**KeySD - KeyStroke Dynamics**

Christian Miccolis matr. 683313

**Abstract:**

Con la rapida crescita delle telecamere di videosorveglianza per monitorare l'attività umana si richiede che tali sistemi siano in grado di riconoscere automaticamente la violenza e gli eventi sospetti. Il rilevamento di azioni anomale e violente è diventata un'area di ricerca attiva della computer vision e dell’elaborazione di immagini che attira sempre più nuovi ricercatori. La letteratura a tal riguardo presenta diverse tecniche per la rilevazione di tali attività rispetto al video proposto. VioSR, acronimo di Violent video Scene Recongition, è un progetto sviluppato per il riconoscimento e eventuale classificazione del tipo di scena violenta all’interno di contenuti video. Il sistema VioSR comprende molteplici sistemi differenti ed indipendenti in grado di far fronte al task di «violence detection», oltre ad una rete in grado di distinguere selettivamente la specifica classe di violenza in corso. VioSR è stato progettato sfruttando le potenzialità di più linguaggi di programmazione e le sue capacità sono state verificate su quattro dataset video e in varie tipologie di test utili per dimostrare come il sistema si comporterebbe in ambienti di uso comune del mondo reale.

**Sezione 1 – Introduzione**

Le tecniche di rilevazione della violenza che utilizzano la computer vision analizzano i video delle telecamere di sorveglianza. Negli ultimi anni, queste telecamere e altre apparecchiature di sorveglianza sono state installate in luoghi diversi per la sicurezza pubblica, ad es. Istituzioni educative, ospedali, banche, mercati, strade ecc. per monitorare le attività delle persone. Il monitoraggio include l'analisi dei comportamenti delle persone, indipendentemente dal fatto che le loro attività siano sospette o normali. A tale scopo, sono stati sviluppati diversi metodi per riconoscere le attività umane nella vita reale. Questi metodi aiutano a rilevare le attività sospette nei video di sorveglianza.

Tra gli eventi che possono verificarsi nelle zone affollate ci sono quelli di aggressione, combattimenti e altre forme di violenza. Se le folle non sono controllate e gestite questo potrebbe avere un effetto negativo sulla sicurezza delle singole persone, creando disagi e situazioni spiacevoli, potenzialmente provocando feriti e nei peggiori dei casi delle morti. I video giocano un ruolo fondamentale per la società di oggi, dando più sicurezza nelle grandi aree dove ci si aspetta che un gran numero di persone si riuniscano in zone limitate. Aumentare la sicurezza in queste aree ha reso l’uso di sistemi di videosorveglianza molto diffuso. Per esempio, ci sono camere installate nelle strade, nei centri commerciali, nei templi, negli stadi e in molti altri contesti. Delle volte, però, l’occhio umano non può predire dove la situazione stia andando fuori controllo: anche avvalendosi di un sistema di video sorveglianza con molteplici camere a circuito chiuso, l’occhio umano non potrebbe guardare simultaneamente tutti i monitor. È di vitale importanza prevenire situazioni potenzialmente dannose e intervenire il prima possibile. Un sistema che riconosca azioni violente all’interno di una folla, potrebbe essere di immenso aiuto per la vigilanza e per ristabilire l’ordine prima che la situazione degeneri. Inoltre, un sistema di questo tipo potrebbe essere applicato immediatamente nelle camere di videosorveglianza.

Anche il rilevamento della violenza dai video di sorveglianza riguarda il rilevamento dell'attività. Diverse tecniche e metodi sono stati sviluppati per rilevare eventi violenti e dannosi nei video [56], [57], [58], [59]. In questi metodi vengono proposti approcci diversi che funzionano con parametri di input diversi. I parametri sono sostanzialmente diversi attributi o caratteristiche del video come accelerazione, flusso, tempo, aspetto ecc. Nel processo di rilevamento dell'attività di violenza, il primo passo è quello di dividere un intero video in segmenti e frame. In secondo luogo, si estraggono le feature secondo il metodo applicato. Infine, rileva l'attività anomala dai frame o dall’intero video. Il passaggio varia in base al metodo applicato per il rilevamento. I passaggi di base delle tecniche di rilevazione della violenza sono mostrati nella Figura (1). Molti ricercatori hanno proposto diverse tecniche per aumentare l'efficienza, l'accuratezza e le prestazioni del processo di rilevazione.

In questo documento, vengono esplorati e discussi diversi metodi di rilevazione della violenza dai video di sorveglianza usando la computer vision. Inoltre il nostro contributo è stato quello di non fermarci alla sola fase di training e testing su singoli dataset adibiti propriamente al task di violence detection, ma di effettuare una sperimentazione cross-dataset su dataset più generali che presentano differenti scene e azioni per verificare che questi metodi siano capaci di generalizzare rispetto a molteplici variabili nei video (luminosità, occlusione, differenze di sfondo, numero di persone a schermo, differenti azioni violente ecc.). Come ultimo step attraverso la rete deep Slow Fast [96] abbiamo cercato di caratterizzare le scene violente (calci, pughi, esplosioni, spari…).

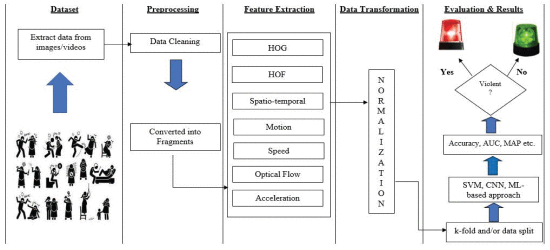


Figura 1. Passaggi base per una tipica tecnica di rilevazione della violenza [48]

**Sezione 2 – Lavori correlati [4]**

Attraverso una indagine approfondita sulla letteratura [22] della violence detection è emerso la possibilità di un duplice approccio al riconoscimento della violenza:

* Approccio tradizionale o shallow learning con la cosiddetta rappresentazione handcrafted, che si basa su rilevatori e descrittori di caratteristiche progettate da esperti come Scale Invariant Feature Transform (SIFT), Histogram of Oriented Gradient (HOG), Speeded Up Robust Feature (SURF) e Local Binary Pattern (LBP). Dopo questa fase un classificatore generico verrà addestrato per il riconoscimento della scena come mostrato in (2).

