Human Action Recognition: Pose-based Attention draws focus to Hands

Fabien Baradel*, Christian Wolf*, Julien Mille**

* Univ Lyon, INSA-Lyon, CNRS, LIRIS, F-69621, Villeurbanne, France

*** Laboratoire d'Informatique de l'Université de Tours (EA 6300), INSA Centre Val de Loire, 41034 Blois, France

Email : fabien.baradel@liris.cnrs.fr

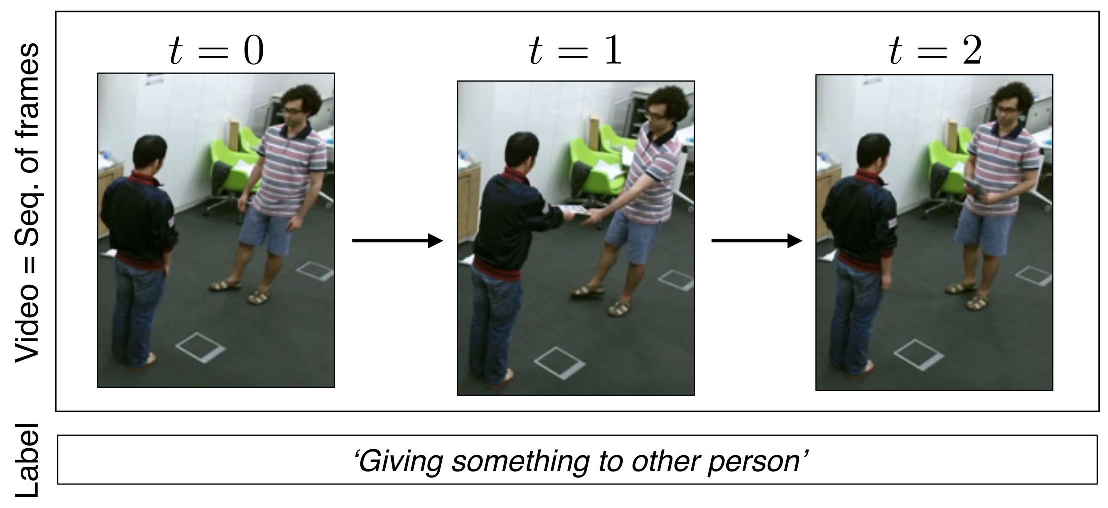
PROBLEM DEFINITION & MOTIVATIONS

Overview

- **Video** Understanding
- **Human** Action Recognition
- Video captured by **Microsoft Kinect3D** (*3D human pose - RGB - Depth*)

Main challenges

- **High dimensional** data
- **Spatio-Temporal** information
- **Noise** in the human pose



Problem statement:

How can an **attention mechanism select the most discriminative parts of the video**?

MAIN IDEA

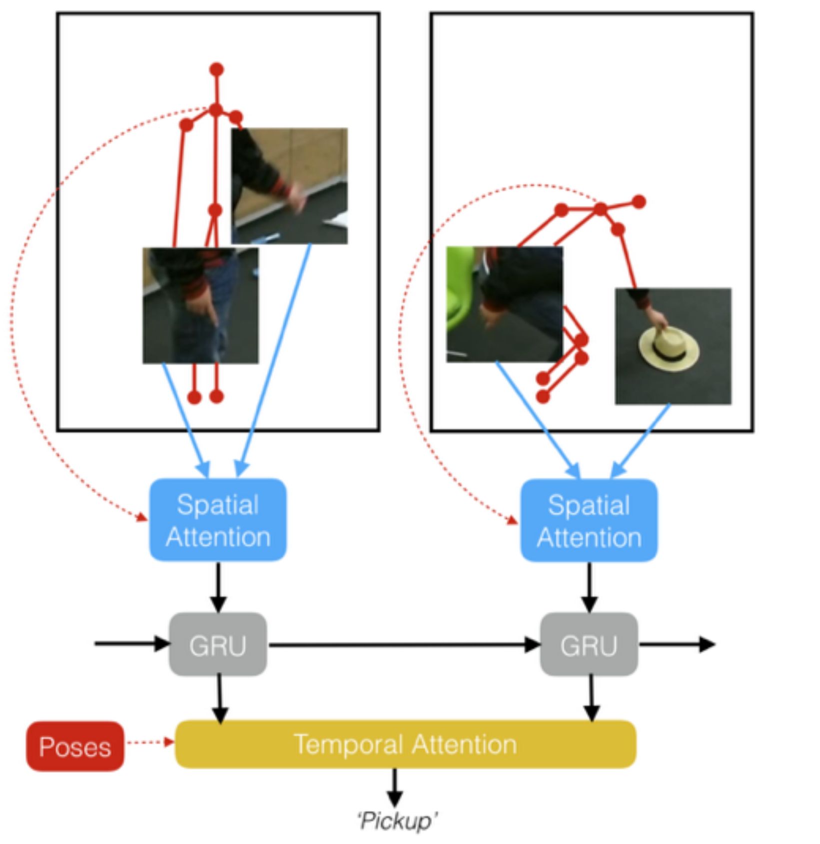
- Two modalities
  - ✓ *3D skeleton coordinates*
  - ✓ *RGB frames*
- **Two stream model**

