WILLOW research group

LEAR research group

# P-CNN: Pose-based CNN Features for Action Recognition

Guilhem Chéron[1,2], Ivan Laptev[1], Cordelia Schmid[2]

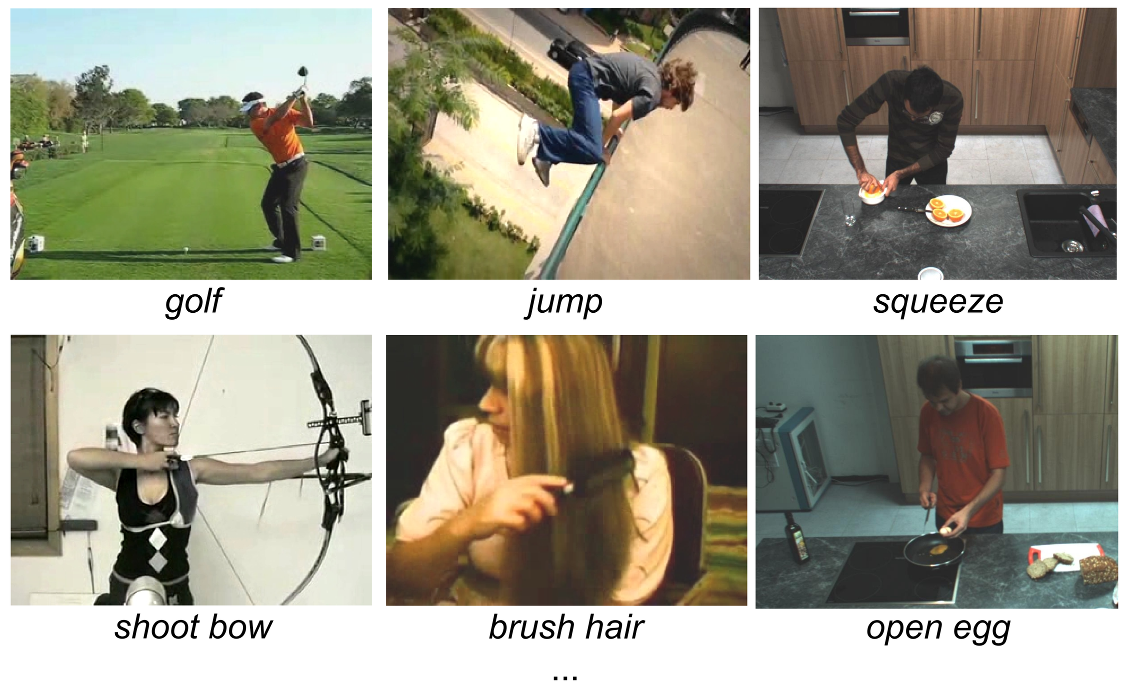
[1]INRIA, Paris, France [2]INRIA, Grenoble Rhône-Alpes, France

INVENTEURS DU MONDE NUMÉRIQUE

Microsoft Research - Inria JOINT CENTRE

## Goal

Recognize human actions in videos using body pose and convolutional neural networks (CNN).




## Motivation

- The structure and dynamics of body poses provide strong cues for action recognition.
- Action recognition has been dominated by local features especially dense trajectories (DT) [2].
- Current video representations based on local features [2] and CNNs [3] lack explicit structure.
- [1] reports significant gains provided by dynamic pose features (HLPF).
- [1] is sensitive to noise in pose estimation and presents results for one dataset only.

## Contribution

- Propose a new CNN-based action descriptor combining appearance and motion of body parts (P-CNN).
- Investigate alternative schemes for temporal aggregation of CNN features.
- P-CNN is complementary to DT [2], combination of P-CNN with DT improves state of the art results on two datasets.
- Our experiments confirm the importance of pose for action recognition.

## References

[1] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Toward understanding action recognition. In ICCV, 2013.

[2] H. Wang and C. Schmid. Action recognition with improved trajectories. In ICCV, 2013.

