Dipartimento di Ingegneria e Scienza dell'Informazione

Corso di Laurea in
Ingegneria dell’informazione e delle comunicazioni

ELABORATO FINALE

ACTION RECOGNITION NELL'AMBITO DELLA PALLACANETRO TRAMITE ZED CAMERA

Supervisore
NICOLA CONCI

Laureando
CASTELLAN SEBASTIANO

Anno accademico 2019/2020

INDICE

Ringraziamenti

Prima di procedere con la trattazione, vorrei dedicare qualche riga a tutti coloro che mi sono stati vicini in questo percorso di crescita personale e professionale.
Un sentito grazie al mio relatore e ai dottorandi per la sua infinita disponibilità ad ogni mia richiesta. Grazie per avermi fornito ogni materiale utile alla stesura dell'elaborato.
Senza il supporto morale di mia madre, non sarei mai potuto arrivare fin qui. Grazie per esserci sempre stata soprattutto nei momenti di sconforto.
Ringrazio i miei colleghi per essermi stati accanto in questo periodo intenso e per gioire, insieme a me, dei traguardi raggiunti.
Grazie a tutti, senza di voi non ce l'avrei mai fatta.

Sommario

L’Action recognition è un caso della pattern recognition, una branca della computer-vision che si propone come obbiettivo quello di rilevare l’avvenimento di certe azioni estrapolando l’informazione da sequenze video.
La modellazione e l'apprendimento delle caratteristiche estratte sono la parte critica del problema, nel migliorare l'accuratezza del riconoscimento. Alcune tecniche popolari includono il rilevamento ottico del movimento, la rappresentazione in 3D del volume, la modellazione temporale, lo sviluppo dell’algoritmo nel riconoscimento della specifica azione rilevandone le peculiarità.

Il presente studio si propone di rilevare azioni specifiche eseguite dai direttori di gara della pallacanestro.
Verranno utilizzate diverse tecnologie per lo sviluppo.

- stereo camere per l’acquisizione di immagini in 3 dimensioni che possano rivelare la profondità dell’ambiente in particolare le camere utilizzate sono le ZEDCAM.
- Openpose una libreria di riconoscimento artificiale per rilevare gli scheletri delle persone inquadrate dalle camere.
- I dati poi verranno analizzati tramite un motore grafico multipiattaforma, chiamato Unity. Esso è principalmente utilizzato per lo sviluppo di videogiochi a livello commerciale, visualizzazioni architettoniche, animazioni 3D in tempo reale e per la ricerca.
- Linguaggio di programmazione utilizzato è il c#, è un linguaggio di programmazione orientato agli oggetti sviluppato da Microsoft all'interno dell’iniziativa. NET, e successivamente approvato come standard dalla ECMA (ECMA-334) e ISO (norma ISO/IEC 23270). Particolarmente adatto per lo sviluppo del progetto in quanto un linguaggio a basso livello che permette l’ottimizzazione del progetto, e compatibile con Unity.

Il problema che si vuole risolvere con questo studio è il riconoscimento dei gesti arbitrali esattamente come ora avviene da parte del tavolo degli ufficiali di campo.
Che hanno l’onere di mettere a referto tutte le decisioni arbitrali che vengono giustappunto comunicate tramite le segnalazioni arbitrali, definite nel regolamento in allegato.
Il fine ultimo od obbiettivo futuro è quello di creare un sistema hardware e software che automatizzi questo procedimento.
I maggiori problemi riscontrati sono stati occlusioni, limitazioni hardware, limitazioni tecniche. Inoltre, è da mettere in evidenza l’impossibilità di accedere al laboratorio per problemi sanitarie nella stesura della parte finale della tesi.
Per la realizzazione dell’software è stato sviluppato dallo studente un visualizzatore degli scheletri estratti tramite open pose, una funzione di identificazione e tracciamento dei diversi scheletri tracciati ed è stata sviluppata una funzione di riconoscimento delle azioni tramite l’algoritmo DTW, ampliamente utilizzato nello stato dell’arte per il riconoscimento delle azioni.
Nell’elaborato verranno illustrati i vari passi compiuti per la realizzazione con la motivazione delle varie scelte fatte e la risoluzione dei problemi presentati.
Verranno infine tratte delle conclusioni sui possibili sviluppi e miglioramenti.

