

Scuola di Scienze Matematiche, Fisiche e Naturali
Corso di Laurea Magistrale in Informatica
Tesi di Laurea

**RICONOSCIMENTO DI AZIONI UMANE
USANDO TECNICHE DI
APPRENDIMENTO PROFONDO
PER LA STIMA DELLA POSA**

Andrea Moscatelli

Relatore: *Marco Bertini*

Anno accademico 2018-2019



UNIVERSITÀ
DEGLI STUDI
FIRENZE

Outline



UNIVERSITÀ
DEGLI STUDI
FIRENZE

- Deep learning
- posa umana
- argomento 3



Concetti chiave



UNIVERSITÀ
DEGLI STUDI
FIRENZE

Riconoscimento di
azioni umane

=

stima della
posa

+

apprendimento
profondo



Concetti chiave



UNIVERSITÀ
DEGLI STUDI
FIRENZE

Riconoscimento di
azioni umane

=

stima della
posa

+

apprendimento
profondo

Saper riconoscere potenzialmente qualsiasi azione umana, sia **individuale** che **di gruppo**, ripresa in video:

- *bere*
- *mangiare*
- *scrivere*
- *telefonare*
- *mettersi le mani in tasca*
- *indicare qualcuno*
- *darsi la mano*
- *picchiarsi*
- *abbracciarsi*
- ...

Concetti chiave



UNIVERSITÀ
DEGLI STUDI
FIRENZE

Riconoscimento di
azioni umane

=

**stima della
posa**

+

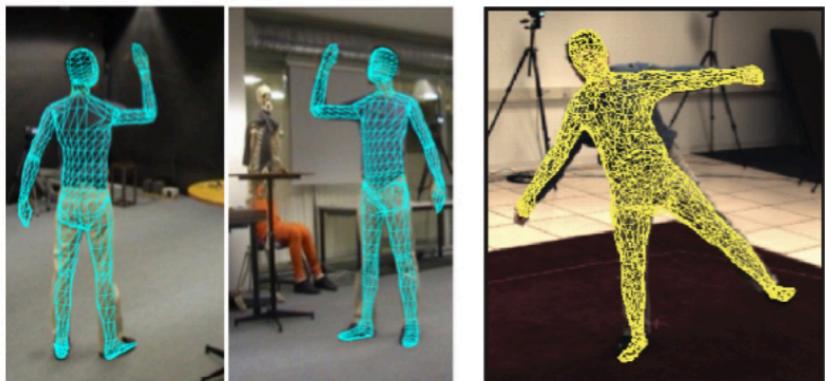
**apprendimento
profondo**

Definizione di **posa** (in computer vision):

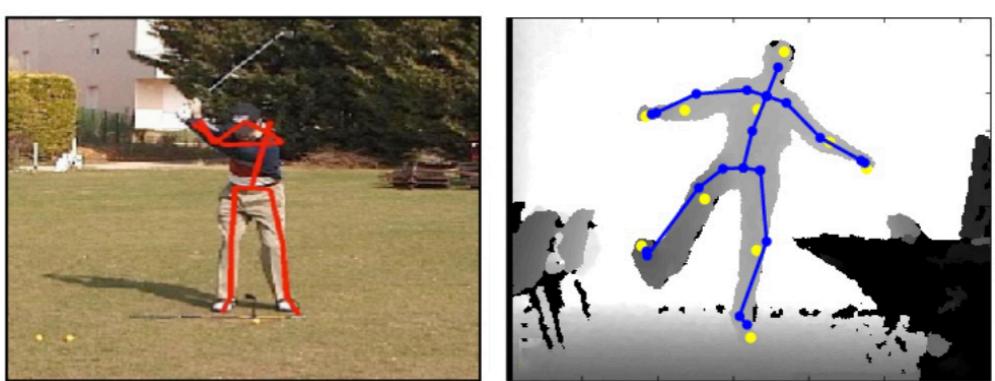
"Combinazione di posizione e orientamento di un oggetto"

La posa umana (o di oggetti) può essere di due tipi:

Volumetrica



Scheletrica



Composta da molti punti e orientata
alla tridimensionalità del soggetto

Composta da pochi punti e orientata ad
una schematizzazione efficace del soggetto