RGB

- **Spatial attention mechanism** over RGB hands crops
- Spatial attention adjusted at each timestep
  - **Conditioned on augmented pose**
- **Temporal Attention** on hidden states
  - **Conditioned on augmented motion**

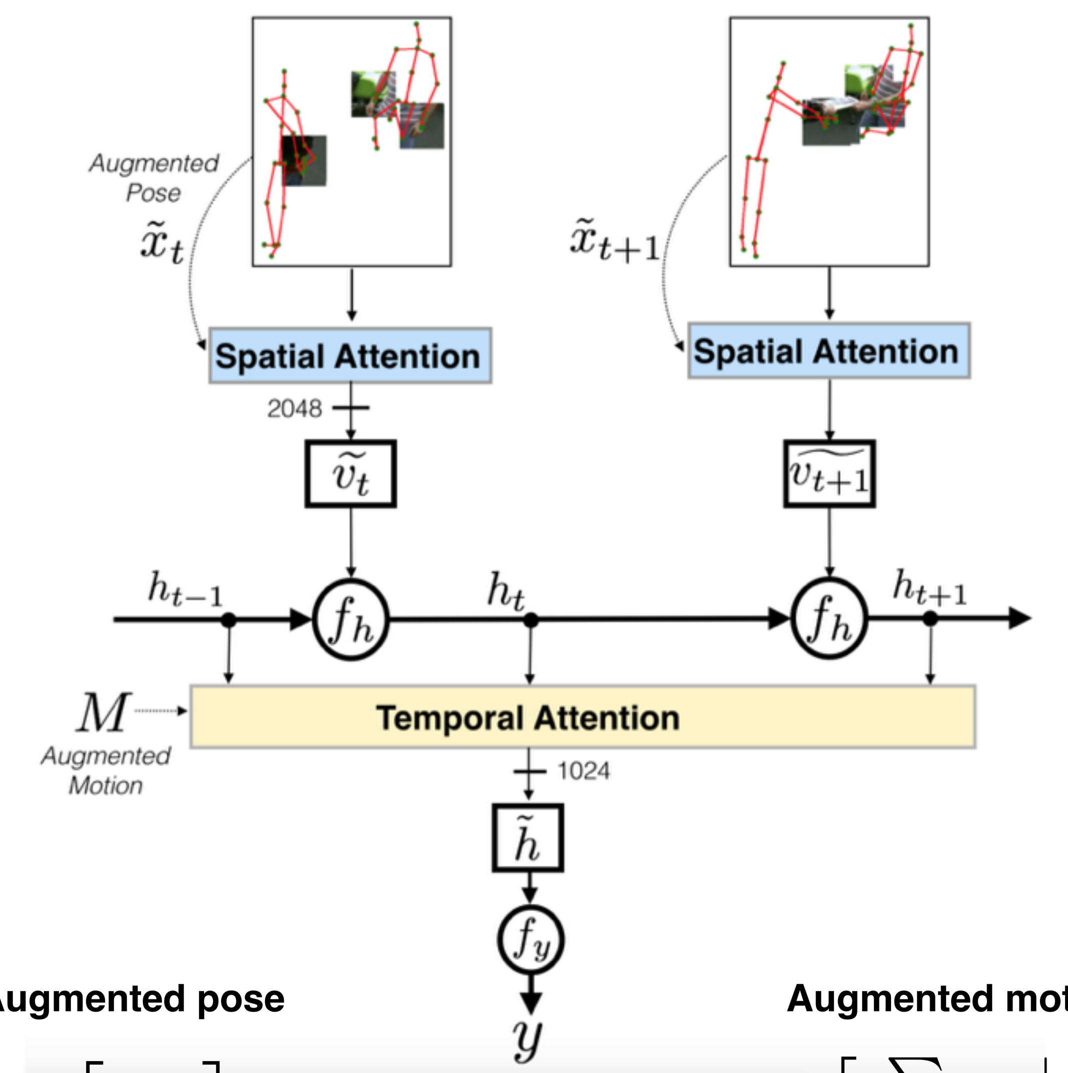
Pose

Standard Deep-GRU



PROPOSED APPROACH

STA-HANDS



Augmented pose

$$\tilde{x}_t = \begin{bmatrix} x_t \\ \dot{x}_t \\ \ddot{x}_t \end{bmatrix}.$$

