JHU vision lab

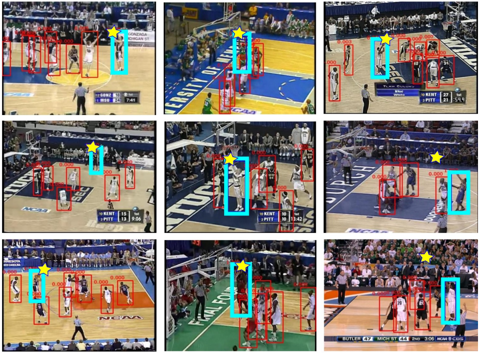
Deep Moving Poselets for Video Based Action Recognition

Effrosyni Mavroudi Lingling Tao René Vidal
Center for Imaging Science, Johns Hopkins University, Baltimore, USA

Why Is Action Classification Important?

Action recognition applications:

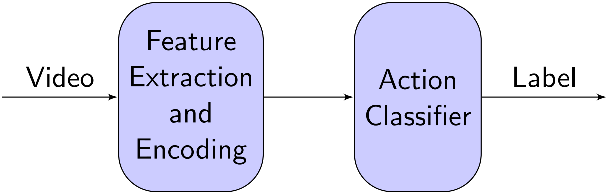
- Human-robot interaction
- Surveillance
- Patient monitoring
- Sports video analysis
- Web video search and retrieval

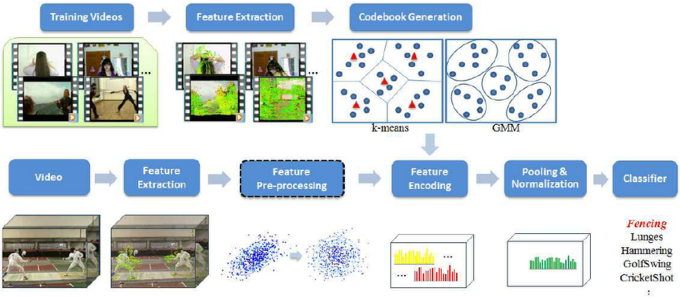


3-pointer success　slam dunk success　free-throw failure

[Ramanathan15]

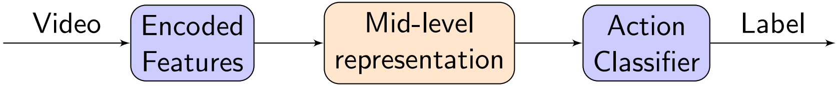
Prior Work: Video Features And SVM Classifier





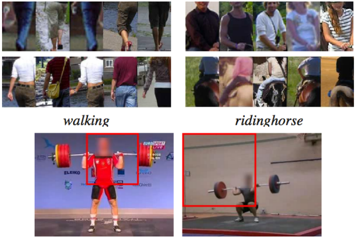
Source: [Peng14]

Previous Work: Midlevel Representations



Previous Work: Midlevel Representations

Video → Encoded Features → Mid-level representation → Action Classifier → Label

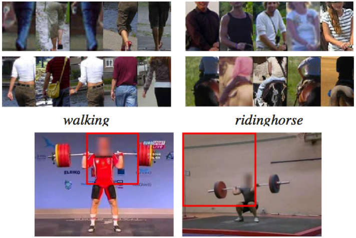


Poselets

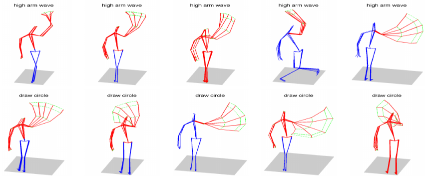
- ✓ discriminative dictionary
- ✗ separate dictionary per action
- ✗ feat from 2D patches/cuboids

Previous Work: Midlevel Representations

Video → **Encoded Features** → **Mid-level representation** → **Action Classifier** → Label



Poselets



Moving Poselets

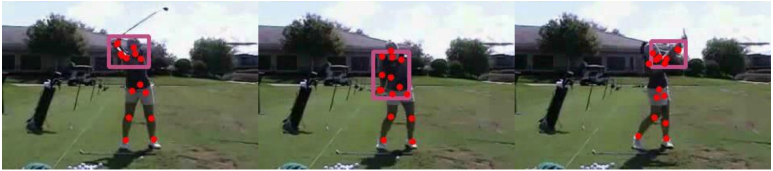
- ✓ discriminative dictionary
- ✗ separate dictionary per action
- ✗ feat from 2D patches/cuboids

- ✓ discriminative dictionary
- ✓ poselets shared between actions
- ✗ feat from 3D locations/velocities

