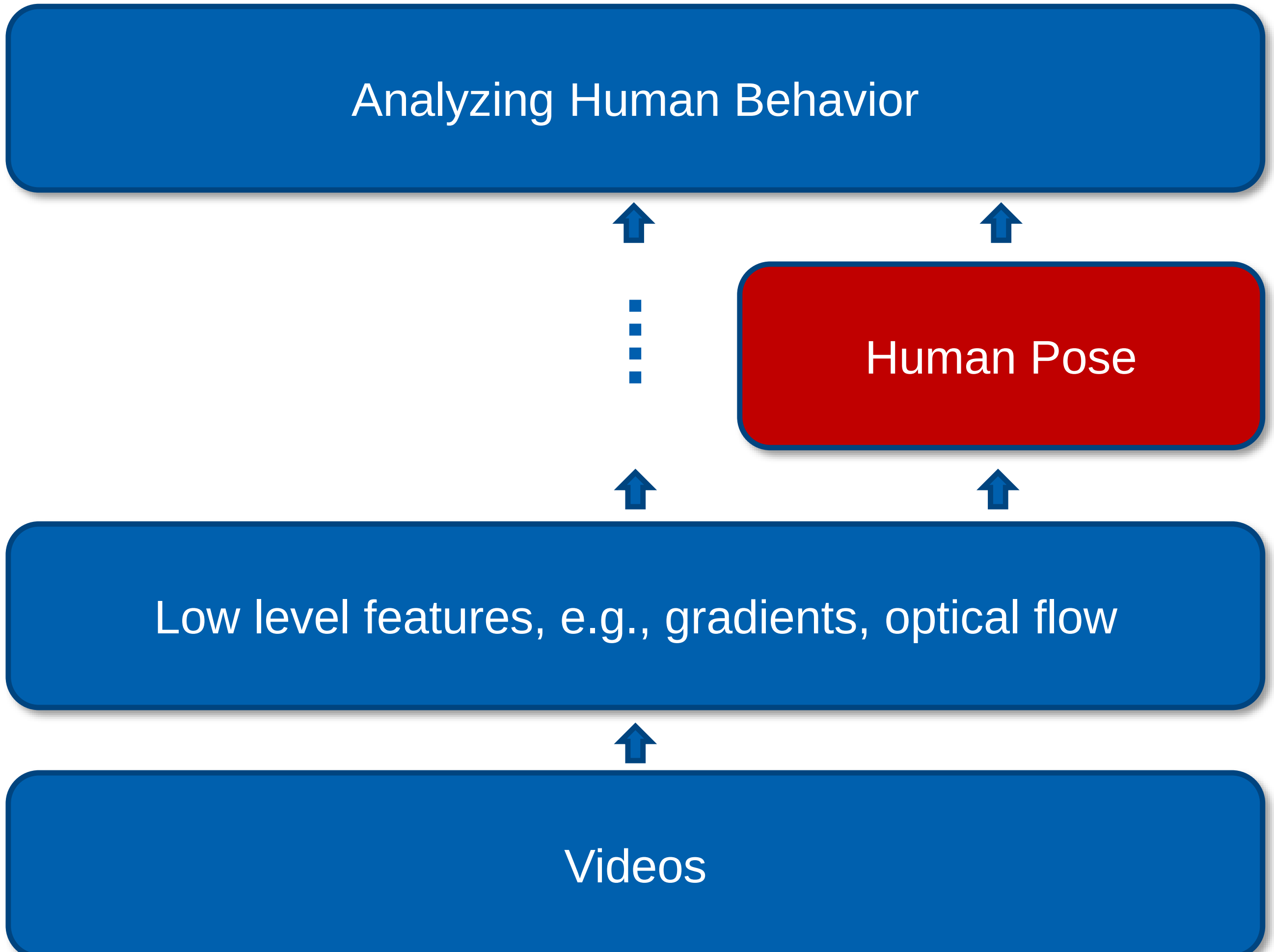
Juergen Gall

Analyzing Human Behavior in Video Sequences

Analyzing Human Behavior



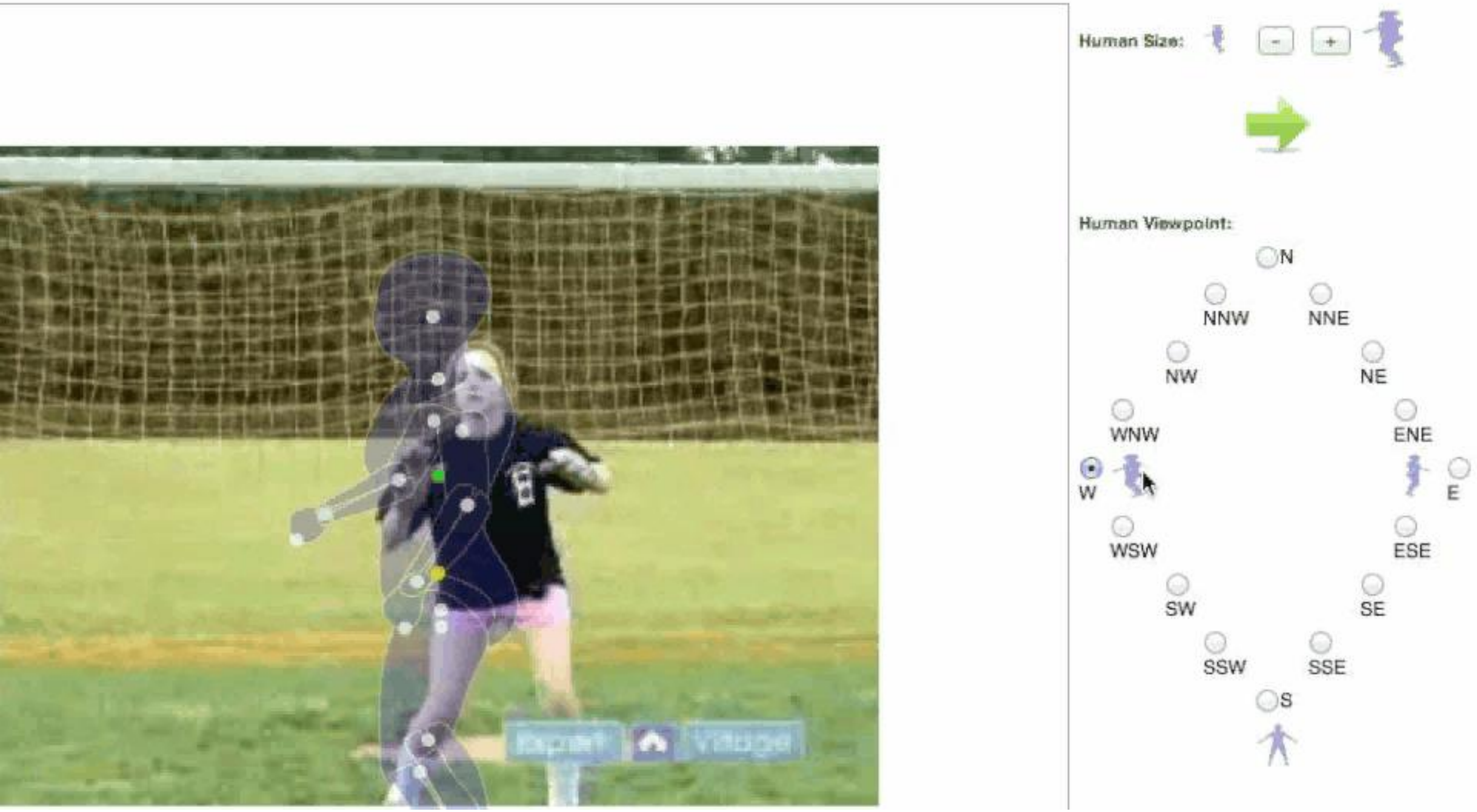
21 Actions from HMDB

HMDB51 (Kuehne et al, ICCV 2011)

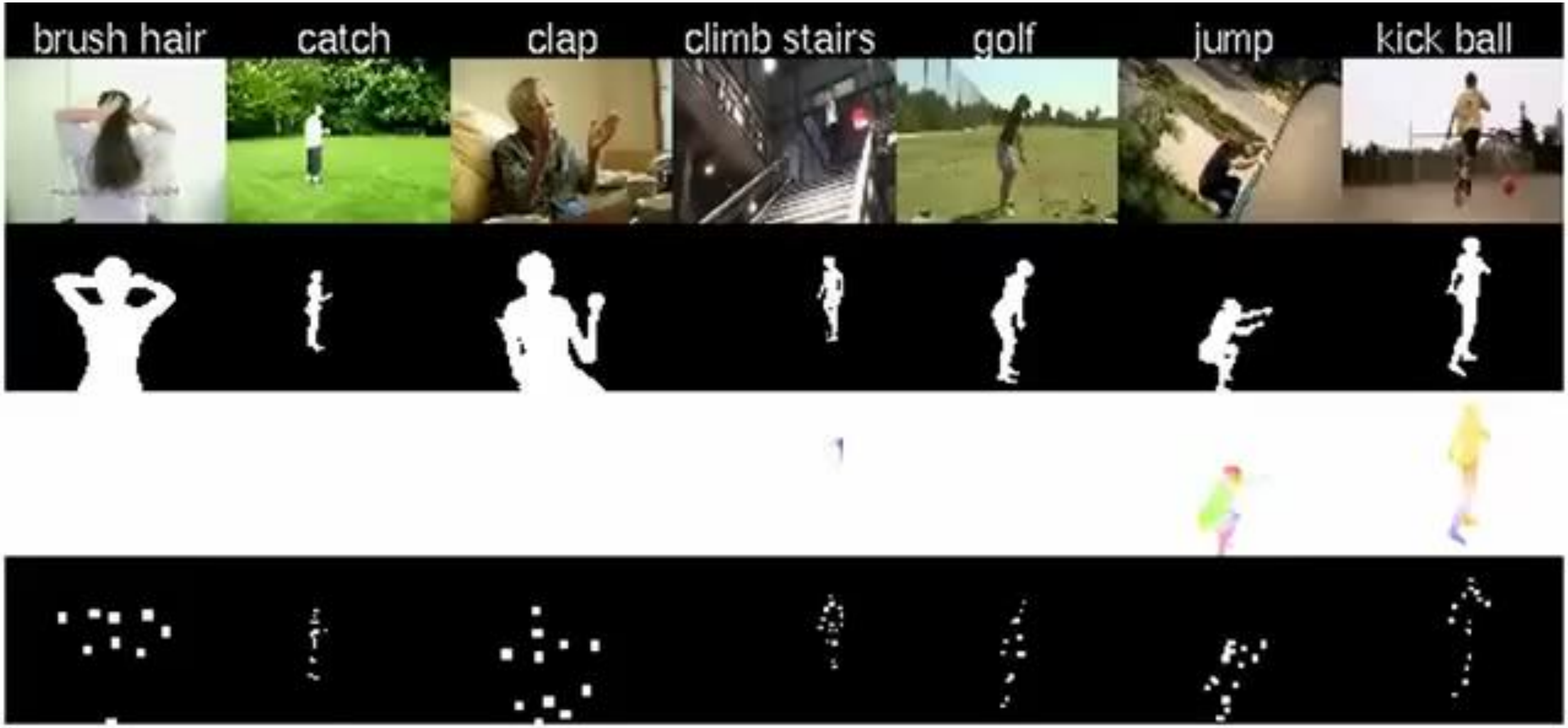
928 clips, 33183 frames

Puppet Annotation

universitätbonn

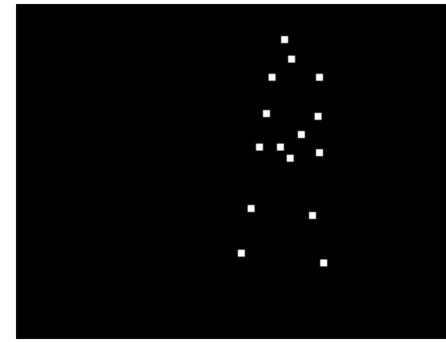


Joint-annotated **HMDB** (JHMDB)



[H. Jhuang et al. **Towards Understanding Action Recognition.** ICCV 2013]
[http://jhmdb.is.tue.mpg.de]

Study with Annotated Data (2013)

	Low	Mid	High
			
baseline	given puppet flow	given puppet mask	given joint positions

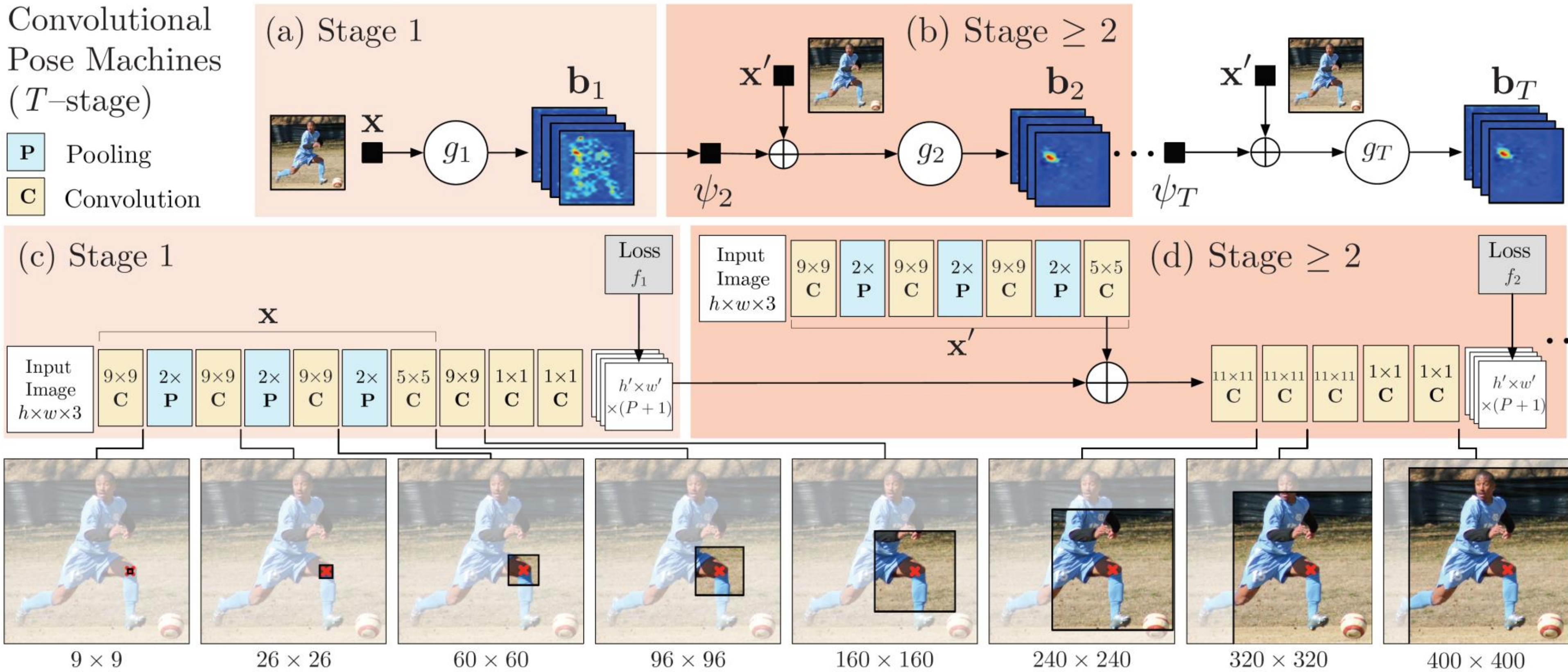
	baseline	given flow	given mask	pose features
GT		+ ~11%	+ ~9%	+ ~20%