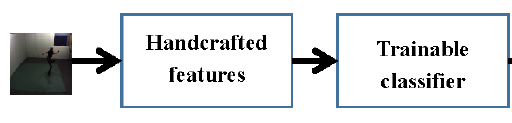


Figura 2. Esempio di approccio basato sulla rappresentazione handcrafted [22].

* Attraverso l’apprendimento profondo (deep learning) approccio più recente emerso nuovamente grazie anche alla possibilità di poter ottenere grandi quantità di dati con capacità di apprendimento automatico delle funzioni dai dati grezzi (raw data). Ciò elimina la necessità di dover adoperare descrittori o rilevatori di funzioni handcrafted come nell’approccio tradizionale (3).

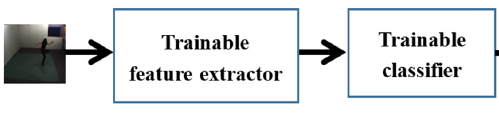


Figura 3. Esempio di approccio basato sul deep leraning [22].

**2.1.1 – Aproccio basato sulla rappresentazione di funzioni handcrafted**

L'approccio tradizionale per il riconoscimento dell'azione si basa sulla rappresentazione dell'azione artigianale (handcrafted). Questo approccio è stato popolare tra la comunità e ha ottenuto risultati notevoli su diversi set di dati pubblici noti. In questo approccio, le caratteristiche importanti di una sequenza di fotogrammi (frame) vengono estratti. Quindi, la classificazione viene eseguita addestrando un classificatore generico come Support Vector Machine (SVM).

**Fast Fight Detection**

Nell'ambito della visione artificiale, il riconoscimento delle azioni è diventato un argomento attivo. Per rilevare le sequenze di violenza, viene proposto un nuovo metodo [97]. Si presume che nelle scene di combattimento, i blob di movimento abbiano una forma e una posizione specifiche. Innanzitutto, viene calcolata la differenza tra fotogrammi consecutivi per immagini assolute. Quindi l'immagine risultante viene binarizzata, portando verso il numero di blob di movimento e contrassegnando il più grande in una sequenza di combattimento e in una scena non di combattimento. Vengono selezionati solo i più grandi BLOB di movimento. Per classificare i BLOB K, vengono calcolati parametri diversi come centroide, area, perimetro e distanza tra i BLOB. Quindi, i BLOB sono caratterizzati come lotta e non lotta. Gli esperimenti vengono eseguiti utilizzando il set di dati Movies con 200 clip, il set di dati Hockey con 1000 clip e il set di dati UCF-101 che consiste in video reali raccolti da YouTube. I risultati dell'esperimento mostrano che il metodo proposto è sovraperformato rispetto ai metodi all'avanguardia considerati BoW (MoSIFT), BoW (SIFT), ViF, LMP, variante v-1 e variante v-2 che utilizzavano SVM, KNN e Ada aumentano come classificatore in termini di precisione e ROC. Ha un tempo di calcolo espressamente più rapido, quindi è piacevole per le applicazioni in tempo reale.

**Riconoscimento e rilevazione della violenza umana**

Il riconoscimento e l'individuazione della violenza è diventato un argomento importante per i video di sorveglianza. Lo scopo di base è determinare se la violenza si verifica. In primo luogo, viene proposta l'estensione di Improved Fisher Vectors (IFV) per i video clip [61]. Vengono utilizzate le funzionalità locali e le loro posizioni spazio-temporali che consentono di rappresentare il video. Quindi, per il rilevamento della violenza, viene studiato l'approccio popolare della finestra scorrevole. Per accelerare l'approccio, viene utilizzata la struttura dei dati della tabella di area sommata e le formule di IFV vengono riformulate. In primo luogo, le funzionalità spazio-temporali locali vengono estratte dai video utilizzando le traiettorie dense migliorate (IDT). Quindi, la rappresentazione video per ciascun descrittore viene calcolata indipendentemente come HOG per rappresentare il video utilizzando IFV. Quindi, i classificatori SVM lineari vengono utilizzati per il riconoscimento della violenza e alla fine, utilizzando un approccio a finestra a scorrimento rapido, viene rilevata la violenza. Una valutazione approfondita viene eseguita utilizzando 4 set di dati all'avanguardia di Violent-Flows, Movies e Hockey Fight. Il set di dati Violence-Flow viene utilizzato per l'attività di rilevamento della violenza. I risultati mostrano che gli approcci proposti si comportano meglio rispetto agli approcci esistenti come HNF, Jerk, HOF, HOG, Violent Flow (ViF) rispetto alle metriche quali AUC e curva ROC.

**Rilevamento della violenza mediante Oriented Violent Flow**

La domanda di mercato per il rilevamento intelligente della violenza è in costante aumento ed è un'area di ricerca stimolante. Innanzi tutto, viene proposto un innovativo metodo di estrazione di funzioni denominato Oriented Violent Flows (OViF) per il rilevamento pratico della violenza nei video [57]. Negli orientamenti statistici del movimento, si sfrutta appieno le informazioni sul cambiamento della magnitudine del movimento. AdaBoost viene utilizzato per la selezione delle features e quindi il classificatore SVM viene addestrato sulle features scelte. Gli esperimenti vengono condotti su set di dati del database Hockey e Violent-Flow per valutare l'utilità del metodo proposto e i risultati mostrano che il metodo proposto funziona meglio dei metodi di base: LTP e ViF in termini di precisione e AUC. In secondo luogo, vengono adottate strategie di combinazione di features e combinazione multi-classificatore ottenendo risultati migliori. I risultati dell'esperimento mostrano che l'utilizzo di features combinate con AdaBoost e Linear-SVM consente di ottenere prestazioni migliorate nel benchmark Violent-Flows.

