Action Classification Fight with Open Pose

Gaetano Cimino
Emilio Mainardi
Bernardino Sagliocca
Dipartimento di Informatica
University of Salerno, Salerno, Italy
{g.cimino2, e.mainardi, b.sagliocca}@studenti.unisa.it

***Abstract* -- Questo paper esplora la possibilità di identificare possibili atti di violenza andando a considerare la distanza tra le varie parti del corpo degli individui coinvolti in un video. I movimenti vengono determinati matematicamente attraverso l'algoritmo di Real time Multi-Person Pose Estimation, Open Pose [2]. In particolare, esso permette di recuperare 18 articolazioni scheletriche del corpo come naso, ginocchia, caviglie. Da tali informazioni vengono calcolate le distanze che permettono al modello di classificare se una coppia di individui è in contrasto. Tale problema affrontato potrebbe essere utilizzato per intensificare la sicurezza in vari contesti per mezzo dei video acquisiti tramite videocamere di sorveglianza.**

I. INTRODUCTION

Una tecnologia in forte sviluppo di cui sentiamo poco parlare è la video analisi, meglio conosciuta come video sorveglianza. I sistemi di ripresa che conosciamo, applicati in ambiti di controllo e prevenzione come le telecamere a circuito chiuso di una banca o i sistemi antifurto, stanno avendo un'evoluzione e un potenziamento notevole grazie all'integrazione tra i sistemi di ripresa e di gestione delle immagini, di nuove intelligenze artificiali. Nell'analisi della sicurezza, ad esempio quando si sorveglia un aeroporto o luoghi in cui si ha la presenza di un numero elevato di persone, è necessario dover individuare eventuali situazioni di violenza che si possono verificare. Ovviamente, ciò permetterebbe di evitare molte situazioni spiacevoli e di intervenire prontamente. Un modo per riconoscere se due persone sono in contrasto è analizzare i movimenti che essi compiono. Tali movimenti sono determinabili dalla postura degli arti e dalle distanze tra le varie parti del corpo. Pertanto, l'andatura può essere utilizzata come biometrica per il riconoscimento di scontri in quanto risulta essere non invasiva e può essere acquisita facilmente a una certa distanza tramite dispositivi di acquisizione, quali telecamere e/o dispositivi mobile. Fino ad ora, l'analisi dell'andatura doveva essere eseguita principalmente da esperti del settore, con l'attuale progresso della tecnologia, i computer possono aiutare a rendere l'analisi dell'andatura più affidabile. Sebbene molte ricerche abbiano espresso preoccupazioni per quanto riguarda la sorveglianza supportata dall'apprendimento automatico, il campo della Computer Vision ha registrato dei notevoli progressi. Per rilevare i combattimenti in modo economico e naturale, viene proposto un approccio basato sull'analisi del movimento. In questo lavoro, abbiamo proposto un metodo per classificare se due persone sono in conflitto tramite l'analisi dei landmarks della postura che le persone assumono nella loro andatura durante i combattimenti. Abbiamo affrontato il problema da una prospettiva geometrica considerando diversi punti del corpo da diversi fotogrammi video. Diversi algoritmi possono essere utilizzati per l'estrazione dei punti del corpo, in questo progetto è stato utilizzato Open Pose. Esso è stato utilizzato per stimare lo scheletro umano 2D. Questi punti vengono determinati per calcolare le distanze tra le parti del corpo al fine di riconoscere se i soggetti sono in contrasto. Per tale scopo è stato utilizzato il dataset "Real-Life-Violence-Situations-Dataset", che contiene video di persone coinvolte in situazioni di "VIOLENCE" e "NONVIOLENCE". Il documento è organizzato come segue: nella sezione successiva ci occuperemo dello stato dell'arte. Nella sezione 3 viene descritto l'approccio utilizzato. Inizialmente, viene introdotto il dataset considerato e viene descritto la pipeline dell'algoritmo, dalla selezione dei punti del corpo agli algoritmi di classificazione utilizzati. Infine, nella sezione 4 vengono presentati i risultati e nella sezione 5 le conclusioni.