- Large potential gain for pose feature
- Not with existing 2d human pose methods

[H. Jhuang et al. **Towards Understanding Action Recognition.** ICCV 2013]
[http://jhmdb.is.tue.mpg.de]

CNNs for Pose Estimation

Stack CNNs:



[S.-E. Wei et al. **Convolutional Pose Machines.** CVPR 2016]

Coupled Action Recognition and Pose Estimation

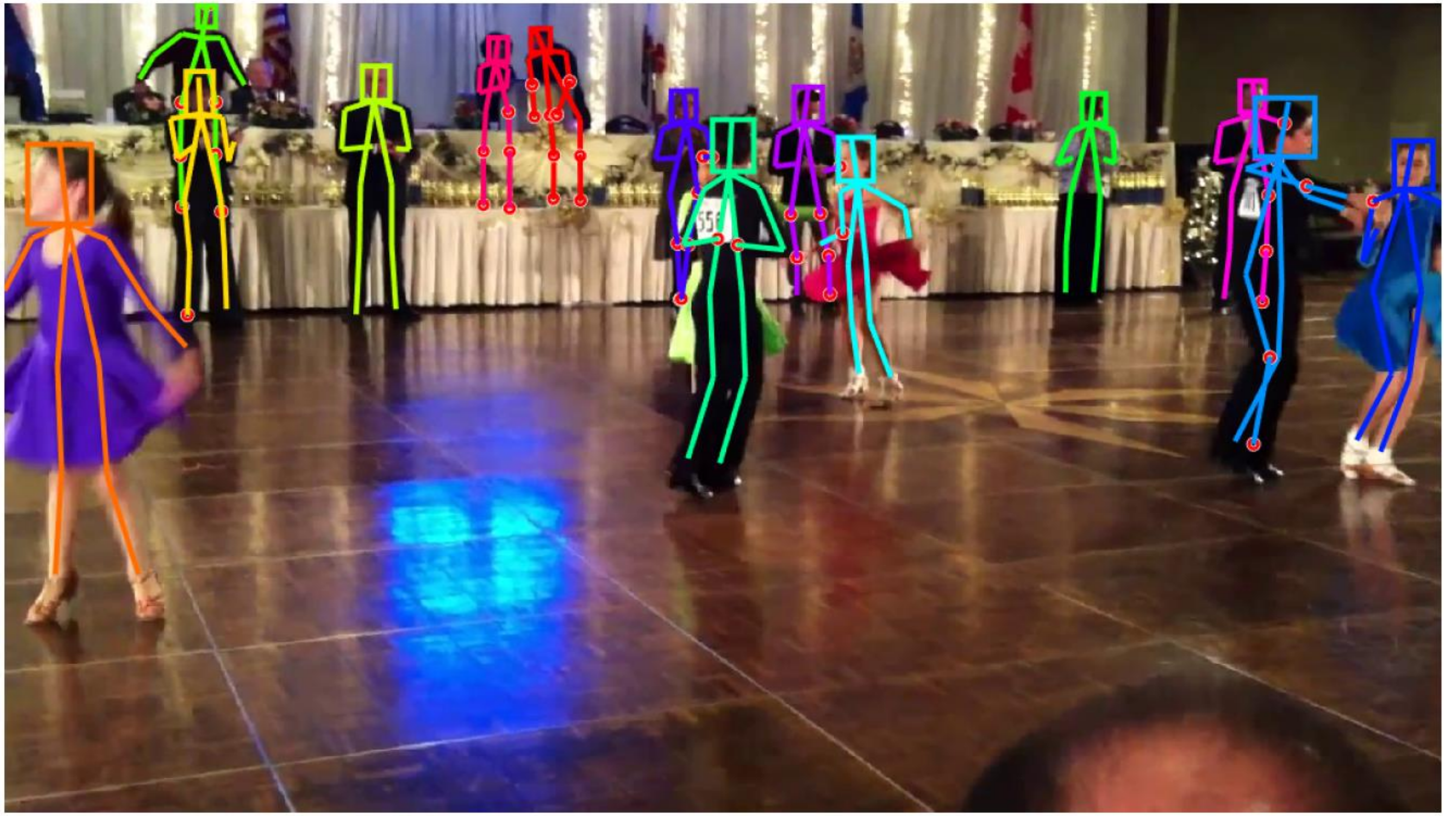
universitätbonn

Method	sub-J-HMDB	Penn-Action
Appearance features only		
Dense [19]	46.0%	—
IDT-FV [55]	60.9%	92.0%
Pose features only		
Pose [19]	54.1%	—
Pose (Ours)	61.5%	79.0%
Pose + Appearance features		
MST [18]	45.3%	74.0%
Pose+Dense [19]	52.9%	—
AOG [15]	61.2%	85.5%
P-CNN [45]	66.8%	—
Pose (Ours)+IDT-FV	**74.6%**	**92.9%**

[U. Iqbal et al. **Pose for Action – Action for Pose.** FG 2017]

Pose Estimation in Videos

Video datasets for human pose in unconstrained videos does not exist.



[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Estimation in Videos

Video datasets for human pose in unconstrained videos does not exist.

Unconstrained means

- Public available content from the Internet (e.g. Youtube)
- Multiple persons in a video (no assumption about position)
- Arbitrary number of visible joints (truncation and occlusion)
- Large scale variations (unknown scale)

[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

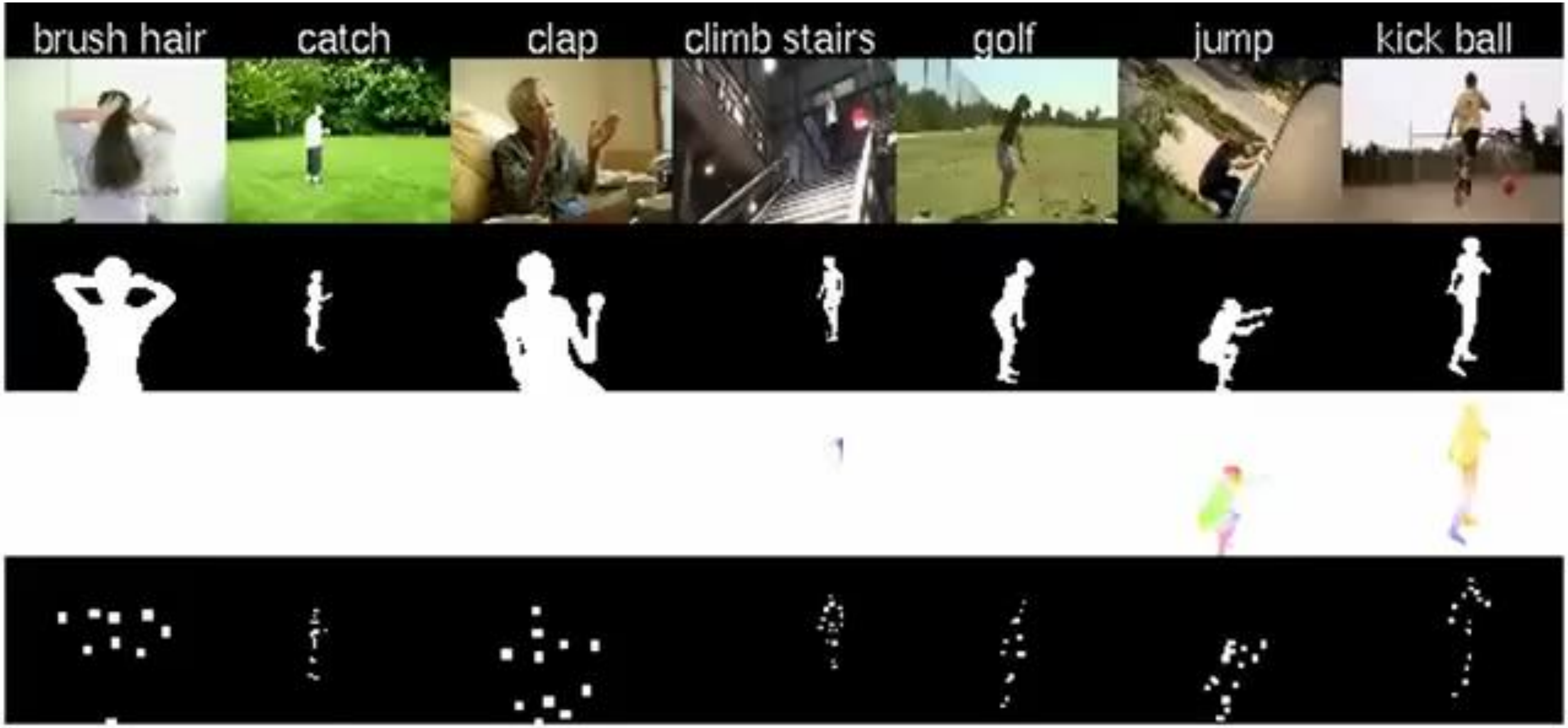
Pose-Track Dataset

universitätbonn

Dataset	videos	multi-person	Large scale variation	variable skeleton size	# of Persons
Leeds Sports [21]					2000
MPII Pose [1]			✓	✓	40,522
We Are Family [12]		✓			3131
MPII Multi-Person Pose [30]		✓	✓	✓	14,161
MS-COCO Keypoints [25]		✓	✓	✓	105,698

[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Joint-annotated **HMDB** (JHMDB)



[H. Jhuang et al. **Towards Understanding Action Recognition.** ICCV 2013]
[http://jhmdb.is.tue.mpg.de]

Pose-Track Dataset

Dataset	videos	multi-person	Large scale variation	variable skeleton size	# of Persons
J-HMDB [20]	✓		✓	✓	32,173
Penn-Action [45]	✓		✓		159,633
VideoPose [35]	✓				1286
Poses-in-the-wild [10]	✓				831
YouTube Pose [8]	✓				5000
FYDP [36]	✓				1680
UYDP [36]	✓				2000
Multi-Person Pose-Track	✓	✓	✓	✓	16,219

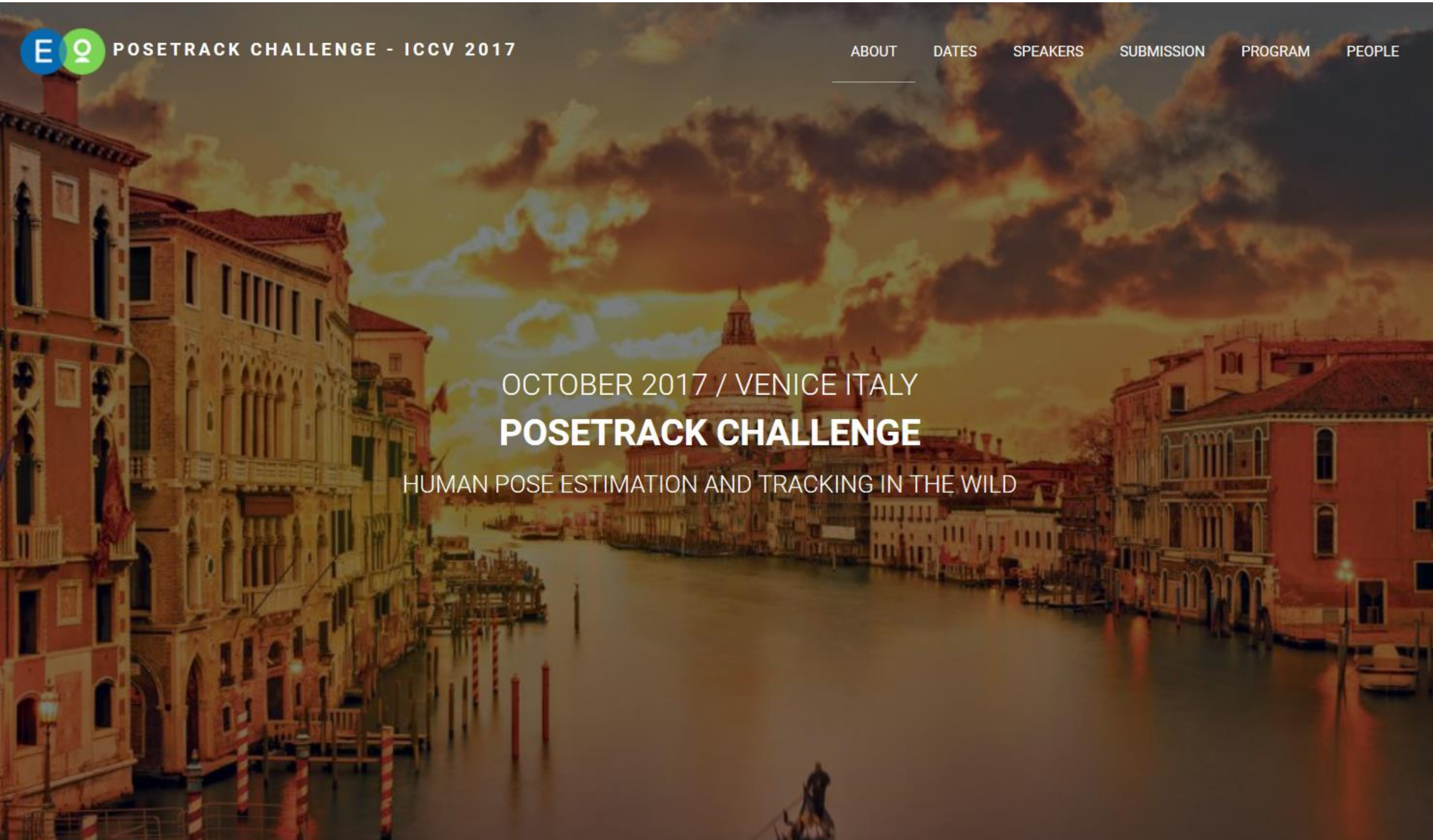
[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose-Track Dataset