1 Introduzione

La visione artificiale (nota anche come computer vision) è l'insieme dei processi che mirano a creare un modello approssimato del mondo reale tramite l'elaborazione di dati acquisiti di tutto lo spettro elettromagnetico. Il principale scopo della computer vision è quello di interpretazione del contenuto dell'area analizzata e non solo della sua raffigurazione virtuale.
Un sistema di visione artificiale è costituito da un sistema che permette di acquisire, registrare ed elaborare immagini.
Il risultato dell'elaborazione è il riconoscimento di determinate caratteristiche dell'immagine o video per varie finalità di controllo, classificazione, selezione, o altro.
In particolare, questo studio va a riconoscere l'avvenimento di certe azioni in un ambiente specifico.
L'action Recognition è sempre più presente nelle vite di tutti i giorni dal mondo dei videogiochi ai social network, e recentemente anche nelle realtà fisiche come musei e mostre.

1.1 introduzione allo studio

Entrando nel merito tesi, si punta a rilevare gli avvenimenti salienti di una partita analizzando i comportamenti degli arbitri è possibile ricreare interamente la partita giocata senza dover tracciare i singoli giocatori, in questo modo si può ritrovare tutta l'informazione necessaria a ricreare la partita analizzando solo i movimenti dei tre arbitri (considerando una partita delle alte leghe) e non tutti e i dieci giocatori in campo.
Con questa metodologia si punta ad avere una ridondanza di informazioni per validare ulteriori software che invece analizzano i giocatori.
La complessità del riconoscimento però aumenta in quanto le possibili segnalazioni arbitrali, come illustrato nel regolamento, sono oltre 60 con possibili variazioni.
Per lo sviluppo della tesi lo studente ha raggiunto un accordo con la squadra "Aquila Basket Trento" per l'utilizzo di tutto il materiale girato durante un allenamento e una partita amichevole.

1.2 peculiarità del basket

Il basket presenta la situazione perfetta per il test in quanto presenta delle gestualità ripetitive e molto accentuate facilmente rilevabili.
Le segnalazioni che si vogliono prendere in esame sono eseguite dagli arbitri che vogliono comunicare con i giocatori o con il tavolo degli ufficiali di campo. In particolare, ci si punta ad analizzare le comunicazioni riferite al tavolo degli ufficiali in quanto più facili da rilevare perché l'arbitro e tenuto ad avvicinarsi al bordo esterno del campo per effettuare la comunicazione, quindi più vicino alla telecamera.
Questi gesti sono molto frequenti durante una partita.
Inoltre, le dimensioni del campo da basket rientrano nel campo di lavoro della bi-camera utilizzata in quanto essa presenta una distanza di rilevamento delle profondità sui 20 metri e angolo di 100° e la dimensione maggiore dal centro del campo all'esterno non supera tale misura.

2 Stato dell'arte

La pattern recognition è il riconoscimento automatico di pattern e regolarità nei dati. Ha applicazioni nell'analisi dei dati statistici, elaborazione dei segnali, analisi delle immagini, recupero delle informazioni, bioinformatica, compressione dei dati, computer grafica e machine learning. Il riconoscimento dei pattern ha le sue origini nella statistica e nell'ingegneria; alcuni approcci moderni al riconoscimento dei pattern includono l'uso del machine learning, a causa della maggiore disponibilità di big data e di una nuova abbondanza di potenza di elaborazione. Una moderna definizione è:

Il campo della pattern recognition riguarda la scoperta automatica delle regolarità nei dati attraverso l'uso di algoritmi informatici e l'uso di queste regolarità per intraprendere azioni come la classificazione dei dati in diverse categorie.

