

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

Concetti chiave



UNIVERSITÀ
DEGLI STUDI
FIRENZE

Riconoscimento di **azioni umane** = **stima della posa** + **apprendimento profondo**



Concetti chiave



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



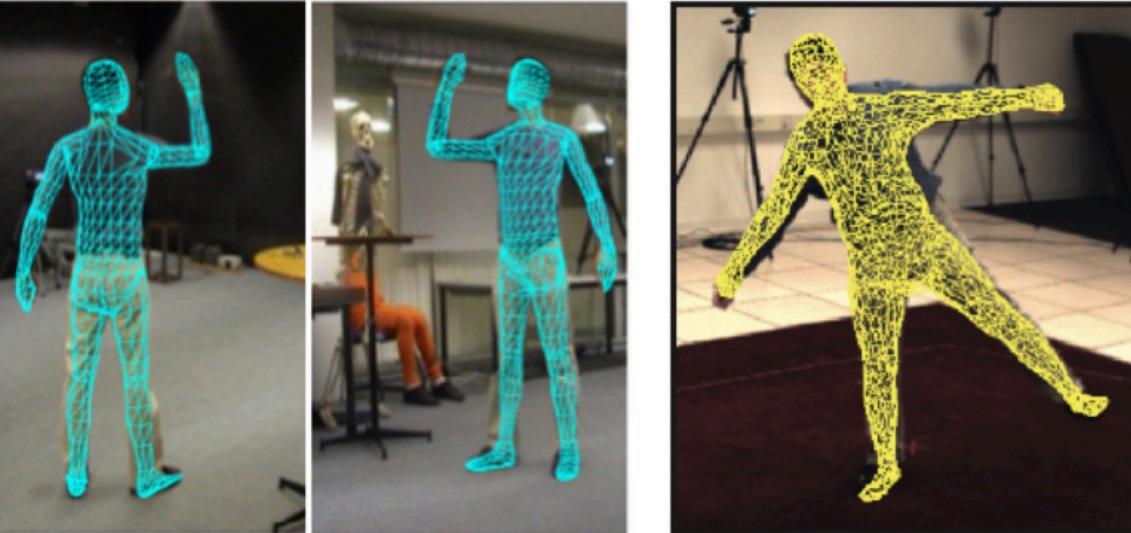
Riconoscimento di **azioni umane** = **stima della posa** + **apprendimento profondo**

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

"Combinazione di posizione ed orientamento di un oggetto"

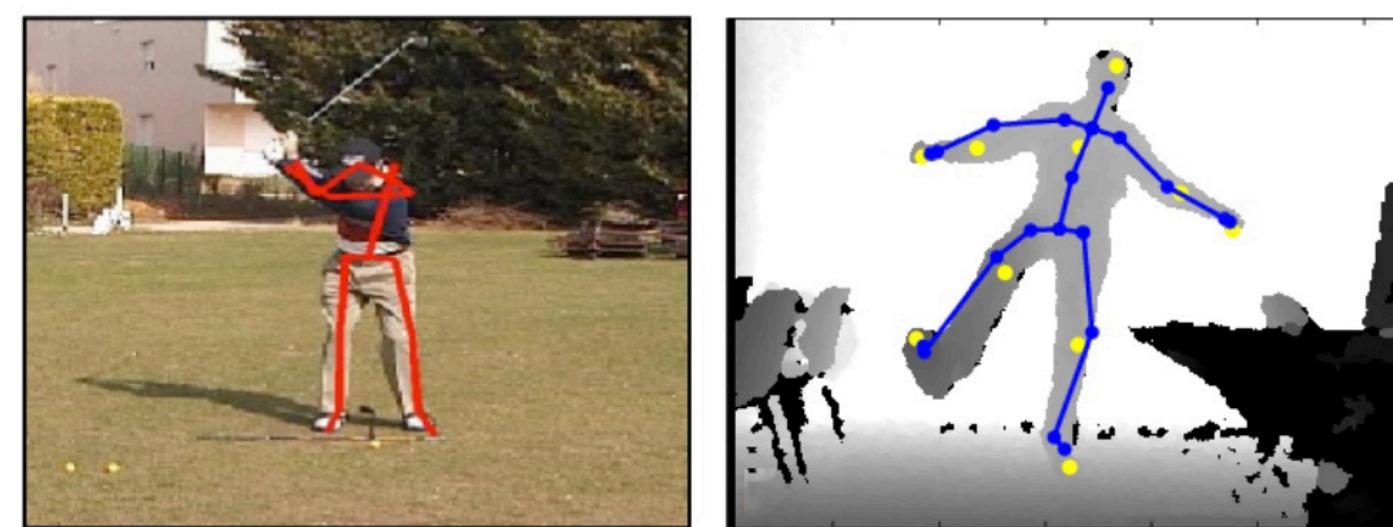
La posa può essere di due tipi:

Volumetrica



Composta da molti punti ed orientata alla tridimensionalità del soggetto

Scheletrica



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

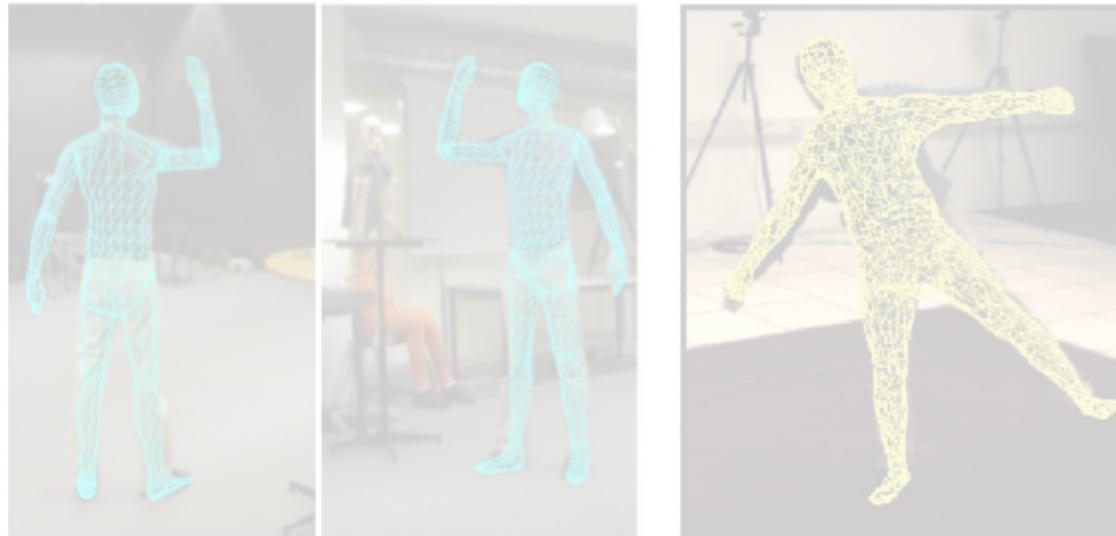
Concetti chiave

Riconoscimento di **azioni umane** = **stima della posa** + **apprendimento profondo**

Definizione di **posa** (in computer vision):
"Combinazione di posizione ed orientamento di un oggetto"

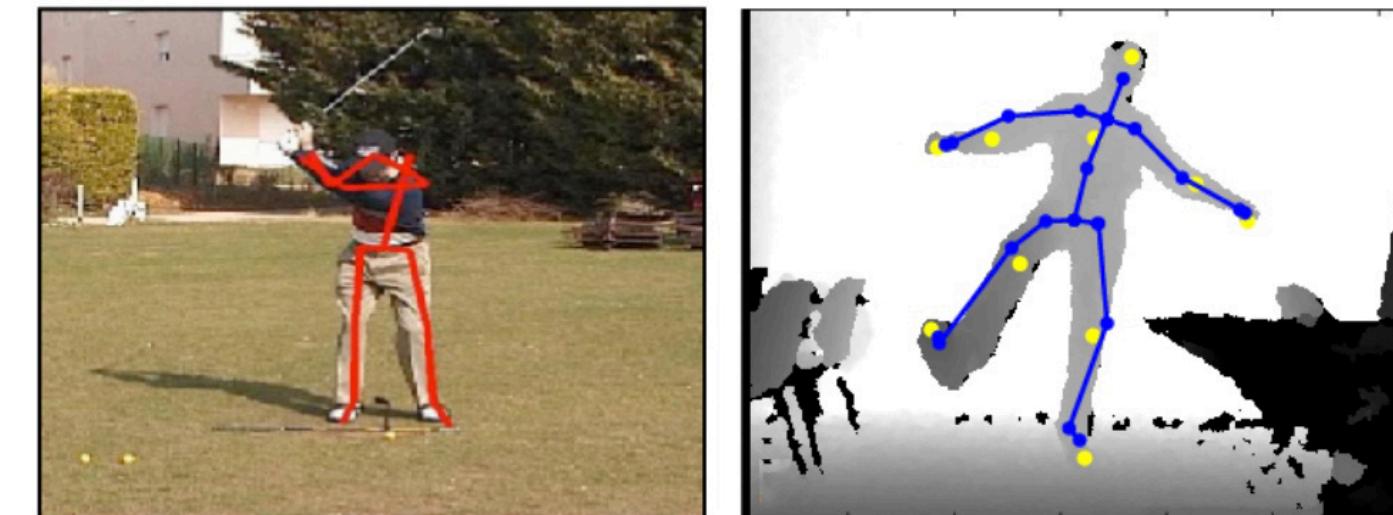
La posa può essere di due tipi:

Volumetrica



Composta da molti punti ed orientata
alla tridimensionalità del soggetto

Scheletrica



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

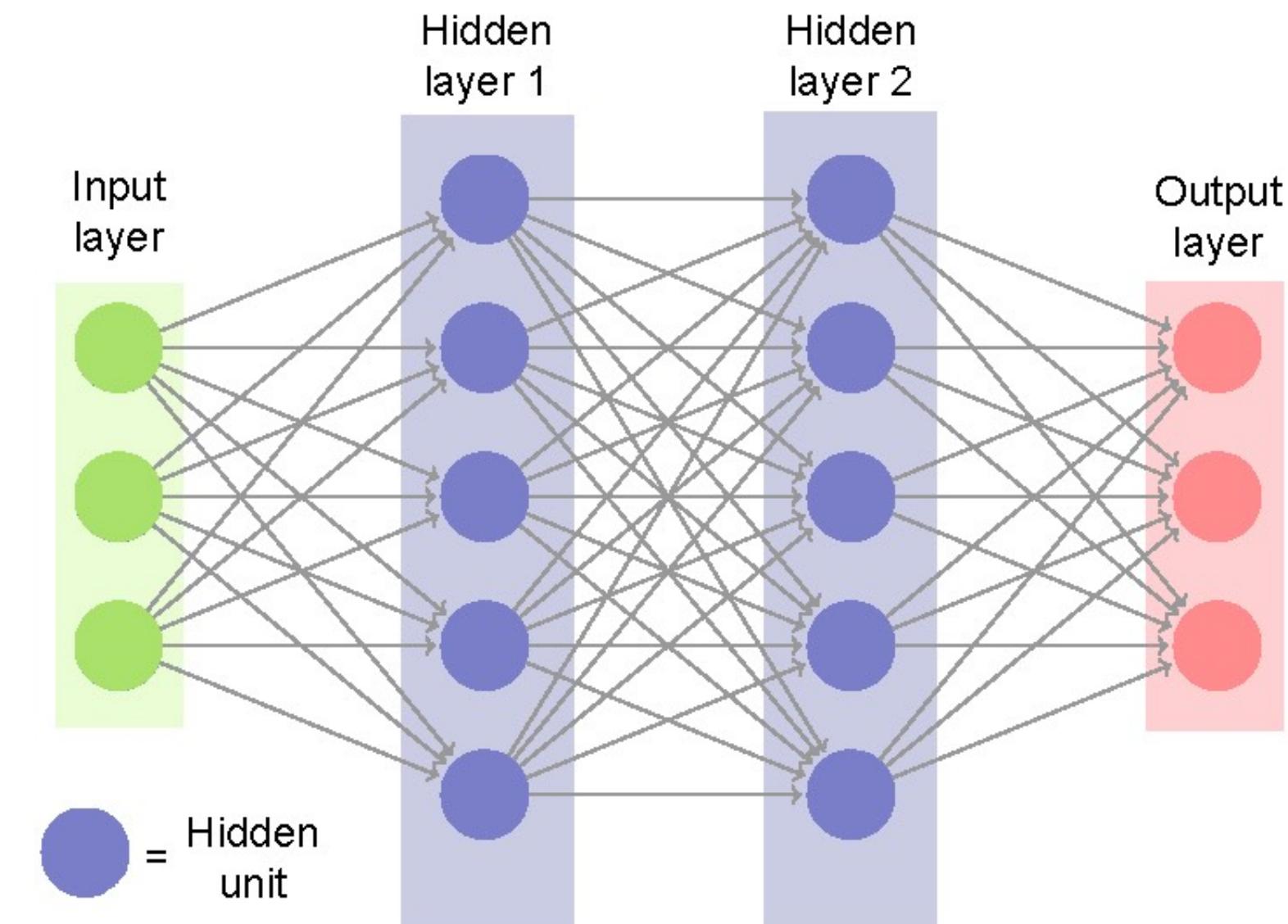
Concetti chiave



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.