Multi-Person Pose-Track Dataset

\# of videos = 60
Training = 30
Testing = 30
\# of annotated persons = 16,219

Challenge ICCV 2017



[http://posetrack.net/workshops/iccv2017]

Pose Track: Simultaneous Pose Estimation and Tracking

Estimate pose + person association over time:

[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking

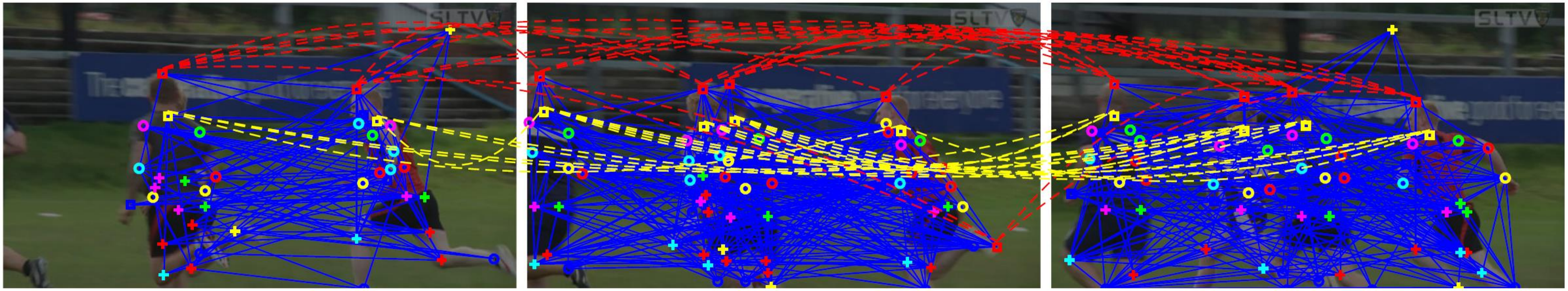
Estimate pose + person association over time:

- Predict body joints (CNN trained on MPII Pose)

Pose Track: Simultaneous Pose Estimation and Tracking

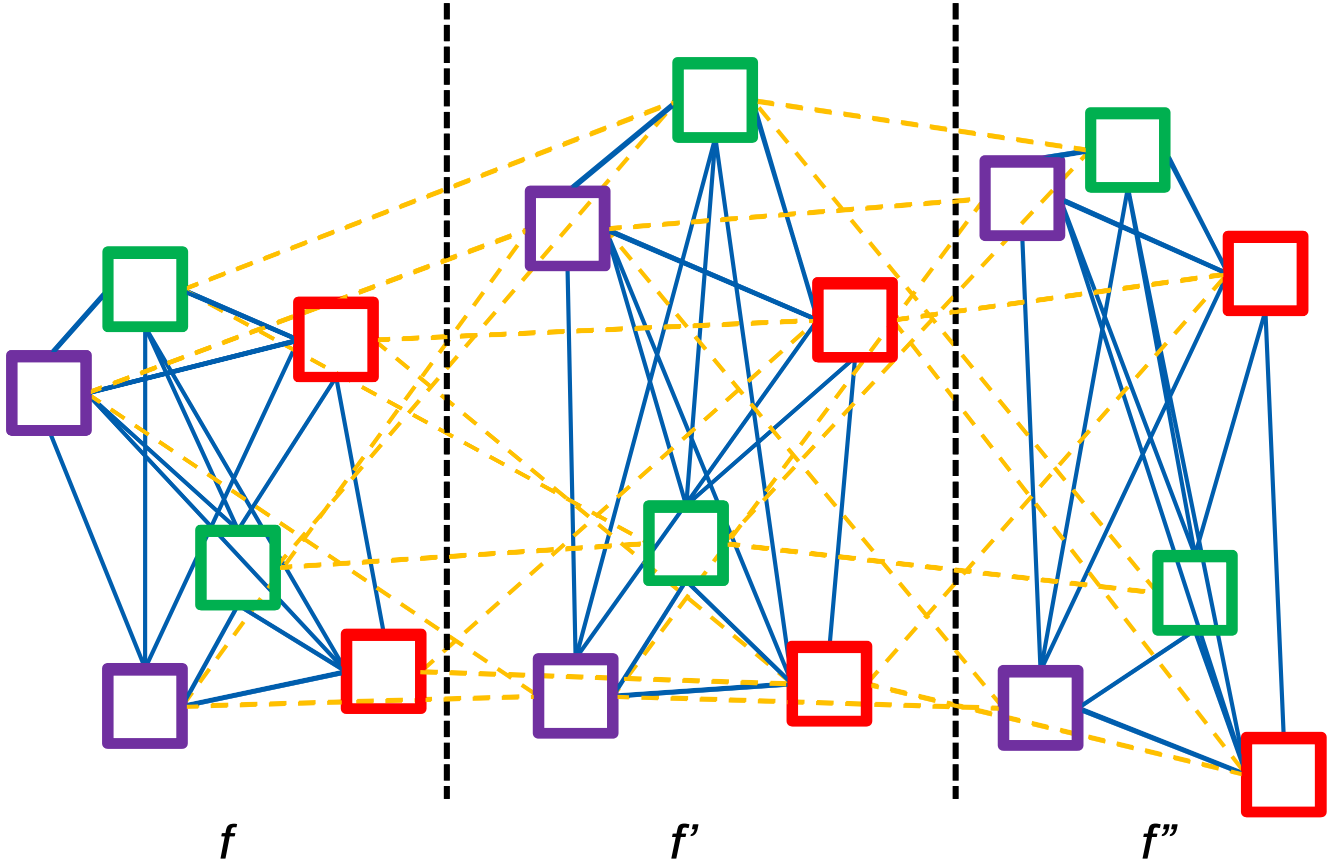
Estimate pose + person association over time:

- Predict body joints (CNN trained on MPII Pose)
- Build a graph with temporal and spatial edges



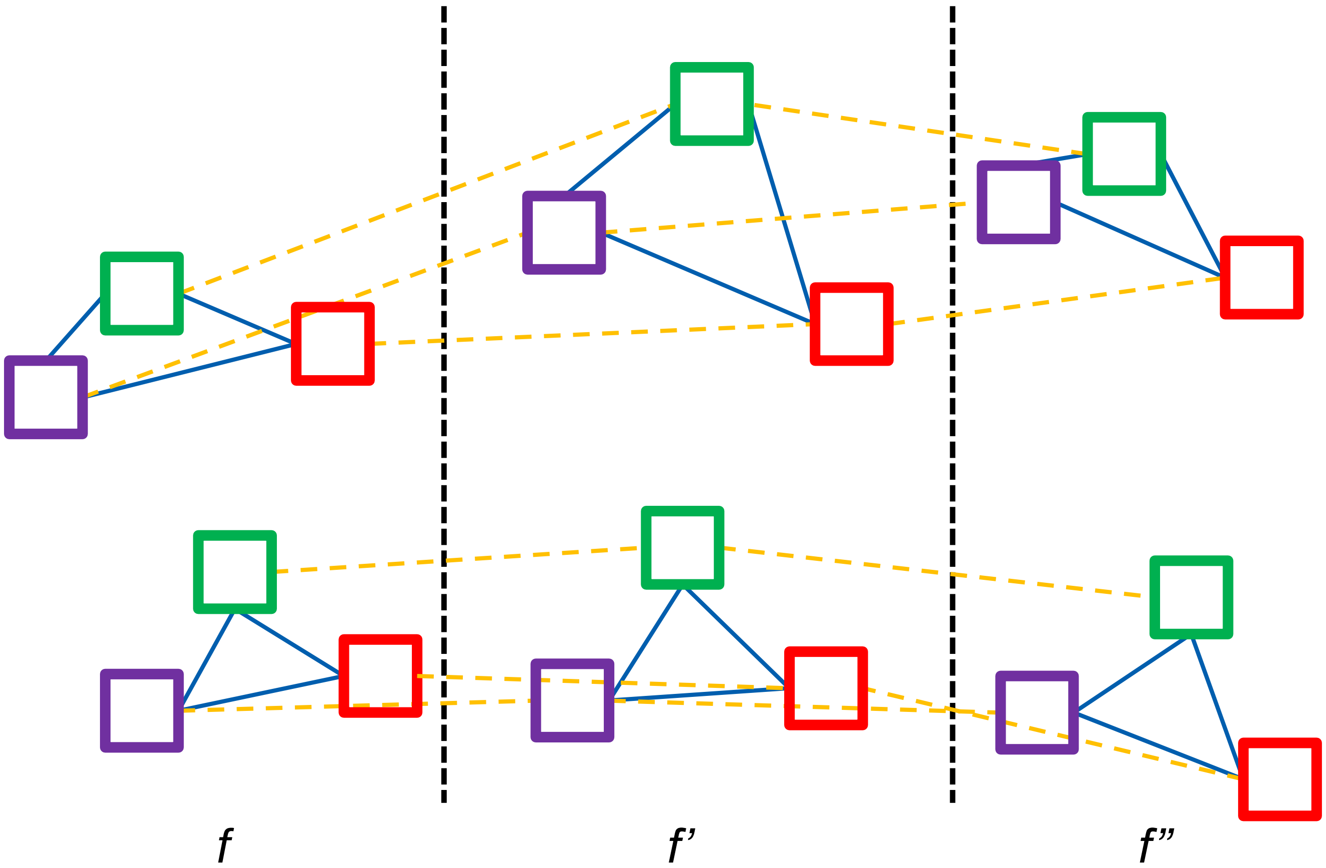
[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking



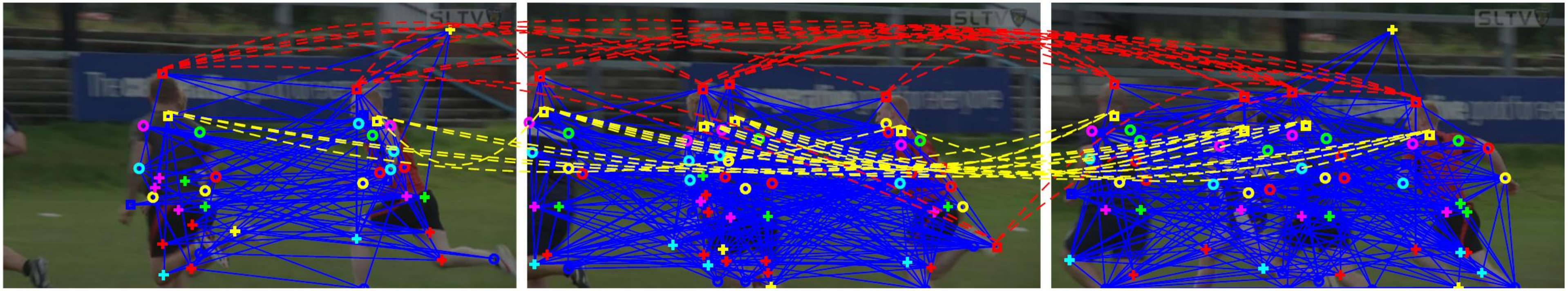
[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking



[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

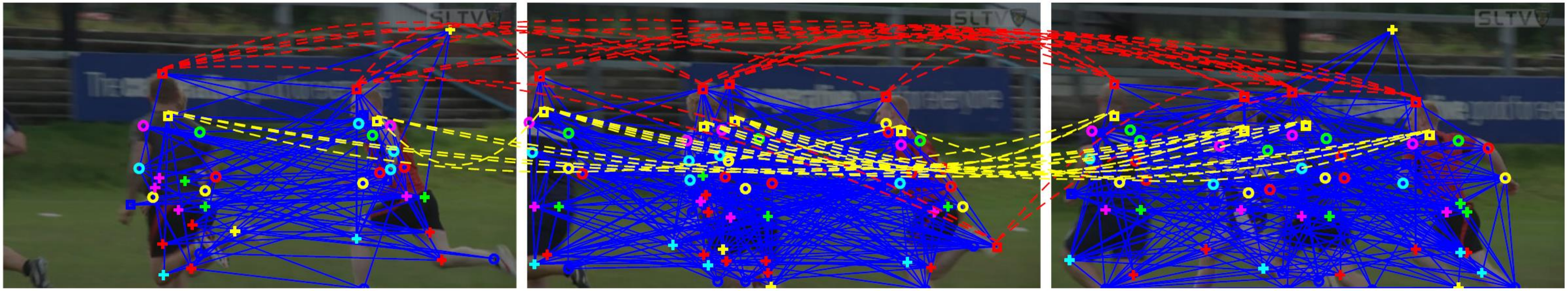
Pose Track: Simultaneous Pose Estimation and Tracking



Unaries: Confidences of detected joints p_d

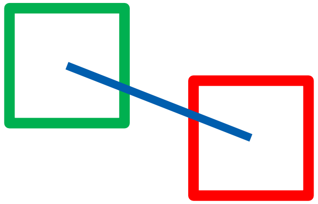
[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking



