UNIVERSITÀ DEGLI STUDI DI FIRENZE
Scuola di Ingegneria - Dipartimento di Ingegneria dell'Informazione

Tesi di laurea in Ingegneria Informatica

Riconoscimento di azioni umane usando tecniche di apprendimento profondo per la stima della posa

Candidato
Andrea Moscatelli

Relatori
Marco Bertini

Secondo Supervisore

Correlatore
Correlatore 1

Anno Accademico 2019 - 2020

Indice

Prefazione

prefazione

Introduzione

Introduzione

Capitolo 1

Stima della posa

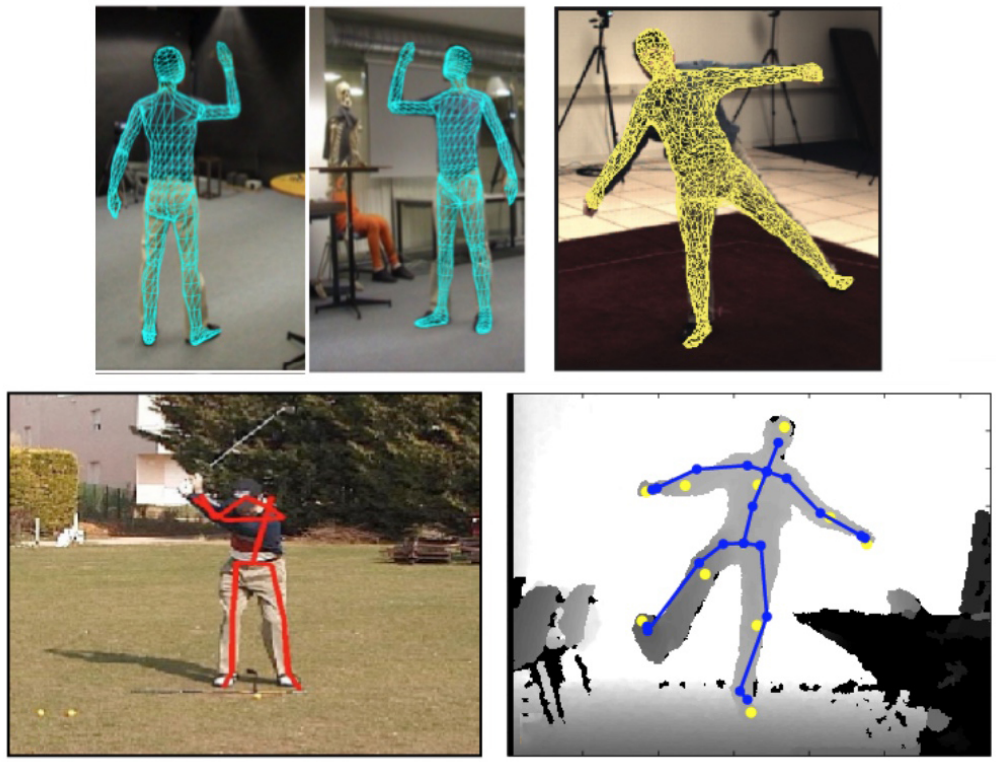

Figura 1.1. Esempi di stima della posa. In alto tre esempi di stima della posa utilizzando modelli di tipo volumetrico. In basso due esempi di stima della posa ottenuti utilizzando modelli di tipo scheletrici.

Cos è la stima della posa?
Quando parliamo di *stima della posa* ci riferiamo ad una tecnica di *computer vision* dedita al riconoscimento di figure umane all'interno di video ed immagini, così da poter riconosce ad esempio dove, all'interno dell'immagine, si trova la testa, il braccio, la gamba destra, etc.. del soggetto inquadrato.
Questa tecnica non va assolutamente confusa con tecniche di riconoscimento di persone, infatti la stima della posa è in grado solo di riconoscere dove le parti del corpo di un individuo sono situate all'interno dell'immagine, non *chi* è inquadrato.

I campi di applicazione della stima della posa sono i più svariati: software interattivi che reagiscono al movimento della posa, robotica, realtà aumentata, animazione, fotoritocco intelligente, fitness, riabilitazione, etc. Stiamo parlando di un problema tutt'altro che semplice, infatti la condizione di luce dell'immagine, la variabilità dell'ambiente circostante, l'inclinazione del soggetto inquadrato, rendono il riconoscimento della posa un problema non affatto banale.

Spinti dal crescente interesse, negli ultimi anni sono stati sviluppati diversi algoritmi per la stima della posa, raggiungendo in molti casi risultati davvero sorprendenti con un'accuratezza prossima alla perfezione.

Figura 1.2. Un esempio di utilizzo in campo medico della stima della posa

La maggior parte dei software in circolazione in grado di stimare in maniera sufficientemente corretta la posa di un individuo non sono liberamente accessibili.

Due fra i migliori algoritmi (ad oggi) di *pose detection* sono sicuramente *Posenet* [1] e *Detectron 2* [4], dei quali ci occuperemo in maniera più approfondita nei capitoli seguenti.

Capitolo 2

PoseNet

I recenti progressi nel campo della visione artificiale hanno permesso alla comunità scientifica di spostarsi verso problemi ancora più articolati rispetto a quelli classici, come ad esempio il riconoscimento facciale, con l'obiettivo di riconoscere figure umane in contesti non vincolati e molto variabili.

