior modello proposto (C3D e gli strati completamente connessi) ottiene la stessa accuratezza dei migliori modelli basati su CNN 2D; i due modelli proposti basati su C3D ottengono prestazioni migliori delle CNN 2D sul Crowd Violence; infine, il modello basato su C3D e il classificatore SVM ha la migliore accuratezza sul dataset Hockey Fight; il modello basato su ConvLSTM, e il modello composto da C3D e gli strati completamente connessi ottengono un'accuratezza migliore di nove modelli basati su CNN 2D su dieci.

6.3.12 Limitazioni

I risultati della ricerca descritta in questo articolo sono promettenti, ma includono alcune limitazioni. Per quanto riguarda il dataset proposto, i video sono stati registrati da attori non professionisti e, pertanto, non includono violenza reale. Per questo motivo, mentre le metriche calcolate per i modelli proposti basati sul deep learning sono promettenti, i risultati non possono essere considerati generali. Tuttavia, i modelli proposti sono stati convalidati sui video reali dei dataset Hockey Fight e Crowd Violence, oltre ad essere radicati nella letteratura sul riconoscimento delle azioni e sul rilevamento della violenza.

Per quanto riguarda i risultati presentati, abbiamo costruito il nostro modello sui risultati del nostro lavoro precedente e sulla ricerca relativa al rilevamento della violenza, come spiegato nella sezione *Stato dell'arte*. Tuttavia, uno studio sistematico su iperparametri e modelli alternativi, così come un confronto su più dataset dovrebbe essere eseguito per ottenere risultati più generali, e quindi convalidare pienamente i nostri metodi.

Inoltre, abbiamo testato il nostro modello su sequenze composte da 16 fotogrammi presi da brevi video clip (la lunghezza media di una clip del dataset AIRTLab è di 5,6 secondi). Infatti, la maggior parte della letteratura si basa su test con video brevi. Tuttavia, l'accuratezza su video completi e reali

dovrebbe essere valutata prima di andare in produzione. Valutare brevi sequenze di fotogrammi presi da video lunghi potrebbe portare a troppi falsi positivi, interferendo negli usi pratici delle tecniche proposte. Pertanto, al fine di massimizzare l'accuratezza su video di lunghezza intera, i risultati sulle sotto-sequenze di fotogrammi dovrebbero essere fusi insieme. A questo proposito, una semplice strategia potrebbe essere quella di etichettare una parte di un video lungo come positiva solo quando un numero fisso di sotto sequenze consecutive di 16 fotogrammi sono etichettate come positive.

6.3.13 Conclusioni

Abbiamo presentato tre modelli basati sul deep learning per il rilevamento della violenza nei video: li abbiamo testati sulle clip del nuovo dataset AIRTLab, specificamente progettato per verificare la robustezza contro i falsi positivi, nonché sui dataset Hockey Fight e Crowd Violence, tradizionalmente utilizzati in letteratura per confrontare le tecniche di rilevamento della violenza. Gli esperimenti presentati in questo articolo consentono di trarre due conclusioni principali:

- i modelli basati su transfer learning proposti (C3D combinato con un classificatore SVM e C3D combinato con nuovi strati completamente connessi) ottengono risultati di accuratezza stabili su tutti e tre i dataset testati, essendo migliori, in molti casi, rispetto ai modelli testati in letteratura su Hockey Fight e Crowd Violence. Ciò suggerisce di persistere con modelli basati sul transfer learning per il compito di rilevamento della violenza;

- i nostri modelli basati su CNN 3D hanno prestazioni migliori rispetto alle ben note CNN 2D pre-addestrate su ImageNet e combinate con un modulo ricorrente per estrarre le caratteristiche spazio-temporali dei video nei dataset, suggerendo di continuare la ricerca sulle architetture 3D per rilevamento della violenza.

Inoltre, tutti i modelli proposti si sono dimostrati capaci di identificare meglio i video violenti che non violenti, dato che la maggior parte degli errori sono falsi positivi. Sebbene questo comportamento sia parzialmente influenzato dal fatto che i campioni delle due classi sono sbilanciati, ciò convalida comunque la progettazione del dataset AIRTLab nel controllo della robustezza rispetto ai falsi positivi.

Oltre ad affrontare i limiti della ricerca descritta, i lavori futuri ospiteranno un confronto più approfondito tra modelli basati su transfer learning e modelli addestrati da zero per il rilevamento della violenza, sia sul dataset AIRTLab che sugli altri dataset disponibili nella letteratura scientifica.

Tabella 6.22: Numero di epoche di addestramento in ogni split (S1-S5) per i due modelli end-to-end cioè C3D con due strati completamente connessi (C3D + FC) e l'architettura basata su ConvLSTM sul dataset Crowd Violence.

	S1	**S2**	**S3**	**S4**	**S5**	**Mean**
C3D + FC	8	8	8	15	14	10.60 ± 3.20
ConvLSTM	8	8	13	12	8	9.80 ± 2.23

6.3.14 *APPENDICE - A - Risultati sul set di dati di combattimento di hockey*

Per testare i modelli proposti sulle clip del dataset Hockey Fight, abbiamo seguito lo stesso protocollo sperimentale applicato al dataset AIRTLab. Pertanto, abbiamo applicato uno schema di convalida incrociata stratificata, randomizzando una divisione 80-20 per 5 volte, con l'80% dei dati usati come insieme di addestramento e il 20% dei dati come insieme di test. Con i due modelli end-to-end, il 12,5% dei dati di addestramento è stato utilizzato come insieme di validazione. I 1000 video del dataset Hockey Fight includono 2007 frammenti di 16 fotogrammi in totale.

La tabella 6.22 elenca il numero di epoche di addestramento in ogni suddivisione dei dati, per ogni modello end-to-end. Mentre il numero medio di epoche è 13,2 per entrambi i modelli, quello basato su C3D ha variato il numero di epoche di addestramento in ogni split in modo più significativo (deviazione standard 6,18) rispetto al modello ConvLSTM (deviazione standard 2,48). Il batch size è per il modello C3D e 8 per il modello ConvLSTM.

La tabella 6.23 include i risultati ottenuti dal modello composto da C3D e dal classificatore SVM in ogni suddivisione del dataset Hockey Fight. Il classificatore è indipendente dalla specifica suddivisione dei dati, poiché i risultati sono simili in tutte le suddivisioni. Con questo set di dati, la specificità è di solito simile o superiore alla sensibilità (tranne nello split numero 5). Infatti, a differenza del dataset AIRTLab, il dataset Hockey Fight non è fatto apposta per sfidare specificamente il rilevamento della violenza nell'etichettare come non violenti movimenti e comportamenti rapidi che potrebbero sembrare violenti. Tuttavia, nello split 5, 8 errori di classificazione su 10 sono falsi positivi, con 8 sequenze di 16 fotogrammi non violente etichettate come violente

La tabella 6.24 mostra le metriche calcolate per il modello end-to-end basato su C3D sulle suddivisioni del dataset Hockey Fight. I risultati sono simili a quelli ottenuti con il modello basato su C3D e SVM. Sensibilità e specificità sono simili in tutti gli split, anche se, nella maggior parte la specificità è leggermente inferiore, evidenziando un maggior numero di falsi positivi.

La tabella 6.25 elenca i risultati del modello end-to-end basato sulla architet-

Tabella 6.23: I risultati del modello composto da C3D e SVM, calcolati per ogni split dello schema di validazione incrociata stratificata shuffle-split sul dataset Crowd Violence.

	Split 1	**Split 2**	**Split 3**	**Split 4**	**Split 5**
Sensitivity	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
Specificity	**100.00%**	98.13%	**100.00%**	**100.00%**	97.20%
FPR	**0%**	1.87%	**0%**	**0%**	2.80%
Accuracy	**100.00%**	99.21%	**100.00%**	**100.00%**	98.81%
F_1 score	**100.00%**	99.32%	**100.00%**	**100.00%**	98.98%
AUC	**100.00%**	99.98%	**100.00%**	**100.00%**	**100.00%**

Tabella 6.24: I risultati del modello composto da C3D e dagli strati completamente connessi per la classificazione, calcolati per ogni split dello schema di validazione incrociata stratificata shuffle-split sul dataset Crowd Violence.

	Split 1	**Split 2**	**Split 3**	**Split 4**	**Split 5**
Loss	0.0909	0.0161	0.0535	0.0107	**0.0070**
Sensitivity	99.32%	**100.00%**	98.63%	**100.00%**	**100.00%**
Specificity	97.20%	98.13%	98.13%	**99.07%**	**99.07%**
FPR	2.80%	1.87%	1.87%	**0.93%**	**0.93%**
Accuracy	98.42%	99.21%	98.42%	**99.60%**	**99.60%**
F_1 score	98.64%	99.32%	98.63%	**99.66%**	**99.66%**
AUC	99.72%	99.98%	99.86%	99.99%	**100.00%**

Tabella 6.25: I risultati del modello basato sull'architettura ConvLSTM, calcolati per ogni split dello schema di validazione incrociata stratificata shuffle-split sul dataset Hockey Fight.

	Split 1	**Split 2**	**Split 3**	**Split 4**	**Split 5**
Loss	0.4845	**0.2876**	0.3780	0.3195	0.2977
Sensitivity	**97.26%**	94.52%	95.21%	93.15%	95.89%
Specificity	41.12%	81.31%	67.29%	72.90%	**83.18%**
FPR	58.88%	18.69%	32.71%	27.10%	**16.82%**
Accuracy	73.52%	88.93%	83.40%	84.58%	**90.51%**
F_1 score	80.91%	90.79%	86.88%	87.46%	**92.11%**
AUC	90.47%	**96.78%**	93.91%	94.85%	95.74%

tura ConvLSTM, sulle suddivisioni del dataset Hockey Fight. La rete basata su ConvLSTM ottiene una sensibilità leggermente inferiore rispetto ai due modelli precedenti. Tale modello potrebbe essere troppo complesso da addestrare da zero sui 1405 chunk da 16 fotogrammi usati come dati di allenamento, non essendo in grado di identificare correttamente i campioni violenti. Come i due modelli precedenti, l'architettura ConvLSTM non evidenzia una differenza significativa tra falsi positivi e falsi negativi, in termini di errori di classificazione.

6.3.15 *APPENDICE - B - Risultati sul set di dati sulla violenza di massa*

Per testare i modelli proposti sulle clip del dataset Crowd Violence, abbiamo seguito lo stesso protocollo sperimentale applicato sui dataset AIRTLab e Hockey Fight: abbiamo applicato uno schema di convalida incrociata stratificato, randomizzando una suddivisione 80-20 per 5 volte, con l'80% dei dati che serviva come insieme di addestramento e il 20% dei dati come insieme di test. Con i due modelli end-to-end, il 12,5% dei dati di addestramento è stato utilizzato come insieme di validazione. Il dataset Crowd Violence include un totale di 1265 pezzi da 16 fotogrammi.

La tabella 6.26 include il numero di epoche di addestramento in ogni divisione dei dati, per ogni modello end-to-end. Il modello basato su C3D mostra il più basso numero medio di epoche di addestramento di tutti i dataset: 10.6 (± 3.2). Infatti, il dataset Crowd Violence ha il più basso numero di campioni, il che si traduce in una più rapida convergenza della rete neurale sull'insieme di addestramento. Il numero medio di epoche di addestramento per il modello basato su ConvLSTM è 9,8 (± 2,23). Il batch size è di 32 per il modello C3D e di 8 per il modello ConvLSTM.

Tabella 6.26: Numero di epoche di addestramento in ogni split (S1-S5) per i due modelli end-to-end cioè C3D con due strati completamente connessi (C3D + FC) e l'architettura basata su ConvLSTM sul dataset Crowd Violence.

	S1	**S2**	**S3**	**S4**	**S5**	**Mean**
C3D + FC	8	8	8	15	14	10.60 ± 3.20
ConvLSTM	8	8	13	12	8	9.80 ± 2.23

Tabella 6.27: I risultati del modello composto da C3D e SVM, calcolati per ogni split dello schema di validazione incrociata stratificata shuffle-split sul dataset Crowd Violence.

	Split 1	**Split 2**	**Split 3**	**Split 4**	**Split 5**
Sensitivity	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
Specificity	**100.00%**	98.13%	**100.00%**	**100.00%**	97.20%
FPR	**0%**	1.87%	**0%**	**0%**	2.80%
Accuracy	**100.00%**	99.21%	**100.00%**	**100.00%**	98.81%
F_1 score	**100.00%**	99.32%	**100.00%**	**100.00%**	98.98%
AUC	**100.00%**	99.98%	**100.00%**	**100.00%**	**100.00%**

La tabella 6.27 mostra i risultati ottenuti dal modello composto da C3D e SVM sul dataset Crowd Violence. A causa del basso numero di campioni, l'architettura si comporta estremamente bene in termini di accuratezza. Tutti i pochi errori di classificazione sono falsi positivi. In particolare, 2 sequenze di 16 fotogrammi nello split 2 e 3 sequenze nello split 5 sono state erroneamente identificate come violente.

La tabella 6.28 elenca i risultati ottenuti sul dataset Crowd Violence dal modello end-to-end basato su C3D. In termini di accuratezza, il modello si comporta in modo simile a quello basato su C3D e SVM, con pochi errori di classificazione. Anche la loss è inferiore a 0,1 sull'insieme di test in tutte le suddivisioni, dimostrando la capacità del modello di classificare sul dataset Crowd Violence. Lo split 1 e lo split 3 mostrano l'accuratezza più bassa 98,42% con 4 errori di classificazione su 353 campioni di test. In particolare, nello split 1, ci sono 3 falsi positivi e 1 falso negativo; invece, nello split 3, entrambi i numeri di falsi negativi e positivi sono uguali a 2.

La tabella 6.29 include i risultati del modello end-to-end basato su ConvLSTM sul dataset Crowd Violence. Mentre i due modelli basati su C3D si sono comportati estremamente bene, al punto che sembrano soffrire di overfit sui dati, il modello basato su ConvLSTM fa fatica in termini di precisione di classificazione. Infatti, il modello è troppo complesso (198 milioni di parametri) per il basso numero di campioni del dataset Crowd Violence, come anche la loss sull'insieme di test è molto alta rispetto ai valori ottenuti dal modello

Tabella 6.28: I risultati del modello composto da C3D e dagli strati completamente connessi per la classificazione, calcolati per ogni split dello schema di validazione incrociata stratificata shuffle-split sul dataset Crowd Violence.

	Split 1	**Split 2**	**Split 3**	**Split 4**	**Split 5**
Loss	0.0909	0.0161	0.0535	0.0107	**0.0070**
Sensitivity	99.32%	**100.00%**	98.63%	**100.00%**	**100.00%**
Specificity	97.20%	98.13%	98.13%	**99.07%**	**99.07%**
FPR	2.80%	1.87%	1.87%	**0.93%**	**0.93%**
Accuracy	98.42%	99.21%	98.42%	**99.60%**	**99.60%**
F_1 score	98.64%	99.32%	98.63%	**99.66%**	**99.66%**
AUC	99.72%	99.98%	99.86%	99.99%	**100.00%**

Tabella 6.29: I risultati del modello basato sull'architettura ConvLSTM, calcolati per ogni split dello schema di validazione incrociata stratificata shuffle-split sul dataset Crowd Violence.

	Split 1	**Split 2**	**Split 3**	**Split 4**	**Split 5**
Loss	0.4845	**0.2876**	0.3780	0.3195	0.2977
Sensitivity	**97.26%**	94.52%	95.21%	93.15%	95.89%
Specificity	41.12%	81.31%	67.29%	72.90%	**83.18%**
FPR	58.88%	18.69%	32.71%	27.10%	**16.82%**
Accuracy	73.52%	88.93%	83.40%	84.58%	**90.51%**
F_1 score	80.91%	90.79%	86.88%	87.46%	**92.11%**
AUC	90.47%	**96.78%**	93.91%	94.85%	95.74%

basato sul C3D e sugli strati completamente connessi. Inoltre, la somiglianza tra clip violente e non violente, così come la bassa risoluzione, potrebbe avere un ruolo in quanto il modello tende ad etichettare i campioni come violenti. Questo è evidente nello split 1: 63 dei 107 campioni negativi sono classificati male, mentre solo 8 dei 146 campioni positivi sono classificati male.

6.4 Combinazione di una rete neurale profonda per dispositivi integrati e di uno strato ricorrente per il rilevamento della violenza nei video

[*pubblicato*][5]

[5] Contardo, P., Tomassini, S., Falcionelli, N., Dragoni, A. F., Sernani, P. (2023). Combining a mobile deep neural network and a recurrent layer for violence detection in videos. In proceedings RTA-CSIT 2023: 5th International Conference Recent Trends and Applications In Computer Science And Information Technology, April 26–27, 2023, Tirana, Albania.

Nella sezione 6.3, abbiamo visto l'importanza dei sistemi di videosorveglianza come fonte d'informazione per le Forze dell'Ordine. Un tipico caso d'uso, potrebbe configurarsi nelle manifestazioni di Ordine Pubblico che interessano ampi spazi e con grande affollamento di persone. L'Autorità di Pubblica Sicurezza, che si identifica nel Questore, dirige e coordina dal punto di vista Tecnico-Operativo la Forza Pubblica per i servizi di Ordine e Sicurezza Pubblica, disponendo se necessario anche la presenza della Polizia Scientifica per la video-documentazione dinamica della manifestazione ed eventualmente l'acquisizione di riprese provenienti da sistemi fissi [216]. Nel caso in cui si verifichino fatti violenti censurabili, un'operatore sarà chiamato ad analizzare i filmati registrati per individuare tali eventi. Ma il complesso meccanismo di visione umana, sfrutta due immagini che raggiungono distintamente le due retine degli occhi e che vedranno quindi singolarmente due immagini diverse a causa della reciproca prospettiva sull'oggetto. Il campo visivo sinistro e destro, sovrapponendosi frontalmente, generano un campo visivo binoculare che il cervello integra e nel punto focale consentirà di vedere un unico oggetto [217], punto in cui si focalizza l'attenzione, scemando ciò che accade intorno. Un compito arduo per il cervello, il quale fa sì che il sistema umano di elaborazione delle informazioni, al crescere della complessità dei problemi, tenda a semplificare l'elaborazione, per esempio attraverso l'attenzione selettiva ecc., condizionando di conseguenza anche la razionalità delle scelte [218]. Ecco che quindi l'opportunità di sfruttare l'Intelligenza Artificiale nell'analisi dei filmati, soprattutto quelli più complessi, come nel caso dell'Ordine Pubblico, grazie anche alle sue proprietà che superano i limiti dell'uomo potrebbe costituire un potente alleato delle Forze dell'Ordine nel riconoscimento e contrasto alla violenza.

Di fatto, le tecniche basate sul Deep Learning hanno dimostrato una migliore accuratezza nel rilevamento della violenza, proponendo di utilizzare le Reti neurali ricorrenti (RNN) e le Reti Neurali Convoluzionali (CNN) per questo compito [219]. Queste tecniche sono in grado di modellare le informazioni spazio-temporali incluse nei filmati CCTV, cioè le caratteristiche che rappresentano le informazioni sul movimento contenute in una sequenza di fotogrammi, oltre alle informazioni spaziali contenute in un singolo fotogramma.

In un nostro precedente lavoro [220], abbiamo testato 13 diverse reti neurali profonde (DNN) per il compito di rilevare la violenza nei video. In particolare, abbiamo confrontato una CNN 3D pre-addestrata, C3D [5], combinata con un classificatore di tipo Support Vector Machine (SVM), con C3D combinata con strati completamente connessi, con una Convolutional Long Short-Term Memory (ConvLSTM) [66] addestrata ex novo e con strati completamente connessi, con altre dieci reti basate su CNN 2D pre-addestrate e distribuite nel tempo, combinate con LSTM bidirezionali (Bi-LSTM) [221] (5 reti) e ConvLSTM (5 reti). I modelli basati su C3D hanno ottenuto i migliori risultati di accuratezza

nel rilevamento della violenza su diversi set di dati, sfruttando l'architettura 3D in grado di modellare le caratteristiche spazio-temporali dei video e l'apprendimento per trasferimento. Tuttavia, le CNN 3D richiedono risorse di calcolo e di archiviazione che di solito non sono compatibili con i dispositivi mobili ed embedded [222], cioè per il cd. "edge computing".

Per affrontare questo problema, in questo lavoro proponiamo due modelli basati sulla combinazione di una CNN specificamente progettata per i dispositivi integrati, MobileNetV2 [223], con uno strato ricorrente per estrarre le informazioni temporali e strati completamente connessi per la classificazione dei video in violenti o meno. In particolare, in un modello abbiamo utilizzato la Bi-LSTM come strato ricorrenti, mentre nell'altro abbiamo utilizzato la ConvLSTM. Per comprenderne l'efficacia e valutare eventuali cali di accuratezza, testiamo le reti proposte sul dataset AIRTLab [68], confrontando i risultati con quelli ottenuti nel nostro precedente lavoro. Per questo motivo, il presente lavoro contribuisce allo stato dell'arte del rilevamento della violenza con:

- la proposta di utilizzare MobileNetV2, pre-addestrata sul dataset Imagenet [214], distribuendola temporalmente sui fotogrammi dei video di sicurezza da classificare come violenti o meno, in combinazione con un modulo ricorrente per modellare le informazioni temporali oltre a quelle spaziali dei video.

- il confronto delle reti proposte con i modelli da noi precedentemente testati [220] per valutare il calo di accuratezza necessario per utilizzare una rete adattata ai dispositivi integrati, ovvero MobileNetV2.

Il resto del documento è organizzato come segue. La sezione *Stato dell'arte*6.4.1 fornisce una rassegna della letteratura sulle tecniche di Deep Learning applicate al rilevamento della violenza. La sezione *Metodologia*6.4.2 descrive le reti proposte, fornendo il background necessario e dettagliando la struttura del dataset utilizzato. La sezione *Risultati e discussione*6.4.6 discute la valutazione sperimentale e presenta i principali risultati. Infine, la sezione *Conclusioni*6.4.10 trae le conclusioni di questo studio.