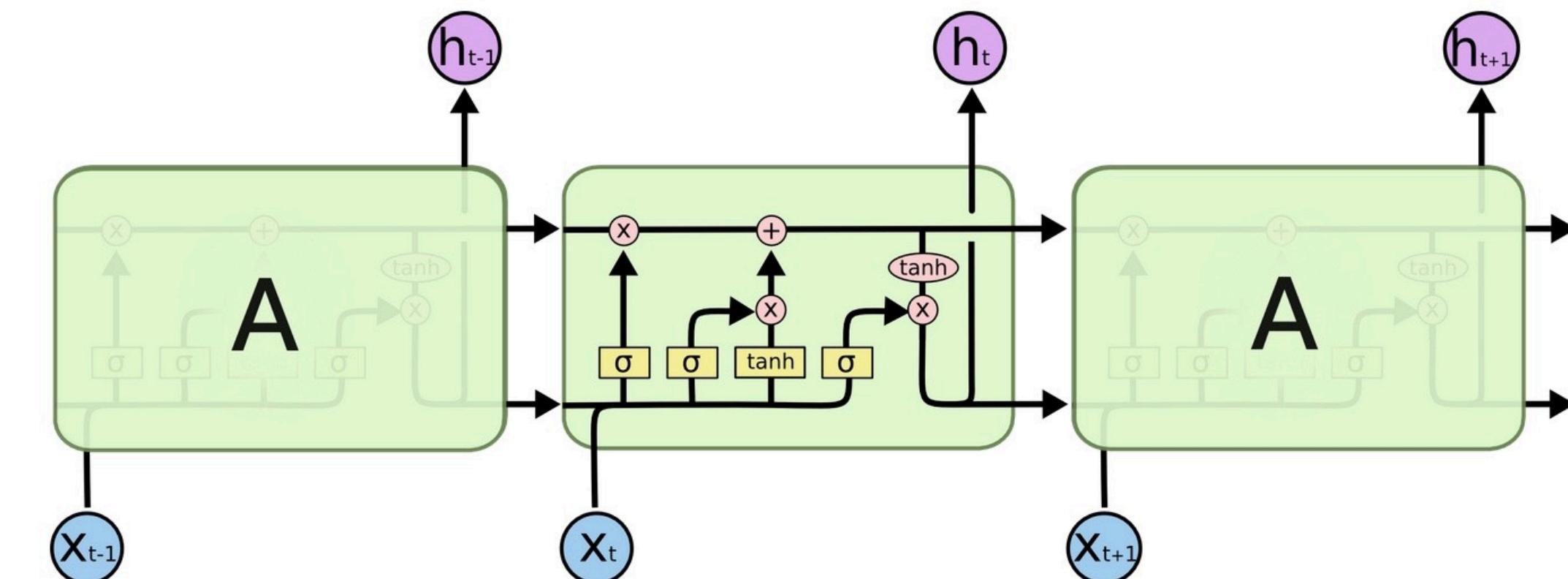
Concetti chiave



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



Detectron2

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

- Accuratezza allo stato dell'arte

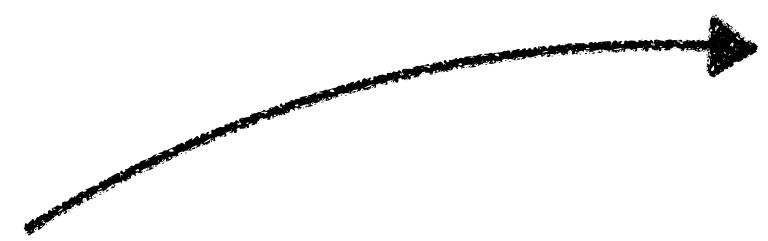
PoseNet

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

- Leggerezza modello
- Rapidità d'inferenza

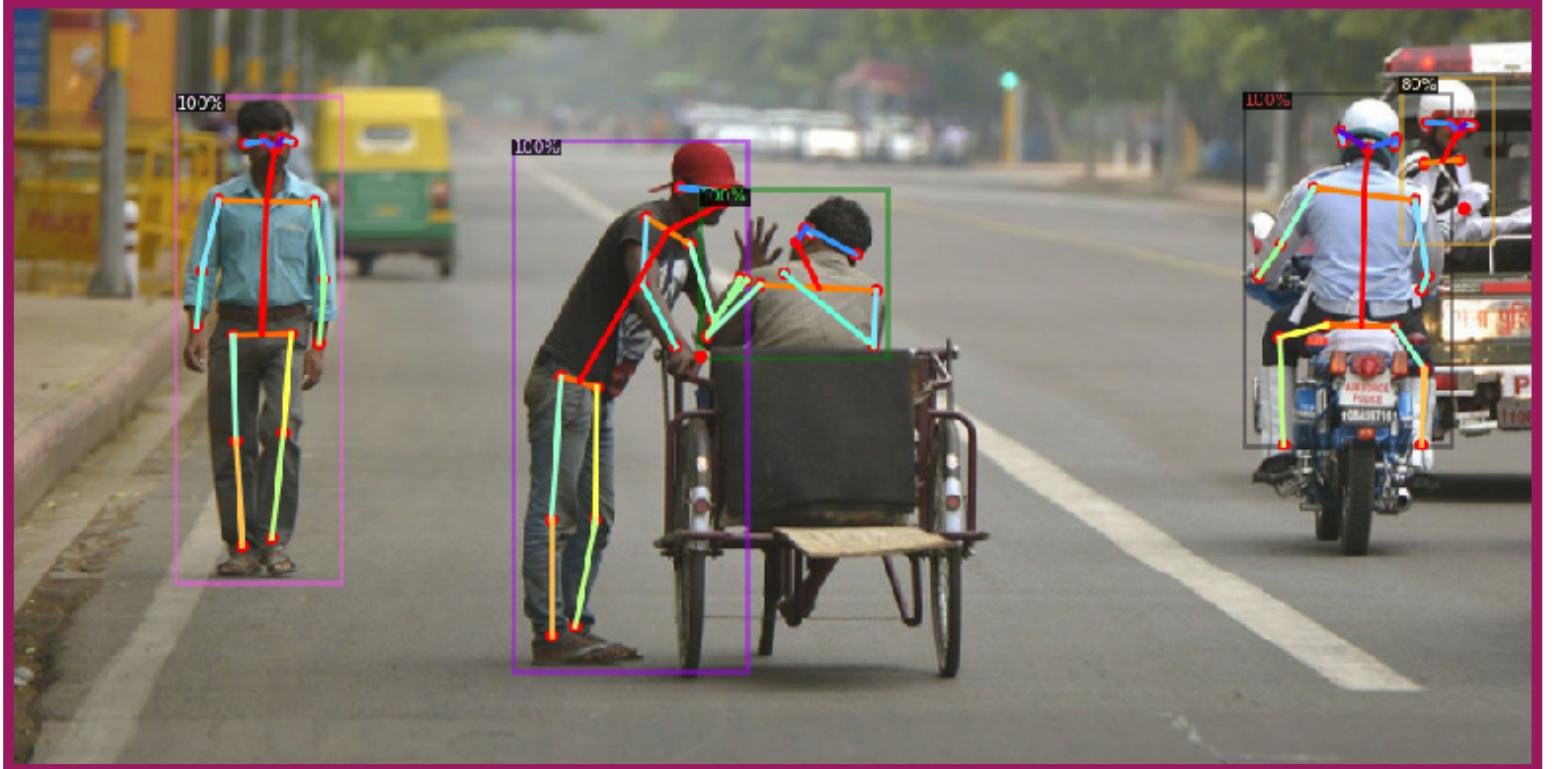


La stima della posa



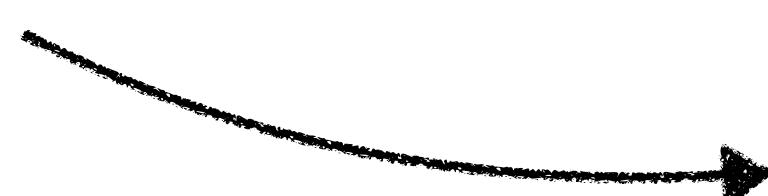
Detectron2

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



PoseNet

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

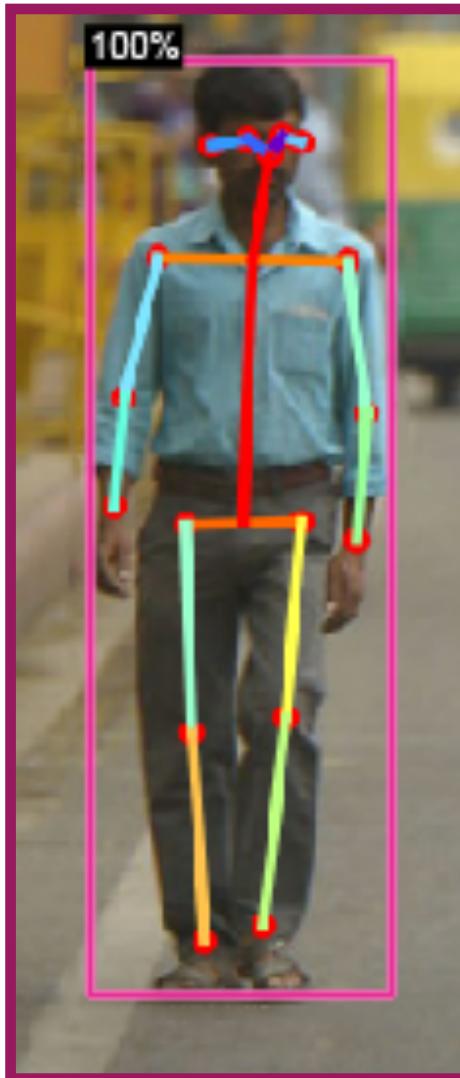


La stima della posa



Detectron2

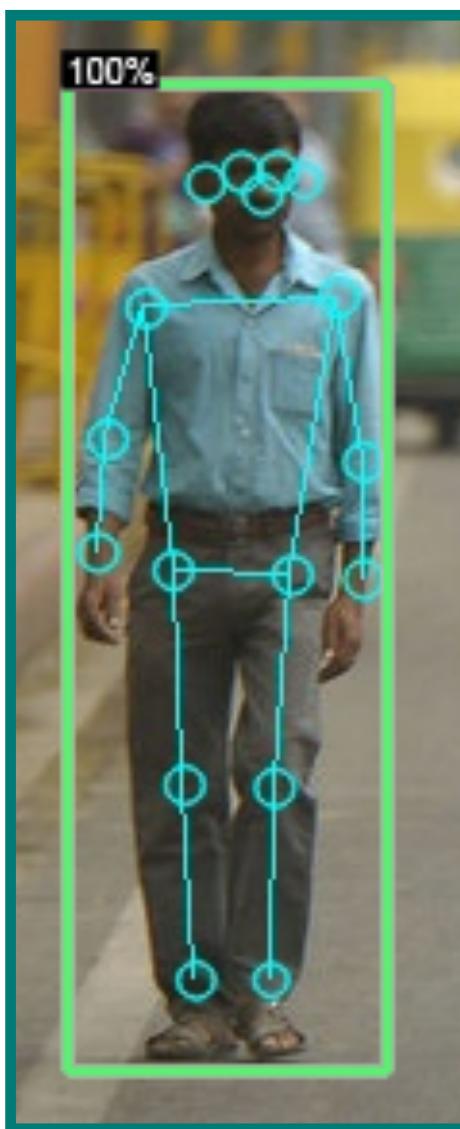
modello : *Mask R-CNN*
backbone : *ResNeXt+FPN*
pre-allenato su : *ImageNet*



- **17 punti** per persona
- **score [0,1]** per persona

PoseNet

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



Dataset utilizzato



NTU-RGB+D

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



Dataset utilizzato



NTU-RGB+D

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



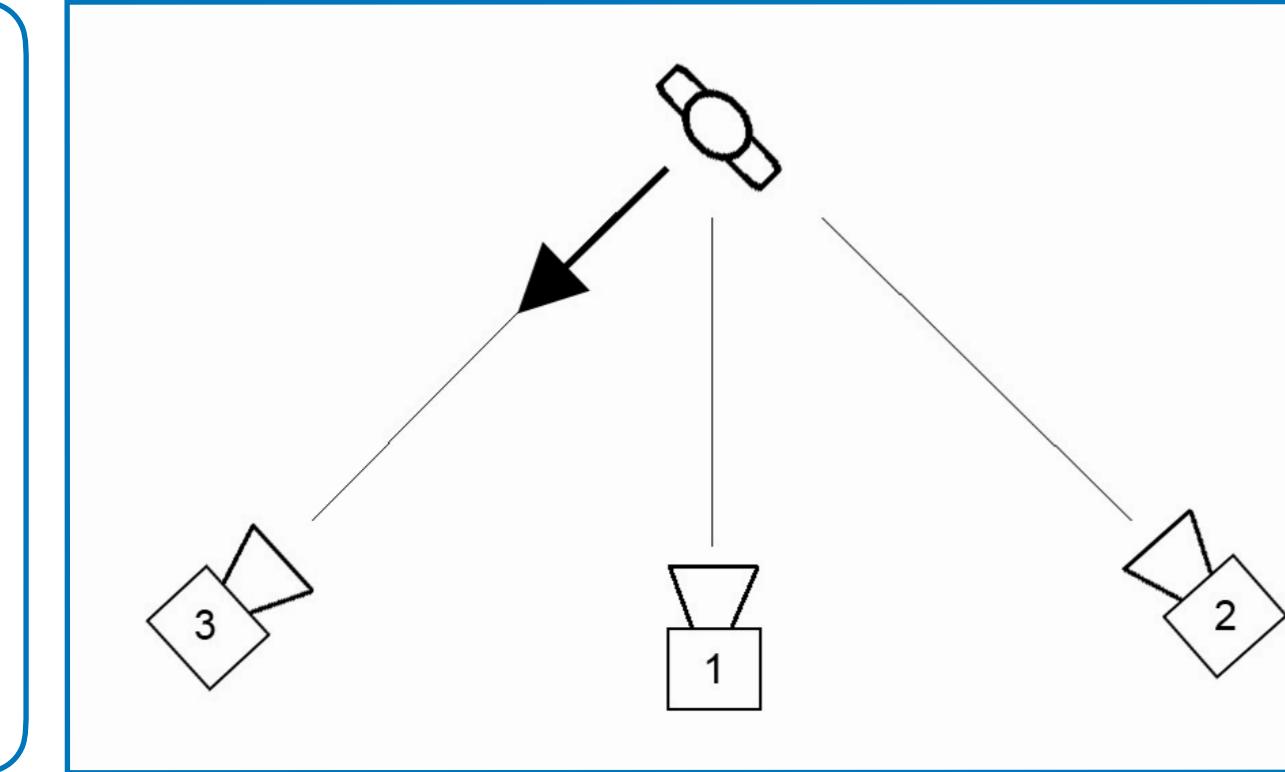
Dataset utilizzato



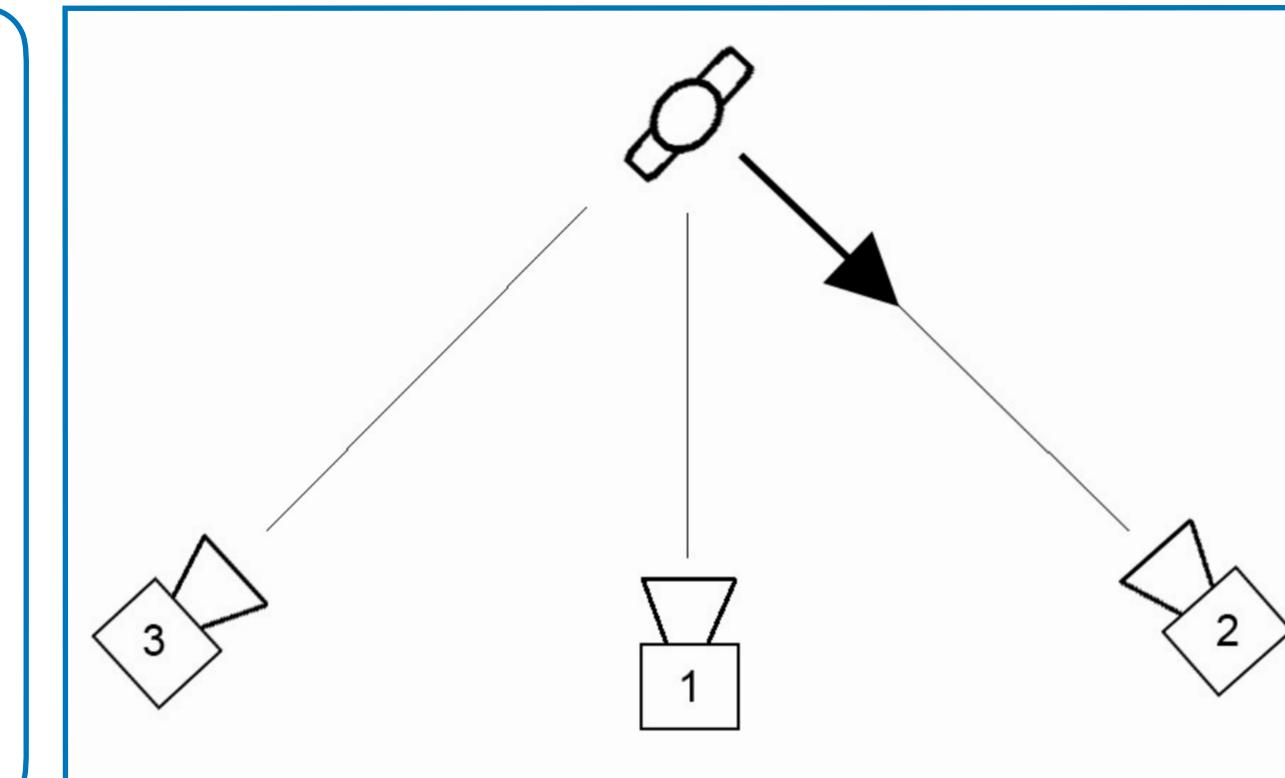
NTU-RGB+D

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

Ripetizione 1

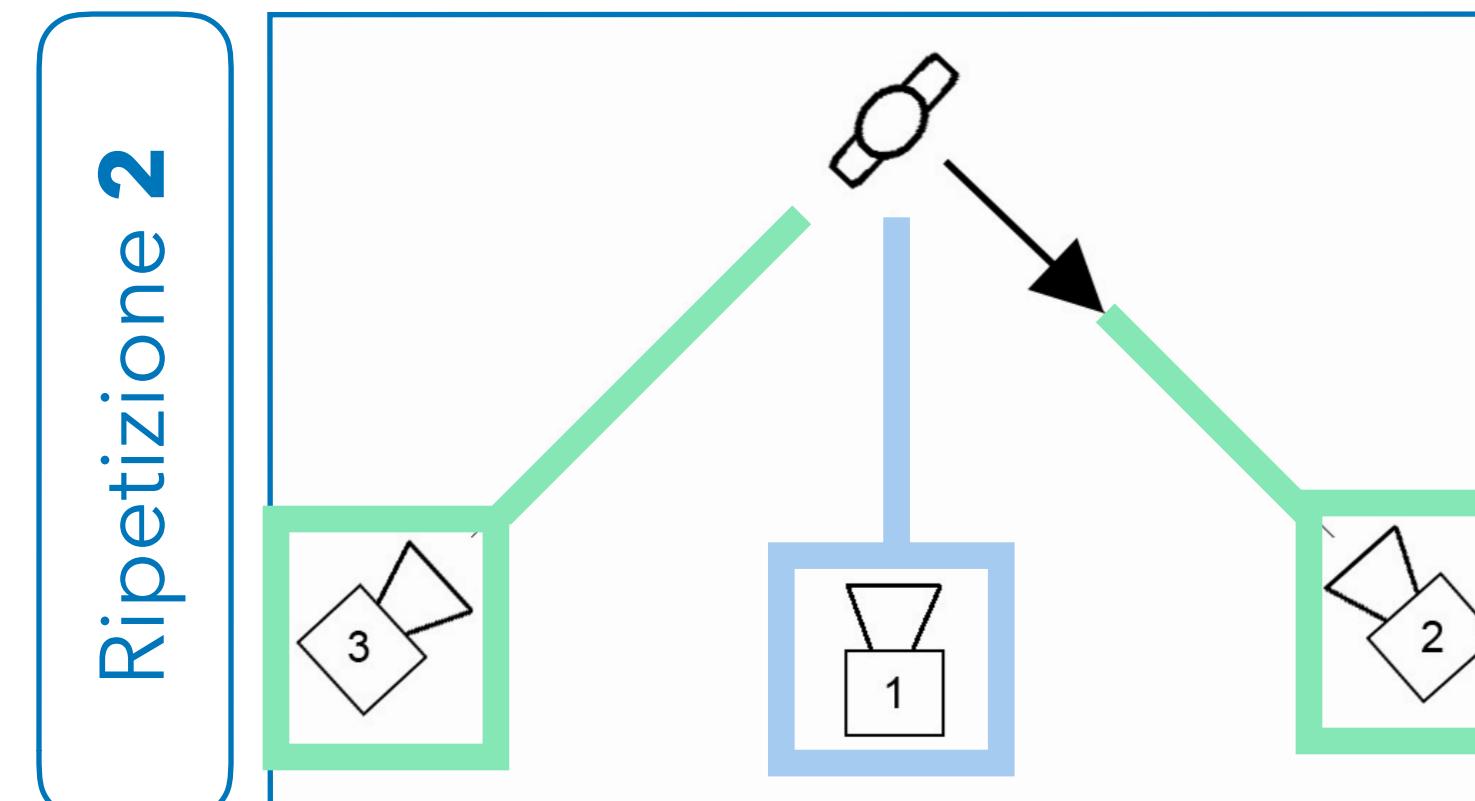
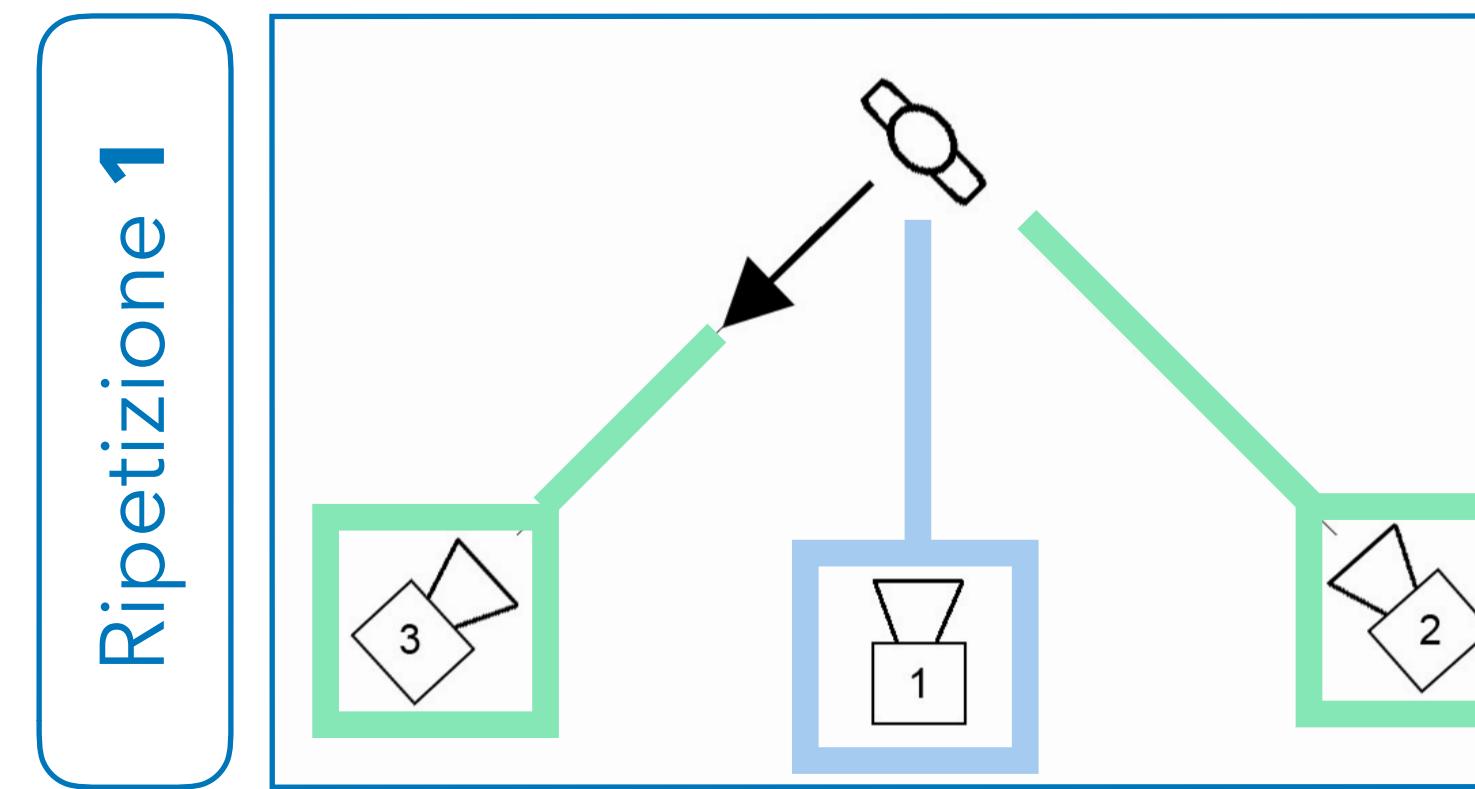


