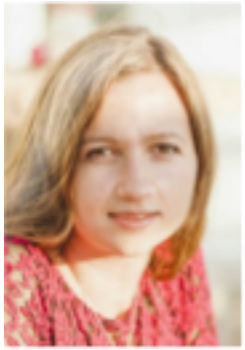
Structured Deep Learning of Human Motion

Christian Wolf


Fabien Baradel

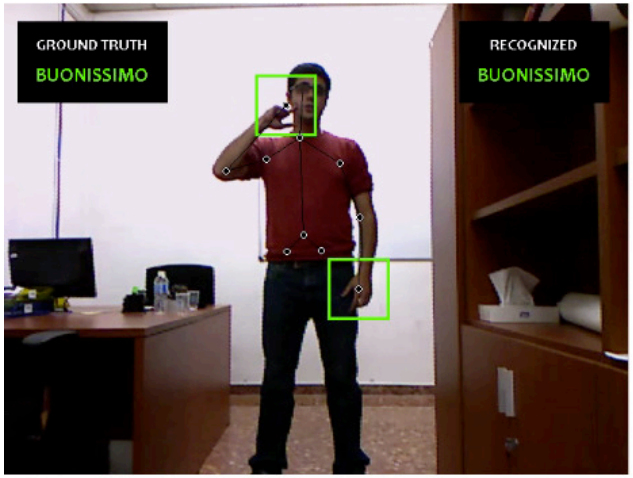

Natalia Neverova

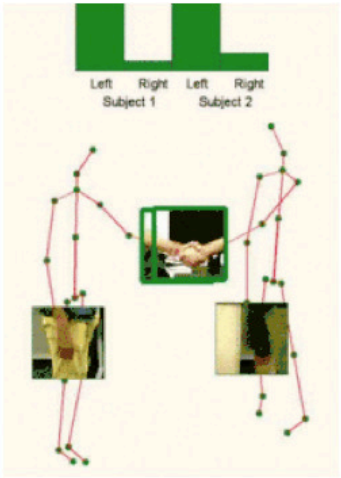

Julien Mille

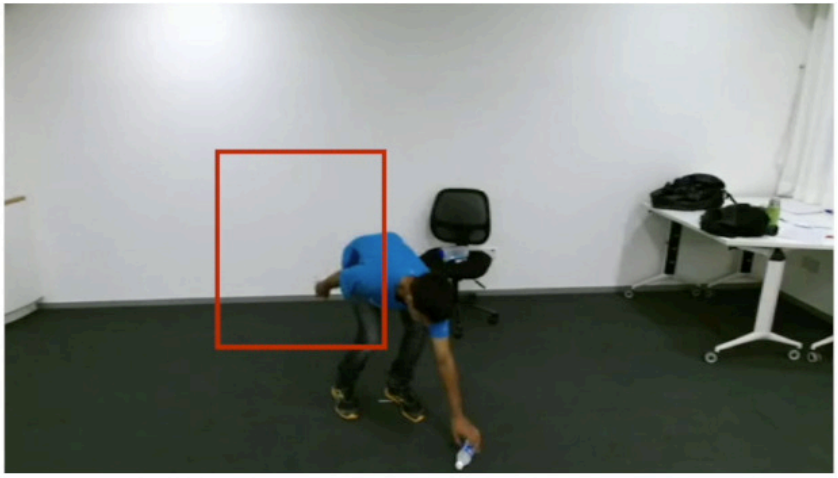

Graham W. Taylor


Greg Mori

Deep Learning of Human Motion

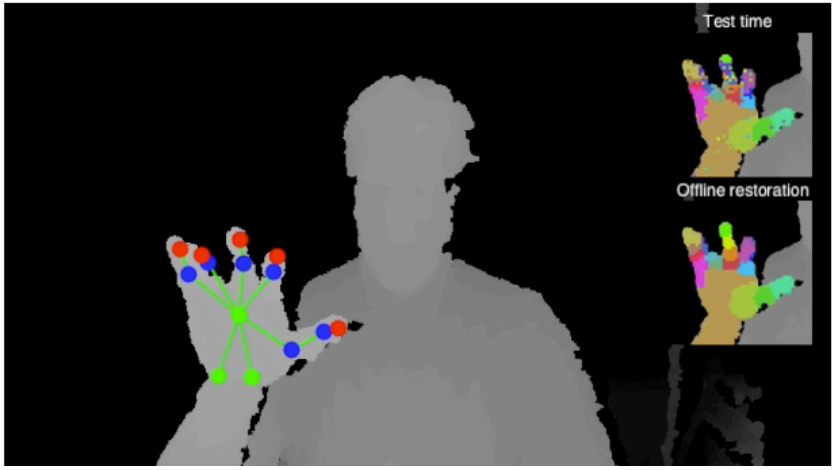


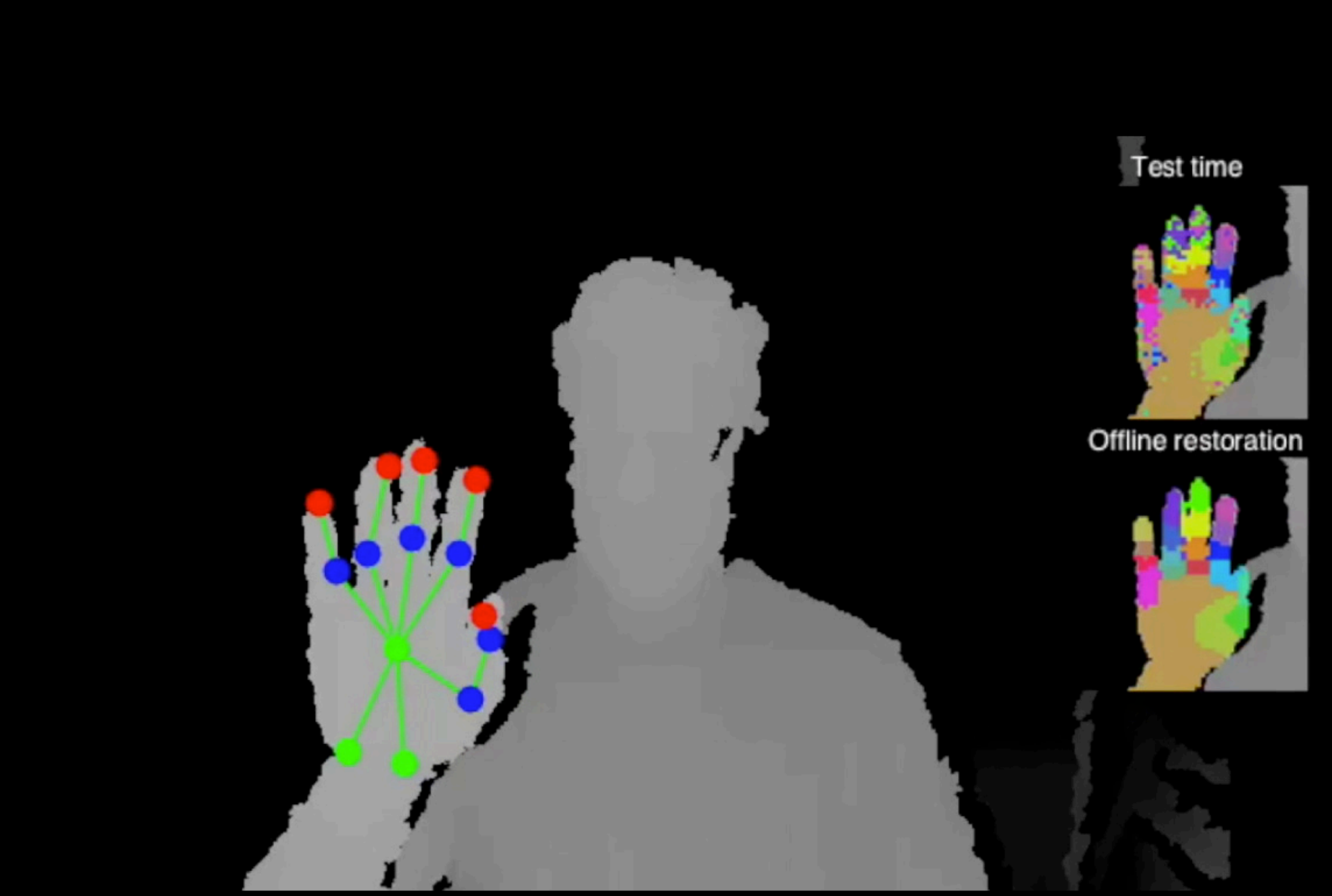




Recognition of group activities

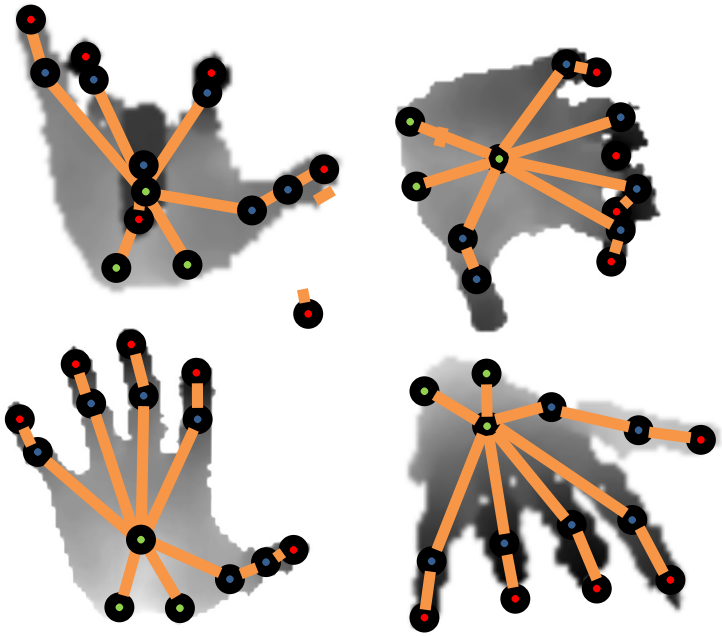
Pose estimation





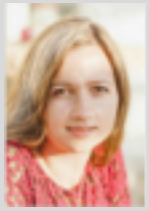
[Neverova, Wolf, Taylor, Nebout. CVIU 2017]

Combining real and simulated data



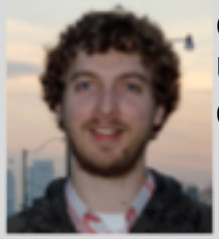
Joint positions (NYU Dataset)

Synthetic data (part segmentation)