6.4.1 Stato dell'arte

Diverse tecniche di rilevamento della violenza basate su reti neurali profonde e, in particolare, su reti neurali ricorrenti (come LSTM, Bi-LSTM, ConvLSTM) e CNN hanno dimostrato la loro efficacia [219]. Ad esempio, Sudhakaran e Lanz. [67] hanno combinato le caratteristiche spaziali calcolate da CNN 2D sui fotogrammi dei video, con una ConvLSTM, per estrarre anche le caratteristiche temporali. Hanno ottenuto il 94,5% di accuratezza sul dataset Crowd Violence [60] e il 97,1% sul dataset Hockey Fight [63]. Li et al. [64] hanno proposto

una CNN 3D composta da 10 strati, aggiungendo strati densi e di transizione dopo gli strati convoluzionali. Hanno ottenuto un'accuratezza del 97,2% sul dataset Crowd Violence e del 98,3% sul dataset Hockey Fight. Anche Accattoli et al. [4] e Ullah et al. [65] hanno basato il loro lavoro su una CNN 3D, ma, invece di addestrarla da zero, hanno applicato il transfer learning. Accattoli et al. hanno aggiunto una SVM alla CNN, ottenendo un'accuratezza del 99,2% sull'Hockey Fight e del 98,5% sul Crowd Violence. Ullah et al. hanno invece implementato una rete neurale end-to-end aggiungendo strati completamente connessi alla CNN 3D, ottenendo un'accuratezza del 98% sul Crowd Violence e del 96% sull'Hockey Fight. Sernani et al. [220] hanno confrontato 13 diverse reti neurali profonde sui dataset Hockey Fight, Crowd Violence e AIRTLab. In particolare, sono state testate una CNN 3D pre-addestrata (C3D) combinata con una SVM, una C3D combinata con strati completamente connessi, una ConvLSTM combinata con strati completamente connessi, 5 CNN 2D pre-addestrate distribuite nel tempo combinate con la Bi-LSTM e le stesse CNN 2D combinate con una ConvLSTM. I risultati migliori sono stati ottenuti con le due reti basate su C3D, con un'accuratezza del 96,1% sul dataset AIRTLab, del 97,86% sull'Hockey Fight e del 99,6% sul Crowd Violence. Freire-Obregón et al. [224] hanno utilizzato una Inflated 3D ConvNet per estrarre le caratteristiche spazio-temporali sull'output di due localizzatori di persone per eseguire il rilevamento della violenza senza contesto, cioè il rilevamento della violenza applicato solo ai soggetti nei video, scartando qualsiasi informazione riguardante lo sfondo o il contesto della scena. Hanno combinato questo estrattore di caratteristiche con diversi classificatori, ottenendo i migliori risultati con la Regressione lineare, con un'accuratezza del 99,45% sul dataset Crowd Violence, del 99,43% sull'Hockey Fight e del 97,54% sull'AIRTLab.

Sebbene le tecniche sopra citate si siano dimostrate efficaci nel compito di rilevare automaticamente la violenza in diversi database di video, sono tutte molto esigenti in termini di risorse di calcolo e di archiviazione, il che le rende inadeguate per l'esecuzione in dispositivi integrati, cioè per l'edge computing. Nel nostro precedente lavoro [220], abbiamo dimostrato che le CNN 2D pre-addestrate, distribuite temporalmente sui fotogrammi dei video di sicurezza e combinate con uno strato Bi-LSTM, raggiungono un'accuratezza inferiore rispetto alle CNN 3D. Ad esempio, VGG16 [174], combinato con una Bi-LSTM, ha ottenuto un'accuratezza del 94,92% sul dataset AIRTLab, del 95,47% sull'Hockey Fight e del 97,39% sul Crowd Violence. Tuttavia, una tale accuratezza nel rilevare la violenza potrebbe essere ancora accettabile, per ottenere un compromesso per eseguire il rilevamento della violenza su dispositivi integrati ed evitare così la trasmissione dei dati nell'ottica di preservare la privacy. Pertanto, alla luce di tali risultati e della necessità di modelli in grado di eseguire il rilevamento della violenza su dispositivi integrati, a differenza dei lavori elen-

cati proponiamo di "distribuire nel tempo" MobileNetV2 [223], una CNN 2D specificamente progettata per dispositivi mobili, sui fotogrammi dei video di sicurezza. La combiniamo con uno strato ricorrente e strati completamente connessi per eseguire la classificazione della violenza e ne testiamo due versioni diverse, una basata su Bi-LSTM e una su ConvLSTM.

Oltre alla ricerca della migliore accuratezza, la letteratura scientifica relativa all'uso di tecniche di Deep Learning per il rilevamento automatico di scene di violenza comprende altri studi. Ad esempio, Ciampi et al. [225] hanno testato alcune delle tecniche citate, come le CNN 3D e le ConvLSTM, su un nuovo dataset, il Bus Violence, per studiare il comportamento delle metodologie di rilevamento della violenza basate sul Deep Learning quando le informazioni sullo sfondo e sul contesto variano in modo significativo. Silva et al. [226] hanno proposto l'uso di un approccio di apprendimento federato per distribuire il processo di apprendimento su diversi dispositivi, preservando la privacy, con un server che combina il modello addestrato localmente in un modello globale. Tuttavia, invece di basarsi su video o su porzioni di video, hanno applicato CNN 2D a singoli fotogrammi, ottenendo i migliori risultati con MobileNet (99,4% di accuratezza sul dataset AIRTLab). Yang et al. [227] hanno proposto un approccio multimodale (Multimodal Contrastive Learning – MCL) per utilizzare sia il video che l'audio per il rilevamento automatico della violenza. Hanno ottenuto una precisione media dell'84,03% sul dataset XD-Violence [228], contro l'83,19% dell'uso del solo video e il 76,07% dell'uso del solo audio.

6.4.2 Metodologia

Come illustrato nelle Sezioni iniziale 6.4 e *Stato dell'arte* 6.4.1, molti studi sull'uso delle reti neurali profonde per il rilevamento della violenza nei video hanno proposto architetture complesse, come le CNN 3D, che richiedono risorse computazionali e di memoria solitamente non compatibili con i dispositivi integrati. A tal fine, proponiamo l'uso di MobileNetV2, "time-distributed" su porzioni di 16 fotogrammi dei video, combinata con uno strato ricorrente per modellare le informazioni temporali della sequenza di fotogrammi, oltre a quelle spaziali. Di seguito vengono fornite alcune informazioni di base su MobileNetV2, sull'architettura LSTM e sull'architettura ConvLSTM (6.4.3). Successivamente, presentiamo le reti neurali proposte (6.4.4) e descriviamo il dataset utilizzato per i test (6.4.5).

6.4.3 MobileNetV2, LSTM e ConvLSTM

Secondo la definizione originale di LeCun e Bengio [206], un'unità di uno strato in una CNN riceve input da un insieme di unità nel campo recettivo locale, tramite un'operazione di convoluzione con kernel composti da pesi con-

divisi. In MobileNetV2 [223] questo concetto viene esteso per far fronte alle limitate risorse computazionali dei dispositivi integrati. Invece della tradizionale operazione di convoluzione delle CNN, MobileNetV2 decompone gli strati convoluzionali in due strati separati:

- il livello di convoluzione "depthwise" che applica un filtro separato a ciascun canale di ingresso;
- lo strato di convoluzione "pointwise" che viene applicato all'uscita dello strato di convoluzione depthwise utilizzando una convoluzione 1x1.

Inoltre, in MobileNetV2, i "linear bottleneck" e le connessioni residue seguono la convoluzione. In particolare, i linear bottleneck utilizzano una funzione di attivazione lineare invece di una funzione di attivazione non lineare, riducendo il costo computazionale della rete

Come una CNN tradizionale, MobileNetV2 modella le informazioni spaziali delle immagini, cioè i fotogrammi dei video. Pertanto, abbiamo aggiunto uno strato ricorrente all'uscita di MobileNetV2 per modellare le informazioni temporali disponibili nei video, utilizzando una Bi-LSTM e una ConvLSTM. Nell'architettura LSTM originale [209], un'unità nascosta è composta da una cella autoricorrente, detta cella di memoria, il cui ingresso/uscita è regolato da tre porte moltiplicative, ossia la porta di ingresso, la porta di uscita e il "forget gate" [210]. In particolare, l'output h_t al tempo t di un'unità nascosta LSTM è dato dalle seguenti equazioni [210]:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{6.6}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{6.7}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{6.8}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{6.9}$$

$$h_t = o_t \tanh(c_t) \tag{6.10}$$

dove i_t, f_t, o_t e c_t sono i vettori di attivazione della porta di ingresso, del forget gate, della porta d'uscita e della cella di memoria al tempo t, σ è la funzione sigmoide, b è la soglia di attivazione di ciascuna porta/cella, e W sono le matrici diagonali dei pesi.

Nella formulazione originale, una LSTM elabora i dati in ingresso in ordine temporale crescente. Tuttavia, il riconoscimento di un certo pattern potrebbe essere più efficace con l'uso del contesto futuro. A tal fine, sono state proposte le RNN bidirezionali [229] e, in particolare, le LSTM bidirezionali [221]. L'idea di base di questi modelli è quella di presentare le sequenze di addestramento sia in avanti che all'indietro, utilizzando due reti ricorrenti separate, collegate

allo stesso strato di uscita. Per questo motivo, abbiamo basato uno dei nostri modelli sul Bi-LSTM, in quanto i video vengono elaborati una volta registrati, sfruttando sia il contesto precedente che quello futuro.

Per il ConvLSTM, utilizziamo la formulazione di Shi et al. [66], che ha esteso l'architettura LSTM aggiungendo strutture convoluzionali alla transizione di stato. Come hanno spiegato Shi et al., l'architettura LSTM è adeguata per estrarre caratteristiche temporali, ma include troppa ridondanza per le caratteristiche spaziali. A questo proposito, hanno proposto di aggiungere strutture convolutive nelle transizioni tra la porta di ingresso e la cella di memoria e nell'autoricorrenza della cella di memoria, regolata dal forget gate. Pertanto, in un ConvLSTM, l'uscita di un'unità nascosta è regolata dalle seguenti equazioni:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{6.11}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{6.12}$$

$$x_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{6.13}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{6.14}$$

$$h_t = o_t \tanh(c_t) \tag{6.15}$$

dove le attivazioni della porta d'ingresso, del forget gate, della porta di uscita e della cella di memoria (i_t, f_t, o_t e c_t), così come l'input e l'output (x_t, h_t) sono tensori 3D. Per questo, abbiamo utilizzato il ConvLSTM nel secondo dei modelli proposti.

6.4.4 Architettura di classificazione proposta

Come illustrato nella figura 6.9, per classificare i video in violenti o meno, proponiamo due classificatori basati sul Deep Learning e su MobileNetV2, pre-addestrata sul dataset Imagenet [214], seguiti da uno strato ricorrente e da uno strato completamente connesso. I pesi di MobileNetV2 sono congelati all'apprendimento su Imagenet. Invece, lo strato Bi-LSTM o lo strato ConvLSTM e gli strati completamente connessi vengono addestrati da zero sul dataset AIRTLab, come spiegato nella *Sezione Risultati e discussione* 6.4.6 (Sottosezione 6.4.7). Dato che nel nostro lavoro precedente abbiamo eseguito la classificazione su porzioni di video da 16 fotogrammi, in questo lavoro utilizziamo le stesse porzioni da 16 fotogrammi, per consentire un confronto equo tra i classificatori. I video del dataset AIRTLab sono stati ridimensionati a $224 \times 224 pixel$, poiché questa è la forma di input dell'implementazione originale di MobileNetV2.

La tabella 6.30 include gli strati che compongono il primo modello proposto. MobileNetV2, con i suoi 2.257.984 pesi congelati, è distribuita temporalmente

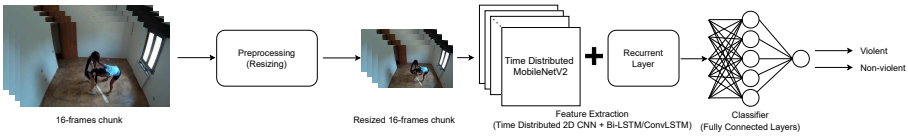


Figura 6.9: Schema dei modelli proposti. Ognuno dei modelli elabora sequenze composte da 16 fotogrammi ridimensionati a 224 x 224 pixel. Per applicare MobileNetV2 ai video (cioè un input 3D), dato che si tratta di una CNN 2D, la rete è distribuita temporalmente sui 16 fotogrammi dei video di sicurezza utilizzati in questo studio. Per estrarre le caratteristiche temporali dei video in aggiunta a quelle spaziali estratte da MobileNetV2, la CNN distribuita nel tempo è seguita da uno strato ricorrente (una Bi-LSTM o una ConvLSTM). Infine, gli strati completamente connessi eseguono la classificazione dei video in violenti o non violenti.

Tabella 6.30: Il primo modello di classificazione proposto. La Bi-LSTM e due strati completamente connessi sono stati aggiunti a MobileNetV2, pre-addestrata su ImageNet. MobileNetV2 è stata distribuita nel tempo per essere applicata a un input 3D, cioè le clip del dataset AIRTLab.

Layer	**Architecture**	**Output Shape**	**Params #**
Time Distr. MobileNetV2	-	(16, 7, 7, 1280)	2257984
Time Distr. Flatten	-	(16, 62720)	0
Bi-LSTM	128 units	(256)	64357376
Dropout	0.5 rate	(256)	0
Dense	128 units, ReLU	(128)	32896
Dropout	0.5 rate	(128)	0
Dense	1 units, Sigmoid	(1)	129

sui 16 fotogrammi utilizzati come input. La Bi-LSTM è composta da 128 unità nascoste, seguite da un dropout di 0,5 per limitare l'overfitting, uno strato completamente connesso con 128 neuroni ReLU, un altro dropout di 0,5 e un neurone sigmoide completamente connesso per la classificazione finale.

La Tabella 6.31 elenca gli strati che compongono il secondo modello proposto. Una ConvLSTM composta da 64 3 filtri *times* 3 con la funzione di attivazione *tanh* segue la MobileNetV2 distribuita nel tempo. La rete è completata da un dropout di 0,5, uno strato completamente connesso con 256 neuroni ReLU, un altro dropout di 0,5 e un neurone sigmoide completamente connesso per eseguire la classificazione finale in violento o meno.

Il modello basato su Bi-LSTM ha un totale di 66.648.385 parametri. I pesi di MobileNetV2 sono congelati, il che significa che il numero totale di parametri addestrabili è 64.390.401 (corrispondenti alle 128 unità nascoste della

Tabella 6.31: Il secondo modello di classificazione proposto. La ConvLSTM e due strati completamente connessi sono stati aggiunti a MobileNetV2, pre-addestrata su ImageNet. MobileNetV2 è stata distribuita nel tempo per essere applicata a un input 3D, cioè le clip del dataset AIRTLab.

Layer	**Architecture**	**Output Shape**	**Params #**
Time Distr. MobileNetV2	-	(16, 7, 7, 1280)	2257984
ConvLSTM	64 3x3 filters, tanh	(5, 5, 64)	3096832
Flatten	-	(1600)	0
Dropout	0.5 rate	(1600)	0
Dense	256 units, ReLu	(256)	409856
Dropout	0.5 rate	(256)	0
Dense	1 unit, Sigmoid	(1)	257

Bi-LSTM, ai 128 neuroni ReLU del primo strato completamente connesso e al neurone sigmoide dell'ultimo strato). Invece, nel modello basato su ConvLSTM ci sono 5.764.929 parametri (3.506.945 sono addestrabili, corrispondenti ai 64 filtri dello strato ConvLSTM, ai 256 neuroni ReLU del primo strato completamente connesso e al neurone sigmoide finale per la classificazione). Pertanto, il modello basato sul ConvLSTM richiede meno memoria rispetto al modello basato sul Bi-LSTM, risultando più adeguato per l'uso in dispositivi integrati.

6.4.5 Dataset usato per i test

Per testare le prestazioni dei classificatori proposti e confrontarli con il nostro lavoro precedente, abbiamo eseguito dei test di accuratezza sul dataset AIRTLab. Esso contiene 350 video (file MP4 con codec H.264, lunghezza media di 5,63 secondi). La frequenza dei fotogrammi è di 30 fps e la risoluzione è di $1920 \times 1080 pixel$. Il dataset comprende 230 video violenti e 120 video non violenti. I 230 video violenti rappresentano 115 azioni violente registrate da due diverse telecamere posizionate in due punti diversi. Allo stesso modo, i 120 video non violenti rappresentano 60 azioni non violente, registrate da due diverse telecamere posizionate in due punti diversi. Tutti i video sono stati girati all'interno della stessa stanza. Una telecamera è stata posizionata nell'angolo in alto a sinistra, davanti alla porta di ingresso nella stanza. La seconda telecamera si trovava nell'angolo in alto a destra, sul lato della porta.

Un gruppo di attori non professionisti ha interpretato le azioni violente e non violente. Il numero di attori varia da 2 a 4 per video. Nei video violenti, gli attori hanno simulato azioni frequenti nelle risse, come pugni, calci, percosse con bastoni, schiaffi, colpi di pistola e accoltellamenti. Nei video non violenti, gli attori hanno simulato azioni che possono dare luogo a falsi positivi a causa della somiglianza con le azioni violente (ad esempio per la presenza di movimenti

veloci). In particolare, i video non violenti contengono azioni come esultare, abbracciare, gesticolare, battere le mani e dare il cinque.

6.4.6 Preambolo

Abbiamo testato i due modelli proposti con lo stesso protocollo utilizzato nel nostro precedente lavoro [220], ossia misurando i risultati di classificazione sul dataset AIRTLab. L'obiettivo è confrontare le prestazioni di accuratezza dei classificatori basati su una CNN 2D progettata per dispositivi integrati con quelle dei classificatori che richiedono maggiori risorse. Pertanto, nelle sottosezioni seguenti, descriviamo il *Protocollo sperimentale e metriche di valutazione* (6.4.7), discutiamo i *Risultati* (6.4.8) e presentiamo i *Limiti della nostra valutazione* (6.4.9).

6.4.7 Protocollo sperimentale e metriche di valutazione

Mentre MobileNetV2 è stato pre-addestrato su Imagenet e i suoi pesi sono stati congelati, gli strati Bi-LSTM e ConvLSTM, insieme agli strati completamente connessi, hanno dovuto essere addestrati da zero. Pertanto, per eseguire l'addestramento e il test sul dataset AIRTLab, abbiamo applicato uno schema di convalida incrociata stratificata (shuffle split). A tal fine, abbiamo ripetuto per 5 volte una suddivisione randomizzata 80-20, utilizzando l'80% dei dati come set di addestramento e il 20% come set di test, mantenendo la percentuale di campioni di ciascuna classe in ogni suddivisione. Le suddivisioni dei dati sono state le stesse sia per i modelli proposti sia per i modelli del nostro lavoro precedente, per realizzare un confronto equo. Dato che gli input per i modelli sono sequenze composte da 16 fotogrammi e che i video del dataset includono un totale di 3537 sequenze di questo tipo, sono stati utilizzati 2829 campioni (cioè porzioni di 16 fotogrammi) per l'addestramento e 708 per il test, in ogni divisione. Il 12,5% dei dati di addestramento, ovvero il 10% dell'intero set di dati, è stato utilizzato come dati di validazione.

Entrambi i modelli proposti hanno utilizzato la funzione di *loss Binary Cross Entropy*, minimizzata con l'ottimizzatore Adam. Abbiamo interrotto l'addestramento dopo 5 epoche senza alcun miglioramento della loss minima di validazione, ripristinando i pesi corrispondenti all'epoca migliore. A tal fine, la Tabella 6.32 elenca il numero di epoche di addestramento in ciascuna suddivisione dello schema di convalida incrociata stratificata, per ciascun modello. Il numero medio di epoche di addestramento è stato di 22,4 (*pm* 5,68) per il modello che utilizza lo strato Bi-LSTM e di 17,8 (*pm* 4,21) per il modello basato su ConvLSTM. La dimensione del batch era di 8 per entrambe le reti neurali.

I test sono stati eseguiti su Google Colab Pro con il runtime GPU (la GPU utilizzata per i test era una Nvidia A100 SXM4 con 40 GB di RAM) e RAM

Tabella 6.32: Numero di epoche di addestramento in ciascuna suddivisione (S1-S5) dello schema di convalida incrociata stratificata.

	S1	S2	S3	S4	S5	Mean
MobileNetV2 + Bi-LSTM	20	18	23	19	32	22.40 ± 5.68
MobileNetV2 + ConvLSTM	11	20	22	19	17	17.80 ± 4.21

estesa (83,5 GB), utilizzando Keras 2.11.0, TensorFlow 2.11.0 e Scikit-learn 1.2.1.

Etichettando come negativi i frammenti di 16 fotogrammi delle clip non violente e come positivi i frammenti delle clip violente, abbiamo calcolato le seguenti metriche sull'insieme di test in ogni suddivisione dello schema di convalida incrociata:

- sensibilità (True Positive Rate – TPR), cioè la porzione dei positivi correttamente identificata (tra tutti i positivi disponibili nell'insieme di test);
- specificità (True Negative Rate – TNR), cioè la porzione dei negativi correttamente identificata (tra tutti i negativi disponibili nell'insieme di test);
- l'accuratezza, cioè i campioni correttamente identificati (tra tutti i campioni disponibili nell'insieme di test);
- lo score F_1, cioè la media armonica di precisione (il rapporto fra i positivi correttamente identificati e tutti i positivi individuati) e sensibilità.

Queste metriche possono essere formulate in termini di veri positivi (TP), veri negativi (TN), falsi positivi (FP) e falsi negativi (FN) secondo le equazioni 6.1, 6.2, 6.4 e 6.5.

Inoltre, in ogni suddivisione, abbiamo calcolato la curva Receiver Operating Characteristic (ROC) e l'Area Under the Curve (AUC), mostrando il TPR contro il tasso di falsi positivi (FPR) al variare della soglia di classificazione, al fine di comprendere la capacità diagnostica di ogni modello.

6.4.8 Risultati e discussione

La Tabella 6.33 elenca le metriche ottenute dal modello composto da MobileNetV2 e dalla Bi-LSTM nei cinque split della convalida incrociata eseguita sul dataset AIRTLab. Le metriche variano significativamente tra gli split, mostrando una scarsa capacità di generalizzazione. Ad esempio, nello split 1 tutti i 708 campioni del set di test sono etichettati come violenti, causando 232 falsi positivi. La sensibilità è del 100%, mentre la specificità è dello 0%. Lo split in

Tabella 6.33: I risultati del modello composto da MobileNetV2 e Bi-LSTM, calcolati per ogni split dello schema di validazione incrociata stratificata shuffle-split.