Ripetizione 2



Dataset utilizzato

NTU-RGB+D



Cross View

Training

id camera: 2,3
video: 37920

Test

id camera: 1
video: 18960

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

contiene tutte le
prospettive a **45°**

Dataset utilizzato



NTU-RGB+D

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
video: 16560

Cross View

Training

id camera: 2,3
video: 37920

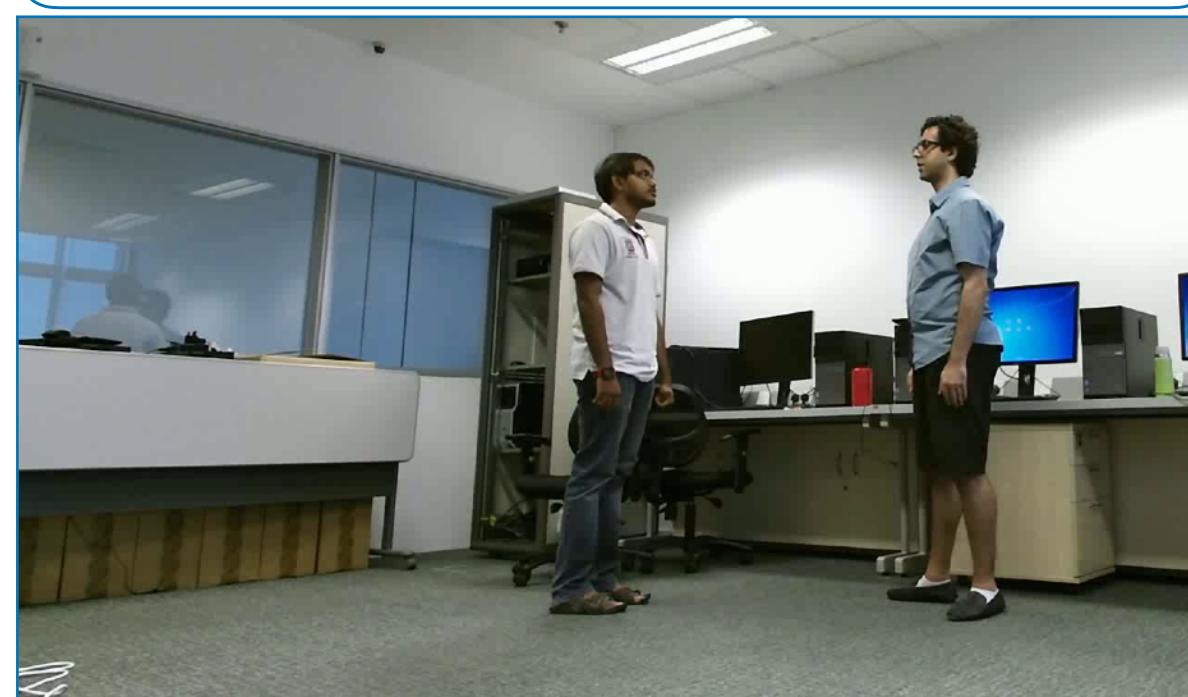
Test

id camera: 1
video: 18960

Metodo proposto - Preprocessing

Assegnazione coerente delle pose

Frame 1



Frame 12



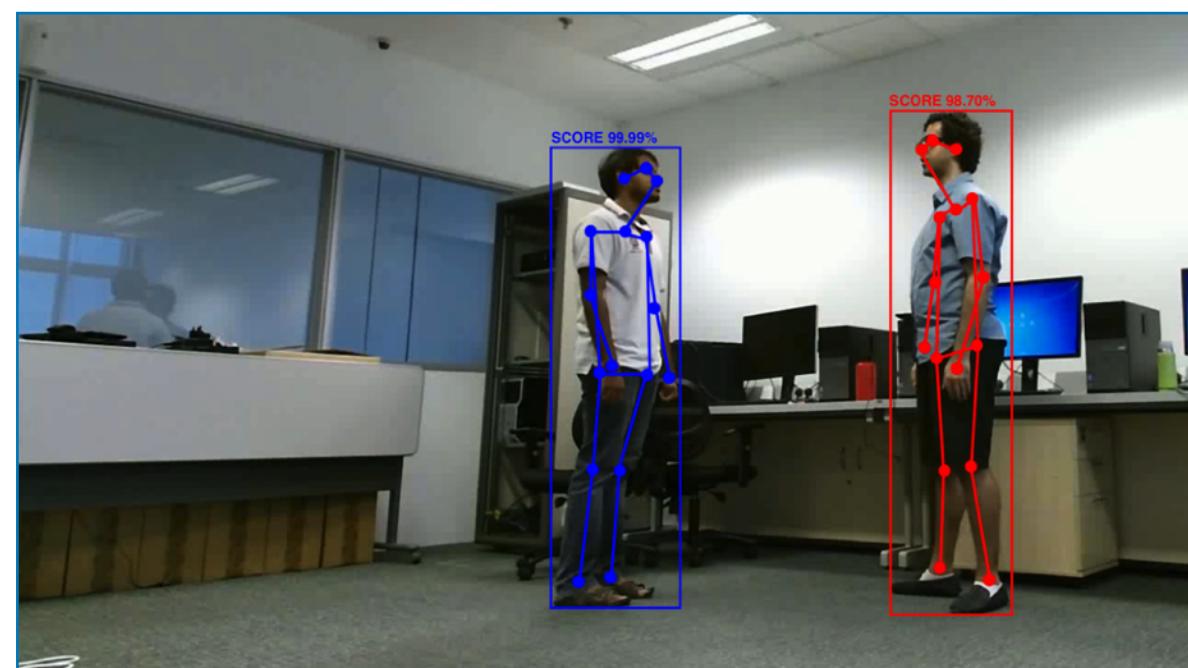
Frame 43



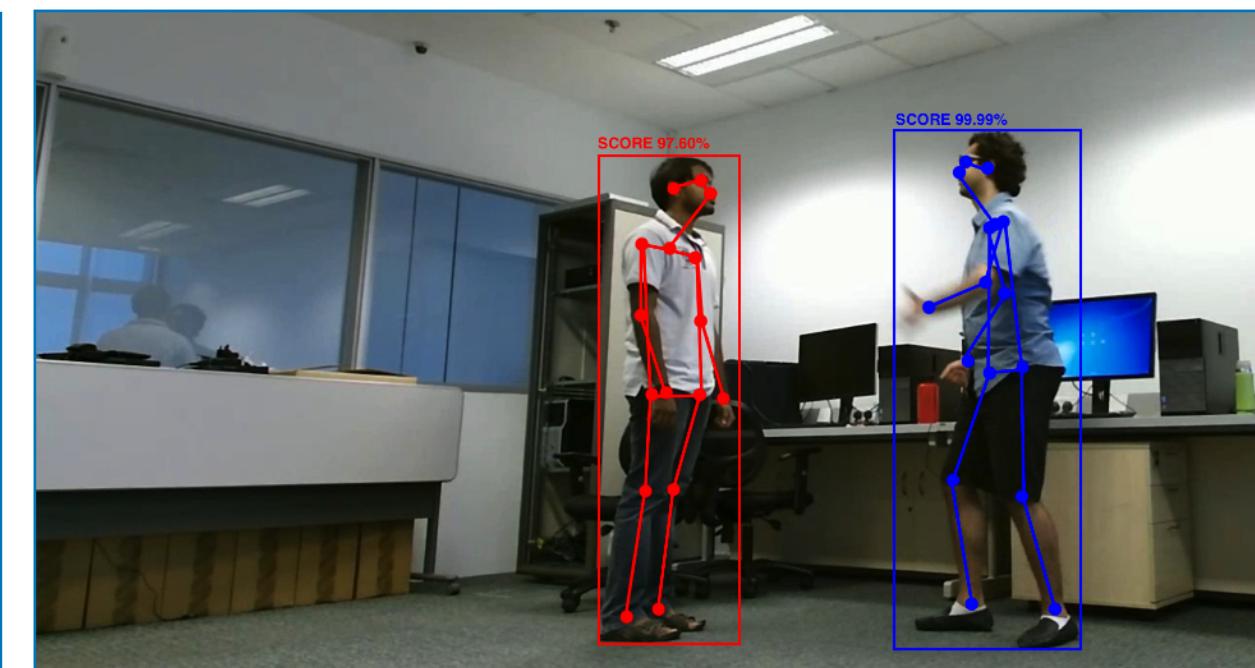
Frame 62



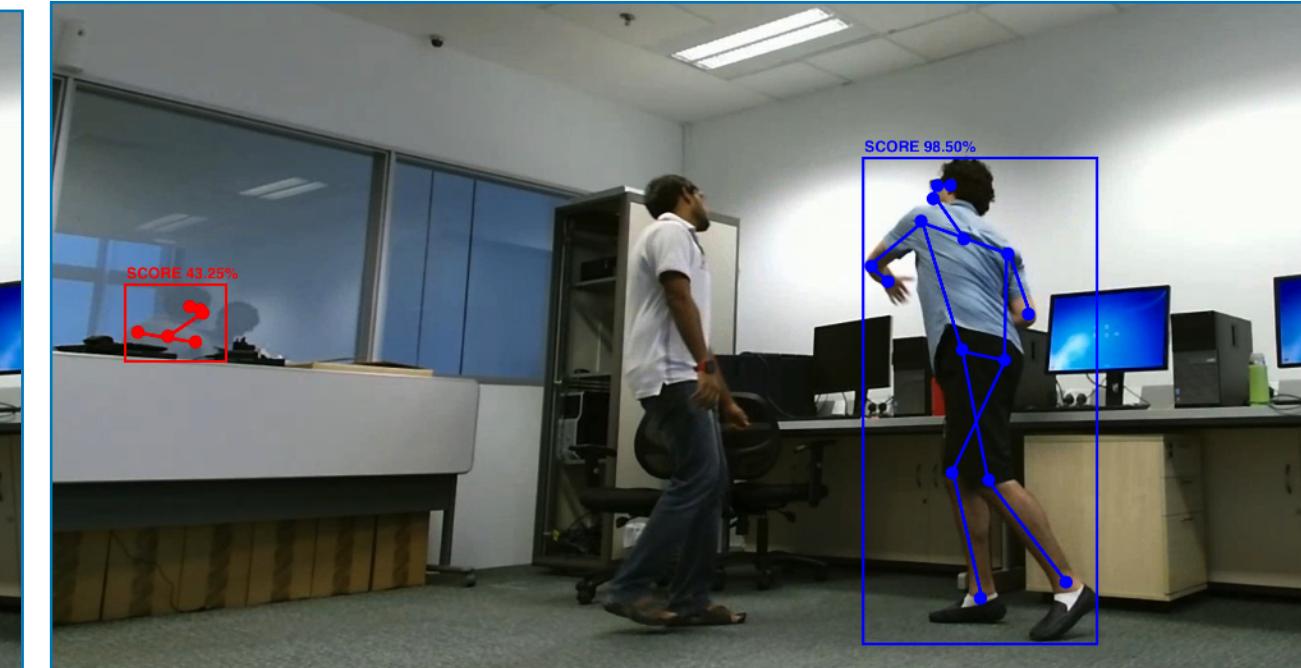
Frame 1



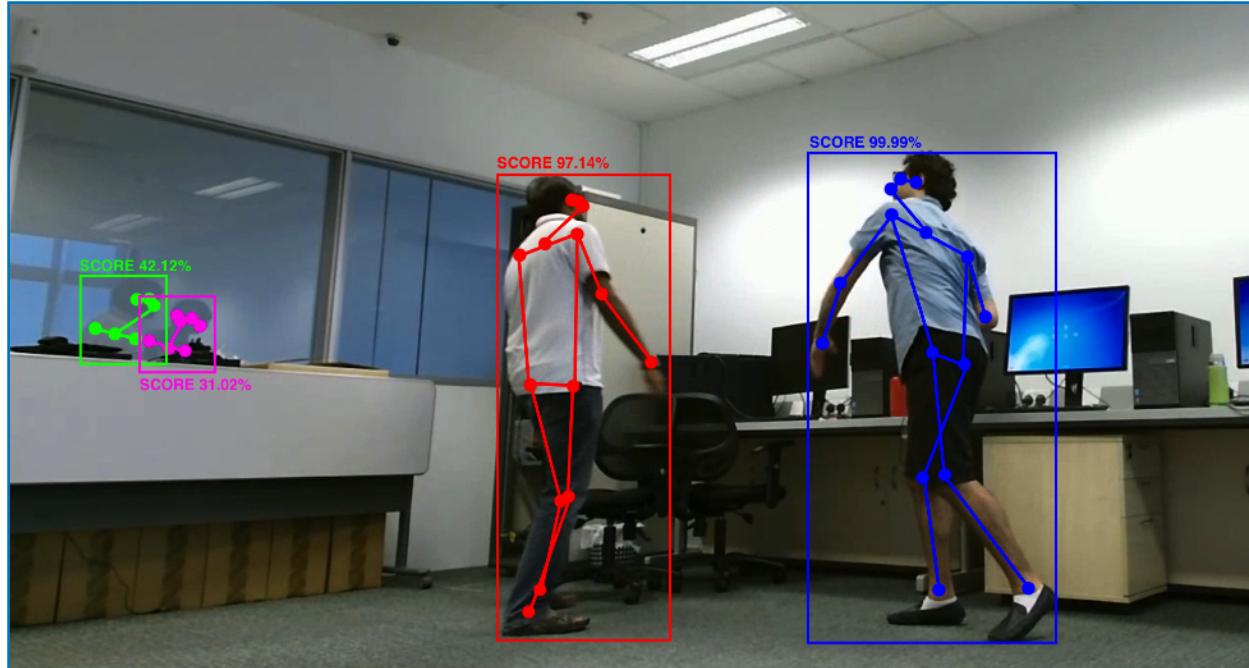
Frame 12



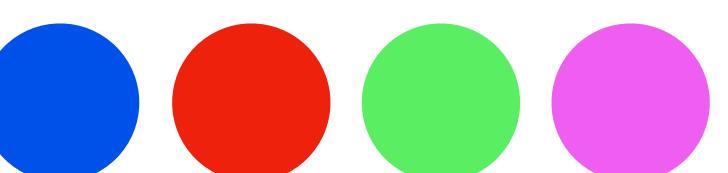
Frame 43



Frame 62



Ordine delle pose restituite dall'algoritmo:



Metodo proposto - Preprocessing

Assegnazione coerente delle pose

K = tipi di punti chiave (giunti)

Posa $\mathcal{P}_i = \{p_{i,k}\}_{k=1}^K$

$$p_{i,k} = (x_{i,k}, y_{i,k})$$

$I(p) = 1$ se p è definito, 0 altrimenti

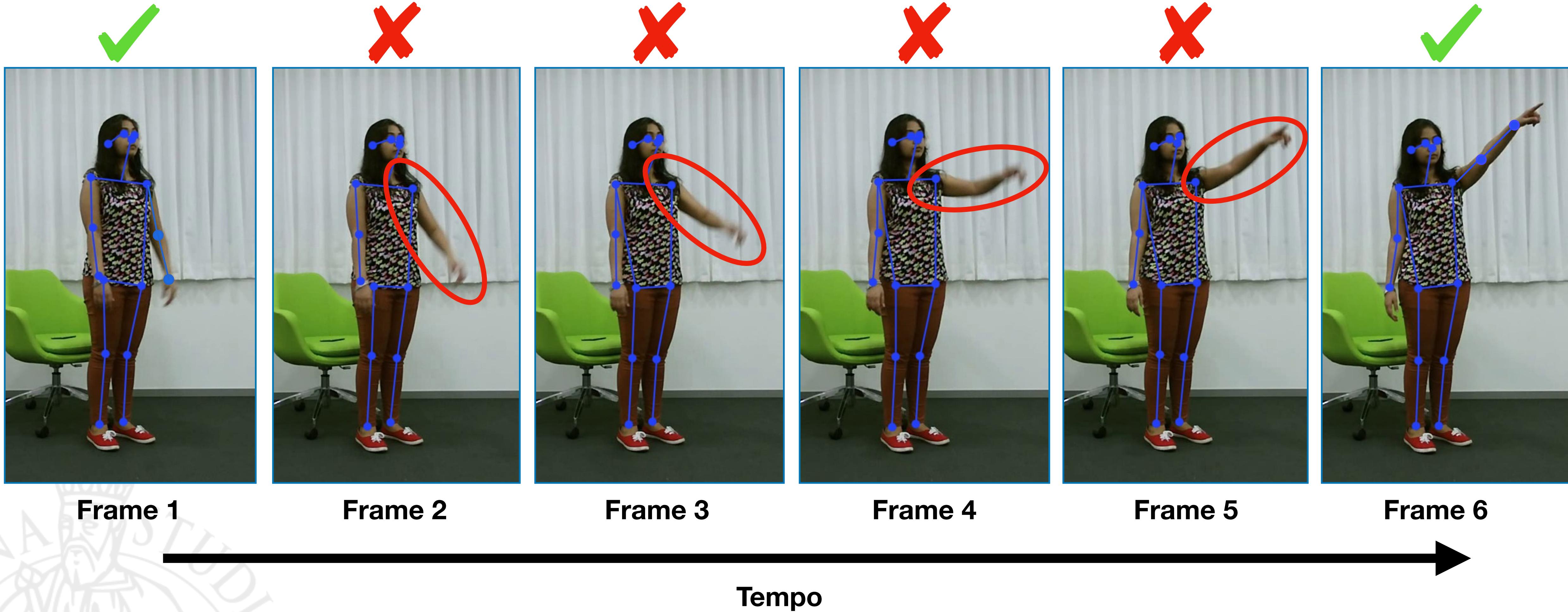
Distanza $\mathcal{D}_{ij} = \frac{\sum_{k=1}^K I(p_{i,k}) \cdot I(p_{j,k}) \cdot eucl(p_{i,k}, p_{j,k})}{\sum_{k=1}^K I(p_{i,k}) \cdot I(p_{j,k})}$

Regole di selezione

- 1) Le sequenze di pose vengono create seguendo le **distanze minime** fra pose di frame consecutivi
- 2) Per ogni video, solo le **2** sequenze di pose col **maggior score medio** sono state selezionate