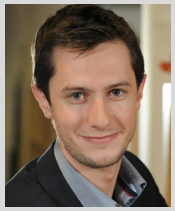


Natalia Neverova
Phd @ LIRIS,
Now at Facebook

Christian Wolf
LIRIS
INSA-Lyon

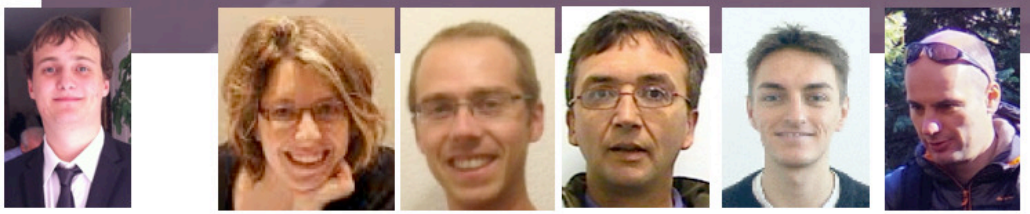


Graham W. Taylor
University of Guelph
Canada



Florian Nebout
Awabot

Semantic Segmentation with GridNetworks

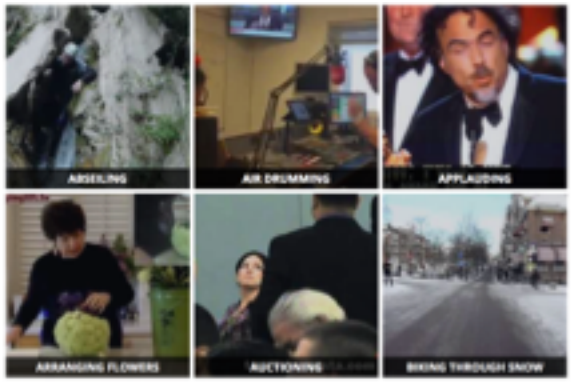


Damien Fourure E. Fromont, R. Emonet, A. Trémeau, D. Muselet, C. Wolf

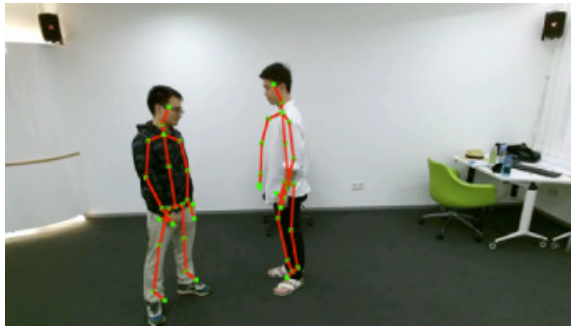
[Fourure, Emonet, Fromont, Muselet, Tremeau, Wolf, BMVC 2017]

Activity recognition

Unconstrained internet/youtube videos
No acquisition
E.g. Youtube-8M dataset: 7M videos, 4716 classes, ~3.4 labels per video. > 1PB of data.



Videos with human activities, from youtube
No acquisition
E.g. ActivityNet/Kinetics dataset: ~300k videos, 400 classes.

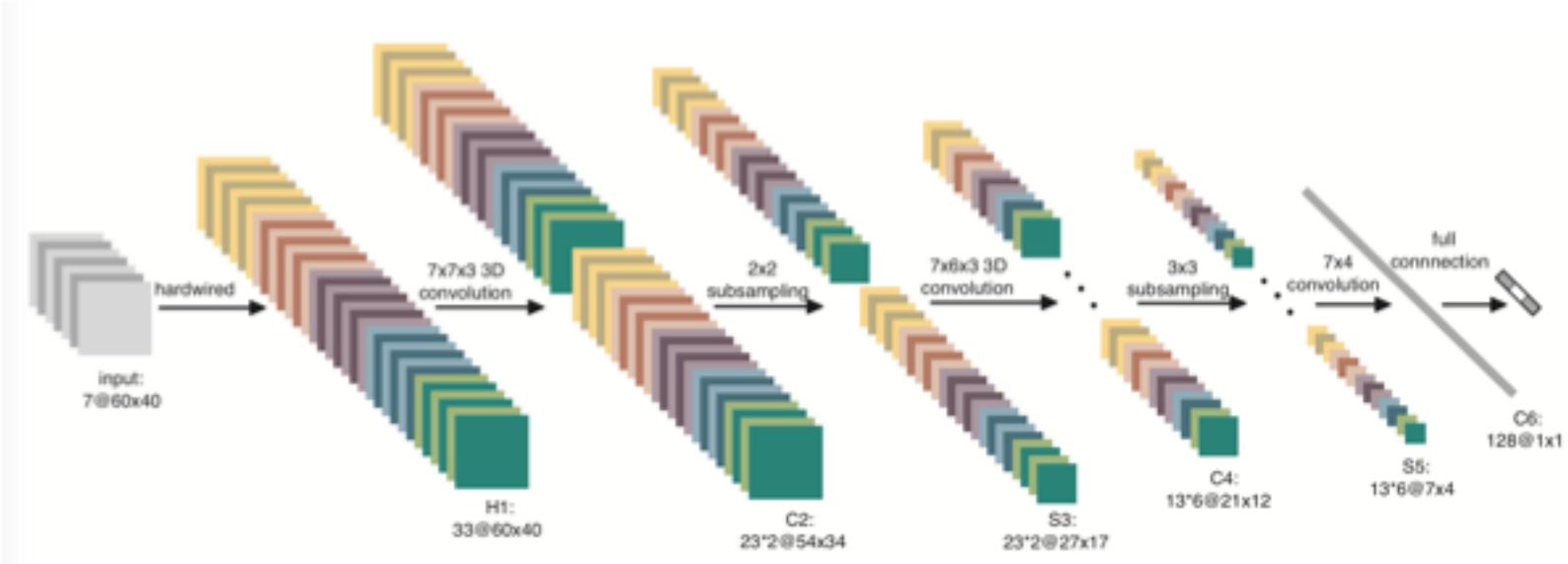


Human activities shot with depth sensors
Acquisition is time consuming!
E.g. NTU RGB-D dataset, MSR dataset, ChaLearn/Montalbano dataset, etc.

Deep Learning (Global)

(Mostly after 2012)

Deep Learning is mostly based on global models.



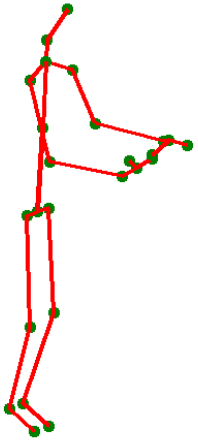
[Ji et al., ICML 2010]

[Carreira and Zisserman, CVPR 2017]

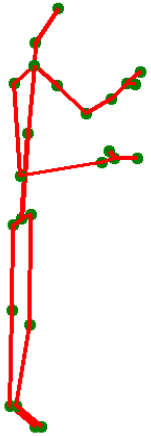
[Baccouche, Mamalet, Wolf, Garcia, Baskur, HBU 2011]

[Baccouche, Mamalet, Wolf, Garcia, Baskur, BMVC 2012]

The role of articulated pose

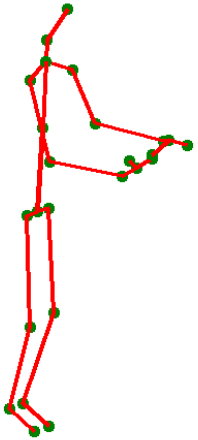


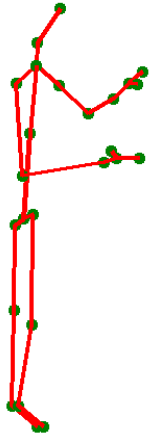
Reading

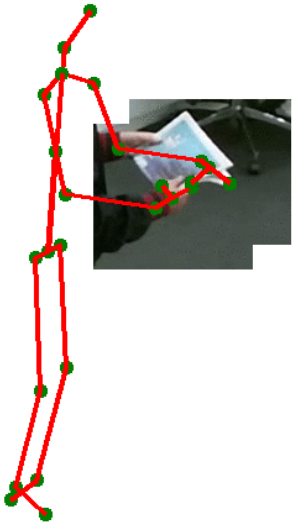


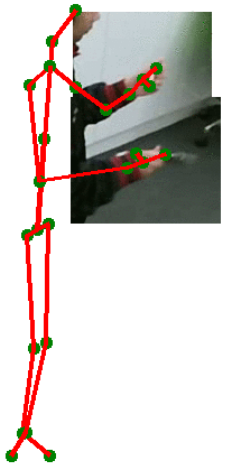
Writing

The role of articulated pose









Reading

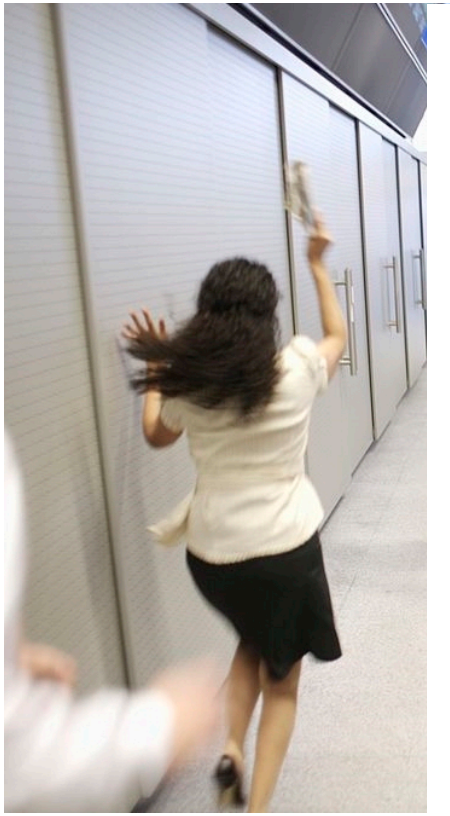
Writing

Appearance is helpful

[Neverova, Wolf, Taylor, Nebout, PAMI 2016]

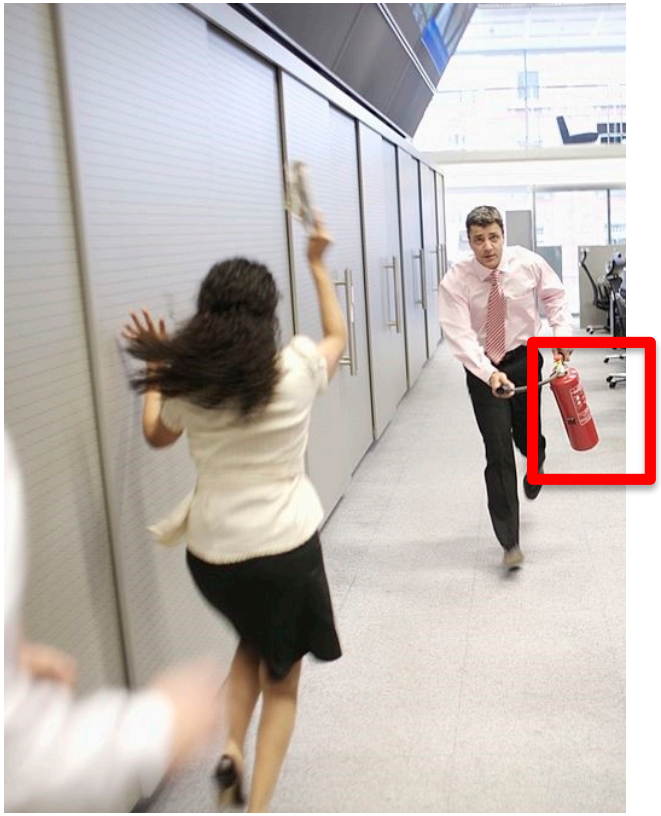
[Baradel, Wolf, Mille, Taylor, BMVC 2018]