Spatial binaries: Extract quadratic bounding box around detection

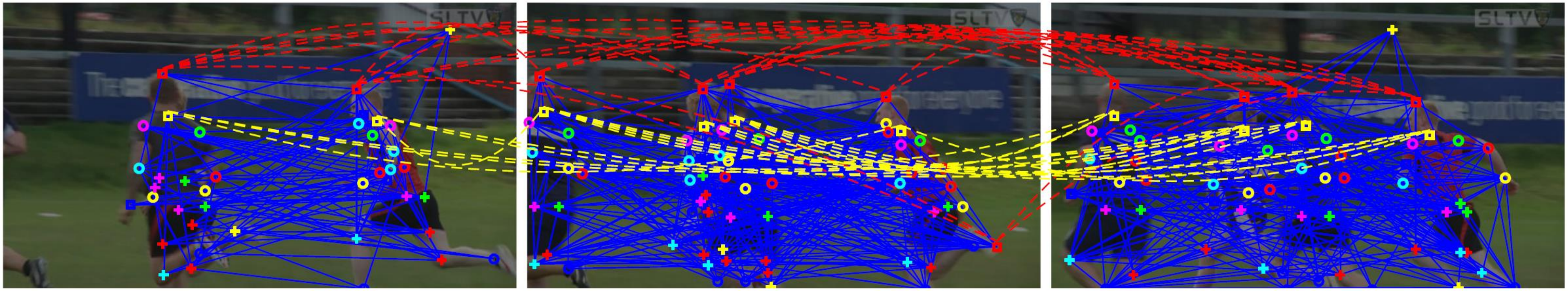
Two cases:



- Different joint type: $p^{s}_{(d_f, d'_f)}$
- Logistic regression based on distance and orientation

[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

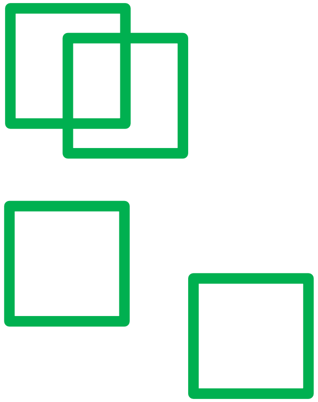
Pose Track: Simultaneous Pose Estimation and Tracking



Spatial binaries: Extract quadratic bounding box around detection

Two cases:

- Same joint type: $p^s_{(d_f, d'_f)} = \mathrm{IoU}(B_d, B_{d'})$

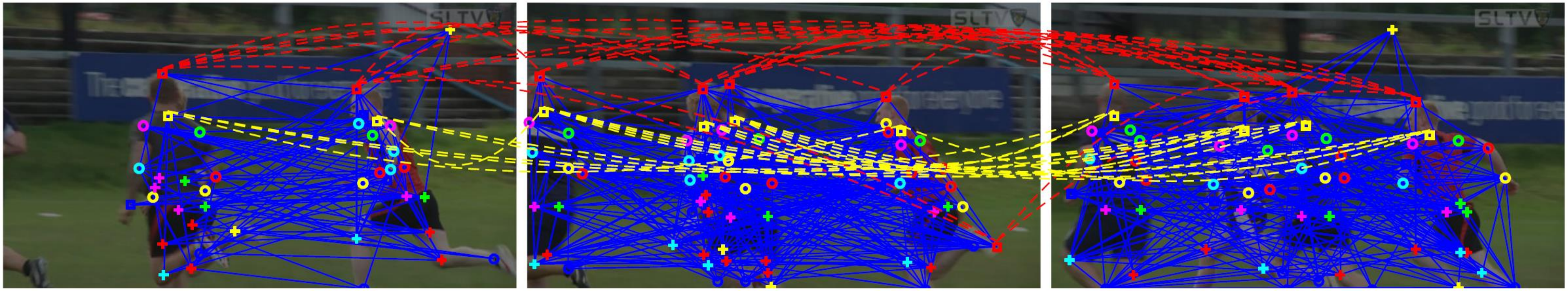


[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

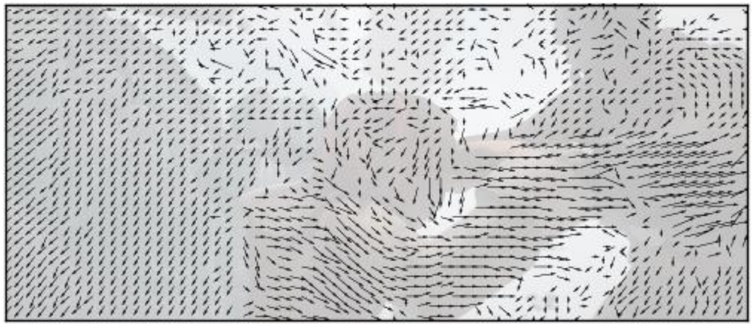
Pose Track: Simultaneous Pose Estimation and Tracking

[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking



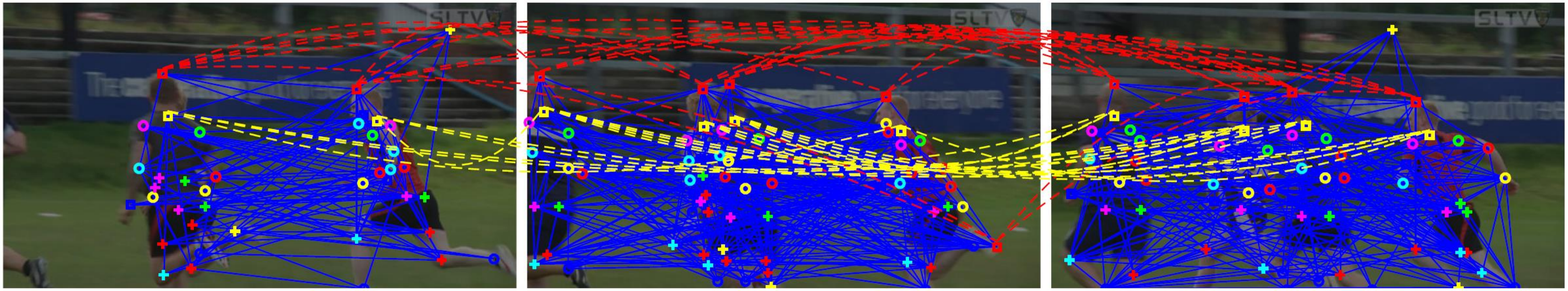
Temporal binaries: Compute optical flow (DeepMatching)



[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking

universitätbonn

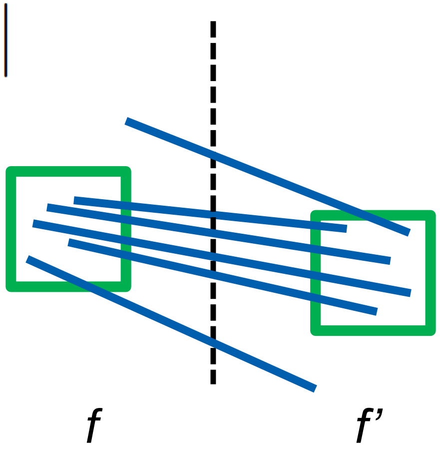


Temporal binaries: Compute optical flow (DeepMatching)

$$\underline{K}_{dd'} = |K_{d_f} \cup K_{d'_{f'}}| \text{ and } \overline{K}_{dd'} = |K_{d_f} \cap K_{d'_{f'}}|$$

$$\{\overline{K}/\underline{K}, \min(p_d, p_{d'}), \Delta\mathbf{x}_{dd'}, \|\Delta\mathbf{x}_{dd'}\|\}$$

Logistic regression: $p^t_{(d_f, d'_{f'})}$



[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking

Solve integer linear program:

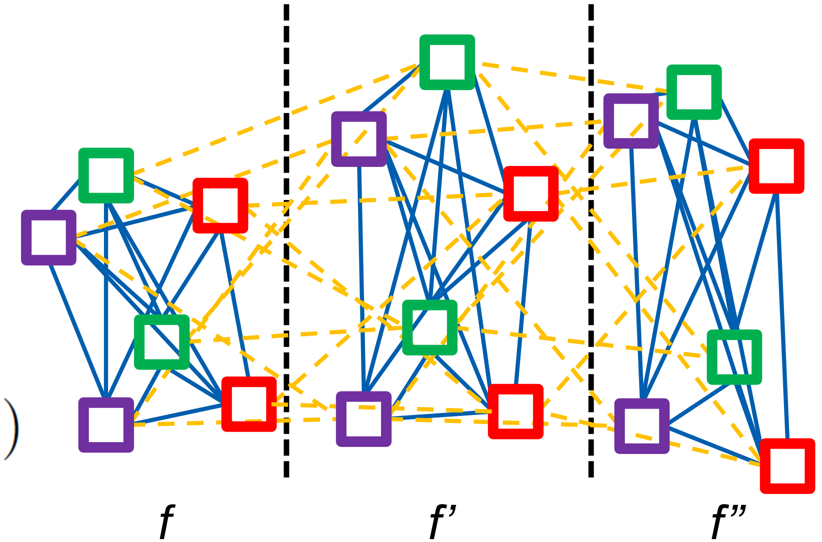
$$v \in \{0,1\}^{|D|},\ s \in \{0,1\}^{|E_s|},\ \text{and}\ t \in \{0,1\}^{|E_t|}$$

$$\underset{v,s,t}{\operatorname{argmin}} \left(\langle v, \phi \rangle + \langle s, \psi_s \rangle + \langle t, \psi_t \rangle \right)$$

$$\langle v, \phi \rangle = \sum_{d \in D} v_d \phi(d)$$

$$\langle s, \psi_s \rangle = \sum_{(d_f, d'_f) \in E_s} s_{(d_f, d'_f)} \psi_s(d_f, d'_f)$$

$$\langle t, \psi_t \rangle = \sum_{(d_f, d'_{f'}) \in E_t} t_{(d_f, d'_{f'})} \psi_t(d_f, d'_{f'})$$



[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking

Solve integer linear program:

$$v \in \{0,1\}^{|D|},\ s \in \{0,1\}^{|E_s|}, \text{ and } t \in \{0,1\}^{|E_t|}$$

$$\underset{v,s,t}{\operatorname{argmin}} \left(\langle v, \phi \rangle + \langle s, \psi_s \rangle + \langle t, \psi_t \rangle \right)$$

$$\langle v, \phi \rangle = \sum_{d \in D} v_d \phi(d)$$

$$\langle s, \psi_s \rangle = \sum_{(d_f, d'_f) \in E_s} s_{(d_f, d'_f)} \psi_s(d_f, d'_f)$$

$$\langle t, \psi_t \rangle = \sum_{(d_f, d'_{f'}) \in E_t} t_{(d_f, d'_{f'})} \psi_t(d_f, d'_{f'})$$

$$\phi(d) = \log \frac{1 - p_d}{p_d}$$

$$\psi_s(d_f, d'_f) = \log \frac{1 - p^s_{(d_f, d'_f)}}{p^s_{(d_f, d'_f)}}$$

$$\psi_t(d_f, d'_{f'}) = \log \frac{1 - p^t_{(d_f, d'_{f'})}}{p^t_{(d_f, d'_{f'})}}$$

[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking

Solve integer linear program:

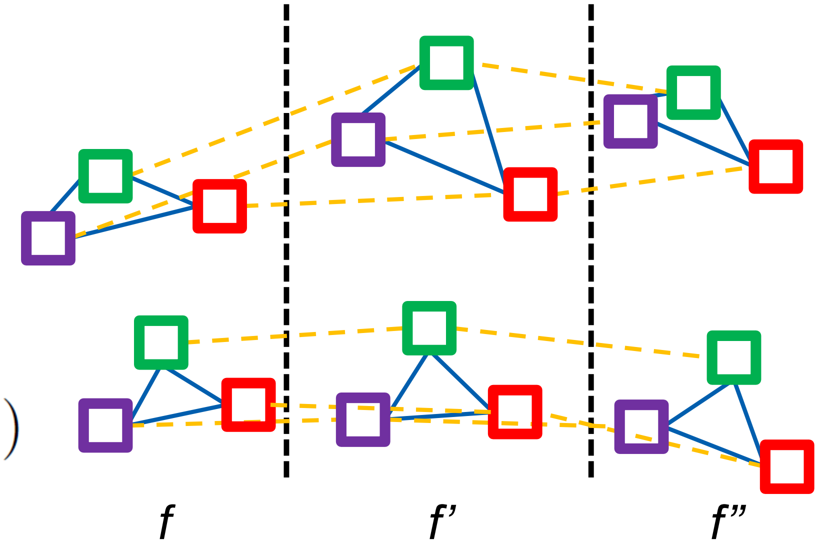
$$v \in \{0,1\}^{|D|},\ s \in \{0,1\}^{|E_s|}, \text{ and } t \in \{0,1\}^{|E_t|}$$

$$\underset{v,s,t}{\operatorname{argmin}} \left(\langle v, \phi \rangle + \langle s, \psi_s \rangle + \langle t, \psi_t \rangle \right)$$

$$\langle v, \phi \rangle = \sum_{d \in D} v_d \phi(d)$$

$$\langle s, \psi_s \rangle = \sum_{(d_f, d'_f) \in E_s} s_{(d_f, d'_f)} \psi_s(d_f, d'_f)$$

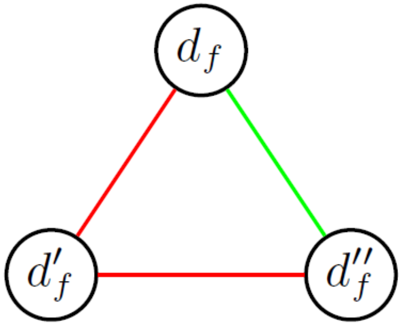
$$\langle t, \psi_t \rangle = \sum_{(d_f, d'_{f'}) \in E_t} t_{(d_f, d'_{f'})} \psi_t(d_f, d'_{f'})$$