II. RELATED WORK

Il metodo proposto in [1] sfrutta l'essenziale comunanza tra le lotte animali e umane come l'accelerazione fisica delle parti del corpo in movimento. Il metodo proposto prende input dei video di lotta tra animali e alcuni video di lotta tra persone. Viene proposto un set basato su Local Motion Feature (LMF) che include statistiche di movimento, correlazione dei segmenti seguendo il paradigma dell'analisi del movimento. LMF sono estratti da ogni video. Vengono estratte le caratteristiche temporali basate sull'euristica umana e per rilevare la lotta vengono adottati algoritmi tradizionali di machine learning come SVM. Vengono proposti classificatori di ensemble per eseguire il rilevamento di combattimenti tra specie diverse. Gli esperimenti vengono eseguiti utilizzando campioni di videoclip, set di dati di hockey e film. In [3] viene proposto un innovativo metodo di estrazione di funzioni denominato Oriented Violent Flows (OViF) per il rilevamento pratico della violenza nei video. Negli orientamenti statistici del movimento, sfrutta appieno le informazioni sul cambiamento della magnitudine del movimento. AdaBoost viene utilizzato per la selezione delle feature e quindi il classificatore SVM viene addestrato sulle feature scelte. Vengono condotti esperimenti sui set di dati del database Hockey e ViolentFlow per valutare l'utilità del metodo proposto. In secondo luogo, vengono adottate strategie di combinazione di feature e combinazione multi-classificatore. I risultati dell'esperimento dimostrano che l'utilizzo di feature combinate con AdaBoost e Linear-SVM consente di ottenere prestazioni migliorate attraverso i metodi all'avanguardia nel benchmark Violent-Flows.

III. APPROACH

L'approccio di questo progetto è stato quello di estrarre prima i fotogrammi consecutivi da video preregistrati recuperati dal dataset "Real-Life-Violence-Situations-Dataset", quindi eseguire una stima della posa su ciascun fotogramma per ottenere le coordinate delle articolazioni scheletriche di ciascun soggetto e infine di calcolare le distanze tra determinate parti del corpo. Viene quindi studiata l'andatura di ciascun soggetto in un fotogramma e tramite il calcolo delle distanze viene effettuato un task di classificazione per determinare se in tale fotogramma è presente un combattimento.

A. Real-Life-Violence-Situations-Dataset

Negli ultimi anni, le telecamere di sorveglianza sono ampiamente utilizzate in luoghi pubblici e il tasso di criminalità generale è stato ridotto in modo significativo a causa di questi dispositivi onnipresenti. Di solito, queste telecamere forniscono spunti e prove dopo che i crimini sono stati condotti, mentre sono usati raramente per prevenire o fermare le attività violente in tempo. Monitorare manualmente una grande quantità di dati video dalle telecamere di sorveglianza richiede tempo e manodopera. Pertanto, il riconoscimento automatico dei comportamenti violenti dai segnali video diventa essenziale. Tale set di dati contiene 1000 video di violenza e 1000 video non violenti raccolti da YouTube. I video di violenza presenti contengono molte situazioni di combattimenti di strada reali in diversi ambienti e condizioni. Allo stesso modo, anche i video di non violenza sono raccolti da molte diverse azioni come sport, alimentazioni, passeggiate, etc.

Sono stati annotati diversi tipi di combattimenti:

- 1vs1: solo due persone in lotta
- Small: per un piccolo gruppo di persone (il numero di persone non è stato conteggiato, corrisponderà approssimativamente a meno di 10)
- Large: per un folto gruppo di persone (> 10)