	Split 1	**Split 2**	**Split 3**	**Split 4**	**Split 5**
Sensitivity	**100.00%**	90.97%	97.48%	95.59%	97.69%
Specificity	**100.00%**	87.93%	62.50%	76.72%	82.76%
FPR	**0%**	12.07%	37.50%	23.28%	17.24%
Accuracy	67.23%	89.97%	86.02%	89.41%	**92.80%**
F_1 score	80.41%	92.42%	90.36%	92.39%	**94.80%**

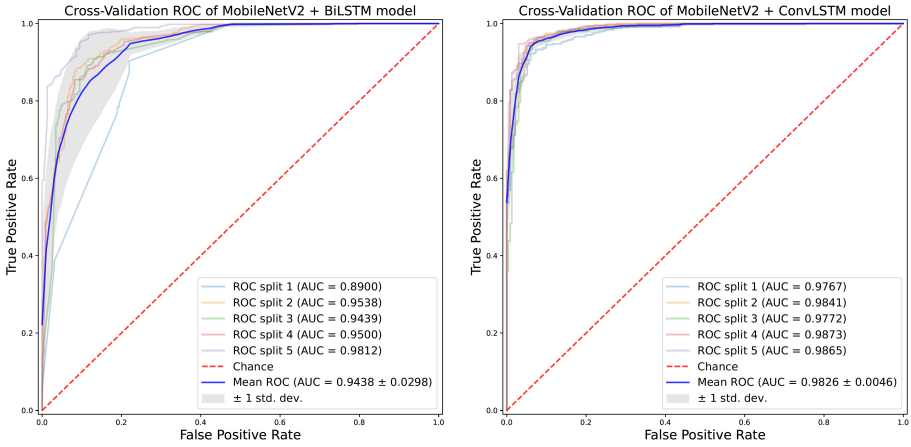
Tabella 6.34: I risultati del modello composto da MobileNetV2 e ConvLSTM, calcolati per ogni split dello schema di validazione incrociata stratificata shuffle-split.

	Split 1	**Split 2**	**Split 3**	**Split 4**	**Split 5**
Sensitivity	94.54%	95.80%	**97.90%**	94.96%	96.22%
Specificity	88.36%	92.24%	85.34%	**93.10%**	**93.10%**
FPR	11.64%	7.76%	14.66%	**6.90%**	**6.90%**
Accuracy	92.51%	94.63%	93.79%	94.35%	**95.20%**
F_1 score	94.44%	96.00%	95.49%	95.76%	**96.42%**

cui la maggior parte dei negativi viene identificata correttamente è il numero 2: in questo caso, 204 negativi su 232 vengono classificati correttamente (specificità 87,93%). Nella stessa divisione, 433 pezzi violenti su 476 sono classificati correttamente. L'accuratezza è quindi del 92,8%.

Invece, il modello basato su MobileNetV2 e ConvLSTM mostra una migliore capacità di generalizzazione rispetto al precedente modello, come mostrato nella Tabella 6.34. La sensibilità è superiore al 94% in tutti gli split e la specificità più bassa è quella dello split 3 (85,34%). Il miglior split è il numero 5, dove lo score F_1 è del 96,42%.

La differenza nella capacità di generalizzazione dei due modelli proposti è evidenziata dalle curve ROC in Figura 6.10. Infatti, il modello che utilizza la Bi-LSTM come strato ricorrente ottiene una AUC media pari al 94,38% (*pm* 2,98%), mentre il modello che utilizza la ConvLSTM ottiene il 98,26% (*pm* 0,46%). Questo comportamento potrebbe essere dovuto al diverso numero di parametri addestrabili dei due modelli. Nel modello basato su Bi-LSTM ci sono 64.390.401 parametri addestrabili. Nel modello basato su ConvLSTM, invece, il numero di parametri addestrabili è pari a 3.506.945. Pertanto, il modello basato su Bi-LSTM potrebbe essere sovradimensionato per il compito di rilevamento della violenza sul dataset AIRTLab, faticando a convergere verso prestazioni



(a) (b)

Figura 6.10: Curva ROC e AUC per i modelli MobileNetV2 + Bi-LSTM (a) e MobileNetV2 + ConvLSTM (b).

Tabella 6.35: Confronto dei valori medi delle metriche per i due modelli proposti, basati su MobileNetV2, con le metriche calcolate per i modelli del nostro lavoro precedente, basati su C3D, sulle cinque suddivisioni della validazione incrociata.

	Sensitivity	**Specificity**	**Accuracy**	**F_1 score**	**AUC**
MobileNetV2 + Bi-LSTM	96.34 ± 3.03%	61.98 ± 32.14%	85.08 ± 9.18%	90.08 ± 5.04%	94.38 ± 2.98%
MobileNetV2 + ConvLSTM	95.88 ± 1.17%	90.43 ± 3.09%	94.10 ± 0.91%	95.62 ± 0.67%	98.26 ± 0.46%
C3D + SVM	97.06 ± 0.80%	94.14 ± 1.51%	96.10 ± 0.71%	97.10 ± 0.53%	99.30 ± 0.23%
C3D + FC	97.82 ± 0.69%	91.12 ± 2.03%	95.62 ± 0.42%	96.78 ± 0.30%	98.94 ± 0.31%

di classificazione accettabili. Pertanto, il modello basato su ConvLSTM, che è il più leggero in termini di risorse richieste tra i due proposti in questo lavoro, mostra prestazioni migliori in termini di accuratezza di classificazione e capacità di generalizzazione.

La Tabella 6.35 confronta le prestazioni dei due modelli proposti in questo lavoro con quelli basati su C3D testati nel nostro precedente lavoro. Anche se più leggero in termini di risorse computazionali richieste, il modello basato su MobileNetV2 e ConvLSTM ottiene un'AUC media del 98%, contro il 99% dei modelli basati su C3D. L'accuratezza media e il punteggio F_1 del modello basato su ConvLSTM sono pari al 94,1% (*pm* 0,91%) e al 95,62% (*pm* 0,67%), inferiori di circa il 2% rispetto al modello C3D + SVM del nostro lavoro precedente. Pertanto, risorse limitate come quelle dei dispositivi integrati potrebbero giustificare l'uso di MobileNetV2 combinato con ConvLSTM, poiché la diminuzione delle metriche di accuratezza è limitata.

6.4.9 Limiti della valutazione

I risultati delle ricerche descritte in questo articolo sono promettenti, ma presentano alcune limitazioni. Infatti, ci siamo concentrati sull'accuratezza di due modelli basati su MobileNetV2, che è stato progettato per dispositivi integrati. Abbiamo però eseguito i nostri test comparativi nel cloud, utilizzando una GPU. Sebbene la diminuzione dell'accuratezza sia limitata e giustifichi l'uso del migliore tra i modelli proposti, per ottenere conclusioni più generali è necessario eseguire test su dispositivi integrati reali, come nel caso tipico dell'edge computing. Inoltre, i nostri test si basano su un set di video in cui la violenza è simulata da attori. Per confermare i risultati di accuratezza sono necessari test su video di telecamere di sorveglianza reali.

Inoltre, abbiamo raccolto le metriche su porzioni di 16 fotogrammi presi dai video brevi del dataset AIRTLab (la lunghezza media è di 5,6 secondi), per rendere questo lavoro confrontabile con le nostre ricerche precedenti. Mentre la maggior parte della letteratura correlata esegue test su video brevi, è necessario valutare l'accuratezza su video reali di lunghezza completa. Infatti, l'utilizzo di brevi frammenti di fotogrammi tratti da video lunghi, come nel nostro studio, potrebbe dare luogo a un numero eccessivo di falsi positivi. Pertanto, i risultati sui frammenti dovrebbero essere uniti a una strategia adeguata per massimizzare l'accuratezza sui video completi. A tal fine, una soluzione semplice consiste nell'etichettare una parte di un video lungo come violento solo quando un numero fisso di spezzoni consecutivi di 16 fotogrammi viene etichettato come violento.

6.4.10 Conclusioni

Per essere utilizzate in applicazioni reali, le tecniche basate sull'Intelligenza Artificiale e sul Deep Learning devono tenere conto delle prestazioni in tempo reale ed essere in grado di funzionare su dispositivi integrati, in modalità edge computing. Infatti, una risposta intelligente conserva la sua importanza solo se viene fornita in tempo, come osservato in [71]. Per tale motivo, in questo lavoro abbiamo proposto due reti neurali profonde per la classificazione dei video in violenti o meno. Entrambe le reti sono basate su MobileNetV2, una CNN progettata specificamente per dispositivi integrati. Tale CNN è responsabile dell'estrazione delle caratteristiche spaziali nei video. Abbiamo combinato MobileNetV2 con uno strato ricorrente per l'estrazione delle caratteristiche temporali. Uno dei due modelli proposti utilizza uno strato Bi-LSTM come modulo ricorrente. L'altro utilizza invece un ConvLSTM.

Abbiamo eseguito test comparativi sul dataset AIRTLab. Il modello che utilizza la ConvLSTM, il più leggero in termini di risorse computazionali e di memoria tra i due proposti in questo lavoro, ha ottenuto la migliore accuratez-

za, con una AUC media pari al 98,26% (*pm* 0,46%). Rispetto ai modelli del nostro lavoro precedente, basati su una CNN 3D, la diminuzione delle prestazioni in termini di AUC è di circa l'1% e del 2% in termini di accuratezza di classificazione sugli split del dataset AIRTLab. Questi risultati incoraggiano l'uso di modelli leggeri per dispositivi integrati. Ad esempio, potrebbero essere utili per elaborare i dati direttamente vicino alla telecamera che sta registrando il video di sicurezza e, quindi, preservare la privacy garantendo allo stesso tempo la sicurezza pubblica.

I lavori futuri affronteranno le limitazioni individuate. In particolare, è necessario eseguire test su dispositivi mobili o embedded reali per ottenere risultati più conclusivi e generali.

6.5 Tecniche di Deep Learning per il riconoscimento automatico della violenza nei file audio

La ricerca dei comportamenti violenti negli audio-file, è una porzione del riconoscimento automatico della violenza che stiamo affrontando nell'ambito dell'Accordo d'Intesa in atto tra il centro di ricerca interdipartimentale CARMELO e la Polizia di Stato. Per la strategia da adottare nella ricerca, abbiamo deciso di essere aderenti alle indicazioni suggerite in ambito Europeo per quanto concerne il trattamento dei file audio a scopi forensi, in modo da rimanere vicini agli standard di garanzia della qualità.

A tal riguardo, l'ENFSI ha rilasciato quattro manuali per l'analisi del parlato e file audio, di cui due manuali di buone pratiche e due linee guida molto utili per definire un flusso di lavoro efficace e condivisibile, che sono:

1. Best Practice Manual for the Methodology of Forensic Speaker Comparison [230];
2. Best Practice Manual for Digital Audio Authenticity Analysis [231];
3. Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition, including Guidance on the Conduct of Proficiency Testing and Collaborative Exercises [232];
4. Forensic Speech And Audio Analysis Working Group Best Practice Guidelines For Enf Analysis In Forensic Authentication Of Digital Evidence [233].

Il primo manuale, riguarda l'analisi delle registrazioni audio di persone sconosciute da riconoscere attraverso il confronto con registrati audio di persone

note. Non integra il riconoscimento automatico o semi automatico del parlante, ma si basa sull'uso di esperti nel campo della fonetica e delle scienze correlate all'analisi del parlato che dovranno effettuare un'analisi uditiva e acustica del materiale vocale e fare una interpretazione dei risultati, al termine della quale dovrà redigere una perizia da presentare in tribunale [230]. Il secondo manuale, si occupa dell'analisi forense dell'autenticità delle registrazioni audio digitali, con particolare riguardo alle risorse, alla disponibilità dei metodi di analisi scientificamente validati, alla garanzia di qualità, la gestione delle registrazione durante l'analisi e la guida alle interpretazioni. Tralasciando i metodi di raccolta delle registrazioni, che presuppone seguano metodi forensi, gestisce tutte le operazioni dal momento della consegna del materiale, munito di relativa richiesta di esame, in poi [231]. Il terzo manuale, si occupa di *Forensic Semiautomatic Speaker Recognition* (FSASR) e *Forensic Automatic Speaker Recognition* (FASR), in cui il perito forense è chiamato a deporre in tribunale dimostrando la fondatezza delle prove, verbali e ipotesi definite, basandosi sull'analisi e il confronto delle registrazioni dei vari casi giuridici. Il livello di fondatezza viene determinato attraverso il rapporto di verosimiglianza (LR) [232]. Il quarto manuale, si occupa di individuare quei fattori che possono introdurre anomalie durante la registrazione di un audio, come discontinuità, la probabile contaminazione accidentale derivante dalla frequenza della rete elettrica (ENF) che alimenta il registratore ecc. Per stabilire l'autenticità e integrità delle prove registrate, il manuale pone l'attenzione su:

- Originalità della registrazione: cioè se il supporto era vergine, clonata, trasferita ecc.
- Originalità del supporto: nuovo o sovrascritto;
- Alimentazione del registratore;
- Impostazioni del registratore;
- Procedura di registrazione e arresto;
- Modalità di salvataggio;
- Dati temporali;
- Sequenza degli eventi;
- Modifiche o riparazioni dei file registrati

L'introduzione di contaminazione ENF, invece, dipende da molti fattori tra cui le specifiche tecniche del registratore, il formato della registrazione, l'ubicazione del registratore ecc [233].

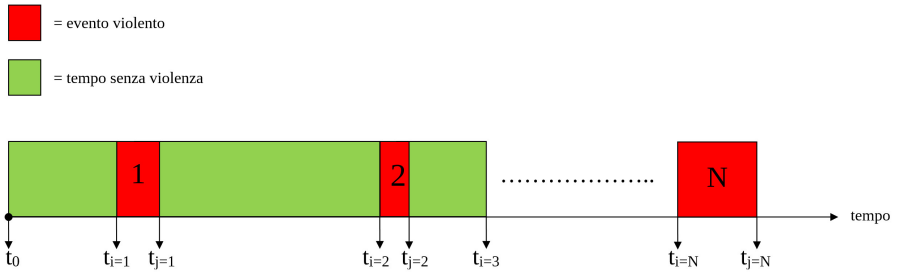


Figura 6.11: Esempio di finestra temporale in cui è evidenziata l'alternanza tra momenti di quiete, in cui non si manifestano atti di violenza, e eventi violenti.

L'obiettivo della nostra ricerca nell'ambito dell'analisi audio, è quello di realizzare un software (ad esempio potrebbe essere un'app scaricabile sul proprio telefono smartphone) da utilizzare non come monitoraggio di massa, ma come strumento di difesa personale, che sia in grado di:

- attivarsi autonomamente, in concomitanza di un evento violento (colpi, grida, parole chiave, richieste d'aiuto, ecc.;
- riconoscere dei dialoghi associabili a comportamenti volenti;
- registrare l'audio e video durante l'atto violento, per testimoniare l'attività in corso;
- attivare le fotocamere proprie e dei dispositivi associati, per esempio interagendo con le centraline della videosorveglianza domestica, ecc.;
- acquisire tutti i segnali bluetooth nel raggio d'azione del dispositivo;
- calcolare un coefficiente di aggressività c_a 6.19 degli eventi;
- depositare in un repository tutti i dati acquisiti, per non perdere lo storico delle violenze subite, utili per chiarire il quadro indiziario in sede di processo.

Il coefficiente di aggressività, è un valore numerico che non è indice del livello di violenza degli atti, ma ne deve evidenziare la pericolosità come espressione della dinamica dell'istinto aggressivo sulla base della frequenza degli eventi, della tendenza all'aumento o alla diminuzione dei fenomeni nel tempo, della durata degli stessi ecc.

Simulando una finestra temporale di monitoraggio come riportato in figura 6.11, dato il numero degli eventi violenti N, iniziati agli istanti di tempo t_i e subiti fino a t_j, con $i = j = (1, 2, \cdots, N)$ a partire dall'istante iniziale del

monitoraggio t_0, possiamo definire una serie di coefficienti utili a capire le dinamiche temporali degli eventi violenti:

- la frequenza f 6.16, è il numero degli eventi violenti N accaduti nel tempo di osservazione $\Delta t_{(0,j=N)}$; tale valore, auspicabilmente molto minore di 1, tanto più è alto, tanto più gli eventi violenti sono frequenti;

- il coefficiente di tregua c_t 6.17, è il rapporto tra il tempo senza violenze trascorso tra gli ultimi due eventi e il tempo medio tra tutti gli eventi consecutivi precedenti; un valore di $c_t > 1$ tra gli ultimi eventi è indice di positività in quanto mediamente c'è sempre più distanza tra gli ultimi eventi, probabilmente qualcosa sta incidendo positivamente sulla tendenza alla violenza; invece un valore di $c_t \leq 1$ indica costanza o un ravvicinamento degli eventi e quindi un probabile segnale di peggioramento della situazione;

- il coefficiente di durata c_d 6.18, è il rapporto tra la durata dell'evento N e la durata media degli eventi precedenti; un valore di $c_d < 1$ tra gli ultimi eventi è indice di positività perché sta a segnalare mediamente una diminuzione della durata degli atti violenti nell'ultimo periodo, contrariamente il ripetersi di un valore $c_d \geq 1$ indica invarianza o aumento della durata dei comportamenti violenti e quindi un probabile peggioramento della situazione.

$$f = \frac{N}{\Delta t_{(0,j=N)}} \tag{6.16}$$

$$c_t = \frac{\Delta t_{(i=N,j=(N-1))}}{\frac{\sum_{i=j=1}^{N-1} \Delta t_{(i,j-1)}}{N-1}} \tag{6.17}$$

$$c_d = \frac{\Delta t_{(i=N,j=N))}}{\frac{\sum_{i=j=1}^{N-1} \Delta t_{(i,j)}}{N-1}} \tag{6.18}$$

Con i coefficienti appena definiti possiamo determinare un coefficiente di aggressività come:

$$c_a = \frac{f \cdot c_d}{c_t} \tag{6.19}$$

In questo modo si può disporre di un pratico indicatore tendente ad aumentare con gli effetti negativi della frequenza, della tregua e della durata degli eventi violenti.

I dati acquisiti, attraverso un approccio di tipo risk-based, potrebbero essere interpretati ed in funzione dei criteri di rischio associati, potrebbero attivare

delle azioni in difesa della vittima. Un esempio molto sintetico di interpretazione e azioni correlate è riportato in figura 6.12, in cui l'audio "catturato" potrebbe essere interpretato nel seguente modo:

1. **Rischio elevato**: c'è un'azione molto violenta in atto (la vittima urla, chiede aiuto, implora di smettere, si avvertono urti, colpi, oggetti che si rompono, bambini che piangono, c_a alto, ecc.) bisogna intervenire tempestivamente per verificare la concretezza del pericolo:
   - Attivare il GPS dello smartphone della vittima;
   - Attivare tutte le videocamere dello smartphone della vittima;
   - Allertare le Forze dell'Ordine indicando la posizione google-maps;
   - Trasmettere in real-time l'audio e video alle Forze dell'Ordine;
   - Registrare tutti i segnali bluetooth raggiungibili dallo smartphone della vittima;
   - Depositare le registrazioni etichettate e i dati correlati su un repository in cloud.
2. **Rischio medio**: c'è un litigio violento in corso (la vittima urla, è vessata, si sentono bambini che piangono, c_a medio, ecc), può essere attivata una campagna di coscienza e assistenza alla vittima:
   - Segnalare il fenomeno agli istituti preposti che, entrando in contatto con la vittima, possono suggerire le azioni migliori da intraprendere attraverso un percorso di rassicurazione e presa di coscienza della realtà dei fatti e suggerire percorsi e trattamenti destinati agli autori di condotte violente (come la *Casa di Accoglienza delle Donne Maltrattate*) [234];
   - Segnalare alle Forze dell'Ordine per valutare eventuali azioni di prevenzione/repressione (anche attraverso app dedicate come *Youpol*[6])[235];
   - Depositare le registrazioni etichettate su un repository in cloud.
3. **Rischio basso**: c'è un litigio in corso (la vittima urla, è vessata, c_a basso, ecc), può essere l'inizio di un atteggiamento violento destinato a crescere, la vittima potrebbe sottovalutarlo:
   - Attivare un servizio di messaggistica che informi la vittima sui rischi correlati alla violenza di genere e suggerisca eventuali possibili soluzioni per rassicurarla e proteggerla (come il chatbot *NONPOSSOPARLARE* [236];

[6]Scaricabile tramite Google Play su: https://play.google.com/store/apps/details?id=it.poliziadistato.youpol&hl=it&gl=US

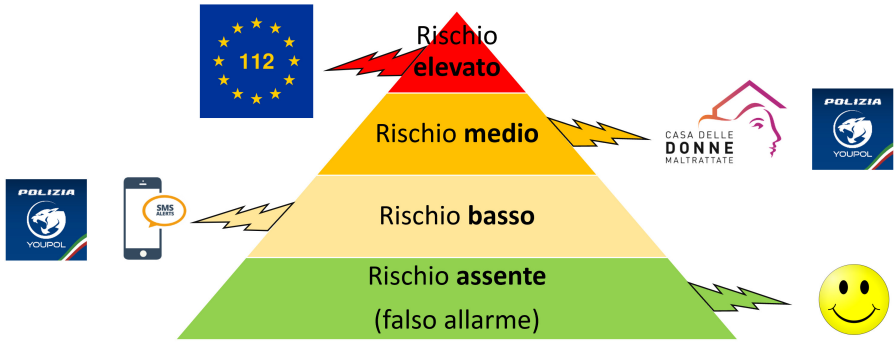


Figura 6.12: Esempio di approccio Risk-based per valutare eventuali azioni automatiche derivanti dall'interpretazione del livello di rischio.

- Segnalare alle Forze dell'Ordine per valutare eventuali azioni di prevenzione (anche attraverso app dedicate come *Youpol*[6])[235];
- Depositare le registrazioni etichettate su un repository in cloud.

4. **Rischio assente**: c'è una discussione animata in corso, ma la durata è breve, c_a molto basso, non si ravvedono né comportamenti violenti ne azioni da intraprendere.

Per capire l'importanza e l'urgenza di affrontare questo argomento, vale la pena di individuare uno dei possibili target che potrebbe trarne beneficio, riportando qui di seguito un breve spaccato di realtà purtroppo attuale.