Augmented motion

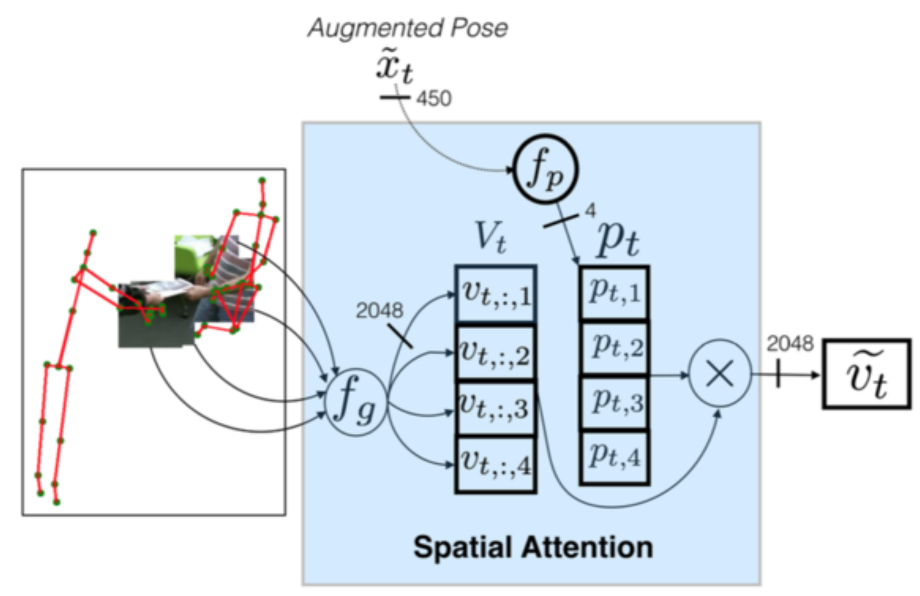
$$\tilde{m}_t = \begin{bmatrix} \sum_{j \in J} |\dot{x}_{t,j}| \\ \sum_{j \in J} |\ddot{x}_{t,j}| \end{bmatrix}.$$

$$M = \{\tilde{m}_t\}_{t=1...T}$$

ATTENTION ON HANDS

SA-Hands: Spatial Attention around Hands crops

- *Inception features* from **RGB crops around hands**
- **Attention weights** computed given
  - ✓ *augmented pose*
- Fully differentiable