Context



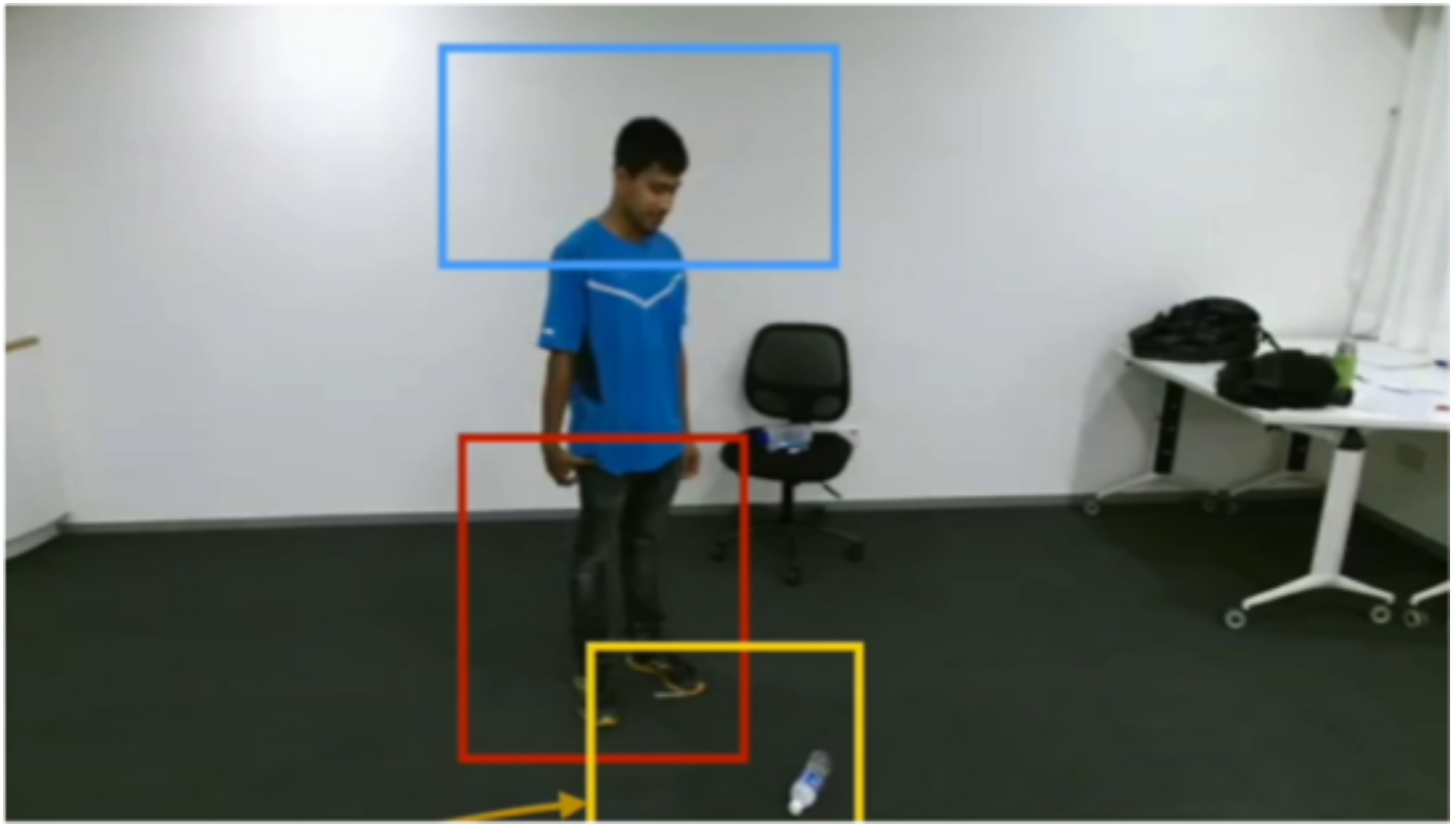
We need put attention to places which are not always determined by pose

Context



We need put attention to places which are not always determined by pose

Context



Frame from the NTU
RGB-D Dataset

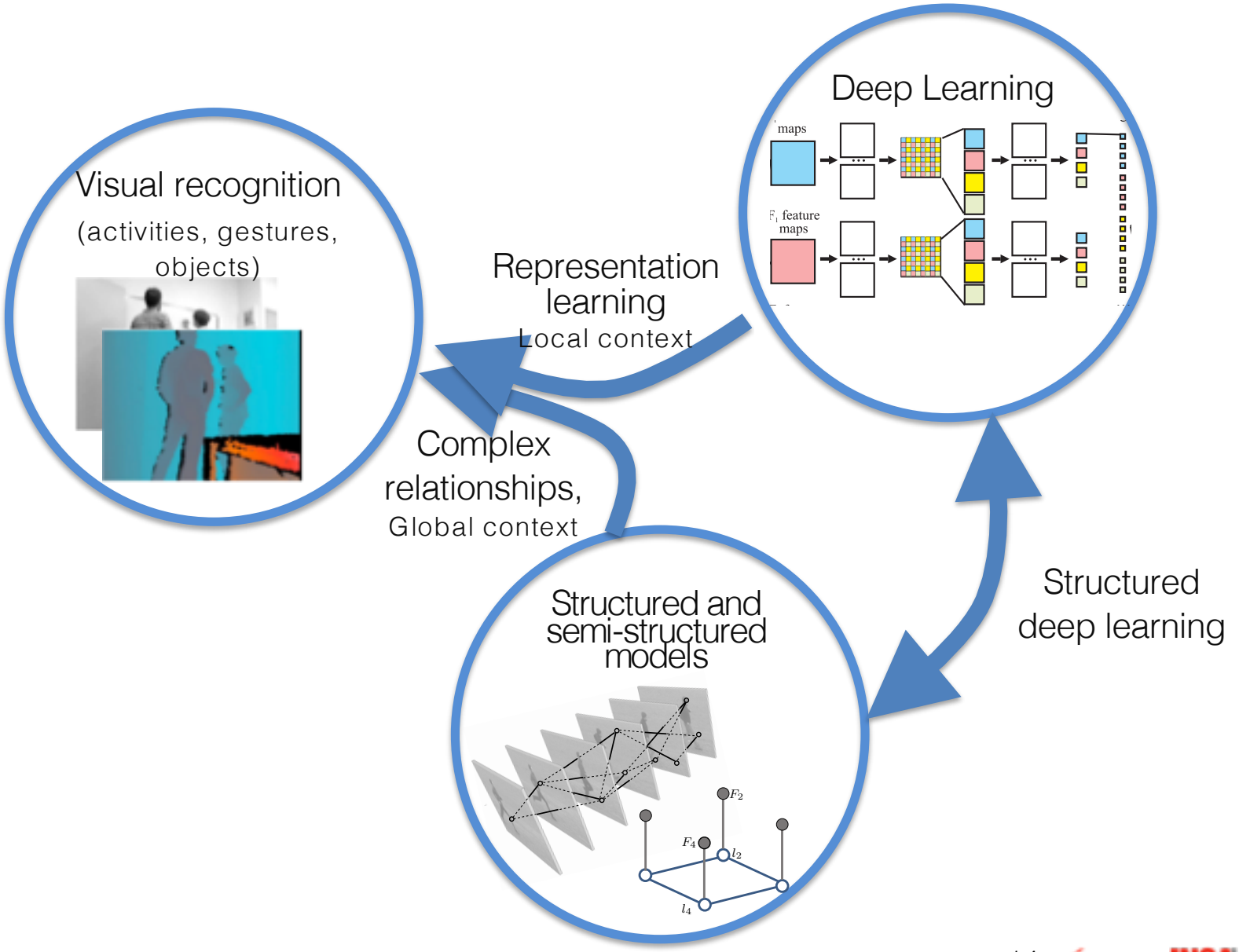
Images, objects and **activities** have often been represented as collections of local features, e.g. through DPMs.

[Felzenszwalb et al., PAMI 2010]

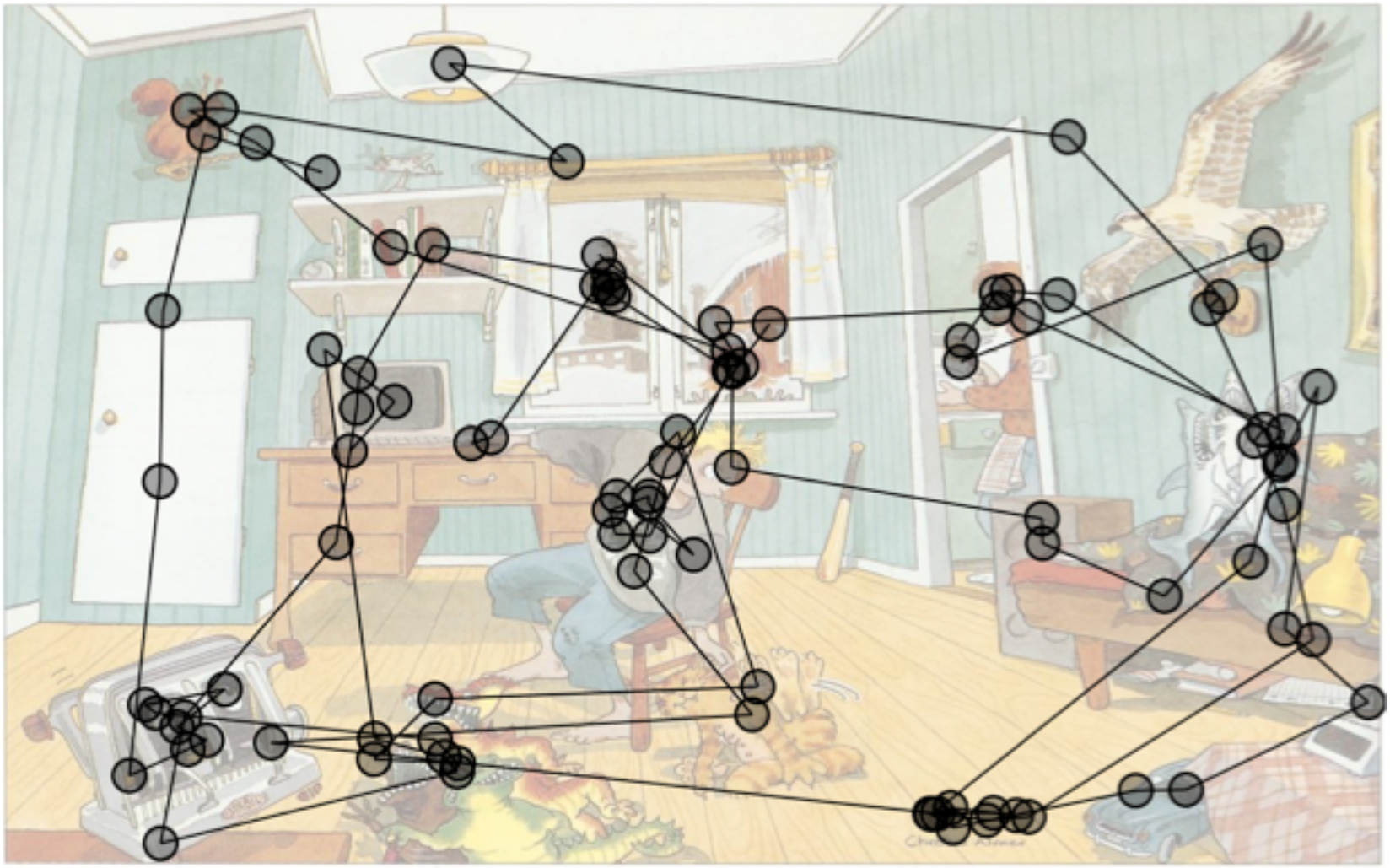
$$\sum_{i=0}^{n} F_i' \cdot \phi(H, p_i) - \sum_{i=1}^{n} d_i \cdot \phi_d(dx_i, dy_i) + b,$$

Local appearance Deformation

Structured Deep Learning



Human attention: gaze patterns



[Johansson, Holsanova, Dewhurst, Holmqvist, 2012]