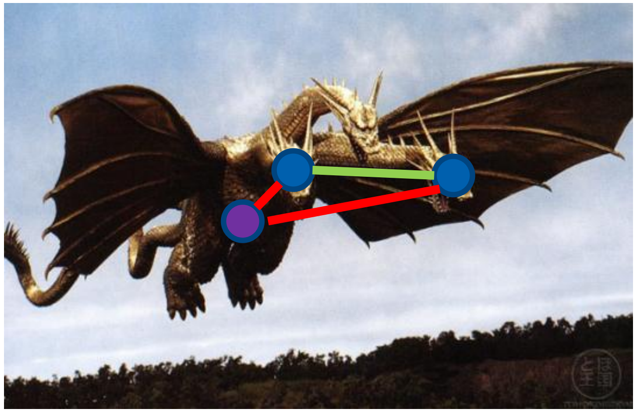


[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking

To obtain plausible pauses, constraints are added:





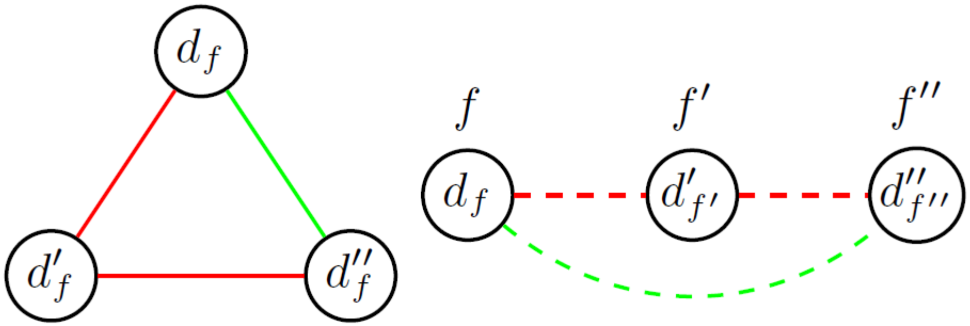
- Spatial transitivity: $$s_{(d_f, d'_f)} + s_{(d'_f, d''_f)} - 1 \leq s_{(d_f, d''_f)}$$

[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking

universitätbonn

To obtain plausible pauses, constraints are added:



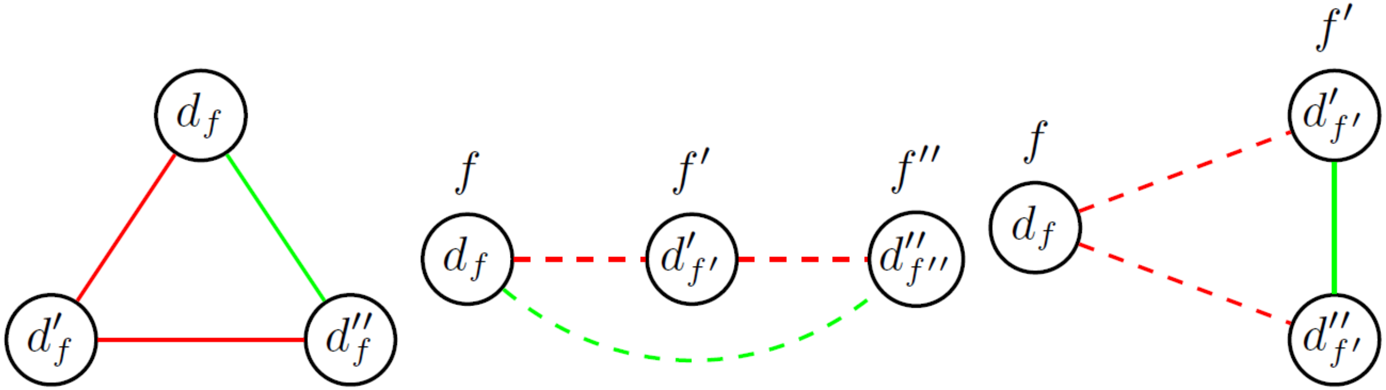
- Spatial transitivity: $s_{(d_f,d'_f)} + s_{(d'_f,d''_f)} - 1 \leq s_{(d_f,d''_f)}$
- Temporal transitivity: $t_{(d_f,d'_{f'})} + t_{(d'_f,d''_{f''})} - 1 \leq t_{(d_f,d''_{f''})}$

[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking

universitätbonn

To obtain plausible pauses, constraints are added:



- Spatial transitivity: $$s_{(d_f, d'_f)} + s_{(d'_f, d''_f)} - 1 \leq s_{(d_f, d''_f)}$$
- Temporal transitivity: $$t_{(d_f, d'_{f'})} + t_{(d'_f, d''_{f''})} - 1 \leq t_{(d_f, d''_{f''})}$$
- Spatio-temporal trans.: $$t_{(d_f, d'_{f'})} + t_{(d_f, d''_{f'})} - 1 \leq s_{(d'_{f'}, d''_{f'})}$$

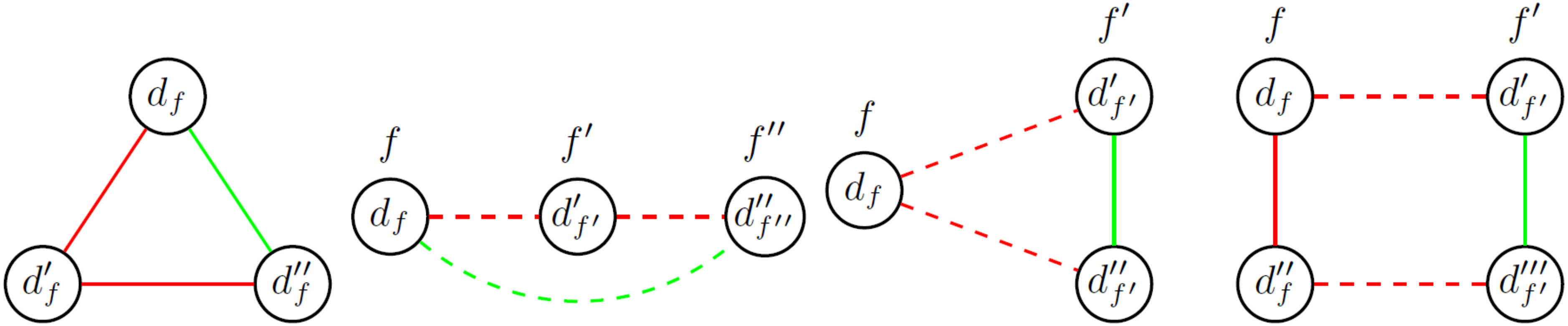
$$t_{(d_f, d'_{f'})} + s_{(d'_{f'}, d''_{f'})} - 1 \leq t_{(d_f, d''_{f'})}$$

[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking

Pose Track: Simultaneous Pose Estimation and Tracking

To obtain plausible pauses, constraints are added:



Spatio-temporal consistency:

$$t_{(d_f, d'_{f'})} + t_{(d''_f, d'''_{f'})} + s_{d_f, d''_f} - 2 \leq s_{d'_{f'}, d'''_{f'}}$$

$$t_{(d_f, d'_{f'})} + t_{(d''_f, d'''_{f'})} + s_{d'_{f'}, d'''_{f'}} - 2 \leq s_{d_f, d''_f}$$

[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

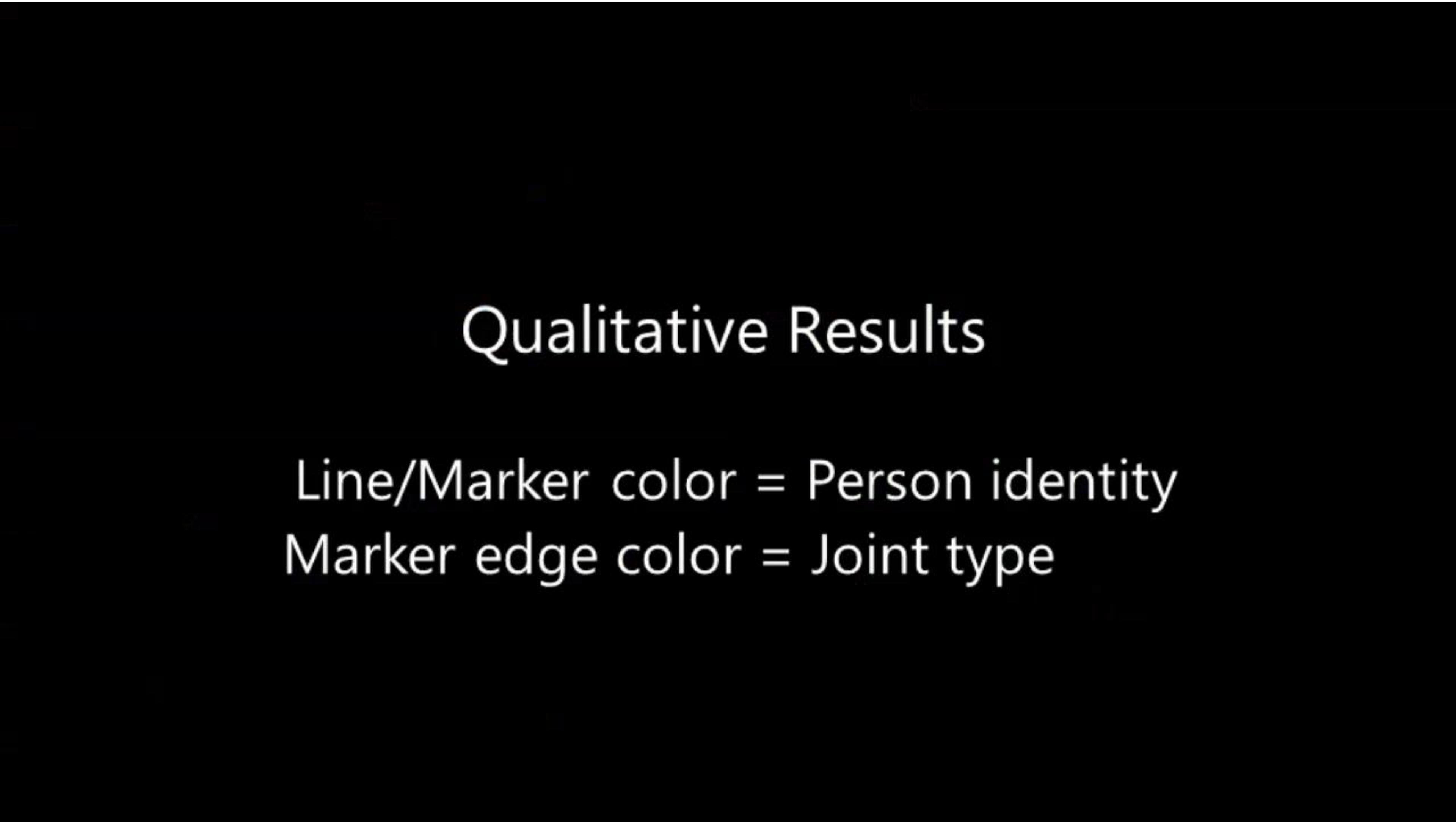
Pose Track: Simultaneous Pose Estimation and Tracking

Estimate pose + person association over time:

- Predict body joints (CNN trained on MPII Pose)
- Build a graph with temporal and spatial edges
- Partition spatio-temporal graph

[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Pose Track: Simultaneous Pose Estimation and Tracking



Pose Track: Evaluation

- Pose estimation accuracy (mAP)
- Person association (MOTA)

Method	Rcll ↑	Prcn ↑	MT ↑	ML ↓	IDs ↓	FM ↓	MOTA ↑	MOTP ↑
Ours	**63.0**	**64.8**	**775**	**502**	431	5629	**28.2**	**55.7**
BBox-Tracking [38, 34]								
+ LJPA [17]	58.8	64.8	716	646	**319**	**5026**	26.6	53.5
+ CPM [40]	60.1	57.7	754	611	347	4969	15.6	53.4

[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

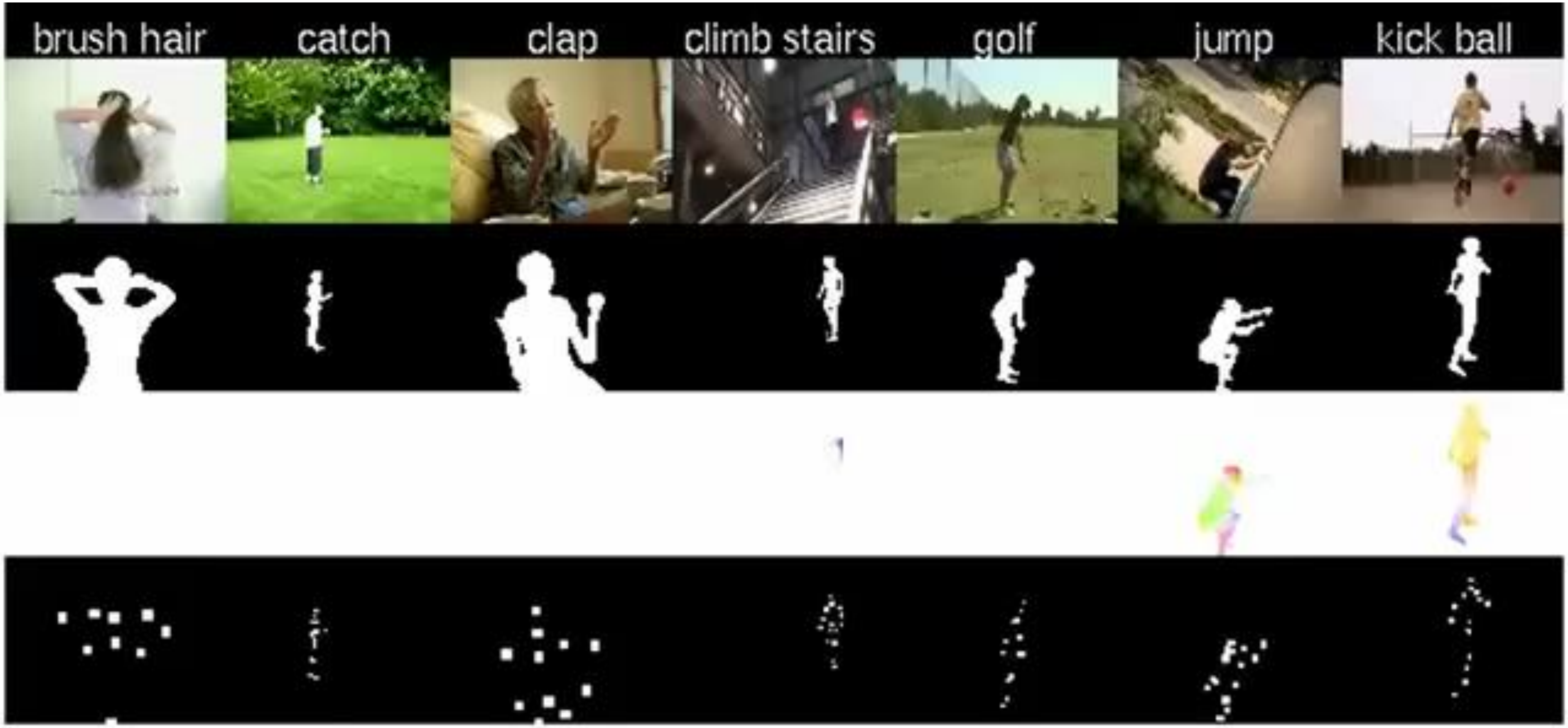
Pose Track: Evaluation

- Pose estimation accuracy (mAP)
- Person association (MOTA)