B. Preprocessing

Innanzitutto, dal dataset sono stati selezionati solo i video contenenti 2 persone al fine di addestrare al meglio il classificatore sui movimenti che vengono effettuati per il combattimento. Successivamente, tale contesto è stato generalizzato anche per video contenenti più di 2 persone. Dai 2000 video presenti nel dataset ne sono stati selezionati 50 di violenza e 50 di non violenza. Fatto ciò, è stato effettuato il processo di estrazione dei frame dai video la quale sono stati tutti normalizzati ad una dimensione di 1080x1080. Quest'ultima, anche se risulta essere abbastanza elevata, è necessaria in quanto molti video presentano una risoluzione

molto bassa e quindi anche l'algoritmo Open Pose sarebbe stato poco performante. Durante il test, è stato notato un notevole calo delle prestazioni dell'algoritmo di stima della posa 2D quando nei fotogrammi vi sono 2 persone sovrapposte, in questo caso l'algoritmo perde una parte delle coordinate delle articolazioni. Per la gestione dei *missing values* si è stabilito di assegnare un valore pari a 0 quando uno o entrambi i punti necessari per il calcolo della distanza non vengono rilevati.

C. 2D Pose Estimation

Al fine di eseguire l'algoritmo di stima della posa 2D di Cao, Simon, Wei e Sheikh del 2017, la demo per l'elaborazione dell'immagine di github dei loro documenti viene utilizzata all'interno di una funzione per calcolare tutte le coordinate X e Y di 18 giunti di scheletro per ogni persona presente nel fotogramma estratto. Le coordinate vengono salvate in una matrice 18 dimensionale con elementi che descrivono le articolazioni nel seguente ordine: naso, collo, spalla destra, gomito destro, polso destro, spalla sinistra, gomito sinistro, polso sinistro, anca destra, ginocchio destro, caviglia destra, anca sinistra, ginocchio sinistro, caviglia sinistra, occhio sinistro, occhio destro, orecchio sinistro, orecchio destro. Per ciascun fotogramma vengono salvati l'array di coordinate e un'immagine del fotogramma con gli scheletri dei soggetti disegnati.

D. Proposed method

In questa sottosezione, affrontiamo il problema ACF attraverso l'uso di una tecnica geometrica basata sulle pose estratte dall'andatura umana. La pipeline è divisa in tre fasi principali:

1) estrazione dei dati

2) creazione delle caratteristiche

3) selezione dei classificatori

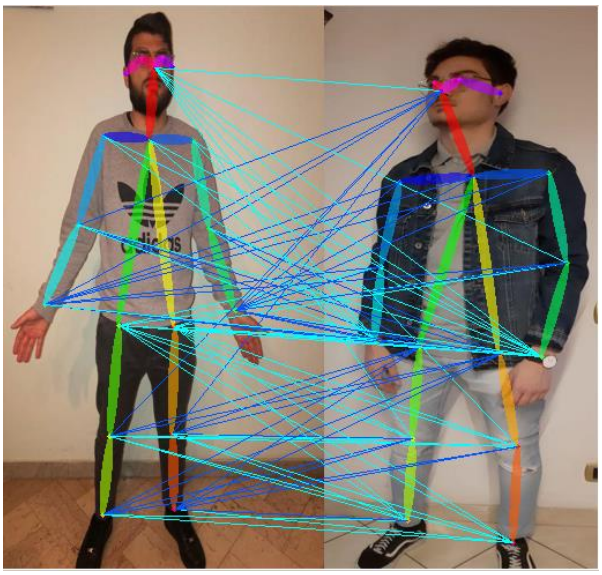

Figura 1: distanze considerate

Utilizziamo l'algoritmo descritto nel paragrafo precedente per l'estrazione dei dati necessari per la determinazione delle feature. Infatti, da tali dati estratti si è proseguiti al calcolo delle distanze tra i vari punti associati a determinate parte del corpo. Per il calcolo della distanza tra due punti è stata utilizzata la distanza euclidea. Nella Figura 1 si possono notare tutte le distanze prese in considerazione. Infatti, non sono state oggetto di analisi le distanze più insolite (ad esempio distanza naso-caviglia). Dopo aver effettuato un'analisi delle distanze più pertinenti, il numero risultante di feature considerate per ciascun individuo presente in un frame è 47 e quindi analizzando il confronto tra 2 persone, nel dataset costruito risultano 94 feature per ogni istanza, che rappresenta un confronto tra 2 persone prese in analisi. Inizialmente, il problema è stato focalizzato sull'individuazione di un combattimento all'interno di un frame contenente solo 2 soggetti. Successivamente, il problema è stato generalizzato andando invece ad analizzare frame contenenti anche più di 2 soggetti. Per ogni frame analizzato sono state prese in considerazione tutte le possibili combinazioni di coppie di persone, e quindi nel momento in cui anche solo una coppia viene etichettata come "fight" allora il frame verrà etichettato come tale. Tale processo ha generato un nuovo problema: andando a considerare tutte le possibili combinazioni di coppie di persone presenti in un frame, l'algoritmo avrebbe potuto portare un costo computazionale oneroso. Per questo motivo è stato utilizzato l'algoritmo LSH (locality sensitive hashing), che è un metodo per la