Gli algoritmi per il riconoscimento dei pattern dipendono dal tipo di output, dal fatto che l'apprendimento sia supervisionato o non supervisionato e dal fatto che l'algoritmo sia di natura statistica o non statistica. Gli algoritmi statistici possono essere ulteriormente classificati come generativi o discriminatori.
Data la grande diversità di algoritmi per il riconoscimento di pattern analizzeremo solo la metodologia utilizzata.
In particolare, appartenente al sottogruppo di algoritmi di etichettatura di sequenze.
Questa è un tipo di operazione di riconoscimento di pattern che comporta l'assegnazione algoritmica di un'etichetta categorica ad ogni membro di una sequenza di valori osservati. Un esempio comune di un'operazione d'etichettatura di sequenza è parte di marcatura di discorso, che cerca di assegnare una parte di discorso ad ogni parola in una frase o in un documento di input. L'etichettatura delle sequenze può essere trattata come un insieme di compiti di classificazione indipendenti, uno per membro della sequenza. Tuttavia, l'accuratezza è generalmente migliorata rendendo l'etichetta ottimale per un dato elemento dipendente dalle scelte degli elementi vicini, utilizzando algoritmi speciali per scegliere il miglior insieme globale di etichette per l'intera sequenza in una sola volta.
L'algoritmo in esame è il Dynamic Time Warping, nell'analisi delle serie temporali, è uno degli algoritmi per misurare la somiglianza tra due sequenze temporali, che possono variare in velocità. Per esempio, le somiglianze nel camminare potrebbero essere rilevate usando il DTW, anche se una persona stesse camminando più velocemente dell'altra, o se ci fossero accelerazioni e decelerazioni durante il corso di un'osservazione. Il DTW è stato applicato alle sequenze temporali dei dati video, audio e grafici - effettivamente, tutti i dati che possono essere trasformati in una sequenza lineare possono essere analizzati con DTW. Un'applicazione ben nota è stata il riconoscimento vocale automatico, per far fronte a diverse velocità di conversazione. Altre applicazioni includono il riconoscimento degli altoparlanti e il riconoscimento della firma online. Può anche essere utilizzato in un'applicazione di corrispondenza parziale della forma.

In generale, DTW è un metodo che calcola una corrispondenza ottimale tra due sequenze date (ad es. serie temporali) con alcune restrizioni e regole:

- Ogni indice della prima sequenza deve essere abbinato ad uno o più indici dell'altra sequenza, e viceversa
- Il primo indice della prima sequenza deve corrispondere al primo indice dell'altra sequenza (ma non deve essere l'unico corrispondente)

- L'ultimo indice della prima sequenza deve corrispondere all'ultimo indice dell'altra sequenza (ma non deve essere l'unico corrispondente)
- La mappatura degli indici dalla prima sequenza agli indici dall'altra sequenza deve aumentare monotonamente, e viceversa.

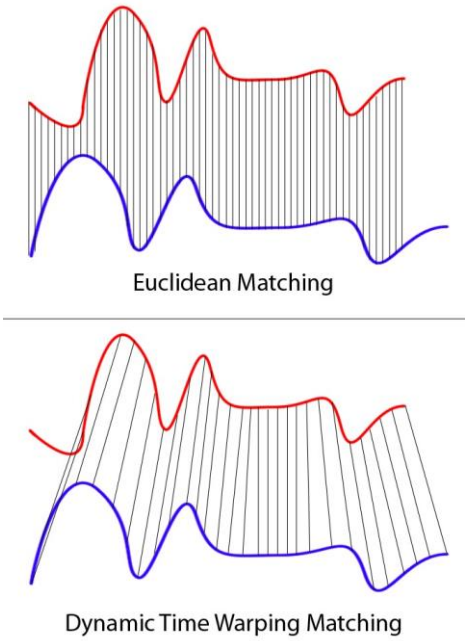


Figura 1 rappresentazione associazione euclidea e DTW

Quindi a differenza di un'associazione euclidea la DTW riesce a trovare la relazione tra sequenze di elementi anche se avvengono con velocità e accelerazioni differenti. Quindi ottimo nel nostro caso in cui dobbiamo rilevare la quando un certo movimento avviene sulla base a dei pattern di riferimento.

3 Metodologia

Il primo passaggio per lo sviluppo del progetto è stato: la richiesta di disponibilità di poter eseguire le rilevazioni alla società Aquila Basket Trento. Con la quale abbiamo determinato oltre alle date ed eventi possibili anche le postazioni possibili da dove possibile fare le rivelazioni.

L'hardware con cui è stata effettuata l'acquisizione era composto da un computer portatile a cui erano connesse due zed-cam e un hard-disk esterno dove venivano immagazzinati i dati prodotti dalle camere.

