Deep Moving Poselets For Video Based Action Recognition

Vision Lab @ JHU
http://www.vision.jhu.edu

Effrosyni Mavroudi Lingling Tao René Vidal

Center for Imaging Science, Johns Hopkins University, Baltimore, USA

JOHNS HOPKINS BIOMEDICAL ENGINEERING

Motivation

- Actions consist of spatio-temporal configurations of body parts.
- There has been a huge success of using discriminative, interpretable body part configurations for skeleton-based action recognition [2].
  - 3D body parts are described in terms of the positions/velocities of their joints.
- Finding body part configurations in a video is challenging because:
  - Joint positions are not measured → they have to be annotated or estimated.
  - Local features do not capture the appearance and movement of body parts.

Contributions

- Propose a video representation based on shared and discriminative mid-level classifiers (deep moving poselets) that capture characteristic spatio-temporal configurations of body parts during different phases of an action.
  - We describe a video of an action with an "activation vector", which captures the degree to which each configuration is present in the video.
  - Activation vectors provide a distributed representation of pose, movement, appearance and context.
- Propose a method for learning the deep moving poselets representation.
  - Extract deep features from short tubelets around a hierarchy of body parts.
  - Max-margin approach to learn both deep moving poselets and action classifiers.

Deep Feature Extraction from Short Tubelets

Fig. 1: Bounding boxes around body parts are inferred based on joint locations.

- Find 2D bounding box containing all joints defining a body part.
- A *tubelet* is a temporal sequence of L bounding boxes containing the joint trajectories of a bodypart.
- A tubelet is represented with a vector of max-pooled deep features.