riduzione della dimensionalità dello spazio vettoriale di un insieme di dati. Si è quindi stabilito di andare a selezionare per la classificazione solo quelle coppie di persone che hanno una probabilità più alta di essere in combattimento. Per tale scopo, è stata selezionata una soglia di distanza minima da rispettare. Tale soglia è stata ricavata andando ad effettuare la media di tutte le distanze calcolate per frame etichettati "fight". Questo per andare a determinare una distanza "solita" che si verifica quando 2 persone sono in combattimento. Inoltre, è stato necessario stabilire una seconda soglia che indica il numero minimo di distanze che devono essere sotto la soglia precedentemente descritta. Si è infine stabilito che, se almeno il 30% delle distanze calcolate per una coppia di persone sono sotto la soglia, allora tale coppia viene candidata per la classificazione. Ovviamente, l'utilizzo di una soglia del 30%, essendo essa abbastanza bassa, può introdurre molti falsi positivi, cioè coppie di persone abbastanza distanti, e che quindi potrebbero essere etichettate direttamente "notfight", ma che vengono comunque selezionate per la classificazione. Appunto, tale soluzione è stata necessaria per ridurre il numero di classificazioni da effettuare, andando a etichettare direttamente come "notfight" coppie di persone all'interno di un frame molto distanti tra loro. Allo stesso tempo, tale tecnica potrebbe introdurre anche dei falsi negativi, cioè persone che dovrebbero essere etichettate come "fight" ma che vengono scartate per la classificazione ed etichettate direttamente come "notfight". Successivamente, il problema si è spostato sul classificare se un video fosse "fight" o "notfight". Da ciascun video dato in input vengono estratti tutti i frame che vengono classificati singolarmente. Si è stabilito che se il numero di frame etichettati come "fight" risulta essere maggiore del numero di frame etichettati come "notfight" allora il video verrà classificato come "fight", e viceversa. I classificatori utilizzati sono: SVM, Random Forest, KNN e Neural Network.

E. Experiments

Come già anticipato precedentemente, per la costruzione del dataset da utilizzare per effettuare il training dei classificatori, sono stati presi in considerazione solo video contenenti 2 persone e quindi sono stati analizzati solo frame con tale caratteristica: ciò per far sì che i classificatori apprendessero i movimenti tipici che vengono effettuati in un combattimento. Nell'esperimento sono state costruite 26.000 istanze, rispettivamente etichettate "fight" o "notfight", ognuna delle quali descrive le distanze tra i landmark, precedentemente mostrati, in un confronto tra 2 persone in un frame. Queste istanze sono divise: 13000 frame di violenza e 13000 frame di non violenza. Ovviamente, la distribuzione di frequenza bilanciata risulta essere essenziale per evitare distorsioni nella fase di training. Abbiamo diviso il dataset usando il 70% dei dati per il training set e il 30% dei dati per il test set. I dati non sono stati normalizzati in quanto risultano essere sulla stessa scala.

IV. RESULTS

Nel seguente paragrafo vengono illustrati i risultati ottenuti dai vari classificatori.