Local representations

(Before 2012)

Deep Learning (Global)

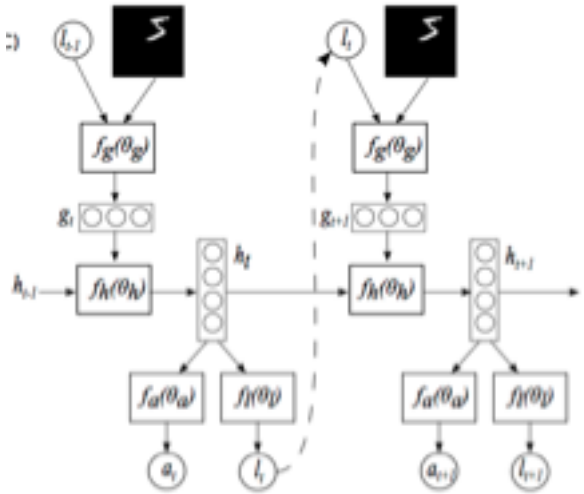
(Mostly after 2012)

Deep Learning (attention maps)

(~2016)

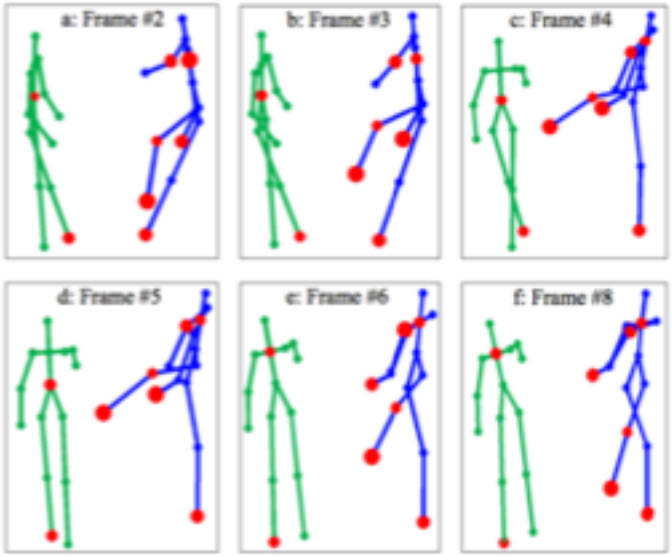
Deep Learning
(Local representations)

Hard attention



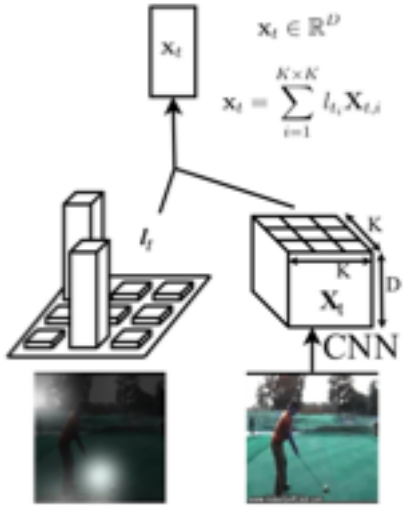
[Mnih et al., NIPS 2015]

Attention on joints



[Song et al., AAAI 2016]

Soft attention in feature maps



[Sharma et al., ICLR 2016]

Local representations

(Before 2012)

Deep Learning (Global)

(Mostly after 2012)

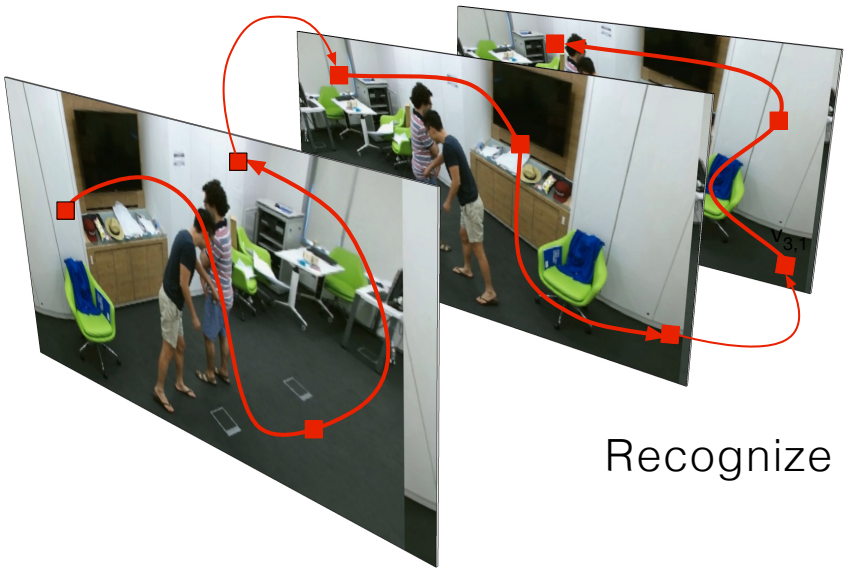
Deep Learning (attention maps)

(~2016)

Deep Learning
(Local representations)

Objective: fully trainable high-capacity local representations

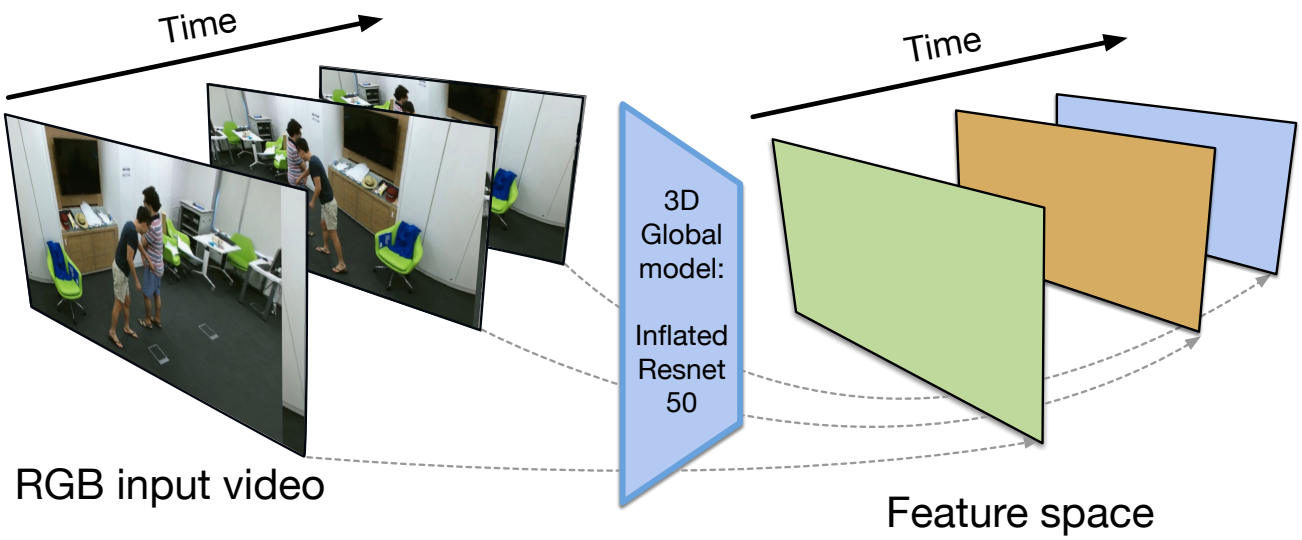
1. Learn where to attend
2. Learn how to track attended points
3. Learn how to recognize from a local distributed representation



Recognize activity

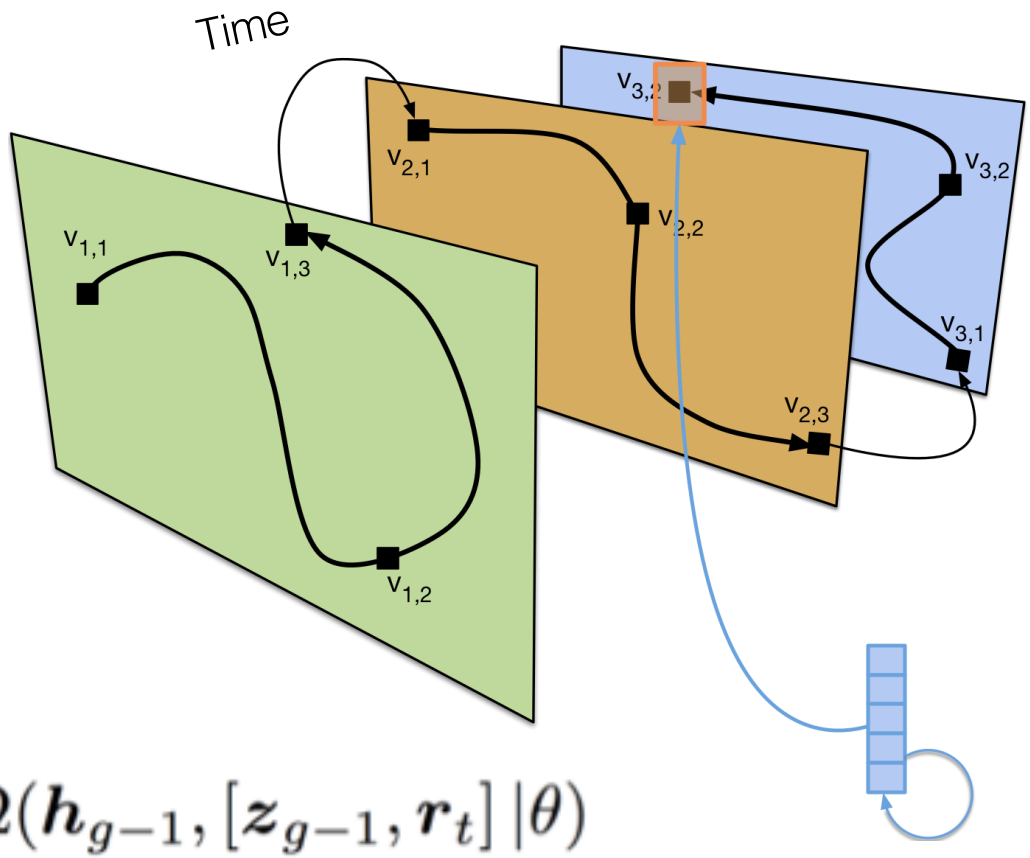
[Baradel, Wolf, Mille, Taylor, CVPR 2018]

Attention in feature space



[Baradel, Wolf, Mille, Taylor, CVPR 2018]