Our Work: Deep Moving Poselets

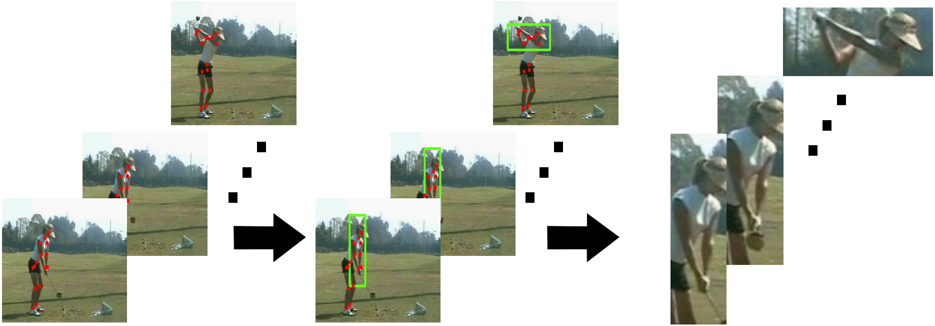
- A new mid-level representation for action recognition
  - ✓ deep moving poselet = appearance of body part in motion
  - ✓ discriminative, shared and interpretable mid-level representation
  - ✓ features from tubelets along a hierarchy of body parts

- A new end-to-end learning method
  - ✓ joint max-margin learning of moving poselets and action classifiers
  - ✓ elastic-net regularization encourages sharing and discriminability



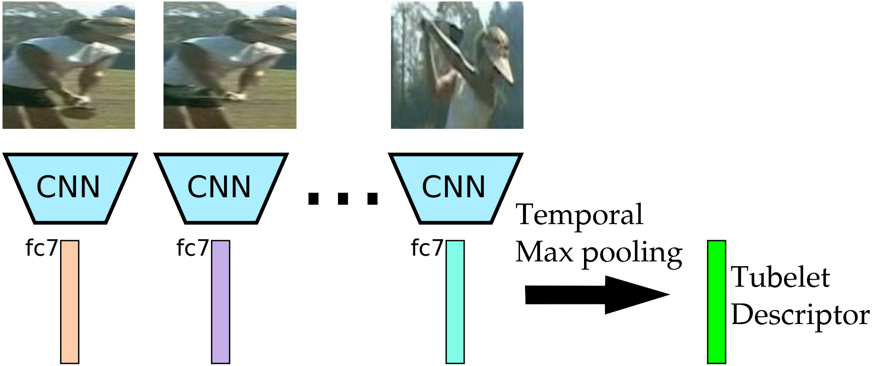
Deep Moving Poselets Representation

Extracting tubelet around the upper body from the first 15 frames.



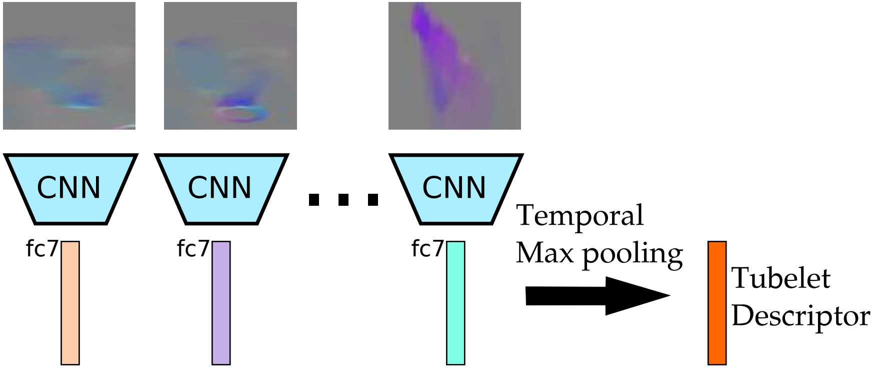
Deep Moving Poselets Representation

Appearance feature extracted from tubelet.



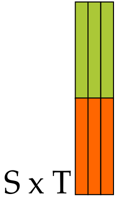
Deep Moving Poselets Representation

Motion feature extracted from tubelet.



Deep Moving Poselets Representation

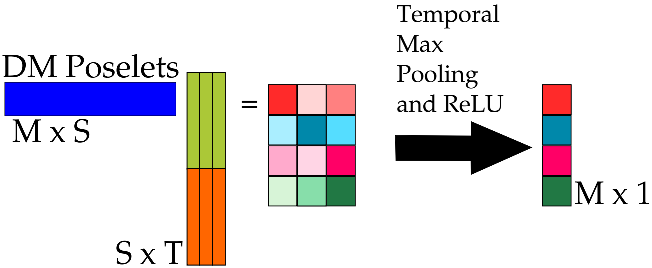
Appearance and motion features are extracted from all temporal windows.