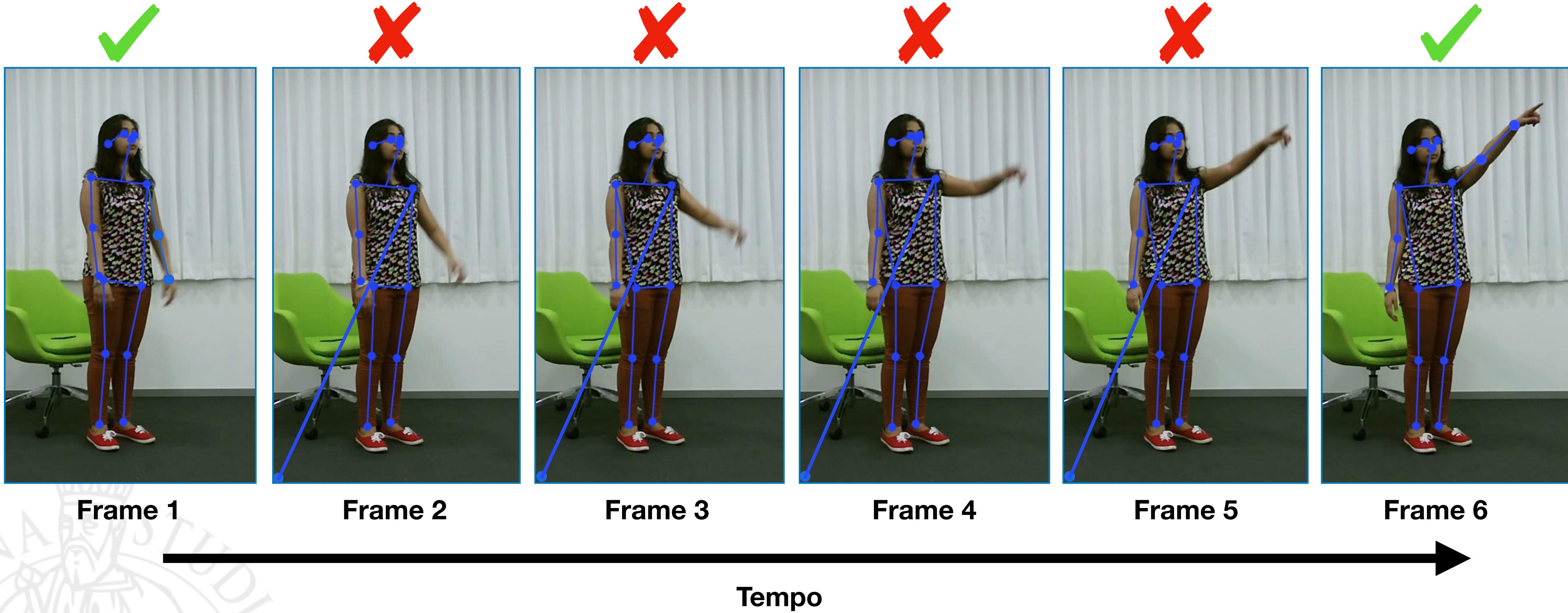
Metodo proposto - Preprocessing

Rimozione degli zeri



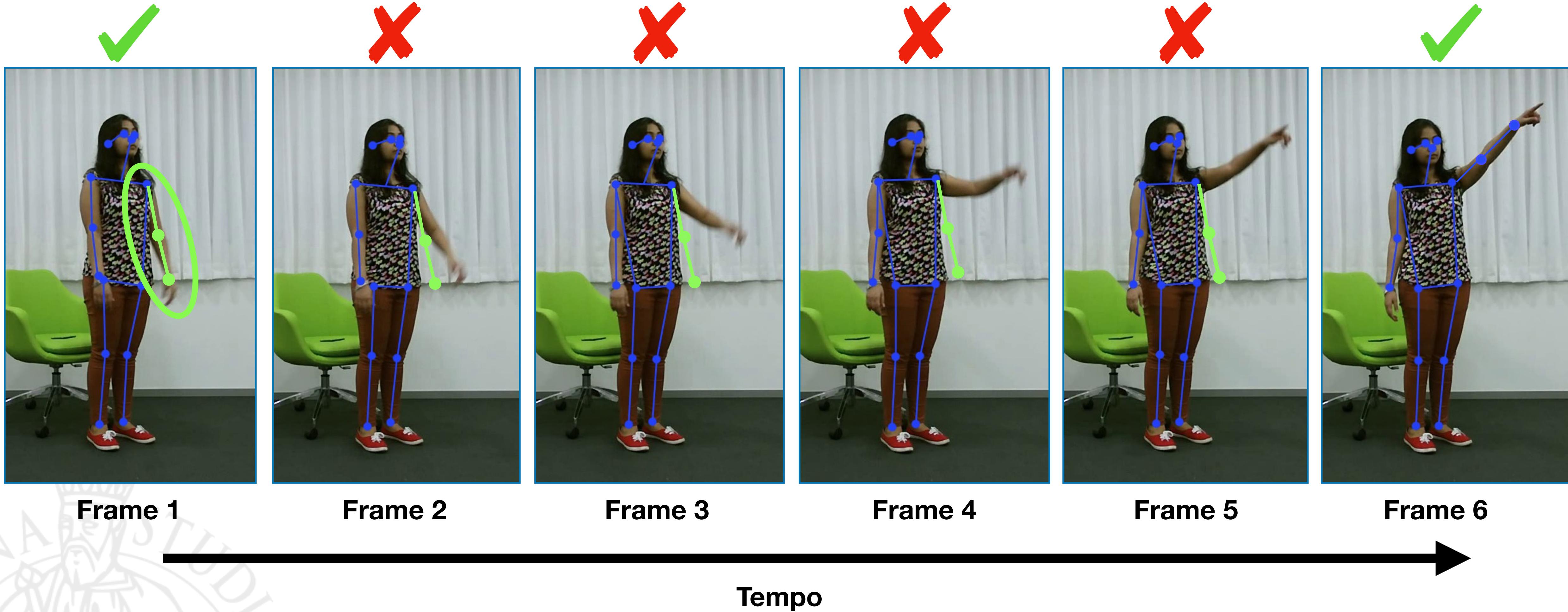
Metodo proposto - Preprocessing

Rimozione degli zeri



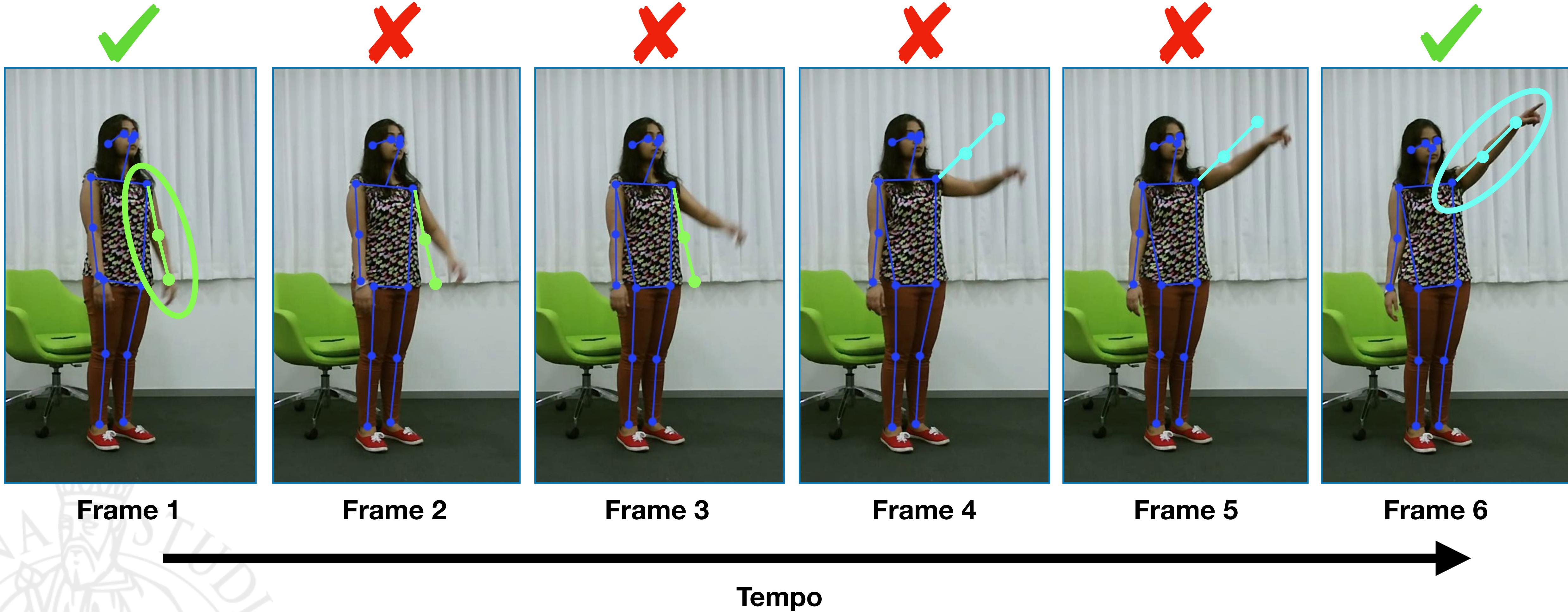
Metodo proposto - Preprocessing

Rimozione degli zeri



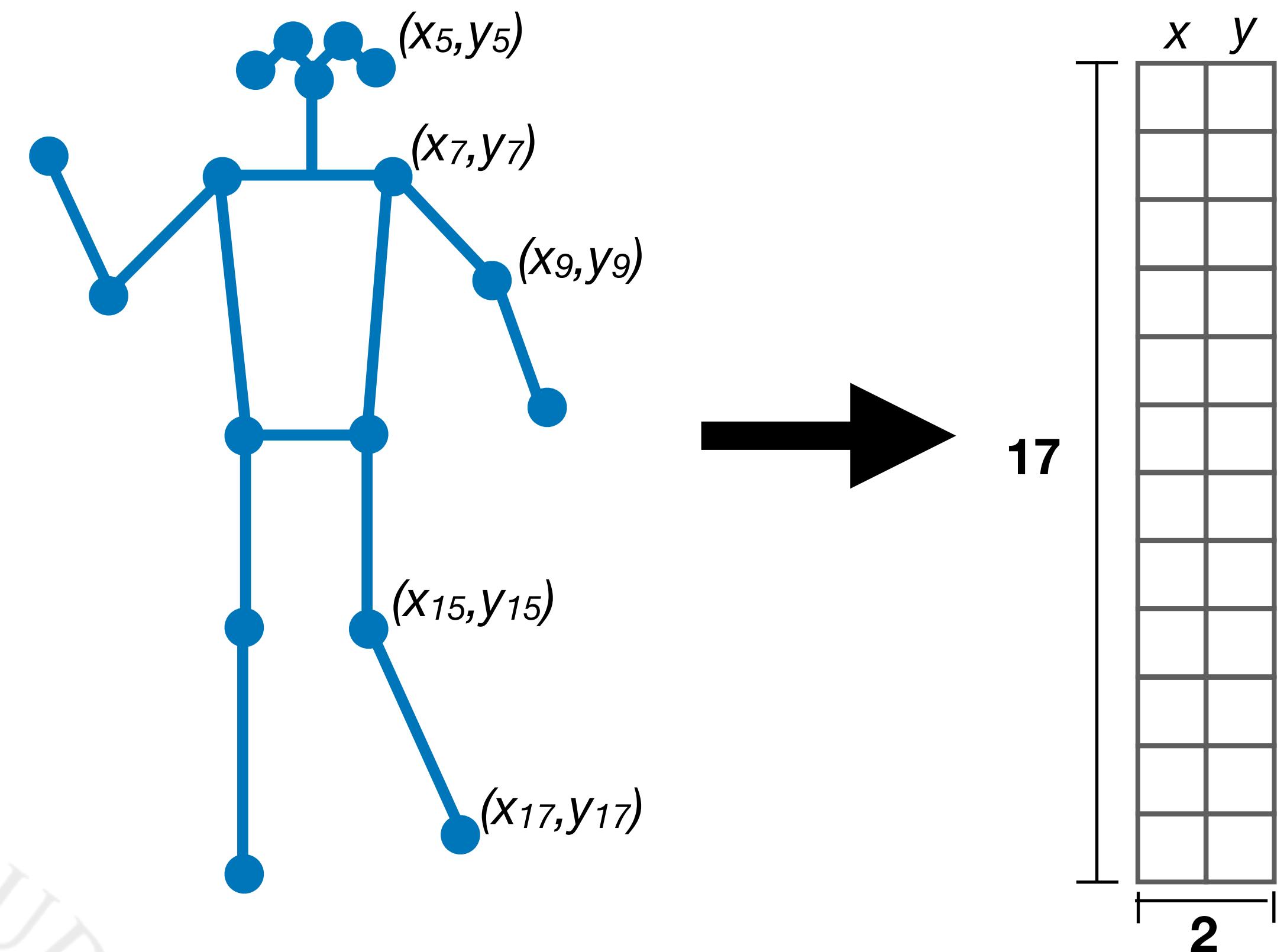
Metodo proposto - Preprocessing

Rimozione degli zeri



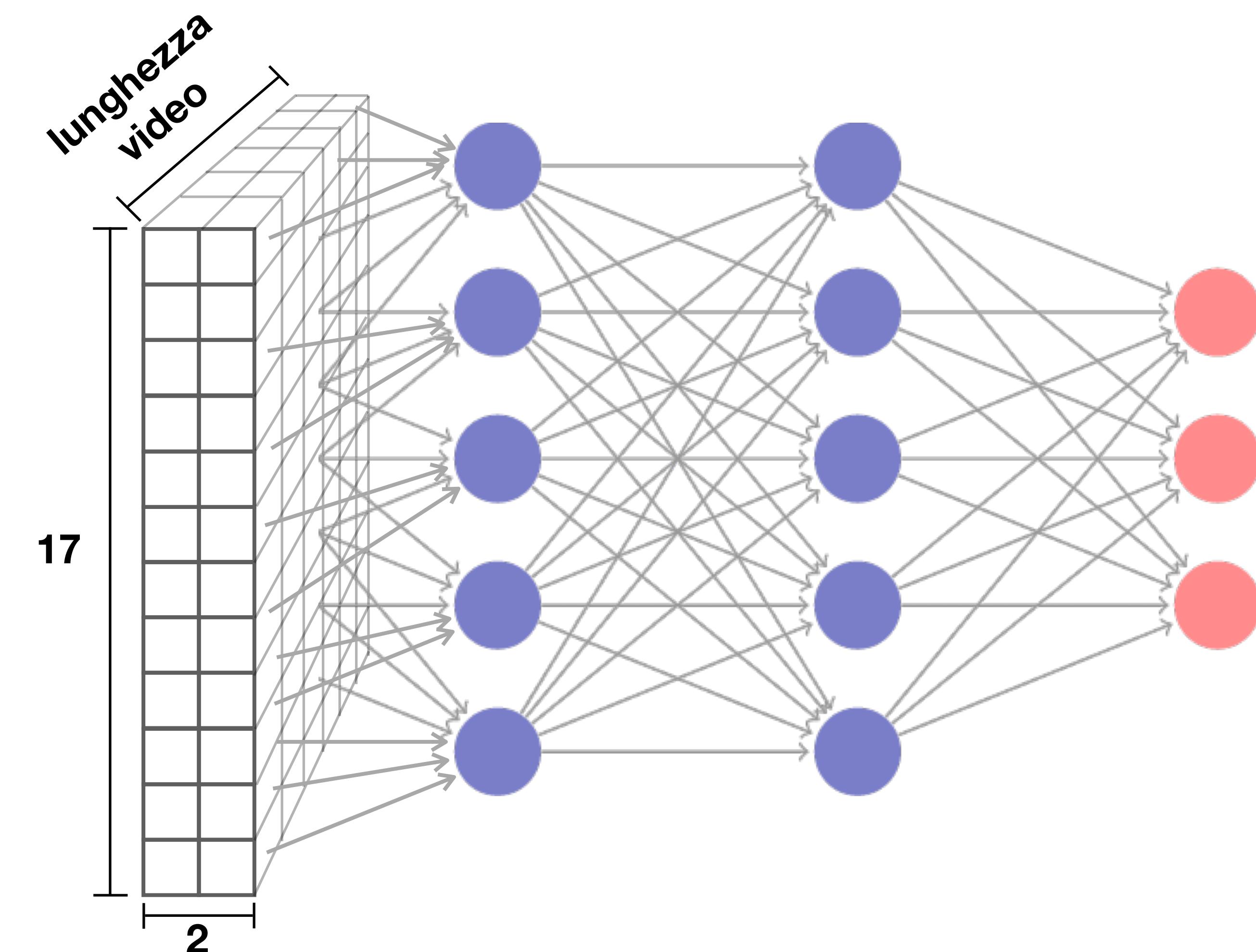
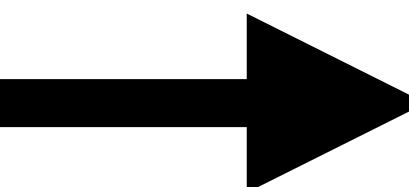
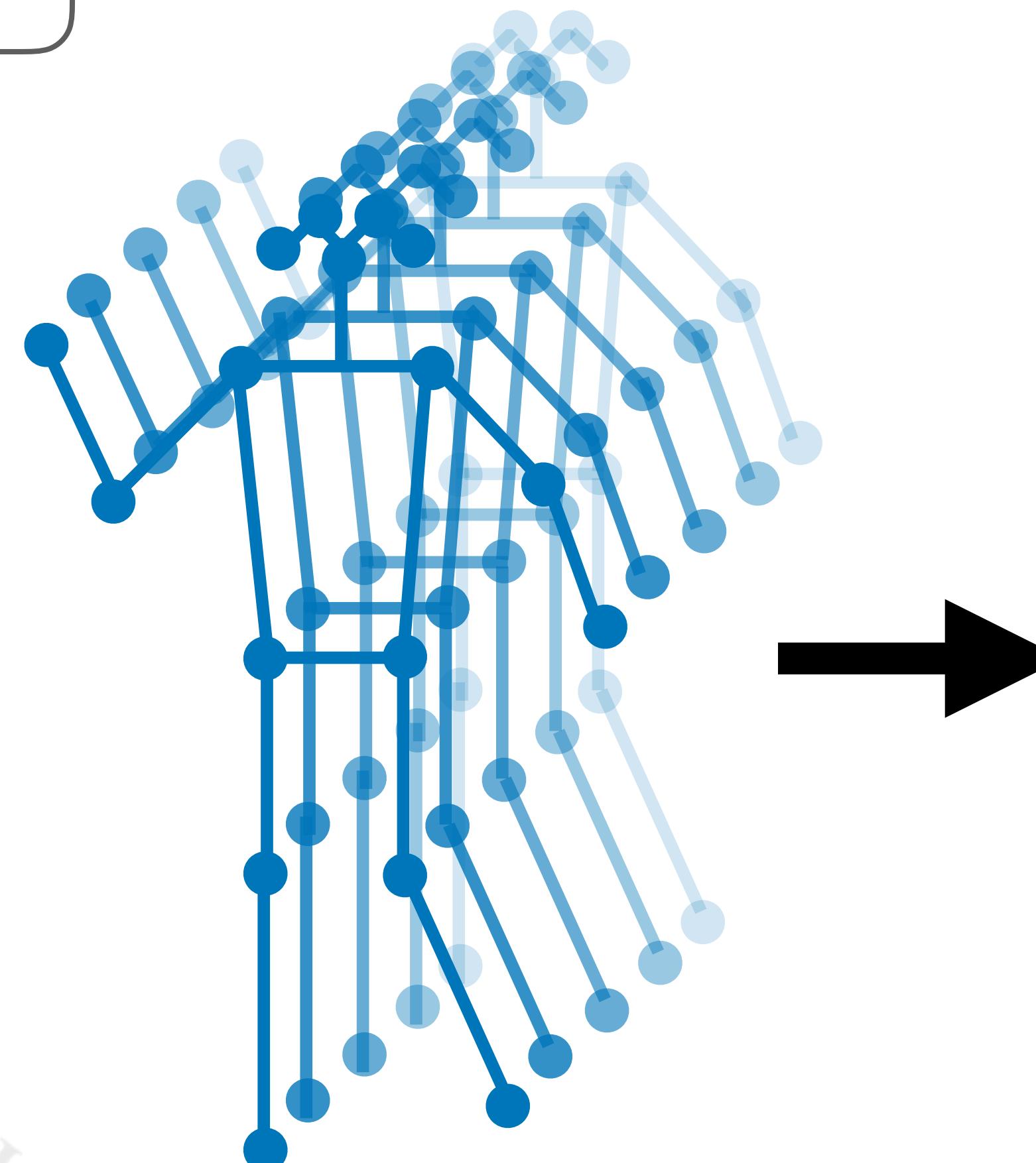
Metodo proposto - Tecniche di rielaborazione