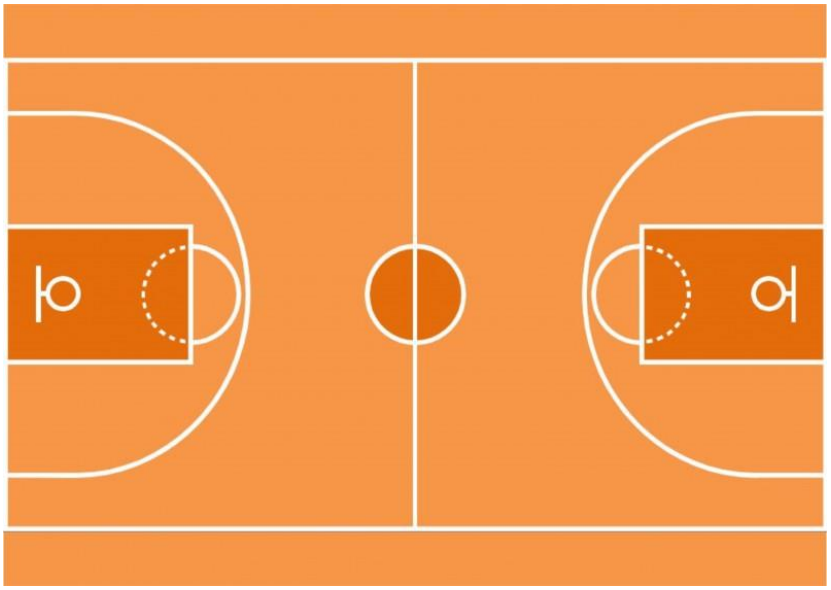


Figura 2 Diagramma del campo da basket.

Le postazioni scelta per la disposizione delle telecamere sono state a due metri fuori al campo corrispettivamente al centro di un lato lungo con le camere posizionate a visualizzare le due metà campo.

3.1 ZED-CAM

La ZED è una telecamera stereo che fornisce alta definizione
immagini e misura accurata della profondità dell'ambiente. È stato progettato
per le applicazioni più impegnative, compreso il controllo autonomo del veicolo,
mappatura mobile, mappatura aerea, sicurezza e sorveglianza. ZED crea una mappa tridimensionale della scena confrontando lo spostamento dei pixel tra le immagini sinistra e destra.

Figura 3 Zed-cam

Con l'hardware disponibile sono state provate differenti configurazioni sia nell'ambiente di indagine che in laboratorio soprattutto per testare e valutare la giusta configurazione.

È stato deciso di utilizzare le zed-cam con risoluzione a 2560x720 in quanto sufficiente per la rilevazione in alta risoluzione dell'evento da analizzare ma soprattutto per mantenere una frame rate di 60 frame per secondo.

Elemento indispensabile per rilevare i movimenti degli arbitri.

Quindi è stato testato in laboratorio sia il salvataggio dei dati per periodi prolungati comparabili con le situazioni d'esame (circa 15 minuti continuativi) alla risoluzione fissata con il salvataggio dei dati di una zed-cam su SSD interno al portatile mentre la seconda sull'HDD esterno collegato.

I risultati erano soddisfacenti e la frame rate non scendeva mai sotto i 50 frame per secondo.

3.2 OpenPose

Openpose rappresenta il primo sistema multi-persona in tempo reale per rilevare punti chiave del corpo umano, della mano, del viso e dei piedi (in totale 135 punti chiave) su singole immagini.

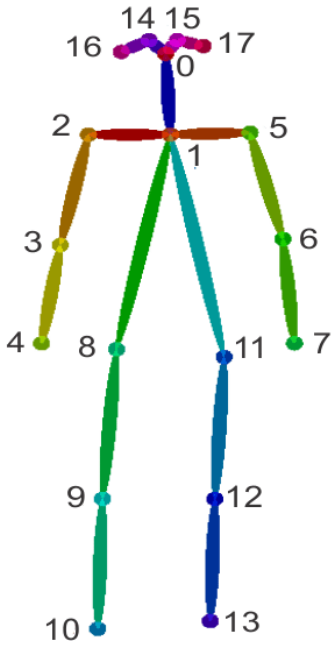


Figura 4 scheletro di punti rilevati da OpenPose

Fondamentale per la finalizzazione del progetto, una particolare configurazione estrapola i file svo delle zed-cam e restituisce un file json dove vengono rappresentati tutti i frame del video con i

relativi punti rilevati di ogni persona.