$\mathbf{X}^{(n)}$

Deep Moving Poselets Representation

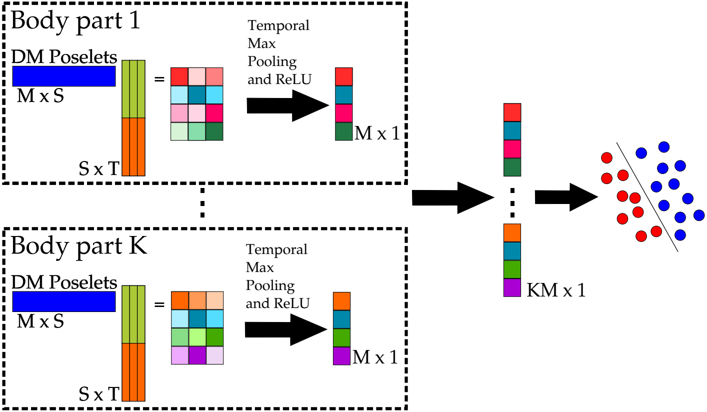
Compute response map and activation vector associated with body part.



$$\underbrace{\mathrm{ReLU}(\mathbf{F}(\mathbf{X}^{(n)}, \mathbf{D}, \mathbf{d}))}_{\text{activation vector}}$$

Deep Moving Poselets Representation

Activation vector from all body parts is fed to action classifiers.



$$\overbrace{\mathbf{W}_c{}^{\top}\,\underbrace{\mathrm{ReLU}(\mathbf{F}(\mathbf{X}^{(n)},\mathbf{D},\mathbf{d}))}_{\text{activation vector}} + b_c}^{\text{action classifier}}))$$

Learning Deep Moving Poselets

- Max-margin formulation for joint learning of action classifiers and moving poselets

$$\min_{\substack{\mathbf{D},\mathbf{d},\\ \mathbf{W},\mathbf{b}}} \sum_{c=1}^{C}\sum_{n=1}^{N} \underbrace{\max(0, 1 - \underbrace{Y_{cn}}_{\text{label}}(\mathbf{W}_c^\top \mathrm{ReLU}(\mathbf{F}(\mathbf{X}^{(n)}, \mathbf{D}, \mathbf{d})) + b_c))}_{\text{hinge loss}} + \lambda \underbrace{(R_W(\mathbf{W}) + R_D(\mathbf{D}))}_{\text{regularization}} \tag{1}$$

- Choices for regularization on $\mathbf{W}$: ℓ_2 or elastic net.

- Joint training by Stochastic Gradient Descent.

Results On sub-JHMDB

- 12 action classes
  - e.g., *catch*, *golf*, *run* and *walk*
- Realistic videos
  - clips from movies, YouTube etc.
- Visible full body
- Joint annotations available
  - human annotated joints (GT)
  - pose estimated (PE)

Results On sub-JHMDB

- State-of-the-art for both GT and PE joints.

Method	Features	Accuracy (%)	
		GT Joints	PE Joints
DT[Jhuang13]	RGB	46.0	46.0
NTraj[Jhuang13]	2D Pose	75.1	54.1
DT + NTraj[Jhuang13]	RGB + 2D Pose	75.5	52.9
MST-AOG[Wang14]	RGB + 2D Pose	-	45.3
AOG[Nie15]	RGB + 2D Pose	-	61.2
Lillo[Lillo16]	RGB + 3D Pose	77.5	-
P-CNN[Cheron15]	RGB + Bps	72.5	66.8
Ours	RGB + Bps	**79.2**	**70.2**

Results On sub-JHMDB

- Body parts + mid-level representation outperform 2D pose features.

Method	Features	Accuracy (%)	
		GT Joints	PE Joints
DT[Jhuang13]	RGB	46.0	46.0
NTraj[Jhuang13]	2D Pose	75.1	54.1
DT + NTraj[Jhuang13]	RGB + 2D Pose	75.5	52.9
MST-AOG[Wang14]	RGB + 2D Pose	-	45.3
AOG[Nie15]	RGB + 2D Pose	-	61.2
Lillo[Lillo16]	RGB + 3D Pose	77.5	-
P-CNN[Cheron15]	RGB + Bps	72.5	66.8
Ours	RGB + Bps	**79.2**	**70.2**

Results On sub-JHMDB

- Body-part based methods are less sensitive to pose estimation errors.

Method	Features	Accuracy (%)		
		GT Joints	PE Joints	
DT[Jhuang13]	RGB	46.0	46.0	
NTraj[Jhuang13]	2D Pose	75.1	54.1	
DT + NTraj[Jhuang13]	RGB + 2D Pose	75.5	52.9	-22%
MST-AOG[Wang14]	RGB + 2D Pose	-	45.3	
AOG[Nie15]	RGB + 2D Pose	-	61.2	
Lillo[Lillo16]	RGB + 3D Pose	77.5	-	
P-CNN[Cheron15]	RGB + Bps	72.5	66.8	
Ours	RGB + Bps	**79.2**	**70.2**	-9%

More experiments and qualitative results in paper/poster.

Discriminative Poselets Visualization

catch

swing baseball

Conclusions

- A new mid-level representation for action recognition
  - ✓ deep moving poselet = appearance of body part in motion
  - ✓ discriminative, shared and interpretable mid-level representation
  - ✓ features from tubelets along a hierarchy of body parts
- A new end-to-end learning method
  - ✓ joint max-margin learning of moving poselets and action classifiers
  - ✓ group-sparse regularization encourages sharing and discriminability