**2.1.2 – Approccio basato sul deep learning**

Studi recenti [18], [20], [22] mostrano che non ci sono descrittori di caratteristiche handcrafetd universali per tutti i set di dati, pertanto l'apprendimento delle funzionalità direttamente dai dati non elaborati può essere più vantaggioso. L'apprendimento profondo (deep learning) è un'area importante dell'apprendimento automatico finalizzata all'apprendimento a più livelli di rappresentazione e astrazione che possono dare un senso a dati come discorso (speech), immagini e testo. Questi approcci hanno la capacità di elaborare le immagini / video nelle loro forme grezze e automatizzare il processo di estrazione, rappresentazione e classificazione delle caratteristiche. Questi approcci utilizzano estrattori di funzionalità e modelli computazionali con più livelli di elaborazione per la rappresentazione dell'azione e riconoscimento di una scena. Secondo l’articolo [18] può essere fatta una categorizzazione delle architetture deep:

* Space-time Networks [1], [10], [14], [18], [20], [22], [25], [26], [33], [34], [35], [36], [37], [38], [43], [45], [48], [52]: catturano informazioni temporali usando convoluzioni su frame adiacenti e quindi estraggono le feature in entrambe e dimensioni spaziali e temporali
* Multi-Stream Networks [1], [18], [8], [9], [20], [22], [33], [34], [48], [52], [54], [51]: contengono due flussi separati spaziale e temporale. Il primo impara azioni da immagini statiche e il secondo esegue il riconoscimento in base al campo di flusso ottico.
* Hybrid Networks [1], [6], [8], [9], [12], [13], [15], [18], [20], [22], [33], [34], [48], [52]: aggiungere uno strato ricorrente sopra la CNN come gli LSTM. Hanno dimostrato risultati promettenti nel catturare modelli di movimento spaziale, temporale.

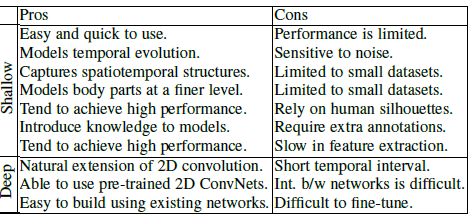


Figura 4. Pro e contro degli approcci Shallow e Depp [18].

**Sezione 3 – Dataset**

In questa sezione vengono discussi set di dati pubblici noti per il riconoscimento delle azioni violente e della rilevazione di una scena violenta.

**UCF101**

È un dataset [80] di azioni realistiche prese da YouTube che comprende 101 categorie di azioni. Ha una grande diversità in termini di azioni e una grande variazione nel movimento della camera, nelle comparse e pose di oggetti, nei punti di vista, nel background disordinato e nelle varie condizioni di illuminazione. Comprende anche 50 azioni derivanti da diversi sport. La risoluzione dei video è di 320x240p. L’obbiettivo di questo dataset è quello del riconoscimento delle azioni.

**HMDB51**

Collezione di circa 7000 clip video provenienti da diverse sorgenti per lo più film e una piccola parte da ridotti dataset pubblici. I video sono suddivisi in 51 categorie ognuna contenente un minimo di 101 clip. Le categorie possono essere raggruppate come segue:

* Azioni facciali quale sorridere, ridere, masticare, parlare.
* Azioni facciali con manipolazione di oggetti: fumare, mangiare, bere.
* Movimenti generali del corpo: battere le mani, arrampicarsi, salire le scale, tuffarsi, cadere sul pavimento, verticale, saltare, tirare su, spingere su, correre, sedersi, capriola, alzarsi, girare, camminare.
* Movimenti del corpo con interazione di oggetti: spazzolare i capelli, catturare, dribblare, giocare a golf, colpire qualcosa, calciare la palla, scegliere, versare, spingere qualcosa, andare in bicicletta, andare a cavallo, uso di arco, sparare con armi da fuoco, utilizzo di mazze da baseball, esercizio con la spada, lancio.
* Movimenti del corpo per l'interazione umana: scherma, abbraccio, calciare qualcuno, bacio, pugno, stringere la mano, combattimento con la spada.

**UT-Interaction**

Il set di dati UT-Interaction suddiviso in due parti contiene video di esecuzioni continue di 6 classi di interazioni uomo-uomo: stringere la mano, indicare, abbracciare, spingere, calciare e colpire. Vengono fornite etichette di base per queste interazioni, inclusi intervalli di tempo e riquadri di delimitazione. Vi sono in totale 20 sequenze video la cui durata è di circa 1 minuto. Ogni video contiene almeno un'esecuzione per interazione, fornendo in media 8 esecuzioni di attività umane per video. Numerosi partecipanti con più di 15 diverse condizioni di abbigliamento appaiono nei video. I video sono ripresi con una risoluzione di 720x480, 30 frame per secondo (fps) e l'altezza di una persona nel video è di circa 200 pixel. Il set 1 è composto da 10 sequenze video riprese in un parcheggio. I video del set 1 sono ripresi con una velocità di zoom leggermente diversa e i loro sfondi sono per lo più statici. Il set 2 (ovvero le altre 10 sequenze) viene portato su un prato in una giornata ventosa. Lo sfondo si muove leggermente (ad es. i movimenti dell'albero) e contengono più nervosismi della videocamera. Ogni set ha uno sfondo, una scala e un'illuminazione diversa.

**UCF Crime**

UCF-Crime è un recente set di dati su larga scala nel suo genere con 128 ore di video. Consiste in 1900 video di sorveglianza del mondo reale lunghi e non tagliati, con 13 anomalie realistiche tra cui: abuso, arresto, incendio doloso, assalto, incidente stradale, furto con scasso, esplosione, combattimento, furto, taccheggio e vandalismo. Queste anomalie sono selezionate perché hanno un impatto significativo sulla sicurezza pubblica. Questo set di dati può essere utilizzato per due attività. Innanzitutto, il rilevamento generale delle anomalie considerando tutte le anomalie in un gruppo e tutte le normali attività in un altro gruppo. In secondo luogo, per riconoscere ciascuna delle 13 attività anomale.

