Glimpse Clouds:
Human Activity Recognition from Unstructured Feature Points

Fabien Baradel[1], Christian Wolf[1,2], Julien Mille[3], Graham W. Taylor[4,5]

1 Univ. Lyon, INSA-Lyon, CNRS, LIRIS, F-69621, Villeurbanne, France.
2 INRIA, CITI Laboratory, Villeurbanne, France
3 Laboratoire d'Informatique de l'Univ. de Tours, INSA Centre Val de Loire, 41034 Blois, France.
4 School of Engineering, Univ. of Guelph, Guelph, Ontario, Canada.
5 Vector Institute, Toronto, Ontario, Canada

fabien.baradel@liris.cnrs.fr   github.com/fabienbaradel/glimpse_clouds

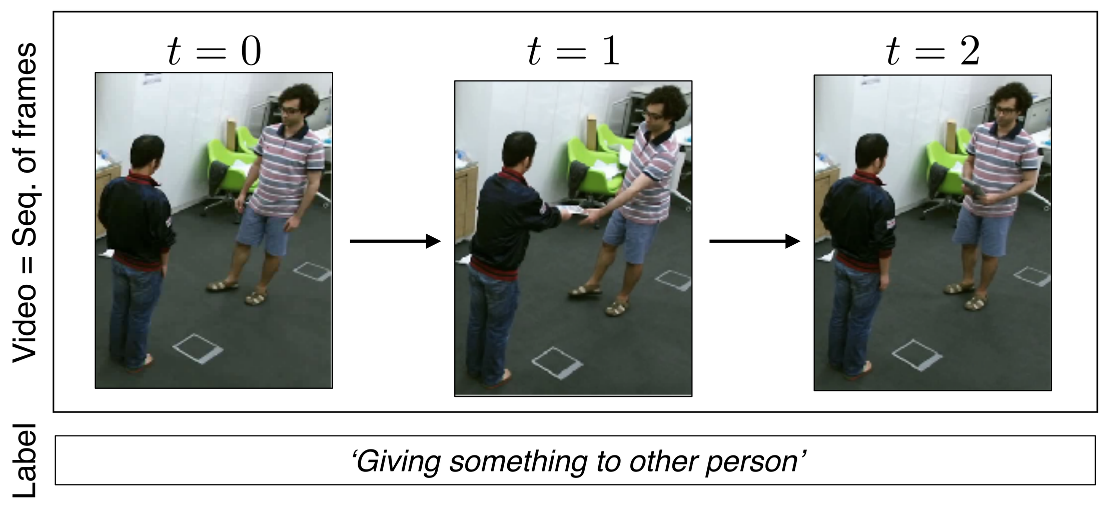
PROBLEM DEFINITION & MOTIVATIONS

Overview

- **Video** Understanding
- **Human** Action Recognition
- **Fine-grained** understanding