Semplice



Metodo proposto - Tecniche di rielaborazione

Semplice

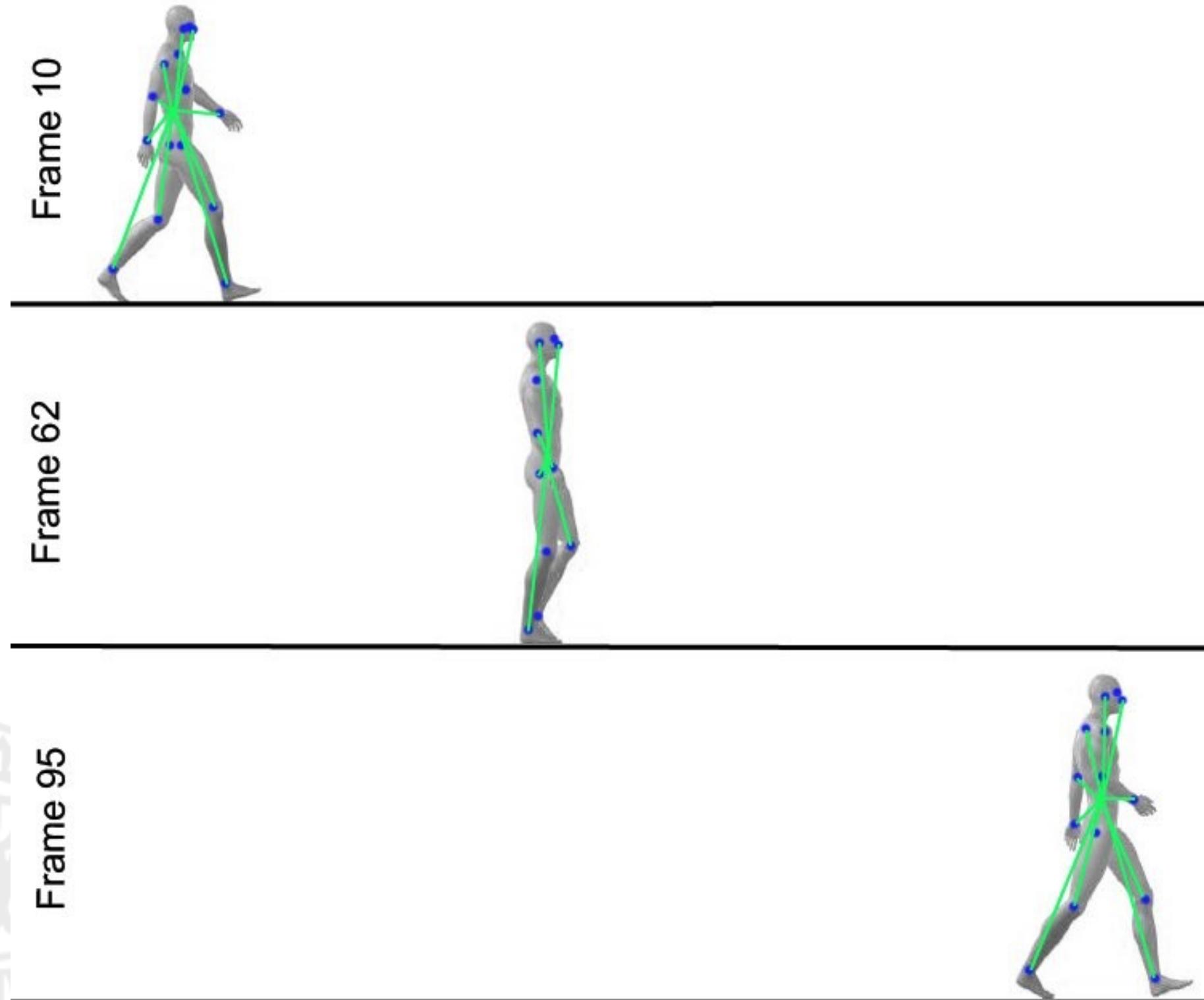


Metodo proposto - Tecniche di rielaborazione

Semplice

Baricentro

Frame



Baricentro posa: $\bar{p}_f = \frac{\sum_{j=1}^K p_{jf}}{K}$

Trasformazione posa: $T_f = \{t_{1f}, \dots, t_{Kf}\}$

Personale

$$t_{if} = p_{if} - \bar{p}_f$$

Globale

$$\bar{g}_f = \frac{\bar{p}_{1,f} + \bar{p}_{2,f}}{2}$$

$$t_{if} = p_{if} - \bar{g}_f$$

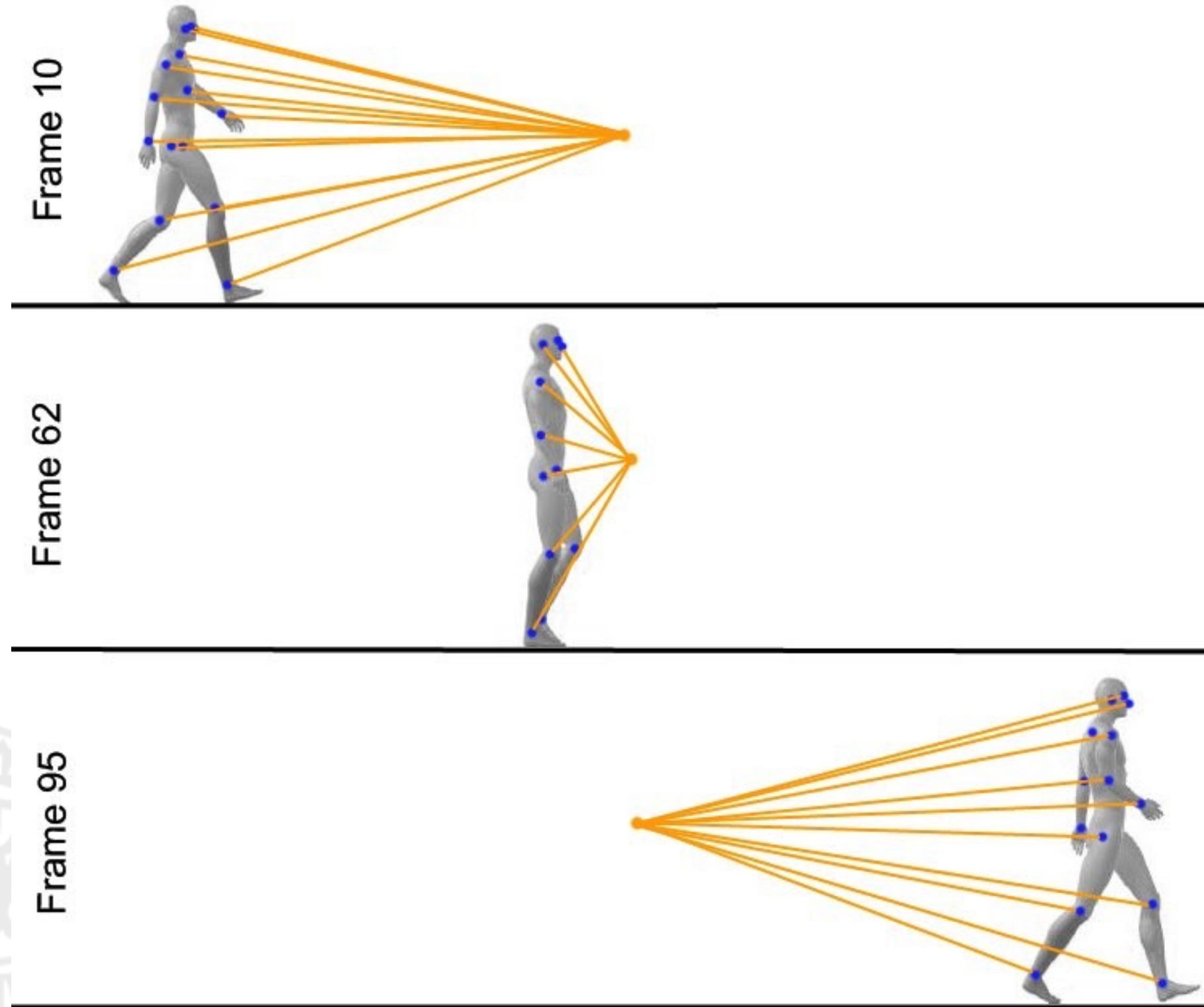
Metodo proposto - Tecniche di rielaborazione

Semplice

Baricentro

Frame

Video



Baricentro video: $\bar{v} = \frac{\sum_{f=1}^L \bar{p}_f}{L}$

Trasformazione posa: $\mathcal{T}_f = \{t_{1f}, \dots, t_{Kf}\}$

Personale

$$t_{if} = p_{if} - \bar{v}$$

Globale

$$\bar{v}_G = \frac{\sum_{f=1}^L \bar{g}_f}{L}$$

$$t_{if} = p_{if} - \bar{v}_G$$

Metodo proposto - Tecniche di rielaborazione

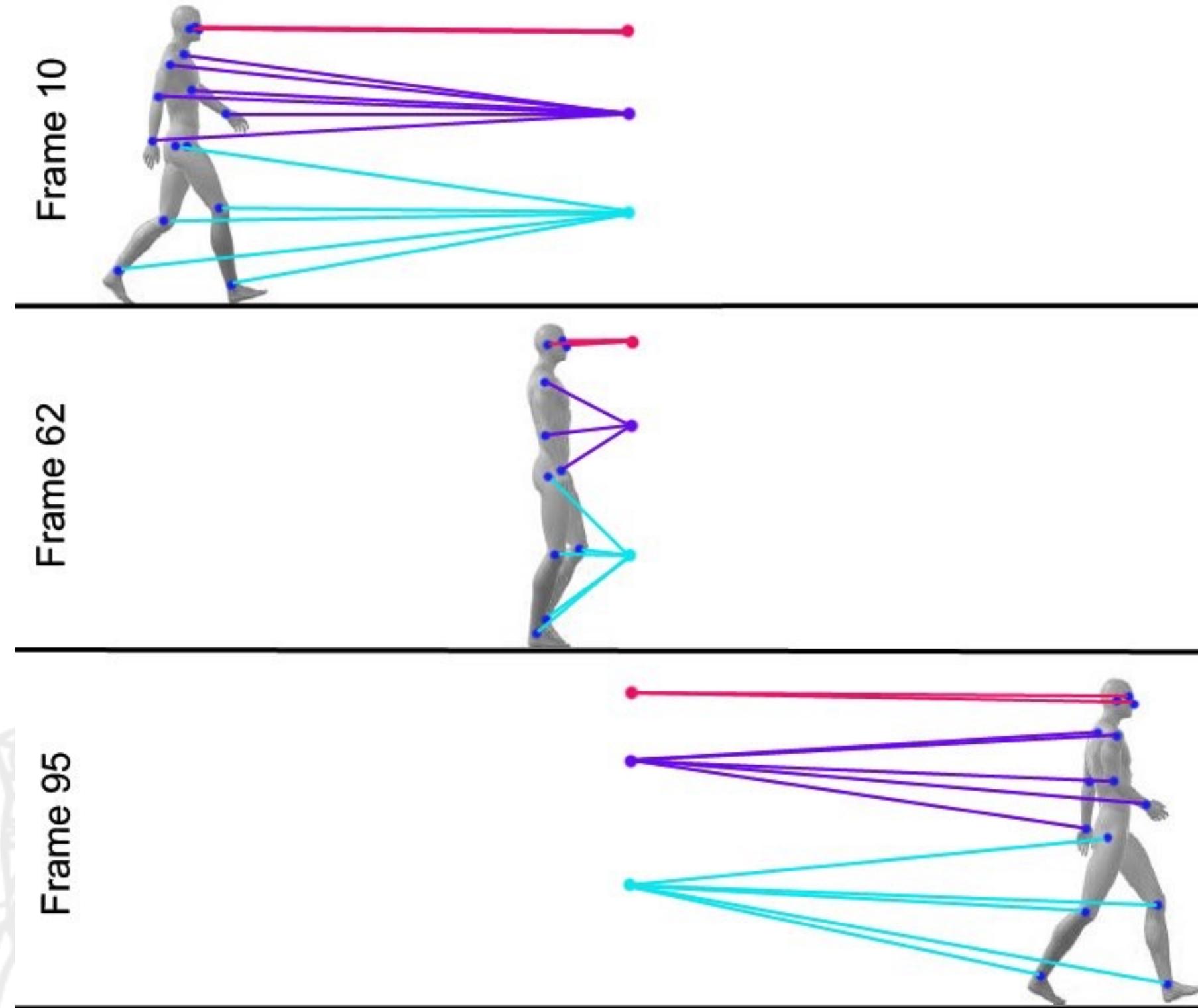
Semplice

Baricentro

Frame

Video

Video multiplo



B sottogruppi: $K = \{\mathcal{K}_1, \dots, \mathcal{K}_B\}, \quad \bigcup_{b=1}^B \mathcal{K}_b = K$

$$\mathcal{K}_i \cap \mathcal{K}_j = \emptyset, \quad \forall \mathcal{K}_i, \mathcal{K}_j \in K$$

Baricentro video: $\bar{\mathcal{B}}_b = \frac{\sum_{i \in \mathcal{B}_b} \sum_{f=1}^L \mathbf{p}_{if}}{L}$

Trasformazione posa: $\mathcal{T}_f = \{\mathbf{t}_{1f}, \dots, \mathbf{t}_{Kf}\}$

Personale

$$\mathbf{t}_{if} = \mathbf{p}_{if} - \bar{\mathcal{B}}_b$$

Assoluta

$$\mathbf{t}_{if} = |\mathbf{p}_{if} - \bar{\mathcal{B}}_b|$$

Globale

$$\bar{\mathcal{G}}_b = \frac{\sum_{i \in \mathcal{B}_b} \sum_{f=1}^L \text{mean}(\mathbf{p}_{1if}, \mathbf{p}_{2if})}{L}$$

$$\mathbf{t}_{if} = \mathbf{p}_{if} - \bar{\mathcal{G}}_b$$

Metodo proposto - Tecniche di rielaborazione

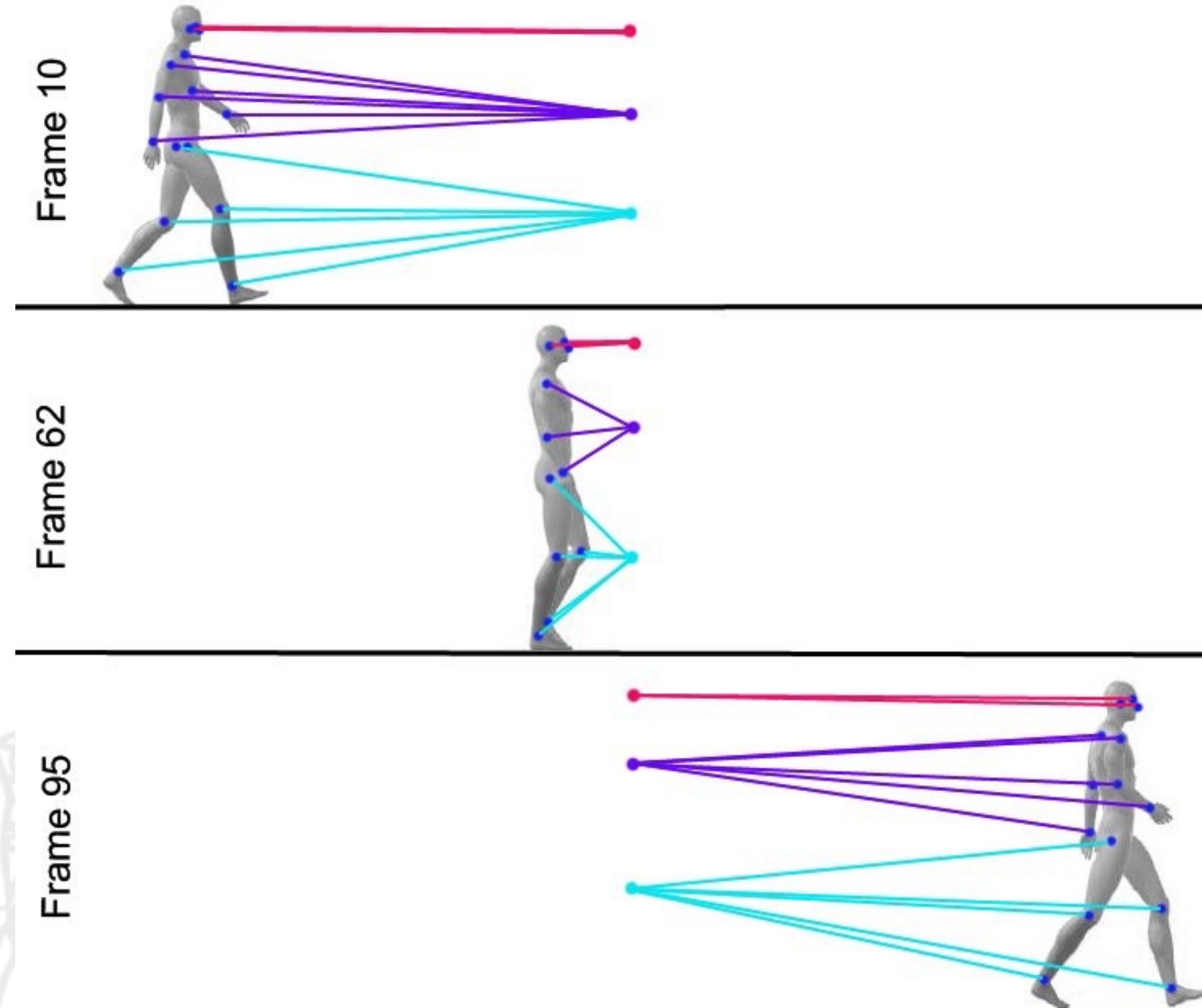
Semplice

Baricentro

Frame

Video

Video multiplo



B sottogruppi: $K = \{\mathcal{K}_1, \dots, \mathcal{K}_B\}, \quad \bigcup_{b=1}^B \mathcal{K}_b = K$

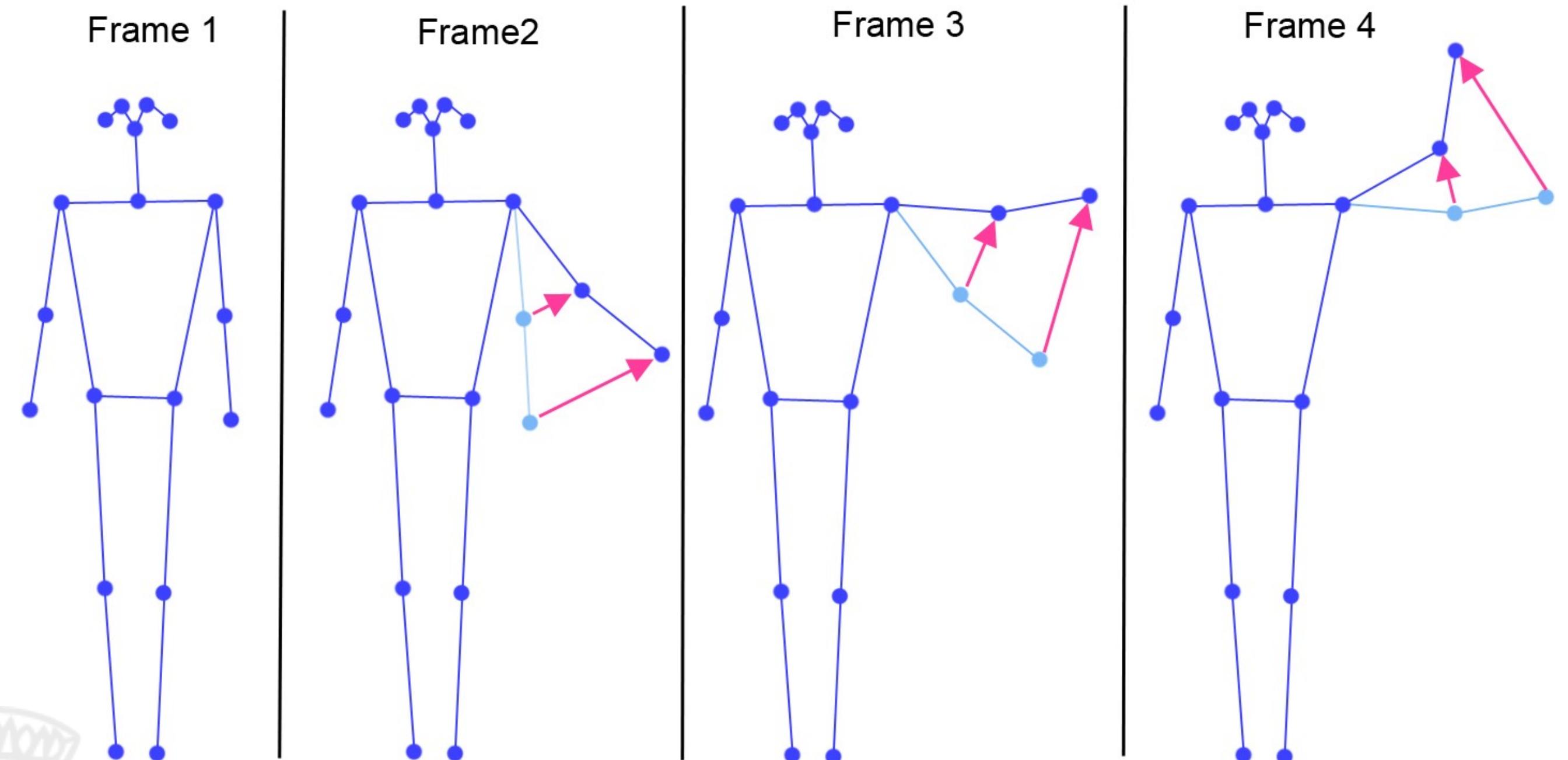
$$\mathcal{K}_i \cap \mathcal{K}_j = \emptyset, \quad \forall \mathcal{K}_i, \mathcal{K}_j \in K$$