Unconstrained differentiable attention



$$\boldsymbol{h}_g = \Omega(\boldsymbol{h}_{g-1}, [\boldsymbol{z}_{g-1}, \boldsymbol{r}_t] \,|\theta)$$

$$\boldsymbol{l}_g = W_l^\top [\boldsymbol{h}_g, \boldsymbol{c}_t]$$

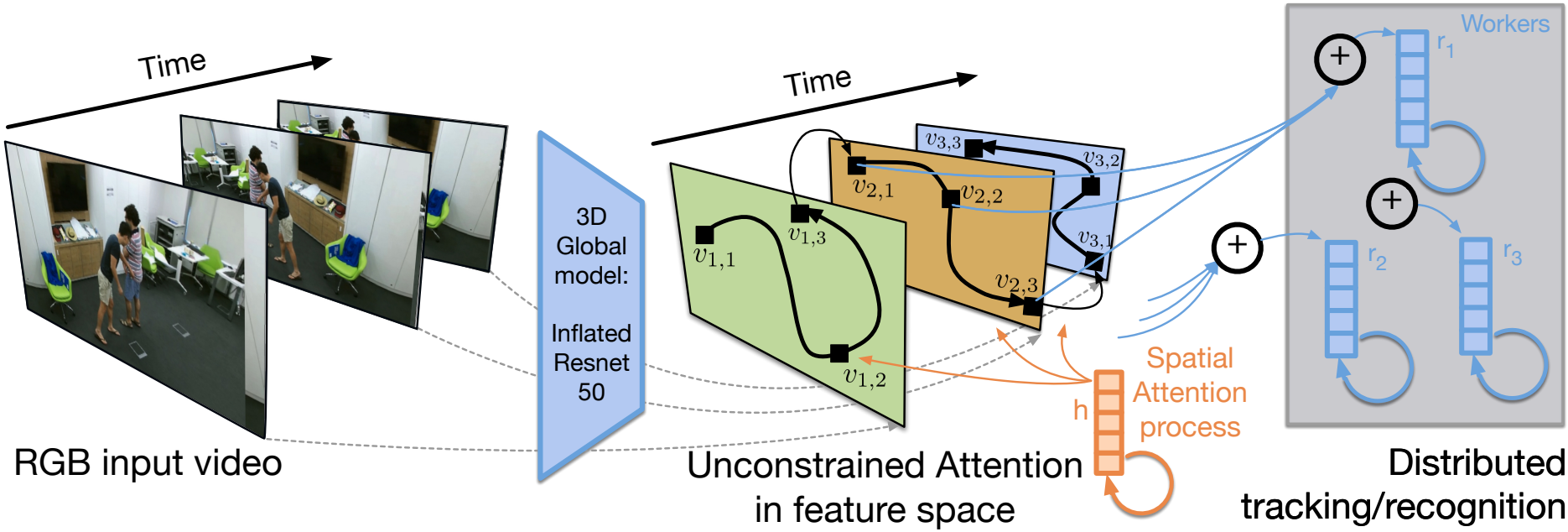
Hidden state from recurrent recognizers (workers)

Frame context

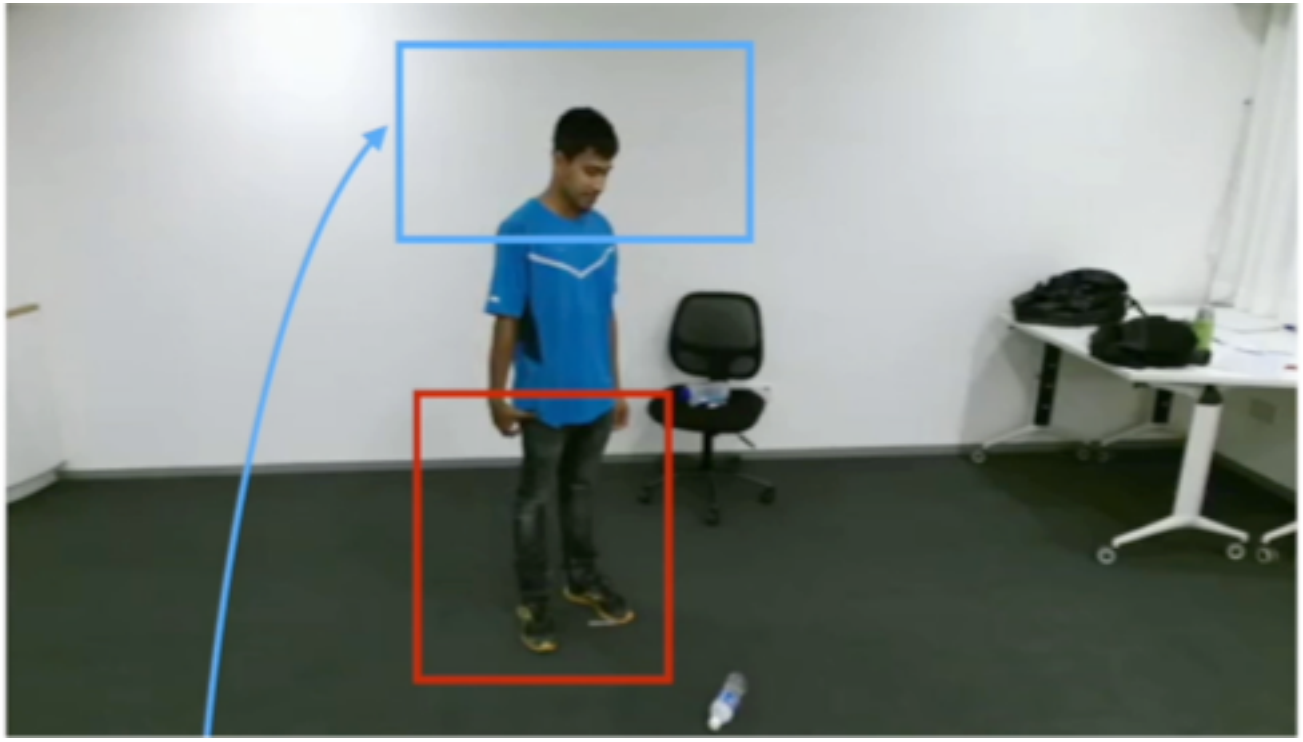
"Differentiable crop »
(Spatial Transformer Network)

[Baradel, Wolf, Mille, Taylor, CVPR 2018]

Distributed recognition



Results



State-of-the-art comparaison

Methods	Pose	RGB	CS	CV	Avg
Lie Group [40]	✓	-	50.1	52.8	51.5
Skeleton Quads [10]	✓	-	38.6	41.4	40.0
Dynamic Skeletons [14]	✓	-	60.2	65.2	62.7
HBRNN [9]	✓	-	59.1	64.0	61.6
Deep LSTM [32]	✓	-	60.7	67.3	64.0
Part-aware LSTM [32]	✓	-	62.9	70.3	66.6
ST-LSTM + TrustG. [26]	✓	-	69.2	77.7	73.5
STA-LSTM [35]	✓	-	73.2	81.2	77.2
Ensemble TS-LSTM [24]	✓	-	74.6	81.3	78.0
GCA-LSTM [27]	✓	-	74.4	82.8	78.6
JTM [41]	✓	-	76.3	81.1	78.7
MTLN [18]	✓	-	79.6	84.8	82.2
VA-LSTM [47]	✓	-	79.4	87.6	83.5
View-invariant [28]	✓	-	80.0	87.2	83.6
DSSCA - SSLM [33]	✓	✓	74.9	-	-
Hands Attention [5]	✓	✓	84.8	90.6	87.7
C3D†	-	✓	63.5	70.3	66.9
Resnet50+LSTM†	-	✓	71.3	80.2	75.8
Glimpse Clouds	-	✓	**86.6**	**93.2**	**89.9**

Table 1. Results on the NTU RGB+D dataset with Cross-Subject and Cross-View settings (accuracies in %); († indicates method has been re-implemented).

Figure 1. Results on Northwestern-UCLA Multiview Action 3D, Cross-View (accuracy in %). V=Visual(RGB), D=Depth, P=Pose.

Methods	Data	$V^3_{1,2}$	$V^2_{1,3}$	$V^1_{2,3}$	Avg
DVV [5]	D	58.5	55.2	39.3	51.0
CVP [11]	D	60.6	55.8	39.5	52.0
AOG [10]	D	45.2	-	-	-
HPM+TM [8]	D	**91.9**	75.2	71.9	79.7
Lie group [9]	P	74.2	-	-	-
HBRNN-L [1]	P	78.5	-	-	-
Enhanced viz. [6]	P	86.1	-	-	-
Ensemble TS-LSTM [3]	P	89.2	-	-	-
Hankelets [4]	V	45.2	-	-	-
nCTE [2]	V	68.6	68.3	52.1	63.0
NKTM [7]	V	75.8	73.3	59.1	69.4
Global model	V	85.6	84.7	79.2	83.2
Glimpse Clouds	V	90.1	**89.5**	**83.4**	**87.6**

SOTA results on two datasets NTU and N-UCLA
Larger difference between Glimpse clouds and global model on N-UCLA

[Baradel, Wolf, Mille, Taylor, CVPR 2018]

Ablation study

Glimpse	Type of attention	CS	CV	Avg
3D tubes	Attention	85.8	92.7	89.2
Seq. 2D	Random sampling	80.3	87.8	84.0
Seq. 2D	Saliency	86.2	92.9	89,5
Seq. 2D	**Attention**	**86.6**	**93.2**	**89.9**

Table 3. Results on the NTU: different attention and alternative strategies.

Methods	L_D	L_P	L_G	CS	CV	Avg
Global model	✓	-	-	84.5	91.5	88.0
Global model	✓	✓	-	85.5	92.1	88.8
Glimpse Clouds	✓	-	-	85.7	92.5	89.1
Glimpse Clouds	✓	✓	-	86.4	93.0	89.7
Glimpse Clouds	✓	-	✓	86.1	92.9	89.5
Glimpse Clouds	✓	✓	✓	86.6	93.2	89.9