Il fenomeno della violenza di genere

Il 25 di novembre di ogni anno, si celebra la *giornata internazionale per l'eliminazione della violenza contro le donne*, istituita dall'ONU nel 1999, forse il fenomeno globale più frequente, diffuso e spregevole di violazione dei diritti umani. Un problema così grave che ha visto costretta anche l'Assemblea Generale delle Nazioni Unite ad emettere nel 1993 una "*Dichiarazione sull'eliminazione della violenza contro le donne*, riconoscendo l'urgente necessità di applicare universalmente alle donne i diritti e i principi riguardanti l'uguaglianza, la sicurezza, la libertà, l'integrità e la dignità di tutti gli esseri umani [237]. In Italia la Polizia di Stato, è da anni impegnata in un progetto denominato "*...questo NON è AMORE*", una campagna di sensibilizzazione sociale in cui anche il capo della Polizia Vittorio Pisani invita alla cooperazione di tutti per contrastare questa forma di violenza, senza fermarsi alla sola indignazione [238]. All'interno della brochure relativa all'edizione 2023[7], sono indicati alcuni dati,

[7]...questo NON è AMORE (2023), disponibile su: https://www.poliziadistato.it/statics/24/opuscolo-2023.pdf

forniti dalla Direzione Centrale Anticrimine, che appare doveroso riportare per cogliere con chiarezza la dimensione del fenomeno:

- **85 donne ogni giorno** nel 2023 sono state vittime di reati di maltrattamenti in famiglia, violenza sessuale e stalking;
  - 4 volte superiore alle vittime di sesso maschile;
  - nel 55% dei casi l'autore è convivente;
- **59 donne nel 1°semestre 2023** sono state vittime di omicidio volontario;
  - 31 in ambito familiare;
  - 16 "femminicidi";
  - 2 i casi in cui l'autore aveva precedenti specifici;
- 52% sono i femminicidi il cui autore è convivente/marito;
- 14% sono i femminicidi il cui autore è ex-convivente/marito;
- 14% sono i femminicidi il cui autore proviene da relazioni extraconiugali;
- 20% sono i femminicidi provenienti da altre relazioni;
- 33% sono i casi in cui la vittima lascia figli piccoli;
- 75% sono i casi in cui la vittima è italiana.

Un esame più ampio e ugualmente preoccupante, è stato presentato l'11 dicembre 2023 presso la Direzione Centrale della Polizia Criminale, in un report dal titolo "*Il Punto - Il pregiudizio e la violenza contro le donne*"[8]. Un'analisi effettuata dal Servizio Analisi Criminale, che esamina il fenomeno attraverso i dati contenuti nella Banca dati delle Forze di polizia. All'interno dell'intervallo di osservazione considerato, sono emersi numeri importanti che riguardano sia i casi di omicidio sia i casi dei cosiddetti "reati spia", in generale è emerso che:

- **109 donne uccise**: nel periodo tra il 1° gennaio e il 3 dicembre 2023, di cui:
  - 90 uccise in ambito familiare/affettivo;
  - 58 assassinate dal partner/ex partner;
- **12.491 vittime di atti persecutori** (stalking): nel periodo compreso tra gennaio e settembre 2023, di cui:
  - 74% vittime donne;

[8] Disponibile su:https://www.interno.gov.it/sites/default/files/2023-12/il_punto_-_il_pregiudizio_e_la_violenza_contro_le_donne.pdf

- **16.599 vittime di maltrattamenti contro familiari e conviventi**: nel periodo compreso tra gennaio e settembre 2023, di cui:
  - 81% vittime donne;
- **4.341 vittime di violenza sessuale**: nel periodo compreso tra gennaio e settembre 2023, di cui:
  - 91% vittime donne;

Numeri emersi da casi accertati, ai quali vanno ad aggiungersi i casi che rimangono silenti[239]. Questi risultati purtroppo non hanno bisogno di commenti per descrivere quanto grave e attuale sia la situazione di questo comportamento ignobile. Anche il presidente della Repubblica Italiana Sergio Mattarella, nella dichiarazione in occasione della giornata Internazionale per l'eliminazione della violenza contro le donne del 25 novembre 2022[240] e durante gli auguri di fine anno del 20 dicembre 2023[9], nel suo discorso ha voluto richiamare la sensibilità dei cittadini ad adoperarsi nella cooperazione per ridurre quanto più rapidamente gli atti di violenza contro le donne. Inoltre, ha promulgato la LEGGE 24 novembre 2023, n. 168, in vigore dal 9 dicembre 2023, recante le disposizioni per il contrasto della violenza sulle donne e della violenza domestica[241]. Una legge che contiene importanti novità giuridiche, una delle quali è disciplinata dall'articolo n.10 che prevede la possibilità di arresto in *flagranza differita* per l'autore del reato, sulla base dei dati videofotografici o altra documentazione informatica o telematica, legittimamente ottenuti[10].

Relativamente alla ricerca scientifica per contrastare questo fenomeno, si è pronunciata anche l'ONU nella Dichiarazione sull'eliminazione della violenza contro le donne, in cui all'art.4 invita gli Stati a condannare i comportamenti violenti contro le donne, suggerendo alla lettera (K)[237] di:

- promuovere la ricerca;
- raccogliere dati e compilare statistiche, in particolare riguardanti la violenza domestica, relative alla prevalenza delle diverse forme di violenza contro le donne;
- incoraggiare la ricerca sulle cause, la natura, la gravità e le conseguenze della violenza contro le donne;

[9] Disponibile su:https://www.youtube.com/watch?v=hj9CxDnarqI

[10] Art. 10 Arresto in flagranza differita 1. Dopo l'articolo 382 del codice di procedura penale è inserito il seguente: « Art. 382-bis (Arresto in flagranza differita). - 1. Nei casi di cui agli articoli 387-bis, 572 e 612-bis del codice penale, si considera comunque in stato di flagranza colui il quale, sulla base di documentazione videofotografica o di altra documentazione legittimamente ottenuta da dispositivi di comunicazione informatica o telematica, dalla quale emerga inequivocabilmente il fatto, ne risulta autore, sempre che l'arresto sia compiuto non oltre il tempo necessario alla sua identificazione e, comunque, entro le quarantotto ore dal fatto ».

- incoraggiare la ricerca sull'efficacia delle misure attuate per prevenire e rimediare alla violenza contro le donne;
- pubblicare i risultati e le statistiche delle ricerche.

Accogliendo questi incoraggiamenti come un appello, si riporta nella sezione che segue lo studio che stiamo affrontando sul rilevamento automatico di situazioni di pericolo dall'analisi audio, attraverso l'uso di reti neurali convoluzionali (CNN), in quanto offrono soluzioni efficaci per questo problema grazie alla loro capacità di estrarre caratteristiche complesse da immagini, video e audio.

6.5.1 Tecniche di filtraggio audio per migliorare il riconoscimento automatico dei comportamenti violenti

Nelle indagini di Polizia Giudiziaria, i tracciati audio generalmente provengono da intercettazioni ambientali che sfruttano i segnali captati da sensori, come microfoni o videocamere, nascosti negli ambienti da controllare. I segnali così ottenuti devono essere poi processati per ottenere le informazioni utili agli scopi forensi per cui sono state raccolte, un'attività che per la Polizia Scientifica è demandata alla Sezione indagini Elettroniche[242]. In ambito di ricerca scientifica, ci sono molte pubblicazioni in letteratura riguardo alle tecniche di estrazione delle caratteristiche di file audio concernenti il linguaggio naturale, che per la sua natura possiede un'infinità di caratteristiche distinte che dipendono da molti fattori come il sesso, l'età, l'origine geografica prevalente ecc. Tra le più recenti, Feng et al. propongono un metodo inclusivo di riconoscimento automatico dell'autore, il cosiddetto *Automatic Speech Recognition* (ASR), in cui utilizzando due diversi modelli di reti neurali basate su architetture SotA ASR, vanno a rilevare e quantificare in modo sistematicamatico i bias nel riconoscimento vocale rispetto al genere, all'età, agli accenti regionali e agli accenti non nativi [243]. Riguardo alle tecniche dedicate a scoprire file artefatti rispetto all'originale, il cosiddetto rilevamento *Deepfake Audio*, di recente Chakravarty et al. hanno proposto un architettura di rete neurale basata su ResNet50 modificata per estrarre le caratteristiche dell'audio, analizzando lo spettrogramma audio Mel dei file. Di fatto, rispetto ai metodi tradizionali, il metodo proposto ha dimostrato di possedere prestazioni migliori per l'estrazione delle caratteristiche, ottenendo valori di *Equal Error Rate* (EER) dello 0,4% e una precisione del 99,7%.[244]. Riscontrando quindi la molteplicità delle caratteristiche che un file audio può intrinsecamente possedere, sorge anche l'esigenza di flessibilità nell'analisi audio, definendo flussi di lavoro differenti, ottimizzati a seconda dell'obiettivo che si intende perseguire nell'analisi. A tale scopo, Puglisi et al. hanno proposto un framework vocale open source, chiamato *Deep Audio Analyser*, in grado di visualizzare le caratteristiche audio del parlato e creare nuovi

flussi di lavoro per l'analisi dei file, personalizzando così il framework secondo i propri obiettivi, attraverso la combinazione agevolata di diversi modelli di reti neurali profonde, con il vantaggio velocizzare la ricerca e la sperimentazione pratica oltre alla possibilità di fare anche valutazioni comparative prestazionali [245].

Relativamente al riconoscimento della violenza negli audio, Fime et al. hanno condotto un recentissimo studio in cui hanno proposto un sistema automatizzato basato su Android per rilevare situazioni violente riconosciute dall'audio ascoltato nell'ambiente circostante la vittima. Un aspetto interessante della loro ricerca ha riguardato la rimozione del rumore circostante, di vario tipo, utilizzando varie tecniche di filtraggio come riduzione del rumore mediano, la riduzione del rumore centroide, Audio DeNoise ecc[246]. Ma soprattutto è emerso dai risultati ottenuti, che il filtraggio incide sulla precisione della classificazione [246]. Alla luce di ciò abbiamo deciso di indagare più approfonditamente sulle tecniche di filtraggio per cercare una strategia di miglioramento della precisione di classificazione nel riconoscimento automatico dei comportamenti violenti, utilizzando gli stessi dati audio prodotti e utilizzati nella ricerca da Fime et al.[247]. Il dataset utilizzato contiene 19375 file audio in totale, di cui 6488 si riferiscono a situazioni di bambini in pericolo, 6427 riguardano donne in pericolo e 6460 sono registrazioni in situazione normali[247]. La struttura è del tipo di seguito indicato:

- DangerDetection
  - test
    - Child
    - Normal
    - Women
  - train
    - Child
    - Normal
    - Women

Al fine garantire un equilibrato bilanciamento tra le diverse classi, la raccolta dei dati è stata accuratamente curata e operata una divisione 70%/30% tra dati di addestramento e dati di test. Tale struttura consente di ottenere modelli di machine-learning affidabili e robusti, in grado di riconoscere con precisione le urla di pericolo.

L'insieme dei file audio utilizzati per l'addestramento e la valutazione delle reti neurali è suddiviso in tre classi:

1. situazioni normali

Tabella 6.36: Suddivisione del dataset utilizzato, con una distribuzione del 70% per i dati di train e del 30% per i dati di test.

Classe	Train	Test	Totale
Situazioni normali	4522	1938	6460
Donne in pericolo	4499	1928	6427
Bambini in pericolo	4544	1944	6488
Totale	13565	5810	19375

2. donne in pericolo

3. bambini in pericolo

In questa prima fase dello studio, si è deciso di eseguire l'addestramento del modello su un campione di soli 1000 audio per ciascuna delle 3 classi, per un totale di 3000 audio. Per i test, è stato utilizzato un campione di 500 audio per classe, per un totale di 1500 audio. La scelta di utilizzare un campione ridotto è stata dettata dalla necessità di velocizzare il tempo di addestramento, sfruttando solo le risorse di Google Colab, avendo eseguito i test in cloud. Il train del modello sull'intero set di dati a disposizione avrebbe richiesto un tempo considerevole. Pertanto la possibilità di utilizzare una GPU dedicata, che avrebbe potuto ridurre significativamente il tempo di addestramento, verrà affrontata in seguito qualora i test effettuati produrranno risultati promettenti e validi a giustificarne l'investimento di fondi. Il campione è stato selezionato in modo da essere bilanciato rispetto alle diverse classi. Questo significa che ogni classe è rappresentata da un numero uguale di audio nel campione. Tuttavia l'utilizzo di un campione ridotto di dati, non ha compromesso considerevolmente l'andamento dei test, di fatto ha comportato una riduzione del tempo di addestramento pur mantenendo un'accuratezza accettabile.

Pre-processamento dei file audio

L'input delle reti neurali, è generalmente costituito da file immagini, pertanto i dati audio devono essere pre-processati e trasformati in un formato adatto all'apprendimento automatico. In questo caso, gli audio non vengono trattati come file audio tradizionali, ma convertiti in immagini sotto forma di spettrogrammi Mel. L'utilizzo di immagini per l'elaborazione audio presenta diversi vantaggi:

- sono più facili da elaborare per le reti neurali rispetto ai segnali audio grezzi;

- possono essere utilizzate con una varietà di reti neurali pre-addestrate;
- possono essere visualizzate e interpretate più facilmente rispetto ai segnali audio grezzi.

Per far ciò come primo layer di ogni rete si utilizza la funzione melspectrogram dalla libreria Kapre [248] [11], tale funzione restituisce lo spettrogramma Mel che permette di considerare l'audio come immagine per addestrare e fare previsioni con le reti. Uno spettrogramma Mel è uno spettrogramma in cui le frequenze vengono convertite nella scala Mel. Il nome deriva dalla parola "melodia", per indicare che la scala si basa sul confronto delle altezze dei suoni. Essa, infatti, è una scala percettiva dell'altezza di un suono. Il punto di riferimento tra questa scala e la normale misurazione della frequenza è definito assegnando un valore percettivo di 1000 Mels ad un tono di 1000 Hz. Al di sopra dei 500 Hz l'orecchio umano non distingue più gli incrementi di frequenza [249]. In figura 6.13 è riportato un esempio di spettrogramma in cui l'asse delle ordinate rappresenta la scala Mel e le variazioni di colore rappresentano la variazione della densità di potenza (proporzionale all'ampiezza) del segnale espressa in dB, la quale si rappresenta con diverse tonalità di colore. Per calcolare lo spettrogramma Mel si sottopone il segnale ad un filtraggio con un opportuno banco di filtri triangolari passa basso, passa banda e passa alto tipicamente utilizzati per la decomposizione spettrale e la composizione dei segnali. Il primo filtro è molto stretto e da un'indicazione di quanta energia è presente vicino alla frequenza continua (0 Hz), man mano che le frequenze aumentano i filtri si allargano, in quanto è importante sapere in maniera approssimata quanta energia si trova in corrispondenza di frequenze più alte [250].

Riduzione del rumore

Prima di alimentare la rete neurale con i dati audio per il calcolo degli spettrogrammi e il successivo addestramento, è opportuno applicare diverse tecniche di riduzione del rumore per "pulire" i dati e ottenere risultati migliori. In questo progetto, vengono utilizzate tre metodologie:

1. Noise Reduce[12];
2. Filtro passa alto
3. Filtro mediano

[11] Disponibile su: https://github.com/keunwoochoi/kapre
[12] Disponibile su: https://github.com/timsainb/noisereduce

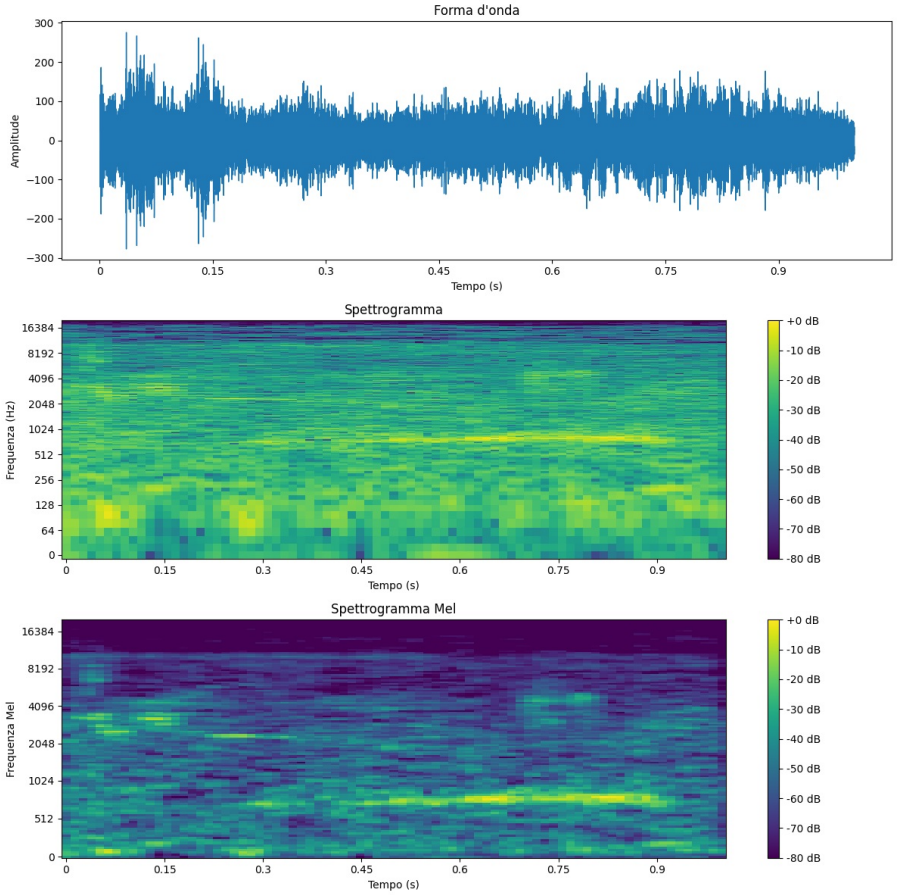


Figura 6.13: Esempio di spettro audio e spettrogramma Mel correlato.

Noise reduce

Noise reduce è un algoritmo di riduzione del rumore implementato in Python che riduce il rumore nei segnali nel dominio del tempo[251]. Si basa su un metodo chiamato "gating spettrale" che è una forma di Noise Gate. Il suo funzionamento si basa sul calcolo dello spettrogramma di un segnale stimandone la soglia di rumore per ciascuna banda di frequenza di quel segnale/rumore. Tale soglia viene utilizzata per calcolare una maschera che porta il rumore al di sotto della soglia di variazione della frequenza[252]. In questo lavoro si è scelto un approccio di riduzione del rumore non stazionario.

Filtro passa alto

Per realizzare un filtro passa alto, è possibile modificare un filtro Butterworth, per sfruttare le sue caratteristiche piatte della sua risposta in frequenza, facendolo lavorare ad una frequenza di taglio specifica in modo da attenuare le basse frequenze, lasciando inalterate le altre. La risposta in frequenza di un filtro analogico Butterworth di ordine N può essere definita matematicamente come:

$$|H(j\omega)| = \frac{1}{\sqrt{1 + (\frac{\omega}{\omega_c})^{2N}}} \tag{6.20}$$

in cui $|H(j\omega)|$ è la risposta in frequenza del filtro assumendo $j = \sqrt{-1}$ e ω la frequenza angolare, mentre ω_c è la frequenza di taglio del filtro [253]. Trà gli effetti visibili nell'uso dei filtri passa alto di Butterworth si osserva l'assenza dell'effetto c.d. *ringing*, i bordi appaiono appaiono con distorsioni minori, inoltre evidenziano un comportamento comune ai filtri passa alto ideali per quanto riguarda i piccoli oggetti e l'effetto spot [254].

Filtro mediano

Il filtro mediano, è generalmente usato per ridurre gli effetti del rumore casuale, il processo di filtraggio si ottiene facendo scorrere una finestra sopra una serie di valori campione di dati, selezionando, come output, il valore mediano [255].

In effetti, l'operazione di filtraggio mediano non è altro che un ordinamento dopo che ogni nuovo elemento viene introdotto nella finestra dei dati. Il valore che si trova a metà strada tra tutti i valori nella finestra è l'output. Il valore di ciascun campione audio verrà sostituito con l'output prodotto calcolando la mediana dei campioni circostanti, all'interno di una finestra di dimensione scelta.

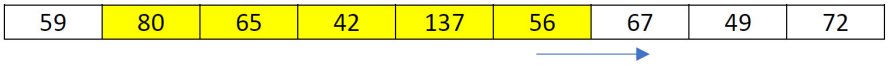


Figura 6.14: Esempio di finestra scorrevole di un filtro mediano su un campione di dati.

Ad esempio, supponiamo che la nostra finestra abbia dimensione 5 e contenga i valori riportati in figura 6.14. La mediana degli elementi della finestra del filtro è 65, poiché all'interno della finestra sono presenti due numeri maggiori di 65 (80 e 137) e due numeri minori di 65 (42 e 56). Mentre la finestra scorre verso destra, il filtro mediano sceglie sempre il valore centrale dei cinque valori come output. Si noti che il numero anomalo in questo esempio, 137, non verrà mai scelto come mediana in nessuno dei raggruppamenti di finestre dati di cui è membro, poiché il suo valore è maggiore di qualsiasi numero con cui è raggruppato. La lunghezza della finestra dei dati viene solitamente scelta in base all'ampiezza prevista per gli impulsi di rumore impulsivo. Nelle applicazioni in cui i picchi di rumore non sono più larghi di uno o due campioni, la finestra con 5 elementi si ritiene adeguata. Nelle applicazioni in cui i picchi di rumore possono essere ampi, sarà necessario utilizzare una finestra dati più lunga. Nel processamento di un'immagine, quando viene attraversato un bordo, un lato o l'altro domina la finestra e l'output cambia bruscamente tra i valori. Grazie a questo fenomeno, si ottiene un bordo meno sfocato. Tra gli svantaggi di tali filtri, però, si rileva che con le immagini che possiedono un basso rapporto segnale-rumore, i bordi tendono ad essere frammentati, producendo così falsi bordi di rumore e non possono sopprimere le distribuzioni di rumore con andamento gaussiano [255].

I modelli di reti proposti

Seguendo un approccio del tipo proposto da Fime et al. [246], nel nostro esperimento sono state utilizzate cinque reti neurali convoluzionali per la classificazione di audio convertiti in immagini tramite spettrogrammi Mel.

Le reti sono:

1. Inception V3: molto efficace per la classificazione di immagini complesse [256];

2. MobileNet V2: modello leggero ed efficiente; particolarmente adatto per dispositivi mobili e applicazioni con risorse limitate [257];

3. ResNet 50: modello profondo con 50 strati convoluzionali: ottima per la classificazione di immagini con un elevato numero di classi [258];

4. ResNet 101: versione più profonda di ResNet 50 con 101 strati convoluzionali; ancora più preciso di ResNet 50, ma richiede più tempo e risorse computazionali per l'addestramento [259];

5. Xception: molto efficace per la segmentazione di immagini e la classificazione di immagini con oggetti di dimensioni variabili [260].