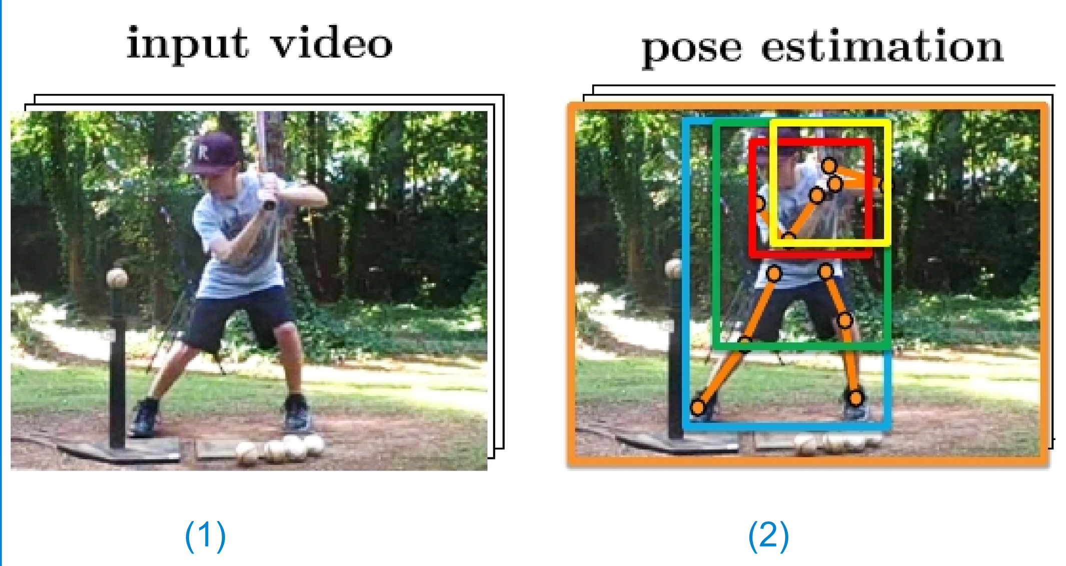
[3] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014.

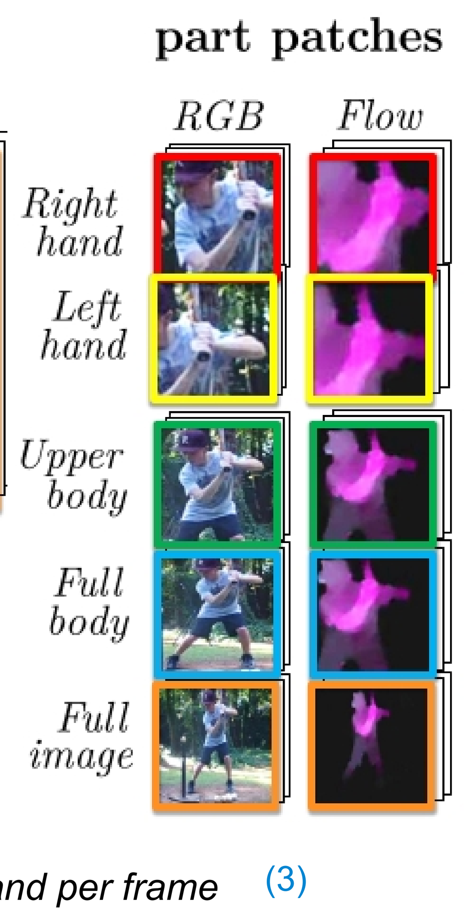
[4] A. Cherian, J. Mairal, K. Alahari, and C. Schmid. Mixing body-part sequences for human pose estimation. In CVPR, 2014.