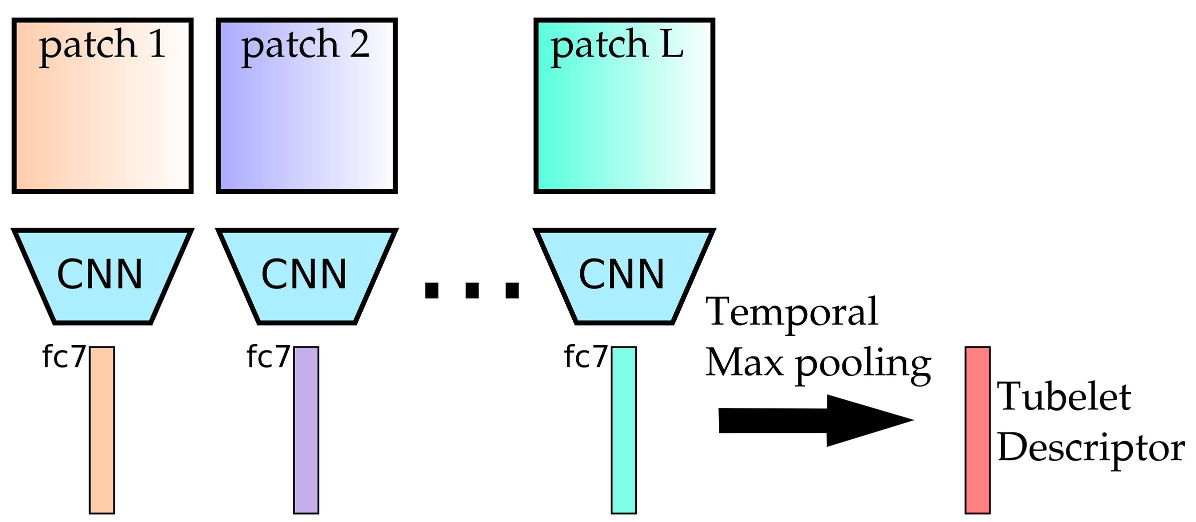
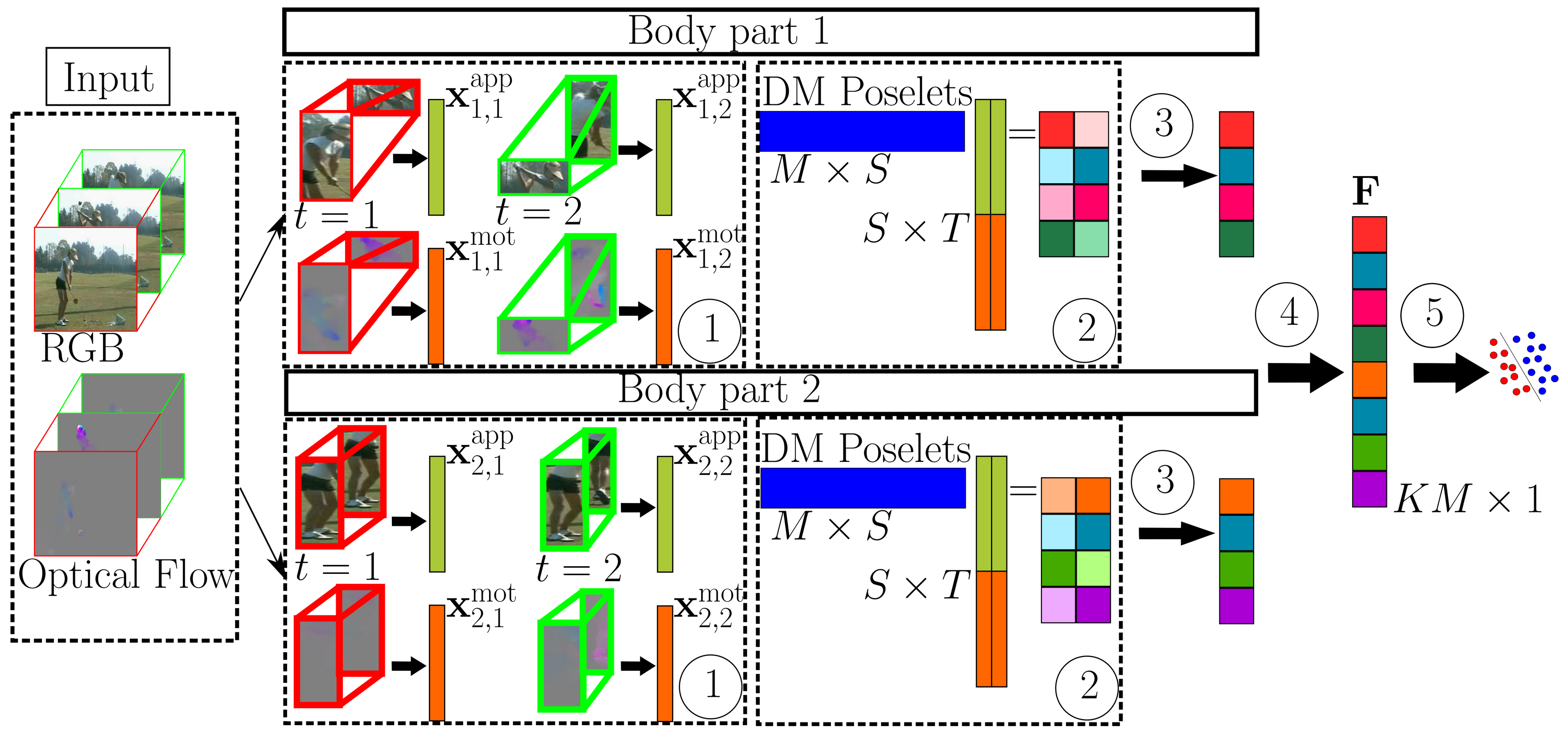


Fig. 2: Deep features [1] are extracted from each bounding box of the tubelet and are temporally max-pooled to obtain a tubelet descriptor.

Learning Deep Moving Poselet Representation



$$\min_{\mathbf{D},\mathbf{d},\mathbf{W},\mathbf{b}} \sum_{c=1}^{C}\sum_{n=1}^{N} \max(0, 1 - \underbrace{Y_{cn}}_{\text{label}} \overbrace{(\mathbf{W}_c^\top \underbrace{\mathrm{ReLU}(\mathbf{F}(\mathbf{X}^{(n)},\mathbf{D},\mathbf{d}))}_{\text{activation vector}} + b_c)}^{\text{action classifier}}) + \lambda \underbrace{(R_W(\mathbf{W}) + R_D(\mathbf{D}))}_{\text{regularizers}}$$

Quantitative Results

Method	Features	Accuracy (%)	
		GT Joints	PE Joints
DT [3]	RGB	46.0	
NTraj [3]	2D Pose	75.1	54.1
DT + NTraj [3]	RGB + 2D Pose	75.5	52.9
MST-AOG [6] [5]	RGB + 2D Pose	-	45.3
AOG [5]	RGB + 2D Pose	-	61.2
[4]	RGB + 3D Pose	77.5	-
P-CNN [1]	RGB + Bps	72.5	66.8
Ours	RGB + Bps	**79.2**	**70.2**

Table 1: sub-JHMDB (GT: human annotated joints, PE: pose-estimated joints, Bps: body parts)