Risultati sperimentali

Eseguito il pre-processamento dei dati audio, si è proceduto all'estrazione dei relativi spettrogrammi Mel e all'addestramento delle diverse reti neurali per la classificazione, valutandone le prestazioni di ciascuna. Relativamente alle metriche impiegate per la valutazione dei modelli sono state determinate:

- *accuratezza*: ovvero la percentuale di predizioni corrette rispetto al numero totale di predizioni; fornisce una misura generale dell'efficacia del modello attraverso l'equazione 6.4;

- *precisione*: ovvero un indicatore della numerosità dei risultati corretti nella classificazione di una classe come positiva, ricavata attraverso l'equazione 6.21;

- *sensitivity*: se siamo interessati a riconoscere quanti più classi positive possibili, allora il nostro modello dovrà avere questo valore alto, determinato attraverso l'equazione 6.1;

- *scoreF1*: che combina precisione e sensitivity, ricavato attraverso l'equazione 6.5;

- *errore quadratico medio* (MSE): una metrica statistica utilizzata per valutare la discrepanza tra i valori predetti da un modello e i valori reali osservati, ricavato attraverso l'equazione 6.22, in cui f è il modello che prende un vettore di feature x in input e genera una previsione y; l'indice i della sommatoria, scorrendo da 1 a N, indica i vari punti nel set di dati. La somma delle previsioni meno i valori reali su N indica la media. Elevando al quadrato rimuoviamo il segno negativo e diamo più peso a differenze maggiori;

$$precisione = \frac{TP}{TP + FP} \tag{6.21}$$

$$MSE(f) = \frac{1}{N}\sum_{i=1}^{n}(f(x_i) - y_i)^2 \tag{6.22}$$

Tabella 6.37: Risultati delle accuratezze sui modelli di reti neurali testate

MODELLO	**FILTRO**	**totale**	**normale**	**child**	**women**
resnet50	ORIGINALI	81,86%	95,29%	97,00%	53,60%
	NOISEREDUCE	87,10%	96,11%	89,80%	75,60%
	ALTO	87,30%	97,54%	94,20%	70,40%
	MEDIANO	86,22%	98,16%	93,60%	67,20%
xception	ORIGINALI	90,32%	98,77%	97,40%	75,00%
	NOISEREDUCE	91,26%	99,80%	90,20%	84,00%
	ALTO	87,23%	98,98%	94,80%	68,20%
	MEDIANO	88,11%	95,90%	82,40%	86,20%
ResNet101	ORIGINALI	81,52%	97,54%	94,80%	52,60%
	NOISEREDUCE	88,17%	94,67%	86,40%	83,60%
	ALTO	89,72%	99,39%	92,00%	78,00%
	MEDIANO	86,56%	97,13%	90,00%	72,80%
MobileNetV2	ORIGINALI	88,51%	97,95%	87,00%	80,80%
	NOISEREDUCE	87,63%	98,36%	90,60%	74,20%
	ALTO	90,59%	99,39%	93,20%	79,40%
	MEDIANO	84,27%	93,44%	89,00%	70,60%
InceptionV3	ORIGINALI	91,87%	99,80%	90,60%	85,40%
	NOISEREDUCE	84,07%	96,31%	90,60%	65,60%
	ALTO	89,79%	99,39%	96,00%	74,20%
	MEDIANO	Test	non	ancora	eseguiti

Tabella 6.38: Risultati delle metriche sui modelli di reti neurali testate

MODELLO	FILTRO	precisione	sensitivity	scoreF1	MSE
resnet50	ORIGINALI	0,851	0,82	0,813	0,679
	NOISEREDUCE	0,876	0,872	0,872	0,468
	ALTO	0,886	0,874	0,873	0,484
	MEDIANO	0,877	0,863	0,861	0,529
xception	ORIGINALI	0,916	0,904	0,903	0,375
	NOISEREDUCE	0,914	0,913	0,913	0,347
	ALTO	0,888	0,873	0,871	0,501
	MEDIANO	0,885	0,882	0,883	0,429
ResNet101	ORIGINALI	0,847	0,816	0,809	0,715
	NOISEREDUCE	0,886	0,882	0,884	0,421
	ALTO	0,903	0,898	0,89	0,405
	MEDIANO	0,873	0,866	0,866	0,509
MobileNetV2	ORIGINALI	0,888	0,886	0,886	0,44
	NOISEREDUCE	0,883	0,877	0,877	0,472
	ALTO	0,911	0,907	0,906	0,37
	MEDIANO	0,852	0,843	0,844	0,565
InceptionV3	ORIGINALI	0,92	0,919	0,919	0,323
	NOISEREDUCE	0,854	0,842	0,84	0,601
	ALTO	0,91	0,899	0,898	0,4
	MEDIANO	Test	non	ancora	eseguiti

Conclusioni

I test condotti hanno evidenziato che diverse reti neurali sono in grado di fornire risultati soddisfacenti per il compito prefissato. In particolare, le reti InceptionV3 (con audio originali), Xception (con riduzione rumore "Noise reduce") e MobileNetV2 (con filtro passa alto) hanno ottenuto un'accuratezza elevata (sopra il 90%) mostrate nella tabella 6.37 e prestazioni promettenti, mostrate nella tabella 6.38. Tuttavia, sono possibili ulteriori sviluppi per migliorare ulteriormente i risultati. Un'opzione potrebbe essere quella di implementare un metodo di "pulizia" dei dati più efficace, in particolare per la classe "Women" che risulta essere la meno accurata. La ricerca di algoritmi di denoising più performanti o l'ottimizzazione dei parametri di quelli già utilizzati potrebbe portare a un miglioramento significativo dell'accuratezza per questa classe. Un altro possibile sviluppo potrebbe essere l'implementazione del sistema su un dispositivo mobile, come proposto da Fime et al. [246]. La rete MobileNetV3, in particolare, si è dimostrata efficiente e richiede risorse computazionali limitate, rendendola adatta per questo tipo di applicazioni. L'utilizzo di questa rete permetterebbe di sfruttare le sue potenzialità a bordo di dispositivi portatili, come gli smartphone, smartwatch e tablet, per l'analisi di audio in tempo reale. In definitiva, i risultati ottenuti sono incoraggianti e aprono la strada a diverse possibilità di sviluppo futuro. L'ottimizzazione del sistema e l'implementazione su dispositivi mobili potrebbero rendere questa tecnologia accessibile a un pubblico più ampio e utile per una varietà di applicazioni come sistema di difesa personale.

Capitolo 7

Etica forense: AI-Act

"*L'Intelligenza Artificiale impara, ma noi cosa le stiamo insegnando?*"

Questo recitava uno spot pubblicitario esposto in autunno alle fermate della metropolitana di Milano (figura 7.1). L'Intelligenza Artificiale si sta imponendo con grande prepotenza in tutto il contesto sociale contemporaneo, promettendo scenari virtuosi con altissime aspettative di profitto, ma che lascia anche profondi dubbi sulla sua abilità di tutela globale durante questo processo di sviluppo e diffusione. Già da qualche anno, utilizziamo servizi e prodotti progettati con tecnologie che sfruttano l'Intelligenza Artificiale, come fare acquisti e accedere alle informazioni online, usandoli con fiducia grazie alla grande utilità e comodità che mostrano ma senza preoccuparci troppo circa le precauzioni necessarie a evitare danni qualora qualcosa dovesse andare storto. In un'epoca storica in cui il tempo sembra aver perso la durata, siamo distratti dalle esigenze crescenti e appagati dalla rapidità ed economicità con cui l'Intelligenza Artificiale ce le risolve. La ragionevolezza che ci distingue dagli automi, porta inevitabilmente a interrogarci sulla capacità umana di gestione e controllo di tali tecnologie e sulla presenza di eventuali "costi latenti". Ma in ambito forense come possono essere tradotti questi interrogativi, e soprattutto come possono essere eliminati i "costi latenti"? Questi ed altri motivi, sono tra quelli che hanno consentito all'Intelligenza Artificiale di entrare nelle preoccupazioni politiche, al punto di individuare e sviluppare una strategia di tutela etico-politica sia nazionale che oltre confine.

L'AI-Act, o Regulation on a European Approach for Artificial Intelligence, è una proposta di legge dell'Unione Europea pensata per regolamentare l'uso dell'Intelligenza Artificiale nell'UE. Questa proposta è stata annunciata nel 2021 dalla Commissione Europea come impegno per affrontare le sfide e garantire un utilizzo etico e sicuro dell'Intelligenza Artificiale all'interno del mercato unico europeo [261]. La proposta si articola intorno a diverse categorie di sistemi AI, classificandoli in base al loro livello di rischio: i sistemi considerati ad alto rischio (come quelli utilizzati nella salute, nella sicurezza e nei trasporti) saranno soggetti a regolamentazioni più stringenti, mentre per i sistemi a basso rischio le norme saranno meno severe. In figura 7.2 è riportata una rappre-

Figura 7.1: Spot pubblicitario esposto alle fermate della metropolitana di Milano a ottobre 2023.

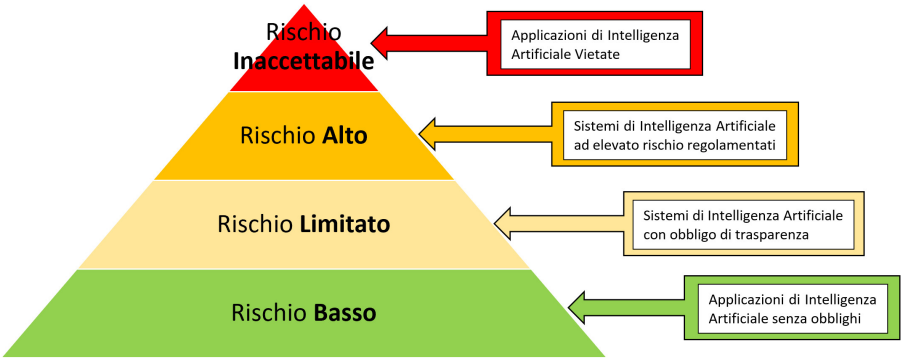


Figura 7.2: Esempio semplificato di approccio Risk-based con cui l'Unione Europea intende regolamentare il ricorso all'Intelligenza Artificiale nella Comunità.

sentazione grafica che mostra schematicamente l'approccio basato sul livello di rischio e le azioni correlate che l'Europa intende adottare nei sistemi che impiegano l'Intelligenza Artificiale. Al vertice della piramide si collocano i sistemi che per loro natura costituiscono una chiara minaccia alla sicurezza, ai mezzi di sussistenza e ai diritti delle persone, pertanto a causa del loro elevato contenuto di rischio sono da considerare inammissibili [261]. Subito dopo, si collocano quei sistemi ad elevato rischio, che hanno un impatto negativo sulla sicurezza delle persone. Relativamente a questa categoria, tendendo alche a minacciare i diritti fondamentali, l'Unione Europea ha pensato di suddividerli in due categorie, una delle quali comprende tra l'altro un ampio settore forense con l'identificazione biometrica e la categorizzazione delle persone fisiche, le Forze dell'Ordine, la gestione della migrazione, dell'asilo, del controllo delle frontiere, dell'Amministrazione della giustizia e dei processi democratici [261]. Relativamente all'attività delle forze dell'ordine, sono previste alcune garanzie ed esenzioni limitate per l'uso dei sistemi di identificazione biometrica, come il riconoscimento automatico dei volti in spazi accessibili al pubblico per scopi di contrasto ai fenomeni illeciti di elevata pericolosità, come per il contrasto al terrorismo, o soggetti ad autorizzazione giudiziaria e per una serie di reati elencati e rigorosamente definiti[262]. Mentre i sistemi di riconoscimento biometrico da remoto sarebbero utilizzati per la ricerca specifica di persone sospettate di aver commesso dei delitti gravi, i sistemi di riconoscimento in tempo reale dovrebbero rispettare delle condizioni ben più stringenti e rigorose, con un impiego che richiede anche una definizione dei parametri temporali e ambientali. Il fine di tali restrizioni sarebbe mirato a :

- ricercare persone sequestrate, vittime di tratta di esseri umani o di sfruttamento della prostituzione;
- svolgere attività di prevenzione in contrasto ad una minaccia terroristica e attuale
- ricercare l'autore di uno dei reati specifici elencati nella norma (es. attentati terroristici o reati correlati al terrorismo, tratta di esseri umani, sfruttamento sessuale, omicidio, sequestro di persona, violenza sessuale, rapina a mano armata, associazione a delinquere, reati ambientali)

Un varco di applicabilità dell'intelligenza artificiale in ambito forense che mira a garantire comunque i diritti fondamentali e la democrazia con senso etico e responsabile [262]. Comunque l'AI Act stabilisce delle regole ben specifiche per la raccolta e il trattamento dei dati, garantendo l'obbligo di documentazione e la trasparenza, inoltre prevede che l'attività dell'Intelligenza Artificiale sia sempre subordinata alla supervisione umana, nonché la certificazione obbligatoria per alcuni tipi di sistemi AI ad alto rischio. [263]. Nell'attività di Polizia

Giudiziaria, che si tratti di prevenzione o repressione di particolari comportamenti antigiuridici, la sfera delle libertà di maggiore impatto per la tutela dei diritti dei cittadini riguarda la libertà personale, per cui affidare a un'intelligenza artificiale il potere di decidere sulla libertà individuale è un aspetto di assoluta criticità. Nella sezione 3.1, abbiamo affrontato il concetto di polizia predittiva, un sistema in grado di predire area e tempo intorno al quale è probabile che si stia per commettere un certo tipo di reato, sulla base di interpretazioni dei dati provenienti da testimonianze e dati investigativi raccolti durante fatti analoghi, accaduti in passato [81]. Stando alla politica dell'AI Act, appare problematica l'applicazione della tecnologia predittiva in quanto le organizzazioni per i diritti civili chiedono il divieto dell'uso indiscriminato o arbitrariamente mirato dei dati biometrici negli spazi pubblici o accessibili al pubblico e restrizioni sull'uso dei sistemi di Intelligenza Artificiale, anche per il controllo delle frontiere e la polizia predittiva [261]. Il motivo risiede nella difficile interpretabilità del concetto di privacy per il monitoraggio di eventi o di particolari comportamenti umani. In questo ambito, il NIST ha proposto una bozza di linee guida sulla valutazione di una tecnica di protezione della privacy per i settori che sfruttano l'Intelligenza Artificiale. In particolare l'idea consiste in un metodo chiamato "Privacy differenziale" pensato per ottenere informazioni utili e precise che potrebbero apportare benefici alla società mantenendo intatta la privacy individuale [264]. L'approccio è di tipo piramidale, in cui la capacità di proteggere la privacy di ogni livello, dipende dal livello sottostante, l'anonimizzazione dei dati si ottiene attraverso un processo di "deidentificazione" in cui non ci sarebbe la possibilità di attacchi di "reidentificazione", o meglio se questi dovessero esserci, il sistema di Privacy Differenziale dovrebbe renderli inefficaci. Tuttavia, come sottolineato nell'articolo del NIST, l'approccio più forte possibile alla privacy è non raccogliere i dati fin dall'inizio [264].

Capitolo 8

Etica nell'industria forense: alcune realtà imprenditoriali

L'uso di tecnologie a elevato contenuto di Intelligenza Artificiale, come abbiamo visto, ha scosso il senso etico dei governi spingendoli ad adoperarsi per promuoverne lo sviluppo ma al tempo stesso tutelando i diritti fondamentali dei cittadini. In ambito forense, essendo un settore che condiziona inevitabilmente la libertà di alcune categorie di individui, il problema è particolarmente critico e necessita di sani princìpi etici da parte dei fornitori di prodotti e servizi ad uso forense.

Per evidenziare come il settore imprenditoriale ha recepito le indicazioni etiche divulgate dai propri governi e come le aziende si adoperano per garantire il loro senso etico, si riportano qui di seguito brevi cenni di alcune affermate realtà industriali che sviluppano interessanti dispositivi ad uso forense.

THALES

Thales è un colosso aziendale che offre soluzioni digitali di sicurezza per infrastrutture critiche, siti sensibili e per il controllo dell'ambiente urbano, operando in molti settori, tra cui i principali sono la Difesa e Sicurezza, Aerospazio e Spazio, Trasporti e Identità digitale. Una interessante tecnologia forense di Thales riguarda il software di riconoscimento facciale LFIS, tale sistema ha ottenuto risultati eccellenti con un tasso di acquisizione del volto del 99,44% in meno di 5 secondi (contro una media del 68%) e con un tasso di errore dell'1% rispetto a una media del 32% [265]. Thales, sulla sua piattaforma on-line, sostiene fermamente un approccio etico e responsabile nell'implementazione dei sistemi di riconoscimento facciale, le sue implicazioni etiche comprendono molti diritti fondamentali, tra cui la privacy, la protezione dei dati personali e il consenso, promettendo di impegnarsi pienamente in un approccio responsabile, in linea con il crescente corpus di norme, regolamenti e leggi che si applicano in questo campo. Relativamente agli aspetti etici più ampi dei sistemi di riconoscimento

facciale nei luoghi pubblici, Thales fissa alcuni interrogativi su cui sviluppare le proprie tecnologie, tra cui:

1. Quali sono i rischi di errori di identificazione e pregiudizi?
2. In che modo i governi e le autorità di regolamentazione proteggono i diritti dei cittadini?
3. Come possono i fornitori garantire che i loro sistemi di riconoscimento facciale siano trasparenti e responsabili?

Oltre alla biometria, l'esperienza di Thales comprende la solida sicurezza informatica necessaria per proteggere i dati biometrici e altri dati personali. Ciò è ulteriormente abbinato a un approccio TrUE AI, un impegno per un'intelligenza artificiale trasparente ed etica che garantisce il rispetto delle leggi e degli standard pertinenti. Come riportato sul suo sito, Thales si propone di sostenere pienamente gli sforzi dei governi e dei cittadini per sviluppare strutture solide ed efficaci che proteggano dagli abusi e forniscano le basi per soluzioni etiche di riconoscimento facciale che offrano sicurezza, comodità e fiducia [265].

INNOVATRICS

Innovatrics è un'altra importante realtà imprenditoriale che fornisce soluzioni biometriche affidabili per governi e imprese. Una delle sue tecnologie più interessanti è sicuramente *Iris*, un sistema di identificazione biometrica basato sulla scansione della retina oculare. Tuttavia, solo una manciata di aziende, rispetto al riconoscimento facciale, forniscono il proprio algoritmo di identificazione dell'iride per la corrispondenza attraverso il confronto. La scansione dell'iride illumina le iridi con luce infrarossa invisibile per scattare un'immagine che descrive motivi unici per ciascun occhio, non visibili a occhio nudo. Una telecamera speciale rileva la posizione della pupilla, dell'iride, delle palpebre e delle ciglia. Ogni occhio ottiene i propri modelli matematici unici, che vengono ulteriormente digitalizzati. Per l'identificazione (1:N) o la verifica (1:1), un modello creato mediante l'imaging di un'iride viene confrontato con il modello memorizzato in un database. L'ultimo algoritmo di identificazione dell'iride di Innovatrics si è posizionato tra i primi 3 sia in termini di precisione che di velocità nei benchmark NIST IREX 10 [266]. Innovatrics nella sua policy etica, in linea con gli obiettivi dell'Europa, ha sposato la ricerca della fiducia, riconoscendo la biometria come portatore di fiducia [267] e fissando la sua politica aziendale ispirandosi a n.6 valori fondamentali:

- qualità;
- agilità;

- ispirare la fiducia attraverso la trasparenza;
- indipendenza;
- lavoro di squadra;
- accettare nuove sfide.

SECOM

La Secom opera, dal 1984, nell'area dei Sistemi Elettronici Computerizzati, ricercando soluzioni originali a tecnologia avanzata, rivolte sia al settore pubblico che al privato. Secom da sempre coopera con le Forze di Polizia italiane, realizzando importanti progetti per la lotta contro il crimine e con una particolare attenzione per l'attività degli operatori che occupano la prima linea nel contrasto alla criminalità [125]. Tra i vari dispositivi di progettazione Secom, a supporto dell'attività delle forze dell'ordine, va rilevata sicuramente la postazione di fotosegnalamento *SPIS/IDENTISYSTEM*, un affermato e complesso sistema di acquisizione che crea un file completo di dati personali e biometrici, foto fronte-profilo, e dettagli rilevanti sul corpo quali: tatuaggi e cicatrici. A partire dal 2024 il nuovo sistema di fotosegnalamento Secom (versione MY23), tra le tante novità introdotte, consente l'acquisizione di 6 foto del soggetto, nr. 5 relative al volto (con inquadrature della fronte, profilo dx, profilo sx, ¾ dx e ¾ sx) più una relativa alla foto in piedi (total body). I file elettronici alimentano un database centrale accessibile via intranet (Weblase e A.F.I.S. Automated Fingerprint Identification System - sistema di identificazione automatica delle impronte) La foto "fronte-profilo" è generata automaticamente con un solo click, nella stessa posizione, nelle stesse condizioni di luce e di distanza, grazie a una tecnologia studiata appositamente e brevettata da Secom. La qualità della foto è direttamente controllata nel rispetto delle normative ICAO attraverso il software SECOM installato sullo SPIS/IDENTISYSTEM [123]. Lo SPIS della Secom acquisisce in tempo reale impronte digitali ruotate, piane e palmari attraverso evoluti scanner elettronici, salvandoli in un formato compatibile con i sistemi di interscambio FBI IAFIS e ANSI-NIST. Una interessante innovazione introdotta da Secom, rispetto al fotosegnalamento canonico, riguarda la possibilità di acquisire il tracciato audio della voce del segnalato. Il sistema genera velocemente la firma vocale del soggetto segnalato, inviandola ad un archivio audio centralizzato. I controlli vocali permettono il confronto della firma vocale del soggetto con quelle presenti in archivio. Oltre alla registrazione vocale, il sistema SPIS-Secom permette di generare tre tipi di IDENTIKIT, attraverso:

- le caratteristiche fisiche grafiche (esistenti) in un database;
- le immagini sezionate presenti nel database generando un identikit fotografico;

- creando un modello 3D, con la possibilità di invecchiamento automatico e cambiamenti somatici [125].

Un'altra interessante progettazione di Secom, riguarda il sistema portatile BIOFAD [268] per il controllo documentale e l'identificazione biometrica in mobilità (volti e impronte digitali), pensato per facilitare le attività di controllo territoriale ma interfacciabile con lo SPIS [269], il centro di competenza (supporto specialistico sul falso documentale) e le banche dati governative (es. AFIS, AMAIS, Agenzie internazionali).

