# Video Action Detection with Relational Dynamic-Poselets

Limin Wang[1,2], Yu Qiao[2], and Xiaoou Tang[1,2]

[1]Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong
[2]Shenzhen Institutes of Advanced Technology, CAS, China

## Introduction

- **Problem**: We aim to not only recognize on-going action class (action recognition), but also localize its spatiotemporal extent (action detection), and even estimate the pose of the actor (pose estimation).
- **Key insights**:

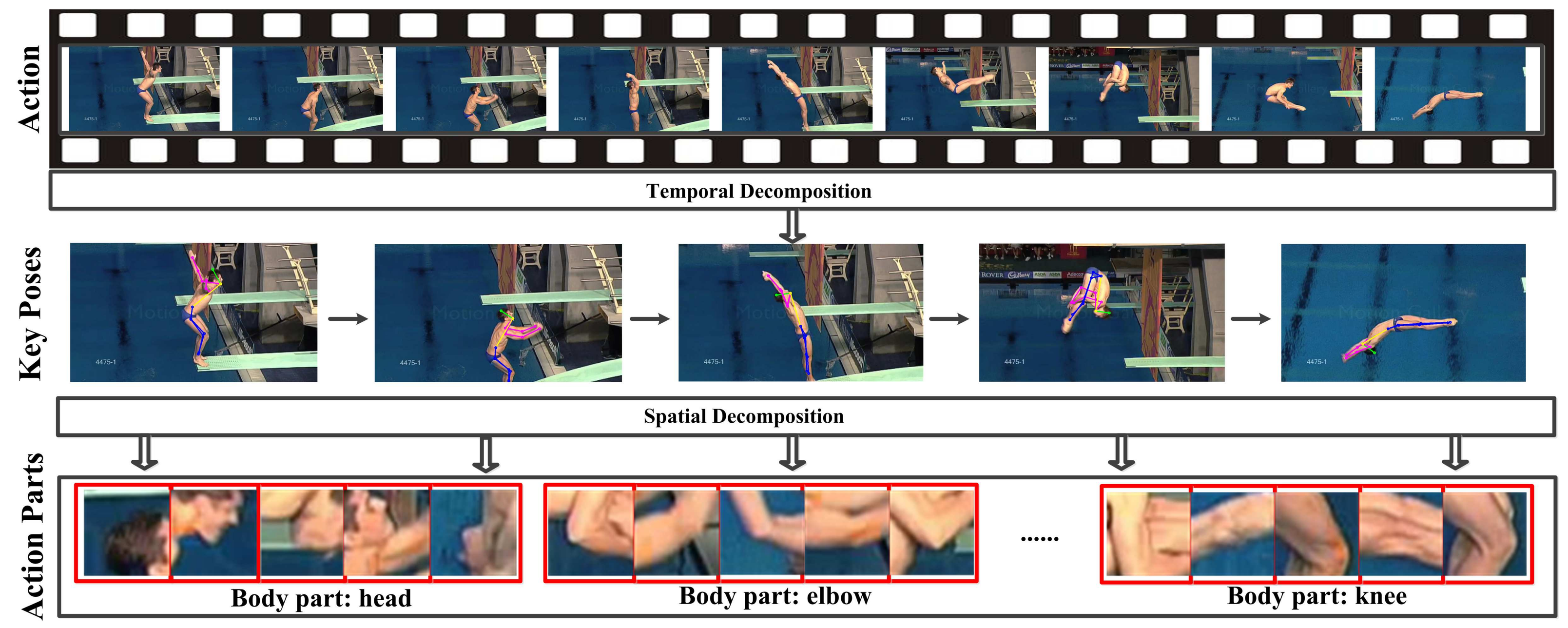



Figure 1: Illustration for motivation.

- An action can be temporally decomposed into a sequence of key poses.
- Each key pose can be decomposed into a spatial arrangement of mixtures of action parts.

- **Main contributions**:
  - We propose to a new pose and motion descriptor to cluster cuboids into dynamic-poselets.
  - We design a sequential skeleton model to jointly capture spatiotemporal relations among body parts, co-occurrences of mixture types, and local part templates.