- **B = 3:** Testa - Bust - Gambe

- **B = 5:** Come suggerita dagli autori di NTU-RGB+D

- **B = 17:** Un baricentro per ogni giunto

Metodo proposto - Tecniche di rielaborazione

[Semplice](#)
[Baricentro](#)
[Next frame](#)


Differenza tra la posa al frame f e quella al frame $f+1$

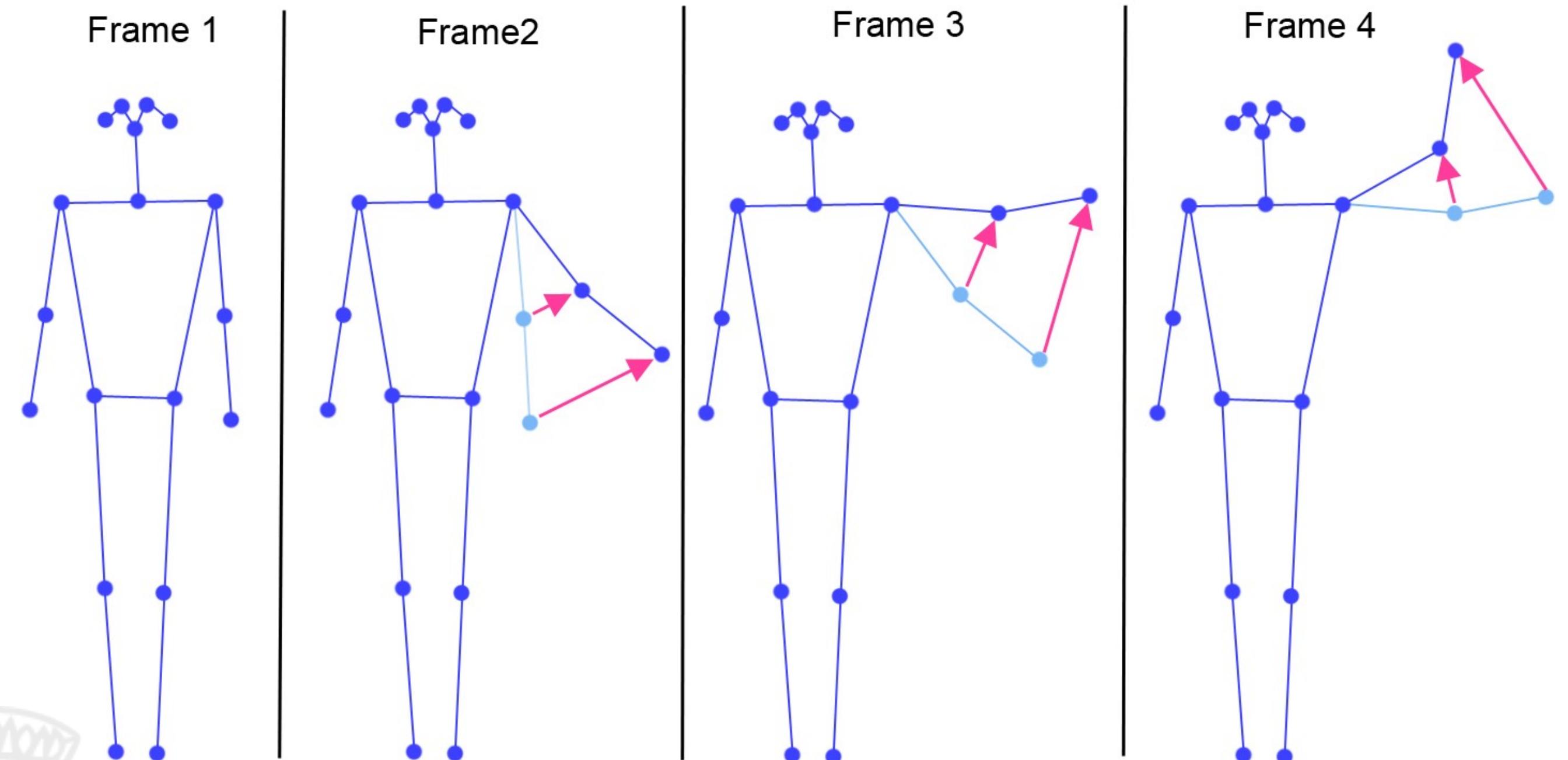
$$\begin{aligned}\mathcal{T}_f &= \mathcal{P}_{f+1} - \mathcal{P}_f \\ &= \bigcup_{k \in K} p_{k,f+1} - p_{k,f}\end{aligned}$$

Metodo proposto - Tecniche di rielaborazione

Semplice

Baricentro

Next frame



Differenza tra la posa al frame f e quella al frame $f+S$

$$\begin{aligned}\mathcal{T}_f &= \mathcal{P}_{f+S} - \mathcal{P}_f \\ &= \bigcup_{k \in K} p_{k,f+S} - p_{k,f}\end{aligned}$$

$$S = \{1, 3, 7, 15\}$$

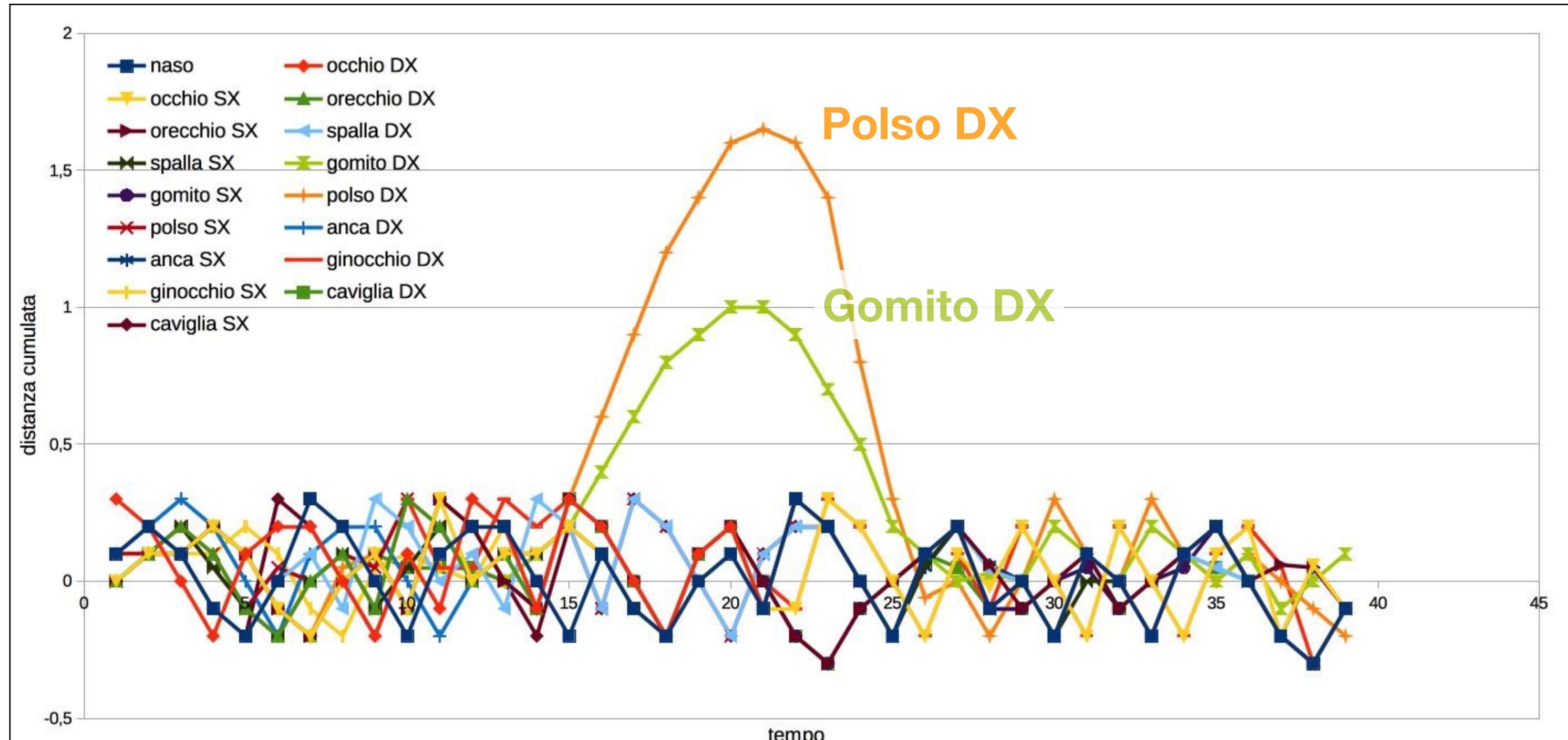
Metodo proposto - Tecniche di rielaborazione

Semplice

Baricentro

Next frame

Distanze cumulate



Evoluzione della posa per l'azione "Alzare e abbassare il braccio destro"

Somma delle differenze di posa
fino al frame f

$$\begin{aligned}\mathcal{T}_f &= \sum_{j=1}^{f-1} \mathcal{P}_{j+1} - \mathcal{P}_j \\ &= \bigcup_{k \in K} \sum_{j=1}^{f-1} p_{k,j+1} - p_{k,j}\end{aligned}$$

Metodo proposto - Tecniche di rielaborazione

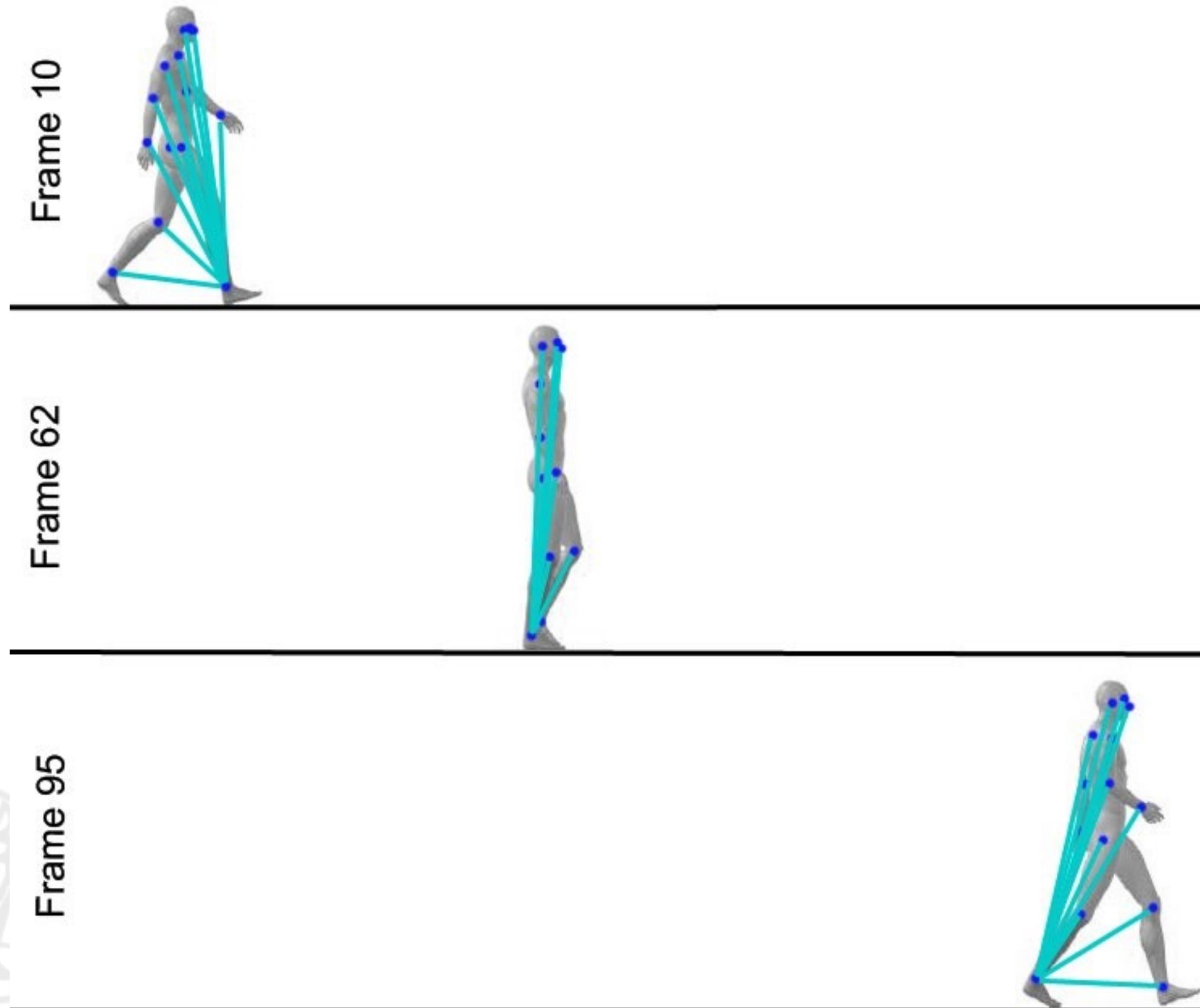
Semplice

Baricentro

Next frame

Distanze cumulate

Distanze relative



Distanza **euclidea** fra ogni giunto e **tutti gli altri** della stessa posa.

$$\begin{aligned} \mathcal{P}_f &\in \mathbb{R}^{K \times 2} \quad \longrightarrow \quad \mathcal{T}_f \in \mathbb{R}^{K \times K} \\ \mathcal{T}_f &= \bigcup_{i \in K} \bigcup_{j \in K} t_{ij} \\ &= \bigcup_{i \in K} \bigcup_{j \in K} \text{eucl}(p_{if}, p_{jf}) \end{aligned}$$

Fase 1 - Esplorazione tecniche



Tecniche di rielaborazione totali: **20** Detectron2 + **20** PoseNet

Database: 8 azioni

- 3 coppie di azioni simili
- 1 azione di coppia

Data augmentation: Mirroring + Jittering

Ottimizzatore: RMSProp

Loss function: Categorical Cross Entropy

$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}} \quad \text{CCE} = - \sum_i^C t_i \cdot \log(f(s)_i)$$

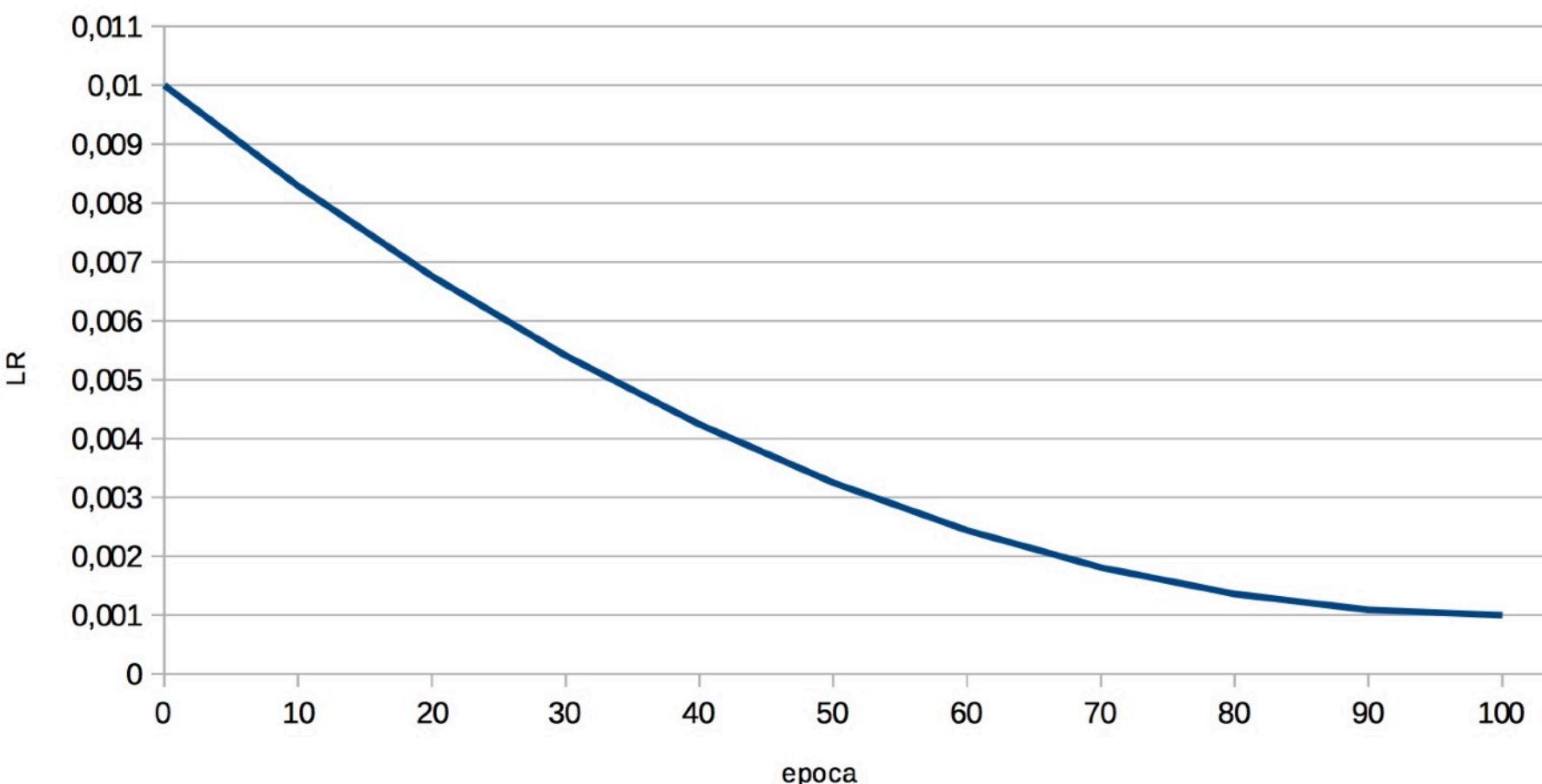
Struttura Rete: 1 livello LSTM

64 hidden units

Batch size: 80

Epoche: 100

LR: Graduale discendente $10^{-2} \rightarrow 10^{-3}$



2 Checkpoint: - Max accuratezza di validazione
- Min loss di validazione

Fase 1 - Esplorazione tecniche



Tecniche di rielaborazione totali: **20** Detectron2 + **20** PoseNet

Database: **8 azioni**

- 3 coppie di azioni simili
- 1 azione di coppia

Data augmentation: Mirroring + Jittering

Ottimizzatore: RMSProp

Loss function: Categorical Cross Entropy

$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}} \quad \text{CCE} = - \sum_i^C t_i \cdot \log(f(s)_i)$$