La parte più complessa del progetto è stata la lettura, raffigurazione e l'isolamento delle persone in quanto si sono presentate molte complicazioni non previste

3.3 Problemi riscontrati

- Il primo problema presentato durante il compimento della tesi si è riscontrato durante l'acquisizione dei dati.
  L'architettura hardware testata in laboratorio non soddisfaceva più i requisii minimi del progetto in quanto la frame rate diminuiva drasticamente al di sotto dei 20 frame per secondo,
  Questo problema era dato dal fatto che la zed-cam salva i dati tramite variazioni di immagini rispetto ai frame precedenti e quindi avendo una ambiente con 13 persone in continuo movimento la variazione è molto alta.

  Quando è stato riscontrato il problema ho limitato l'utilizzo ad una sola zed-cam con scrittura sull'SSD

- La seconda problematica rilevata riguarda l'estrazione dei punti tramite OpenPose. La posizione della camera era a 1,20 metri da terra a soli 2 metri dal campo una posizione ottima per la rilevazione frontale dei movimenti ma che presenta molte difficoltà nell'estrazione degli scheletri in presenza di occlusioni con altre persone cosa che si verifica in più casi di quanti previsti.
- Infine, a seguito della pandemia avvenuta da inizio di quest'anno non è stato possibile accedere al laboratorio ed processare buona parte dei dati per le analisi conclusive, quindi ho dovuto utilizzare i dati delle simulazioni fatte in laboratorio.

3.4 DTW e algoritmo di estrazione dati

Inizialmente i dati sono stati processati e normalizzati per poter essere confrontati.
L'algoritmo di Dynamic Time Warping è stato sviluppato per analizzare il problema in esame.
In particolare il movimento rilevato è l'interruzione per fallo in quanto è una segnalazione arbitrale molto frequente e che implica movimenti molto ampi.

Figura 5 raffigurazione movimento analizzato

L'algoritmo mette a confronto due sequenze di frame. Il primo che chiameremo Q è la sequenza campione che è composta dalla sequenza d'esempio, il caso che contiene l'azione verificata.
Il secondo invece C è la sequenza in cui bisogna verificare la presenza dell'evento.
L'algoritmo come spiegato precedentemente associa ad ogni frame il frame con distanza euclidea minore mantenendo la continuità temporale sono quindi presenti associazioni uno-molti e molti-uno.
Quindi si costruisce una matrice composta dalle distanze euclidee normalizzate di tutti i frame di C con tutti i frame di Q.

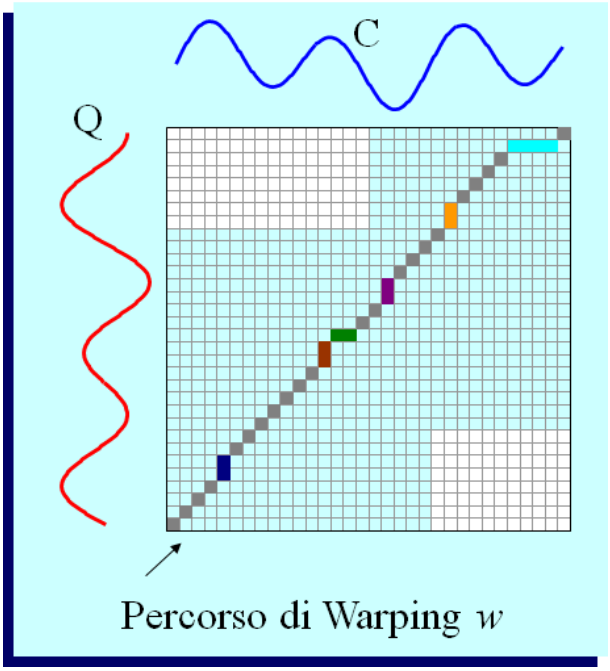


Figura 6 raffigurazione algoritmo DTW

Cominciando dalla posizione [0,0] si somma alla distanza euclidea del primo frame C con il primo frame Q si somma le minori distanze considerando le posizioni [0,1], [1,0],[1,1] e si prosegue fino alla conclusione delle sequenze.
Espresso matematicamente si esprime con la formula:

$$\mathbf{w(i,j) = d(qi,cj) + min\{wp(i-1,j-1) , wp(i-1,j) , wp(i,j-1) \}}$$

la valutazione dell'intero processo e determinato come:

$$\boldsymbol{DTW(Q,C) = min}\left\{\sqrt{\sum_{\boldsymbol{K=1}}^{\boldsymbol{K}} \boldsymbol{w(k)}}\right\}$$

Il calcolo verrà fatto per tutti i punti dello scheletro della figura così da avere una caratterizzazione specifica del movimento.
A questo punto abbiamo tutto ciò che occorre per l'algoritmo, qui riportiamo solo il diagramma complessivo.