## Dynamic-Poselets

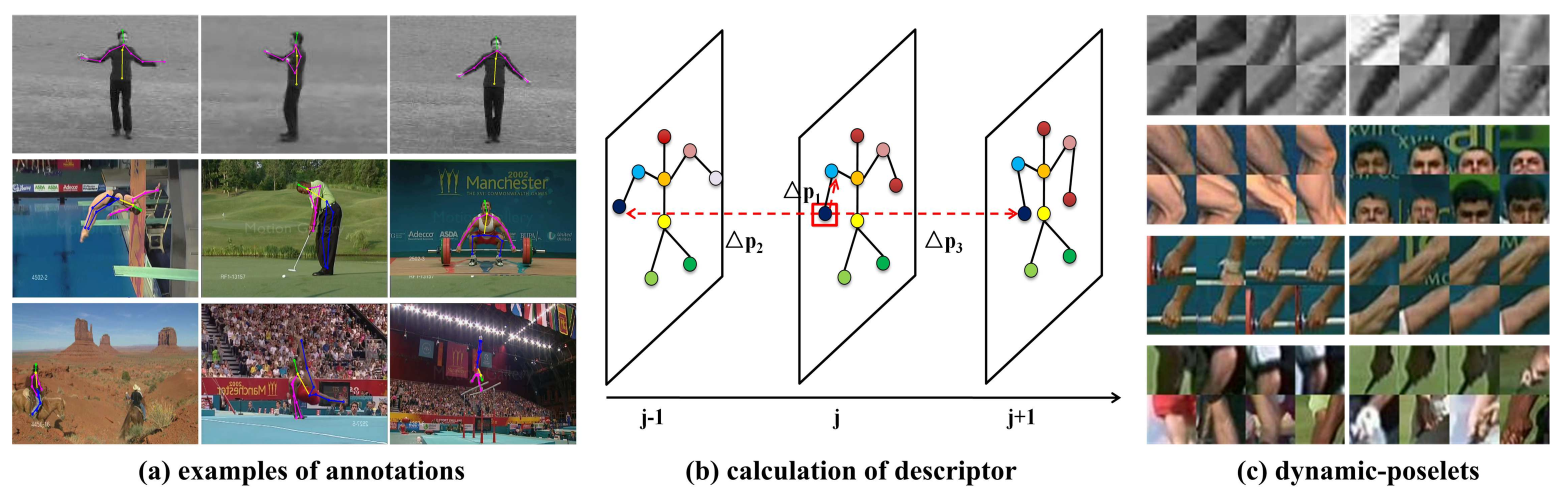



Figure 2: Construction of dynamic-poselets.

- **A pose and motion descriptor:**
  - $f(p^v_{i,j}) = [\Delta p^{v,1}_{i,j}, \Delta p^{v,2}_{i,j}, \Delta p^{v,3}_{i,j}]$.
  - $\Delta p^{v,1}_{i,j} = p^v_{i,j} - p^v_{par(i),j}$, $\Delta p^{v,2}_{i,j} = p^v_{i,j} - p^v_{i,j-1}$, $\Delta p^{v,3}_{i,j} = p^v_{i,j} - p^v_{i,j+1}$.
  - $\overline{f(p^v_{i,j})} = [\overline{\Delta p^{v,1}_{i,j}}, \overline{\Delta p^{v,2}_{i,j}}, \overline{\Delta p^{v,3}_{i,j}}]$.
  - $\overline{\Delta p^{v,k}_{i,j}} = [\Delta x^{v,k}_{i,j}/s^v_{i,j}, \Delta y^{v,k}_{i,j}/s^v_{i,j}]$ $(k = 1, 2, 3)$.
- **Using this descriptor, we run *k*-means algorithm to cluster cuboids into dynamic-poselets.**