Concetti chiave



UNIVERSITÀ
DEGLI STUDI
FIRENZE

Riconoscimento di
azioni umane

=

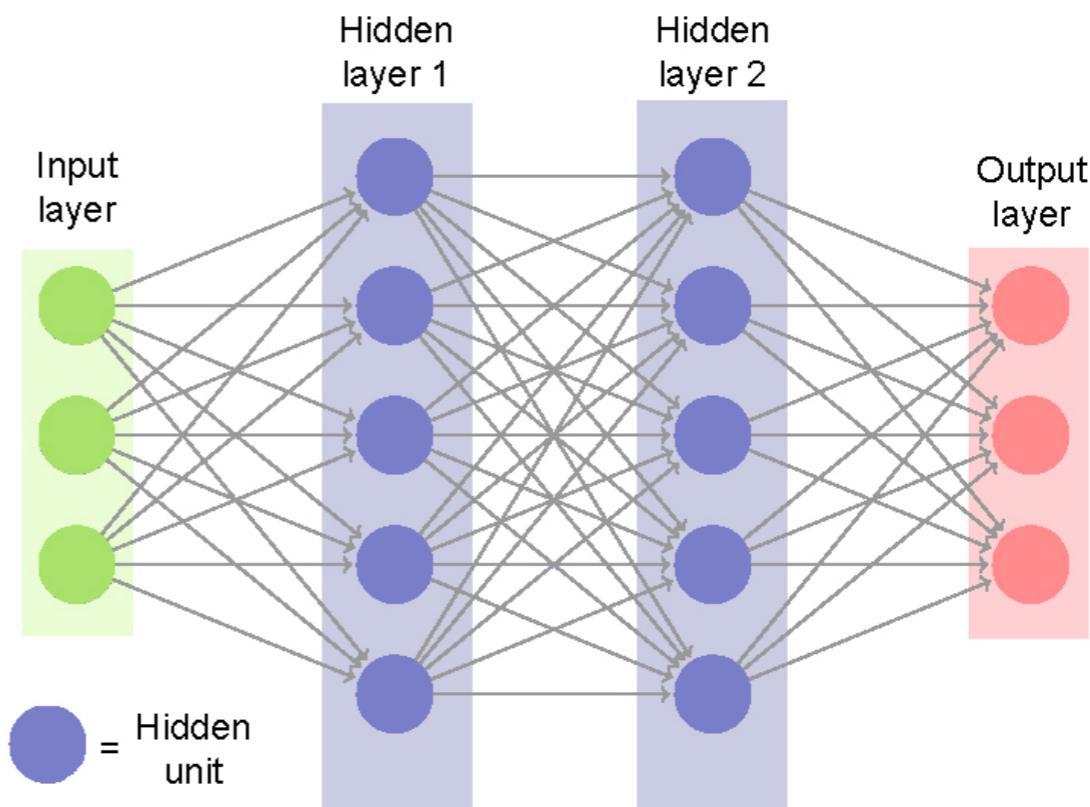
stima della
posa

+

apprendimento
profondo

L'**apprendimento profondo** (o **Deep learning**) è quella branca dell'apprendimento automatico (o *Machine learning*) basata su diversi **livelli** di rappresentazione dove i valori di alto livello sono definiti sulla base di quelli di basso.

Lo strumento utilizzato è la **rete neurale artificiale**, organizzata in diversi strati ognuno dei quali calcola i valori per quello successivo affinché l'informazione venga elaborata in maniera sempre più completa.