**Sezione 4 – Sperimentazione**

Per la sperimentazione sono stati scelti 4 dataset: HMDB51, UCF Crime, UCF101 e UT-Interaction. Le fasi della sperimentazione sono state le seguenti:

* Test diretto: utilizzo dello stesso dataset per train e test.
* Test cross-dataset: 3 dataset per il train e il restante per il test.
* Test generale: unione di tutti i dataset con relativa fase di train e test.
* Test classificazione violenza: utilizzo dello stesso dataset ma suddiviso in molteplici classi di violenza, es: punch, kick, shoot, sumo …
* Test su video di bullismo: utilizzo dei modelli creati con il test generale per classificare i video di bullismo e fase di train e test sul medesimo dataset.

I sistemi utilizzati sono stati 5: Vif, Fast Fight, Fast Video, Inception V3, ResNet 50 e la rete deep SlowFast. I primi due sono una replica di sistemi presenti nello stato dell’arte. La scelta è stata fatta prendendo quei sistemi che più si adattavano ad una implementazione in real time, sia da un punto di vista di accuratezza che di tempo di classificazione dei video.

Il primo sistema è stato replicato secondo [56], questo sistema considera le statistiche di come le magnitudini di vettori optical flow cambino nel tempo. Queste statistiche, collezionate per una piccola sequenza di frame, sono rappresentate sotto il nome di features Violent-Flows (ViF). Questo sistema utilizza per le stesse feature due classificatori differenti e indipendenti: SVM, Adaboost.

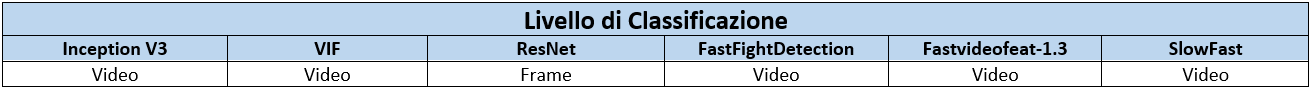
Il secondo metodo si basa sull’articolo [97], le features vengono estratte da quelli che vengono chiamati in computer vision “blobs”, cioè quelle regioni di una scena che differiscono in alcune caratteristiche come il colore o l’illuminazione, che vanno in contrasto con altre regioni dei frame. Il sistema va a rilevare i K blobs più importanti per ogni frame estraendo features che verranno aggregate in un’unica rappresentazione. Questo sistema utilizza per le stesse feature due classificatori differenti e indipendenti: SVM, Adaboost.

Fast Video è stato implementato in python. Le feature scelte per il calcolo dei fisher vector sono state: HOF, HOG, MBH. L’estrazione dei fisher vector è avvenuta tramite l’utilizzo del tool per il calcolo dei vettori messo a disposizione da [66]. Il GMM è stato addestrato sul dataset Holliwood2, tale dataset è stato scelto per la varietà di azioni al suo interno. Questo sistema utilizza per le stesse feature due classificatori differenti e indipendenti: SVM, Random Forest.

Gli ultimi tre sistemi sono deep neural network, addestrando due CNN, una basata sull’architettura Inception V3 mentre l’altra su ReNet 50. In fine SlowFast [96] nonché una rete C3D, essa addestra una rete convoluzionale a tre dimensioni. Quest’ultima rete è stata utilizzata per l’esecuzione del task di classificazione del tipo di violenza; nonché operazione complessa che altre reti più semplici non permettono di svolgere.

Tutte le reti adottate svolgono una classificazione a livello Video, l’unica eccezione è rappresentata da ResNet che svolge una classificazione a livello di frame.

Inception V3 svolge una classificazione a livello di frame, ma in seguito riassume attraverso una media sui valori delle singole previsioni [Violenza (<=0.5) / Non Violenza (>0.5)] per poter classificare l’intero video.



**Test diretto**

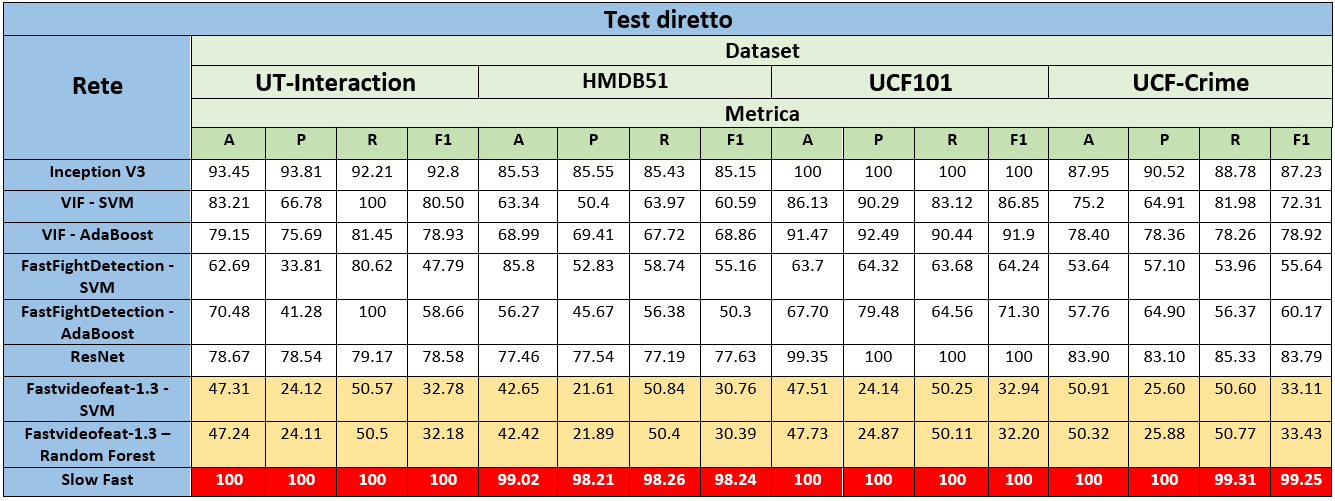
In questa prima fase i vari sistemi sono stati addestrati e testati su singoli dataset utilizzando le medesime proporzioni di suddivisione dei video per il train e test secondo [29], [30], [33], [59], [44], [77], [78]. Per HMDB51 sono stati scelti 103 video della categoria kick, 130 shoot gun, 126 punch [29], [30]. Per UCF101 271 video di combattimento da Punch, SumoWrestling e 271 dalle categorie BooxingSpeedBag e BooxingPunchingBag come esempi negativi di combattimento [30], [33], [59]. Per il dataset UT-Interaction le classi utilizzate sono state handshaking, higging e pointing come esempi negativi di violenza ed kicking, punching e pushing come esempi positivi di violenza. Il dataset è suddiviso in set 1 e 2. Un terzo di set 1 e 2 sono stati utilizzati come test mentre la restante parte come training set [44]. UCF Crime contiene 810 video anormali e 800 normali per il training mentre 140 anormali e 150 normali per il test set [78].