## Action Detection with SSM

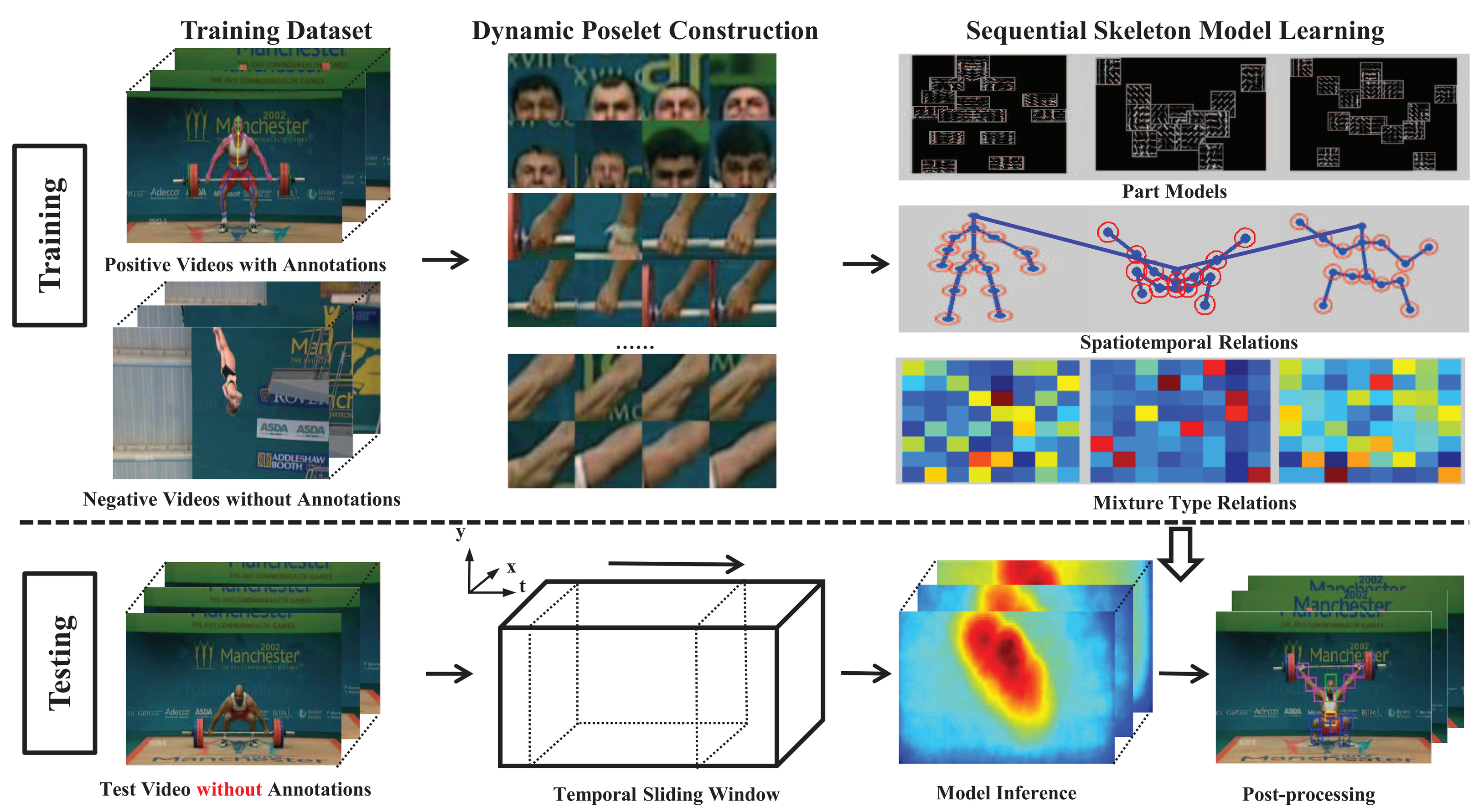



Figure 3: Overview of our approach.

- **Sequential Skeleton Model (SSM):**

$$S(v,p,t) = \underbrace{b(t)}_{\text{Mixture Type Relations}} + \underbrace{\Psi(p,t)}_{\text{Spatiotemporal Relations}} + \underbrace{\Phi(v,p,t)}_{\text{Action Part Models}}$$

$v$ is a video clip, $p$ and $t$ are the pixel positions and the mixture types of dynamic-poselets, respectively .

- Mixture Type Relations:

$$b(t) = \sum_{j=1}^{N}\sum_{i=1}^{K} b^{t_{i,j}}_{i,j} + \sum_{(i,j)\sim(m,n)} b^{t_{i,j}t_{m,n}}_{(i,j),(m,n)}$$

$b^{t_{i,j}}_{i,j}$ encodes the mixture prior, $b^{t_{i,j}t_{m,n}}_{(i,j),(m,n)}$ captures the compatibility of mixture types.

- Spatiotemporal Relations:

$$\Psi(p,t) = \sum_{(i,j)\sim(m,n)} \beta^{t_{i,j}t_{m,n}}_{(i,j),(m,n)} \psi(p_{i,j}, p_{m,n}),$$

$$\psi(p_{i,j}, p_{m,n}) = [dx, dy, dz, dx^2, dy^2, dz^2]$$

$\beta^{t_{i,j}t_{m,n}}_{(i,j),(m,n)}$ represents the parameter of quadratic spring model.

- Action Part Models:

$$\Phi(v,p,t) = \sum_{j=1}^{N}\sum_{i=1}^{K} \alpha^{t_{i,j}}_{i} \phi(v, p_{i,j})$$

$\phi(v, p_{i,j})$ is the feature vector, $\alpha^{t_{i,j}}_{i}$ denotes the feature template.
Note that the body part template $\alpha^{t_{i,j}}_{i}$ is shared among different key poses.