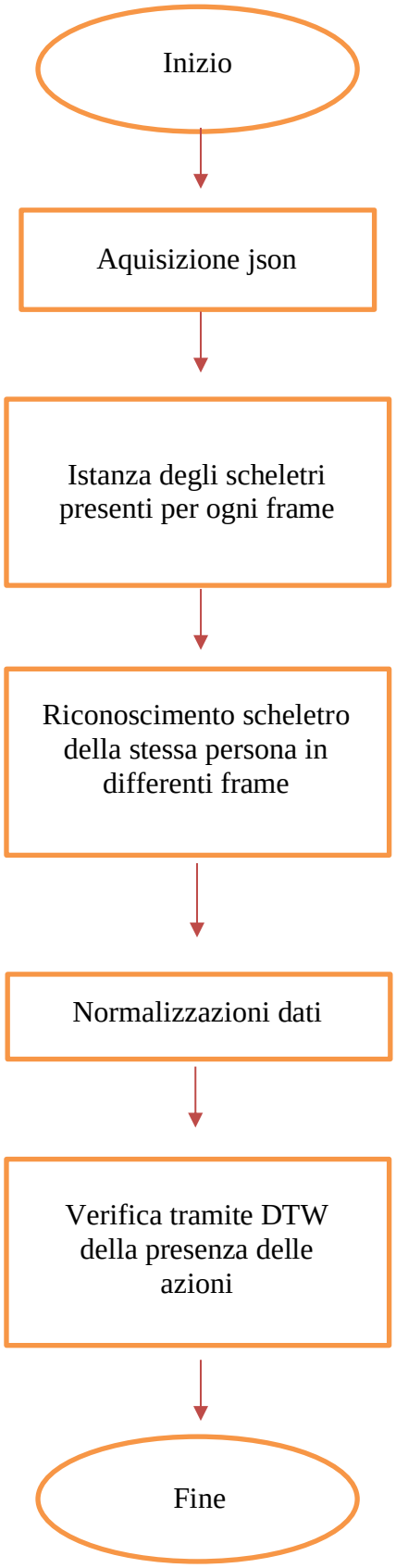


Diagramma 1 flusso del codice

4 Risultati

La finalità del Progetto era sviluppare un prototipo di rilevazione di azioni concentrata nella fattibilità dell'individuazione di alcune particolari segnalazioni fatte dall'arbitro in seguito analizeremo le varie componenti del processo per vedere punti di forza e debolezza.

4.1 Impostazione hardware

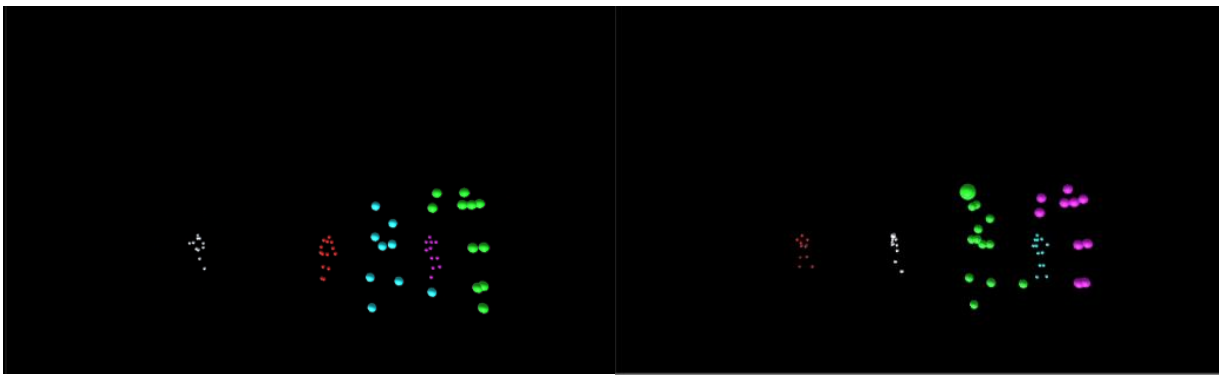
L'impostazione dell'hardware ha dato due importanti implicazioni che hanno impattato sul risultato finale del progetto:
La posizione delle camere non fissa ha presentato delle variazioni della posizine del immagine rilevata nel corso della partita, in quanto essendo posizionate su un tavolo mobile e soggetto a urti e spostamenti dovuti al gioco.
La posizione non perfettamente centrale ha comportato una minor accuratezza delle acquisizioni avvenute nella parte più lontana del campo