Figura 5. A = accuratezza, P = precisione, R = richiamo, F1 = F1-measure. I valori sono espressi in %.

Figura 5. Tabella rappresentante il test diretto. Le metriche adottate sono: A accuratezza, P precisione, R richiamo, F1 media armonica. I valori sono espressi in percentuale

I valori evidenziati in rosso indicano i test che hanno dimostrato performance superiori rispetto agli altri sistemi adottati. Invece i valori evidenziati in giallo indicano i test con i sistemi che hanno raggiunto performance insufficienti rispetto lo stato dell’arte al task della Violence Detection.

**Tempi di esecuzione delle reti**

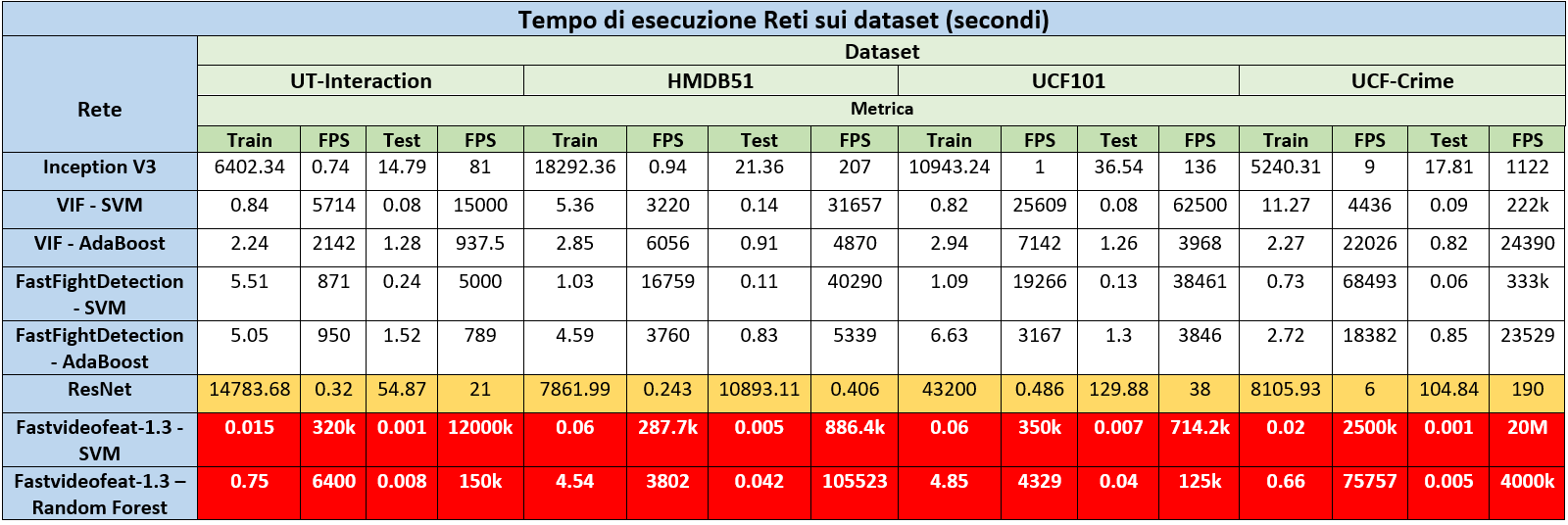
****In questa tabella sono riassunti i tempi di esecuzione di test e di train dei singoli sistemi adottati sui singoli dataset. Ogni valore è indicato in secondi.

Figura 6. FPS = frame/secondi Train = secondi esecuzione fase di train Test: secondi esecuzione fase di test

*UT-Interaction: è costituito da 4800 frame per il set di Train e 1200 frame per il set di Test.*

*HMDB51: è costituito da 17262 frame per il set di Train e 4232 frame per il set di Test.*

*UCF101: è costituito da 21000 frame per il set di Train e 5000 frame per il set di Test.*

*UCF Crime: è costituito da 50000 frame per il set di Train e 20000 frame per il set di Test.*

I valori evidenziati in rosso indicano i test che hanno dimostrato performance temporali superiori rispetto agli altri sistemi adottati. Invece i valori evidenziati in giallo indicano i test con i sistemi che hanno raggiunto tempi di esecuzioni molto elevati