Concetti chiave



UNIVERSITÀ
DEGLI STUDI
FIRENZE

Riconoscimento di
azioni umane

=

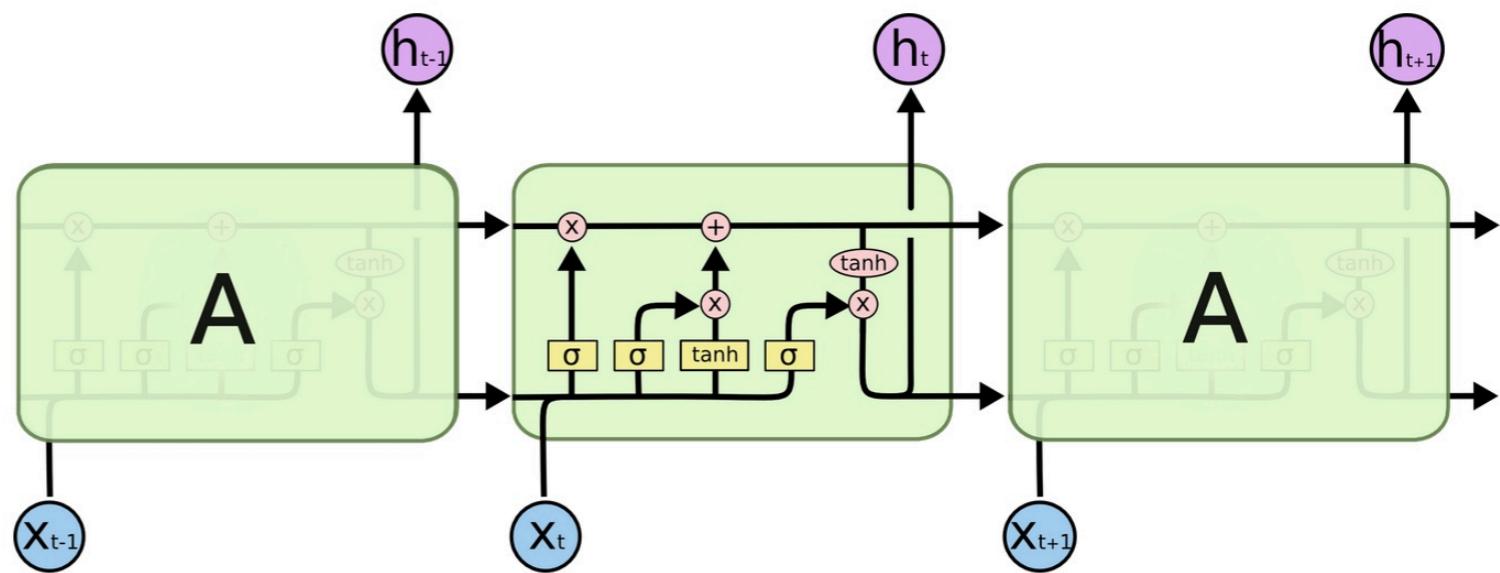
stima della
posa

+

apprendimento
profondo

L'**apprendimento profondo** (o **Deep learning**) è quella branca dell'apprendimento automatico (o *Machine learning*) basata su diversi **livelli** di rappresentazione dove i valori di alto livello sono definiti sulla base di quelli di basso.

In questo lavoro di tesi la rete neurale usata è una combinazione di livelli **Long Short-Term Memory (LSTM)**, della famiglia delle *reti neurali ricorrenti*, particolarmente adatte alla classificazione di sequenze temporali