La stessa soluzione BIOFAD è in grado di gestire localmente (off-line) una black-list con informazioni biometriche provenienti dai database (Weblase, HI-Secom e AFIS), funzionalità utile per operare in scenari non coperti da rete intranet/internet; identificare rapidamente una persona attraverso il sensore di impronte digitali e la lettura dei documenti d'identità elettronici e biometrici (es. ePASSPORT) [268, 270].

Si può generare una black-list con informazioni biometriche provenienti dai database (Weblase, HI-Secom e AFIS); identificare rapidamente una persona attraverso il sensore di impronte digitali e con la Smart Card e il lettore di passaporto, vi è la possibilità di controllo dei documenti di identità in tempo reale.

Come è evidente, si tratta di sistemi ad elevato contenuto tecnologico, che coinvolgono in modo massiccio il rilevamento, il trattamento, il trasferimento e la condivisione di dati sensibili. In linea con i fondamenti etici trattati nel capitolo 7, la SECOM dispone di un codice etico proprietario in cui fissa n.9 principi fondamentali:

- infondere fiducia;
- valorizzazione delle risorse umane;
- onestà e senso etico interno;
- trasparenza d'immagine;
- riservatezza dei dati personali;
- imparzialità senza discriminazioni;
- sicurezza e salute nei luoghi di lavoro;
- tutela ambientale;
- concorrenza leale.

Per ragioni di semplicità, sono state citate solo tre realtà imprenditoriali in ambito forense, in cui è stata evidenziala la loro politica etica in ragione della

particolarità dei dispositivi prodotti e dei relativi dati trattati. Di fatto si è rilevato, in via generale, che tutte le aziende stanno recependo le inclinazioni politiche, sia nazionali che oltre frontiera, volte a stimolare il più alto senso etico nello sviluppo dell'Intelligenza Artificiale a bordo dei propri dispositivi. Questo atteggiamento industriale positivo, fissa solide fondamenta formate da una robusta miscela composta da ricerca scientifica, industria e utente finale, proiettando con fiducia lo sviluppo dell'Intelligenza Artificiale verso il futuro.

Capitolo 9

Pubblicazioni scientifiche

In questi tre anni di dottorato, l'attività di ricerca scientifica portata avanti all'interno del laboratorio AIRTLab, in collaborazione con il laboratorio di misure meccaniche e termiche dell'UNIVPM e in stretto contatto con il Gabinetto Interregionale di Polizia Scientifica per le Marche e l'Abruzzo, ha consentito di affermare le potenzialità dell'Intelligenza Artificiale come strumento di implementazione della sicurezza al servizio della comunità, in particolar modo riguardo ai tre ambiti di ricerca rientranti nell'Accordo d'Intesa tra l'UNIVPM, attraverso il centro di ricerca interdipartimentale CARMELO, e la Polizia di Stato attraverso il Gabinetto Interregionale di Polizia Scientifica per le Marche e l'Abruzzo. Senza attingere a dedicati fondi di finanziamento, grazie alle risorse fisiche e umane messe a disposizione dall'UNIVPM, sono stati progettati utili sistemi di acquisizione dati e costruiti i relativi prototipi, realizzati per lo più con materiale di recupero, che si sono dimostrati efficaci a produrre i database di dati necessari alla ricerca. Grazie ai dati acquisiti è stato possibile condurre efficacemente gli studi sugli argomenti dell'Accordo d'Intesa, che nel corso del triennio di dottorato hanno consentito la pubblicazione scientifica dei seguenti papers, alcuni dei quali su rivista scientifica e altri come proceedings in conferenze internazionali:

1. **A dataset for automatic violence detection in videos**, Miriana Bianculli, Nicola Falcionelli, Paolo Sernani, Selene Tomassini, Paolo Contardo, Mara Lombardi, Aldo Franco Dragoni; Data in brief, volume 33, anno 2020; https://www.sciencedirect.com/science/article/pii/S2352340920314682

2. **Deep learning for law enforcement. A survey about three application domains**, Nicola Falcionelli, Paolo Sernani, Paolo Contardo, Aldo Franco Dragoni; CEUR Work Shop Proceedings, volume 2872, anno 2021; https://u-pad.unimc.it/handle/11393/301474

3. **Deep learning for automatic violence detection: Tests on the AIRTLab dataset**, Nicola Falcionelli, Paolo Sernani, Selene Tomassini,

Paolo Contardo, Aldo Franco Dragoni; IEEE Access, volume 9, anno 2021; https://ieeexplore.ieee.org/abstract/document/9627980

4. **Analyzing the impact of police mugshots in face verification for crime investigations**, Nicola Falcionelli, Paolo Sernani, Paolo Contardo, Emanuele Di Lorenzo, Aldo Franco Dragoni; 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), anno 2022; https://ieeexplore.ieee.org/document/9967671

5. **FRMDB: Face recognition using multiple points of view**, Nicola Falcionelli, Paolo Sernani, Selene Tomassini, Paolo Contardo, Milena Martarelli, Paolo Castellini, Aldo Franco Dragoni; Sensors,volume 23, numero 4,anno2023; https://www.mdpi.com/1424-8220/23/4/1939

6. **Evaluating Deep Neural Networks for Face Recognition with Different Subsets of Mugshots From the Photo-Signaling Procedure**, Nicola Falcionelli, Paolo Sernani, Selene Tomassini, Paolo Contardo,Nicolò Rossini, Aldo Franco Dragoni; 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), anno 2023;https://ieeexplore.ieee.org/abstract/document/10405736

Tuttavia, come già detto, il citato Accordo d'Intesa è ancora in corso ed è stato rinnovato fino al 2026. Pertanto continuerà l'attività di ricerca nei tre ambiti, oltre al nuovo ambito aggiunto di recente, a cui seguiranno auspicabilmente ulteriori pubblicazioni scientifiche.

Capitolo 10

Conclusioni

L'intelligenza Artificiale, in tutte le sue forme, si sta via via insediando sostituendosi ai metodi tradizionali, un po in tutti i settori. E anche introducendola nei sistemi investigativi delle Forze dell'Ordine, può apportare grandi benefici a garanzia del diritto e per migliorare il servizio alla collettività. Ma come abbiamo visto, certe applicazioni potrebbero entrare in contrasto con alcune forme di tutela. In questo la ricerca scientifica gioca un ruolo fondamentale e di prim'ordine, ma ogni qualvolta un cambiamento repentino invade un certo settore, come quello forense, deve sforzarsi di lavorare su un duplice binario:

- una parte di ricerca scientifica deve preoccuparsi di sviluppare gli aspetti tecnologici mirati a far evolvere il settore d'interesse;
- una parte di ricerca scientifica deve accertarsi che l'evoluzione del settore d'interesse sia scevro da rischi che minacciano l'integrità dei diritti fondamentali;

A volte però, come abbiamo avuto anche modo di constatare tra i capitoli di questo lavoro, i due binari entrano in contrasto reciproco, un problema di cui i portatori di "interessi sporchi" non se ne curano molto poiché il loro unico scopo riguarda l'evoluzione dei loro affari, spesso senza dare peso alle conseguenze negative. La criminalità, si evolve velocemente introducendo innovazione anche nei loro affari illeciti, ed essendo priva di senso etico, lo fa ad una velocità difficilmente raggiungibile. Il Procuratore della Repubblica presso il Tribunale di Napoli Nicola Gratteri, attivo nella lotta contro la 'Ndrangheta fin dagli inizi della sua carriera, in un recente articolo de "Il Fatto Quotidiano"[1], commentato sul sito "Antimafia2000" nel novembre 2023, ha spiegato che con l'avvento dei social network, la proliferazione delle criptovalute e la navigazione non indicizzata nel Dark Web, gli affari mafiosi hanno avuto l'opportunità di sviluppare i loro affari sfruttando le potenzialità comunicative, di anonimato e svincolate di questi strumenti digitali. Attraverso Dark Web, le associazioni criminali giovano di un ambiente d'affari che consente loro di non essere intercettati delle forze

[1] Disponibile su: https://www.ilfattoquotidiano.it/in-edicola/articoli/2023/11/07/la-mafia-usava-il-faxe-ora-vive-online/7345366/

dell'ordine. Con l'ascesa della "Google Generation Criminale" [271] a partire dal 2016, le mafie hanno iniziato ad affacciarsi al mondo virtuale correlandolo e facendolo interagire con la loro realtà criminale. Sia in termini di spionaggio che di azione, tutte le piattaforme digitali vengono sfruttate come palcoscenico delle loro attività illecite [272].

Di riflesso, le Amministrazioni che si occupano di sicurezza, devono inseguire il progresso tecnologico avvalendosi della Ricerca Scientifica, auspicabilmente con strategie più efficaci che ricomprendano anche il vincolo del senso etico. Una sfida prestigiosa all'interno della quale ben si inquadra anche l'accordo d'Intesa tra il centro di ricerca interdipartimentale CARMELO e la Polizia di Stato, in cui si innesta il mio dottorato di ricerca, siglato nel 2019. La ricerca scientifica portata avanti in questi tre anni, ha dimostrato che in ambito forense c'è ancora molto da indagare, ed è un settore così importante che ha convinto le due Amministrazioni interessate a rinnovare l'accordo fino al 2026, incrementandolo con un ulteriore campo di ricerca. In questo lavoro di tesi, è riportata l'attività di ricerca che ho condotto nei tre anni di dottorato ove, con il gruppo di ricerca del laboratorio AIRTLab dell'Università Politecnica delle Marche, abbiamo inizialmente presentato una breve panoramica circa le applicazioni di deep learning sui tre domini applicativi rientranti nelle missioni dell'accordo d'Intesa.

Il metodo di identificazione dattiloscopica classico trattato nel capitolo 4, per quanto tuttora efficace, oggettivamente trascura elementi misurabili di verifica dell'affidabilità e dell'accertamento dattiloscopico. Alcune valutazioni metriche, come la valutazione statistica delle composizioni di minuzie, il calcolo del rapporto di verosimiglianza e la stima dell'errore, potrebbero essere dei nuovi parametri che se indagati opportunamente, con l'ausilio delle più avanzate tecnologie di computer vision e il Deep Learning, unite all'interpretazione dei risultati delle misurazioni, potrebbero garantire comunque la certezza dell'identificazione anche con minori corrispondenze. Come citato anche nella sezione 4.1.2, l'Unione Europea suggerisce di adottare alcune metodologie nell'accertamento dattiloscopico che sono ritenute le migliori pratiche verso una standardizzazione metodologica condivisa tra gli stati membri. Una guida che sarà presa in considerazione per gli sviluppi futuri di questo campo di ricerca nel tempo residuo in convenzione, integrandola con l'utilizzo del Deep Learning nel riconoscimento dei frammenti di impronte digitali latenti, come potrebbero essere ad esempio quelli rinvenuti sulla scena di un crimine, attraverso il confronto con impronte digitali note.

Relativamente ai rilievi segnaletici, trattati nel capitolo 5, l'obiettivo era indagare se, attraverso un nuovo protocollo operativo di fotosegnalamento, si poteva trovare una risposta convincente al problema "QFQ". L'ostacolo principale da superare è stata la mancanza di dati segnaletici adeguati per affrontare

il problema con le reti neurali, risolto attraverso la produzione del database FRMDB, unico nel suo genere con immagini di segnaletiche multi-prospettive sia su piano zenitale che azimutale e videoriprese degli stessi soggetti. Per questo, il primo passo è stato costruire un prototipo di acquisizione multi-prospettiva, l'MCMPrototype, con cui abbiamo generato il dataset di immagini del database FRMDB. L'MCMPrototype, costruito con materiale per lo più di recupero e a costo quasi zero, per quanto all'apparenza rudimentale, si è dimostrato tecnologicamente all'avanguardia e flessibile al punto giusto per definire accuratamente tutti i parametri necessari all'acquisizione ottimale dei volti. Parallelamente all'MCMPrototype, sono stati costruiti due sistemi di simulazione di videosorveglianza, uno a bassa risoluzione e uno ad alta risoluzione, per simulare gli impianti di videosorveglianza disseminati sul territorio, con cui abbiamo generato il dataset di video del database FRMDB. Con i primi test, abbiamo dimostrato la validità dell'FRMDB come dataset utile per testare il riconoscimento automatico dei volti, attraverso tecniche di Deep Learning in cui la selezione delle immagini era casuale. Successivamente abbiamo cercato di agevolare il riconoscimento automatico, come suggerito dai manuali di buone pratiche dell'ENFSI [110], selezionando manualmente dalle videoriprese della videosorveglianza i fotogrammi che più si avvicinavano alle pose del fotosegnalamento. Questa operazione ha aumentato notevolmente le metriche dei test, consentendoci di individuare tra i test proposti il n.3 come configurazione con il minor numero di immagini del volto che possiede un elevato valore di accuratezza, riscontrando invece nelle pose canoniche del test n.1 il peggior risultato. Un risultato molto importante, che fa luce sul problema "QFQ" e che nel tempo rimanente in convenzione sarà ulteriormente approfondito per definire, in modo automatico, l'individuazione dei fotogrammi delle videosorveglianze che contengono le immagini di volti con pose più vicine alle immagini del fotosegnalamento. Inoltre sarà studiato un nuovo prototipo di acquisizione simultanea e multi-prospettiva, per studiare la minima dimensione delle immagini di fotosegnalamento che conserva le qualità necessarie per un riconoscimento automatico efficace e la ricostruzione tridimensionale del volto, conservando la legacy del fotosegnalamento attuale.

Relativamente al rilevamento automatico della violenza, l'obiettivo principale era studiare un'architettura di rete neurale in grado di riconoscere on-line e off-line scene di violenza. Una tecnologia utile da portare a bordo macchina, per esempio nelle telecamere di videosorveglianza, per il controllo del territorio in aree particolarmente pericolose o per il processamento di lunghi filmati registrati. Abbiamo visto che esistono in letteratura molte ricerche scientifiche che si occupano proprio di questo argomento, ma ci siamo chiesti quale potesse essere il potere discriminante del Deep Learning nei confronti dei falsi positivi. Nei vari repository in rete, vi sono parecchi dataset contenenti video di compor-

tamenti violenti pensati per addestrare le reti neurali al riconoscimento della violenza, ma non abbiamo trovato dei dataset adatti a testare la robustezza delle reti neurali verso i falsi positivi. Per questo, grazie alla partecipazione di attori non professionisti che hanno simulato comportamenti violenti e comportamenti confondibili come violenti, abbiamo prodotto l'AIRTLab dataset. Nella prima fase della ricerca, abbiamo validato il dataset AIRTLab che si è dimostrato adatto a testare le reti neurali. Successivamente abbiamo testato reti convoluzionali 3D per verificare la robustezza delle stesse verso i falsi positivi, valutando i risultati derivati dal confronto tra i test effettuati sul dataset AIRTLab e altri dataset contenenti video violenti. L'esito è stato positivo, di fatto le reti CNN 3D che abbiamo testato si sono rivelate adatte a discriminare correttamente le scene violente da quelle non violente, ma hanno richiesto elevate prestazioni computazionali e capacità di memoria, due caratteristiche che poco si adattano a dispositivi portatili o di piccole dimensioni. Per risolvere il problema, al modello proposto abbiamo sostituito la rete convoluzionale 3D con una rete 2D progettata per dispositivi mobili, mettendo in cascata ad essa un modulo ricorrente. Ripetendo i test sulla nuova architettura, i risultati sono stati molto promettenti, con perdite minimali rispetto alle metriche ottenute con le reti 3D. Pertanto nel tempo residuo in convenzione, continueremo a perseguire questa direzione per individuare l'architettura migliore da installare a bordo macchina. Tuttavia, seppure il controllo del territorio attraverso sistemi di videosorveglianza e il processamento di lunghi filmati per le attività investigative delle forze dell'ordine rimane un argomento di elevata importanza, ma va affrontata seguendo i vincoli imposti dall'AI Act [263], vista la capacità delle reti neurali nel riconoscere la violenza, ci sembrava doveroso condurre anche uno studio per contrastare uno dei più deplorevoli fenomeno di violenza, purtroppo di assoluta attualità e di diffusione globale, che riguarda la violenza di genere. Uno studio che rientra pienamente nei punti dell'accordo d'intesa e che potrebbe avere effetti positivi sia sulle vittime dirette, ad esempio le donne che subiscono atti violenti, sia sulle vittime indirette che potrebbero essere i figli che assistono alle violenze. I maltrattamenti, che derivino da un disturbo mentale o da un comportamento cosciente, sono atti vigliacchi e censurabili che tutti abbiamo il dovere di contrastare, per restituire la dignità alle vittime che subiscono ingiustamente tali violenze ma anche per gli autori di atti violenti affinché, attraverso percorsi di sostegno psicologico, prendano coscienza dei loro comportamenti spregevoli, accettando le offerte di aiuto per rieducare il loro istinto violento. Una comunità che si fonda sulla civile convivenza e sani principi morali, non può prescindere sui comportamenti censurabili, ma deve adoperarsi affinché tutti i consociati possano vivere serenamente, stimolando la cooperazione individuale verso il più elevato senso civico.

Bibliografia

[1] "Fingerprint whorld, Vol.28, No 107, pag.19," jenuary 2002.

[2] S. Pankanti, S. Prabhakar, and A. K. Jain, "On the individuality of fingerprints," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 8, pp. 1010–1025, 2002.

[3] S. Marascio, "Distribuzione dei punti caratteristici delle impronte digitali di 74 italiani di sesso maschile suddivise per figure dattiloscopiche," Rassegna dell'Arma dei Carabinieri n.2, febbraio 2009.

[4] S. Accattoli, P. Sernani, N. Falcionelli, D. N. Mekuria, and A. F. Dragoni, "Violence detection in videos by combining 3D convolutional neural networks and support vector machines," *Applied Artificial Intelligence*, vol. 34, no. 4, pp. 329–344, 2020.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[6] "Costituzione italiana → Parte II - Ordinamento della repubblica → Titolo IV - La magistratura → Sezione I - Ordinamento giurisdizionale," art. 109.

[7] "Funzioni della polizia giudiziaria: Codice di procedura penale → libro primo - soggetti → titolo iii - polizia giudiziaria," art.55.

[8] Università Politecnoica delle Marche, Polizia Scientifica Marche Abruzzo e UnivPM, presentazione nuovo protocollo d'intesa Polizia Scientifica Univpm, Comunicato stampa, 6 giugno 2023. [Online]. Available: https://www.univpm.it/Entra/Magazine_online/Polizia_Scientifica_Marche_Abruzzo_e_UnivPM

[9] Suprema Corte di Cassazione, *II* Sezione Penale, sentenza n.2559 del 14.11.1959.

[10] P. Contardo, *Dattiloscopia 2.0 Nuove prospettive in materia d'identificazione dattiloscopica - Dactyloscopy 2.0 new perspectives in fingerprint identification.* Università Politecnica delle Marche, 2018.

[11] A. Giuliano, *Dieci e tutte diverse. Studio sui dermatoglifi umani*. Tirrenia-Stampatori, 2004.

[12] P. Contardo, P. Sernani, N. Falcionelli, A. F. Dragoni *et al.*, "Deep learning for law enforcement: A survey about three application domains." in *RTA-CSIT*, 2021, pp. 36–45.

[13] J. Haugeland, *Artificial intelligence: The very idea*. MIT press, 1989.

[14] N. Falcionelli, P. Sernani, A. Brugués, D. N. Mekuria, D. Calvaresi, M. Schumacher, A. F. Dragoni, and S. Bromuri, "Indexing the event calculus: Towards practical human-readable personal health systems," *Artificial Intelligence in Medicine*, vol. 96, pp. 154–166, 2019.

[15] N. Falcionelli, P. Sernani, A. Brugués, D. N. Mekuria, D. Calvaresi, M. Schumacher, A. F. Dragoni, and S. Bromuri, "Event calculus agent minds applied to diabetes monitoring," in *Autonomous Agents and Multiagent Systems*. Cham: Springer International Publishing, 2017, pp. 258–274.

[16] A. F. Dragoni and S. Animali, "Maximal consistency, theory of evidence, and bayesian conditioning in the investigative domain," *Cybernetics and Systems*, vol. 34, no. 6-7, pp. 419–465, 2003.

[17] N. Falcionelli, P. Sernani, D. Mekuria, and A. F. Dragoni, "An event calculus formalization of timed automata," in *Proceedings of the 1st International Workshop on Real-Time compliant Multi-Agent Systems co-located with the Federated Artificial Intelligence Meeting*, ser. CEUR Workshop Proceedings, vol. 2156, 2018, pp. 60–76. [Online]. Available: http://ceur-ws.org/Vol-2156/paper5.pdf

[18] A. F. Dragoni, P. Giorgini, and L. Serafini, "Mental states recognition from communication," *Journal of Logic and Computation*, vol. 12, no. 1, pp. 119–136, 2002.

[19] P. Sernani, A. Claudi, and A. F. Dragoni, "Combining artificial intelligence and netmedicine for ambient assisted living: A distributed bdi-based expert system," *International Journal of E-Health and Medical Communications*, vol. 6, no. 4, pp. 62–76, 2015.

[20] P. Sernani, M. Biagiola, N. Falcionelli, D. Mekuria, S. Cremonini, and A. F. Dragoni, "Time aware task delegation in agent interactions for video-surveillance," in *Proceedings of the 1st International Workshop on Real-Time compliant Multi-Agent Systems co-located with the Federated Artificial Intelligence Meeting*, ser. CEUR Workshop

Proceedings, vol. 2156, 2018, pp. 16–30. [Online]. Available: http://ceur-ws.org/Vol-2156/paper2.pdf

[21] D. N. Mekuria, P. Sernani, N. Falcionelli, and A. F. Dragoni, "Reasoning in multi-agent based smart homes: A systematic literature review," in *Ambient Assisted Living*. Cham: Springer International Publishing, 2019, pp. 161–179.

[22] D. N. Mekuria, P. Sernani, N. Falcionelli, and A. Dragoni, "Smart home reasoning systems: a systematic literature review," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–18, 2019.

[23] E. Serral, P. Sernani, A. F. Dragoni, and F. Dalpiaz, "Contextual requirements prioritization and its application to smart homes," in *Ambient Intelligence*. Cham: Springer International Publishing, 2017, pp. 94–109.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[25] S. Raaijmakers, "Artificial intelligence for law enforcement: Challenges and opportunities," *IEEE Security Privacy*, vol. 17, no. 5, pp. 74–77, 2019.