L'algoritmo *PoseNet* è stato ideato proprio con lo scopo di identificare una o più figure umane in qualsiasi contesto, anche in contesti "affollati", ed essere in grado di identificare l'istanza di ogni persona stimandone i suoi *punti chiave* (o *key-points*).

Esistono due approcci principali per affrontare il rilevamento di più persone, la stima della posa e la segmentazione. L'approccio *top-down* inizia identificando e localizzando approssimativamente le singole istanze di persona identificando il riquadro dell'immagine dentro le quali sono contenute, seguito da una fase di stima della posa o di separazione "primo piano-sfondo" nell'area identificata. Al contrario, l'approccio *bottom-up* inizia localizzando entità semantiche individuali, come ad esempio gambe, braccia, mani, etc, seguito dal loro raggruppamento in istanze di persone complete. PoseNet adotta questo secondo approccio.

In particolare PoseNet utilizza una rete neurale convoluzionale nella quale il costo computazionale del riconoscimento delle pose è essenzialmente indipendente dal numero di persone raffigurate nella scena ma dipende esclusivamente dalla scelta delle features della rete.

L'approccio adottato in PoseNet è quello di identificare dapprima tutti i punti

chiave di ogni persona nell'immagine e successivamente raggrupparli in istanze utilizzando un processo "greedy", ovvero partendo dal rilevamento "più sicuro", e non come spesso accade da un punto fisso di riferimento (ad esempio il naso), avendo come vantaggio quello di funzionare bene anche se in disordine.

Oltre a stimare punti chiave sparsi, PoseNet stima anche maschere di segmentazione per ogni persona. Per fare ciò, viene allenata una seconda rete neurale con la quale viene associato ad ogni pixel x_i dell'immagine la probabilità di appartenenza di quel pixel ad ogni candidato j identificato. Se la probabilità è sufficientemente alta allora viene associato il pixel x_i al candidato j.

Questo algoritmo è stato allenato utilizzando il dataset COCO [2] che annota molte persone con 17 punti chiave (12 del corpo e 5 del volto), migliorando l'*AP* (average-precision) dal precedente miglior risultato da 0,655 a 0,687.

Questo metodo essendo molto semplice è anche quindi molto rapido, poiché non richiede alcuna fase supplementare di raffinamento dei risultati con tecniche di tipo *box-based* o *clustering*, facendo di PoseNet uno degli algoritmi più facilmente installabili su rete mobile.

2.1 Stima dei key-points

L'obiettivo di questa fase è quello di rilevare, in modo indipendente dall'istanza, tutti i key-points visibili appartenenti a qualsiasi persona dell'immagine. A tale scopo vengono prodotte delle *heatmaps*, ovvero dei canali della rete neurale dediti al riconoscimento di particolari caratteristiche dell'immagine (una canale per ogni key-point) e degli *offset* (due canali per ogni key-point per gli spostamenti in orizzontale e verticale). Sia x_i la posizione 2-D nell'immagine, dove $i = 1, ..., N$ e N è il numero di pixels; $D_R(y) = \{x : \|x - y\| \leq R\}$ un disco di raggio R centrato in y e $y_{j,k}$ la posizione 2-D del k-esimo key-point della j-esima istanza di persona, con $j = 1, ..., M$, dove M è il numero di istanze nell'immagine.

Per ogni tipo di key-point $k = 1, ..., K$, viene impostato un task di classificazione binaria come segue. Viene generata una heatmap $p_k(x)$ tale che $p_k(x) = 1$ se $x \in D_R(y_{j,k})$ per qualsiasi istanza j, altrimenti $p_k(x) = 0$. Abbiamo quindi K tasks di classificazione binaria indipendenti, una per ogni tipo di key-point.

Ciascuno equivale a prevedere un disco di raggio R attorno a un tipo di key-point specifico di qualsiasi persona nell'immagine.

Oltre alle heatmaps, vengono anche usati vettori di offset a *corto raggio* $S_k(x)$ il cui scopo è quello di migliorare l'accuratezza della localizzazione dei key-points. Per ogni punto x all'interno dei dischi ricavati al passo precedente, il vettore di offset 2-D a corto raggio $S_k(x) = y_{j,k} - x$ rappresenta la distanza fra il punto x e il k-esimo key-point della j-esima persona. Vengono cosi generati K vettori per ogni punto x all'interno del disco definito che, combinati insieme, miglioreranno l'accuratezza della posizione predetta per ogni key-point.

2.2 Raggruppamento dei key-points in istanze di persona

A questo è necessario però capire come associare ogni key-point stimato ad ogni persona nell'immagine (nel caso ce ne sia più di una).

Capitolo 3

Detectron2

Descrizione di cosè Detectron2

Capitolo 4

Classificazione

4.1 Struttura della rete

4.2 Tecniche

4.2.1 Semplice

4.2.2 Tecnica dei centri

4.2.3 Tecnica delle differenze

Capitolo 5

Risultati ottenuti

Capitolo 6

Conclusioni

Capitolo 7

Sviluppi futuri

Bibliografia

[1] *"PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model"* - George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, Kevin Murphy - 2018

[2] Lin, T.Y., Cui, Y., Patterson, G., Ronchi, M.R., Bourdev, L., Girshick, R., Dollr,P. - Coco 2016 keypoint challenge. - 2016

[3] PoseNet with TensorFlow.js - *https://medium.com/tensorflow/real-time-human-pose-estimation-in-the-browser-with-tensorflow-js-7dd0bc881cd5*

[4] Detectron2 - *https://research.fb.com/wp-content/uploads/2019/12/4.-detectron2.pdf*