**Test cross-dataset (3vs1)**

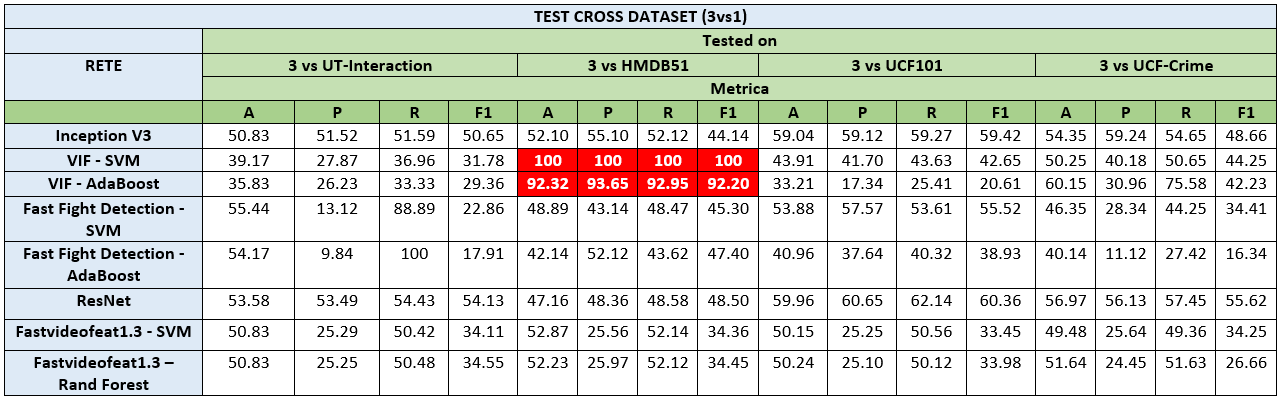
Per questa face invece è stato adottato il metodo 3 vs 1. Simile alla tecnica del k-fold cross validation. A turno un dataset veniva utilizzato per la fase di test mentre gli altri tre come train set. 

Figura 7. Risultati per la fase di sperimentazione cross-dataset, i risultati sono espressi in percentuale.

Come si evince dalla tabella i risultati non sono stati soddisfacenti tranne per un singolo caso, quello evidenziato di rosso. La ragione è da attribuire alla natura differente delle scene dei diversi dataset. Il singolo sistema non riesce a generalizzare perfettamente e non riesce a classificare correttamente scene violente di un dataset sul quale non è stato addestrato.

**Test cross-Dataset (1vs1)**

In questa tabella sono riassunti i test cross dataset dei singoli sistemi addestrati su dataset differenti rispetto a quelli su cui sono stati in seguito testati. Ogni valore è espresso in percentuale.

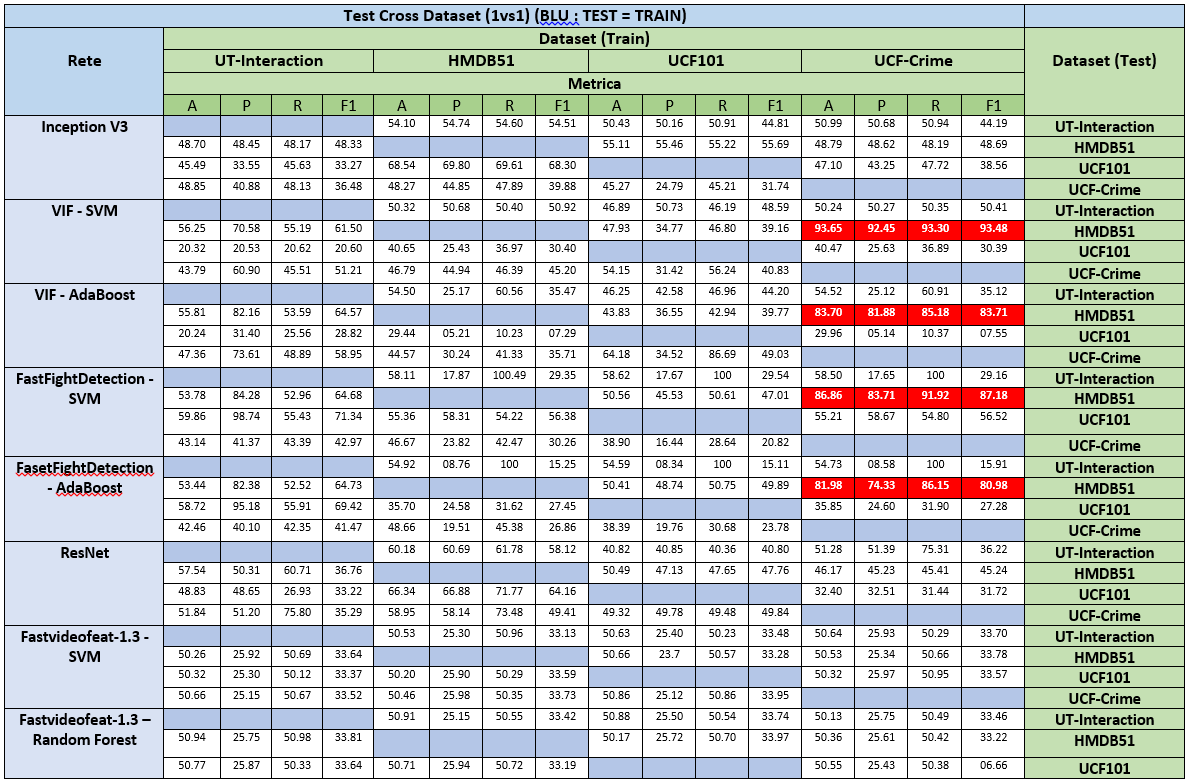
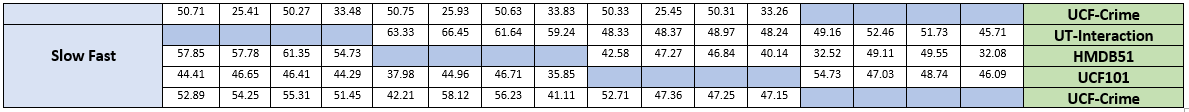
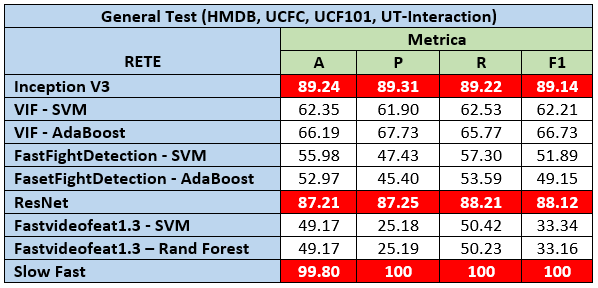


Figura 8. La tabella illustra i risultati relativi al test cross dataset 1vs1. Le metriche adottate sono: A accuratezza, P precisione, R richiamo, F1 media armonica. I valori sono espressi in percentuale

I valori evidenziati in rosso indicano i test che hanno dimostrato performance temporali superiori rispetto agli altri sistemi adottati.