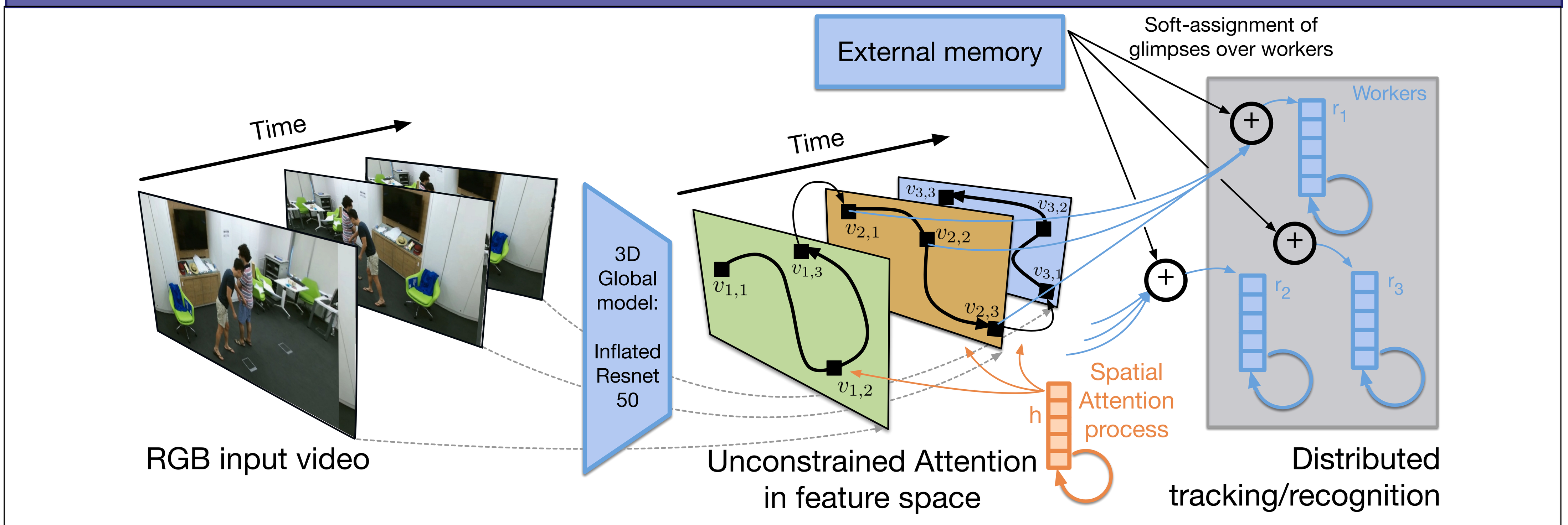
Main challenges

- **High dimensional** data
- **Only few parts of the video are important**
- **Combining spatial and temporal information**



Problem statement:
How can **an attention process selects local glimpse points in a video**?

MAIN IDEA



Main features

1. Video is mapped to features
2. Model selects important Local Parts
3. Local features assigned to a set of workers (distributed recognition)

Advantages

1. Fine-grained features
2. Workers automatically focus on discriminative entities
3. Attention process can be visualized (see below)

PROPOSED APPROACH

DIFFERENTIABLE GLIMPSE

How to extract more fine-grained features in an automatic way?
Extraction of local crops

Recurrent predictions of the glimpse locations

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

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

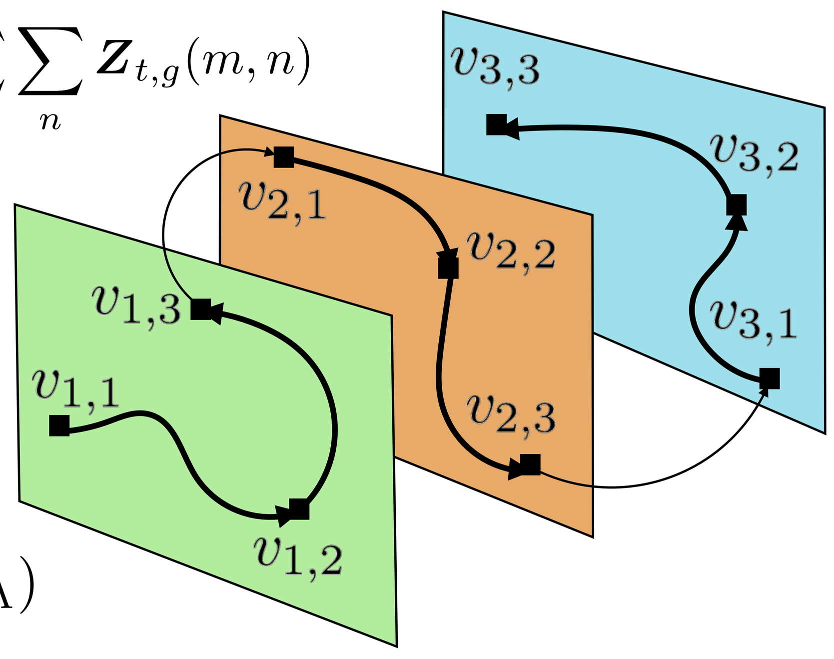
Advantages

1. *Automatic attention*
2. *Learn where to look*
3. *Fully-differentiable crops*

Zoom with Spatial Transformer

$$\boldsymbol{Z}_{t,g} = \mathrm{STN}(\boldsymbol{Z}_t, \boldsymbol{l}_{t,g})$$

$$\boldsymbol{z}_{t,g} = \Gamma(\boldsymbol{Z}_{t,g}) = \frac{1}{H'W'} \sum_m \sum_n \boldsymbol{Z}_{t,g}(m,n)$$