Struttura Rete: **1 livello LSTM**

64 hidden units

Batch size: 80

TECNICA	1 Livello LSTM					
	DETECTRON			POSENET		
	Media	CV	CS	Media	CV	CS
rimoz 0 + VIDEO + norm	63,53	66,14	60,92	53,88	59,22	48,55
rimoz 0 + FRAME + norm	63,29	65,98	60,60	57,96	59,89	56,02
rimoz 0 + FRAME PERS + norm	62,48	64,04	60,92	58,11	58,78	57,43
rimoz 0 + FRAME + norm (MIN LOSS)	62,17	65,23	59,10	57,18	59,02	55,34
rimoz 0 + NEXT 3 + norm	61,00	62,22	59,78	59,63	62,74	56,52
rimoz 0 + FRAME PERS + norm (MIN LOSS)	60,92	63,05	58,79	55,63	56,69	54,57
rimoz 0 + DIST REL + norm	60,22	64,56	55,89	51,28	48,89	53,67
rimoz 0 + 3BAR + norm	60,17	66,81	53,53	57,39	53,40	61,37
rimoz 0 + 3BAR ASS + norm (MIN LOSS)	59,60	63,85	55,34	62,03	62,46	61,59
rimoz 0 + 3BAR ASS + norm	59,47	64,64	54,30	63,04	61,67	64,40
rimoz 0 + VIDEO + norm (MIN LOSS)	59,31	61,27	57,34	52,53	59,22	45,83
rimoz 0 + 5BAR + norm	59,06	56,29	61,82	60,59	62,34	58,83
rimoz 0 + DIST CUM + norm	58,64	61,75	55,53	32,93	12,50	53,35
rimoz 0 + DIST REL + norm (MIN LOSS)	58,46	61,71	55,21	51,61	49,64	53,58
rimoz 0 + VIDEO PERS + norm (MIN LOSS)	58,12	59,14	57,11	57,87	60,52	55,21
rimoz 0 + 3BAR + norm (MIN LOSS)	57,60	64,44	50,77	55,76	49,96	61,55
rimoz 0 + 17BAR ASS + norm	57,03	62,30	51,77	54,71	58,19	51,22
rimoz 0 + 5BAR + norm (MIN LOSS)	55,41	53,13	57,70	54,83	55,58	54,08
normXY	55,04	60,13	49,96	57,15	55,10	59,19
normXY (MIN LOSS)	54,69	59,65	49,73	52,89	54,55	51,22
norm	53,96	55,66	52,26	53,88	55,46	52,31
norm (MIN LOSS)	53,76	55,62	51,90	53,46	55,10	51,81
rimoz 0 + norm (MIN LOSS)	53,56	53,44	53,67	52,64	53,24	52,04
rimoz 0 + VIDEO PERS + norm	52,65	57,12	48,19	14,99	12,54	17,44
rimoz 0 + DIST CUM + norm (MIN LOSS)	52,41	59,26	45,56	30,16	12,54	47,78
rimoz 0 + norm	52,29	54,98	49,59	51,98	53,24	50,73
rimoz 0 + 5BAR ASS + norm	51,17	47,23	55,12	53,42	53,80	53,03
rimoz 0 + 17BAR ASS + norm (MIN LOSS)	49,93	55,34	44,52	50,52	56,21	44,84
rimoz 0 + 5BAR ASS + norm (MIN LOSS)	48,80	46,56	51,04	47,78	45,69	49,86
rimoz 0 + NEXT 3 + norm (MIN LOSS)	47,16	46,72	47,60	12,57	12,54	12,59
Semplice	39,49	43,47	35,51	42,09	43,51	40,67
Semplice (MIN LOSS)	39,38	43,16	35,60	42,21	42,88	41,53
rimoz 0 + 17BAR + norm	34,96	12,54	57,38	33,18	12,50	53,85
rimoz 0 + 17BAR + norm (MIN LOSS)	33,99	12,54	55,44	32,86	12,54	53,17
rimoz 0 + NEXT 7 + norm	12,54	12,58	12,50	29,50	13,85	45,15
rimoz 0 + NEXT 15 + norm	12,52	12,54	12,50	21,40	30,30	12,50
rimoz 0 + NEXT 15 + norm (MIN LOSS)	12,52	12,54	12,50	21,48	30,46	12,50
rimoz 0 + NEXT 7 + norm (MIN LOSS)	12,52	12,54	12,50	25,40	13,85	36,96
rimoz 0 + NEXT 7 + norm (MIN LOSS)	12,52	12,54	12,50	12,52	12,50	12,55

Fase 2 - Esplorazione livelli

Tecniche di rielaborazione totali: **9** Detectron2 + **9** PoseNet

Database: 8 azioni

- 3 coppie di azioni simili
- 1 azione di coppia

Data augmentation: Mirroring + Jittering

Struttura Rete: 1-2-3 livelli LSTM

64 hidden units

Batch size: 80

TECNICA	1 Livello LSTM						2 Livelli LSTM						3 Livelli LSTM					
	DETECTRON			POSENET			DETECTRON			POSENET			DETECTRON			POSENET		
	Media	CV	CS	Media	CV	CS	Media	CV	CS	Media	CV	CS	Media	CV	CS	Media	CV	CS
rimoz 0 + VIDEO + norm	63,53	66,14	60,92	53,88	59,22	48,55	60,43	61,71	59,15	57,00	57,83	56,16	59,24	62,82	55,66	56,86	60,09	53,62
rimoz 0 + FRAME + norm	63,29	65,98	60,60	57,96	59,89	56,02	63,43	64,72	62,14	56,91	58,03	55,80	33,02	12,50	53,53	12,50	12,50	12,50
rimoz 0 + FRAME PERS + norm	62,48	64,04	60,92	58,11	58,78	57,43	57,52	59,65	55,39	57,88	59,65	56,11	36,16	59,81	12,50	55,30	56,80	53,80
rimoz 0 + FRAME + norm (MIN LOSS)	62,17	65,23	59,10	57,18	59,02	55,34	61,28	62,10	60,46	55,28	57,20	53,35	32,77	12,50	53,03	12,50	12,50	12,50
rimoz 0 + NEXT 3 + norm	61,00	62,22	59,78	59,63	62,74	56,52	60,83	64,91	56,75	58,52	58,35	58,70	61,14	61,91	60,37	32,81	12,50	53,13
rimoz 0 + FRAME PERS + norm (MIN LOSS)	60,92	63,05	58,79	55,63	56,69	54,57	56,71	59,61	53,80	56,76	56,05	57,47	33,90	55,30	12,50	54,77	56,69	52,85
rimoz 0 + DIST REL + norm	60,22	64,56	55,89	51,28	48,89	53,67	59,88	59,97	59,78	55,96	57,16	54,76	59,00	62,66	55,34	58,64	62,66	54,62
rimoz 0 + 3BAR + norm	60,17	66,81	53,53	57,39	53,40	61,37	62,63	61,71	63,54	60,35	62,74	57,97	59,44	61,27	57,61	58,64	58,03	59,24
rimoz 0 + 3BAR ASS + norm (MIN LOSS)	59,60	63,85	55,34	62,03	62,46	61,59	59,11	62,07	56,16	61,70	61,99	61,41	61,63	60,72	62,55	59,18	60,84	57,52
rimoz 0 + 3BAR ASS + norm	59,47	64,64	54,30	63,04	61,67	64,40	61,86	59,42	64,31	60,89	61,59	60,19	62,22	61,35	63,09	61,22	60,52	61,91
rimoz 0 + VIDEO + norm (MIN LOSS)	59,31	61,27	57,34	52,53	59,22	45,83	59,49	60,64	58,33	55,54	55,38	55,71	59,32	60,52	58,11	55,04	57,99	52,08
rimoz 0 + 5BAR + norm	59,06	56,29	61,82	60,59	62,34	58,83	63,92	68,87	58,97	57,86	56,21	59,51	34,85	12,58	57,11	33,13	53,76	12,50
rimoz 0 + DIST CUM + norm	58,64	61,75	55,53	32,93	12,50	53,35	56,70	58,82	54,57	56,33	57,67	54,98	36,00	59,49	12,50	54,45	54,91	53,99
rimoz 0 + DIST REL + norm (MIN LOSS)	58,46	61,71	55,21	51,61	49,64	53,58	59,84	59,53	60,15	56,25	57,16	55,34	57,34	53,06	55,62	57,60	60,80	54,39
rimoz 0 + VIDEO PERS + norm (MIN LOSS)	58,12	59,14	57,11	57,87	60,52	55,21	60,83	64,91	56,75	58,52	58,35	58,70	61,14	61,91	60,37	32,81	12,50	53,13
rimoz 0 + 3BAR + norm (MIN LOSS)	57,60	64,44	50,77	55,76	49,96	61,55	60,10	60,32	59,87	58,37	62,03	54,71	56,30	56,88	55,71	58,06	58,47	57,65

Fase 3 - Dropout e Regolarizzatore

Tecniche di rielaborazione totali: **9 Detectron2 + 9 PoseNet**

Database: 60 azioni

Data augmentation: Mirroring + Jittering

Epoche: 100

LR: 10^{-3}

Struttura Rete: 2-3 livelli LSTM
64 hidden units

Batch size: 600

Dropout: 5% - 50%

Regolarizzatore: $10^{-1} - 10^{-8}$

Detectron2

Media	Cross view	Cross Subject	Nome	Drop (%)	Reg	Liv LSTM
84,30	88,89	79,71	3 BAR	15	10^{-5}	3
84,23	88,66	79,81	3 BAR	15		3
83,86	87,64	80,08	NEXT 3	15	10^{-5}	3
83,76	88,54	78,98	3 BAR (ML)	15		3
83,56	88,04	79,08	NEXT 3	10		3
:	:	:	:	:	:	:

PoseNet

Media	Cross view	Cross Subject	Nome	Drop (%)	Reg	Liv LSTM
76,05	80,09	72,00	3 BAR	15		3
75,99	79,82	72,15	NEXT 3	15	10^{-6}	3
75,84	78,91	72,77	3 BAR (ML)	10	10^{-6}	3
75,83	78,69	72,97	NEXT 3	10	10^{-6}	3
75,81	79,95	71,68	5 BAR	5	10^{-6}	3
:	:	:	:	:	:	:

Fase 4 - Addestramento completo

Tecniche di rielaborazione totali: **6** Detectron2 + **6** PoseNet

Database: 60 azioni

Data augmentation: Mirroring + Jittering

Epoche: ∞

LR: 10^{-3}

Struttura Rete: 2-3 livelli LSTM

64 hidden units

Batch size: 600

Dropout: 5% - 50%

Regolarizzatore: 10^{-1} - 10^{-8}

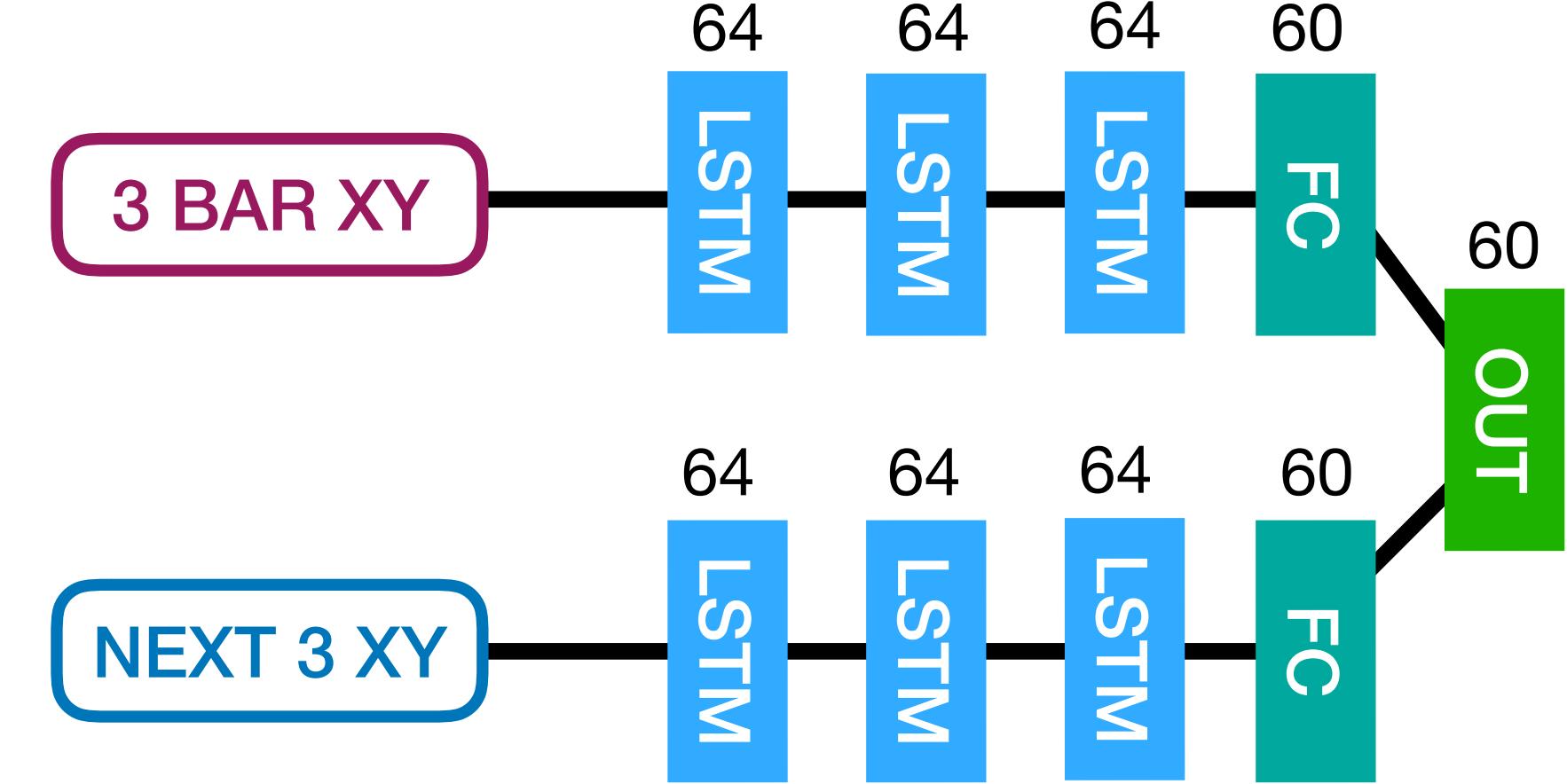
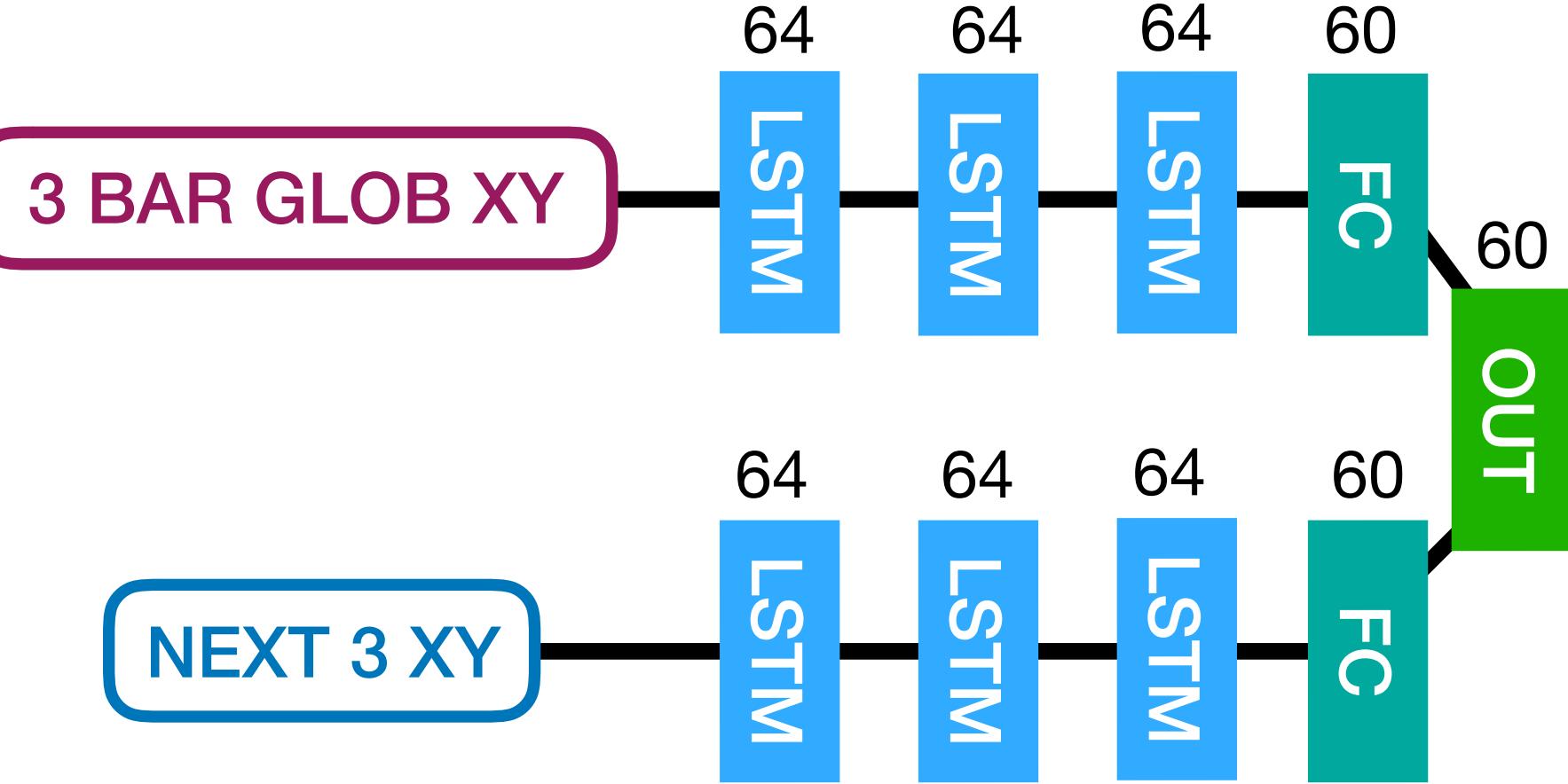
Detectron2

Media	Cross view	Cross Subject	Nome	Drop (%)	Reg	Liv LSTM
86,84	91,89	81,78	3 BAR GLOB XY	15	10^{-5}	3
86,62	91,72	81,52	3 BAR GLOB	15	10^{-5}	3
86,10	90,72	81,47	NEXT 3 XY	15	10^{-5}	3
85,70	89,80	81,59	3 BAR	15		3
85,35	89,22	80,74	NEXT 3	15	10^{-5}	3
85,18	89,02	81,33	3 BAR XY	15		3

PoseNet

Media	Cross view	Cross Subject	Nome	Drop (%)	Reg	Liv LSTM
82,51	87,47	77,55	3 BAR XY	10		3
82,15	86,98	77,32	3BAR	10		3
81,55	85,60	77,50	3 BAR GLOB	15		3
80,01	84,64	75,38	3 BAR GLOB XY	15		3
78,90	82,74	75,06	NEXT 3 XY	15	10^{-5}	3
77,64	81,74	73,53	NEXT 3	15	10^{-5}	3