La stima della posa



UNIVERSITÀ
DEGLI STUDI
FIRENZE

Detectron2

modello : *Mask R-CNN*
backbone : *ResNeXt+FPN*
pre-allenato su : *ImageNet*

- Accuratezza allo stato dell'arte

PoseNet

modello : *PersonLab*
backbone : *MobileNetV1*
pre-allenato su : *COCO-2016*

- Leggerezza modello
- Rapidità d'inferenza

La stima della posa



UNIVERSITÀ
DEGLI STUDI
FIRENZE



Detectron2

modello : *Mask R-CNN*
backbone : *ResNeXt+FPN*
pre-allenato su : *ImageNet*



PoseNet

modello : *PersonLab*
backbone : *MobileNetV1*
pre-allenato su : *COCO-2016*



La stima della posa

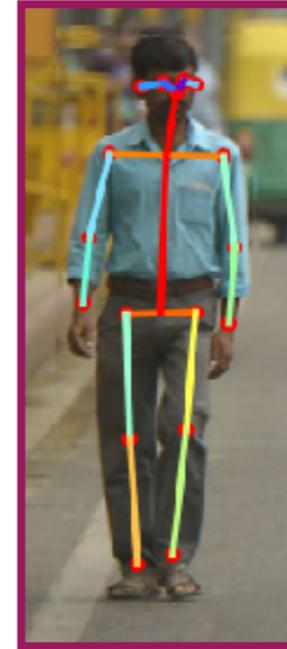


UNIVERSITÀ
DEGLI STUDI
FIRENZE



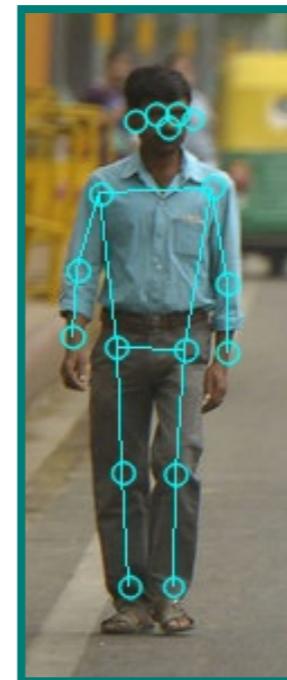
Detectron2

modello : *Mask R-CNN*
backbone : *ResNeXt+FPN*
pre-allenato su : *ImageNet*



PoseNet

modello : *PersonLab*
backbone : *MobileNetV1*
pre-allenato su : *COCO-2016*



- **17 punti** per persona
- **score** per persona

Dataset utilizzato



UNIVERSITÀ
DEGLI STUDI
FIRENZE

NTU-RGB+D

- anno **2016**
- **56880** video
 - RGB
 - frame di profondità
 - pose
 - frame ad infrarossi
- **3** angolazioni di ripresa
- **60** azioni
 - 50 individuali
 - 10 di coppia
- **40** attori (10-35 anni)
- **2** ripetizioni

Dataset utilizzato



UNIVERSITÀ
DEGLI STUDI
FIRENZE

NTU-RGB+D

- anno **2016**
- **56880** video
 - RGB
 - ~~- frame di profondità~~
 - ~~- pose~~
 - ~~- frame ad infrarossi~~
- **3** angolazioni di ripresa
- **60** azioni
 - 50 individuali
 - 10 di coppia
- **40** attori (10-35 anni)
- **2** ripetizioni

Dataset utilizzato

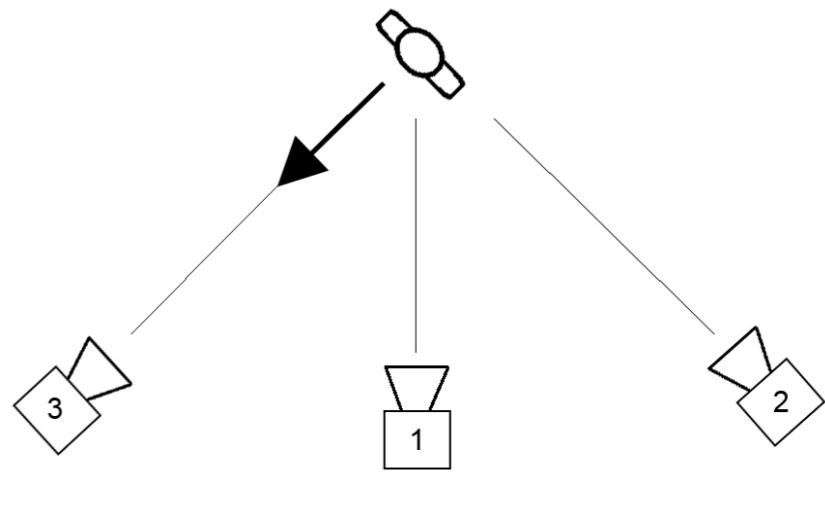


UNIVERSITÀ
DEGLI STUDI
FIRENZE

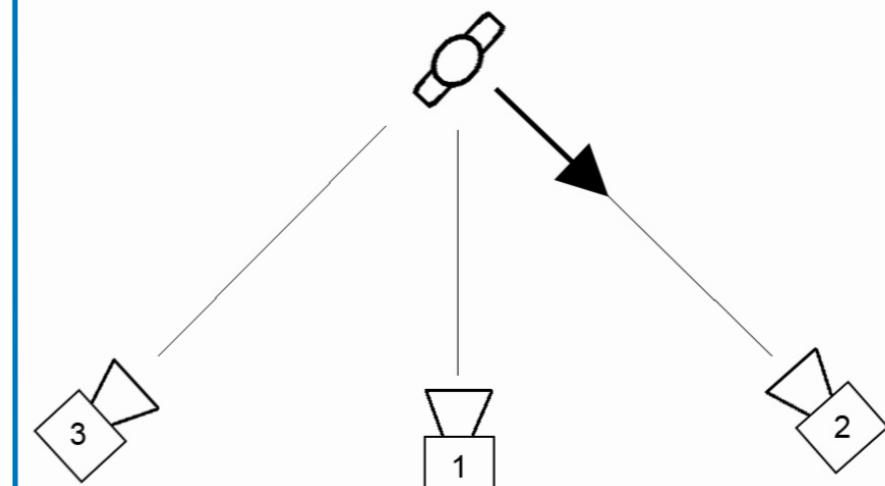
NTU-RGB+D

- anno 2016
- **56880** video
 - RGB
 - ~~frame di profondità~~
 - ~~pose~~
 - ~~frame ad infrarossi~~
- **3** angolazioni di ripresa
- **60** azioni
 - 50 individuali
 - 10 di coppia
- **40** attori (10-35 anni)
- **2** ripetizioni