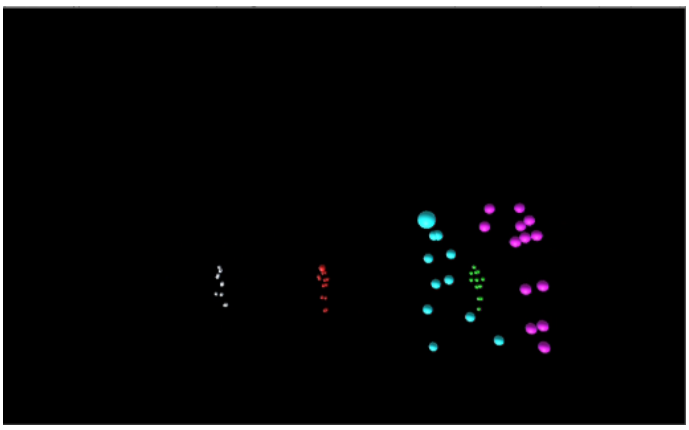
Le zed-cam sono telecamere leggere e non fisse questo ha aumentato lo spostamento continuo prima citato, la rilevazioni di immagini non è stata pienamente accurata come nelle simulazioni in laboratorio enormi cali di frame rate hanno comportato a dei vuoti nella registrazione dell'evento. Una posizione più rialzata avrebbe evitato molte occlusioni semplificando anche la qualità.

Un singolo portatile non è stato sufficiente per l'utilizzo simultaneo di due camere con risoluzione 2560x720 in una situazione così dinamica.

4.2 OpenPose

OpenPose è risultato abbastanza accurato sia i in situazioni statiche che in situazioni di movimento. Alcuni punti vengono difficilmente rilevati come il punto 0(vedi immagine) ma i punti delle braccia, interessati dal punto di vista dell'analisi sono sempre rilevati.
Con l'aumento delle persone nel campo analizzato si presentano delle difficoltà a riconoscere le sagome oltre ai 7 metri di distanza dato il fondale con molto rumore e non uniforme.

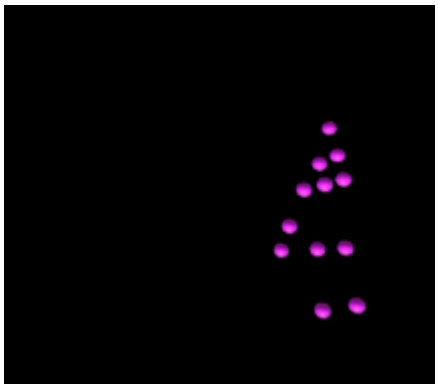




Video 1 frame esplicativi output OpenPose

Nelle immagini precedenti si susseguono 3 frame di output di OpenPose e come è facile notare in ogni immagine il colore delle persone raffigurate cambia, il colore è raffigurante dell'Id che OpenPose attribuisce a ogni scheletro. Questo comporta dei dati difficili da analizzare ed è stato necessario sviluppare un algoritmo che riconosca e mantenga gli id.
Inoltre, a complicare la situazione è stato il fatto che molti scheletri mancano di alcuni frame complicando ulteriormente il processo di gestione dati.

Infine, avendo separato adeguatamente gli id un frame della sequenza risultante presa come riferimento per l'algoritmo di riconoscimento è la seguente



Video 2 raffigurazione del caso da rilevare dopo essere estratto e normalizzato

I confronti sono stati fatti con casi simili registrati in laboratorio che presentavano il medesimo movimento. La soglia di confronto però perché il movimento venga riconosciuto risulta molto alta il che comporta moti falsi positivi. Non avendo ulteriori dati su cui testare l'algoritmo non si è potuto procedere.