[26] M. Kücken and A. C. Newell, "Fingerprint formation," *Journal of Theoretical Biology*, vol. 235, no. 1, pp. 71–83, 2005.

[27] G. Bebis, T. Deaconu, and M. Georgiopoulos, "Fingerprint identification using delaunay triangulation," in *Proceedings 1999 International Conference on Information Intelligence and Systems (Cat. No.PR00446)*, 1999, pp. 452–459.

[28] D. Maio and D. Maltoni, "Direct gray-scale minutiae detection in fingerprints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 1, pp. 27–40, 1997.

[29] A. Farina, Z. M. Kovács-Vajna, and A. Leone, "Fingerprint minutiae extraction from skeletonized binary images," *Pattern Recognition*, vol. 32, no. 5, pp. 877–889, 1999.

[30] H. Fronthaler, K. Kollreider, and J. Bigun, "Local features for enhancement and minutiae extraction in fingerprints," *IEEE Transactions on Image Processing*, vol. 17, no. 3, pp. 354–363, 2008.

[31] D. Peralta, M. Galar, I. Triguero, D. Paternain, S. García, E. Barrenechea, J. M. Benítez, H. Bustince, and F. Herrera, "A survey on fingerprint

minutiae-based local matching for verification and identification: Taxonomy and experimental evaluation," *Information Sciences*, vol. 315, pp. 67–87, 2015.

[32] R. Cappelli, M. Ferrara, and D. Maltoni, "Minutia cylinder-code: A new representation and matching technique for fingerprint recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2128–2141, 2010.

[33] K. Cao, E. Liu, and A. K. Jain, "Segmentation and enhancement of latent fingerprints: A coarse to fine ridgestructure dictionary," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1847–1859, 2014.

[34] Y. Tang, F. Gao, J. Feng, and Y. Liu, "Fingernet: An unified deep network for fingerprint minutiae extraction," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 108–116.

[35] J. Li, J. Feng, and C.-C. J. Kuo, "Deep convolutional neural network for latent fingerprint enhancement," *Signal Processing: Image Communication*, vol. 60, pp. 52–63, 2018.

[36] K. Cao and A. K. Jain, "Automated latent fingerprint recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 788–800, 2019.

[37] C. Lin and A. Kumar, "Contactless and partial 3D fingerprint recognition using multi-view deep representation," *Pattern Recognition*, vol. 83, pp. 314–327, 2018.

[38] V. Anand and V. Kanhangad, "Porenet: Cnn-based pore descriptor for high-resolution fingerprint recognition," *IEEE Sensors Journal*, vol. 20, no. 16, pp. 9305–9313, 2020.

[39] F. Liu, Y. Zhao, G. Liu, and L. Shen, "Fingerprint pore matching using deep features," *Pattern Recognition*, vol. 102, p. 107208, 2020.

[40] H.-U. Jang, H.-Y. Choi, D. Kim, J. Son, and H.-K. Lee, "Fingerprint spoof detection using contrast enhancement and convolutional neural networks," in *Information Science and Applications 2017*. Springer Singapore, 2017, pp. 331–338.

[41] D. M. Uliyan, S. Sadeghi, and H. A. Jalab, "Anti-spoofing method for fingerprint recognition using patch based deep learning machine," *Engineering Science and Technology, an International Journal*, vol. 23, no. 2, pp. 264–273, 2020.

[42] A. Khairwa, K. Abhishek, S. Prakash, and T. Pratap, "A comprehensive study of various biometric identification techniques," in *2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12)*, 2012, pp. 1–6.

[43] R. Jafri and H. R. Arabnia, "A survey of face recognition techniques," *Journal of Information Processing Systems*, vol. 5, no. 2, pp. 41–68, 2009.

[44] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, 1991, pp. 586–591.

[45] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 711–720, 1997.

[46] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 12, pp. 2037–2041, 2006.

[47] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2018, pp. 471–478.

[48] A. Dragoni, G. Vallesi, and P. Baldassarri, "A continuos learning for a face recognition system," in *ICAART 2011 - Proceedings of the 3rd International Conference on Agents and Artificial Intelligence*, vol. 1, 2011, pp. 541–544.

[49] P. Sernani, A. Claudi, G. Dolcini, L. Palazzo, G. Biancucci, and A. F. Dragoni, "Subject-dependent degrees of reliability to solve a face recognition problem using multiple neural networks," in *Proceedings ELMAR-2013*, 2013, pp. 11–14.

[50] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[51] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, *Labeled Faces in the Wild: A Survey.* Cham: Springer International Publishing, 2016, pp. 189–248.

[52] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[53] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 67–74.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[55] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Computer Vision and Image Understanding*, vol. 189, p. 102805, 2019.

[56] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao, "Violent video detection based on mosift feature and sparse coding," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3538–3542.

[57] M. Y. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," Carnegie Mellon University, Tech. Rep. CMU-CS-09-161, 2009. [Online]. Available: http://ra.adm.cs.cmu.edu/anon/usr/anon/home/ftp/2009/CMU-CS-09-161.pdf

[58] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

[59] O. Deniz, I. Serrano, G. Bueno, and T. Kim, "Fast violence detection in video," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2, 2014, pp. 478–485.

[60] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 1–6.

[61] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," *Image and Vision Computing*, vol. 48-49, pp. 37–41, 2016.

[62] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence detection in video by using 3d convolutional neural networks," in *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, R. McMahan,

J. Jerald, H. Zhang, S. M. Drucker, C. Kambhamettu, M. El Choubassi, Z. Deng, and M. Carlson, Eds. Springer International Publishing, 2014, pp. 551–558.

[63] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Computer Analysis of Images and Patterns*, P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano, and W. Kropatsch, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 332–339.

[64] J. Li, X. Jiang, T. Sun, and K. Xu, "Efficient violence detection using 3d convolutional neural networks," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2019, pp. 1–8.

[65] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence detection using spatiotemporal features with 3D convolutional neural network," *Sensors*, vol. 19, no. 11, p. 2472, 2019.

[66] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *CoRR*, vol. abs/1506.04214, 2015. [Online]. Available: http://arxiv.org/abs/1506.04214

[67] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2017, pp. 1–6.

[68] M. Bianculli, N. Falcionelli, P. Sernani, S. Tomassini, P. Contardo, M. Lombardi, and A. F. Dragoni, "A dataset for automatic violence detection in videos," *Data in Brief*, vol. 33, p. 106587, 2020.

[69] M. Cheng, K. Cai, and M. Li, "RWF-2000: an open large scale video database for violence detection," *CoRR*, vol. abs/1911.05913, 2019. [Online]. Available: http://arxiv.org/abs/1911.05913

[70] E. Sacchetto, "Face to face: il complesso rapporto tra automated facial recognition technology e processo penale," *La legislazione penale*, pp. 1–14, 2020. [Online]. Available: https://iris.unito.it/retrieve/handle/2318/1758754/668686/Sacchetto-finale.pdf

[71] A. F. Dragoni, P. Sernani, and D. Calvaresi, "When rationality entered time and became real agent in a cyber-society," in *Proceedings of the 3rd International Conference on Recent Trends and Applications in*

Computer Science and Information Technology, ser. CEUR Workshop Proceedings, vol. 2280, 2018, pp. 167–171. [Online]. Available: http://ceur-ws.org/Vol-2280/paper-24.pdf

[72] D. Doran, S. Schulz, and T. R. Besold, "What does explainable AI really mean? a new conceptualization of perspectives," in *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017*, ser. CEUR Workshop Proceedings, vol. 2071, 2017, pp. 15–22. [Online]. Available: http://ceur-ws.org/Vol-2071/CExAIIA_2017_paper_2.pdf

[73] D. Prota, *Sistemi esperti nel diritto: riflessioni su Compas e sul caso Loomis*. Università degli Studi di Padova, 2021/2022.

[74] L. A. Zadeh, "Fuzzy logic," *Computer*, vol. 21, no. 4, pp. 83–93, 1988.

[75] G. Taddei Elmi and A. Contaldo, *Intelligenza artificiale. Algoritmi giuridici Ius condendum o "fantadiritto"?* Pacini Giuridica, 2020.

[76] P. L. M. Lucatuorto, "Artificial intelligence and law: Judicial applications of expert systems (intelligenza artificiale e diritto: Le applicazioni giuridiche dei sistemi esperti)," *Cyberspace and Law (Ciberspazio e Diritto)*, vol. 7, no. 2, pp. 219–242, 2006.

[77] G. Sartor, *Le applicazioni giuridiche dell'intelligenza artificiale*. Dott. A. Giuffrè Editore, 1990.

[78] C. Asaro, E. Nissan, and A. Martino, "Daedalus, a procedural-support tool for the italian examining magistrate and prosecutor," *Nissan (2012), Sec*, vol. 4, no. 3, 2012.

[79] M. Iaselli, le nuove prospettive nel campo dell'informatica giudiziaria, in Diritto e Diritti, maggio 2001.

[80] "Governo Italiano, Programma Strategico Intelligenza Artificiale 2022-2024, p.15, Roma," 24.11.2021. [Online]. Available: https://assets.innovazione.gov.it/1637777289-programma-strategico-iaweb.pdf

[81] "C. Morabito, Polizia Moderna, La chiave del crimine," luglio 2015. [Online]. Available: https://poliziamoderna.poliziadistato.it/articolo/3535eaff4b96eead312456971

[82] "Commissione Europea, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions A Digital Agenda for Europe, Documento 52010DC0245," 19.05.2010.

[Online]. Available: https://eur-lex.europa.eu/legal-content/IT/ALL/?uri=celex%3A52010DC0245

[83] "Commissione Europea, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, L'intelligenza artificiale per l'Europa, Documento 52018DC0237," 25.04.2018. [Online]. Available: https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX%3A52018DC0237

[84] High-Level Expert Group on Artificial Intelligence, ETHICS GUIDELINES FOR TRUSTWORTHY AI, European Commission,B-1049 Brussels, 8 April 2019. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

[85] "Commissione Europea, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Creare fiducia nell'intelligenza artificiale antropocentrica, Documento 52019DC0168," 08.04.2019. [Online]. Available: https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:52019DC0168&from=IT

[86] "Commissione Europea, Proposta di regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'Intelligenza Artificiale (legge sull'intelligenza artificiale) e modifica alcuni atti legislativi dell'unione," 21.04.2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=celex%3A52021PC0206

[87] Commissione Europea, Direzione generale della Comunicazione, Leyen, U., Un'Unione più ambiziosa : il mio programma per l'Europa : orientamenti politici per la prossima Commissione europea 2019-2024, Ufficio delle pubblicazioni, 2019. [Online]. Available: https://data.europa.eu/doi/10.2775/6010

[88] "Commissione Europea, LIBRO BIANCO sull'intelligenza artificiale - Un approccio europeo all'eccellenza e alla fiducia, Documento 52020DC0065," 19.02.2020. [Online]. Available: https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX%3A52020DC0065

[89] "Regolamento (ue) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/ce, document 32016r0679," 27.04.2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=celex%3A32016R0679

[90] D. Maltoni, D. Maio, A. K. Jain, S. Prabhakar *et al.*, *Handbook of fingerprint recognition*. Springer, 2009, vol. 2.

[91] C. Ryu, S. G. Kong, and H. Kim, "Enhancement of feature extraction for low-quality fingerprint images using stochastic resonance," *Pattern Recognition Letters*, vol. 32, no. 2, pp. 107–113, 2011.

[92] M. Gao, Y. Tang, H. Liu, and R. Ma, "Statistics of fingerprint minutiae frequency and distribution based on automatic minutiae detection method," *Forensic Science International*, vol. 344, p. 111572, 2023.

[93] ORGANISMO DI CERTIFICAZIONE DELLA POLIZIA DI STATO, 07/10/2022, modificato il 18/01/2024. [Online]. Available: https://www.poliziadistato.it/articolo/407633fdfb91186c074874089

[94] Ministero dell'Interno, Dipartimento Pubblica Sicurezza, Assistenza tecnico-operativa e consulenziale sul sistema APFIS per le esigenze del Servizio di Polizia Scientifica. [Online]. Available: https://www.poliziadistato.it/statics/38/capitolato-assistenza-applicativa-apfis.-b.pdf

[95] Polizia Scientifica, Sistemi di indagine; 10/05/2013. [Online]. Available: https://www.poliziadistato.it/articolo/sistemi-di-indagine-1

[96] 2° Divisione, art. 107 c. 2 lett. b, Direzione Centrale Anticrimine della Polizia Di Stato, Servizio Polizia Scientifica. [Online]. Available: https://www.poliziadistato.it/articolo/27761430713671b3286193180

[97] "Presidenza del Consiglio dei Ministri, Dipartimento per gli Affari Europei; EURODAC," 11 dicembre 2000, regolamento (CE) n. 2725/2000. [Online]. Available: https://www.affarieuropei.gov.it/it/comunicazione/euroacronimi/eurodac/

[98] "Regolamento (UE) n. 603/2013 del Parlamento Europeo e del Consiglio; Gazzetta Ufficiale dell'Unione europea," 26 giugno 2013. [Online]. Available: https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32013R0603&from=DE

[99] "REGOLAMENTO (UE) N. 604/2013 DEL PARLAMENTO EUROPEO E DEL CONSIGLIO; Gazzetta ufficiale dell'Unione europea," 26 giugno 2013. [Online]. Available: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:180:0031:0059:IT:PDF

[100] "Best Practice Manual for Fingerprint Examination, ENFSI-BPM-FIN-01, version 01," november 2015. [Online]. Available: https://enfsi.eu/about-enfsi/structure/working-groups/documents-page/documents/best-practice-manuals/

[101] "Commissione Europea, Programma di Prevenzione e Lotta contro il Crimine– Direzione Generale Affari Interni, codice: PROJECT HOME/2012/ISEC/MO/4000004278," 2012. [Online]. Available: https://enfsi.eu/wp-content/uploads/2016/09/mp2012_-_project_summary_0.pdf

[102] S. Orandi, C. Watson, J. M. Libert, G. P. Fiumara, and J. D. Grantham, "Contactless fingerprint capture and data interchange best practice recommendation," *NIST Special Publication*, vol. 500, p. 334, 2021.

[103] S. Orandi, S. Orandi, J. Libert, B. Bandini, K. Ko, J. Grantham, and C. Watson, *Evaluating the operational impact of contactless fingerprint imagery on matcher performance*. US Department of Commerce, National Institute of Standards and Technology, 2020.

[104] L. Ruzicka, D. Söllinger, B. Kohn, C. Heitzinger, A. Uhl, B. Strobl *et al.*, "Improving sensor interoperability between contactless and contact-based fingerprints using pose correction and unwarping," *IET Biometrics*, vol. 2023, 2023.

[105] K. Rajaram, B. A. NG, and A. S. Guptha, "Clnet: a contactless fingerprint spoof detection using deep neural networks with a transfer learning approach," *Multimedia Tools and Applications*, pp. 1–20, 2023.

[106] Orandi, Shahram and Libert, John M and Grantham, John and Ko, Kenneth and Bandini, Bruce and Watson, Craig I: Specification for Certification Testing of Contactless Fingerprint Acquisition Devices, v1. 0; Special Publication (NIST SP), National Institute of Standards and Technology, 2023.

[107] C. Weyermann, S. Willis, P. Margot, and C. Roux, "Towards more relevance in forensic science research and development," *Forensic Science International*, p. 111592, 2023.

[108] ENFSI, Vision of the European Forensic Science Area 2030, "Improving the Reliability and Validity of Forensic Science and Fostering the Implementation of Emerging Technologies".

[109] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: A survey," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8647–8695, 2023.

[110] Guideline for Facial Recognition System End Users; ENFSI-DI-GDL-001; Version 001 – November 2022. [Online]. Available: https://enfsi.eu/wp-content/uploads/2023/02/DI-GDL-001_GDL-for-Facial-Recognition-System-End-Users_20221111.pdf

[111] "Testo unico delle leggi di pubblica sicurezza → Titolo I - Dei provvedimenti di polizia e della loro esecuzione → Capo I - Delle attribuzioni dell'autorità di pubblica sicurezza e dei provvedimenti d'urgenza o per grave necessità pubblica," art.4.

[112] "Identificazione della persona nei cui confronti vengono svolte le indagini e di altre persone: Codice di procedura penale → LIBRO QUINTO - Indagini preliminari e udienza preliminare → Titolo IV - Attività a iniziativa della polizia giudiziaria," art.349.

[113] "Testo unico sull'immigrazione → Titolo II - Disposizioni sull'ingresso, il soggiorno e l'allontanamento dal territorio dello stato → Capo I - Disposizioni sull'ingresso e il soggiorno; D.lgs. 25 luglio 1998, n. 286," 10-9-2002, art.5.

[114] "REGOLAMENTO (UE) N. 603/2013 DEL PARLAMENTO EUROPEO E DEL CONSIGLIO; Gazzetta ufficiale dell'Unione europea," 26 giugno 2013, art.9. [Online]. Available: https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32013R0603&from=DE

[115] "REGOLAMENTO (UE) N. 603/2013 DEL PARLAMENTO EUROPEO E DEL CONSIGLIO; Gazzetta ufficiale dell'Unione europea," 26 giugno 2013, art.14. [Online]. Available: https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32013R0603&from=DE

[116] "REGOLAMENTO (UE) N. 603/2013 DEL PARLAMENTO EUROPEO E DEL CONSIGLIO; Gazzetta ufficiale dell'Unione europea," 26 giugno 2013, art.17. [Online]. Available: https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32013R0603&from=DE

[117] Appendice documentaria, Troiani Rocco Luigi, vers. 1992, b. 88 fasc. 19, Ministero dell'Interno, Direzione generale di .S., Scuola Superiore di Polizia, Servizio Centrale di segnalamento e identificazione, scheda segnaletica anno 1937.

[118] Cipollini Francesco, vers. 1992, b. 23 fasc. 29, Ministero dell'Interno, Direzione generale di P.S., Scuola Superiore di Polizia, Servizio Centrale di segnalamento e identificazione, scheda segnaletica anno 1942.

[119] M. R. Fiori, Archivio storico della Questura di Ascoli Piceno, Individui sottoposti a misure di vigilanza politica divisione 1° - Gabinetto categoria A8, vigilati politici inventario, Introduzione e Strumenti di Consultazione, FAS Editore, 2018-2021.

[120] "Rifiuto d'indicazioni sulla propria identità personale: Codice Penale → LIBRO TERZO - Delle contravvenzioni in particolare → Titolo I - Delle contravvenzioni di polizia → Capo I - Delle contravvenzioni concernenti la polizia di sicurezza → Sezione I - Delle contravvenzioni concernenti l'ordine pubblico e la tranquillità pubblica," art.351.

[121] "Regio Decreto 6 maggio 1940, n. 635: Regolamento per l'esecuzione del Testo Unico 18 giugno 1931, n. 773 delle Leggi di Pubblica Sicurezza: Gazz. Uff. del 26 giugno 1940, n. 149, Suppl. Ord." art.294.

[122] "DECRETO DEL PRESIDENTE DELLA REPUBBLICA 28 dicembre 2000, n. 445: Testo unico delle disposizioni legislative e regolamentari in materia di documentazione amministrativa. (Testo A)." 7-3-2001, art.35.

[123] International Civil Aviation Organization ICAO, Security and Facilitation, Technical reports, Annex A Photograph Guidelines: 2008, modified 2011.

[124] M. Ferrara, A. Franco, D. Maio, and D. Maltoni, "Face image conformance to iso/icao standards in machine readable travel documents," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 4, pp. 1204–1213, 2012.

[125] SPIS/IDENTISYSTEM SISTEMA DI INDAGINE E FOTOSEGNALAMENTO : sistema per generare velocemente un file completo di dati di identificazione, caratteristiche fisiche, foto, voce e impronte digitali: acquisizione della fotosegnaletica. [Online]. Available: https://www.secomitalia.com/spis2

[126] Sicurezza Pubblica per Ministero dell'interno e Polizia di Stato: Reco3.26, sistema SARI, 2018. [Online]. Available: https://www.reco326.com/it/storie-di-successo/29-sicurezza-pubblica-per-ministero-dell-interno-e-polizia-di-stato

[127] 4° Divisione, art. 107 c. 2 lett. b, Direzione Centrale Anticrimine della Polizia Di Stato, Servizio Polizia Scientifica. [Online]. Available: https://www.poliziadistato.it/articolo/277614308f7bc511420114914

[128] Direzione Centrale Anticrimine della Polizia Di Stato, Servizio Polizia Scientifica, Organizzazione. [Online]. Available: https://www.poliziadistato.it/articolo/organizzazione

[129] Best Practice Manual for Facial Image Comparison; ENFSI-BPM-DI-01; Version 01 - January 2018. [Online]. Available: https://enfsi.eu/wp-content/uploads/2017/06/ENFSI-BPM-DI-01.pdf

[130] Best Practice Manual for Forensic Image and Video Enhancement; ENFSI-BPM-DI-02; Version 01 – June 2018. [Online]. Available: https://enfsi.eu/wp-content/uploads/2017/06/Best-Practice-Manual-for-Forensic-Image-and-Video-Enhancement.pdf

[131] Best Practice Manual for Digital Image Authentication; ENFSI-BPM-DI-03; Issue 01 – October 2021. [Online]. Available: https://enfsi.eu/wp-content/uploads/2022/12/1.-BPM_Image-Authentication_ENFSI-BPM-DI-03-1.pdf

[132] FACIAL IDENTIFICATION SCIENTIFIC WORKING GROUP; Standard Guide for Capturing Facial Images for Use with Facial Recognition Systems; Version 2.0 ; 2019.05.10. [Online]. Available: https://fiswg.org/FISWG_Guide_for_Capturing_Facial_Images_for_FR_Use_v2.0_20190510.pdf

[133] 2011, national Institute of Standards and Technology: Data Format for the Interchange of Biometric and Forensic Information. [Online]. Available: https://www.nist.gov/system/files/documents/2021/03/18/ansi-nist_archived_2010_draft_2_an-2011_v1.pdf

[134] C. Watson and P. Flanagan, "Nist special database 18. nist mugshot identification database (mid)," National Institute of Standards and Technology, Tech. Rep., 2016.

[135] P. Grother and M. Ngan, "Face recognition vendor test (frvt): Performance of face identification algorithms," *NIST Interagency report*, vol. 8009, no. 5, p. 14, 2014.

[136] Federal Bureau of Investigations (FBI); FACIAL IDENTIFICATION SCIENTIFIC WORKING GROUP; 2009. [Online]. Available: https://www.fiswg.org/index.html

[137] T. Rademacher, *Artificial Intelligence and Law Enforcement*. Springer International Publishing, 2020, pp. 225–254.