Fase 5 - Combinazione



Detectron2						
Media	Cross view	Cross Subject	Nome	Drop (%)	Reg	Liv LSTM
88,51 (+1.67)	93,69	83,32	3 BAR GLOB XY + NEXT 3 XY	15	10 ⁻⁵	3
85,05	90,23	79,87	3 BAR GLOB XY + 3 BAR XY	15	10 ⁻⁵	3

PoseNet						
Media	Cross view	Cross Subject	Nome	Drop (%)	Reg	Liv LSTM
83,43 (+0,92)	87,79	79,06	3 BAR XY + NEXT 3 XY	10		3
81,56	85,39	77,74	3 BAR GLOB XY + NEXT 3 XY	15	10 ⁻⁵	3

Loss function:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_1 + \alpha \cdot \mathcal{L}_2$$

\mathcal{L}_1 = loss function
prima tecnica

\mathcal{L}_2 = loss function
seconda tecnica

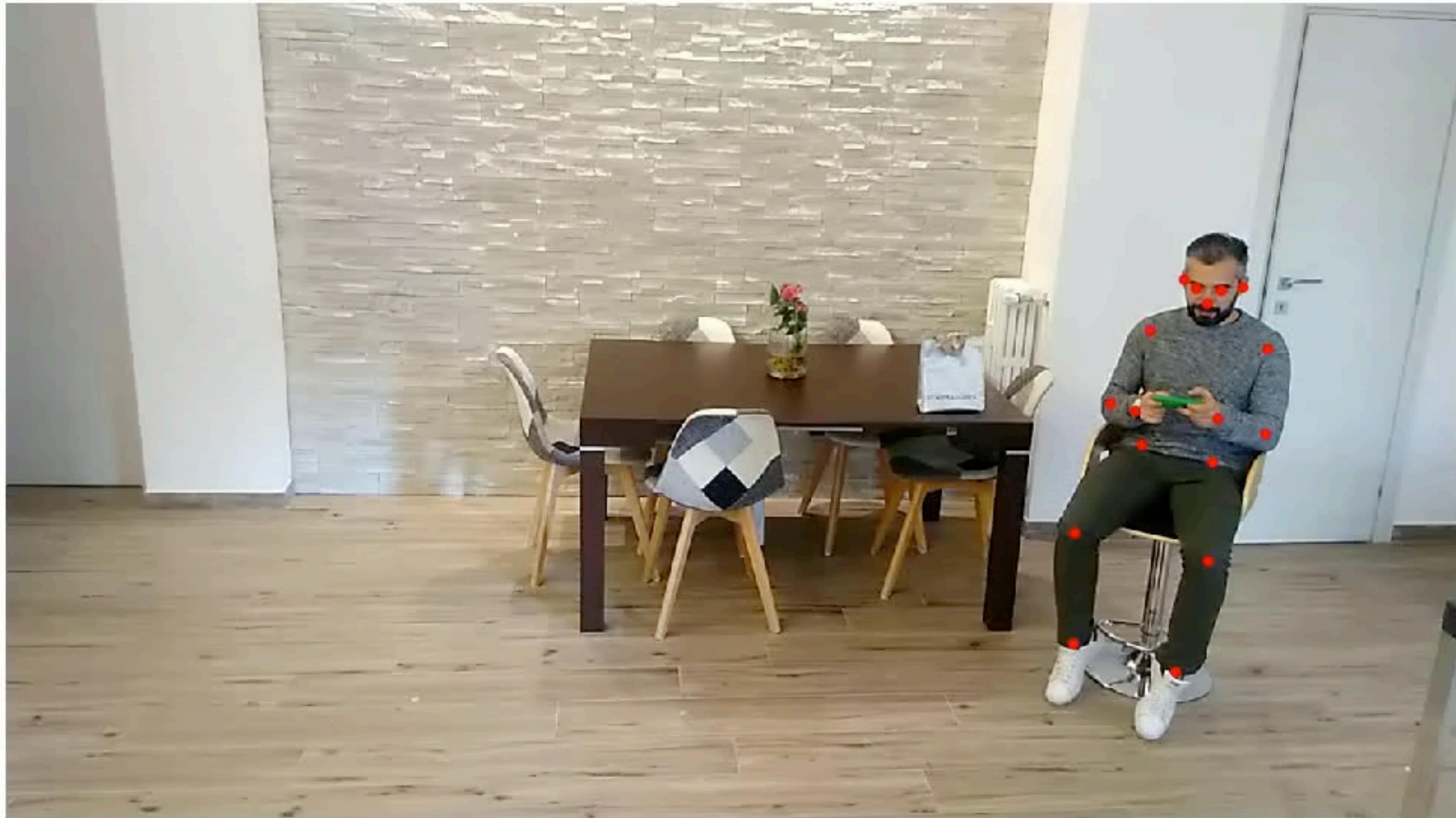
$$\alpha = 0.5$$

Conclusioni



Articolo	Posa	RGB	IR	Cross View (%)	Cross subject (%)	Media
Action machine		X		97,2	94,3	95,75
		X		93,2	86,6	89,90
		X		-	85,5	-
3BAR-NEXT-Detectron		X		93,69	83,32	88,51
Chained		X		-	80,8	-
3BAR-NEXT-PoseNet		X		87,79	79,06	83,43
DSSCA-SSLM		X		-	74,9	-
	X	X		-	91,99	-
	X	X		95,2	91,7	93,45
	X		X	94,5	91,6	93,05
	X	X		-	90,04	-
	X			93,2	87,5	90,35
	X			92,4	84,8	88,60
	X			88,3	81,5	84,90
	X			87,6	79,4	83,5
	X			84,8	79,6	82,2
	X			84,7	76,5	80,6
	X			83,1	74,3	78,78
	X			77,7	69,2	73,45
	X			70,3	62,9	66,60
	X			67,3	60,7	64,00
	X			64,0	59,1	61,55
Lie Group	X			52,8	50,1	51,45

Test sul campo





UNIVERSITÀ
DEGLI STUDI
FIRENZE

Grazie!

Spare slides

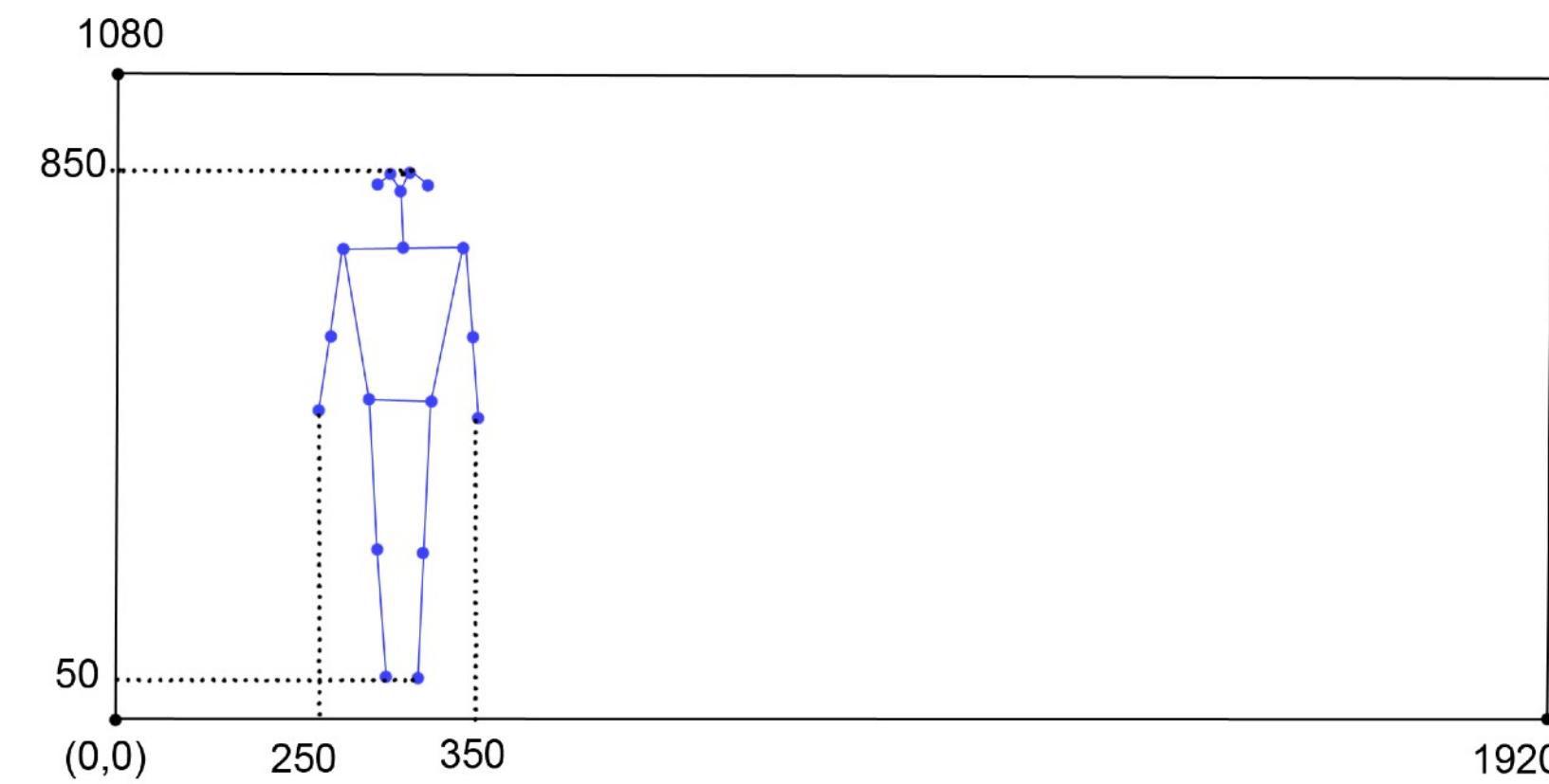
Sviluppi futuri

- Utilizzare le informazioni provenienti anche da una classificazione ottenuta direttamente dai dati RGB
- Testare funzioni loss combinate diverse fra le migliori tecniche (massimo score, α diverso da 0.5, ...)
- Testare l'algoritmo anche sui dataset *NTU-RGB+D120*, *MSR Daily Activity3D* e *Northwestern-UCLA Multiview Action 3D* (NUCLA)

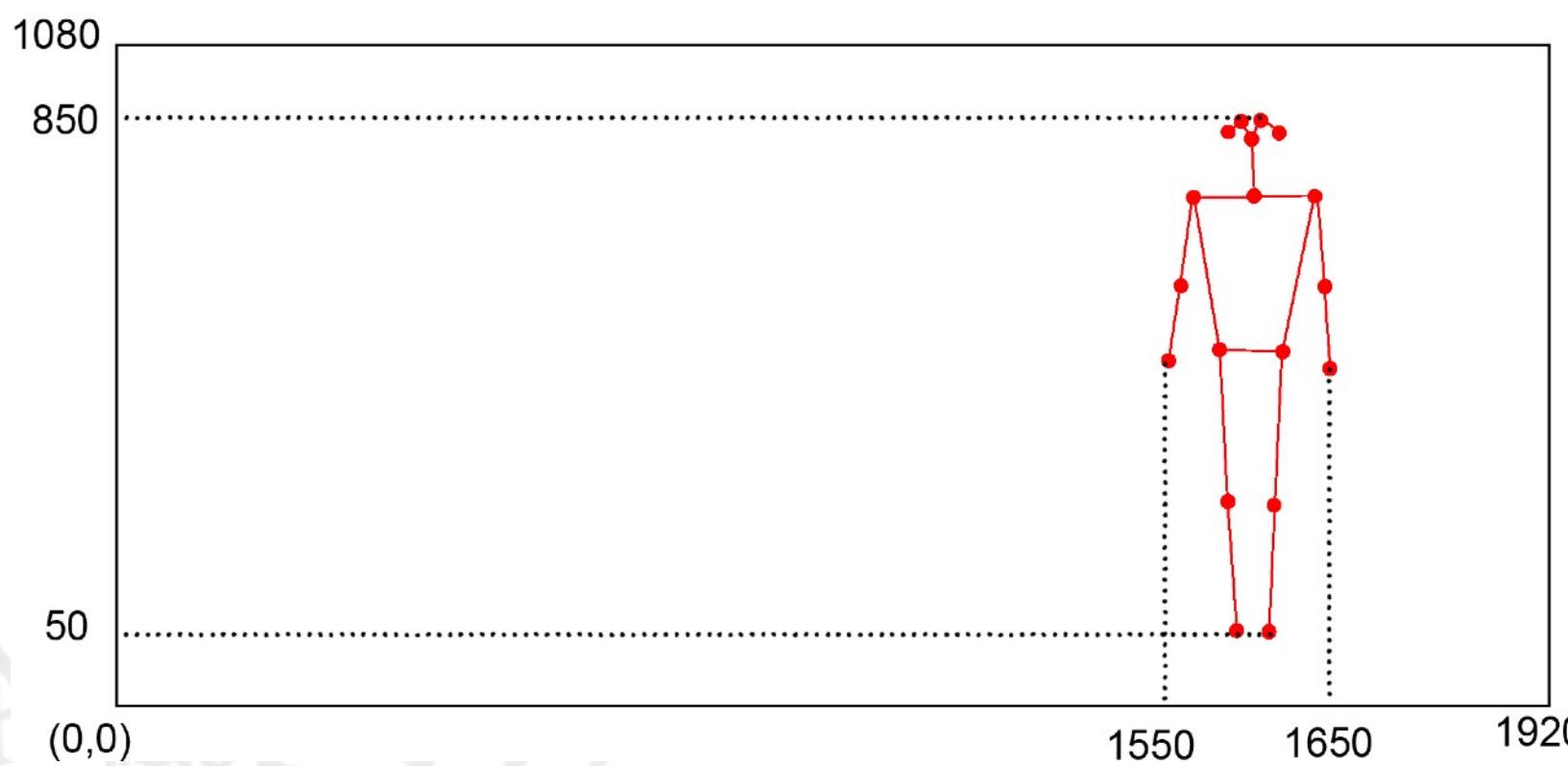
- | | | |
|---|--|--|
| 1 - drink water | 21 - take off a hat/cap | 41 - sneeze/cough |
| 2 - eat meal/snack | 22 - cheer up | 42 - staggering |
| 3 - brushing teeth | 23 - hand waving | 43 - falling |
| 4 - brushing hair | 24 - kicking something | 44 - touch head (headache) |
| 5 - drop | 25 - reach into pocket | 45 - touch chest (stomachache/heart pain) |
| 6 - pickup | 26 - hopping (one foot jumping) | 46 - touch back (backache) |
| 7 - throw | 27 - jump up | 47 - touch neck (neckache) |
| 8 - sitting down | 28 - make a phone call/answer phone | 48 - nausea or vomiting condition |
| 9 - standing up (from sitting position) | 29 - playing with phone/tablet | 49 - use a fan (with hand or paper)/feeling warm |
| 10 - clapping | 30 - typing on a keyboard | 50 - punching/slapping other person |
| 11 - reading | 31 - pointing to something with finger | 51 - kicking other person |
| 12 - writing | 32 - taking a selfie | 52 - pushing other person |
| 13 - tear up paper | 33 - check time (from watch) | 53 - pat on back of other person |
| 14 - wear jacket | 34 - rub two hands together | 54 - point finger at the other person |
| 15 - take off jacket | 35 - nod head/bow | 55 - hugging other person |
| 16 - wear a shoe | 36 - shake head | 56 - giving something to other person |
| 17 - take off a shoe | 37 - wipe face | 57 - touch other person's pocket |
| 18 - wear on glasses | 38 - salute | 58 - handshaking |
| 19 - take off glasses | 39 - put the palms together | 59 - walking towards each other |
| 20 - put on a hat/cap | 40 - cross hands in front (say stop) | 60 - walking apart from each other. |

Metodo proposto - Tecniche di rielaborazione

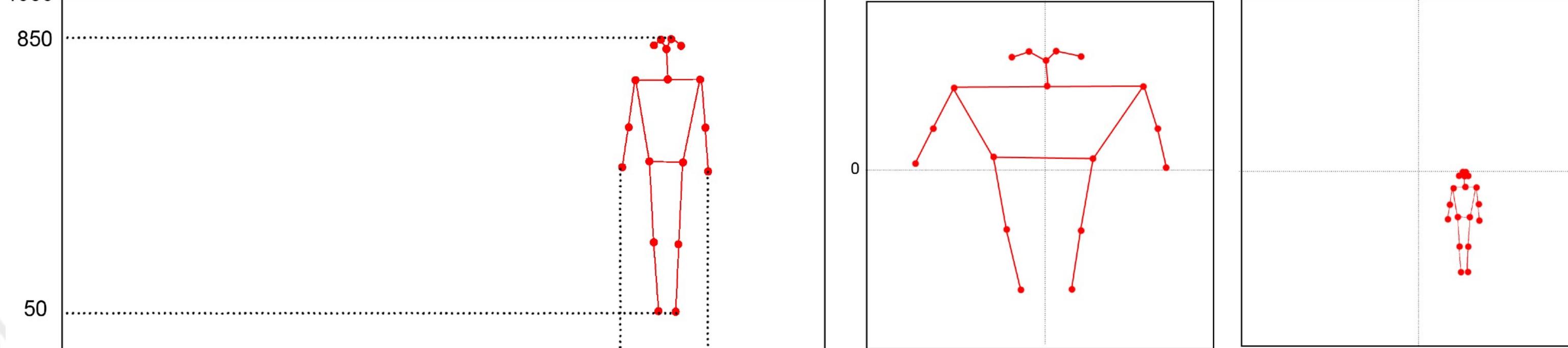
Normalizzazione



Norm SEPARATA



Norm GLOBALE



Normalizzazione Separata:

le dimensioni in input vengono normalizzate separatamente.

Particolarmente adatta quando l'input rappresenta **punti nel piano o vogliamo separare l'importanza delle dimensioni**

Normalizzazione Globale:

le dimensioni in input vengono normalizzate globalmente

Metodo proposto - Validation set

Cross Subject

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

Training

setup: 80%
video: 32256

Validation

setup: 20%
video: 8064

Test

id attori: tutti gli altri video: 16560

Cross View

id camera: 2,3

Training

prospettive frontali: 40%
prospettive laterali: 100%
video: 30336

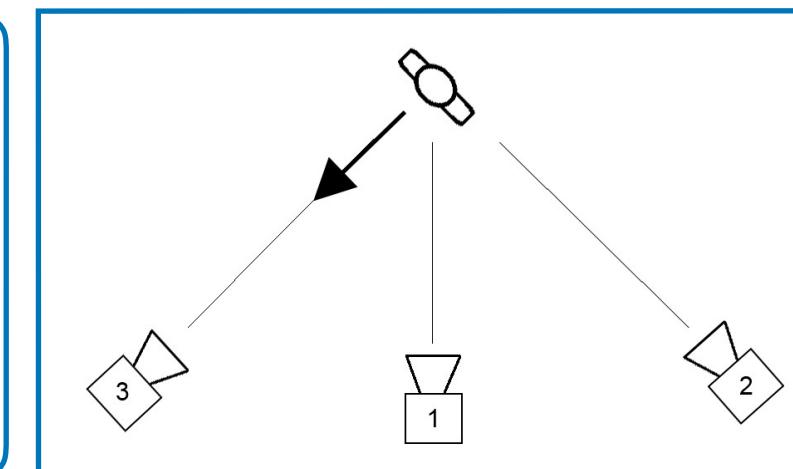
Validation

prospettive frontali: 60%
prospettive laterali: 0%
video: 7584

Test

id camera: 1 video: 18960

Ripetizione 1



Ripetizione 2