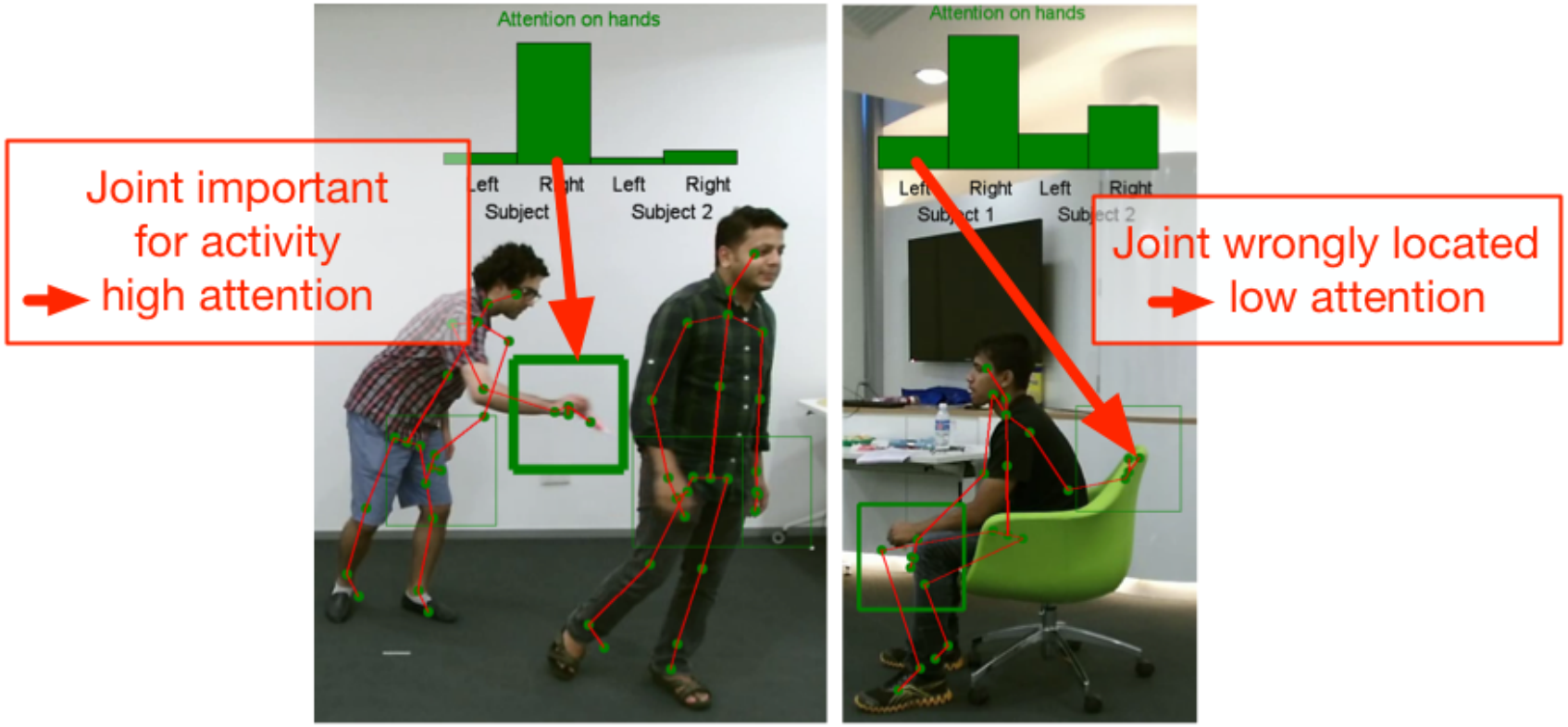
Table 1. Results on NTU: ablation study

Methods	Global model	Spatial Attention	Soft Workers	Loss on Pose	CS	CV	Avg
Global model only	✓	-	-	-	84.5	91.5	88.0
Global model only	✓	-	-	✓	85.5	92.2	88.8
$\sum$ Glimpses + GRU	-	✓	-	✓	85.8	92.4	89.1
Glimpse clouds	-	✓	✓	✓	**86.6**	**93.2**	**89.9**
Glimpse clouds + Global model	-	✓	✓	✓	86.6	93.2	89.9

Table 2. Results on NTU: ablation study.

[Baradel, Wolf, Mille, Taylor, CVPR 2018]

Pose conditioned attention



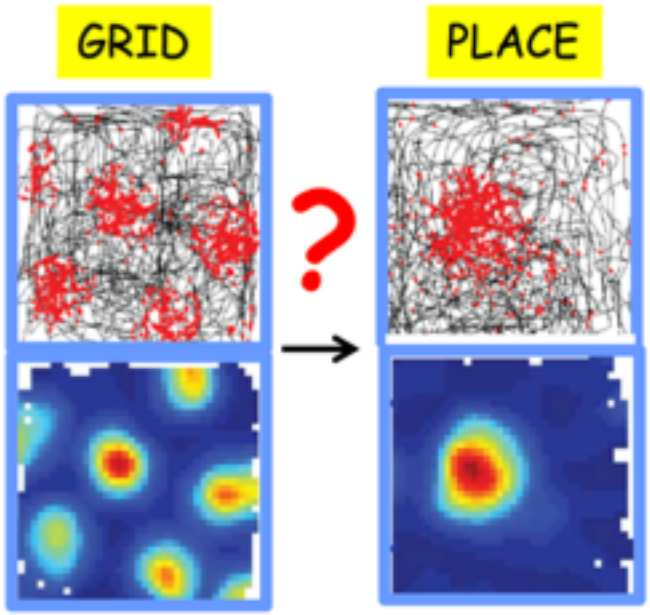
[Baradel, Wolf, Mille, Taylor, BMVC 2018]

AI vs. NI

Photo: A. Mahmoud
May-Britt Moser
Prize share: 1/4

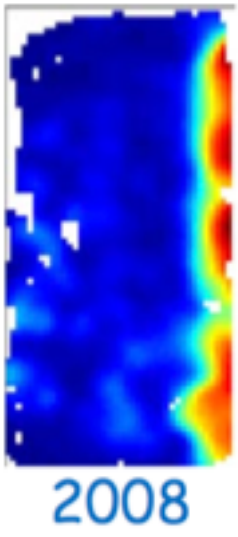
Photo: A. Mahmoud
Edvard I. Moser
Prize share: 1/4

2014 Nobel Prize in Medecine



Head direction

Border cells

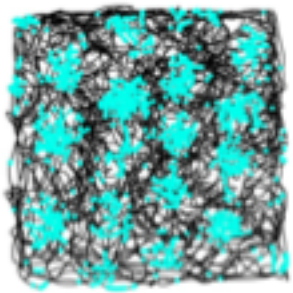


AI vs. NI

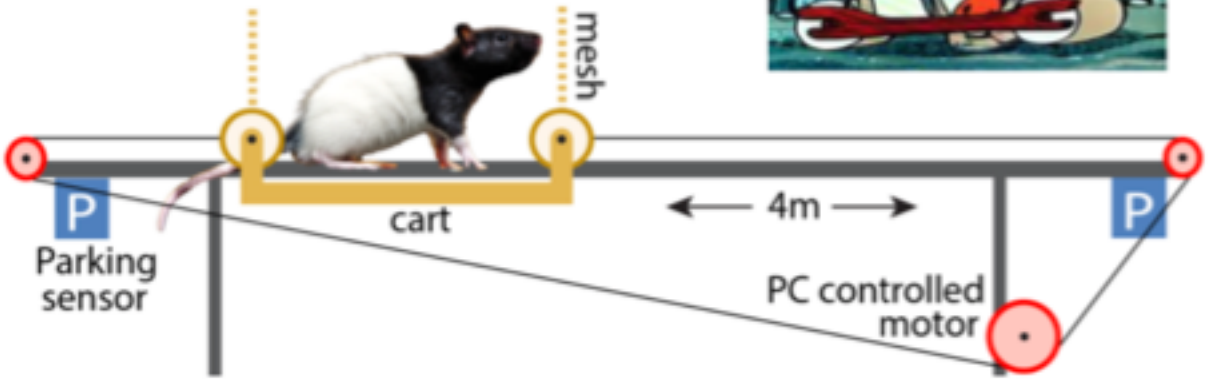
Photo: A. Mahmoud
May-Britt Moser
Prize share: 1/4

Photo: A. Mahmoud
Edvard I. Moser
Prize share: 1/4

2014 Nobel Prize in Medecine

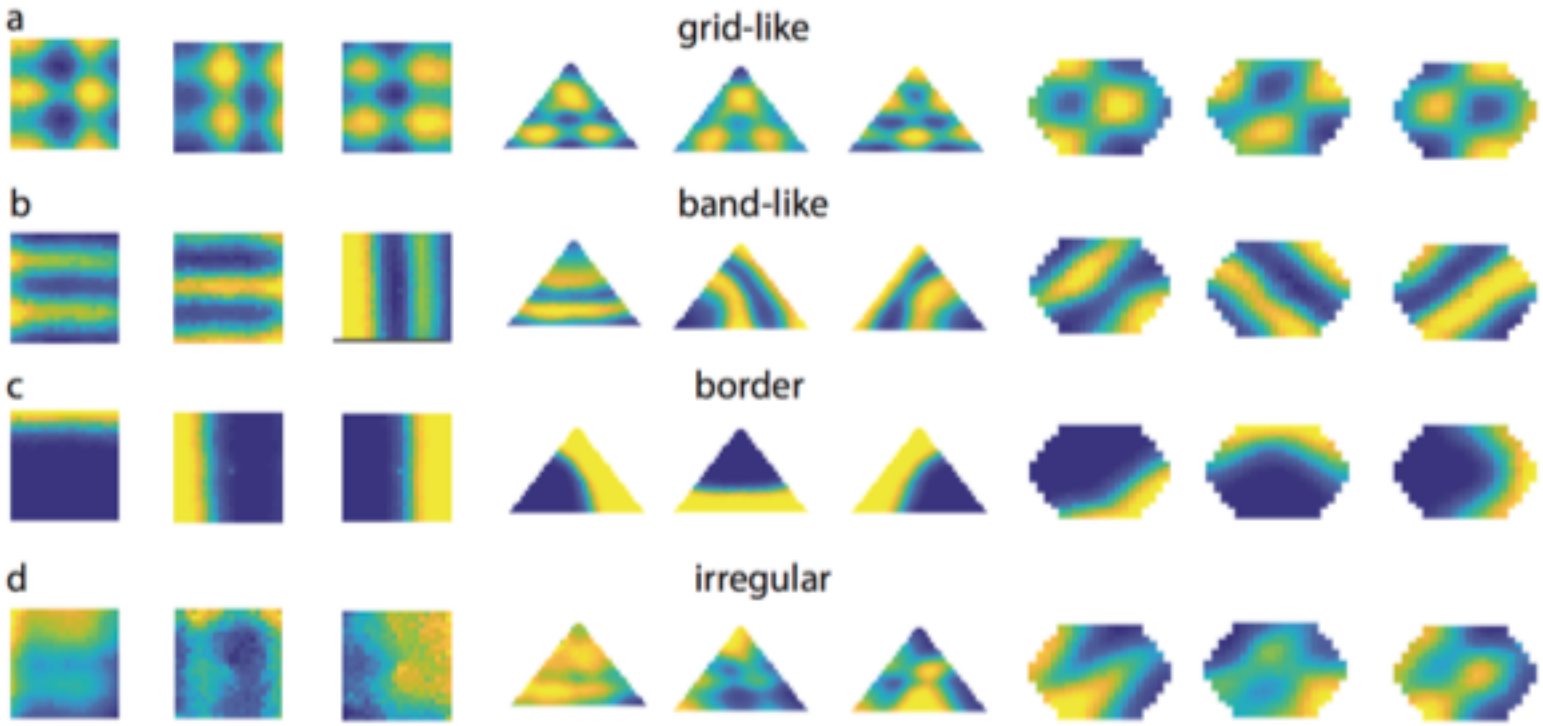


Speed cells are necessary for updating the grid pattern in accordance with the animal's movement (distance=speed x time)



AI vs. NI

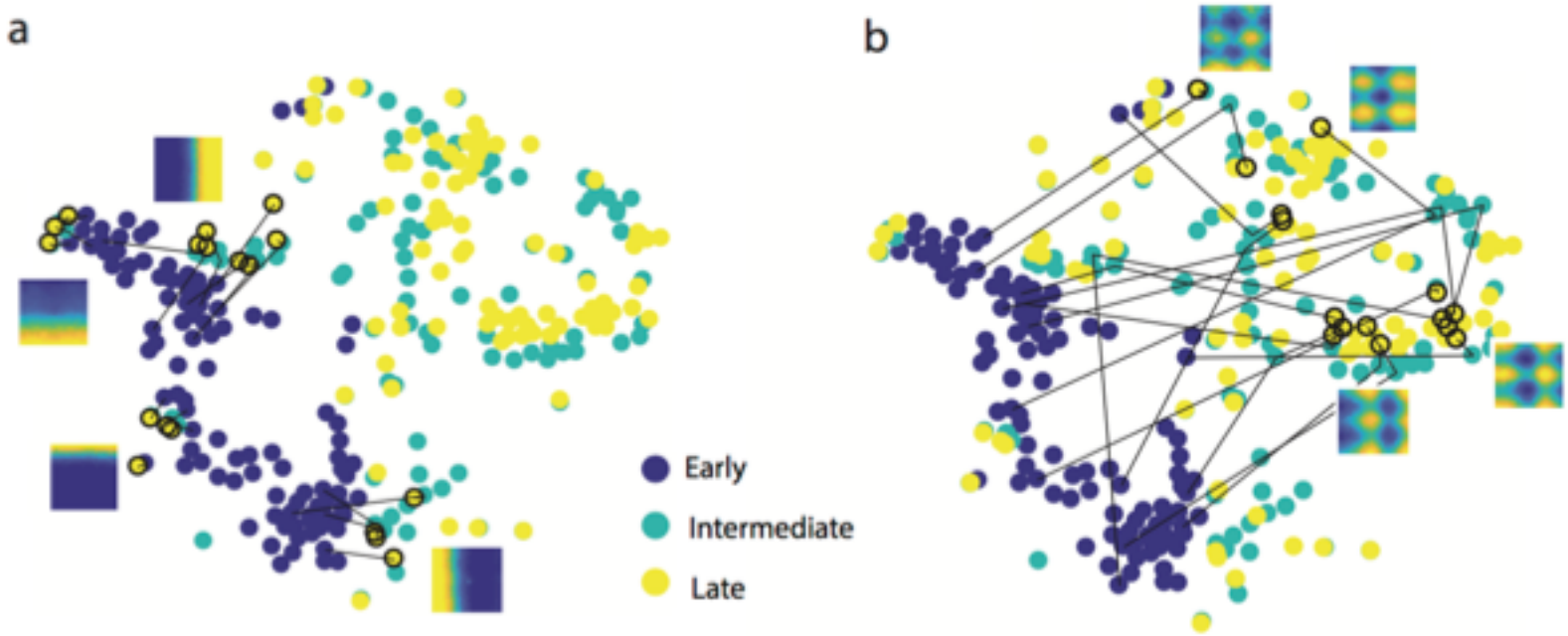
2018 : discoverty of the same cells in neural networks trained on similar tasks.