[138] M. T. Bhatti, M. G. Khan, M. Aslam, and M. J. Fiaz, "Weapon detection in real-time cctv videos using deep learning," *IEEE Access*, vol. 9, pp. 34 366–34 382, 2021.

[139] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, and A. F. Dragoni, "Deep learning for automatic violence detection: Tests on the airtlab dataset," *IEEE Access*, vol. 9, pp. 160 580–160 595, 2021.

[140] Z. Yuan, X. Zhou, and T. Yang, "Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal

data," in *24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '18. Association for Computing Machinery, 2018, p. 984–992.

[141] Z. Xu, C. Hu, and L. Mei, "Video structured description technology based intelligence analysis of surveillance videos for public security applications," *Multimedia Tools and Applications*, vol. 75, no. 19, pp. 12 155–12 172, 2016.

[142] M. Gomez-Barrero, P. Drozdowski, C. Rathgeb, J. Patino, M. Todisco, A. Nautsch, N. Damer, J. Priesnitz, N. Evans, and C. Busch, "Biometrics in the era of covid-19: Challenges and opportunities," *IEEE Transactions on Technology and Society*, vol. 3, no. 4, pp. 307–322, 2022.

[143] D. Crouse, H. Han, D. Chandra, B. Barbello, and A. K. Jain, "Continuous authentication of mobile user: Fusion of face image and inertial measurement unit data," in *2015 International Conference on Biometrics (ICB)*, 2015, pp. 135–142.

[144] A. Opitz and A. Kriechbaum-Zabini, "Evaluation of face recognition technologies for identity verification in an egate based on operational data of an airport," in *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2015, pp. 1–5.

[145] B. Ammour, L. Boubchir, T. Bouden, and M. Ramdani, "Face–iris multimodal biometric identification system," *Electronics*, vol. 9, no. 1, 2020.

[146] M. Forti, "Ai-driven migration management procedures: fundamental rights issues and regulatory answers," *BioLaw Journal*, vol. 2021, no. 2, pp. 433–451, 2021.

[147] M. Hassaballah and S. Aly, "Face recognition: challenges, achievements and future directions," *IET Computer Vision*, vol. 9, no. 4, pp. 614–626, 2015.

[148] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 3, pp. 1–42, 2016.

[149] S. Ahmed, S. Ali, J. Ahmad, M. Adnan, and M. Fraz, "On the frontiers of pose invariant face recognition: a review," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 2571–2634, 2020.

[150] P. Contardo, E. Di Lorenzo, N. Falcionelli, A. F. Dragoni, and P. Sernani, "Analyzing the impact of police mugshots in face verification for

crime investigations," in *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*. IEEE, 2022, pp. 236–241.

[151] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, present, and future of face recognition: A review," *Electronics*, vol. 9, no. 8, 2020.

[152] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[153] M. Grgic, K. Delac, and S. Grgic, "SCface — surveillance cameras face database," *Multimedia Tools and Applications*, vol. 51, no. 3, p. 863–879, 2011.

[154] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *British Machine Vision Association*, 2015. [Online]. Available: https://www.robots.ox.ac.uk/~vedaldi/assets/pubs/parkhi15deep.pdf

[155] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *1994 IEEE Workshop on Applications of Computer Vision*, 1994, pp. 138–142.

[156] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain, "Unconstrained face recognition: Identifying a person of interest from a media collection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2144–2157, 2014.

[157] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, 2007.

[158] G. B. Huang and E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2014-003, 2014.

[159] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *2011 IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 529–534.

[160] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[161] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *CoRR*, vol. abs/1411.7923, 2014. [Online]. Available: http://arxiv.org/abs/1411.7923

[162] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4873–4882.

[163] A. Nech and I. Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3406–3415.

[164] P. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[165] P. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[166] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '99. ACM Press/Addison-Wesley Publishing Co., 1999, p. 187–194.

[167] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[168] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[169] H. Bon-Woo, H. Byun, R. Myoung-Cheol, and L. Seong-Whan, "Performance evaluation of face recognition algorithms on the asian face database, kfdb," in *Audio- and Video-Based Biometric Person Authentication*, J. Kittler and M. S. Nixon, Eds. Springer Berlin Heidelberg, 2003, pp. 557–565.

[170] W. Gao, B. Cao, S. Shan, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL large scale chinese face database and evaluation protocols," ICT-ISVISION Joint Research & Development Laboratory for Face Recognition, Chinese Academy of Sciences, Tech. Rep. JDL-TR_04_FR_001, 2004.

[171] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[172] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops*, 2011, pp. 81–88.

[173] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*. Curran Associates, Inc., 2012, vol. 2, pp. 1097–1105. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[174] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[175] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[176] M. You, X. Han, Y. Xu, and L. Li, "Systematic evaluation of deep face recognition methods," *Neurocomputing*, vol. 388, pp. 144–156, 2020.

[177] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4295–4304.

[178] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1283–1292.

[179] ——, "Representation learning by rotating your faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 3007–3021, 2019.

[180] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, S. Yan, and J. Feng, "Towards pose invariant face recognition in the wild," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2207–2216.

[181] J. Xiang and G. Zhu, "Joint face detection and facial expression recognition with mtcnn," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, 2017, pp. 424–427.

[182] D. Hazra and Y.-C. Byun, "Upsampling real-time, low-resolution cctv videos using generative adversarial networks," *Electronics*, vol. 9, no. 8, 2020.

[183] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, and A. F. Dragoni, "Deep learning for automatic violence detection: Tests on the AIRTLab dataset," *IEEE Access*, vol. 9, pp. 160 580–160 595, 2021.

[184] D. Bright, R. Brewer, and C. Morselli, "Using social network analysis to study crime: Navigating the challenges of criminal justice records," *Social Networks*, vol. 66, pp. 50–64, 2021.

[185] A. Castillo, S. Tabik, F. Pérez, R. Olmos, and F. Herrera, "Brightness guided preprocessing for automatic cold steel weapon detection in surveillance videos with deep learning," *Neurocomputing*, vol. 330, pp. 151–161, 2019.

[186] A. Franco, D. Maio, and D. Maltoni, "2d face recognition based on supervised subspace learning from 3d models," *Pattern Recognition*, vol. 41, no. 12, pp. 3822–3833, 2008.

[187] P. Contardo, P. Sernani, N. Falcionelli, and A. F. Dragoni, "Deep learning for law enforcement: A survey about three application domains," in *4th International Conference on Recent Trends and Applications in Computer Science and Information Technology*, ser. CEUR Workshop Proceedings, vol. 2872, 2021, pp. 36–45. [Online]. Available: http://ceur-ws.org/Vol-2872/paper06.pdf

[188] L. Spreeuwers, A. Hendrikse, and K. Gerritsen, "Evaluation of automatic face recognition for automatic border control on actual data recorded of travellers at schiphol airport," in *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, 2012, pp. 1–6.

[189] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2268–2283, 2016.

[190] E. Vazquez-Fernandez and D. Gonzalez-Jimenez, "Face recognition for authentication on mobile devices," *Image and Vision Computing*, vol. 55, pp. 31–33, 2016.

[191] M. Zulfiqar, F. Syed, M. J. Khan, and K. Khurshid, "Deep face recognition for biometric authentication," in *2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 2019, pp. 1–6.

[192] M. Ferrara, A. Franco, D. Maltoni, and Y. Sun, "On the impact of alterations on face photo recognition accuracy," in *Image Analysis and Processing–ICIAP 2013: 17th International Conference, Naples, Italy, September 9-13, 2013. Proceedings, Part I 17*. Springer, 2013, pp. 743–751.

[193] P. Contardo, P. Sernani, S. Tomassini, N. Falcionelli, M. Martarelli, P. Castellini, and A. F. Dragoni, "FRMDB: Face recognition using multiple points of view," *Sensors*, vol. 23, no. 4, 2023.

[194] P. Contardo, E. D. Lorenzo, N. Falcionelli, A. F. Dragoni, and P. Sernani, "Analyzing the impact of police mugshots in face verification for crime investigations," in *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, 2022, pp. 236–241.

[195] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," *CoRR*, vol. abs/1502.00873, 2015. [Online]. Available: http://arxiv.org/abs/1502.00873

[196] A. B. Mabrouk and E. Zagrouba, "Spatio-temporal feature using optical flow based distribution for violence detection," *Pattern Recognition Letters*, vol. 92, pp. 62–67, 2017.

[197] E. Fenil, G. Manogaran, G. Vivekananda, T. Thanjaivadivel, S. Jeeva, A. Ahilan *et al.*, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional lstm," *Computer Networks*, vol. 151, pp. 191–200, 2019.

[198] Z. Meng, J. Yuan, and Z. Li, "Trajectory-pooled deep convolutional networks for violence detection in videos," in *Computer Vision Systems*, M. Liu, H. Chen, and M. Vincze, Eds. Springer International Publishing, 2017, pp. 437–447.

[199] S. Dinesh Jackson, E. Fenil, M. Gunasekaran, G. Vivekananda, T. Thanjaivadivel, S. Jeeva, and A. Ahilan, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM," *Computer Networks*, vol. 151, pp. 191–200, 2019.

[200] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2017, pp. 1800–1807.

[201] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2018, pp. 8697–8710.

[202] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail, M. Ahmed, M. Ilyas, and A. Mahmood, "A review on state-of-the-art violence detection techniques," *IEEE Access*, vol. 7, pp. 107 560–107 575, 2019.

[203] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A novel violent video detection scheme based on modified 3D convolutional neural networks," *IEEE Access*, vol. 7, pp. 39 172–39 179, 2019.

[204] I. Serrano Gracia, O. Deniz Suarez, G. Bueno Garcia, and T.-K. Kim, "Fast fight detection," *PloS one*, vol. 10, no. 4, p. e0120448, 2015.

[205] A. Hanson, K. PNVR, S. Krishnagopal, and L. Davis, "Bidirectional convolutional LSTM for the detection of violence in videos," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 280–295.

[206] Y. LeCun and Y. Bengio, *Convolutional Networks for Images, Speech, and Time Series*. Cambridge, MA, USA: MIT Press, 1998, p. 255–258.

[207] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[208] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[209] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[210] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 273–278.

[211] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'14, 2014, p. 568–576.

[212] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: unsupervised learning using temporal order verification," in *Computer Vision – ECCV 2016*. Springer, 2016, pp. 527–544.

[213] S. Wager, S. Wang, and P. Liang, "Dropout training as adaptive regularization," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'13, 2013, p. 351–359.

[214] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[215] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, 2017.

[216] Polizia di Stato: Il Questore, 25.08.2011 mod. 28.01.2013. [Online]. Available: https://www.poliziadistato.it/articolo/il-questore#:~:text=In%20base%20all'art.,pubblica%20nel%20capoluogo%20di%20pertinenza.

[217] James Garrity, "Struttura e funzione degli occhi", MD, Mayo Clinic College of Medicine and ScienceManuale MSD, marzo 2022. [Online]. Available: https://www.msdmanuals.com/it-it/casa/disturbi-oculari/biologia-degli-occhi/muscoli-nervi-e-vasi-sanguigni-degli-occhi

[218] E. Viceconte, "Il processo decisionale e la razionalita'limitata," *core.ac.uk*, 2004.

[219] F. U. M. Ullah, M. S. Obaidat, A. Ullah, K. Muhammad, M. Hijji, and S. W. Baik, "A comprehensive review on vision-based violence detection in surveillance videos," *ACM Comput. Surv.*, vol. 55, no. 10, 2023.

[220] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, and A. F. Dragoni, "Deep learning for automatic violence detection: Tests on the airtlab dataset," *IEEE Access*, vol. 9, pp. 160 580–160 595, 2021.

[221] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[222] W. Niu, M. Sun, Z. Li, J.-A. Chen, J. Guan, X. Shen, Y. Wang, S. Liu, X. Lin, and B. Ren, "RT3D: Achieving real-time execution of 3D convolutional neural networks on mobile devices," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, pp. 9179–9187, 2021.

[223] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[224] D. Freire-Obregón, P. Barra, M. Castrillón-Santana, and M. D. Marsico, "Inflated 3d convnet context analysis for violence detection," *Machine Vision and Applications*, vol. 33, pp. 1–13, 2022.

[225] L. Ciampi, P. Foszner, N. Messina, M. Staniszewski, C. Gennaro, F. Falchi, G. Serao, M. Cogiel, D. Golba, A. Szczęsna, and G. Amato, "Bus violence: An open benchmark for video violence detection on public transport," *Sensors*, vol. 22, no. 21, 2022.

[226] V. E. D. S. Silva, T. B. Lacerda, P. B. Miranda, A. C. Nascimento, and A. P. C. Furtado, "Federated learning for physical violence detection in videos," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8.

[227] L. Yang, Z. Wu, J. Hong, and J. Long, "MCL: A contrastive learning method for multimodal data fusion in violence detection," *IEEE Signal Processing Letters*, pp. 1–5, 2022.

[228] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Computer Vision – ECCV 2020*. Cham: Springer International Publishing, 2020, pp. 322–339.

[229] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[230] December 2022, best Practice Manual for the Methodology of Forensic Speaker Comparison; ENFSI-FSA-BPM-003 Version 01. [Online]. Available: https://enfsi.eu/wp-content/uploads/2022/12/5.-FSA-BPM-003_BPM-for-the-Methodology-1.pdf

[231] December 2022, best Practice Manual for Digital Audio Authenticity Analysis; ENFSI-FSA-BPM-002 Version 01. [Online]. Available: https://enfsi.eu/wp-content/uploads/2022/12/FSA-BPM-002_BPM-for-Digital-Audio-Authenticity-Analysis.pdf

[232] 2015, andrzej Drygajlo, Michael Jessen, Stefan Gfroerer, Isolde Wagner, Jos Vermeulen and Tuija Niemi, Methodological Guidelines for Best

Practice in Forensic Semiautomatic and Automatic Speaker Recognition, including Guidance on the Conduct of Proficiency Testing and Collaborative Exercises; A project funded by the EU ISEC 2011 Agreement Number: HOME/2011/ISEC/MO/4000002384. [Online]. Available: https://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf

[233] 2 giugno 2009, forensic Speech And Audio Analysis Working Group Best Practice Guidelines For Enf Analysis In Forensic Authentication Of Digital Evidence; Ref. code: FSAAWG-TOR-FSA-001, Issue No. 001 – Version approved by FSAAWG Steering Committee. [Online]. Available: https://enfsi.eu/wp-content/uploads/2016/09/forensic_speech_and_audio_analysis_wg_-_best_practice_guidelines_for_enf_analysis_in_forensic_authentication_of_digital_evidence_0.pdf

[234] CADMI - Casa di Accoglienza delle Donne Maltrattate; Milano. [Online]. Available: https://cadmi.org/chi-siamo/

[235] Polizia di Stato, YouPol: l'app per il contrasto alla violenza di genere. [Online]. Available: https://www.poliziadistato.it/articolo/youpol--l-app-per-il-contrasto-alla-violenza-di-genere

[236] Save the woman: NONPOSSOPARLARE, un chatbot per tutte le donne in difficoltà. [Online]. Available: https://www.savethewoman.org/il-chatbot/

[237] Declaration on the Elimination of Violence against Women, General Assembly resolution 48/104, 20 December 1993. [Online]. Available: https://www.ohchr.org/en/instruments-mechanisms/instruments/declaration-elimination-violence-against-women

[238] Ministero dell'Interno, Polizia di Stato, Questo non è amore 2023: uniti contro la violenza di genere; 21/11/2023, modificato il 25/11/2023. [Online]. Available: https://www.poliziadistato.it/articolo/2475655cc52f32872891041042

[239] Ministero dell'Interno, Servizio Analisi Criminale, Il Punto - Il pregiudizio e la violenza contro le donne; Roma 11/12/2023,. [Online]. Available: https://www.interno.gov.it/it/notizie/punto-pregiudizio-e-violenza-contro-donne-presentato-roma-report-servizio-analisi-criminale

[240] Presidenza della Repubblica, Dichiarazione del Presidente Mattarella in occasione della Giornata Internazionale per l'eliminazione della Violenza contro le Donne: Roma, 25/11/2022. [Online]. Available: https://www.quirinale.it/elementi/74136

[241] GU n.275 del 24-11-2023, LEGGE n. 168: Disposizioni per il contrasto della violenza sulle donne e della violenza domestica. (23G00178). [Online]. Available: https://www.gazzettaufficiale.it/eli/gu/2023/11/24/275/sg/pdf

[242] Indagini hi-tech, Inserto di POLIZIAMODERNA - Gennaio 2024; Cristiano Morabito. [Online]. Available: https://poliziamoderna.poliziadistato.it/statics/40/indagini-elettroniche.pdf

[243] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, p. 101567, 2024.

[244] N. Chakravarty and M. Dua, "A lightweight feature extraction technique for deepfake audio detection," *Multimedia Tools and Applications*, pp. 1–25, 2024.

[245] V. F. Puglisi, O. Giudice, and S. Battiato, "Deep audio analyzer: a framework to industrialize the research on audio forensics," in *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*. IEEE, 2023, pp. 537–542.

[246] A. A. Fime, M. Ashikuzzaman, and A. Aziz, "Audio signal based danger detection using signal processing and deep learning," *Expert Systems with Applications*, vol. 237, p. 121646, 2024.

[247] Aziz, Abdul; Fime, Awal Ahmed; Ashikuzzaman, Md.; Hossain, Sk Imran (2023), "Audio Signal Dataset for Danger Detection of Women and Children", Mendeley Data, V1, doi: 10.17632/gfvsdtnf3v.1. [Online]. Available: https://data.mendeley.com/datasets/gfvsdtnf3v/1

[248] K. Choi, D. Joo, and J. Kim, "Kapre: On-gpu audio preprocessing layers for a quick implementation of deep neural network models with keras," in *Machine Learning for Music Discovery Workshop at 34th International Conference on Machine Learning*. ICML, 2017.

[249] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The journal of the acoustical society of america*, vol. 8, no. 3, pp. 185–190, 1937.

[250] A. Das, S. Guha, P. K. Singh, A. Ahmadian, N. Senu, and R. Sarkar, "A hybrid meta-heuristic feature selection method for identification of indian spoken languages from audio signals," *IEEE Access*, vol. 8, pp. 181 432–181 449, 2020.

[251] T. Sainburg, "timsainb/noisereduce: v1.0," jun 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3243139

[252] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, p. e1008228, 2020.

[253] R. G. Mello, L. F. Oliveira, and J. Nadal, "Digital butterworth filter for subtracting noise from low magnitude surface electromyogram," *Computer methods and programs in biomedicine*, vol. 87, no. 1, pp. 28–35, 2007.

[254] S. Battiato, filtraggio nel Dominio della Frequenza Parte II; Università degli Studi di Catania. [Online]. Available: https://www.dmi.unict.it/~battiato/mm1112/Parte%206_2%20-%20Filtraggio%20nel%20Dominio%20della%20Frequenza.pdf

[255] C. T. Leondes, *Multidimensional Systems: Signal Processing and Modeling Techniques: Advances in Theory and Applications*. Elsevier, 1995.

[256] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Szegedy_Going_Deeper_With_2015_CVPR_paper.pdf

[257] T. Lawrence and L. Zhang, "Iotnet: An efficient and accurate convolutional neural network for iot devices," *Sensors*, vol. 19, no. 24, p. 5541, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/24/5541

[258] S. Mukherjee, Aug 18, 2022, the Annotated ResNet-50. [Online]. Available: https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758

[259] A. Khan, M. A. Khan, M. Y. Javed, M. Alhaisoni, U. Tariq, S. Kadry, J.-I. Choi, and Y. Nam, "Human gait recognition using deep learning and improved ant colony optimization." *Computers, Materials & Continua*, vol. 70, no. 2, 2022.

[260] P. Sumari, W. M. A. W. Ahmad, F. Hadi, M. Mazlan, N. A. Liyana, R.-W. Bello, A. S. A. Mohamed, and A. Z. Talib, "A precision agricultural application: Manggis fruit classification using hybrid deep learning." *Rev. d'Intelligence Artif.*, vol. 35, no. 5, pp. 375–381, 2021.

[261] T. Madiega, "Artificial intelligence act," *European Parliament: European Parliamentary Research Service*, 2021. [Online]. Available: https://superintelligenz.eu/wp-content/uploads/2023/07/EPRS_BRI2021698792_EN.pdf

[262] European Parliament, News, Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI; Press Releases 09-12-2023. [Online]. Available: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai

[263] Parlamento Europeo, Normativa sull'IA: la prima regolamentazione sull'intelligenza artificiale, Data di pubblicazione: 13-06-2023 Ultimo aggiornamento: 10-01-2024. [Online]. Available: https://www.europarl.europa.eu/topics/it/article/20230601STO93804/normativa-sull-ia-la-prima-regolamentazione-sull-intelligenza-artificiale

[264] J. P. Near, D. Darais, N. Lefkovitz, G. Howarth *et al.*, "Guidelines for evaluating differential privacy guarantees," National Institute of Standards and Technology, Tech. Rep., 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-226.ipd.pdf

[265] Thales, Key biometric matching technology providers, results of sponsored tests at the Maryland Test Facility, 2018. [Online]. Available: https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/biometrics/facial-recognition

[266] NIST, IREX 10: Identification Trac, Last Updated: February 14, 2024. [Online]. Available: https://pages.nist.gov/IREX10/

[267] Innovatrics, Welcome to the World of Instant Trust. [Online]. Available: https://www.innovatrics.com/about-us/

[268] Secom BIOFAD, analisi visiva abbinata all'autenticazione elettronica dei documenti di identità (ePassport). [Online]. Available: https://www.secomitalia.com/biofad

[269] Secom: SPIS/IDENTISYSTEM,funzionalità. [Online]. Available: https://www.secomitalia.com/spis2

[270] Secom, Sistema mobile di controllo del territorio. [Online]. Available: https://www.secomitalia.com/sistemimobili

[271] M. Ravveduto, "La google generation criminale: i giovani della camorra su facebook," *Rivista di Studi e Ricerche sulla criminalità organizzata*, vol. 4, no. 4, pp. 57–78, 2018.

[272] Gratteri e Nicaso: le mafie e la nuova frontiera della criminalità digitale, Antimafia 2000, 29/08/2023. [Online]. Available: https://www.antimafiaduemila.com/home/mafie-news/309-topnews/97944-gratteri-e-nicaso-le-mafie-e-la-nuova-frontiera-della-criminalita-digitale.html