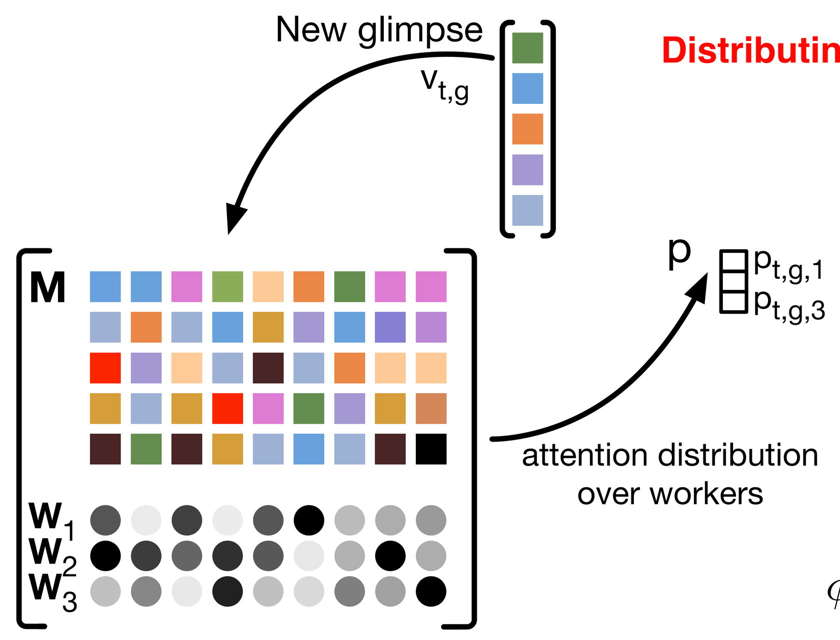


What and where features

$$\boldsymbol{v}_{t,g} = \boldsymbol{z}_{t,g} \otimes \Lambda(\boldsymbol{l}_{t,g}|\theta_\Lambda)$$

MEMORY NETWORK

How to aggregate local features?
Distributing the recognition task over several workers



Independent Recurrent Workers

$$\boldsymbol{r}_{t,c} = \Psi_c(\boldsymbol{r}_{t-1,c}, \tilde{\boldsymbol{v}}_{t,c}|\theta_{\Psi_c})$$

$$\boldsymbol{r}_t = \sum_c \boldsymbol{r}_{t,c}$$

Soft-assignment of glimpses to workers

$$\tilde{\boldsymbol{v}}_{t,c} = \boldsymbol{V}_t \boldsymbol{p}_{t,c}$$

External memory

$$\boldsymbol{M} = \{\boldsymbol{m}_k\}$$

Glossary

C : worker
g : glimpse

Distance function

$$\phi(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^\top \boldsymbol{D} (\boldsymbol{x}-\boldsymbol{y})}$$

Advantages

1. *Recognition distributed*
2. *Each worker specialized on a subtask*
3. *Fully-differentiable operations through External Memory*

Importance of a glimpse for a worker

$$p_{t,c,g} = \sigma_\alpha \left(\sum_k e^{-t^{\boldsymbol{m}_k}} \times w_{c,k} \left[1 - \phi(\boldsymbol{v}_{t,g}, \boldsymbol{m}_k)\right] \right)$$