Method	Head	Sho	Elb	Wri	Hip	Knee	Ank	mAP
Ours	**56.5**	51.6	**42.3**	**31.4**	22.0	**31.9**	**31.6**	**38.2**
BBox-Detection [34]								
+ LJPA [17]	50.5	49.3	38.3	33.0	21.7	29.6	29.2	35.9
+ CPM [40]	48.8	47.5	35.8	29.2	20.7	27.1	22.4	33.1
DeeperCut [16]	56.2	**52.4**	40.1	30.0	**22.8**	30.5	30.8	37.5

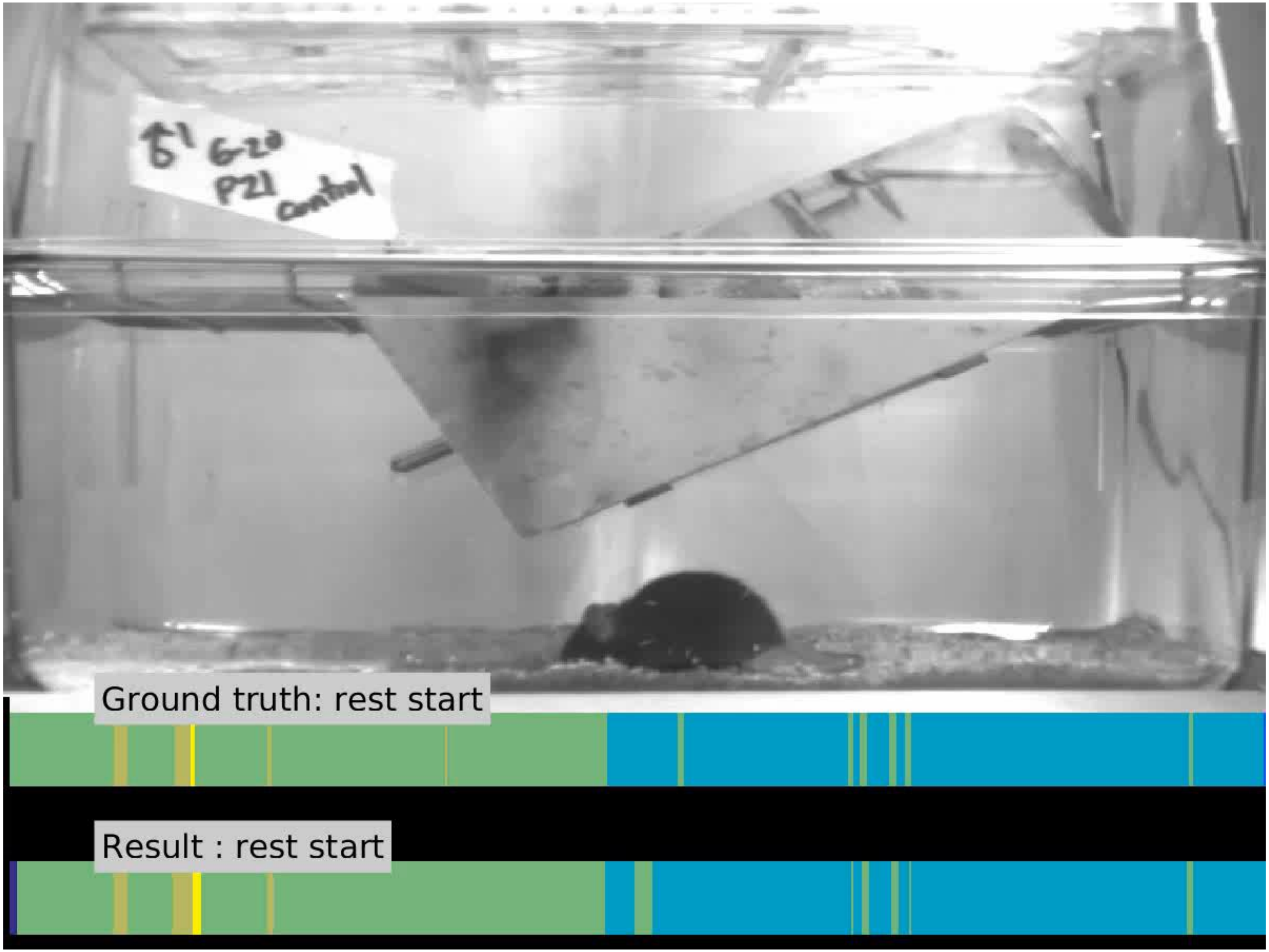
[U. Iqbal et al. **Pose-Track: Joint Multi-Person Pose Estimation and Tracking.** CVPR 2017]

Joint-annotated **HMDB** (JHMDB)



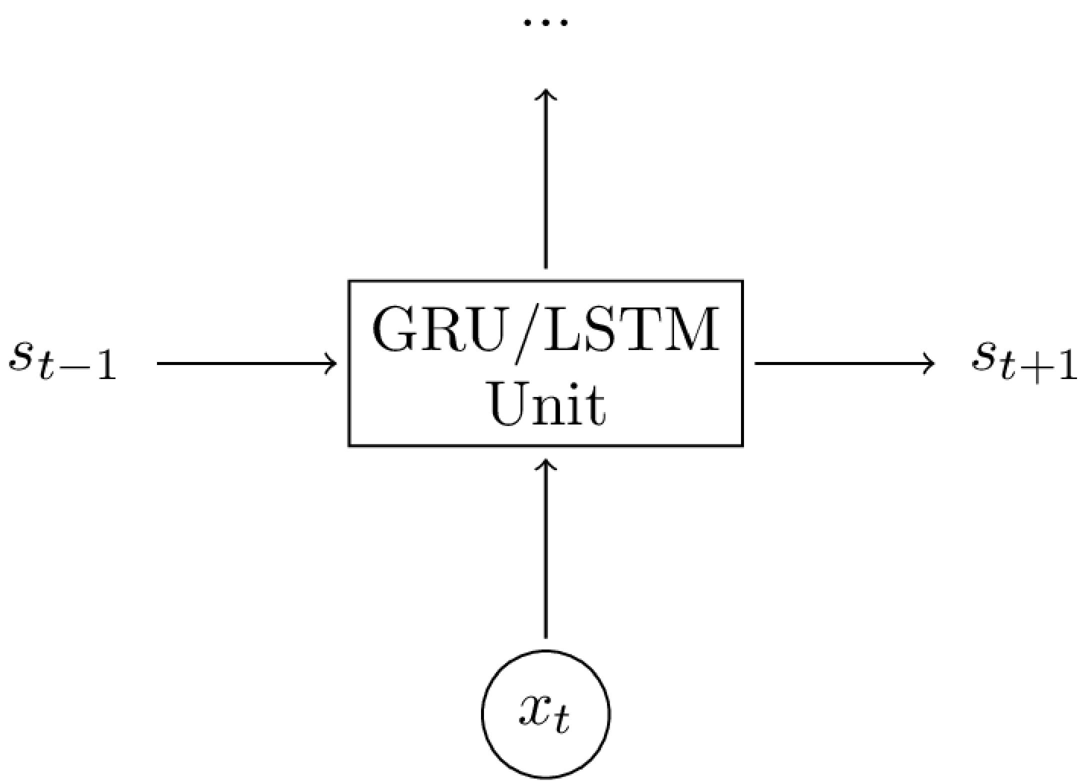
[H. Jhuang et al. **Towards Understanding Action Recognition.** ICCV 2013]
[http://jhmdb.is.tue.mpg.de]

Video Analysis for Studying the Behavior of Mice



Recurrent Neural Networks

- Gated units (LSTM/GRU)



Weakly Supervised Learning

- Fully supervised:

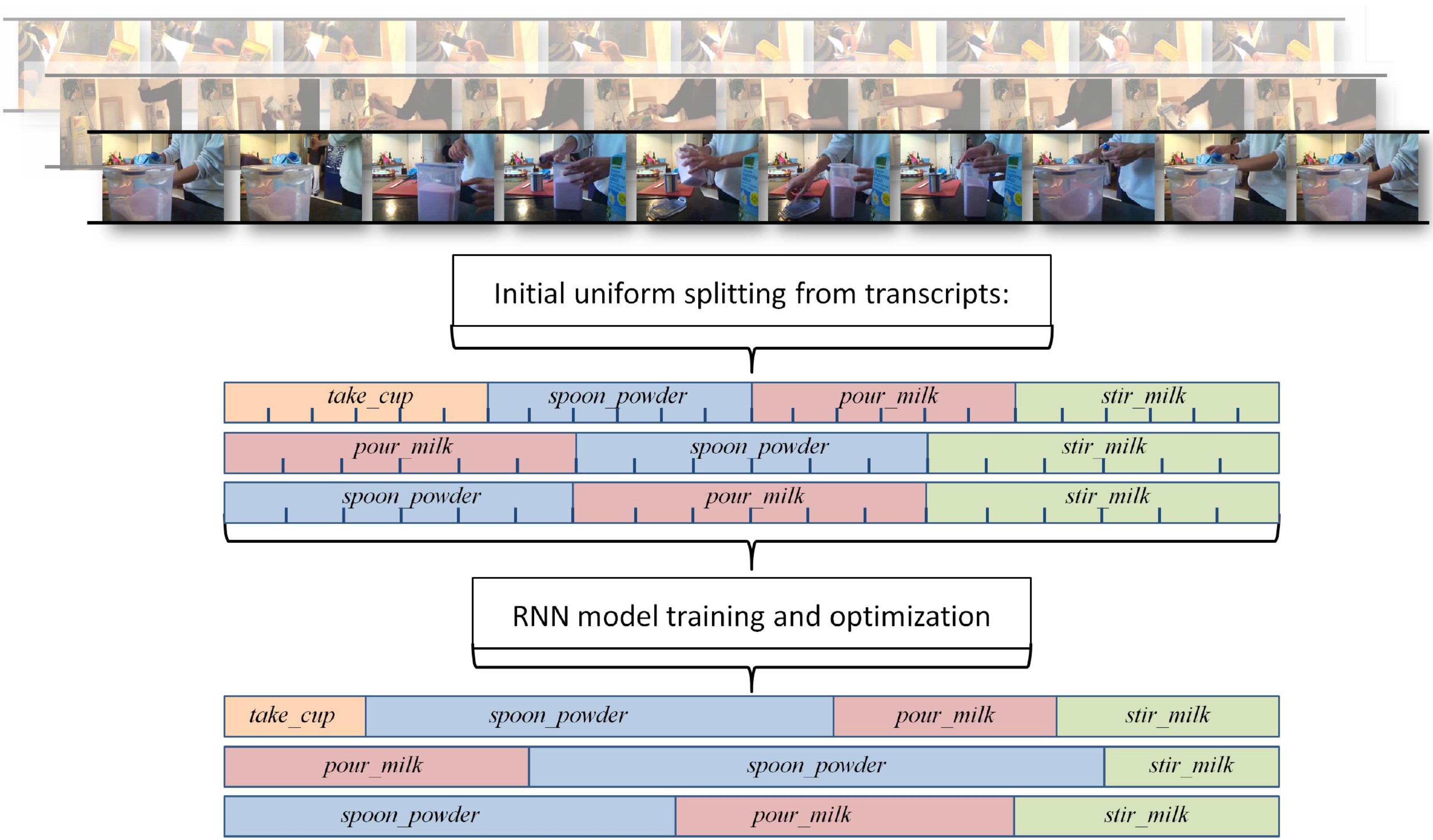
- Weakly supervised (transcripts)

action_A → action_B → action_A → action_C

[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

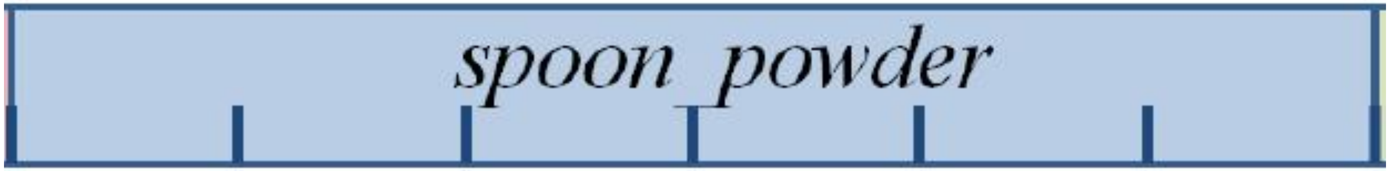
Weakly Supervised Learning

universitätbonn



Weakly Supervised Learning

- Represent an activity a like "spoon_powder" by latent sub-activities $s_1^{(a)}, s_2^{(a)}, s_3^{(a)}, \ldots$