[Cueva, Wei, ICLR 2018]

AI vs. NI

Emergence of the different types of cells in the same order.

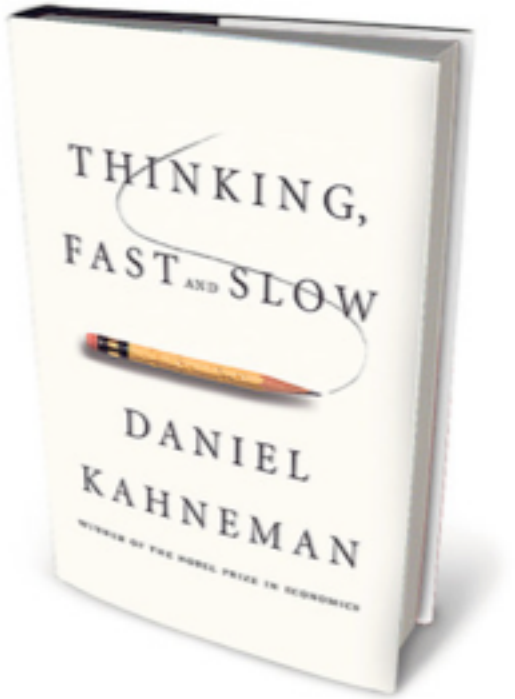


[Cueva, Wei, ICLR 2018]

Reasoning : what happened?

Human psychology

- Daniel Kahnemann (Nobel prize in 2002)
- Book: "Thinking Fast and Slow"



Cognitive tasks

24*17 = ?

Two systems

System 1

- Continuously monitors environment (and mind)
- No specific attention
- Continuously generates assessments / judgments w/o efforts, even in the presence of low data. Jumps to conclusions
- Prone to errors. No capabilities for statistics

System 2

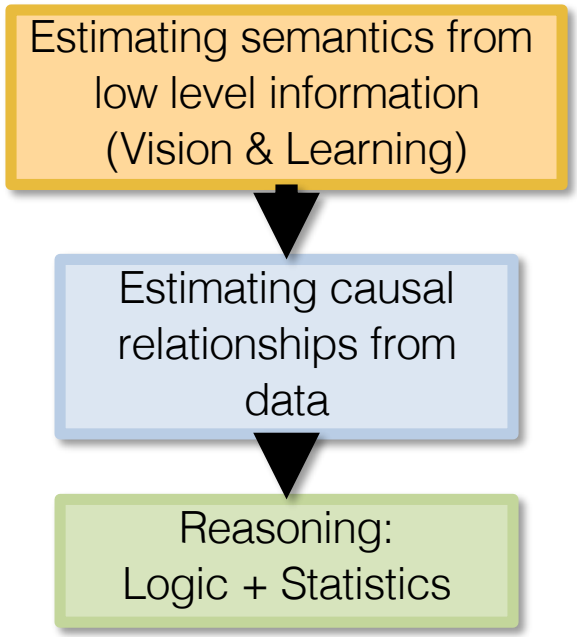
- Receives questions or generates them
- Directs attention and searches memory to find answers
- Requires (eventually a lot of) effort
- More reliable

Where is ML today?

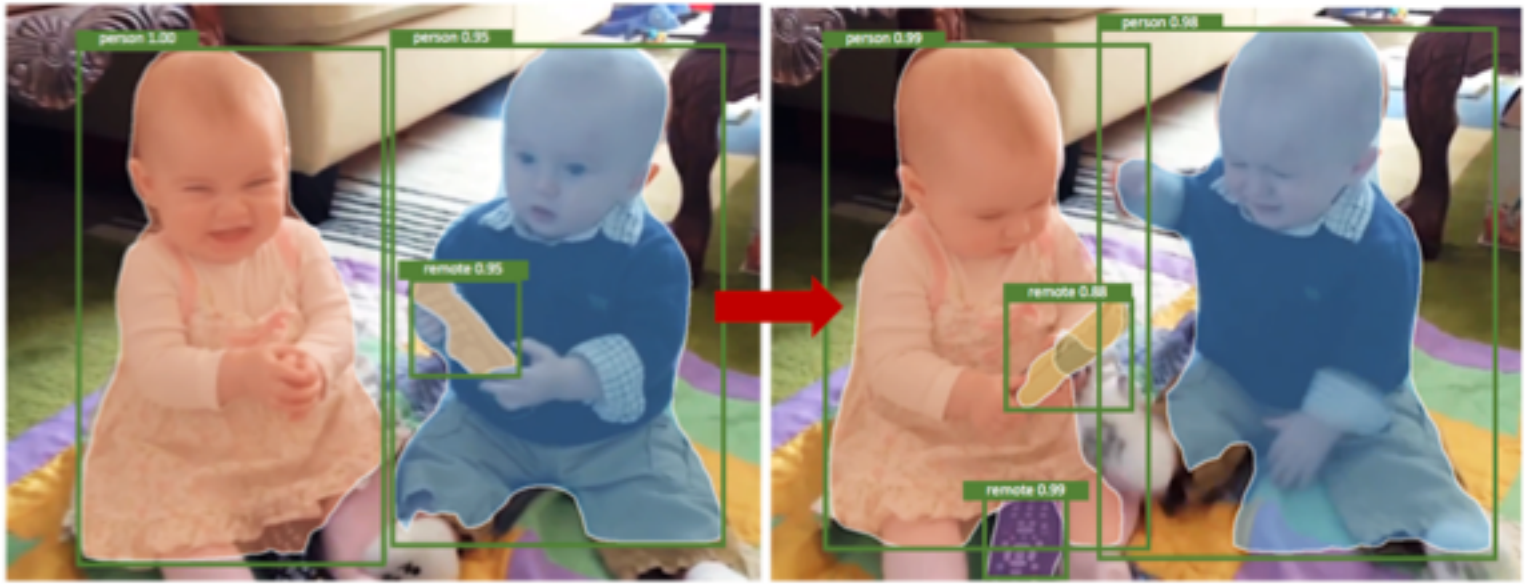
Claim: AI requires a combination of

- Extraction of high-level information from high-dimensional input (visual, audio, language): **machine learning**
- High-level reasoning: **compare, assess, focus attention**, perform **logical deductions**

Roadmap:

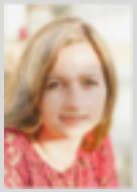


Object level Visual Reasoning



[Baradel, Neverova, Wolf, Mille, Mori, ECCV 2018]

Fabien Baradel
Phd @ LIRIS, INSA-Lyon



Natalia Neverova
Facebook AI Research, Paris

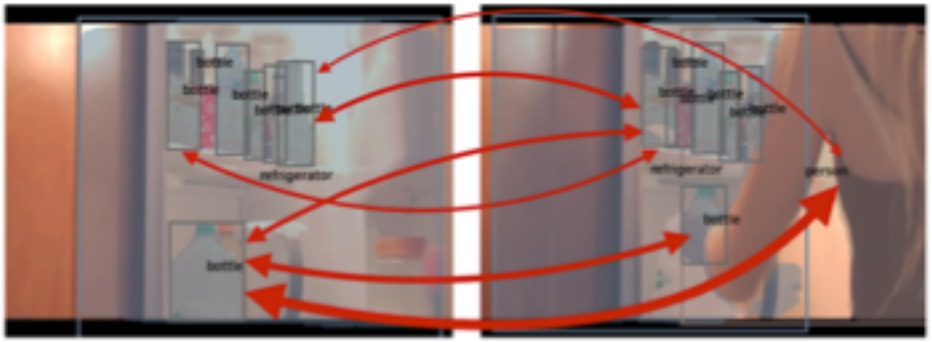
Christian Wolf
INRIA Chroma

Julien Mille
LI, INSA VdL

Greg Mori
Simon Fraser University, Canada

Object level Visual Reasoning

hand-cup interaction (VLOG)



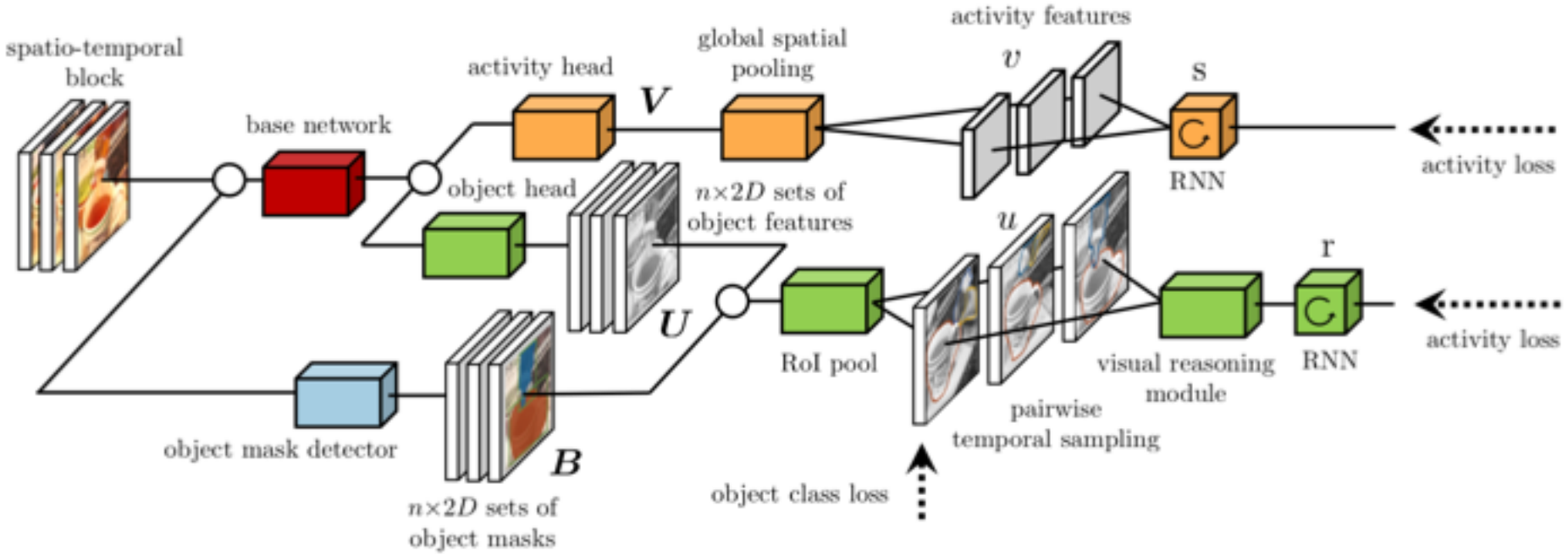
hand-bottle interaction (VLOG)

hand-bed interaction (VLOG)

tilting something with something on it until it falls off (SS)