Method	Acc (%)
RGB	
STIP [6]	54.5
DT [7]	71.7
MST-AOG [6]	73.1
IPM [7]	**83.3**
RGB + Pose	
IPM+Joints [7]	**89.3**
RGB + Body parts	
Ours	**84.4**

Table 2: MSR Daily Activity 3D

Ablation Analysis

Method	Accuracy (%)
app, full body, no sliding window	60.3
mot, full body, no sliding window	66.1
app+mot, full body, no sliding window	74.3
app+mot, all bps, no sliding window	77.7
app+mot, all bps, with sliding window	79.2

Method	Accuracy (%)
P-CNN + SVM [1]	72.5
P-CNN + DMPs (no sliding window)	74.3
P-CNN + DMPs (with sliding window)	76.9

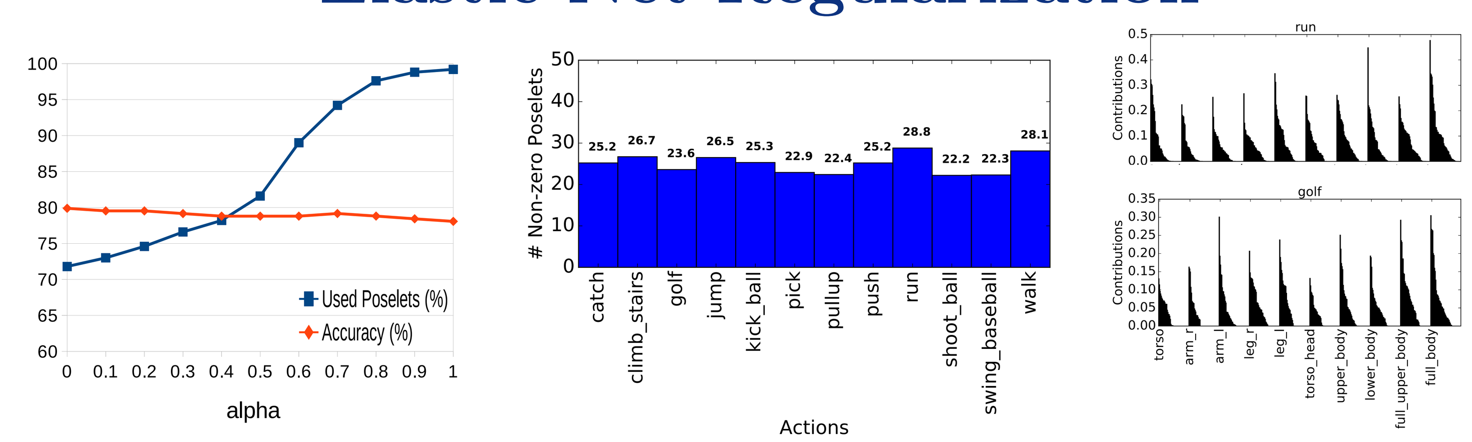
- Experiments on sub-JHMDB with annotated joints.
- ✓ Appearance and motion streams are complementary.
- ✓ Hierarchical body part structure improves over full body.
- ✓ Extracting short tubelets further improves performance.
- ✓ Adding a mid-level representation improves over P-CNN + SVM.

Qualitative Results

Fig. 3: From left to right: each column shows the 5 most significant deep moving poselets for action classes *catch*, *swing baseball*, *pullup*.

Fig. 4: Examples of deep moving poselets shared among action classes in the sub-JHMDB dataset (split 2). Each row shows 3 tubelets from different classes with high activations for a specific poselet.

Elastic Net Regularization



(a) (b) (c)

References

[1] G. Chéron et al., P-CNN: Pose-Based CNN Features For Action Recognition. ICCV'15.
[2] L. Tao et al., Moving poselets: A discriminative and interpretable skeletal motion representation for action recognition. ICCVW'15.
[3] H. Jhuang et al., Towards Understanding Action Recognition ICCV'13.
[4] I. Lillo et al., A Hierarchical Pose-Based Approach To Complex Action Understanding Using Dictionaries of Actionlets and Motion Poselets. CVPR'16.
[5] B. X. Nie et al., Joint action recognition and pose estimation from video. CVPR'15.
[6] J. Wang et al., Cross-View Action Modeling, Learning and Recognition. CVPR'14.
[7] Y. Zhou et al., Interaction part mining: A mid-level approach for fine-grained action recognition. CVPR'15.

This work was supported by NIH grant R01HD87133-01.