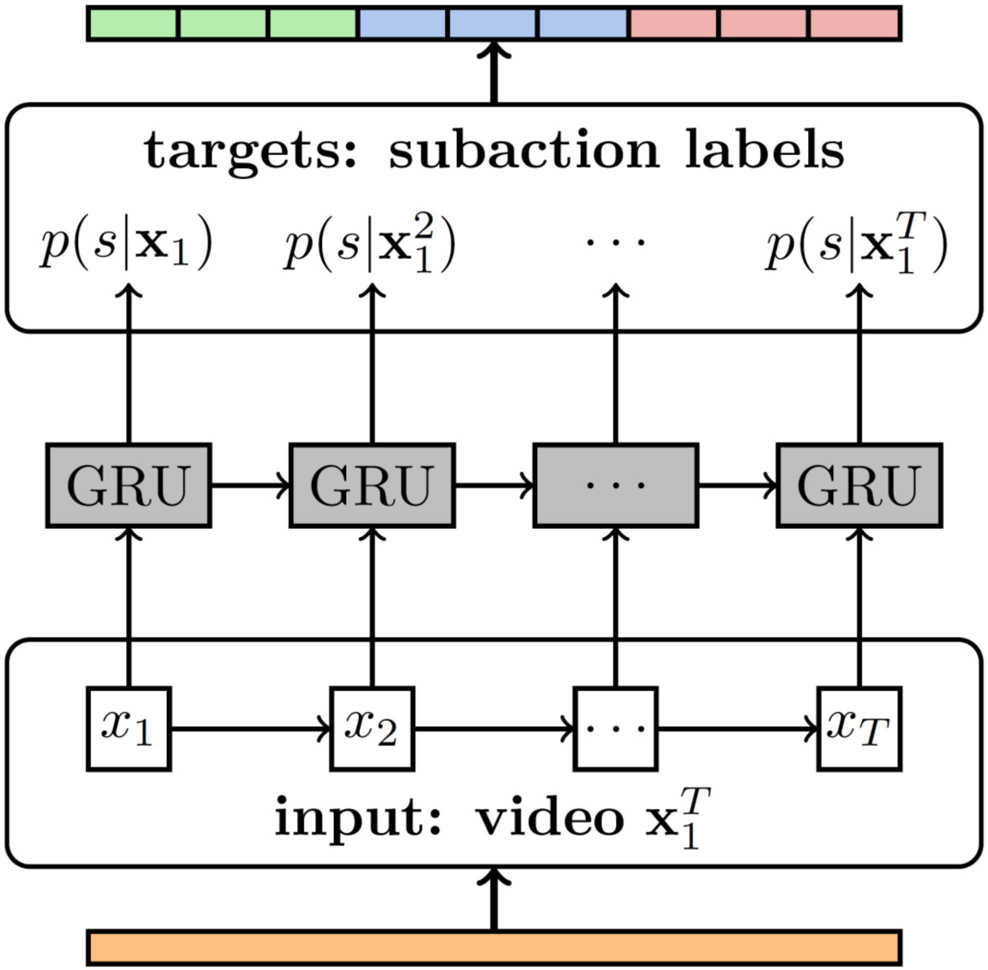


$s_1^{(a)}$ $s_2^{(a)}$ $s_3^{(a)}$ $s_4^{(a)}$ $s_5^{(a)}$ $s_6^{(a)}$

- Optimal number of sub-activities is unknown:
  - Many sub-activities for long activities
  - Few sub-activities for short activities

Model

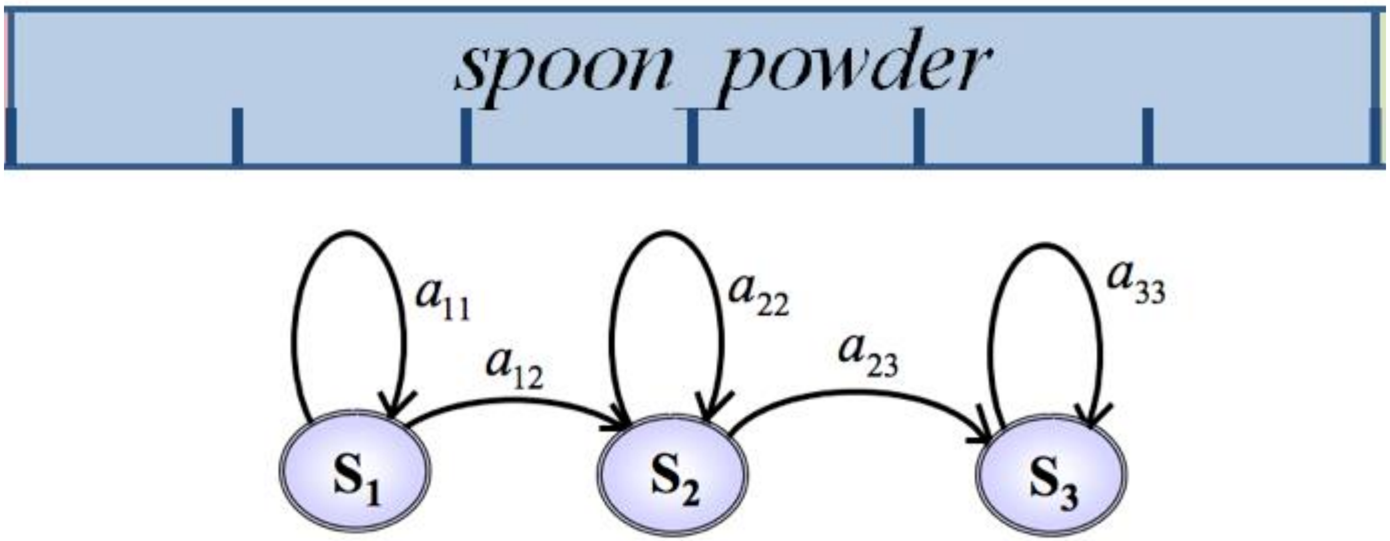
- RNN with Gated Recurrent Units (GRU)



$$p(x_t|s) = \text{const} \cdot \frac{p(s|x_t)}{p(s)}$$

Model

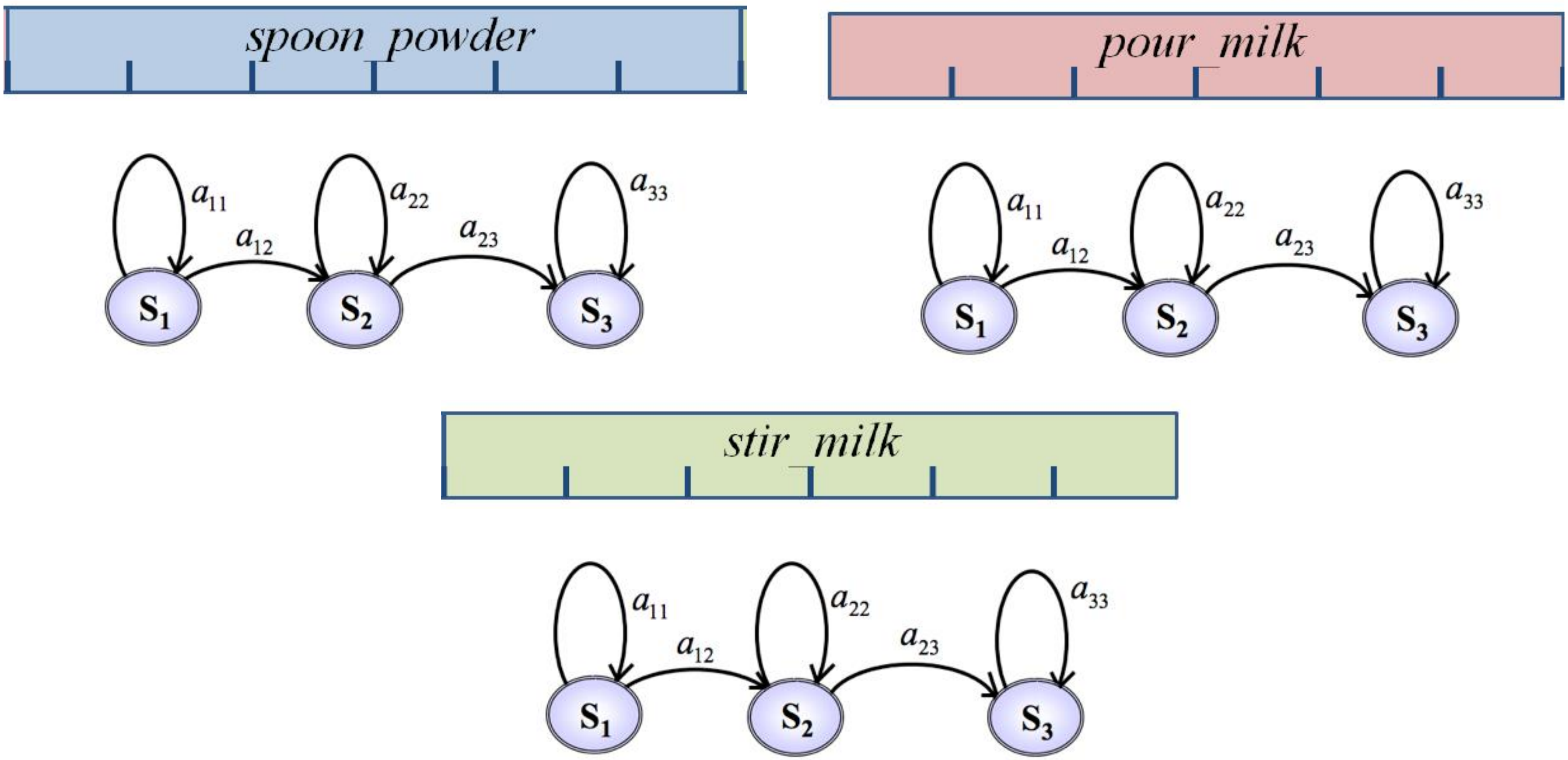
- Hidden Markov Model (HMM) enforce fixed order of sub-activities: $s_1^{(a)}, s_2^{(a)}, s_3^{(a)}, \ldots$



- HMMs use probabilities of RNN as input

Model

- Hidden Markov Model (HMM) for each activity

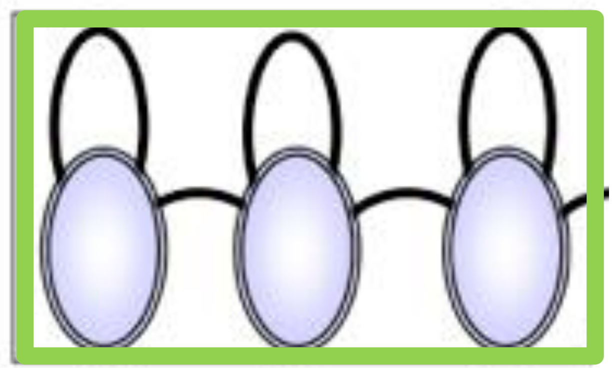


universitätbonn

- The transcripts define the order of activities:

Action transcript:

action_1 action_2 action_3



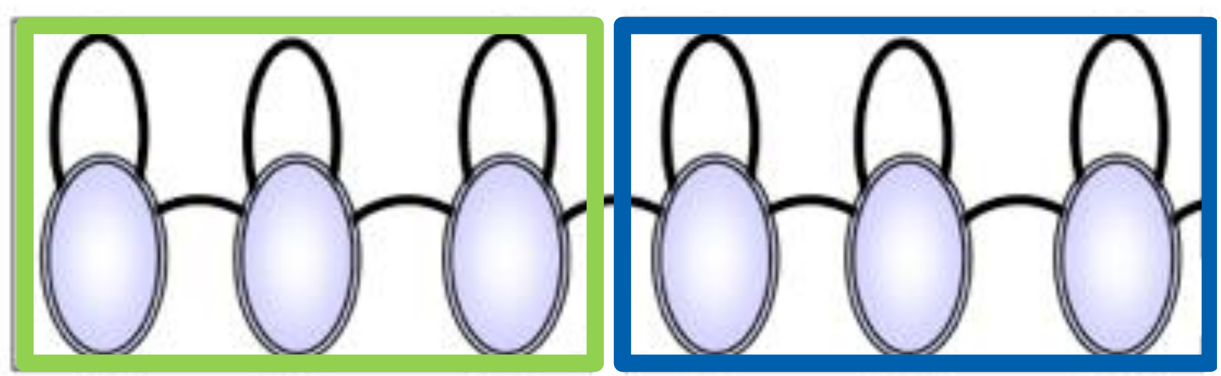
$s_1^{(a_1)}$ $s_2^{(a_1)}$ $s_3^{(a_1)}$

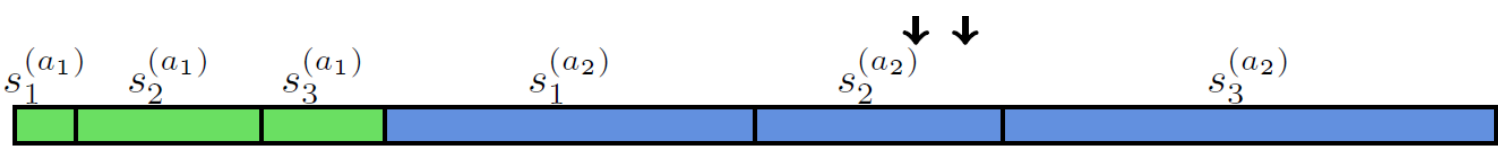
Model

- The transcripts define the order of activities:

Action transcript:

action_1 action_2 action_3



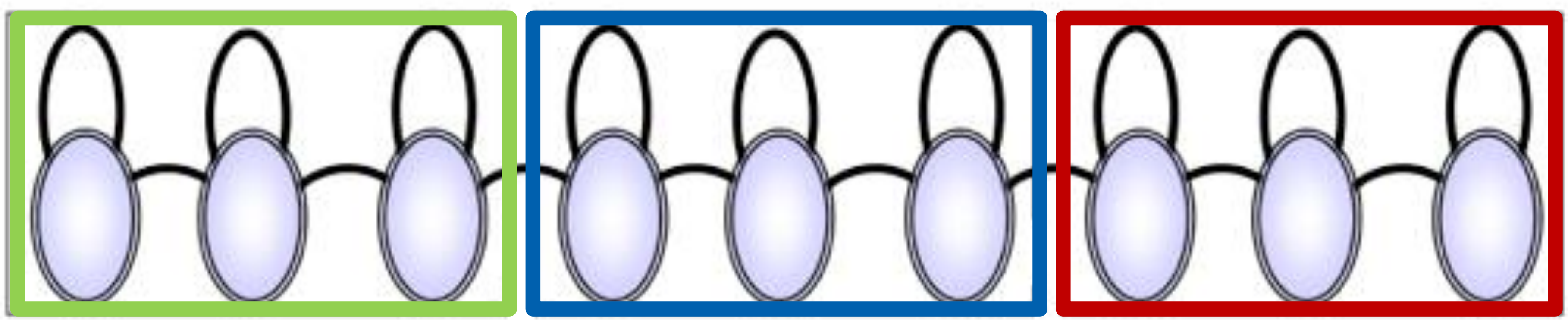


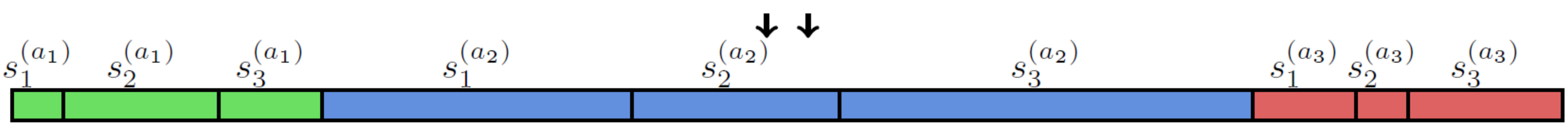
Model

- The transcripts define the order of activities:

Action transcript:

action_1 action_2 action_3





Weakly Supervised Learning

universitätbonn

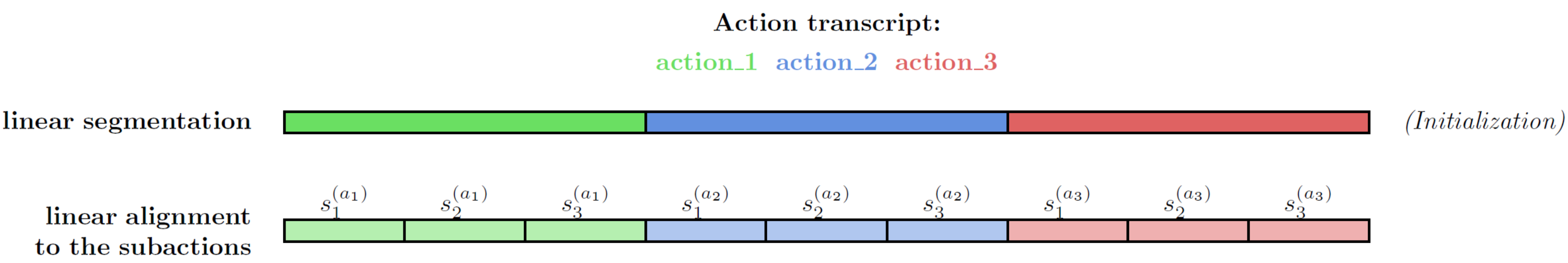
Action transcript:

action_1 action_2 action_3

linear segmentation  *(Initialization)*

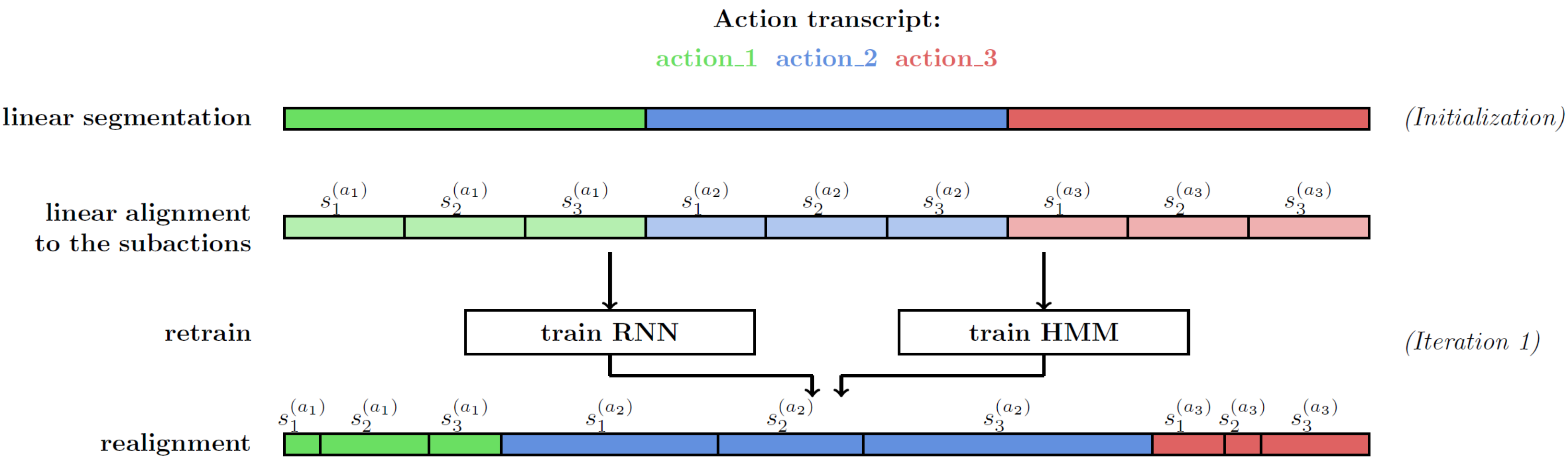
[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

Weakly Supervised Learning



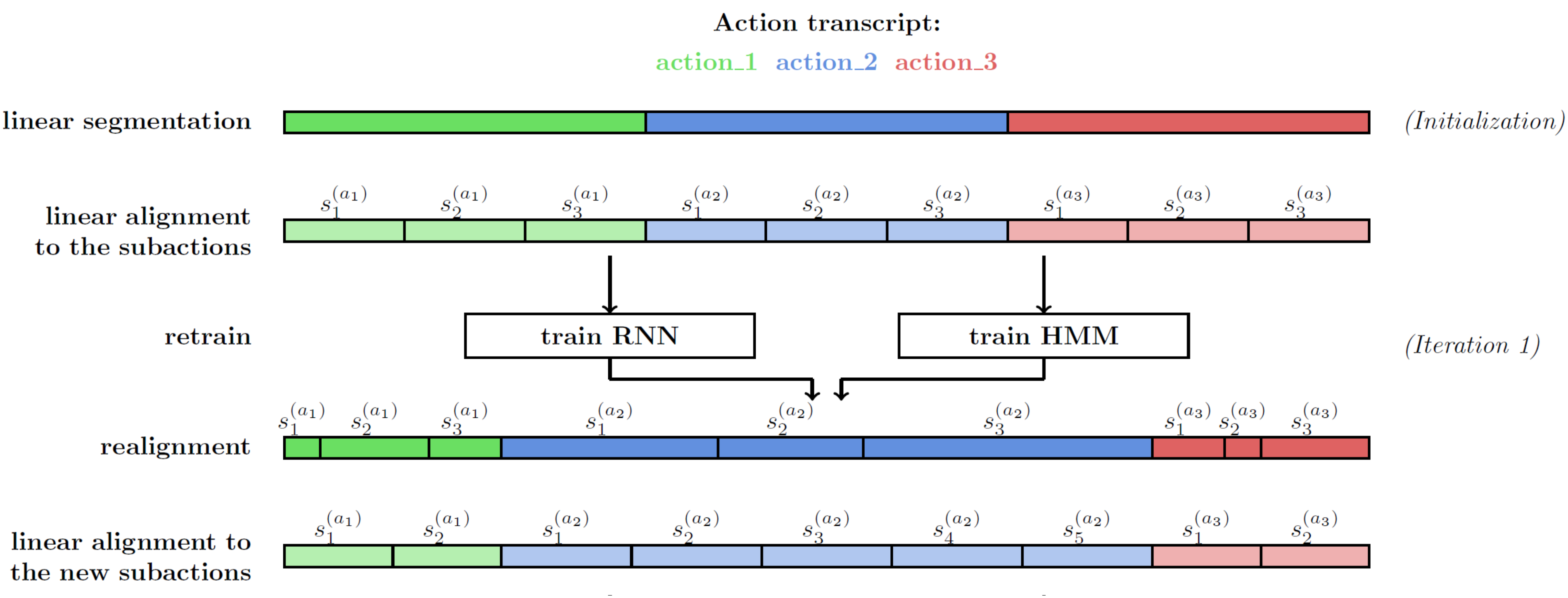
[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

Weakly Supervised Learning



[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

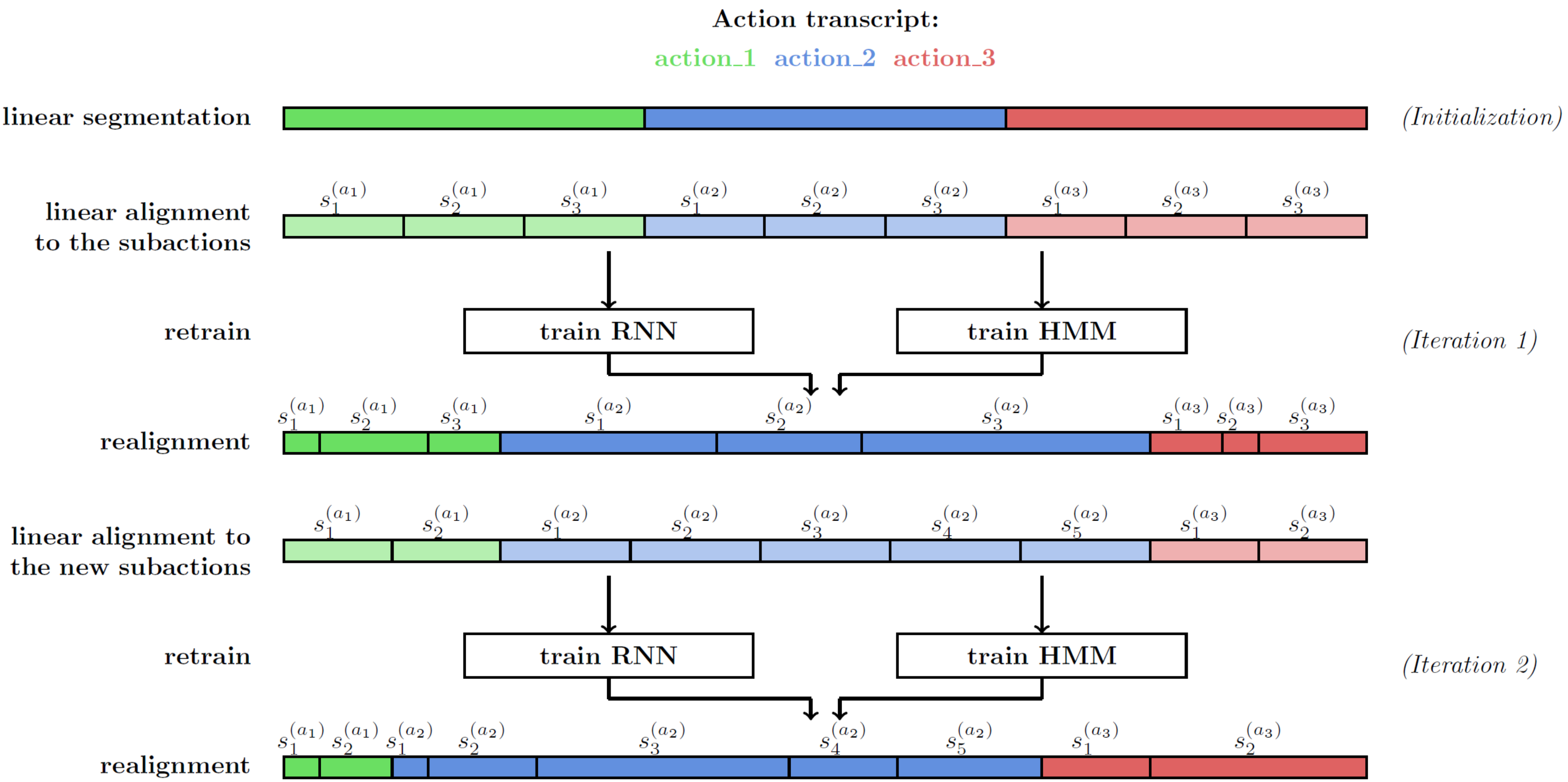
Weakly Supervised Learning



[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

Weakly Supervised Learning

universitätbonn



[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

Results

Results

- Accuracy on unseen sequences (video without transcript)

Breakfast	Accuracy (Mof)
GRU no subactions	22.4
GRU w/o reestimation	28.8
GRU + reestimation	33.3
GRU + GT length	51.3

[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

Results

- Accuracy on unseen sequences (video without transcript)

Breakfast	Iter 1	Iter 2	Iter 3	Iter 4	Iter 5
GMM w/o reest.	15.3	23.3	26.3	27.0	26.5
MLP w/o reest.	22.4	24.0	23.7	23.1	20.3
GRU w/o reest.	25.5	29.1	28.6	29.3	28.8
GRU w/o HMM	21.3	20.1	23.8	21.8	22.4

[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

Results

- Accuracy on unseen sequences (video with transcript)

	Breakfast	Hollywood Ext.
Model	Jacc. (IoD)	Jacc. (IoD)
OCDC [3]	23.4	43.9
HTK [16]**	40.6	42.4
ECTC [9]**	-	41.0
GRU w/o reestimation	41.5	50.1
GRU + reestimation	**47.3**	**51.1**

[A. Richard et al. **Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling.** CVPR 2017]

Research Unit - Anticipating Human Behavior

[https://pages.iai.uni-bonn.de/FOR2535]

Research Unit - Anticipating Human Behavior

Thank you for your attention.

European
Research
Council

Deutsche
Forschungsgemeinschaft