Glossary:

f_g : Inception feature vector

p_t : Spatial Attention weights for each hand

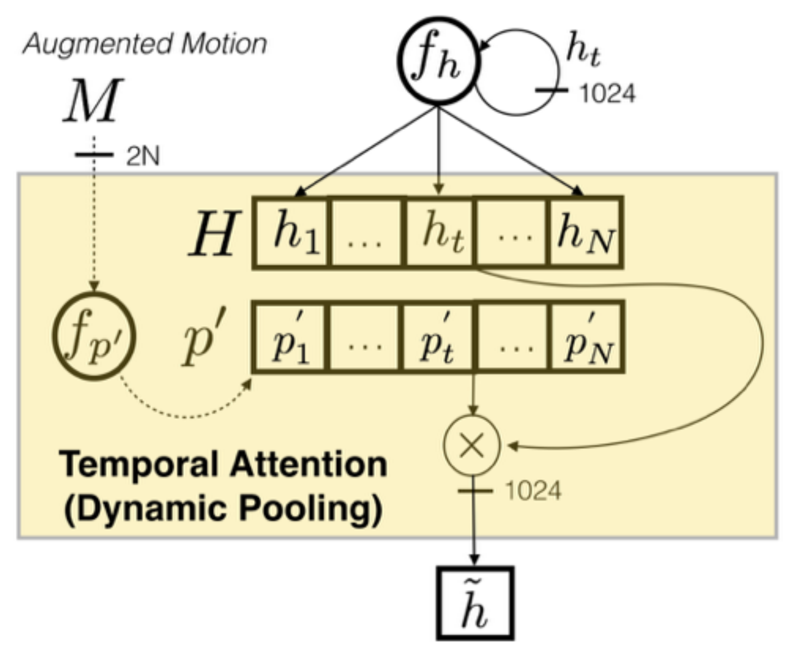
$\tilde{v}_t$: Output of the Spatial Attention framework - Input of the LSTM

f_h : GRU

$\tilde{x}_t$: Augmented Pose

Temporal Attention on LSTM features

- Can be seen as a **dynamic pooling**
- **Weighted average** of hidden states
- Given *augmented motion*
- Fully differentiable



Glossary:

h_t : hidden state at timestep t

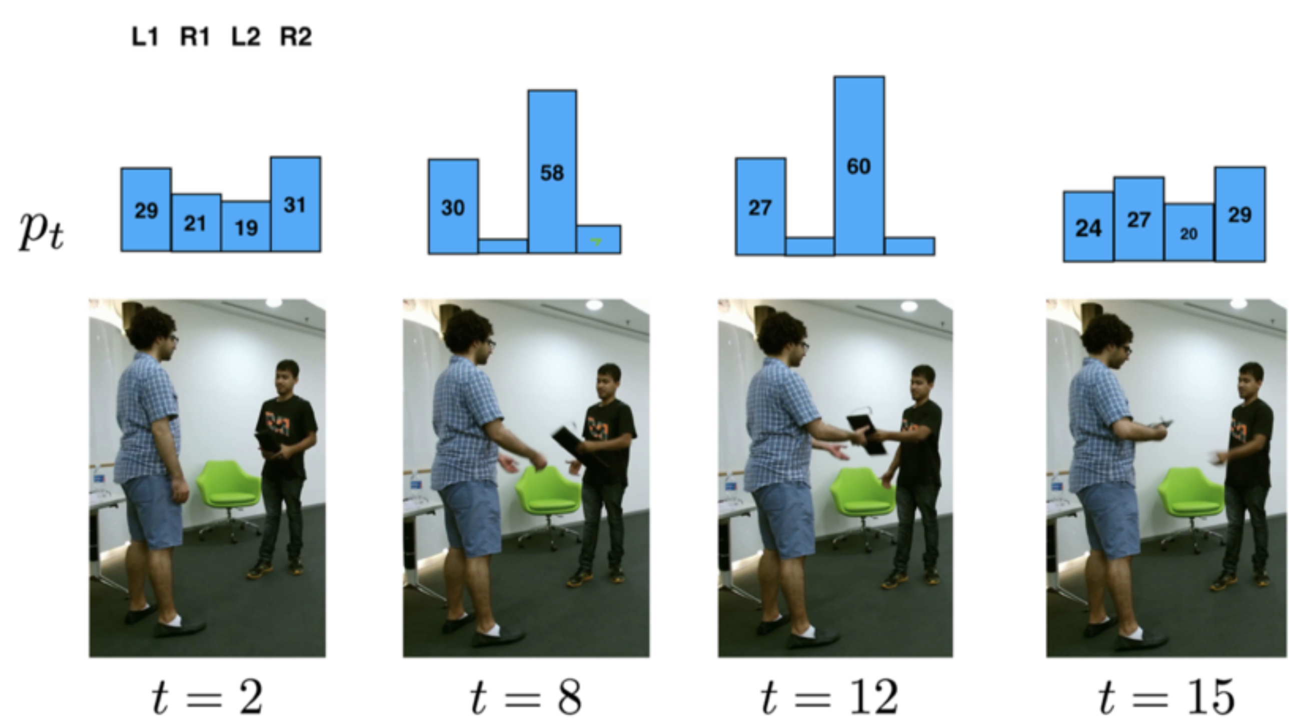
p' : Temporal Attention weights

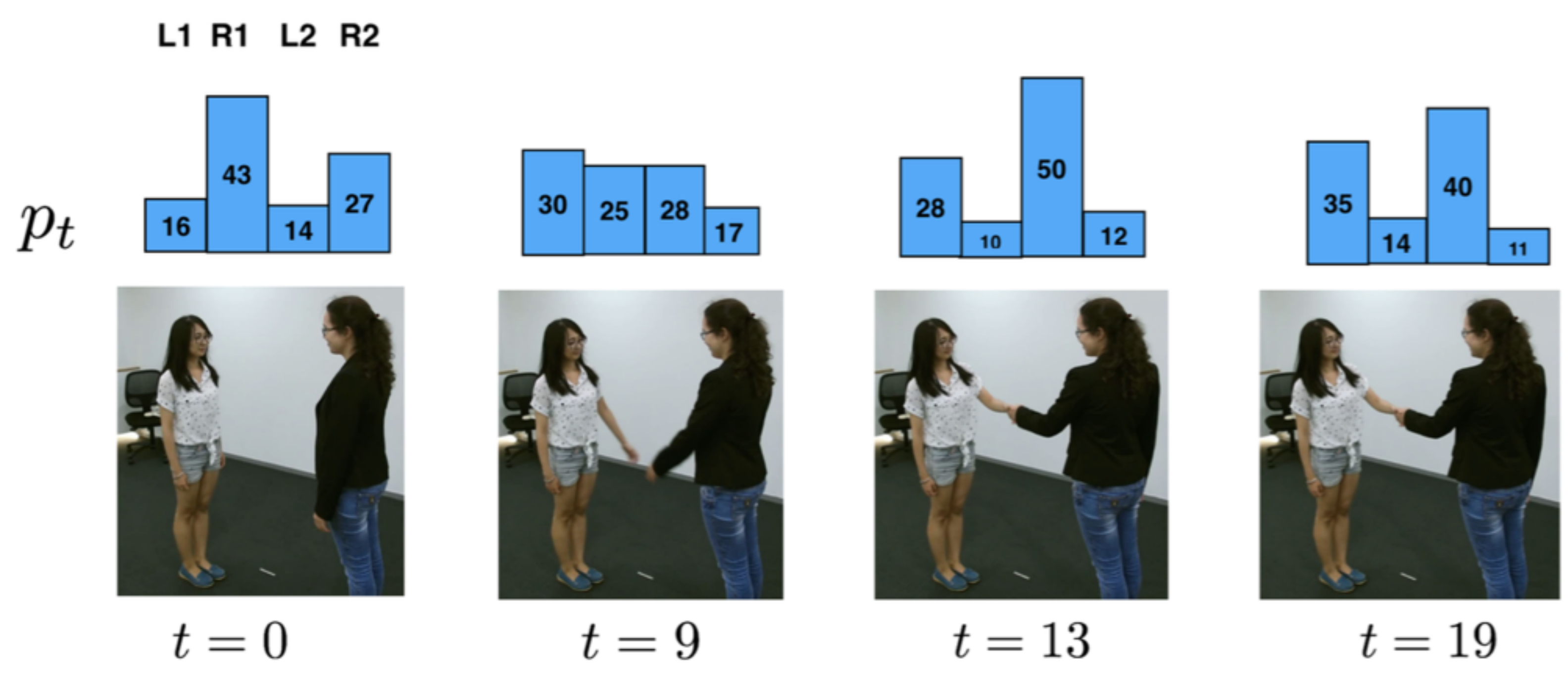
$\tilde{h}$: Final features vector

f_y : Classifier

M : Augmented Motion

VISUALIZATION OF THE ATTENTION PROCESS





EXPERIMENTAL RESULTS

Methods	Pose	RGB	CS	CV	Avg
Lie Group [35]	X	-	50.1	52.8	51.5
Skeleton Quads [9]	X	-	38.6	41.4	40.0
Dynamic Skeletons [13]	X	-	60.2	65.2	62.7
HBRNN [8]	X	-	59.1	64.0	61.6
Deep LSTM [30]	X	-	60.7	67.3	64.0
Part-aware LSTM [30]	X	-	62.9	70.3	66.6
ST-LSTM + TrustG. [24]	X	-	69.2	77.7	73.5
STA-LSTM [33]	X	-	73.2	81.2	77.2
GCA-LSTM [25]	X	-	74.4	82.8	78.6
JTM [36]	X	-	76.3	81.1	78.7
MTLN [17]	X	-	79.6	84.8	82.2
DSSCA - SSLM [31]	X	X	74.9	-	-
Deep GRU [A]	X	-	**68.0**	**74.2**	**71.1**
STA-Hands [B]	○	X	**73.5**	**80.2**	**76.9**
A+B	X	X	**82.5**	**88.6**	**85.6**

Table 1: Results on the NTU RGB+D dataset with Cross-Subject (CS) and Cross-View (CV) settings (accuracies in %, ○ means that pose is only used for the attention mechanism).

Comparison

- **State of the art on NTU RGB+D** (NTU) (~57'000 videos - 60 classes)
- First to combine **3D skeleton data and RGB frames** on NTU

Methods	Spatial Attention		Temporal Attention	CS	CV	Avg
	Hidden state	Augmented Pose	Augmented Pose			
Sum	-	-	-	68.3	74.6	71.5
Concat	-	-	-	68.9	75.2	72.0