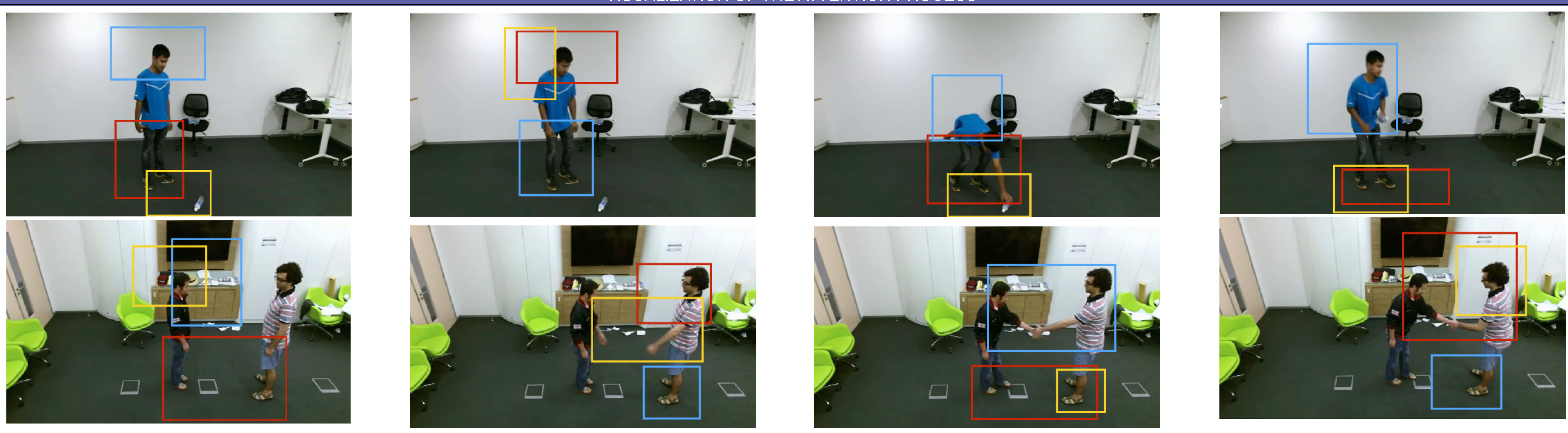
LOSS FUNCTION

$$\mathcal{L} = \mathcal{L}_D(\hat{\boldsymbol{y}}, \boldsymbol{y}) + \mathcal{L}_P(\hat{\boldsymbol{y}}^p, \boldsymbol{y}^p) + \mathcal{L}_G(\boldsymbol{l}, \boldsymbol{y}^p)$$

Activity prediction — **Pose prediction**

Attracting glimpses to humans
- **encourages diversity**
- **not too far from humans**

VISUALIZATION OF THE ATTENTION PROCESS



EXPERIMENTAL RESULTS

State-of-the-art

- Two datasets: NTU and N-UCLA
- RGB only (no D, no pose) during testing
- outperforms multi-modal approaches
- +1.9 and +4.4 vs Global Model

Northwestern-UCLA

Methods	Data	$V^3_{1,2}$	$V^2_{1,3}$	$V^1_{2,3}$	Avg
DVV	D	58.5	55.2	39.3	51.0
CVP	D	60.6	55.8	39.5	52.0
AOG	D	45.2	-	-	-
HPM+TM	P	91.9	75.2	71.9	79.7
Lie group	P	74.2	-	-	-
HBRNN-L	P	78.5	-	-	-
Enhanced viz.	P	86.1	-	-	-
Ensemble TS-LSTM	P	89.2	-	-	-
Hankelets	V	45.2	-	-	-
nCTE	V	68.6	68.3	52.1	63.0
NKTM	V	75.8	73.3	59.1	69.4
Global model	V	85.6	84.7	79.2	83.3
Glimpse Clouds	**V**	**90.1**	**89.5**	**83.4**	**87.6**

NTU-RGB+D

Methods	Pose	RGB	CS	CV	Avg
Lie Group	✓	-	50.1	52.8	51.5
Skeleton Quads	✓	-	38.6	41.4	40.0
Dynamic Skeletons	✓	-	60.2	65.2	62.7
HBRNN	✓	-	59.1	64.0	61.6
Deep LSTM	✓	-	60.7	67.3	64.0
Part-aware LSTM	✓	-	62.9	70.3	66.6
ST-LSTM + TrustG.	✓	-	69.2	77.7	73.5
STA-LSTM	✓	-	73.2	81.2	77.2
Ensemble LSTM	✓	-	74.6	81.3	78.0
GCA-LSTM	✓	-	74.4	82.8	78.6
JTM	✓	-	76.3	81.1	78.7
MTLN	✓	-	79.6	84.8	82.2
VA-LSTM	✓	-	79.4	87.6	83.5
View-invariant	✓	-	80.0	87.2	83.6
DSSCA-SSLM	✓	✓	74.9	-	-
STA-Hands	x	x	82.5	88.6	85.6
Hands Attention	✓	✓	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**

Different attention strategies

Glimpses	Type of attention	CS	CV	Avg
3D tubes	Attention	85.5	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**

Ablation study

- Coherent attention matters
- Recurrent action > Saliency
- Distributed workers > GRU

Ablation study

Methods	Spatial Attention	Soft Workers	L_D	L_P	L_G	CS	CV	Avg
GM	-	-	✓	-	-	84.5	91.5	88.0
GM	-	-	✓	✓	-	85.5	92.1	88.8
GM + Glimpses + GRU	-	-	✓	✓	-	85.8	92.4	89.1
GC	✓	✓	✓	-	-	85.7	92.5	89.1
GC	✓	✓	✓	✓	-	86.4	93.0	89.7
GC	✓	✓	✓	-	✓	86.1	92.9	89.5
GC	✓	✓	✓	✓	✓	**86.6**	**93.2**	**89.9**
GC + GM	✓	✓	✓	✓	✓	86.6	93.2	89.9

Code