**Test generale**

Nel test generale tutti i dataset sono stati riuniti per formare un dataset globale che presenta scene violente e non violente eterogenee tra loro.

I valori evidenziati in rosso indicano i test che hanno dimostrato performance superiori rispetto agli altri sistemi adottati.

Figura 9. Le metriche adottate sono: A accuratezza, P precisione, R richiamo, F1 media armonica. I valori sono espressi in percentuale

**Test** **Classificazione del tipo di Violenza**

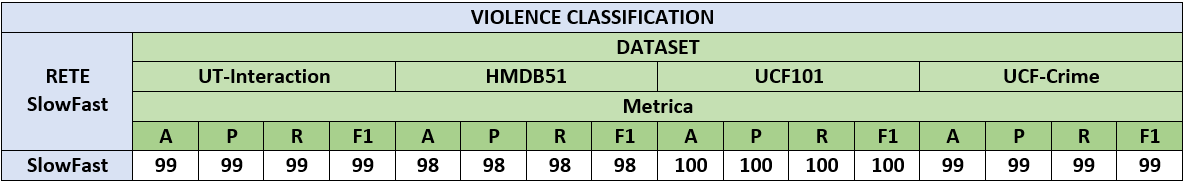
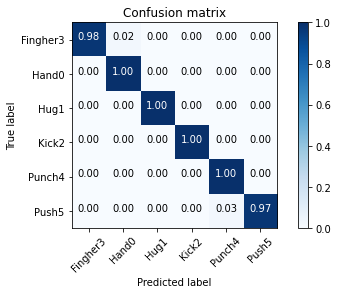
****Nel test di classificazione del tipo di violenza tutti i dataset sono stati testati utilizzando la rete deep SlowFast [96], nonché unica rete in grado di ottenere ottime prestazioni su questa tipologia di task.

Figura 10. La tabella illustra i risultati del test di classificazione della violenza. Le metriche adottate sono: A accuratezza, P precisione, R richiamo, F1 media armonica. I valori sono espressi in percentuale



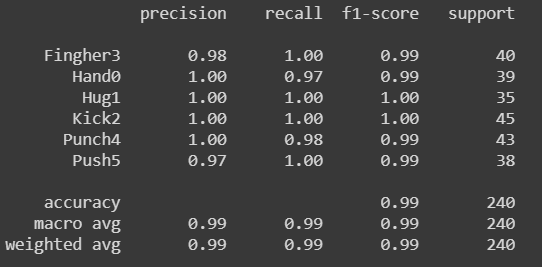
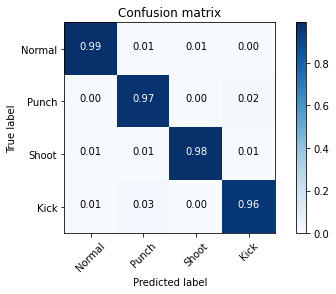
****

Figura 11. Report dataset UT-Interaction



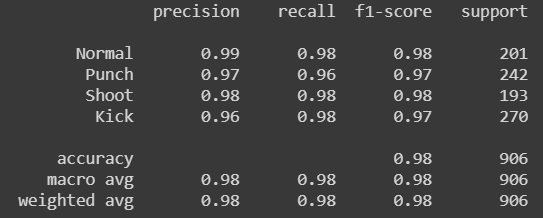
****

Figura 12. Report dataset HMDB51

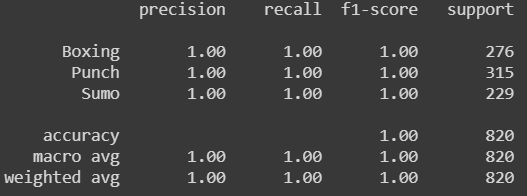
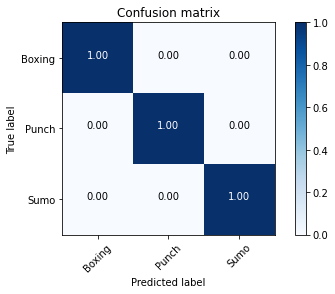
****

Figura 13. Report dataset UCF101

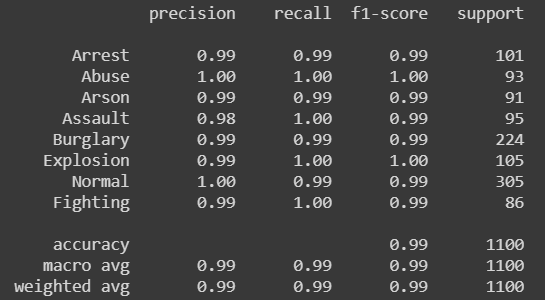
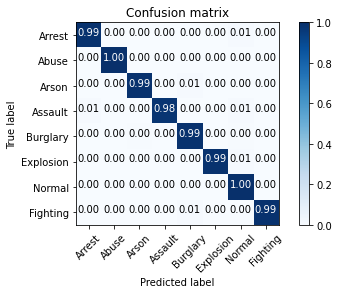
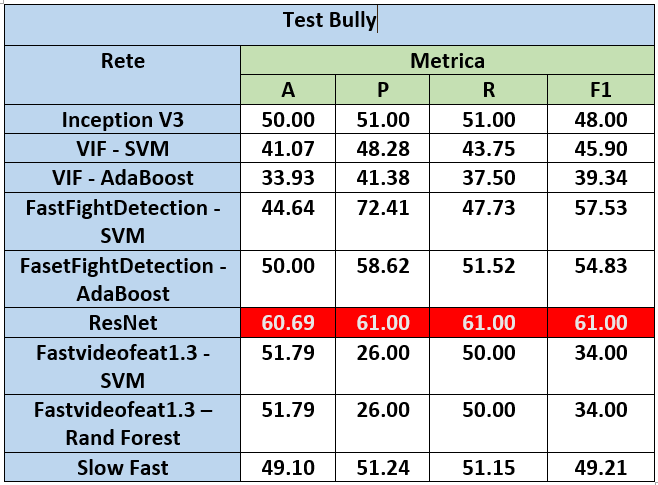
****