SA-Hands	X	-	-	69.8	76.2	73.0
	-	X	-	**71.0**	**78.9**	**75.0**
	X	X	-	70.5	76.6	73.6
ST-Hands	-	-	X	**71.1**	**78.5**	**74.8**
STA-Hands	X	-	X	72.2	77.8	75.0
	-	X	X	**73.5**	**80.2**	**76.9**
	X	X	X	72.8	78.3	75.6

Table 2: Effects of the conditioning on the spatial attention and the temporal attention (RGB stream only, accuracies in %).

Ablation Study

- **Attention Conditioning: pose features > hidden state**
- **Attention mechanism has a high impact** on RGB only stream
  - ✓ Spatial Attention : + 3.5 points
  - ✓ Temporal Attention : + 3.2 points
  - ✓ Spatio-Temporal Attention : + 5.4 points
- Still a significant impact on the **two stream model**
  - ✓ Spatial Attention : + 1.6 points
  - ✓ Temporal Attention : + 1.4 points
  - ✓ Spatio-Temporal Attention : + 2.8 points

RGB stream methods	Spatial Attention		Temporal Attention	CS	CV	Avg
	Hidden state	Augmented Pose	Augmented Motion			
Sum-Hands	-	-	-	79.5	85.9	82.8
SA-Hands	X	-	-	80.5	86.8	83.7
	-	X	-	81.4	87.4	84.4
	X	X	-	81.0	86.9	84.0
ST-Hands	-	-	X	80.8	87.6	84.2
STA-Hands	X	-	X	81.4	87.4	84.4
	-	X	X	**82.5**	**88.6**	**85.6**
	X	X	X	81.6	88.0	84.8

Table 3: Effects of conditioning the spatio-temporal attention on different latent variables in the RGB stream for the two-stream model (accuracies in % on NTU). The pose stream is always the same: (*Deep GRU*) for every row.