5 Conclusioni

L'obbiettivo principale del progetto era quello di ottenere un software in grado di riconoscere l'avvenimento di certe azioni da parte degli arbitri di basket.
A seguito dei vari problemi riscontrati e illustrati nei vari punti posso affermare che la il fine ultimo del progetto non è riuscito.
Il sistema ha riscontrato varie problematiche sia software che hardware e molte dovute all'inesperienza e alla complessità dei problemi presentati tra i quali la possibilità di non poter fare una seconda rilevazione dati, che avrebbe risolto i problemi di movimento delle camere e una frame rate stabile per tutte le riprese.
In seguito, la posizione della camera non copriva gran parte dell'area di gioco questo dovuto alle disposizioni della società e da limitazioni tecniche.
In ultimo grave crisi medica che non ha permesso l'utilizzo del laboratorio escludendo l'utilizzo di alcune parti di riprese.

A seguito di questo però ritengo ancora possibile un'attuazione del sistema prendendo i dovuti accorgimenti e facendo le dovute correzioni.
In particolare, applicare l'algoritmo DTW non a tutti i punti rilevati dal movimento ma in particolare creare delle metriche solo per gli arti coinvolti nel movimento in analisi dando dei pesi diversificati ad ogni punto dello scheletro così da accentuare il peso degli arti coinvolti.
Un ulteriore possibile approccio al problema potrebbe utilizzare una rete neurale per il riconoscimento delle azioni.

5.1 Considerazioni personali sul progetto

Il progetto è stato stimolante e molto innovativo il che spiegherebbe le innumerevoli problematiche presentate.
La complessità dello sviluppo software era elevata rispetto alle capacità. In questo modo ho potuto sviluppare le mie conoscenze hardware e software soprattutto per quanto riguarda la programmazione date le oltre 3000 righe di codice scritte.

Bibliografia

- Liliana Lo Presti, Marco La Cascia 3D skeleton-based human action classification: A survey, December 2015
- Shih-En Wei and Varun Ramakrishna and Takeo Kanade and Yaser Sheikh, CVPR, Convolutional pose machines, 2016
- Paschalis Panteleris, Iason Oikonomidis, Antonis Argyros, Using a single RGB frame for real time 3D hand pose estimation in the wild, Dic 2017
- *Tomas Simon, Hanbyul Joo, Iain Matthews, Yaser Sheikh*; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Hand Keypoint Detection in Single Images using Multiview Bootstrapping, CVPR, 2017
- Tomas Simon and Hanbyul Joo and Iain Matthews and Yaser Sheikh, CVPR, Hand Keypoint Detection in Single Images using Multiview Bootstrapping, 2017
- Zhe Cao and Tomas Simon and Shih-En Wei and Yaser Sheikh, CVPR, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, 2017
- Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, Xiaokang Yang, Fine-grained Video Captioning for Sports Narrative Shanghai Institute for Advanced Communication and Data Science, Shanghai Key Laboratory of Digital Media Processing and Transmission, Shanghai Jiao Tong University, 2018
- Federico Angelini, Student Member, IEEE, Zeyu Fu, Student Member, IEEE, Yang Long, Senior Member, IEEE, Ling Shao, Senior Member, IEEE, and Syed Mohsen Naqvi, Senior Member, IEEEActionXPose: A Novel 2D Multi-view Pose-based Algorithm for Real-time Human Action Recognition, 2018
- Angjoo Kanazawa, Michael J. Black, David W. Jacobs, Jitendra Malik University of California, Berkeley MPI for Intelligent Systems, Tubingen, Germany, University of Maryland, End-to-end Recovery of Human Shape and Pose, 2018
- Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, Yebin Liu, DoubleFusion: Real-time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor, 2018
- Rıza Alp Guler, Natalia Neverova, Iasonas Kokkinos, DensePose: Dense Human Pose Estimation In The Wild, 2018
- Z. Cao and G. Hidalgo Martinez and T. Simon and S. Wei and Y. A. Sheikh, IEEE Transactions on Pattern Analysis and Machine Intelligence, OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, 2019