Figura 14. Report dataset UCF Crime

**Test su video di bullismo**

In questa fase finale abbiamo utilizzato i modelli creati dalla fase di test generale per classificare i video di un dataset artigianale contenenti video relativi ad atti di bullismo. Questo dataset contiene 56 video con situazioni completamente sconosciute ai modelli creati precedentemente, infatti le scene sono ambientate in strutture scolastiche.

Figura 15. Riepilogo dei test sul dataset bully con i rispettivi sistemi adottati. I risultati sono in %. I valori evidenziati in rosso sono stati i migliori. A = accuratezza, P = precisione, R = richiamo, F1 = media armonica.

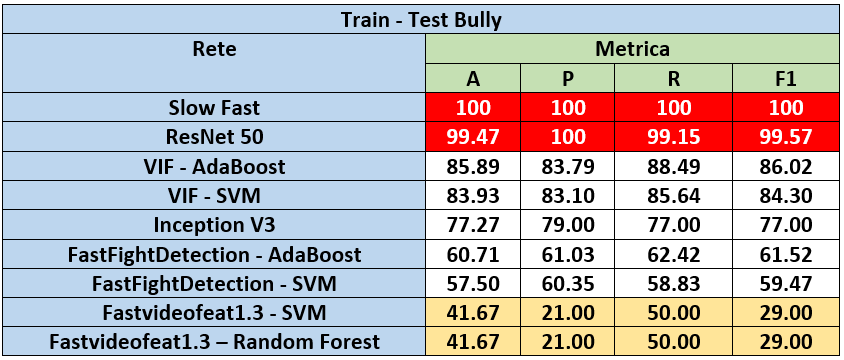


Figura 16. Tabella riepilogativa dei risultati sul dataset bully con fase di train e test sul medesimo dataset. I risultati sono in %. I valori evidenziati in rosso sono stati i migliori quelli in giallo i peggiori. A = accuratezza, P = precisione, R = richiamo, F1 = media armonica.

**Sezione 5 – Confronto per classificazione violenza**

|  |  |  |
| --- | --- | --- |
| **UT-Interaction** | | |
| **Metodo** | **Metrica (%)** | **Train / Test** |
| Neural Network[44] | A = 78 | Test = 33%  Train = 66% |
| Slow Fast | A = 99 |

|  |  |  |
| --- | --- | --- |
| **HMDB51** | | |
| **Metodo** | **Metrica (%)** | **Train / Test** |
| SVM gerarchica (Fourier descriptor, shape moments) [29] | A = 95.3 | Test = 50 / 50 / 50  Train = 130 /130 /126 |
| Naive Bayes (Fourier descriptor, shape moments) [29] | A = 94 |
| Slow Fast | A = 98 |

|  |  |  |
| --- | --- | --- |
| **UCF101** | | |
| **Metodo** | **Metrica (%)** | **Train / Test** |
| SVM (HOA, HOP, HOD) [59] | A = 93.4 | Test = 20 %  Train = 80 % |
| Adaboost (HOA, HOP, HOD) [59] | A = 92.8 |
| Random Forest (Blobs area, centroids, compactness, distance blob centroids) [87] | A = 79.5 |
| Slow Fast | A = 100 |

|  |  |  |
| --- | --- | --- |
| **UCF Crime** | | |
| **Metodo** | **Metrica (%)** | **Train / Test** |
| VGG 16 [71] | A = 72.66 | 50 per classe (7+1 classi)  Test = 20 %  Train = 80 % |
| VGG 19 [71] | A = 71.66 |
| Flow Net [71] | A = 71.33 |
| Dearest [71] | A = 76.66 |
| VGG 16 + LSTM [76] | A = 85 |
| I3D [77] | A = 67.9 |
| Slow Fast | A = 99 |

**Conclusioni e Sviluppi futuri**

Per un’applicazione real-time utile al riconoscimento della violenza all’interno di contenuti video è consigliabile l’adozione di reti con un ottimo rapporto tempo/prestazioni. Ad esempio la rete deep Inception V3 ha dimostrato di poter classificare video come violenti o non violenti nella fasi di testing a velocità di FPS molto elevate, circa: 386 FPS in media tra i 4 dataset.

Per quanto riguarda la classificazione selettiva del tipo di violenza, l’unica rete consigliabile è SlowFast, poiché ha dimostrato tempi di esecuzione molto elevati e prestazioni superiori allo stato dell’arte.

Le reti shallow implementate in matlab: (Fast Fight Detection e Vif) hanno dimostrato prestazioni inferiori alle reti deep, ma notevolmente superiori dal punto di vista dei tempi di esecuzione, essendo in grado di classificare la presenza di violenza di un singolo video in pochi istanti.

In generale, i test cross-dateset hanno dimostrato l’incapacità dei sistemi di ottenere prestazioni elevate in scenari nei quali si sottopone il testing su sistemi addestrati su dataset differenti. Per tali scenari di utilizzo sarebbe consigliabile l’implementazione di reti in grado di estrarre informazioni spazio-temporali relative allo scheletro dei soggetti a schermo; permettendo di poter ottenere una forma di generalizzazione assoluta dell’azione svolta dall’individuo.

**BIBLIOGRAFIA**