Ripetizione 1



Ripetizione 2



Dataset utilizzato



UNIVERSITÀ
DEGLI STUDI
FIRENZE

NTU-RGB+D

- anno 2016
- **56880** video
 - RGB
 - ~~frame di profondità~~
 - ~~pose~~
 - ~~frame ad infrarossi~~
- **3** angolazioni di ripresa
- **60** azioni
 - 50 individuali
 - 10 di coppia
- **40** attori (10-35 anni)
- **2** ripetizioni

Cross View

Training

id camera: 2,3
video: 37920

Test

id camera: 1
video: 18960

Dataset utilizzato



UNIVERSITÀ
DEGLI STUDI
FIRENZE

NTU-RGB+D

- anno 2016
- **56880** video
 - RGB
 - frame di profondità
 - pose
 - frame ad infrarossi
- **3** angolazioni di ripresa
- **60** azioni
 - 50 individuali
 - 10 di coppia
- **40** attori (10-35 anni)
- **2** ripetizioni

Cross View

Training

id camera: 2,3
video: 37920

Test

id camera: 1
video: 18960

contiene tutte
le riprese
frontali e a **90°**

Dataset utilizzato



UNIVERSITÀ
DEGLI STUDI
FIRENZE

NTU-RGB+D

- anno 2016
- **56880** video
 - RGB
 - ~~frame di profondità~~
 - ~~pose~~
 - ~~frame ad infrarossi~~
- **3** angolazioni di ripresa
- **60** azioni
 - 50 individuali
 - 10 di coppia
- **40** attori (10-35 anni)
- **2** ripetizioni

Cross View

Training

id camera: 2,3
video: 37920

Test

id camera: 1
video: 18960

contiene tutte le riprese **frontali** e a **90°**

contiene tutte le riprese a **45°**

Dataset utilizzato



UNIVERSITÀ
DEGLI STUDI
FIRENZE

NTU-RGB+D

- anno 2016
- **56880** video
 - RGB
 - ~~frame di profondità~~
 - ~~pose~~
 - ~~frame ad infrarossi~~
- **3** angolazioni di ripresa
- **60** azioni
 - 50 individuali
 - 10 di coppia
- **40** attori (10-35 anni)
- **2** ripetizioni

Cross View

Training

id camera: 2,3
video: 37920

Test

id camera: 1
video: 18960

Cross Subject

Training

id attori: 1, 2, 4, 5, 8, 9,
13, 14, 15, 16, 17, 18, 19,
25, 27, 28, 31, 34, 35, 38
video: 40320

Test

id attori: tutti gli
altri

Deep learning

- reti neurali (funzionamento, GD, loss function,...)
- reti ricorrenti
- reti LSTM
- struttura rete lavoro di tesi

questa slide falla per ultima, guarda cosa devi spiegare di una rete per introdurre quello che dirai dopo

Preprocessing

- assegnazione coerente delle pose
- rimozione degli zeri

Tecniche di rielaborazione

- semplice
- next frame
- baricentri
- distanze relative
- distanze cumulate
- normalizzazione

tecnica semplice

tecnica dei baricentri

NEXT frame

distanze cumulate

distanze relative

fase 1 struttura della rete: prima esplorazione tecniche

- ottimizzatore
- criterio d'arresto
- dataset validazione
- ...

fase 2 struttura della rete: esplorazione livelli tecniche migliori

- ottimizzatore
- criterio d'arresto
- dataset validazione
- ...

fase 3 struttura della rete: esplorazione regolarizzatore e dropout

- ottimizzatore
- criterio d'arresto
- dataset validazione
- ...

fase 4 struttura della rete: addestramento tecniche migliori

- ottimizzatore
- criterio d'arresto
- dataset validazione
- ...

fase 5 struttura della rete: combinazione tecniche migliori

- ottimizzatore
- criterio d'arresto
- dataset validazione
- ...

**risultati finali comparati
allo stato dell'arte**

applicazione algoritmo a video

conclusioni e sviluppi futuri

Grazie!