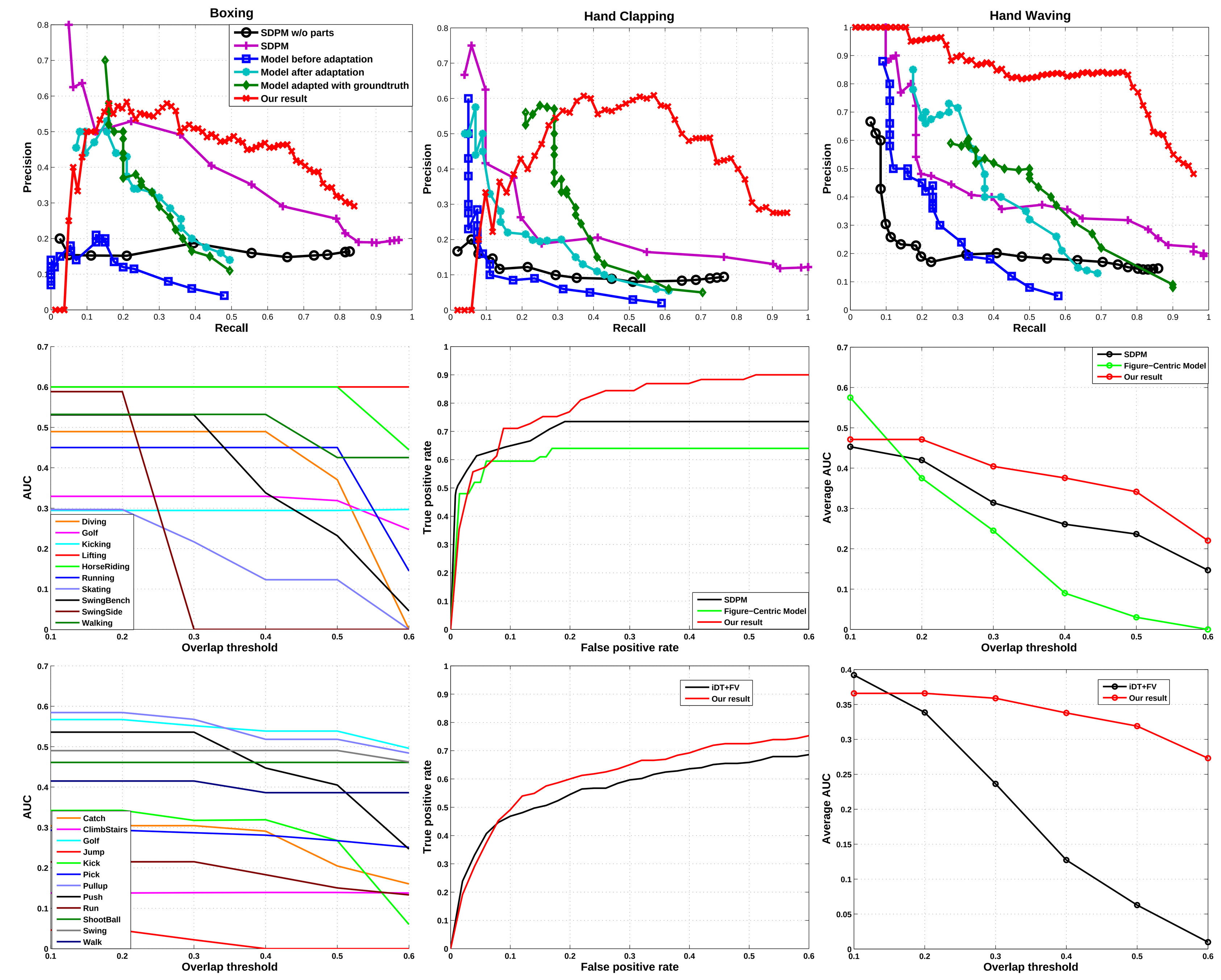
- **Action detection pipeline:**
  - temporal sliding window → model inference → non-maximum suppression.

## References

1. Cao, L., Liu, Z., Huang, T.S.: Cross-dataset action detection. In: CVPR (2010).
2. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR (2011)
3. Lan, T., etc.: Discriminative figure-centric models for joint action localization and recognition. In: ICCV (2011).
4. Tian, Y., Sukthankar, R., Shah, M.: Spatiotemporal deformable part models for action detection. In: CVPR (2013).
5. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV (2013).

## Experiments

- **Quantitive results on MSR-II, UCF Sports, J-HMDB:**




- **Detection examples on MSR-II, UCF Sports, J-HMDB:**