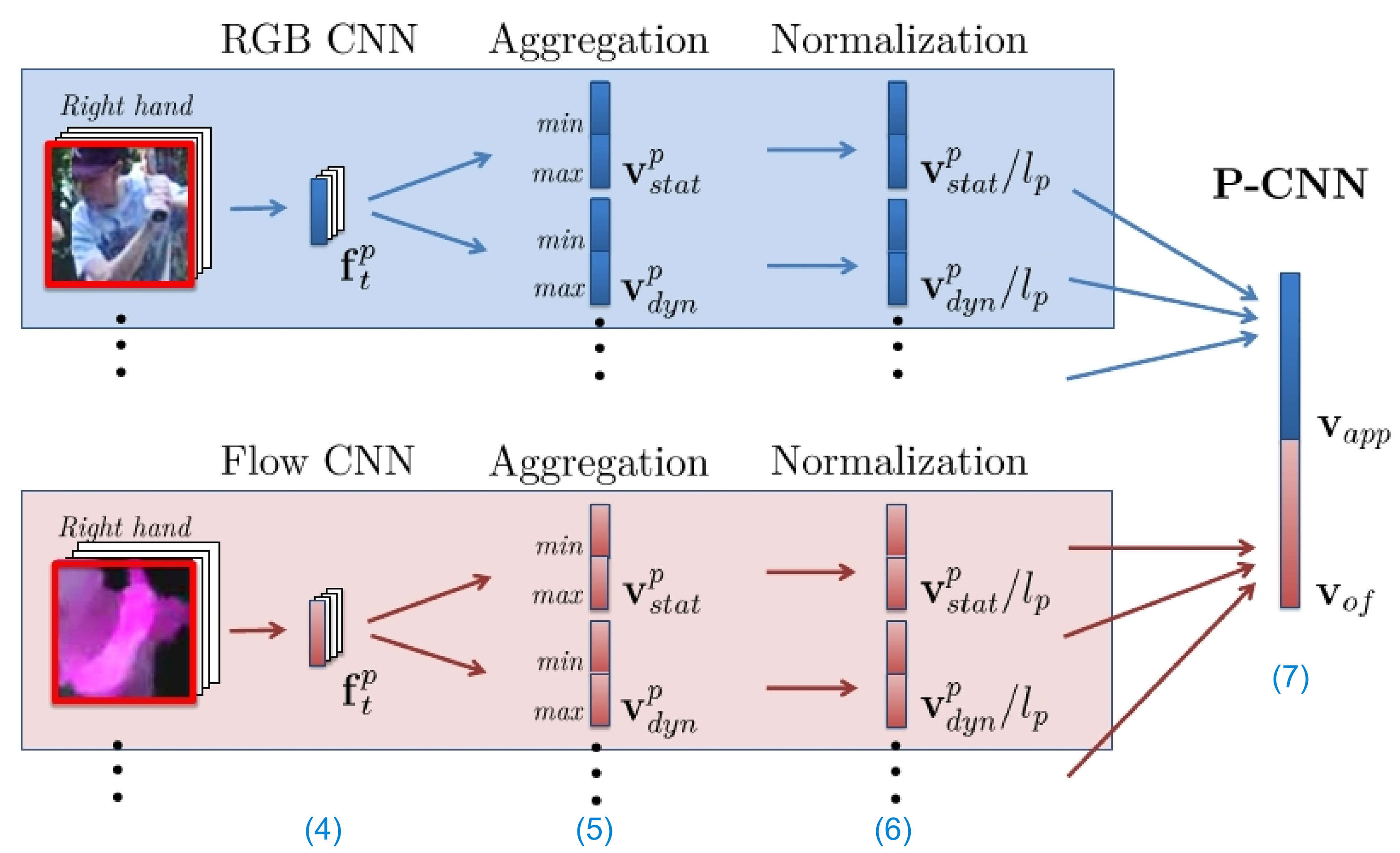
[5] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A Database for Fine Grained Activity Detection of Cooking Activities. In CVPR, 2012.

[6] Y. Zhou,B. Ni, S.Yan, P.Moulin, and Q.Tian. Pipelining localized semantic features for fine-grained action recognition. In ECCV, 2014.

## Approach




(1) *input video*
(2) *human pose estimation*
(3) *crop RGB and Optical Flow patches of body parts*
(4) *extract CNN features (appearance and motion) per part and per frame*
(5) *aggregate per-frame descriptors over time (max/min)*
(6) *normalize aggregated descriptors*
(7) *concatenate appearance and motion descriptors from all body parts*

*P-CNN code available at:* http://www.di.ens.fr/willow/research/p-cnn/

### Method details

- Compute temporal differences of CNN features $\mathbf{f}_t^p$:

$$\Delta\mathbf{f}_t^p = \mathbf{f}_{t+\Delta t}^p - \mathbf{f}_t^p \quad \text{with} \quad \Delta t = 4 \text{ frames.}$$

- Aggregation (max and min) of frame descriptors:

$$m_i = \min_{1\le t\le T} \mathbf{f}_t^p(i) \qquad \Delta m_i = \min_{1\le t\le T} \Delta\mathbf{f}_t^p(i)$$

$$M_i = \max_{1\le t\le T} \mathbf{f}_t^p(i) \qquad \Delta M_i = \max_{1\le t\le T} \Delta\mathbf{f}_t^p(i)$$

- Concatenation to get static and dynamic video descriptors:

$$\mathbf{v}_{stat}^p = [m_1, ..., m_k, M_1, ..., M_k]^\top$$

$$\mathbf{v}_{dyn}^p = [\Delta m_1, ..., \Delta m_k, \Delta M_1, ..., \Delta M_k]^\top$$

- Normalization of video descriptor: normalize by the average L2-norm of the $\mathbf{f}_t^p$s from the training set ($l_p$)

## Results

**Datasets:** JHMDB [1]: 21 sport oriented human actions. MPII Cooking [5]: 64 fine-grained cooking actions.

**Human poses:** Pose: automatic pose estimation using [4] / GT: manually annotated (ground truth) pose.

### Effect of body parts

| Parts | JHMDB-GT App | JHMDB-GT OF | JHMDB-GT App + OF | MPII Cooking-Pose App | MPII Cooking-Pose OF | MPII Cooking-Pose App + OF |
|---|---|---|---|---|---|---|
| Hands | 46.3 | 54.9 | 57.9 | 39.9 | 46.9 | 51.9 |
| Upper body | 52.8 | 60.9 | 67.1 | 32.3 | 47.6 | 50.1 |
| Full body | 52.2 | 61.6 | 66.1 | - | - | - |
| Full image | 43.3 | 55.7 | 61.0 | 28.8 | 56.2 | 56.5 |
| All | **60.4** | **69.1** | **73.4** | **43.6** | **57.4** | **60.8** |

*human parts CNN features appearance/flow, max-aggregation.*

- Combination of parts improves action classification.
- Appearance and flow descriptors are complementary.