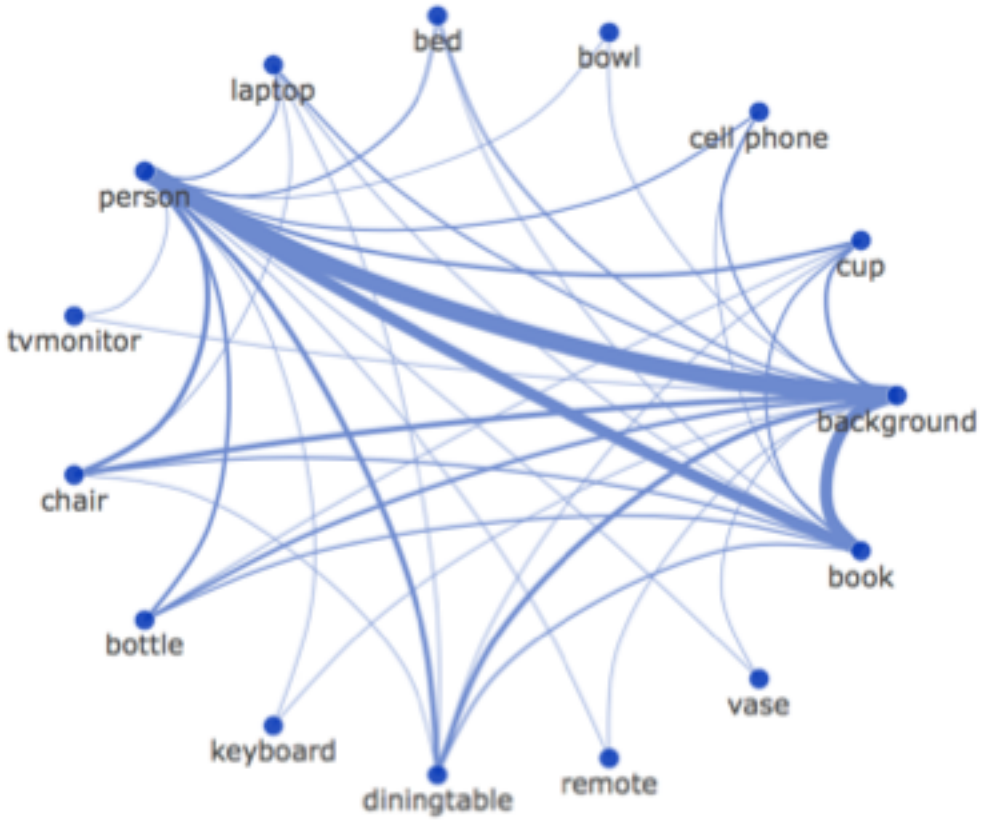
[Baradel, Neverova, Wolf, Mille, Mori, ECCV 2018]

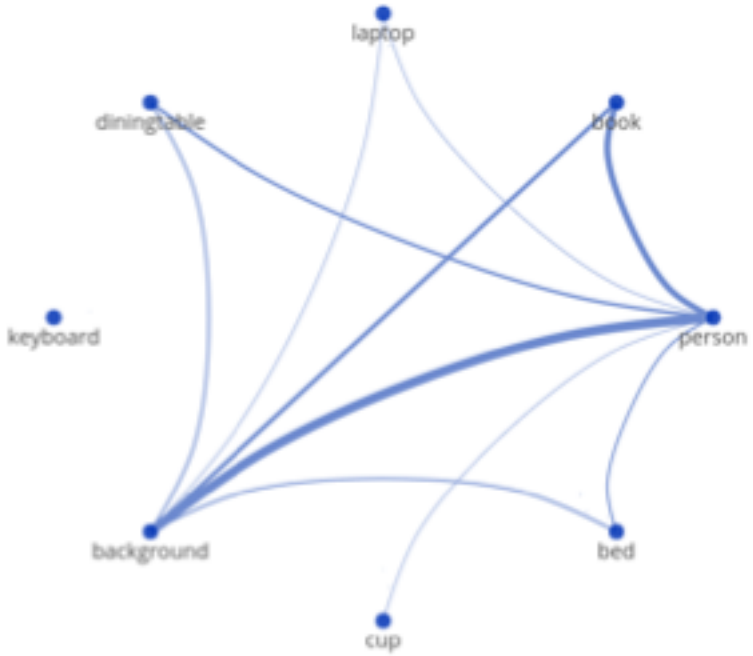
Object level Visual Reasoning



[Baradel, Neverova, Wolf, Mille, Mori, ECCV 2018]

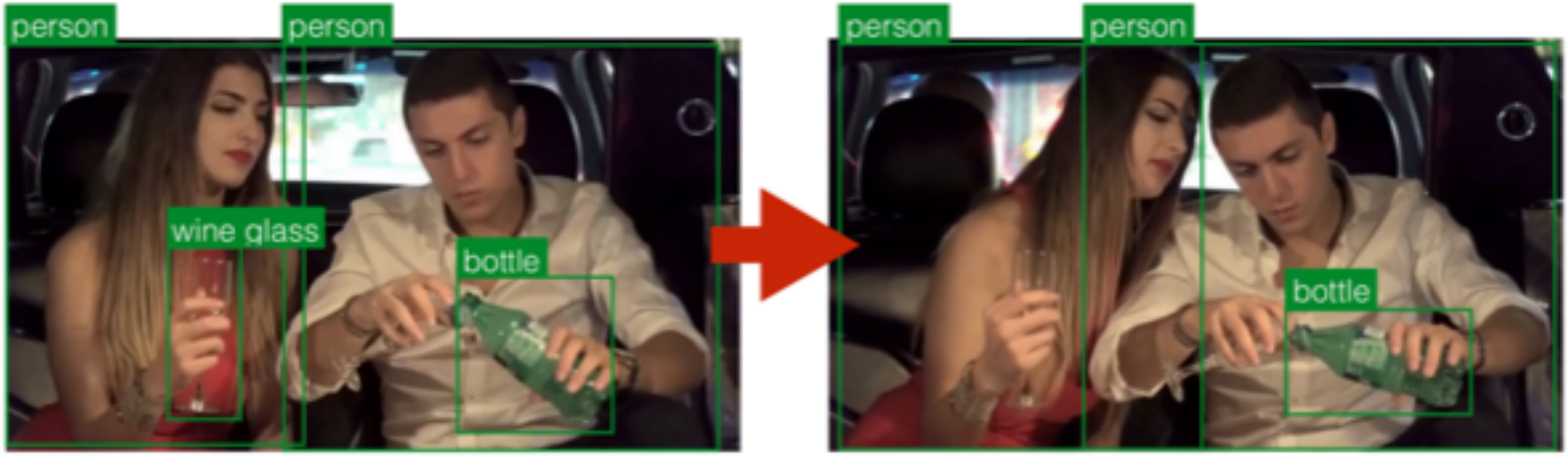
Learned interactions





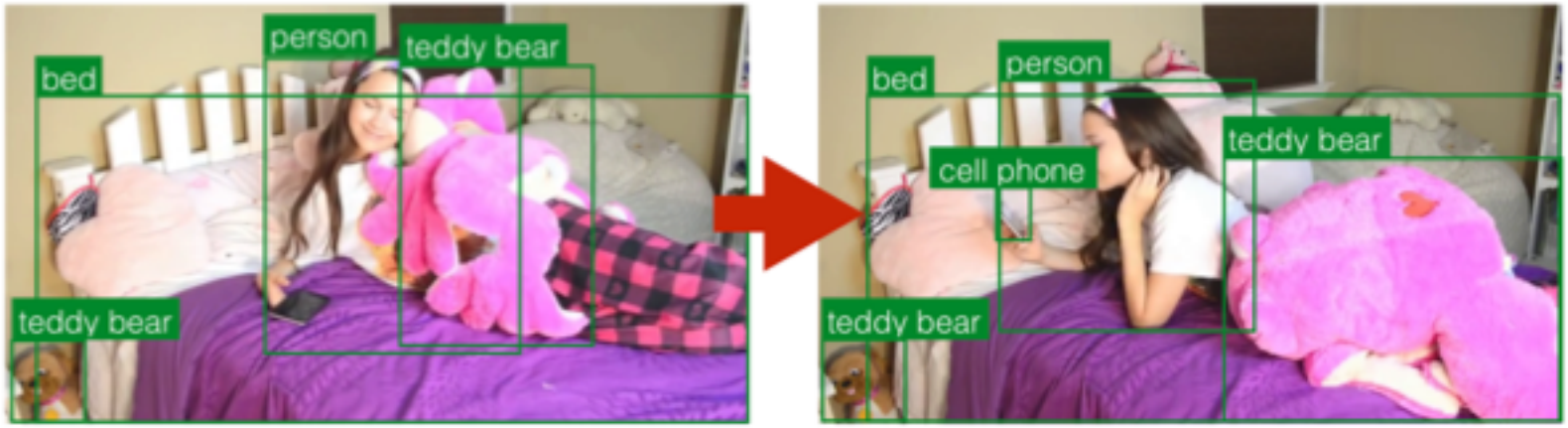
Class: person-book interaction

Failure cases



Confusion between semantically similar objects

- prediction of *hand-cup-contact* instead of *hand-glass-contact*



Small size object

- *hand-cell-phone* contact not detected

Results

Methods	Top1
C3D + Avg [5]	21.50
I3D [5]	27.63
MultiScale TRN [39]	33.60
Ours	**34.32**

Something-something dataset

R50 [45]	40.5
I3D [3]	39.7
Ours	**41.7**

VLOG dataset

Methods	Top1
R18 [44]*	32.05
I3D-18 [3]*	34.20
Ours	**40.89**

EPIC Kitchen dataset

	Nb. head		Object type		f_ϕ		Pairwise relations	Results	
	1	2	Pixel	COCO	RNN	MLP		VLOG	Something
Baseline	-	-	-	-	-	-	-	29.92	33.43
Variant 1	✓	-	-	✓	✓	-	✓	32.01	35.09
Variant 2	-	✓	✓	-	✓	-	✓	31.36	35.15
Variant 3	-	✓	-	✓	-	✓	✓	32.38	34.15
Variant 4	-	✓	-	✓	✓	-	-	31.82	34.65
Ours	-	✓	-	✓	✓	-	✓	**33.75**	**36.12**

Conclusion

- We propose a models which recognize activities from
  - a cloud of unconstrained feature points
  - Interactions between spatially well defined objects

- Visual spatial attention is useful and competitive compared to pose

- State of the art performance on 5 datasets (NTU RGB-D, Northwestern UCLA, VLOG, Something-Something, Epic Kitchen)

- Reasoning is key component of human cognition, also important for IA systems