Classifier	Overall Accuracy
Random Forest	0.803
SVM	0.759
Neural Network	0.753
KNN	0.762

Tabella 1: Overall Accuracy

Neural Network:

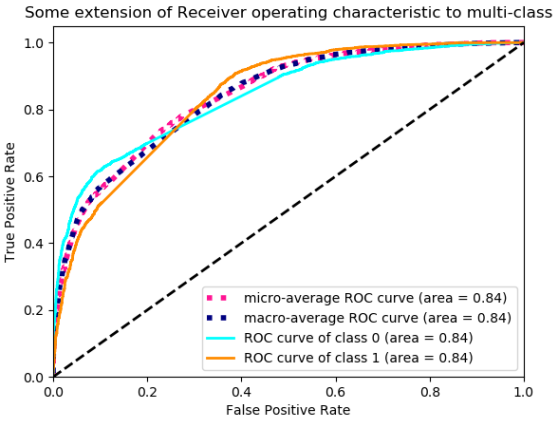

Figura 2: ROC curve

La ROC curve [7] è uno schema di rappresentazione grafica che permette di valutare la bontà di un classificatore binario. Definite le due classi che possono essere associate ad un frame, definiamo "true positive" il numero di casi in cui il classificatore identifica correttamente la classe di un frame, e "false positive" il numero di casi in cui

il classificatore sbaglia la predizione. La ROC curve viene costruita considerando tutti i possibili valori del test e, per ognuno di questi, si calcola la proporzione di true positive e la proporzione di false positive. L'area sottostante alla ROC curve è una misura di accuratezza diagnostica. Tanto maggiore è l'area sotto la curva (cioè tanto più la curva si avvicina al vertice del grafico) tanto maggiore è il potere discriminante del test. Nella Figura 2 viene illustrata la ROC curve ottenuta con la Neural Network.

KNN

Per l'algoritmo K-Nearest Neighbors il valore del parametro K scelto è 5; in base a quanto si evince dal seguente grafico, è il valore in corrispondenza del quale si presenta il tasso di errore più basso, come mostrato nella Figura 3.

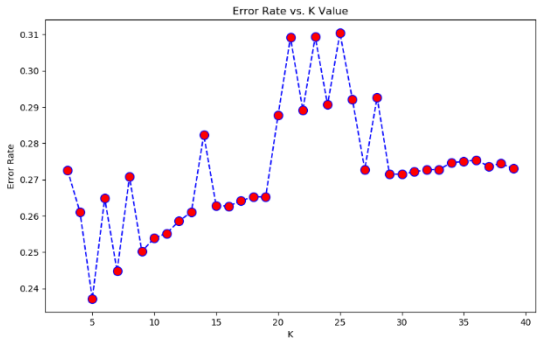

Figura 3: Scelta del valore K

Confusion Matrix

Nel campo dell'apprendimento automatico e in particolare del problema della classificazione statistica, una matrice di confusione [6] è un layout di tabella specifico che consente di ricavare una serie di parametri utili per descrivere le performance di un classificatore. L'output della confusion matrix è una matrice di due righe e due colonne, dove le righe corrispondono alla true classes e le colonne corrispondono alla predicted classes. Ogni voce conta la frequenza con cui un campione appartenente alla classe corrispondente alla riga è stato classificato come la classe corrispondente alla colonna. Essa è stata utilizzata per valutare e confrontare le performance dei classificatori KNN, SVM e Random Forest, e viene illustrata nelle figure seguenti.

KNN:

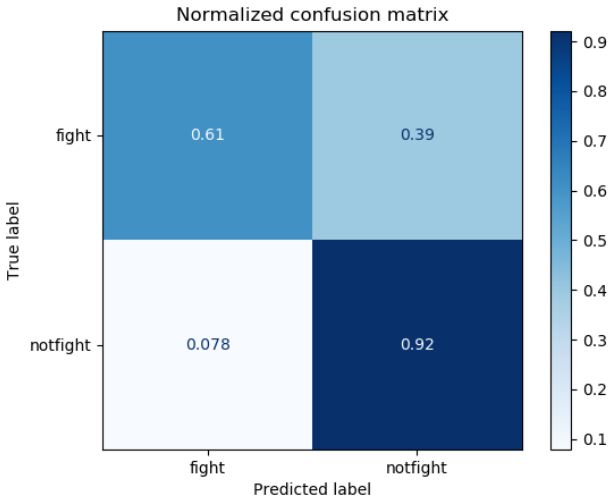

Figura 4: Confusion Matrix KNN

Random Forest:

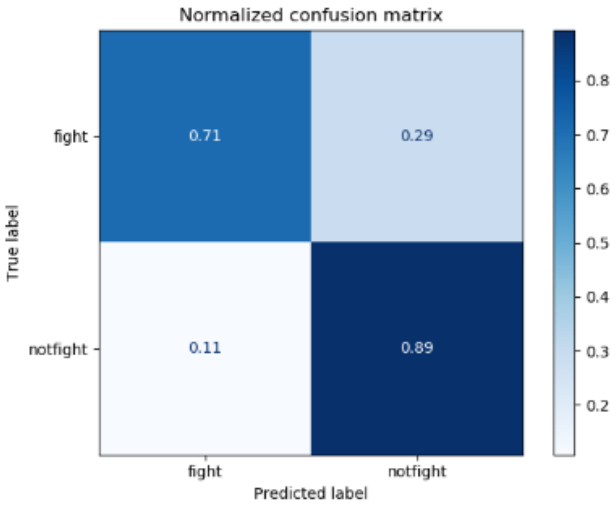

Figura 5: Confusion Matrix Random Forest

SVM:

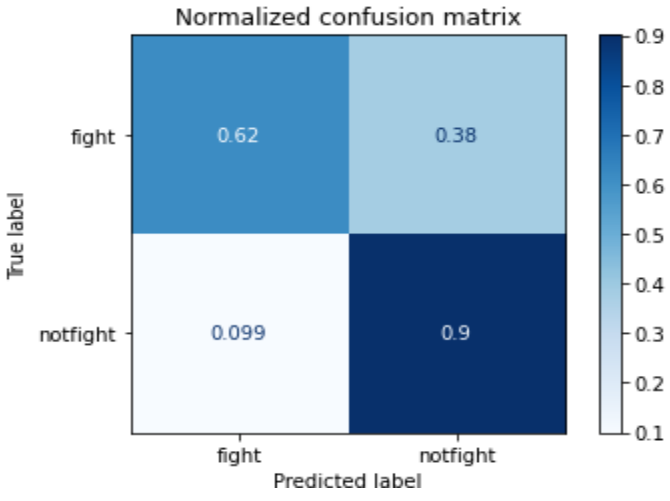

Figura 6: Confusion Matrix SVM

V. CONCLUSIONS

Il documento presentato fornisce un'interessante proposta di analisi dell'andatura per l'ACF. I risultati ottenuti dai vari classificatori sono soddisfacenti, in particolare si può notare come il Random Forest sia il più performante. Come hanno dimostrato i nostri risultati, il metodo richiede tuttavia una grande quantità di dati di training affinché l'algoritmo apprenda la tipica postura che le persone assumono in casi di combattimento. I frame estratti dai video dovrebbero essere raccolti in modo tale che i soggetti siano chiaramente visibili senza rumore sullo sfondo al fine di catturare le loro articolazioni in modo affidabile quando si utilizza Open Pose. La ricerca futura potrebbe mirare a migliorare ulteriormente l'accuracy aggiungendo nuovi dati di training e sviluppando modelli più robusti.

REFERENCES

[1]https://www.researchgate.net/publication/334778564_A_Review_on_state-of-the-art_Violence_Detection_Techniques/Cross-Species Fight Detection

[2]GithubRepository,https://github.com/michalfaber/keras Realtime MultiPerson Pose Estimation (last checked on 12/01/2020)

[3]https://www.researchgate.net/publication/334778564_A_Review_on_state-of-the-art_Violence_Detection_Techniques/Violence Detection using Oriented Violent Flow

[4]https://www.kaggle.com/mohamedmustafa/real-life-violence-situations-dataset/download

[5]Z. Cao, G. Hidalgo, T. Simon, S. Wei and Y. Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. arXiv preprint arXiv:1812.08008. 2018.

[6]Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow by Aurélien Géron

[7]https://it.wikipedia.org/wiki/Receiver_operating_characteristic