### Effect of aggregation

| Aggregation scheme | App | OF | App+OF |
|---|---|---|---|
| All (Stat, Max-aggr) | 60.4 | 69.1 | 73.4 |
| All (Stat, Max/Min-aggr) | 60.6 | 68.9 | 73.1 |
| All (Stat+Dyn, Max-aggr) | 62.4 | **70.6** | 74.1 |
| All (Stat+Dyn, Max/Min-aggr) | **62.5** | 70.2 | **74.6** |

- Max and Min aggregations combined with static and dynamic features improve action classification.

### Automatic vs. GT pose

JHMDB

| | sub-JHMDB GT | sub-JHMDB Pose | sub-JHMDB Diff | JHMDB GT | JHMDB Pose | JHMDB Diff |
|---|---|---|---|---|---|---|
| P-CNN | 72.5 | 66.8 | -5.7 | 74.6 | 61.1 | -13.5 |
| HLPF [1] | 78.2 | 51.1 | -27.1 | 77.8 | 25.3 | -52.5 |

- P-CNN are significantly more robust to errors in the automatic Pose.
- HLPF [1] outperforms P-CNN in the case of GT pose.

MPII Cooking

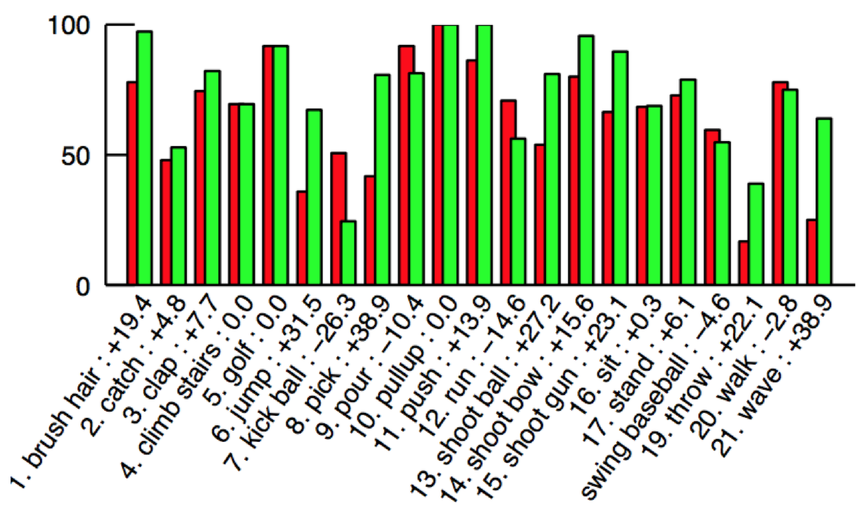
| | sub-MPII Cooking GT | sub-MPII Cooking Pose | sub-MPII Cooking Diff | MPII Cooking Pose |
|---|---|---|---|---|
| P-CNN | 83.6 | 67.5 | -16.1 | 62.3 |
| HLPF [1] | 76.2 | 57.4 | -18.8 | 32.6 |

- P-CNN significantly outperforms HLPF [1] for automatic Pose and GT pose.

### Comparison to other methods

| Method | JHMDB GT | JHMDB Pose | MPII Cook. Pose |
|---|---|---|---|
| Holistic + Pose[5] | - | - | 57.9 |
| Semantic Features[6] | - | - | 70.5 |
| P-CNN | 74.6 | 61.1 | 62.3 |
| DT-FV | 65.9 | 65.9 | 67.6 |
| P-CNN + DT-FV (our) | **79.5** | **72.2** | **71.4** |

- P-CNN outperforms state-of-the-art DT-FV with manually annotated human pose (GT).
- With GT and Pose: P-CNN and DT-FV are complementary and improve state-of-the-art results on JHMDB and MPII Cooking.




*Per class accuracy on JHMDB-GT: P-CNN (green) and DT-FV [2] (red).*

### Qualitative results

r: # : *P-CNN ranking* r: # : *DT+FV[2] ranking*

MPII Cooking *Improved ranking* *worsen ranking*

JHMDB *Improved ranking* *worsen ranking*